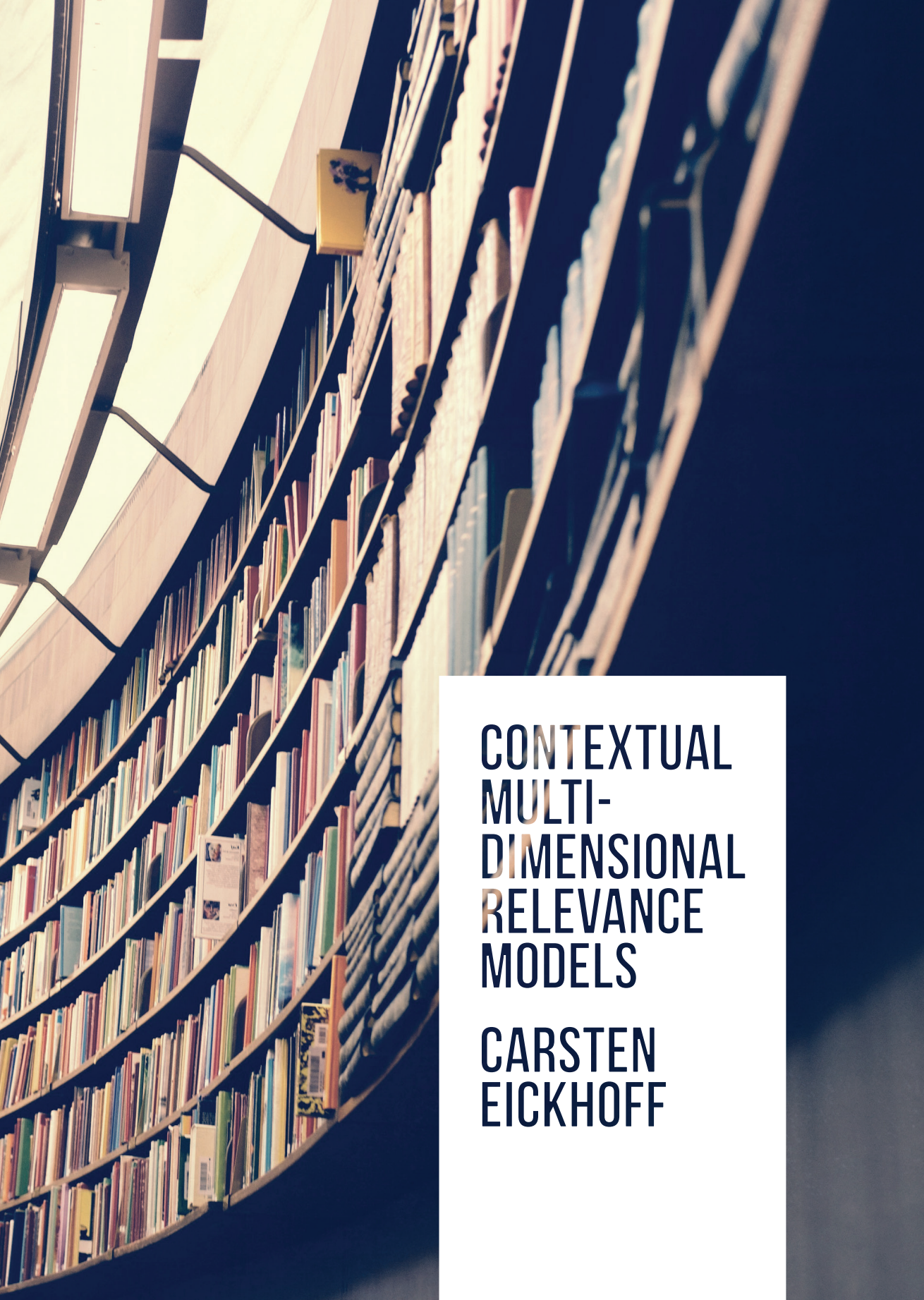

CONTEXTUAL MULTI-DIMENSIONAL RELEVANCE MODELS

CARSTEN EICKHOFF

Contextual Multidimensional Relevance Models

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K. C. A. M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op dinsdag 14 oktober 2014 om 10:00 uur

door

Carsten EICKHOFF

informaticus
geboren te Twistringen, Duitsland.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. A. P. de Vries

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Prof. dr. ir. A. P. de Vries,	Technische Universiteit Delft, CWI, promotor
Prof. dr. K. Collins-Thompson,	University of Michigan
Prof. dr. ir. W. Kraaij	Radboud Universiteit Nijmegen, TNO
Dr. M. Lalmas	Yahoo Labs London
Prof. dr. F. M. G. de Jong	Universiteit Twente, Erasmus Universiteit Rotterdam
Prof. dr. C. M. Jonker	Technische Universiteit Delft
Dr. L. Aroyo	Vrije Universiteit Amsterdam
Prof. dr. A. Hanjalic,	Technische Universiteit Delft, reservelid

SIKS Dissertation Series No. 2014-42
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Keywords: Information Retrieval, Relevance Modelling

Printed by: PrintEnBind.nl

Front & Back: Esther Smit

ISBN 978-0-692-30796-0

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To Esther

Contents

I

Motivation

1

Introduction

Space is big.
You just won't believe how vastly, hugely, mind-bogglingly big it is.
I mean, you may think it's a long way down the road to the chemist's,
but that's just peanuts to space.

-Douglas Adams, Author

Finding a small number of desired documents[1] in a large collection has always been a challenging task. Some even like to compare it to the metaphorical search for needles in a haystack. In my high school days, the local library held 14,000 volumes and the main way of finding anything was asking the librarian, who had her ways of divining the whereabouts of a desired book. As we turn to larger collections, such as the university library in Delft with more than 800,000 titles, or the British Library that boasts a catalogue of more than 15 million items, approaches that require domain knowledge and involve indirections via search experts become less attractive. Over the last decades, the Internet has developed into a document collection the dimensions of which easily dwarf the haystack comparison with size estimates ranging between 15 and 60 billion publicly accessible Web pages.

Search engines enable users to efficiently search and browse through their collections by specifying information needs in the form of a *query*. Queries can take very

[1] In the course of this dissertation, we will frequently encounter the notion of documents. In an information retrieval context, we will adopt a very wide definition of the term and consider it to be any kind of media item, e.g., books, Web sites, images, videos, or, music tracks, to name a few.

different forms, including boolean expressions, concrete document examples, or, sequences of keywords. Search engines measure the expected utility of documents, their so-called *relevance*. They try to highlight those documents that have the highest estimated likelihood of satisfying the user's information need (*i.e.*, the most relevant ones). For most collections and search settings, this estimation step is much more complex than simple pattern matching between queries and documents. Hardly any user can be expected to precisely know the exact documents they are searching for at query formulation time. According to Belkin [26], this so-called *anomalous state of knowledge* (ASK) requires users to envision and describe relevant documents without knowing the full range of available information. In this particular setting, relevance models have to account for a considerable degree of uncertainty in the user-provided query and carefully interpret the available sources of evidence.

Over the decades, numerous definitions of relevance have been described in the literature, e.g., by Bates [25], Mizzaro [154], or Borlund [37]. Considerable effort has been dedicated to creating consistent and universally applicable descriptions of relevance in the form of relevance frameworks. Examples of such frameworks are given by [184], [185], [190], [90], [154], and [37].

Across these various formal systems, a wide range of criteria indicating document relevance has been identified. *Topicality* is the single most frequently encountered ranking criterion that, in informal usage, especially in related domains, has often been used as a synonym for relevance. It expresses a document's topical overlap with the user's information need. For textual resources, it is often estimated based on term co-occurrences between query and document. There are however a significant number of further noteworthy relevance dimensions. Prominent specimen as for example compiled by [189] and [168] are:

Recency determines how timely and up to date the document is. Outdated information may have become invalid over time.

Availability expresses how easy it is to obtain the document. Users might not want to invest more than a threshold amount of resources (e.g., disk space, downloading time or money) to get the document.

Readability describes the document's readability and understandability. A document with a high topical relevance towards a given information need can become irrelevant if the user is not able to extract the desired information from it.

Credibility contains criteria such as the document author's expertise, the publication's reputation and the document's general trustworthiness.

Novelty describes the document's contribution to satisfying an information need with respect to the user's context. E.g., previous search results or general knowledge about the domain.

It is evident that these criteria can have very different scopes. Some of them are static characteristics of the document or the author, others depend on the concrete information need at hand or even the user's search context. Many of the previously introduced

frameworks describe inherently multivariate representations of relevance, accounting for the complex and volatile nature of relevance. As the variety of information services, offered for example on the Internet, grows, the need for highly specialised retrieval models with concisely tailored relevance estimation schemes arises. The one-search-fits-all paradigm loses more and more of its viability.

We can find a wide range of empirical studies investigating the distribution, nature and dynamics of relevance and how people assess it. Examples include: [168], [24], [229], [218], [187], and, [240]. These studies unanimously describe relevance as a composite notion, agreeing with [76]'s fundamental finding that topicality on its own is not sufficient to reliably judge document relevance.

The growing importance of multivariate notions of relevance is also reflected in the retrieval model literature where we find numerous examples of applied multidimensional frameworks, including [55], [221], [200], [117], and, [56].

Employing multivariate notions of relevance into the retrieval process has shown two distinct advantages over conflated univariate notions: Lavrenko and Croft [126] report consistently better performance as a consequence of explicitly modelling the IR process and the involved decision criteria. Nallapati [160] notes greater trust in, and more intuitive usability of systems that explicitly model the variety of criteria humans consider when assessing document relevance.

A recently established alternative way of catering for multi-dimensional models is given by the *learning to rank* (L2R) family. This data-driven class of approaches learns a ranking model based on a, typically large, number of shallow document, session and user features, as well as annotated historic interaction data as for example presented by [41], [136], and, [176]. While industrial applications often rely on learning-based methods in order to harness the considerable power of their massive amounts of usage data, they largely remain empirical "black boxes". Formal models for IR, on the other hand, compute the probability of relevance of a document towards a query and rank results accordingly [181]. Models that formally combine multiple relevance dimensions may be more valuable for human interpretation. In this work, we will employ large-scale data-driven methods for estimating concrete model parameters but will ultimately rely on a well-grounded formal modelling of the probability of relevance.

This dissertation strives to answer three high-level research questions, each of which is comprised of a number of auxiliary insights and studies:

Q1: **How is relevance influenced by subjective and situational factors?** To highlight the subjective nature of relevance, we will investigate in a concrete example, **(a)** if and how the search behaviour of individual user groups differs, and **(b)** whether search success can be predicted as a function of search behaviour. Subsequently, we will turn to studying situational relevance by researching **(c)** whether there is evidence for situational influence on relevance and behaviour, and, how such cases can be identified. Finally, **(d)** we are interested in which way we can still make use of previously collected usage and profile information to improve atypical search sessions.

Q2: **How to estimate individual relevance dimensions?** Numerous alternative ways of achieving this have been proposed and we will visit several prominent examples.

We investigate the use of fully automatic estimation schemes **(a)** based on document content or **(b)** traces of user interaction with the content. As a stark contrast, we demonstrate the use of human computation to show **(c)** how relevance estimation can be crowdsourced explicitly, or, **(d)** implicitly in the form of games.

Q3: **How to obtain a single probability of relevance?** We will investigate copulas, an example of a robust framework for multivariate probability estimates with complex interdependencies. In particular, we will research **(a)** how copulas compare to established multidimensional relevance frameworks, **(b)** we will demonstrate how the merit of using copulas can be predicted based on a number of properties of the concrete retrieval scenario at hand. Finally, **(c)** we will investigate the performance of copulas for the task of score fusion and how robust their fused scores are to low-quality outliers.

The research presented in this thesis was partially conducted within the European Commission's FP7 project PuppyIR. The project's objective is to investigate and develop child-friendly means of information access. Children are not exempt from the growing ubiquity of electronic information systems, but they are often at a disadvantage as systems tend to be designed with the specific interests, needs, cognitive and motor skills of adults in mind. Employing a wide array of techniques from the domains of information retrieval, data mining and natural language processing, the project delivered an open source library that allows for the creation of age-appropriate, flexible and easily maintainable search services for children[2]. The implications and insights drawn throughout the following chapters aspire to demonstrate generality and applicability across a wide number of tasks, information types and settings. However, a number of concrete examples (in particular in Chapters 2 and 4-6) will derive from the domain of information services dedicated to an audience of very young users.

This dissertation is organized in three parts. In this first part, we motivate the challenge of representing document relevance as a contextual, multidimensional property. To this end, earlier in this chapter, we briefly revisited an introductory overview of theoretical relevance frameworks and real world industry-scale solutions, and discussed their ability to capture the notion of multidimensional relevance. Chapter 2 underlines the subjective nature of relevance by showing how preferences and requirements towards an information system differ fundamentally across individual users and user groups. To give a concrete example, we discuss information services for children and their use in classroom or hospital settings. In Chapter 3, we expand the search context beyond the user model and include situational influences on document relevance. We show that even for the same searcher, relevance can strongly depend on the context in which the search was performed.

The first step towards a contextual multivariate relevance model is to estimate the contextual constituents that, in combination, describe the overall relevance of a document towards a query. In Part 2 of this thesis, we discuss two major classes of relevance estimation schemes: Chapter 4 describes automatic estimation methods based on document content and user activity. In Chapter 5, we turn to the use of human computa-

[2] More information about PuppyIR can be found on the project Web page (http://www.puppyir.eu) as well as on Sourceforge (http://sourceforge.net/projects/puppyir/)

tion, either explicitly in the form of crowdsourcing assignments, or implicitly by means of gamified tasks, for relevance estimation.

The third and final part of the thesis takes over the previously estimated relevance scores and combines them into a single, univariate probability of relevance according to which we will rank results. To this end, Chapter 6 introduces *copulas*, a probabilistic framework which allows for convenient estimation of joint relevance distributions from their individual marginals. The validity of the method is demonstrated for both, direct estimation of relevance as well as score fusion from multiple independent rankings.

References

[24] Carol L. Barry. "User-defined relevance criteria: an exploratory study". In: *Journal of the American Society for Information Science* 45.3 (1994), pp. 149–159.

[25] Marcia J. Bates. "Information search tactics". In: *Journal of the American Society for information Science* 30.4 (1979), pp. 205–214.

[26] Nicholas J. Belkin. "Anomalous states of knowledge as a basis for information retrieval". In: *Canadian journal of information science* 5.1 (1980), pp. 133–143.

[37] Pia Borlund. "The concept of relevance in IR". In: *Journal of the American Society for information Science and Technology* 54.10 (2003), pp. 913–925.

[41] Chris Burges et al. "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 89–96.

[55] Nick Craswell et al. "Relevance weighting for query independent evidence". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 416–423.

[56] Fabio Crestani et al. ""Is this document relevant?... probably": a survey of probabilistic models in information retrieval". In: *ACM Computing Surveys (CSUR)* 30.4 (1998), pp. 528–552.

[76] Thomas J. Froehlich. "Relevance Reconsidered - Towards an Agenda for the 21st Century: Introduction to Special Topic Issue on Relevance Research". In: *Journal of the American Society for Information Science* 45.3 (1994), pp. 124–134.

[90] Stephen P. Harter. "Psychological relevance and information science". In: *Journal of the American Society for Information Science* 43.9 (1992), pp. 602–615.

[117] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. "The importance of prior probabilities for entry page search". In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, pp. 27–34.

[126] Victor Lavrenko and W. Bruce Croft. "Relevance based language models". In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 120–127.

[136] Tie-Yan Liu. "Learning to rank for information retrieval". In: *Foundations and Trends in Information Retrieval* 3.3 (2009), pp. 225–331.

[154] Stefano Mizzaro. “Relevance: The whole history”. In: *Journal of the American Society for Information Science* 48.9 (1997), pp. 810–832.

[160] Ramesh Nallapati. “Discriminative models for information retrieval”. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2004, pp. 64–71.

[168] Taemin Kim Park. “The nature of relevance in information retrieval: An empirical study”. In: *The library quarterly* (1993), pp. 318–351.

[176] Filip Radlinski and Thorsten Joachims. “Query chains: learning to rank from implicit feedback”. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 239–248.

[181] Stephen E. Robertson. “The probability ranking principle in IR”. In: *Journal of documentation* 33.4 (1977), pp. 294–304.

[184] Tefko Saracevic. “Relevance: A review of and a framework for the thinking on the notion in information science”. In: *Journal of the American Society for Information Science* 26.6 (1975), pp. 321–343.

[185] Tefko Saracevic. “Relevance reconsidered”. In: *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*. 1996, pp. 201–218.

[187] Reijo Savolainen and Jarkko Kari. “User-defined relevance criteria in web searching”. In: *Journal of Documentation* 62.6 (2006), pp. 685–707.

[189] Linda Schamber and Judy Bateman. “User Criteria in Relevance Evaluation: Toward Development of a Measurement Scale.” In: *Proceedings of the ASIS Annual Meeting*. Vol. 33. ERIC. 1996, pp. 218–25.

[190] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. “A re-examination of relevance: toward a dynamic, situational definition”. In: *Information processing & management* 26.6 (1990), pp. 755–776.

[200] Ilmério Silva et al. “Link-based and content-based evidential information in a belief network model”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2000, pp. 96–103.

[218] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. “How users assess web pages for information seeking”. In: *Journal of the American society for Information Science and Technology* 56.4 (2005), pp. 327–344.

[221] Howard Turtle and W. Bruce Croft. “Evaluation of an inference network-based retrieval model”. In: *ACM Transactions on Information Systems (TOIS)* 9.3 (1991), pp. 187–222.

[229] Peiling Wang and Marilyn Domas White. “A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages”. In: *Journal of the American Society for Information Science* 50.2 (1999), pp. 98–114.

[240] Yunjie Calvin Xu and Zhiwei Chen. “Relevance judgment: What do information users consider beyond topicality?” In: *Journal of the American Society for Information Science and Technology* 57.7 (2006), pp. 961–973.

2

Relevance in Personal Context

Everything we hear is an opinion, not a fact.
Everything we see is a perspective, not the truth.

-Marcus Aurelius, Emperor

People, quite naturally, are different. Individual preferences and tastes let us favour certain books, pieces of music, TV programmes, *etc.* over others. While, in some domains, we can find a broad consensus on what makes quality, in others, the field is more diverse. Document relevance makes no exception to this rule. In the previous chapter, we discussed a range of multi-dimensional relevance frameworks that were introduced in the literature. Many of them include aspects that reflect the user's personal context in order to better adjust document selection to the searcher's individual preferences and needs. In this chapter, we will try to understand subjective properties of relevance that vary between individual users and groups of users. As a concrete example, we will address the use case of very young searchers whose cognitive and motor skills, as well as their behaviour when interacting with technology is very different from those of adult users.

Over the previous decades, children have been growing considerably more acquainted with technologies such as computers and the Internet. This trend can be observed across all age levels, starting with children as young as 3-5 years. As a consequence, the age of first contact with said technologies is decreasing while the overall time spent using Web-based information systems rises significantly. According to the EU Kids on-line report [139], 83% of European children in the age of 6 to 10 and 96% of the 11 to 14 year-

olds use the Internet at least on a weekly basis. Similar figures are reported by national counterparts in countries such as for example the UK [163] or Germany [58].

This new media-centric development is reflected in modern school curricula that encourage and support children, already at very young ages, in their use of computers. In this way, schools aim to prepare children for dramatic changes in skillsets demanded by both society and the labour market. Most prominently, Web information search has nowadays become an integral part of the preparation phase of essay writing or the creation of classroom presentations.

Popular Web search engines such as Google, Bing, Yandex and their various competitors, however, are designed with the needs and capabilities of adult users in mind. Needless to say, this assumption is not always correct when catering for an audience of 10 year-olds. We will dedicate the first part of this chapter to understanding the differences between adult and child searchers as an example of the subjective nature of relevance, before detailing two potential solutions for supporting young searchers.

2.1. Children as information seekers – A case study

In the literature, we can find a wide range of work dedicated to general models of information seeking. In one of the early studies on human search behaviour, Saracevic and Kantor [186] conducted a qualitative survey in which 40 participants were interviewed concerning their search habits. The authors tested a number of hypotheses of human information search, finding significant differences between individual searchers in terms of search behaviour, preferences and strategies even for identical information needs. In a later study, Kuhlthau [118] revisits the information search process from a user perspective, arguing for the necessity of users understanding IR systems in order to interact with them faithfully. According to the author's proposed model, this hypothesis is confirmed in a user survey. In their survey article, Belkin and Croft [27] establish a close relationship between information retrieval and information filtering. The latter of the two is crucially required to be user aware in order to provide suitability of retrieved results given the user's context and information need. Marchionini [147] describes the information seeking process as constructed of 3 partially parallel phases. Each search starts with the user *understanding* their information need. Followed by *planning and executing* the query and a subsequent *evaluation and use* of the returned results. Choo et al. [50] incorporated the search session's driving motivation into the model, finding it to have clear influence on the observed search strategies such as undirected browsing or targeted searching. In 2005, Ingwersen and Järvelin [101] highlighted the importance of searcher context for models of information seeking. In a dedicated cognitive model, they establish various levels on which context crucially influences search processes. A particular aspect of information seeking models that was recently highlighted is concerned with the notion of search success. Several studies, including [242] and [146], define search success as the rate of search sessions that result in satisfying the user's information need. Stronge et al. [209] relate a user's search strategies to their likelihood of search success in Web search.

Given the dramatic changes and developments that children undergo on their way to adulthood physically, emotionally and cognitively, it makes intuitive sense to demand a dedicated formal framework of children's information seeking. Shenton and Dixon [195]

present a range of seeking models developed based on empirical survey results among children in the age of 4 to 18. Their grounded model of information seeking via the Internet consists of 11 different actions or influences. Before the start of the actual search process, a multi-step framework accounts for factors such as the origin of the information need, the directness of use or the place of information access. While most of these factors are generic enough to apply to adult seekers as well, some aspects are specific to children only. Examples include the degree of parental support during the search or whether the information need is related to school. Bilal et al. [34] presented a study of Arabic children's interaction with a digital library system. Based on their search behaviour on the International Children's Digital Library (ICDL)[1], a Web interface that introduces children to various cultures with books, she formulated an alternative model of children's information seeking. Her model is centred around three fundamental concepts:

Browsing. A child scans the list of book thumbnails and moves to the next page with thumbnails.

Backtracking. A child uses the back arrows of ICDL or the back button of their browser to return to an earlier stage of their search.

Navigating. A child uses the ICDL's functionality for page-internal navigation such as zooming in on particular page aspects.

A number of particular challenges are frequently reported to frustrate young searchers and prevent them from achieving search success. **(A)** Moore and St George [158] report query formulation as difficult due to insufficiently-developed writing skills and small active vocabularies. **(B)** Borgman et al. [36] note that identifying relevant organic search results often overwhelms children as they struggle to judge which results will satisfy their information needs. **(C)** Finally, Large et al. [124] found the overall number of results presented by a typical Web search engine to impose a high cognitive load on children that often leads to confusion. Bilal [32] investigated the search success rates of children using Yahoo!'s child-oriented platform Yahooligans[2]. Later in this chapter, our user survey will employ search success as a key surrogate for determining children's need for assistance with information search assignments in school settings.

An important line of related work central to our case study is led by Allison Druin. In a pilot study, she and her colleagues investigate how children of an age between 7 and 11 years old search the Internet using keyword interfaces at home [64]. The study highlights a number of barriers that hinder children from successfully searching the Web using technologies designed for adult users. The particular challenges include spelling, typing, query formulation and deciphering results. In a subsequent qualitative home study among 83 US children, [63] establish a taxonomy of young searchers. After an initial interview, children were encouraged to search for information using a Web search engine of their choice. Qualitative analyses revealed a number of characteristics that motivated a framework of the following 7 searcher roles:

[1] http://www.childrenslibrary.org

[2] At a later stage, the platform was rebranded and is now known as Yahoo! Kids (http://kids.yahoo.com)

Developing searchers tend to use natural language queries, “asking" the search engine questions. They are able to complete simple queries, but have trouble with complex ones.

Domain-specific searchers limit searches to finding content specific to a domain of personal interest. They repeatedly return to a small number of specific Websites which they have accepted as authoritative or informative.

Power searchers display advanced search skills and are able to use keywords instead of natural language in the query formulation step. They do not suffer from breakdowns and are able to solve more complex search assignments.

Non-motivated searchers are not persistent when searching. They lack motivation to find alternative problem solutions or query reformulations and easily give up after set-backs.

Distracted searchers have trouble staying focused on the current search task. They frequently side track into investigating other information needs and are easily distracted by external stimuli.

Visual searchers guide their information search along visual clues. They often start search sessions from image and video search engines, identifying visual concepts relevant to their assignment.

Rule-bound searchers follow a strict set of rules from which they are not easily able to deviate.

Motivated by this role framework, we have conducted a user study with 29 children from different grade-levels at the Dutch elementary school “De Kroevendonk" in Roosendaal in the Netherlands. Children ranging from 8-12 years of age are particularly interesting for Web search experiments, as they already have well-developed reading skills, while still displaying significantly different behaviour from adult searchers [3, 65]. The experiment was carried out during regular school hours with informed consent by the students' legal guardians and conforming to the Dutch and European Data Protection Acts [129, 128, 130]. To introduce the researcher as well as the research goal to the children and to make them comfortable with the experiment, we gave an explanatory presentation in all participating classes with the possibility for asking questions.

The sessions of 5 participants were collected in a pilot run used to refine the experiment set-up. This leaves 24 participants in the final data collection. Figure 2.1 shows key properties such as age, gender or class distribution of the participants. We could observe a generally high degree of computer skills, with the majority of participants reporting regular computer and Internet contacts.

Experimental set-up

To limit external distractions, experiments were conducted in a separate room in the school building and children were individually participating while their peers would continue with regular classroom activities. One researcher was always present during the experiment to take notes as well as to assist if the child had questions concerning the

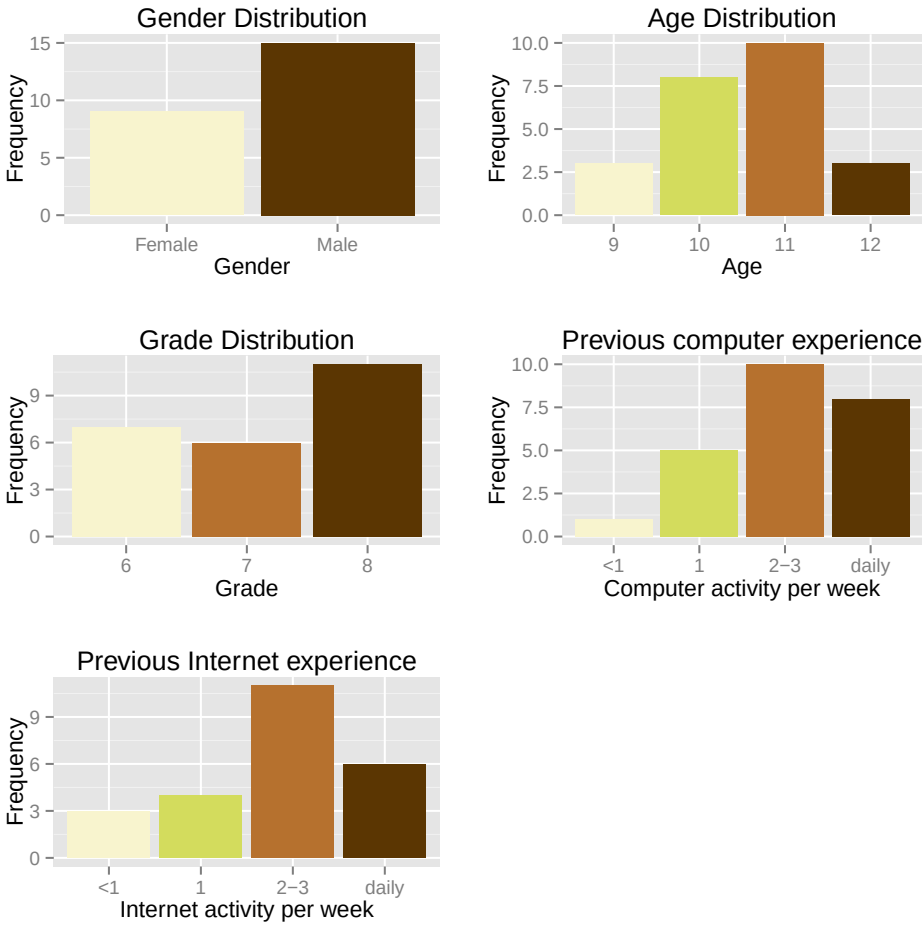

Figure 2.1: Participant demographics.

general experiment. The research did not interfere with or comment on the search processes. A browser-based survey system guided the participant through the experiment. After a brief introduction, we asked three initial questions about well-being and prior experience:

"How do you like participating in this study?"

"How often do you use a computer?"

"How often do you use the Internet?"

After this collection of personal background information, the actual search tasks started. We use three types of questions: **(a)** *Factual questions* can be answered with a single sentence. Tasks like this can typically be answered with a single query. **(b)** *Open-ended questions* express exploratory information needs that aim towards acquiring broad knowledge about a given topic. **(c)** *Multi-step questions* require advanced reasoning to combine information acquired over multiple queries in a session. To create an initial feeling of success and to enable the participant to adjust to the search interface, we started with a simple fact-based question, before moving on to the actual search assignments (one per question type).

"What do whales eat?" (a)

"How many brothers and sisters does Queen Beatrix have?" (a)

"What can you find out about the first car ever built? Write down some facts about it." (b)

"Which day of the week is the birthday of the Dutch prime minister in 2011?" (c)

The questions were shown one by one. Only after answering the current assignment, the next one would be made available. After completing the experiment the researcher asked about the participant's opinion on the questions and how she or he liked the overall experience. To prevent frustration in the case of struggling searchers who could not find an answer to a question, we introduced a time limit to the tasks. For the first two questions the participants had 6 minutes each, for the second one 8 minutes and for the last one 10 minutes. After this time, the researcher ended the task and encouraged the participant to move on to the next step. The default search engine shown in the survey interface was Google which [63] previously reported to be popular among young searchers. We did, however, not restrict the use of other search facilities. After they completed the final search task, participants were once more asked to indicate how much they enjoyed the experiment, before they were guided back to the classroom.

Data collection

Besides the qualitative observations made by the researcher who takes notes on physical signals of motivation, confidence and immersion, we exploit a range of additional data sources in order to accurately capture relevant session properties. To facilitate manual annotation of search sessions, we used CamStudio 2.0, an open source screen capturing

software[3] to be able to revisit all screen activity in the form of video files. Additionally, to create a more machine-readable representation of search sessions, we employed a Firefox add-on, the HCI Browser [45]. This program can be used to log HTTP requests, mouse movements, keyboard input and click data. Instead of logging events on page level by injecting javascript as is done by popular tools such as Usaproxy [20], this add-on makes it possible to log every action within the browser. Consequently, we can also capture signals as for example the use of the browser's back button that would otherwise have eluded recording. Figure 2.2 shows an example of the data recorded by the HCI Browser.

[3] http://camstudio.org

1292494008865 16-12-2010 11:06:48 Focus http://www.google.nl/ http://www.google.nl/ clientx=1366 clienty=605
1292494008929 16-12-2010 11:06:48 LoadCap http://www.google.nl/ http://www.google.nl/ clientx=1366 clienty=605
1292494009544 16-12-2010 11:06:49 MouseMove http://www.google.nl/ x=596 y=254
1292494010592 16-12-2010 11:06:50 MouseMove http://www.google.nl/ x=606 y=253
1292494010822 16-12-2010 11:06:50 LClick x=606 y=253 undefined http://www.google.nl/
1292494012166 16-12-2010 11:06:52 KeyPress key=D keycode=68 combi= http://www.google.nl/
1292494012503 16-12-2010 11:06:52 KeyPress key=E keycode=69 combi= http://www.google.nl/
1292494013273 16-12-2010 11:06:53 KeyPress key=Space keycode=32 combi= http://www.google.nl/
1292494013795 16-12-2010 11:06:53 KeyPress key=E keycode=69 combi= http://www.google.nl/
1292494014167 16-12-2010 11:06:54 KeyPress key=E keycode=69 combi= http://www.google.nl/

Figure 2.2: Data sample captured by the HCI Browser.

Data analysis

The previously described user study leaves us with a collection of 96 search sessions (4 per unique participant). For each session, we assigned 2 types of labels: **(1)** A role label, following [63]'s categorization. **(2)** A binary search success label, indicating whether the participant could find a valid answer to the task. The decisions were based on the qualitative notes taken during the search session as well as the screen recordings of the full sessions. We conduct our annotation on session-level rather than user-level to account for individual preferences and abilities for solving different task types. Each session was independently labelled by 2 researchers. As a measure of task feasibility and annotation reliability, we investigate inter-annotator agreement. An overall share of 82% of all sessions received identical labels by both annotators. Table 2.1 shows task-level agreement ratios and Cohen's κ scores. We can observe an interesting tendency of task 1 and 4 agreements being significantly higher than those for tasks 2 and 3. The tasks were designed and ordered by increasing difficulty. This initial overview suggests that very easy or difficult tasks are more beneficial for determining role affiliations. We will take this intuition as one of our hypotheses for later design and evaluation of our automatic classification scheme. To obtain final judgements, the annotators discussed all instances of disagreement, arriving at consensus labels for each.

Table 2.1: Inter-annotator agreement per task.

Task	Agreement	κ
1	0.92	0.83
2	0.71	0.45
3	0.75	0.57
4	0.96	0.92

Table 2.2 shows the distribution of roles in our and Druin et al.'s work [63]. The developing and power searcher roles were found to be dominant in the present data set. All other roles could at most be observed sporadically. The developing role was already frequent in the 2010 study, but many other roles follow significantly different frequency distributions. We see the reason for this difference in the changed setting between information search at home and search assignments in a school setting. Due to the more formal environment, phenomena such as non-motivated searchers are intuitively less likely. Both our and Druin's studies find a strong correlation between the participants' age and their likelihood of being a power searcher. An even stronger connection could be found between the participants' school grade and their power searcher status. Despite the correlation between age and school grade, formal school education seems to explain advanced search proficiency better than mere age.

To give further insight into the effect of prior experience in information search and general computing on search success and power searcher status, we analysed this relationship more deeply. There was no substantial correlation between the participants' background credentials such as their gender or their self-reported computer and Internet experience and their role affiliations and search success. This finding supports our claim that dedicated support and training are valuable even for children who are practised computer users.

Table 2.2: Search role distribution as observed in our school study and Druin et al.'s home setting.

Role	School	Home
Developing searcher	48%	43%
Domain-specific searcher	0%	21%
Power searcher	47%	12%
Non-motivated searcher	2%	9%
Distracted searcher	0%	6%
Visual searcher	3%	5%
Rule-bound searcher	0%	4%

In addition to the previously-discussed questions on personal background, we asked each participant about their emotional state before and after participating in the experiment. The questions offered a 5-point scale ranging from "I really do/did not want to participate." to "I really like/liked to participate.". Based on the findings of Yusoff et al. [244], the answers were supported by a smiley-scale that visually underlined emotional states. The concrete scale used is depicted in Figure 2.3. Table 2.3 shows the emotional state before and after participating in the experiment.

In the majority of sessions, the emotional state changed during the course of the experiment. In order to further understand this observation, we define δ_e as the number of categories by which a participant's emotional state changed before and after the experiment. A negative number indicates a drop in motivation while a positive number represents gains in well-being. We can find a mild correlation between δ_e and a participant's success rate ($\rho = 0.43$), and their likelihood of being power searcher ($\rho = 0.31$). This underlines the assumption, that search failures can have a frustrating effect on young searchers and may even prevent them from indulging in future searches. This emphasises the importance of appropriate search support at this stage of a child's development.

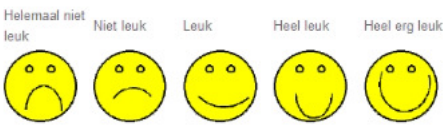

Figure 2.3: Smiley scale.

Table 2.3: Development of participant motivation before (rows) and after (columns) the experiment.

	1	2	3	4	5
1	-	-	-	-	-
2	-	-	-	-	-
3	-	-	1	1	2
4	-	-	4	8	3
5	-	-	-	3	2

2.2. Predicting children's Web search success

Previously, we saw that, both in the literature as well as a dedicated study among elementary school students, children face significant challenges when using Web search engines designed for adults. This scenario is especially relevant in the school setting.

When guiding classroom assignments involving Web information search, elementary school teachers have to supervise groups of 20 - 30 children who are each, individually, searching the Internet, at the same time. A particular challenge lies in the fact that such groups tend to be heterogeneous in their information search capabilities. While some students are coping well with the task, others may struggle. Identifying those few students in a large group who need the most assistance at a given point in time, however, is not easy. Wallace et al. [227] note that time that could have been invested into aiding struggling children may be wasted due to the problem of identifying them in the first place.

In this section, we propose a solution for this problem by devising an automatic scheme for predicting children's search success based on a wide range of cognitive, information theoretic and empirical features of search sessions. We envision a teacher's "dashboard", integrated in electronic learning environments, which highlights where help is needed most. In this way, the teacher can dedicate more time to children who need help.

In the following, we present a detailed outline of an automatic search role and success classification scheme. As a starting point, we will describe a wide range of features motivated by empirical observations as well as by literature in cognitive and behavioural science. We can identify 3 types of features accessible during a search session: **(1)** *task-independent features* are static properties of the participant such as age or gender. **(2)** *task-dependent direct features* are directly extractable from the interaction log but may vary across tasks for the same participant. Examples include the number of Web page visits or the number of mouse moves. **(3)** *task-dependent inferred features*, finally, cannot be directly read from the search log but require further processing steps that may involve external data. Think for example of the percentage of natural-language queries issued in the session. We identified a total of 37 individual features.

Task-independent features

Age. The age of the participant at the time he or she participated in the experiment. Druin already showed that for example older children are more likely to be power searchers than younger children. Bilal [33] investigated the differences between children and adults as Web users.

Grade. The participant's grade level according to the Dutch school system. While being correlated with participant age, this feature aims more at the amount and level of formal school education the participant has received.

Gender. The participant's gender was included for completeness's sake. Previous research did, however, not find significant gender-specific differences in children's search proficiency.

Computer Experience. The participant's self-reported previous experience with computers can be expected to give good indications of his or her likelihood of success.

Internet Experience. In analogy, we also include the self-reported Internet experience.

Task-dependent direct features

Total number of mouse movements. Qualitative analysis of our search sessions showed a good correlation between motor skills with the mouse and power searcher status. The ability to navigate the search interface with only the necessary user actions (e.g., a low number of mouse moves) is therefore seen as an indicator of operational competence.

Mouse movement patterns. Previous work by Pusara and Brodley [173] employed mouse movement characteristics for user authentication. Instead of identifying specific users, we try to generalize mouse movement patterns of groups of users by a number of additional mouse input features beyond the targeted ones that were previously introduced. Concretely, this encompasses: (1) The number of mouse moves per second (2) average mouse move distance (3) move distance standard deviation (4) average horizontal distance (5) average vertical distance (6) the ratio of vertical / horizontal distances.

Total number of page visits. Capable searchers are able to accurately decide on result relevance based on Web page titles and snippets displayed on the search engine result list. High counts of visited and abandoned pages indicate headless browsing.

Number of visits per unique page. Druin et al. [63] found inexperienced young searchers to seemingly arbitrarily revisit Web pages multiple times. We exploit this observation by measuring the average number of times each unique page was visited within a session.

Average display time per Webpage. Query log analyses of children's interaction with a popular search engine showed, that young children often experience difficulties judging the relevance of search result snippets which manifests in a high number of very brief visits that are quickly abandoned once the participant realizes that the page was not what he or she was looking for [65]. We capture similar behaviour in terms of average display times per Web page.

Total number of mouse clicks. In analogy to the previous features, click events are counted and employed as an additional indicator of operational confidence.

Total number of issued queries. Experienced searchers are expected to be able to phrase their information need more accurately in keywords than beginners who have to rely on subsequent rephrasings. We count the number of such reformulations per session.

Number of query term additions. We count the overall number of times additional terms are added to a previous query. An example of such an operation would be the step from “prime minister birthday" to “prime minister Netherlands birthday".

Number of query term removals. Query generalizations by means of removing query terms are counted. E.g., from “prime minister Netherlands birthday" to “prime minister Netherlands".

Averaged query distance Drastic query reformulations are an indicator of low confidence in the original search terms and, more generally, in the participant's search skills. In this work, we measure the distance in terms of query term overlap between issued queries. We use the Jaccard Coefficient as distance metric. Finally, the computed distances are averaged.

Average query length. Kumaran and Carvalho [119] found long queries to be problematic for modern search engines. We count the number of terms per query and average across all queries within a session.

Query length standard deviation. To give another alternative measure of query reformulation activities, we include the query length standard deviation across a session's queries.

Typing speed. Hutchinson et al. [100] reported interaction with keyboard interfaces to be one of the major sources of frustration for inexperienced searchers. We measure typing speed for each sequence of keyboard inputs without any interruptions by mouse moves or clicks. Finally, the number of keystrokes per minute is averaged across all such sequences.

Time spent on search engine pages. We measure the absolute time per session the participant lingers on search engine pages. This represents the combined efforts of query formulation and result inspection.

Number of back button clicks. Inspired by Bilal's model of children's information seeking [34], we inspect backtracking activities by means of counting the number of times the participant makes use of the browser's back button.

Session length. The total time a participant invests into solving a task is recorded and can be seen as a surrogate for search proficiency.

Number of backspace keystrokes. Spelling has been frequently observed (e.g., by [100]) to be one of the specific challenges of children's query formulation steps. We report the total number of back space keystrokes per session as an indication of the participant's orthographic competence.

Number of scroll actions. During our manual inspection of search sessions, we saw that not every child knew how to use the mouse wheel for navigation. To capture the participant's ability to use this advanced control mechanism, we record the total number of mouse scroll actions per session.

Question words. Inexperienced searchers tend to “ask” the search engine for information. We check for the presence of question words such as “why”, “when”, “who” etc.

Stop words. Modern search engines are designed and optimized for accepting keyword queries. Excessive usage of stop words indicates low search experience. We report the averaged number of stop words per query.

Task-dependent inferred features

Query-task distance. Query formulation is a crucial and cognitively-expensive step in the information search process. Inexperienced searchers have been found to take the "shortcut" of copying the assignment question as a query. We measure the Jaccard distance between tasks and observed queries. We expect this distance to be minimal for developing searchers and significantly larger for experienced users.

Average number of verbs|nouns|adjectives per query. According to Druin et al. [63], developing searchers tend to issue natural language queries. We apply part-of-speech tagging to identify different token type distributions.

Based on this selection of 37 features, we trained a number of different machine learning techniques for the task of automatic role classification based on the collected session data and evaluate classification performance in a 10-fold cross-validation setting. Our experiments are based on the WEKA implementation of the respective learning methods [87]. Table 2.4 shows the best classification performance per method averaged across all classes. To set our results into context, we include a dominant class baseline that assigns the most frequent label to all sessions. All evaluated methods performed significantly and consistently better than the baseline intuition. The overall strongest approach was a support vector machine (SVM). Statistical significance was tested using a Wilcoxon signed rank test with $\alpha < 0.05$.

Table 2.4: Role classification performance by method.

Method	P	R	F_1
Dominant class baseline	0.23	0.48	0.31
Naive Bayes	0.65	0.68	0.66
Logistic Regression	0.59	0.57	0.58
MLP	0.74	0.75	0.74
SVM	**0.76**	**0.80**	**0.78**
Decision Table	0.61	0.65	0.63
Decision Tree	0.59	0.62	0.60
Random Forest	0.66	0.69	0.67

A reliable means of identifying individual search role affiliations makes an important contribution to educating children regarding their Web search abilities. Knowledge about their specific deficits (e.g., those of a visual searcher) helps teachers and parents to give targeted advice on how to improve. For the problem at hand, however, we can reformulate our task to finding those children in the classroom that fall into one of the defective search roles (i.e., all except for power searchers). We will refer to this lower-order classification problem as *Deficit detection*. Adjusting to this new setting, we achieve significantly higher scores than for dedicated role prediction. The strongest models ap-

Table 2.5: Deficit detection performance by method.

Method	P	R	F_1
Dominant class baseline	0.28	0.53	0.37
Naive Bayes	0.66	0.82	0.73
Logistic Regression	0.67	0.71	0.69
MLP	0.79	0.73	0.76
SVM	**0.84**	**0.80**	**0.82**
Decision Table	0.67	0.67	0.67
Decision Tree	0.63	0.63	0.63
Random Forest	0.69	0.69	0.69

proximate the agreement ratio of our human annotators. Table 2.5 reports the resulting performance figures.

When working with search roles as defined by Druin et al., we noticed a conceptual disparity between some of the categories. While some roles are essentially performance oriented (power and developing searchers), others are based on the employed search strategy (visual, rule-bound and domain-specific searchers) and a third group is concerned with notions of attentiveness (non-motivated and distracted searchers). While a manual qualitative analysis of searcher behaviour may benefit from such a broad categorization scheme, it appears to be problematic for automatic methods designed for classroom teacher support. In our dataset the distribution of search roles was so skewed towards power and developing searchers that they effectively formed a proxy for search competency. A closer investigation of the data set, however, showed that power searcher status is only loosely correlated to the likelihood of search success ($\rho = 0.4$). It appears as if Druin's roles cannot be seen as direct surrogates for search competency. Further evidence was given in Section 2.1, where we observed a relationship between searcher motivation and search success.

In our final classification experiment, we abandon Druin's class hierarchy and turn to directly predicting search success. Table 2.6 compares the performance of a number of classifiers for this task. The best overall performance could be achieved using an SVM approach with polynomial kernel ($\epsilon = 10^{-12}$, $c = 0.6$ and $e = 1$). This prediction model was able to correctly identify 3 out of 4 successful search sessions. Based on these performance figures, a classroom teacher could prioritise the order in which she or he visits its students, based on their likelihood of search success as determined by an automatic classifier running in the background of the school's computers. Given the substantial performance gains over baseline intuitions (i.e., checking with every child), our method can be expected to result in less time being invested into identifying struggling students. This, in turn, frees up resources for actual assistance and teaching.

In order to gain a deeper understanding of the domain, we identified the best-performing features according to our SVM model. Table 3.5 shows the top 7 features for the tasks of deficit and success prediction. We find a high overlap between both sets, confirming the central role of those notions. In both scenarios, being in a higher school grade, phrasing short queries with only few nouns and refraining from substantial query shortenings, are

Table 2.6: Success prediction performance.

Method	P	R	F_1
Dominant class baseline	0.27	0.52	0.36
Naive Bayes	0.63	0.78	0.70
Logistic Regression	0.59	0.60	0.59
MLP	0.66	0.72	0.69
SVM	**0.77**	**0.75**	**0.76**
Decision Table	0.64	0.80	0.71
Decision Tree	0.58	0.60	0.59
Random Forest	0.66	0.66	0.66

indicators of successful searches and power searcher status. The ranking of features is consistent across task types with only minor differences in relative contribution weights. The relative contribution of features decreases rapidly with rising rank. Models based on the 3 highest-ranking features were able to approximate the performance of those incorporating the full feature space, showing no significant differences in performance.

Our findings have two key implications on the educational sector: **(1)** Search success prediction can be reliably used to aid teachers to quickly identify those children that struggle with a search assignment and that would therefore benefit from assistance. **(2)** Role prediction appears to be a valuable method for identifying children's search strategies. Some of these strategies are better suited for use in Web search scenarios than others. Gaining knowledge about children's search strategies enables teachers and educators to provide targeted guidance, highlighting difficult aspects of the search process and how to best address them. The concrete roles drawn from previous work may, however, need to be revised for application in the classroom setting.

Table 2.7: Best deficit / success prediction features.

Rank	Deficit	Success
1	# query term removals	grade
2	grade	# query term removals
3	# query nouns	# query nouns
4	horiz. mouse distance	# back buttons uses
5	mouse move interval	avg. query length
6	# back button uses	# query adjectives
7	avg. query length	# of visited pages

2.3. Designing interfaces for young searchers

Previously, we aimed at predicting children's search success based on session level features, in order to direct an educator's attention to struggling students. As we observed previously, a number of key obstacles hampering children's Web search success can be related to the search engine's user interface. Typically, these interfaces are designed with

adult users in mind and do not cater for children's specific needs. We will now switch domains from the school setting that was described previously, to young searchers with medical information needs.

For children, illnesses and other undesirable medical conditions can be very confusing and frightening. Children faced with such problems will often express an interest in learning about their medical case, what is happening to them, and what to expect. However, finding information related to medical conditions is often a difficult and sensitive task. Consequently, designing and developing search services for children presents a number of challenges, including: children's problems expressing complex information needs, finding and identifying relevant information, and ensuring that information is understandable, appropriate, and sensitive to the child's physical and emotional state.

There is a substantial body of work dedicated to children's behaviour when interacting with search engines, both over local collections as well as on the Web. To motivate the design decisions taken in this work, we will briefly summarize their main findings. **(1)** Bilal [32] found query formulation to be a major source of frustration for young searchers. Due to their smaller active vocabularies, children are not as proficient as adults in finding the right keywords to express their information needs. This becomes even more important in domains with inherently difficult terminology such as medical topics. **(2)** After query formulation, the returned result page needs to be evaluated. Distinguishing organic search results from advertisements and sponsored results is an easy task for adults. Children, however, have been shown to struggle with the step [162]. **(3)** Even without the problem of advertisements, identifying relevant search results has been found to be a challenging task for children. They were observed to significantly more often click on irrelevant high-ranked results without critically questioning the presented material [65, 62]. **(4)** A specific habit that has often been observed for children is a preference for browsing over searching [63, 62]. Where adult users often explore a given topic by iteratively refining the search query, children tend to browse through the results of the initial query to find the desired pieces of information. This finding can be related to their previously-mentioned difficulties in query formulation. **(5)** Dedicated children's search engines tend to be loaded with playfulness and involved metaphors that abstract the search process to a more child suitable level. Jochmann [104] found that playfulness and entertainment can impede search success if the interface is not clear any more.

To address these challenges, we developed the Emma Search engine (EmSe) for the Emma Kinderziekenhuis (EKZ) at the Amsterdam Medical Centre[4]. The goal of the EmSe service is to improve the accessibility of information, in particular from the medical domain, along with the services provided by the patient information centre by:

1. providing an engaging interface that encourages children to explore,
2. facilitating query formulation,
3. improving the understandability of content, and,
4. enabling moderated and trusted Web and medical site search services.

[4] http://www.emmakids.nl/

EmSe is built using the PuppyIR Framework described by Glassey et al. [81], which provides a flexible suite of components that can be combined to build child-specific search services. Component types range from interfaces to various search resources (e.g., Bing, YouTube, Twitter) to a collection of information processing components that filter and modify both queries and results to support the user and their search tasks. Fundamentally, EmSe enables searching the information centre's local information repository, trusted medical sites as well as the Web, over which the following services are built: **(1)** the Body Browser, a novel visual querying interface, which lets children explore the patient information centre, **(2)** a multi-site search service of recommended and related medical sites, which lets children find out more about medical conditions from high-quality sites, and **(3)** a moderated Web search service, which lets children safely search the Web via moderated queries and results. To help children understand difficult medical terms, returned documents are augmented such that **(4)** known medical terms are annotated with simplified explanations. The interface is kept clean and minimalistic to avoid overwhelming the child with information or distractions. EmSe only shows a limited number of graphically enhanced search results, incorporating cover images of books and DVDs from the information centre. Figure 2.5 shows the EmSe user interface. In order to make the interface more appealing and to enhance the user experience, the search process is guided by a number of comic avatars representing different search metaphors. The boy and girl characters represent the child operating the search engine while the puppy character represents the search system. The child engages the system through a dialogue box which prompts, "I would like to find .." The puppy retrieves result sets and continues the dialogue by providing search and spelling suggestions (through the use of a "dogalogue" box as shown in Figure 2.7). As a fallback, Emma (an adult) is also included and enables the child to contact the staff of the information centre for further guidance and support, if the puppy cannot fulfil the child's request, or if they would like to physically borrow an item from the information centre. Figure 2.6 shows the different avatars in EmSe. In the following section, we provide a brief description of the different components with the Emse Search Service.

The Body Browser The Body Browser enables exploration of the patient information centre's repository of books, DVDs and other media via an interactive illustration of a body, where users can zoom to various levels of detail from the entire body to specific organs [202], which triggers medical Web searches related to the body parts and organs in focus. This metaphorical interface is expected to reduce vocabulary difficulties in the query formulation step, which are particularly salient with children and health information. Furthermore, it emphasises a browsing paradigm rather than requiring the user to search for information. For each selected body part, sub parts, related diseases and treatments are shown as query suggestions. The interaction has been kept deliberately simple: a point-and-click paradigm is used for both selecting and zooming in, with immediate feedback to make the functionality intuitive. In order to allow for greater personalisation of the search experience, the Body Browser, along with the child avatar can be selected to be either male or female. Figure 2.4 shows the Body Browser in its default zoom level.

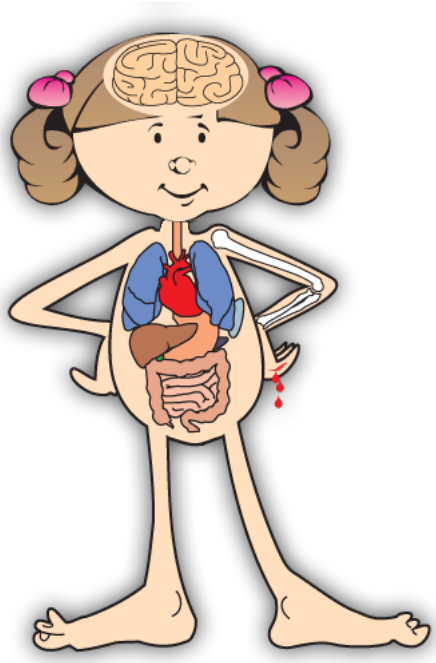

Figure 2.4: The Body Browser.

Query moderation The query moderation component identifies and enhances queries that are unlikely to yield content that is child-oriented by applying simple, real-time technology [85]. The focus is not explicitly on removing mature content, but rather on making the results of general queries more suitable for children. The component breaks up queries into n-grams and examines the children's appeal of these grams by checking Google Suggest for the co-occurrence of these n-grams with child-oriented modifier terms (e.g., "for kids"). We adapted this technique for the Dutch language ("voor kinderen"). In case of n-gram co-occurrences with known children's terms, the query is dispatched unmodified. Otherwise, it is appended with Dutch kids modifier terms ("voor kinderen") before being dispatched. Despite its simplicity, empirical evidence suggests this technique to result in good performance of ensuring child-friendly results. Additional filtering components can be configured to reject queries containing explicit language or undesired content requests.

Result moderation Even given moderated query formulation, returned result sets from a Web search engine may contain unsuitable items. A particular example from the medical domain are non-objective and sponsored pages authored or influenced by pharmaceutical or medical companies [206]. In order to address this challenge, EmSe offers the possibility to maintain a black list of known biased hosts or suspicious terms. Pages originating from such undesired hosts, or mentioning certain keywords can automatically be excluded from being displayed to the children.

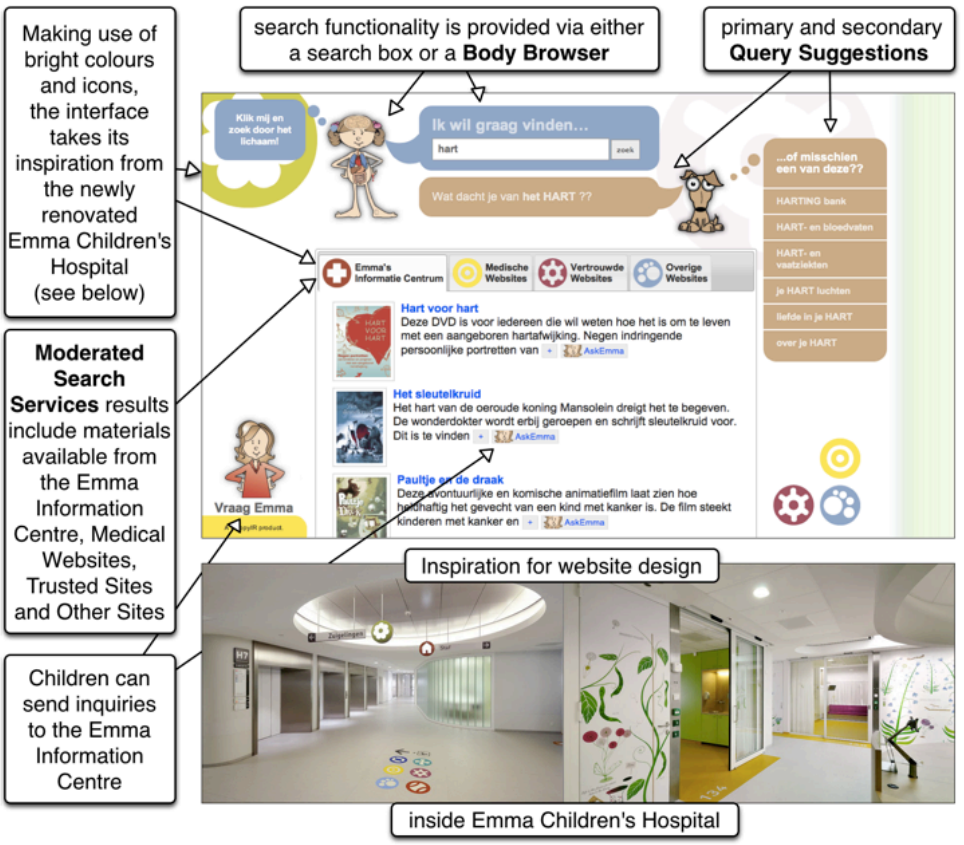

Figure 2.5: An annonated screenshot of the EmSe search service.

Query suggestions To further mitigate children's problems with the query formulation step, EmSe offers spelling corrections and query suggestions. They help children explore and query the recommended and related medical sites by providing suggestions that reflect the specific content on these sites. They are generated by extracting meaningful and informative phrases from the anchor text of the recommended resources.

Content Simplification Children often face significant difficulties understanding texts of advanced reading levels. Previous work described in [68] has developed techniques for identifying and subsequently filtering out difficult material. This approach, however, can result in poor coverage of certain, inherently difficult, topics. Manual inspection of resources with high reading level showed that often the text was understandable except for few occurrences of specialised terms. Instead of *a priori* rejecting difficult content, requested pages are checked for difficult terms, which the system augments with brief definitions. When hovering over the term with the mouse cursor, a tooltip containing the definition appears. In this non-intrusive way, the text can be read without the child having to leave the page to look up external information. The simplifications are accompanied by hyper links to local and Web searches if the child should be interested in finding out more about the term. The definitions can either be manually maintained in a server-side dictionary or can be dynamically extracted from external resources such as Wikipedia or Wiktionary. Figure 2.7 shows the interactive content simplification dialogue available for result page snippets as well as external pages.

Evaluation

The first version of EmSe was released in early 2012. It is accessible to staff and patients within the hospital (via bedside computers and other terminals), and also to out-patients via the Web[5]. The system evaluation consists of two main stages. **(1)** The first stage is to collect feedback from the staff at the hospital. From this initial study we refined the demonstrator to incorporate suggested changes before **(2)** obtaining patient feedback.

Stage 1: Staff survey

In March 2012, we visited the children's hospital and interviewed a total of 11 staff members in 8 individual sessions. Of these eleven, three were nursing staff, two were in-house

[5] http://wickham.dcs.gla.ac.uk:8080/hospital/

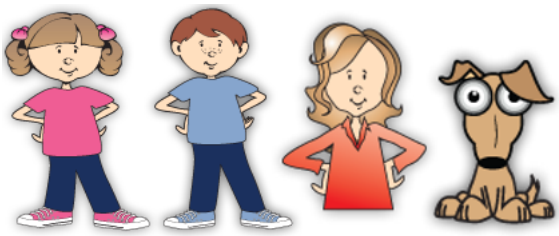

Figure 2.6: EmSe avatars: Girl, Boy, Emma and the Puppy.

Figure 2.7: Textual content simplification of “migraine” as “a bad headache”.

teachers, three pedagogues, two paediatricians and one member of the information centre staff. At the start of each session, the participants were shown a brief explanatory video, introducing the features and functionality of EmSe. Afterwards, they were encouraged to use the system themselves. Due to scheduling reasons, some participants interacted with the interface in groups rather than individually. The participants were encouraged to think aloud while operating the system and the interviewer did not interfere with their search activities unless directly asked for assistance or clarification. Afterwards, all participants were asked for their general opinion on the system and any aspects that they particularly liked, disliked, or missed. Finally, they were asked to comment on major aspects such as search result quality, interface design, the Body Browser and the content simplifications. Sessions typically took 15-20 minutes each.

There were a number of minor bugs as well as localization issues (the interface is Dutch and not all developers were Dutch native speakers) that were spotted in this early stage of evaluation. They have been addressed in the current version of EmSe. In the following, we will discuss the participants’ feedback concerning EmSe’s main components as well as additional features that were suggested and that remain to be investigated in the future.

Search result Quality The participant’s all felt the provided results were of high quality and presented in appropriate manner (i.e., results from the patient information centre first, as these are specifically designed for children, then results from recommended medical sites, before backing off to results from trusted Web sites and then all the Web). Two participants suggested that the distinction between results for children and adults should be more explicit. Either by separating out the general Web results not specifically designed for children from those that were, or by highlighting the more child-friendly results. Additionally, there was a concern that some topics, such as death or cancer, should only be reachable when explicitly being queried for. This form of search moderation represents one of the major challenges for future additions as it involves **a)** identifying sensitive topics that require special treatment, and, **b)** based on query, user context and session information, deciding whether or not to display sensitive information.

Interface design While all participants appreciated the clean minimalistic interface, they encouraged further personalization and integration of the user into the search process by adopting a login concept so that the user could be individually addressed by the system. This could for example happen in the form of the puppy avatar (representing

the search engine) calling the child by its name in the search prompt and the interactive steps such as query modification. We envisage that the avatars could be used to help establish a rapport between the child and the puppy – so that it could illicit information such as the child's name, how they are feeling at particular point in time, find out their favourite things (through a guessing game, etc.) so that the puppy could go off and find interesting and/or entertaining information for the child. An additional request for mobile device compatibility of EmSe was already partially addressed in the latest version which conforms with systems like the iPad. However, full functionality on small-screen mobile devices is currently not supported.

The Body Browser All participants liked the idea and implementation of the Body Browser. They considered it one of the key features to facilitate browsing-driven information discovery. However, there were a number of issues with this component. They suggested several additional anatomical structures (e.g., the appendix and the genital tracts of the children) to be included for greater coverage. The Body Browser opens as an overlay over the default EmSe screen. Three users were confused by this interaction style and would have preferred to have been able to more seamlessly switch between the Body Browser and the search results. We appreciate their concerns as the current solution aims to preserve the layout, whereas trying to integrate the Body Browser within the page (instead of on top of it) would require moving elements around and down. This would mean that to access results the child would have to scroll up and down (which introduces a similar problem).

Content simplifications All participants found in-line content simplifications helpful and non-intrusive. It was suggested to offer the simplification service independently of a term's expected difficulty in order to account for different user preferences and needs. More concretely, this would mean to not highlight difficult terms but also offer an interface through which arbitrary term simplifications could be requested on-line.

Stage 2: Patient Survey

The second stage of the EmSe evaluation was conducted on May 15th, 2012, involving patients of the EKZ. 6 patients of ages 10-17 were presented with EmSe and Microsoft Bing in order to solve a number of medical information finding tasks. All participants were Dutch native speakers and the interaction with the researchers as well as the search efforts were conducted in Dutch. At first, the participants received a brief introduction into the EmSe search system and its functionalities such as the Body Browser, suitable query suggestions or content simplifications. Subsequently, we compared the usefulness of EmSe with Microsoft Bing. Starting on one of the systems (The order was alternated, half of the participants started with Bing, the other half with EmSe) they were given first a closed-class question (simple factual "What is x?" type question), then an open-ended one ("List all information that you can find about x."). The same procedure was repeated on the second system (with different questions). Finally, each participant was asked to search for information concerning their own medical condition using their preferred choice among the two compared systems. Using a set of fall-back questions from both types, we made sure that no participant encountered questions concerning

their own medical conditions prior to this final question. To conclude each session, the participants were asked to fill a brief form in which they could describe their search experience. Individual sessions lasted for approximately 40 minutes. The researchers took notes concerning search behaviour and success during the full duration of the sessions and provided support where necessary, otherwise avoiding interference with the participants' search efforts. The questions offered belonged to the following set:

Closed class:

1. *"What is a port-a-carth?"*

2. *"What is a mic key button?"*

3. *"What is a PEG?"*

Open class:

1. *"What information can you find about cystic fibrosis?"*

2. *"What information can you find about Crohn's disease?"*

3. *"What information can you find about the Sickle-cell disease?"*

During the six sessions of the study, participants were generally positive about EmSe's look and feel as well as the quality of the presented results. The Body Browser was unanimously liked and in-line content simplifications were considered useful. 100% of participants chose EmSe to answer the final question about their own medical condition, irrespective of the order in which the two search engines were presented to them. Due to network problems, the server on which EmSe was offered, was not always responsive. Two participants commented on the system being slow. Query suggestions were frequently (50% of the sessions) confused with search results. The participants afterwards stated that they would have expected suggestions to be offered in form of automatic completions as done by several commercial search engines. Older participants (aged 13+) often found the interface too focused on young children. Since the target audience of EmSe are children of ages 8-12, this was to be expected. Especially for children with less developed reading and writing skills, the content simplifications were helpful. Interestingly enough, this feature was frequently used by older participants as well. The concluding questionnaire supported a number of previous assumptions, such as older children being more likely to be attracted by conventional search engines such as Bing. At the same time, a number of surprising observations could be made: **(1)** Older participants liked the in-line content simplifications more than their younger peers. **(2)** Bing was reported to be easier to use than EmSe. We attribute this to the participants already being used to the search interfaces and paradigms of conventional Web search engines. Every participant reported using Google for their day-to-day information needs. **(3)** Despite their familiarity with Bing's interface, EmSe consistently received higher agreement scores for the statements "I like this system" and "I would like to use this system again".

Ongoing evaluation efforts
In the first two evaluation stages, we gained a qualitative understanding of professionals' and patients' usage behaviour and wishes for modifications towards EmSe. In both stages, there was a strong consensus about the system's usefulness for children's information search. A number of shortcomings were discovered and either directly addressed or discussed for later integration into subsequent versions of EmSe. Most of these requests concerned the interface as users expected interaction paradigms they are used to from commercial Web search engines. For productive use of EmSe a closer orientation towards well-known behaviour such as displaying query suggestions in the form of automatic completions might be advisable. Greater numbers of participants are aimed to be reached in the analysis of user interaction logs (Evaluation Stage 3). They are needed to strengthen the conclusions drawn from the small-scale evaluations of Stages 1 and 2.

2.4. Conclusion

In Research Question 1.a), we hypothesised that there can be fundamental differences between the preferences and requirements that individual users or user groups pose towards Web search facilities. To demonstrate this effect on the concrete example of children using standard Web search services designed for adult users, we conducted a user study in a Dutch elementary school. Confirming the findings of previous work, we noted a wide range of deficits that hinder children's Web search success. According to the literature, these aspects do not represent challenges for most adults.

We discussed two potential solutions to the problem. Firstly, we present an automatic search success prediction scheme based on a wide range of linguistic and cognitive session features. When integrated into an electronic classroom environment, the teacher's attention can be efficiently guided towards those students that are struggling with an assignment. In an alternative approach, we attempted to construct a more appropriate search interface for children with medical information needs in a hospital setting. The EmSe[6] search service is comprised of a range of support features such as query and content moderation, non-textual query interfaces and on-line content simplification. Initial qualitative evaluation stages among medical staff as well as patients suggest the usefulness and adequacy of the proposed interface.

Both scenarios show document relevance subjected to personal context. We noticed that previously relevant documents were often missed because of the user's inability to adequately operate the search engine interface or to locate and understand the relevant information on the page. Given a different personal context, such as for example that of an adult searcher, search success rates would have been vastly different given the same retrieval system.

In the further course of this dissertation, we will revisit the examples described in this chapter. Especially in Chapters 3 and 4, we will discuss several ways of estimating personalized relevance dimensions, and, in Chapter 6 we will finally combine them in a single multivariate relevance model.

[6] EmSe is an open source project. Its code base (along with the underlying PuppyIR framework) can be downloaded via Sourceforge at: http://sourceforge.net/projects/puppyir/.

References

[3] Denise E. Agosto and Sandra Hughes-Hassell. "Toward a model of the everyday life information needs of urban teenagers, part 1: Theoretical model". In: *Journal of the American Society for Information Science and Technology* 57.10 (2006), pp. 1394–1403.

[20] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. "Knowing the user's every move: user activity tracking for website usability evaluation and implicit interaction". In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 203–212.

[27] Nicholas J. Belkin and W. Bruce Croft. "Information filtering and information retrieval: two sides of the same coin?" In: *Communications of the ACM* 35.12 (1992).

[32] Dania Bilal. "Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks". In: *Journal of the American society for information science* 51.7 (2000), pp. 646–665.

[33] Dania Bilal and Joe Kirby. "Differences and similarities in information seeking: children and adults as Web users". In: *Information processing & management* 38.5 (2002).

[34] Dania Bilal, Sonia Sarangthem, and Imad Bachir. "Toward a model of children's information seeking behavior in using digital libraries". In: *Proceedings of the second international symposium on Information interaction in context*. ACM. 2008, pp. 145–151.

[36] Christine L. Borgman et al. "Children's searching behavior on browsing and keyword online catalogs: the Science Library Catalog project". In: *Journal of the American Society for Information Science* 46.9 (1995).

[45] Robert Capra. "HCI browser: A tool for studying web search behavior". In: *Proceedings of the American Society for Information Science and Technology* 47.1 (2010), pp. 1–2.

[50] Chun Wei Choo, Brian Detlor, and Don Turnbull. "Information seeking on the Web: An integrated model of browsing and searching". In: *First Monday* 5.2 (2000).

[58] Sebastian Czajka and Sabine Mohr. "Internetnutzung in privaten Haushalten in Deutschland". In: *Ergebnisse der Erhebung* (2008).

[62] Pieter Dekker. "Children's roles in web search". In: *Master Thesis, Delft University of Technology* (2011).

[63] Allison Druin et al. "Children's roles using keyword search interfaces at home". In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 413–422.

[64] Allison Druin et al. "How children search the internet with keyword interfaces". In: *Proceedings of the 8th International Conference on Interaction Design and Children*. ACM. 2009.

[65] Sergio Duarte Torres and Ingmar Weber. "What and how children search on the web". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 393–402.

[68] Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries. “A combined topical/non-topical approach to identifying web sites for children”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 505–514.

[81] Richard Glassey, Tamara Polajnar, and Leif Azzopardi. “PuppyIR Unleashed: A Framework for Building Child-Oriented Information Services”. In: *In Proc. of the 11th Dutch-Belgian IR Workshop*. 2011.

[85] Karl Gyllstrom and Marie-Francine Moens. “Clash of the Typings”. In: *Advances in Information Retrieval*. Springer, 2011, pp. 80–91.

[87] Mark Hall et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD Explorations Newsletter* 11.1 (2009).

[100] Hilary Hutchinson et al. “How do I find blue books about dogs? The errors and frustrations of young digital library users”. In: *Proceedings of HCII 2005* (2005).

[101] Peter Ingwersen and Kalervo Järvelin. *The turn: Integration of information seeking and retrieval in context*. Vol. 18. Kluwer Academic Pub, 2005.

[104] Hannah Jochmann-Mannak. *Websites for Children: Search strategies and interface design*. Twente University, 2014.

[118] Carol C. Kuhlthau. “Inside the search process: Information seeking from the user’s perspective”. In: *Journal of the American Society for information Science* 42.5 (1991).

[119] Giridhar Kumaran and Vitor R. Carvalho. “Reducing long queries using query quality predictors”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 564–571.

[124] Andrew Large, Jamshid Beheshti, and Tarjin Rahman. “Design criteria for children’s Web portals: The users speak out”. In: *Journal of the American Society for Information Science and Technology* 53.2 (2002), pp. 79–94.

[128] Dutch Legislation. *Wet bescherming persoonsgegevens*. http://wetten.overheid.nl/BWBR0011468. 2000.

[129] Dutch Legislation. *Wet medisch-wetenschappelijk onderzoek met mensen*. http://wetten.overheid.nl/BWBR0009408. 1998.

[130] European Legislation. *European directive on privacy and electronic communications*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:201:0037:0047:EN:PDF. 2002.

[139] Sonia Livingstone and Leslie Haddon. “EU Kids Online: Final Report”. In: *LSE, London: EU Kinds Online (EC Safer Internet Plus Programme Deliverable D6. 5)* (2009).

[146] Gary M. Marchionini. “Exploratory search: from finding to understanding”. In: *Communications of the ACM* 49.4 (2006).

[147] Gary M. Marchionini. *Information seeking in electronic environments*. 9. Cambridge Univ Pr, 1997.

[158] Penelope A. Moore and Alison St George. “Children as Information Seekers: The Cognitive Demands of Books and Library Systems”. In: *School Library Media Quarterly* 19.3 (1991), pp. 161–68.

[162] Jakob Nielsen. “Kids’ corner: Website usability for children”. In: *Jakob Nielsen’s Alertbox* (2002).

[163] Ofcom. *UK children’s media literacy: Research Document*. http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf. 2010.

[173] Maja Pusara and Carla E. Brodley. “User re-authentication via mouse movements”. In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*. ACM. 2004.

[186] Tefko Saracevic and Paul Kantor. “A study of information seeking and retrieving. III. Searchers, searches, and overlap”. In: *Journal of the American Society for information Science* 39.3 (1988), pp. 197–216.

[195] Andrew K. Shenton and Pat Dixon. “Models of young people’s information seeking”. In: *Journal of Librarianship and Information Science* 35.1 (2003).

[202] Frans Van der Sluis et al. “Visual exploration of health information for children”. In: *Advances in Information Retrieval*. Springer, 2011, pp. 788–792.

[206] Parikshit Sondhi, Vinod Vydiswaran, and ChengXiang Zhai. “Reliability Prediction of Webpages in the Medical Domain”. In: *Advances in Information Retrieval* (2012).

[209] Aideen J. Stronge, Wendy A. Rogers, and Arthur D. Fisk. “Web-based information search and retrieval: Effects of strategy use and age on search success”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48.3 (2006).

[227] Raven M. C. Wallace et al. “Science on the Web: Students online in a sixth-grade classroom”. In: *The Journal of the Learning Sciences* 9.1 (2000), pp. 75–104.

[242] Ka-Ping Yee et al. “Faceted metadata for image search and browsing”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2003, pp. 401–408.

[244] Yusrita M. Yusoff, Ian Ruthven, and Monica Landoni. “The fun semantic differential scales”. In: *Proceedings of the 10th International Conference on Interaction Design and Children*. ACM. 2011.

3

Relevance in Situational Context

When we are no longer able to change a situation,
we are challenged to change ourselves.

-Viktor E. Frankl, Psychiatrist

In the previous two chapters, we showed empirical and literature evidence for the subjective nature of document relevance. In recent years, we have seen a strong emerging tendency towards personalizing users' Web search experiences in order to better account for the searcher's individual context. Context, in this case, is often understood as the searcher's previous search history, geo-spatial position, topical interests, and language or literacy background. Most of these are static or slowly-evolving user-specific properties and are typically captured by means of query and interaction log analyses. While personalization functionality has been shown to benefit retrieval performance e.g., [215], there exist significant situational factors that can influence performance and thus should be taken into account. White et al. [235] introduced domain expertise as one such factor. Depending on the topic searched for, an individual can display significantly different search behaviour based on their previous knowledge of the domain at hand.

In this chapter, we argue for the importance of including the situation-dependent context into the retrieval step. To demonstrate this, we will go beyond personal relevance and investigate instances of users straying from their search profiles to satisfy information needs outside their regular areas of interest or following strategies that differ from their regular behaviour. Such atypical information needs can often be triggered by external events (e.g., pending medical treatments, financial deadlines, or upcoming

vacations) that explain the unprecedented interest in a previously unseen domain. As an example, a user might in general favour easily-readable documents about sports and be confident in querying, selecting, and understanding this type of information. At the same time, they might display significantly different preferences and skills when pursuing a novel task, such as completing a particularly involved tax form. Due to static modelling of user profiles, atypical information needs are currently poorly represented by Web search engines. Personalizing atypical search sessions in the 'regular' way does not seem appropriate as it assumes topical and behavioural consistency with previous search sessions. Often, this is not the case for atypical searches.

Our discussion will be guided by the following three research questions: **(1)** What is the frequency, extent and success rate of users pursuing atypical information needs? Additionally, can we identify common types of information needs for which users diverge from their previous preferences? **(2)** How can we automatically distinguish atypical search sessions from typical ones? **(3)** Can we improve retrieval performance for atypical sessions by re-ranking search results in a typicality-aware fashion? Our investigations are based on manually-annotated log files of the Bing Web search engine.

Previous related work can be grouped into three categories: General search personalization efforts; Query log analysis with the goal of long-term user modelling; Investigations of user expertise and content readability.

Search personalization

A growing number of data sources, such as search history, manually or automatically created preference profiles, and social network information, are being exploited for personalizing the selection of results for users [172, 216]. Approaches to search personalization vary in types of features considered (e.g., language models, topical categories, links or other metadata such as reading level), the time frame chosen (e.g., short-term or long-term profiles), and how the profiles are used (e.g., for ranking or recommendations). Several researchers have shown how profiles that consist of topical representations of users' search interests can be used to personalize search. Gauch et al. [79] learned topical user profiles based on browsing history or search history. Speretta and Gauch [208], as well as Ma et al. [143] used topical profiles that users specified explicitly. In all cases, user profiles were compared with those of search results and used to re-order search results for individuals. Several research groups, including [86], [65], and, [31] have recently shown how topics from the *Open Directory Project's* (ODP) topology of the Web[1] can be used to filter and personalize search ranking for individuals. Haveliwala [91] proposed a topic-sensitive modification of the PageRank algorithm to allow for a direct, more focused scoring of the Web graph given a query's topic or a user's topical preference. Queries and Web content were automatically categorized using the ODP hierarchy in order to facilitate topic-sensitive scoring. Sugiyama et al. [210] employed collaborative filtering on users' observed Web search histories for profile building. They compared their approach to exclusively using browsing history and implicit feedback mechanisms, finding significant merit in the use of profile expansion via their proposed method. Teevan et al. [215] investigated the potential of re-ranking the top 50 search engine results based on previous user profiles. In particular, the authors explored alter-

[1] http://www.dmoz.org

native document representations for search personalization, finding that full-text representations outperformed models based on a selection of pre-selected keywords. Li et al. [132] developed a dynamic graph-based adaptation scheme modelling a user's general preferences for search personalization, while accounting for changes of interest by incorporating short-term browsing information. In a recent study, Goel et al. [82] analyzed the U.S. market's large-scale consumption of films, music, and Web search results in order to quantify the importance of the "long tail" of items in those respective categories of popular media. The authors found that the majority of users largely displayed standard tastes in most categories but showed some degree of eccentricity in choices. In this chapter, we will investigate a related notion, namely, that of atypical search sessions: cases in which users occasionally stray from their "personal mainstream".

Long-term user modelling

A special subclass of research on user modelling and search personalization is based on long-term profiles rather than focusing only on the user's immediate history. While being noisier than short-term profiles, this approach has the advantage of being able to detect niche interests or those that surface in long cycles. Matthijs et al. [150] captured users' 3-month Web history across multiple search engines and sites via a browser plug-in. The full resulting log files were used for result re-ranking and showed significant performance improvements over the native ranking of popular search engines. Furthermore, long-term user profiles served as reliable general descriptors of a user's interests. Tan et al. [213] presented a language modelling approach that interpolated immediate search history and long-term user profiles in order to improve retrieval performance. They found that short-term profiles contained more useful clues as to the current query's intent, but that adequately-weighted long-term information introduced further performance gains. White et al. [234] investigated the usefulness of short-, mid-, and long-term profiles for the task of predicting user interest in Web sites. The authors demonstrated that, depending on the type of information being profiled as well as the type of information need, different profiling durations could be optimal. Finally, Bennett et al. [31] showed how long- and short-term profiles could be optimally combined for effective search personalization. They found that long-term models provided the most benefit at the beginning of a session, while short-term models became more important for longer sessions. In Section 3.4 of this chapter, we will pick up this particular setting by Bennett *et al.* and will demonstrate that knowledge about session typicality can be used to mix long and short term models.

Expertise

Studies of Web search often distinguish between two types of expertise – search expertise (reflecting knowledge of the search process) and domain expertise (reflecting knowledge of the domain or topic of the information need). In one of the early comparisons of Internet information search behaviour and success, Holscher and Strube [96] examined both search and domain expertise. They reported that search experts displayed a richer set of skills, such as selection of tools, query formulation and relevance judgement than novice searchers. Also, experts were found to navigate search interfaces more efficiently. Beyond search skills, [217] showed that experts and non-experts follow different

strategies to obtain search results, depending on the task. White et al. [236] conducted a large-scale log analysis of the differences in search behaviours and success of search experts and novices. The authors found that experts generate different types of queries, have shorter and less branchy post-search browse trails, and are generally more successful than novices. We noticed similar disparities between individual searchers and search strategies in Chapter 2 where we investigated the search roles and success of elementary school students. More recent work has tried to model strategies of successful searchers. Ageev et al. [1] exploited this expertise-dependent difference in search behaviour by using a Markov chain approach to predict search success for a range of pre-defined search tasks based on the sequence of actions the searcher had undertaken in the session. One of their main findings was that searchers who are more successful are generally more active (e.g., more queries issued and results clicked) in a given time window. Aula et al. [21] analyse different characteristics of successful and unsuccessful search sessions. Based on a small qualitative lab study and a subsequent large-scale evaluation, they established a range of indicators for user frustration during search sessions that were not yielding the desired results. Most saliently, the authors report longer sessions, question-type queries, the use of advanced query operators and aimless scrolling on the results page for failing searches. We will revisit these findings in Section 3.3 to employ them for identifying atypical sessions.

Beyond the effect of general search expertise on success rates, other recent work has considered the searcher's familiarity with the search topic. Based on a large-scale query log analysis, White et al. [235] found significant differences between the search behaviour of domain experts and non-experts within the domain of their expertise (but not outside of the domain). The authors found that domain experts generate longer queries with more technical terms, have longer search sessions with more branches, and have greater success in satisfying their information needs than novices. Collins-Thompson et al. [53] investigated the use of reading level metadata for search personalization, finding that search ranking could be improved by taking into consideration the user's previous reading level preferences as well as the reading level coherence between a Web page and its result snippet. Kim et al. [110] followed up in this direction by jointly modelling reading level and topic preferences to describe users. Their so-called RLT profiles were used to distinguish domain experts from non-experts as well as to identify occurrences of "stretch" reading behaviour, i.e., when users go beyond their usual preferences to satisfy information needs. In a similar effort, Tan et al. [214] exploit notions of reading level and text comprehensibility for ranking popular answers on the Web portal *Yahoo! Answers*. According to the searcher's degree of domain expertise, simple vs. more technical answers were ranked higher. Following a related notion, namely that of content quality, Agichtein et al. [2] established an automatic classification framework for determining the quality of user-created information on a community QA platform. The authors demonstrate how a range of community signals and shallow content properties can be employed to estimate answer quality with human judge accuracy.

The research we present in this chapter extends previous results by: characterizing the extent to which searchers diverge from their long-term search profiles, and demonstrating how the ability to detect such atypical sessions can be used to improve search personalization.

3.1. Designing human-readable user profiles

As a natural first step towards our investigation of the importance of situational context on document relevance, we need high-quality assessments of session typicality. Ideally, the actual users who are being targeted for personalization would make the judgements. In practice, however, individual users are rarely available for collaboration or discussion. Instead, the research community typically relies on external annotators who first need to form a mental image of the user before being able to judge the quality of personalised rankings. This step, however, can be difficult and time-consuming as it requires an in-depth inspection of the user's entire search and browsing history in order to accurately account for their interests and preferences. Previously, Amato and Straccia [12] used topical user modelling for content selection in digital libraries. Their profiles focus on users' preferences in a number of domains such as document content or structure. [161] propose a hierarchical profile based on terms extracted from clicked documents. However, at this point, there has not been an in-depth exploration of how to generate compact, human-readable user profile representations. To this end, we propose a means of summarizing a user's Web search history into a compact, yet informative profile. Our profiles combine features that indicate topics of interest, representative queries, search context, and content complexity, to enable external judges to quickly form an accurate model of a user's interests and expertise. Previous work in personalized search motivates the attributes to include in profiles (specific queries, general topics and content complexity), and work in human-computer interaction guides the presentation.

1. A user's interests can be summarized by a set of **topics** – but the topics must have clear and consistent definition, and not be too broad or too specific [12]. Additionally, the **most dominant** topics of a user's interests should be clearly recognisable.

2. Past **queries** should be included in order to provide concrete examples of common information needs [215].

3. The session **context** should be available in order to better understand the intention that motivated a sequence of queries [38].

4. User profiles should be **concise** in order to enable efficient work flows. Additionally, the variation in length between profiles should be limited in order to make the required work load predictable [197].

5. Content **complexity** has recently been shown to be a strong signal for search personalisation [53]. User profiles should reflect the general complexity of content consumed by the user.

6. **Consistency** in how profiles and sessions are shown enables more efficient processing [197].

We aimed to accommodate all of these considerations into the design of our user profile representation. Figure 3.1 shows an example of the resulting user profile. To obtain topics, we classify each clicked Web search result into topical categories based on the Open Directory project hierarchy (ODP), as described by [30]. We use categories at the second level of the ODP tree (e.g. Science/Biology, Computers/Hardware) since

this provides a consistent, sufficient level of specificity. A profile consists of one line per frequently-observed topic in the user's previous search history. We include each category that accounts for at least 5% of the overall amount of clicked pages. In this way, we ensure all profiles have a predictable length of 1-20 lines of text, regardless of how active the user was in the past. For each topic, we also show the 3 most frequent previously issued queries associated with that topic. To assign a topic to a query, we aggregate the topical classification of all clicked search results for that query. For example, for the query "Apple", if a user visited two pages classified as "Computers/Hardware", we would assign that topic to the query. We then display the queries that were most frequently associated with that topic in order to represent typical search patterns given a user and a topic. To further help the annotator form a model of the searcher, all queries are formatted as hyper links leading to the search engine result page for that particular query so that the annotator can see the topical spread of results. Finally, we include an estimate of the complexity of textual content in the form of a heat map of resource reading level. We estimate the reading level for each clicked result on a 12-point scale according to [53] and average the scores of clicked results for each query. We then highlight the query in green if the average reading level is less than or equal to 4, in red if the estimate is greater or equal to 9, and in black if it is between these two levels. The resulting profiles have the added benefit that they can be applied to any profiling duration, ranging from a single query to months of search activity. This ensures conceptual conformity when, for example, comparing a single session with an extended period of previous activity. Figure 3.1 shows an example of such a condensed profile. Given the interests reflected by this profile, it is straightforward to judge the typicality new sessions. For example, a session related to boxing is easily identifiable as typical, while dental hygiene queries had not been encountered before.

3.2. Data Exploration

As a starting point for our investigation, we begin with an analysis of real search sessions to get insights into the problem domain. Our data set originates from the proprietary log files of the commercial Web search engine Bing. Our analysis focuses on a 4-month period of query logs from January to April, 2012 submitted by English-speaking U.S. users. We refer to the respective time spans as M1 (January) through M4 (April). Throughout this paper, we will use M3 as our profiling period and M4 to test for atypical sessions. Later, in Section 3.3, we will also investigate the usefulness of prolonged profiling periods, using M2 and M1 in addition to M3. To gain a first, qualitative insight into the domain, we limit our scope to the 200 most active users (those with the highest number of queries submitted) to ensure that users are well represented by their profiles. Together, they submitted a total of 679,808 queries in 67,812 sessions. Session boundaries are drawn based on a 30-minute threshold of user inactivity as suggested by several previous studies (e.g., by Fox et al. [74] as well as Gao et al. [78]). Since we are concerned with information seeking behaviour, we exclude navigational queries from our inspection in order to get a clearer impression of the difference between normal and atypical *informational* queries. To this end, we employed a list of frequent navigational queries as well as structural heuristics to detect queries encoding domain names or URLs (e.g., those starting with "www." or ending in ".[domain]"). After this pre-processing step, 370,844

Figure 3.1: An example of a condensed topical user profile, a typical session (Session 1) and an atypical session (Session 2).

queries and 44,059 sessions (55% and 65%, respectively, of the full data set) remained for investigation. On average, individual users submitted 464 queries in 55 sessions per month. Figure 3.2 shows the observed distribution of session counts across users for M1.

Annotation

In order to get a deeper understanding of the extent and frequency of atypical search sessions, we collected manual labels for all search sessions from M4. Previous work by Jones and Klinkner [105] and Kotov et al. [116] suggests that sessions tend to be topically coherent, often serving a single information need or task. Based on this observation, we labelled typicality at the session level in order to preserve as much search context as possible. The labelling was facilitated by means of a crowdsourcing effort. Prior to this step, all sessions containing *personal identifiable information* (PII), such as names, phone numbers, addresses, social security numbers, etc., as well as identical sessions for individual users, were manually removed from our data sample in order to protect the users' privacy and anonymity.

At first, each annotator was shown a condensed profile representing the users' previous search history during the profiling duration (M3). The judges were then presented with a single search session from M4. All queries, both in the profile as well as in the session, were presented as Web search hyper links to enable the judges to quickly explore the types of content to which the search led. Analogously to the user profile, session queries were color-encoded by reading level. A short survey probed two main aspects of a session: 1) its typicality for the user, and 2) the degree of importance that the search

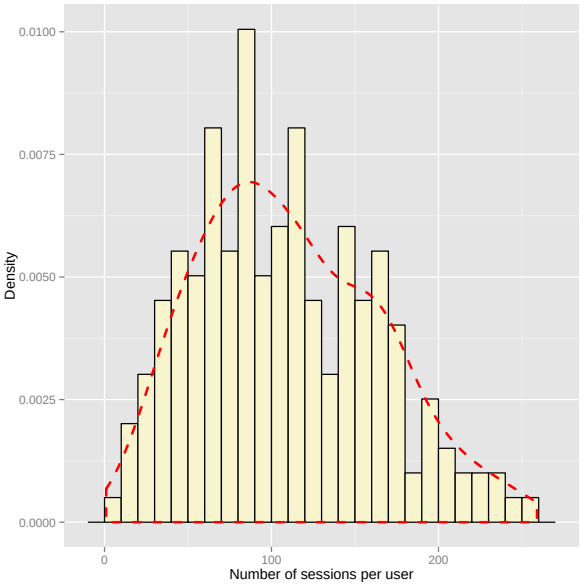

Figure 3.2: Distribution of session counts across unique users in January 2012.

task shows. The latter is interesting as we assume that atypical search sessions may often relate to important tasks or problems. The survey questions were:

1. *"How typical is this session for the user whose profile was shown above?"*

2. *" Now go back to the list of queries in the session above, and select all queries that support your decision."*

3. *"For the queries you selected in #2, do you think the desired information has high importance to this user? (e.g., likely to have lasting value, or help solve an important problem.)"*

Questions 1) and 3) were rated on a 5-point scale. In addition to the main questions, the judges could give further feedback by means of a free text field. To account for subjectivity and inaccuracies of individual workers, each session was labeled by 5 independent judges. The final label was determined by averaging across the constituent judgments. One of the researchers broke ties that could occur when individual judgements were rejected as the worker had flagged their decision as "*unsure*". The task was offered as a privately sourced task on the crowdsourcing platform Clickworker[2] at a pay level of 5 cents per session, a rate comparable to those suggested in related studies [9]. A grand total of $500 was invested in the label acquisition for all 2790 informational search sessions in M4. We followed accepted practices on the design and quality control of crowdsourcing studies by [107] by employing a hand-labeled set of gold standard tasks

[2] http://www.clickworker.com

as well as measuring agreement between judges in order to discourage low-quality submissions. We will discuss crowdsourcing, its potential and challenges as well as means of dealing with the sometimes high amounts of spam in Chapter 5. We computed several profile-based features for each assessed session:

1. The number of queries in a session (sessionQueryCount)
2. The entropy of the profile's topic distribution (userProfileEntropy)
3. Identical query overlap between queries and the full user history (overlapH-S)
4. Term overlap between queries and the full user history (overlapH-S-Terms)
5. Identical query overlap between queries and the condensed profile (overlapP-S)
6. Term overlap between queries and the condensed user profile (overlapP-S-Terms)
7. Term overlap between the condensed profile and the full history (overlapP-H-Terms)

Table 3.1 summarizes the Spearman rank correlations observed between these profile features and judging features. All overlap features had positive correlation with average typicality rating, the strongest correlation was found between profile and session terms (overlapP-S-Terms, +0.39). In addition, increasing the overlap of identical queries between condensed profiles and new sessions improved inter-rater agreement (overlapP-S-Terms is positively correlated with inter-rater agreement +0.24). High-overlap sessions were evaluated faster (-0.24 correlation of overlapP-S-Terms vs. judging time). In general, user profile-based features had a stronger influence on typicality scores and rating efficiency than their counterparts based on the full history. We also found that sessions from highly-focused users, whose profiles were dominated by just a few topics (low userProfileEntropy) were evaluated faster, with higher typicality scores and agreement. That is, the entropy of a user's profile was positively correlated with time spent judging (+0.25), negatively correlated with inter-rater agreement (-0.30), and negatively correlated with typicality (-0.29). Perhaps not surprisingly, the number of queries in a session (sessionQueryCount) was positively correlated (+0.41) with time spent judging.

The results showed substantial agreement between workers for the typicality vote. The standard deviation between each individual crowdsourcing worker and the majority vote among all 5 judges was found to be less than one point (0.854). To give an indication of the general task difficulty, we asked 3 expert judges to create redundant annotations for a subset of 100 sessions in a lab-based study using the same interface as the workers. Among experts, the standard deviation from majority votes was found to be even lower (0.495). Finally, we computed the overlap between majority votes from experts and those from crowdsourcing workers. In the vast majority of cases (82.6%) the two majority votes were identical.

Data analysis

Out of all 2790 informational search sessions labelled in M4, 166 were found to be atypical given the user's previous profile. Based on M3 profiles, 74% of all users showed at

Table 3.1: Spearman rank correlation of user profile/session features (rows) with judging features (columns). Judging features included average typicality scores, agreement on typicality, and average time to judge.

Profile features	Judging features		
	Typicality Average	Typicality Agreement	Average Time Spent Judging
overlapH-S	+0.10	+0.09	-0.14
overlapH-S-Terms	+0.32	+0.28	-0.16
overlapP-S	+0.24	+0.10	-0.17
overlapP-S-Terms	+0.39	+0.24	-0.24
overlapP-H	+0.37	+0.24	-0.19
sessionQueryCount	-0.07	-0.10	+0.41
userProfileEntropy	-0.29	-0.30	+0.25

Table 3.2: Properties of typical and atypical sessions.

Type	freq	$\frac{queries}{session}$	$\frac{terms}{query}$	$\frac{unique\ terms}{session}$	SAT. dwell time	SAT. rank	RL	SAT. RL
typical	2624	6.26	3.10	8.93	209 sec	1.5	5.4	3.9
atypical	166	6.69	5.23*	16.07*	180 sec	1.8	5.8	5.3*

least one atypical search session in M4. On average, each user displayed 5.9 atypical sessions which comprised 7.5% of their overall monthly query volume. Figure 3.3 shows the distribution of atypical session counts across users. While atypical sessions can be observed for most users, their frequency differed across searchers. The average user was largely coherent in search behaviour except for occasional atypical sessions, which is consistent with what [82] observed as well. Some, however, regularly explored different topics, making their search history topically very diverse.

As a next step, we compare typical and atypical sessions based on a number of session-level properties. Table 3.2 shows a juxtaposition across the whole user population. Statistically significant differences between session types are denoted by an asterisk. Significance was tested using a Wilcoxon signed-rank test ($\alpha < 0.05$). Both groups show comparable session lengths with only a slight increase in number of queries submitted in atypical sessions.

As we turn to the queries, however, we observe significant differences. Atypical sessions show longer queries (5.23 vs 3.10 terms/query) and also explore the result space more broadly by employing almost twice as many unique terms as regular sessions (16.07 vs 8.93 unique terms/session). Since explicit relevance judgements are not available, previous work frequently accepts clicked results on which the user dwells for at least a

Table 3.3: Coherence with previous session statistics per user.

Type	freq	$\frac{queries}{session}$	$\frac{terms}{query}$	$\frac{unique\ terms}{session}$	SAT. dwell time	SAT. rank	RL	SAT. RL
typical	2624	-0.21	-0.11	-0.80	+15 sec	+0.19	-0.70	-0.40
atypical	166	+0.43	+1.70*	+1.55*	-21 sec	+0.49	-0.09	+1.80*

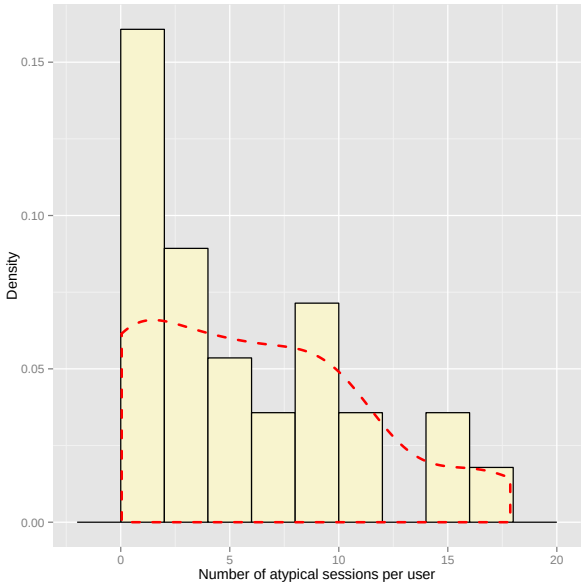

Figure 3.3: Distribution of atypical search session counts across unique users in April 2012 (M4).

threshold amount of time, as relevant given the underlying information need. According to [74] and [78], a commonly-used threshold is a dwell time of at least 30 seconds. We investigated these so-called *satisfied* (SAT) clicks in terms of their dwell time and the rank on which the user clicked. For typical sessions, such clicked pages show significantly longer dwell times (208 vs 180 sec.) and a better user satisfaction in the top ranks (average SAT click rank position of 1.5 vs 1.8) compared to atypical sessions. Finally, turning to textual complexity, we find comparable reading levels of clicked resources for both session types. However, SAT-clicked material in atypical sessions had significantly higher reading difficulty (average level 5.3) than in typical sessions (average level 3.9). In conclusion, atypical sessions show a number of established indicators of sub-optimal search experience such as short dwell times, long clicks in the bottom part of the ranking and significantly increased textual complexity of pages with long clicks.

In order to rule out the effect of individual users' querying style on these numbers, we now compare typical and atypical sessions per user. To enable this, we collected session statistics for all users in M3 and report the divergence from those observations for M4. Table 3.3 shows the population-wide averages of this comparison. Generally, typical sessions conform more closely with the user profile than atypical ones. While the differences are less marked than in the global comparison, we can note similar tendencies. For atypical sessions, queries are longer on average (by 1.7 terms) compared to the user's typical sessions, and use a wider selection of unique terms (an average of 1.55 more). Dwell times shrink and rankings are somewhat less optimal. Again, we note a higher textual complexity of documents that satisfy atypical information needs. These distinct characteristics of typical and atypical search sessions are also reflected in the resulting

Table 3.4: Prominent topics in atypical sessions.

Category	atypical freq.	typical freq.
Medical	49%	3%
Computers	21%	9%
Crafting	7%	3%
Cooking	5%	5%
Pets	4%	2%
Administrative	4%	2%
Travel	3%	7%
Other	7%	69%

retrieval performance. An indication of this trend can be found in the number of sessions that are abandoned without a single user click. This occurs 17% more frequently for atypical sessions than for typical ones.

Finally, we are interested in the content and cause of atypical search sessions. We manually grouped the 166 atypical sessions from M4 according to their high-level topic. Table 3.4 shows an overview of the most prominent resulting categories. Almost half of the atypical search sessions are concerned with health and medical information. Queries in these sessions are often dedicated to getting advice on healthy diets or finding information about causes and cures for certain medical conditions. Technical and computer queries are another major reason for atypical sessions. Typically computer problems such as viruses, or requests for help on diagnostic procedures, can be found in this group. The remaining 30% of atypical queries are distributed across a wide range of topics. Instructions for claiming taxes, preparing foreign recipes, or caring for pets were among the most prominent queries. To set these numbers into perspective, we contrast them with the overall frequencies of the respective categories in the same period. Interestingly enough, we note that many of the dominant topics in atypical sessions occur at significantly different frequencies than in the global collection. For example, medical and technical queries are significantly less common in the set of general queries. On the other hand, generally popular topics such as inquiries about celebrities, sports, or films are rarely found in the set of atypical sessions.

With respect to our first research question, we conclude that atypical Web search sessions are events that affect the majority of users. Often, they occur when users seek advice on unfamiliar subjects outside of their topical area of expertise. As the users struggle with finding the appropriate keywords in the unknown domain, many affected queries are natural language questions, a class of queries that is known to often yield inferior result quality [28]. In Section 3.4, we will demonstrate that state-of-the-art personalization techniques achieve inferior results on atypical sessions.

3.3. Identifying atypical sessions

The previous section summarized the concept and extent of atypical search sessions. In this section, we turn towards automatically identifying atypical search sessions and queries in order to appropriately react to the different nature of the information need.

While, ultimately, it would be desirable to classify ongoing sessions to directly benefit retrieval performance, this first investigation of using typicality information for search personalization addresses sessions in a post-hoc fashion. To this end, we propose a two-step approach: First, we model user interests, preferences and querying style in the form of a profile, e.g. based on past sessions. Then, for new sessions, we measure the coherence with the existing profile. Based on the findings summarized in Section 3.2, as well as previous work, our classification scheme employs 2 distinct types of features: **(1)** Direct observations from the current search session. **(2)** Coherence of the current session with the user profile.

Session-level features

Session length Search sessions within a well-known topical area are typically shorter than those issued by users exploring a novel domain. In the latter case, frequent query reformulations can be expected as the user closes in on the desired information. To measure this effect, we consider the number of queries issued per session. In the previous section, this feature could not be confirmed to indicate session typicality when inspected in isolation. We include it in our classification scheme to test its validity in interplay with other features.

Query length Atypical queries, on average, were found to be longer than typical ones. We use the average number of terms per query as a feature.

Unique terms per session Previously, we saw that there are significant differences in how deeply typical and atypical sessions explore a given topic by using a wide or narrow vocabulary. We measure the number of unique terms per session as an indicator of topic exploration vs. focused search.

Question query ratio In our qualitative analysis, we observed that many atypical sessions contain natural language questions. To account for this fact, we measure the ratio of queries per session that contain at least one of the following question words: *What, Where, When, Why, Who, How*.

Advanced operator ratio Previous work by Aula et al. [21] on the nature of unsuccessful and difficult search sessions, found that struggling searchers tend to make more use of otherwise often neglected advanced querying operators. We denote the ratio of queries per session employing at least one of the following advanced operators: *AND, OR, NOT* and *literal text matches* indicated by quotation marks.

Position of longest query The query editing history has been previously reported to hold information about the success rate of a search session [21]. Successful sessions tend to end in the longest query, as the user has sufficiently narrowed down the scope of the result set. On the other hand, unsuccessful sessions often see several iterations of specifications and generalizations before the search is finally abandoned. In the latter case the longest query can be found in the middle of the search session. We employ this observation by considering the relative position of the longest query as a feature (i.e., the rank of the longest query divided by the overall number of queries in the session).

POS ratios Our analysis of atypical sessions showed a high number of natural language queries. In order to exploit this apparently different syntactic structure of regular and atypical queries, we apply *Part-of-Speech* (POS) tagging and note the relative frequencies of *nouns, verbs, adjectives*, and *miscellaneous constituents* (anything that could not be grouped into one of the previous categories). We assume that natural language queries will display a lower ratio of verbs and nouns but more of the "syntactic glue", such as prepositions, that fall into the miscellaneous category.

Clicks per query Previous work (e.g., [1] and [235]) found domain experts to be more active and to generally explore more results per query than non-experts. We measure the average number of clicks each query receives in order to account for different degrees of user activity and proficiency in the target domain.

SAT clicks per query Similarly to clicks per query, we consider the relative frequency with which SAT clicks (clicks with a dwell time of at least 30 seconds) occur. More frequent SAT clicks can indicate a better ability to formulate successful queries and identify relevant material in the result lists [96].

SAT click ratio In relation to the previous two features, we measure the relative number of satisfied clicks. A high ratio indicates efficient search behaviour with targeted clicks on relevant material. Atypical search sessions are expected to display comparably lower ratios.

SAT click dwell time In Section 3.2, we saw shorter dwell times on the results of atypical sessions. In order to measure the degree to which the current result list satisfies the user, we report the average dwell time of all satisfied clicks in the session.

Median SAT clicked rank Previously, we observed a difference in ranking quality for regular and atypical sessions. For the latter, the user was more often forced to visit lower ranks of the result list. We account for this difference by measuring the median rank per session on which a SAT click was registered.

Reading level When faced with an unfamiliar problem, users are not always able to maintain their usual preferences for (typically lower) textual complexity. Due to the novel domain, they might lack the necessary knowledge for finding adequate, yet easy-to-understand material. Alternatively, the domain might inherently be of a more complex nature. We follow up on this notion by measuring the average reading level (as estimated using the classifier described by Collins-Thompson et al. [53]) of all clicked search results per session.

SAT-clicked RL Similarly to the previous feature, here we only consider SAT-clicked pages. This distinction has been used in the previous section and was observed to separate regular and atypical search sessions better than considering all clicks.

Topical flags In the previous section, we saw that certain topics are more dominant in the group of atypical queries than others. To reflect this, we include a signal that indicates whether the current session serves, e.g. a medical information need. Since the actual distribution of topics underlying the user's information need is

unknown, we employ topical classification of clicked results and note the relative frequency at which we observe the following categories: *Medical, Computers, Crafting, Cooking, Pets, Administrative, Travel.*

Unique topics per session Exploratory sessions in a novel topical domain tend to be more diverse than regular ones. Again, we classify all clicked search results into ODP categories and report the number of unique categories per session as a measure of coherence and focus.

Profile-based features

For each session-based feature above, we compute a corresponding coherence feature with respect to the user's profile. More specifically, we compute the difference between the session feature for the current session, and its historical average value across a user's previous sessions. For example, the length of the current session, minus the average session length across the profiling duration, will give the session length coherence feature. Additionally, two new feature types are considered.

Query term coherence For each user, we collect frequency counts of all query terms during the profiling duration. For each new session we do the same. Both profile and session can now be projected into a vector space with one dimension per unique term and frequency counts as components. We measure vocabulary coherence in terms of cosine distance between previous query terms and the current session.

Topic coherence Analogously, we also measure coherence in terms of topics, using cosine distance across topical vectors to account for sudden changes in the general subject domain.

For a high-level understanding of the problem domain, we computed estimates of the informativeness of all previously presented features. Table 3.5 shows a ranking of the 10 strongest features (out of 34 total) according to *Information Gain* (IG) and a χ^2 test. The feature rankings produced by the respective methods are largely consistent, with a few swaps at lower ranks. The strongest feature overall was the difference in query length from the user's previous profile (query length coherence), followed by the absolute query length. A majority (7 out of 10) of the high-ranking features are directly based on query information. Additional important signals are those based on the reading level of SAT-clicked pages (SAT RL) as well as the estimated difference in page topics (topic coherence). We note a balanced mixture of session- and profile-based features in the top 10.

Classification

We applied the above features to the binary classification task of predicting whether a session was typical or atypical, relative to a given user's profile. We compared several different classification models, including support vector machines and various regression methods, using the Weka toolkit [87]. The data set was split into distinct stratified training (90%) and test (10%) sets, such that no unique user's sessions were present in both

Table 3.5: 10 strongest features for identifying atypical search sessions by information gain and χ^2.

Feature	Rank by IG	Rank by χ^2
query length coherence	1	1
query length	2	2
question ratio	3	4
verb ratio coherence	4	3
topic coherence	5	5
longest query position	6	8
SAT RL	7	6
SAT RL coherence	8	7
adjective ratio coherence	9	9
noun ratio	10	10

sets. The session labels were obtained as described in Sec. 3.2. Consistently, the best results on the training set were achieved by a logistic regression classifier that reached a final performance of $F_1 = 0.84$ ($P = 0.82$ and $R = 0.86$). When moving from the cross-validation setting on the training set to the previously unseen test set, we observed a final score of $F_1 = 0.74$ ($P = 0.8$ and $R = 0.68$). This number is close to a human annotator's accuracy of agreeing with the annotator majority vote label ($F_1 = 0.79$).

We investigated how the amount of previous search history used to compute features affected the classification performance in finding atypical sessions. Figure 3.4 shows cross-validation performance of the logistic regression classifier as a function of the number of search sessions per user that were used for building profiles. While performance rises quickly across the first sessions per user, scores level out between 18 and 20 sessions. At this point, no significant differences from the previously observed overall performance of $F_1 = 0.84$ can be observed. 95% of our users issued this threshold amount of 20 sessions across 14 days. Using longer search histories beyond this two-week threshold (e.g., from M2 and M1) for profiling did not result in statistically significant performance changes.

With respect to our second research question, we conclude that automatic classification methods based on direct session-level features and coherence-from-profile features can be effective at estimating a search session's degree of typicality. Additionally, we note that a few weeks of query logs (between 1 and 2 weeks) were sufficient to make reliable typicality decisions for an individual user.

3.4. Retrieval experiments

After examining properties of atypical search sessions (Section 3.2) and describing an automatic scheme for identifying them (Section 3.3), we now turn towards improving retrieval performance for atypical search sessions. Previously, we conducted our investigations on a sample of the 200 most active users. Now, we apply the insights gained from the qualitative setting to Web-scale retrieval tasks on a much larger dataset, as described next.

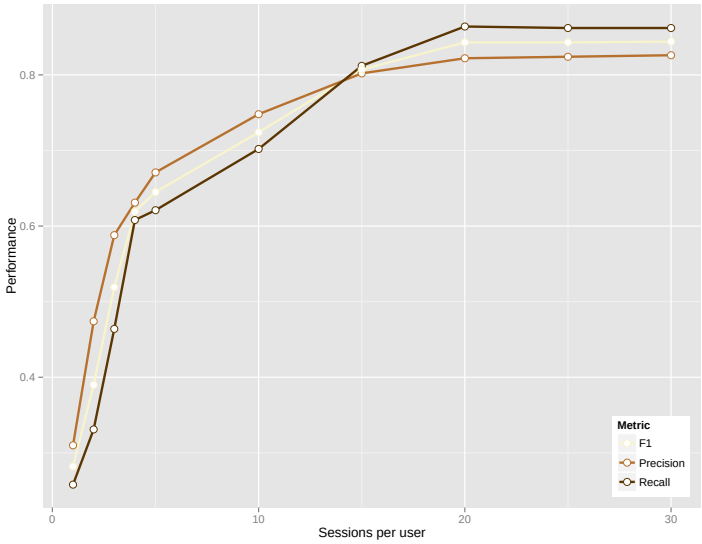

Figure 3.4: Classification performance as a function of the number of sessions per user in training set.

Method

Our study in this section closely follows recent research by Bennett et al. [31], and so we summarize that work briefly. In Bennett et al., the authors examined a rich family of modelling techniques for search personalization, looking at how different scopes of search and interaction history affected search personalization performance. They defined three key scope types: **(1) Session** considers all previous queries and actions from the current session. **(2) Historic** includes all previous actions and queries except for those in the current session. **(3) Aggregate** considers all previous actions before the current query. In addition to temporal extent, they considered several other factors used in earlier personalization research, including related queries and results in the profile. For each related query a score based on the weight of related queries, the similarity of the related results to the current result and the action taken on related results (e.g., click, skip) was computed and used to re-rank the current results. Then, they used a unified re-ranking framework that comprised the following components:

rel(q,u) For each new query q issued by user u, this function returns a set of related past queries q_r from the same user.

view(u) Denotes the temporal view on the user u. Different temporal views can limit the scope of $rel(q,u)$. Possible settings for $view(u)$ are: *session*, *historic* and *aggregate*. Their concrete definitions will be given later in this section.

w(q_r,q,u) Each related query q_r is assigned a relatedness score. We abbreviate it as w_{q_r}.

results(q_r) For each related query q_r, a set of previously returned results d_{q_r} is stored.

sim(d_{q_r}, d) This function determines the similarity between a current result for query q and past results for related query q_r.

action(q_r, d_{q_r}) The action the user took for the previously seen result d_{q_r} for query q_r. Actions, such as SAT clicks, can indicate document relevance (e.g., [74, 78]).

Given the above components, many commonly studied personalization features can be represented as triples of $< q, d, u >$. The feature score $f(q, d, u)$ presented by [31] is:

$$f(q,d,u) = \sum_{q_r \in R} w_{q_r} \sum_{d_{q_r}} sim(d_{q_r}, d)\, action(q_r, d_{q_r}) \tag{3.1}$$

where $R = \{q : q \in rel(q,u) \mid q \in view(u)\}$.

A key finding in [31] was that the *aggregate* context scope achieved better overall improvements in MAP of satisfied clicks than either *session* or *historic*. The authors noted that some sessions showed performance losses, which might be attributable to sessions in which users look for very different material than what they are usually interested in. We hypothesize that the atypical sessions studied in this paper are examples of this class of sessions. Thus, as a first step towards personalization of atypical search needs, we investigate the performance of the above personalization framework on typical and atypical search sessions.

Experiments

To ensure comparability of results, we identically replicated the original experiment setting used in [31], with the same underlying dataset. Our experiments ranged over an 8-week period based on logs collected in July and August 2011. The selection covers 155,000 unique users and 10.4 million sessions, with an average of 174.4 queries per user and 2.61 queries per session. All reported results are mean values across 5 stratified experiment folds.

The features derived from the above framework were used to train a LambdaMART learning algorithm [238] for re-ranking the top 10 returned results. The goal was to produce an optimized ranking, and, following Fox et al. [74], positive judgements were assigned to *satisfied result clicks* (SAT clicks). We estimated session typicality with the logistic regression classifier that was described in Section 3.3.

Table 3.6 compares the MAP re-ranking performance gains using *session*, *historic* and *aggregate* profiles over the original search engine ranking. We are also interested in the proportion of searches that were (positively or negatively) affected by the re-ranking. Consequently, we report the ratio of sessions whose MAP scores improve to those whose MAP score worsened. MAP scores are computed as the mean of average precision across the top 10 retrieved results. Cases in which the performance on atypical sessions differs significantly from that of typical ones are marked with an asterisk (determined via Wilcoxon signed-rank test at $\alpha < 0.05$-level). We confirmed the previous finding of Bennett et al. [31], that aggregate profiles lead to the highest overall performance gains for typical sessions. However, as hypothesised, atypical sessions show a very different trend. Session-level information yields the strongest gains, followed by aggregate information. Interestingly, re-ranking using historic (pre-session) profiles is worse than the original

Table 3.6: Personalization for atypical search sessions.

δ_{MAP}

	session	historic	aggregate
typical	0.0023	0.0047	0.0064
atypical	0.0067*	-0.001*	0.0059*

$\frac{\#improved}{\#worsened}$

	session	historic	aggregate
typical	1.56	1.26	1.48
atypical	1.79*	0.91*	1.5

Table 3.7: Session and historic information for search personalization.

	% improved	% worsened	$\frac{\#improved}{\#worsened}$	δ_{MAP}
session	3.32%	2.1%	1.58	0.00247
historic	3.53%	2.83%	1.25	0.00454
session/historic hybrid	4.11%*	2.6%	1.58	0.0055*

ranking for atypical sessions. All performance differences between different information sources for the same class of sessions (e.g., historic vs. aggregate information for typical sessions) are statistically significant.

We now address this difference between personalization performance for typical and atypical search sessions. Rather than uniformly applying one type of search history for personalization, we propose a hybrid approach that uses an initial classification step to predict whether the user is enacting a typical vs atypical session. Then, for all typical sessions, we apply historic personalization, and for atypical sessions, we rely exclusively on session-level information. Table 3.7 shows the overall performance gains of the proposed hybrid approach compared to both constituent methods in isolation. Significant improvements over **both** constituent methods are marked with an asterisk. Despite the relatively low frequency of atypical sessions, there are substantial gains in overall performance over the original search engine ranking. This tendency is also reflected in the case-based improvement and loss ratios. Atypical sessions see significantly more performance losses than gains when exclusively using historic profiles.

Finally, we conduct the analogous experiment for hybrid session-level and aggregate personalization. Table 3.8 shows the result of this alternative setup. We can note that the improvements over the original search engine ranking are consistently higher

Table 3.8: Session and aggregate information for search personalization.

	% improved	% worsened	$\frac{\#improved}{\#worsened}$	δ_{MAP}
session	3.32%	2.1%	1.58	0.00247
aggregate	4.9%	3.31%	1.48	0.00637
session/aggregate	4.83%	3.19%	1.52	0.00639*

than in the previous case. The gain of the hybrid method over uniform application of aggregate personalization shrinks, yet remains significant. This makes intuitive sense as aggregate histories already inherently contain session-level information. The ratio of improvements and performance losses remains largely stable. With respect to our third research question, we were able to obtain significant personalization improvements when identifying atypical sessions first and treating them differently from typical ones during re-ranking, by applying short-term session-level personalization rather than the historic or aggregated versions.

3.5. Conclusion

While previous work on search personalization has focused on the problem of matching content to a user profile, in Research Question 1.c), we hypothesize that even when a person's individual preferences are known, the situational context can have significant influence on document relevance. The original motivation, urgency of the information need, etc. may critically affect the searcher's relevance criteria. We demonstrate this by investigating atypical information needs. For such sessions, matching against the user's existing profile may *not* be accurate or desirable. Atypical searches are particularly interesting because in many cases they correspond to high-motivation needs in which the user exhibits a willingness to stretch their own boundaries for what is familiar or easy. Based on human labelling of "typical" vs. "atypical" sessions from several months of commercial search logs, we analysed topic, reading level, and session-level properties of atypical sessions. We found significant differences between typical and atypical sessions: certain topics such as medical information and technical support were much more likely to arise in atypical sessions, along with query features such as increased term count, more unique terms, and more natural language-type terms. We showed how atypical sessions could be successfully identified using a classification approach that combined session-level and profile-based features. Finally, addressing Research Question 1.d), we showed that the ability to identify atypical sessions results in significant performance gains for search personalization based on short- and long-term user profiles.

One important implication of the study presented in this chapter is that a user's motivation to succeed at a search, and the corresponding utility they place on finding the information, might be estimated in part *by the effort or risk they are willing to take to get the information*: an application of the classic von Neumann-Morganstern definition of economic utility. By "risk" we have in mind a compound quantity that captures both a) the uncertainty of relevance for the information sources the user is accessing, as measured by proxy quantities such as how "unfamiliar" or "new" a source is for that user, and b) the opportunity cost that the user perceives from accessing these unknown information sources with uncertain pay-off compared to accessing a known source with more certain pay-off. This is a different dimension of user effort than is captured by existing behaviour-oriented measures like user frustration, since it accounts for content-based factors such as the unfamiliarity and difficulty of the material being retrieved, and the quality of alternatives that may be available. We believe these connections to economic utility theory as well as related work on information foraging [171] could be a rich area for further exploration. This could for example be achieved by integrating information about the searcher's situational context into information *scent* models.

References

[1] Mikhail Ageev et al. “Find it if you can: a game for modeling different types of web search success using interaction data”. In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. 2011.

[2] Eugene Agichtein et al. “Finding high-quality content in social media”. In: *Proceedings of the international conference on Web search and web data mining*. ACM. 2008, pp. 183–194.

[9] Omar Alonso and Ricardo Baeza-Yates. “Design and implementation of relevance assessments using crowdsourcing”. In: *Advances in information retrieval*. Springer, 2011, pp. 153–164.

[12] Giuseppe Amato and Umberto Straccia. “User profile modeling and applications to digital libraries”. In: *Research and Advanced Technology for Digital Libraries*. Springer, 1999, pp. 184–197.

[21] Anne Aula, Rehan M. Khan, and Zhiwei Guan. “How does search behavior change as search becomes more difficult?” In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 35–44.

[28] Michael Bendersky and W. Bruce Croft. “Analysis of long queries in a large scale search log”. In: *Proceedings of the 2009 workshop on Web Search Click Data*. ACM. 2009, pp. 8–14.

[30] Paul N. Bennett, Krysta Svore, and Susan T. Dumais. “Classification-enhanced ranking”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 111–120.

[31] Paul N. Bennett et al. “Modeling the impact of short-and long-term behavior on search personalization”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 185–194.

[38] Pia Borlund and Peter Ingwersen. “Measures of relative relevance and ranked half-life: performance indicators for interactive IR”. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, pp. 324–331.

[53] Kevyn Collins-Thompson et al. “Personalizing web search results by reading level”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 403–412.

[65] Sergio Duarte Torres and Ingmar Weber. “What and how children search on the web”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 393–402.

[74] Steve Fox et al. “Evaluating implicit measures to improve web search”. In: *ACM Transactions on Information Systems (TOIS)* 23.2 (2005), pp. 147–168.

[78] Jianfeng Gao et al. “Smoothing clickthrough data for web search ranking”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 355–362.

[79] Susan Gauch, Jason Chaffee, and Alexander Pretschner. “Ontology-based personalized search and browsing”. In: *Web Intelligence and Agent Systems* 1.3 (2003), pp. 219–234.

[82] Sharad Goel et al. “Anatomy of the long tail: ordinary people with extraordinary tastes”. In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM. 2010, pp. 201–210.

[86] Karl Gyllstrom and Marie-Francine Moens. “Wisdom of the ages: toward delivering the children’s web with the link-based agerank algorithm”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM. 2010, pp. 159–168.

[87] Mark Hall et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD Explorations Newsletter* 11.1 (2009).

[91] Taher H. Haveliwala. “Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search”. In: *Knowledge and Data Engineering, IEEE Transactions on* 15.4 (2003), pp. 784–796.

[96] Christoph Hölscher and Gerhard Strube. “Web search behavior of Internet experts and newbies”. In: *Computer networks* 33.1 (2000), pp. 337–346.

[105] Rosie Jones and Kristina Lisa Klinkner. “Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs”. In: *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, pp. 699–708.

[107] Gabriella Kazai. “In search of quality in crowdsourcing for search engine evaluation”. In: *Advances in information retrieval*. Springer, 2011, pp. 165–176.

[110] Jin Young Kim et al. “Characterizing web content, user interests, and search behavior by reading level and topic”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 213–222.

[116] Alexander Kotov et al. “Modeling and analysis of cross-session search tasks”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 5–14.

[132] Lin Li et al. “Dynamic adaptation strategies for long-term and short-term user profile to personalize search”. In: *Advances in Data and Web Management* (2007), pp. 228–240.

[143] Zhongming Ma, Gautam Pant, and Olivia R. Liu Sheng. “Interest-based personalized search”. In: *ACM Transactions on Information Systems (TOIS)* 25.1 (2007), p. 5.

[150] Nicolaas Matthijs and Filip Radlinski. “Personalizing web search using long term browsing history”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 25–34.

[161] Nikolaos Nanas, Victoria Uren, and Anne De Roeck. “Building and applying a concept hierarchy representation of a user profile”. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM. 2003, pp. 198–204.

[171] Peter Pirolli and Stuart Card. “Information foraging in information access environments”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co. 1995, pp. 51–58.

[172] Pitkow, James and Schütze, Hinrich. “Personalized search”. In: *Communications of the ACM* 9.45 (2002), pp. 50–55.

[197] Ben Shneiderman and Catherine Plaisant. *Designing the user interface 4th edition*. 2005.

[208] Micro Speretta and Susan Gauch. “Personalized search based on user search histories”. In: *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*. IEEE. 2005, pp. 622–628.

[210] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. “Adaptive web search based on user profile constructed without any effort from users”. In: *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004, pp. 675–684.

[213] Bin Tan, Xuehua Shen, and ChengXiang Zhai. “Mining long-term search history to improve search accuracy”. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 718–723.

[214] Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. “To each his own: personalized content selection based on text comprehensibility”. In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 233–242.

[215] Jaime Teevan, Susan T Dumais, and Eric Horvitz. “Personalizing search via automated analysis of interests and activities”. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 449–456.

[216] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. “Potential for personalization”. In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 17.1 (2010), p. 4.

[217] Andrew Thatcher. “Web search strategies: The influence of Web experience and task type”. In: *Information Processing & Management* 44.3 (2008), pp. 1308–1329.

[234] Ryen W. White, Peter Bailey, and Liwei Chen. “Predicting user interests from contextual information”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 363–370.

[235] Ryen W. White, Susan T. Dumais, and Jaime Teevan. “Characterizing the influence of domain expertise on web search behavior”. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM. 2009, pp. 132–141.

[236] Ryen W. White and Dan Morris. “Investigating the querying and browsing behavior of advanced search engine users”. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, pp. 255–262.

[238] Qiang Wu et al. “Ranking, boosting, and model adaptation”. In: *Technical Report, MSR-TR-2008-109* (2008).

II

Relevance Measures

4

Automatic Estimation of Relevance

I never predict anything.
And I never will.

-Paul Gascoigne, Footballer

Enabled by the significant amounts of data available on the Internet, data-driven automatic classification methods are applied for a wide number of tasks. In this chapter, we will discuss 3 fundamental automatic means of determining relevance dimensions. In Section 4.1, we will present content-based methods that rely on the actual data in the document (e.g., a Web page's text, the visual content of an image or video, or, the audio signal of a piece of music). As a concrete use case, we will predict how suitable Web sites are for children of a given age based on the Web sites' content. Section 4.2 investigates the use of interaction data for estimating relevance dimensions. Concretely, we use the vast comment streams dedicated to shared videos on YouTube in order to estimate the topical relevance of the respective videos without taking into account the actual audio-visual content of the document. Finally, in Section 4.3, we close the circle by employing information on both content and interaction in the inference process. Here, we will determine the child suitability of shared YouTube videos.

4.1. Content-based Methods

The use of the Internet has become an integral part of most people's lives. This tendency also holds true for children who naturally adopt the behaviour that is displayed by parents and teachers. According to Wartella et al. [231], the age at which children have their first contact with the Internet has become consistently and significantly lower over the last years. Even adult users often struggle to cope with the amounts of information on the Web. Judging which content is relevant for their information need can be hard. Children face the same challenges, but have been reported to are at an even greater disadvantage. Schacter et al. [188] reported very young searchers to unquestioningly believe in any kind of information and to readily absorb new knowledge. While this in general is a desirable and essential characteristic that enables children to easily acquire large amounts of information such as new words, in the case of the Internet it imposes potential dangers. Several popular Web search engines such as Google and Yahoo! even state in their terms of service[1][2] that their search should not be used by minors. It is hardly possible to continuously supervise children's Internet usage in person as would be necessary in order to make sure that they are not exposed to inappropriate content. Recent surveys (e.g., [163]) found that approximately 40% of UK children aged 5-15 years regularly access the Internet without parental guidance or supervision. An automatic means of determining child-appropriateness of Web pages cannot replace a caring adult, but, considering the empirical evidence, any step towards a safer Internet for children can be considered highly desirable.

State-of-the-art Web search engines can greatly benefit from content-based methods of finding children's Web sites. At the moment children's search resources are typically directories of manually selected Web pages. As such, they offer high quality children's pages, but due to the high amount of manual labour involved in their curation, they are less flexible and have only a limited coverage in comparison to what an automatic approach could achieve. Examples of current Web search facilities for children are Ask Kids[3], or the recently discontinued Yahoo! Kids[4]. It is important to note that assuring content appropriateness exceeds filtering offensive material. Most state-of-the-art search engines offer safe search functionalities that reliably remove adult material. This kind of filtering usually takes a topical approach, which is already well-understood. Notions of text difficulty and age-appropriate Web site design however can be largely independent of the page topic and should strongly contribute to the decision of showing a certain page to a child. An automatic solution should take care of various aspects, assessing appropriateness in terms of topical relevance, textual content difficulty and presentation style. In this section, we will show content-based techniques can be used in order to make the multi-faceted suitability decision.

While automatic measures of content appropriateness and suitability are not very well explored, as early as in the 1950s, we can find advances by McLaughlin [152] into designing readability measures that would capture the difficulty of a natural language text. Collins-Thompson and Callan [52] investigated the use of language models to es-

[1] http://www.google.com/accounts/TOS
[2] http://info.yahoo.com/legal/us/yahoo/utos/
[3] http://sp.askkids.com/docs/askkids/
[4] http://Web.archive.org/web/20130115051736/http://kids.yahoo.com/

timate reading difficulty. They employed various modified naive Bayes models to represent different reader age groups. Schwarm and Ostendorf [194] applied support vector machines to determine a text's reading level, using a wider range of features such as for example linguistically-motivated information. The state of the art mainly focused on shallow features. Recent work on readability assessment now also employs grammar- and discourse-level features [42, 70]. Previous work by Bennett et al. [29], reported the Kids & Teens category to yield particularly low result scores. Classification based on the Open Directory Project (ODP) or similar Web topologies typically benefits from exploiting topic-inherent similarities within each topical category. Those occur less frequently in a heterogeneous category such as Kids & Teens which can deal with an arbitrary range of topics as long as the pages satisfy the suitability requirements. For this reason Gabrilovich and Markovitch [77] altogether excluded this branch due to its different structure. This tendency suggests that classification based on purely topical Web site aspects is not sufficient for the task of finding children's resources. The fact that children's search engines still heavily rely on manual classification strongly supports this hypothesis.

Non-topical Web classification has been a thriving field for applications. Kolari et al. [115] automatically identify Weblogs, Castillo et al. [47] use it for Web spam detection, and Liu et al. [135] employ it for sentiment analysis. In the course of this section, we will see that neither readability measures nor language modelling approaches are satisfying predictors of suitability when considered in isolation. We will introduce a more appropriate way of identifying children's pages by combining topical and non-topical aspects taking into account the specific needs of children.

Feature Extraction

Our experimental dataset was acquired using the Open Directory Project. In this Internet directory, Web pages are represented as leaves in a hierarchical topical tree. The Kids & Teens section of the ODP has human annotators categorize Web pages and, if appropriate, set one or several of the age flags {kid (≤12), teen (13-15), mature teen (16-18)} indicating that the page's content is suitable for those age groups. The ODP editors state in their content selection guidelines that a good children's Web page should be:

- informative
- age-appropriate
- non-commercial
- **for** children, not **about** children

We identified parallel categories between the children's branch and the general tree to select Web pages that deal with the same topic from an adult/ child perspective. This way we hope to reduce the topic-specific noise in the pages and to detect age-specific patterns that can be used for classification. The specific information needs of teens and mature teens are in the middle ground between those of children and adults. We kept only pages suitable for children in order to get clearly disjoint categories for this study. Since many of our features are language-dependent we concentrate on English resources by excluding the *World* and *Regional* branches of the ODP as suggested by for example [29]. We extracted a training set of 20,778 Web pages (6225 for children and 14553 for

adults) from a range of 1350 distinct topics. Among the kids pages, a share of 27% were deemed suitable exclusively for children.

Determining the child appropriateness of Web content is a novel task for which we explored a wide range of prospective features that reflect human notions of Web site child-suitability. To better understand what this means for Web pages we will discuss in detail the various criteria that qualify a "good" children's page. The two main dimensions of appropriateness are child-friendliness and focus towards child audiences.

The decision of what qualifies as child-friendly should ultimately be made by children. Based on recent studies on children's preferences towards Web sites [44], [124], [159], and, [162], we propose a range of child-friendliness criteria[5]. We try to formalize their findings, discussing child-friendliness in terms of complexity of text, presentation style, navigational as well as ethical considerations.

Complexity of text

There are significant differences between texts suitable for children who are inexperienced readers at best and those for adults. We expect child-friendly Web pages to rely on a language use and general design of textual resources that respect these specific aspects.

Shallow features (12) State-of-the-art text readability assessment often uses shallow characteristics as a measure of syntactic text complexity. There are commonly assumed to be significant differences in complexity between texts for children and adults. Examples of features in this category are the average number of words per sentence, the number of complex (3+ syllables) words, or the average word length.

Readability scores (8) According to Klare [112], a more high-level notion of syntactic complexity is delivered by automatic readability scores. Most of them are linear combinations of several shallow features that result in an age group whose members should be able to understand the evaluated text. Examples are the well-known ARI or Coleman-Liau measures.

Part-of-speech features (5) The features in this category are based on the notion of syntactic differences between adult and children's texts on a higher linguistic level than could be captured by mere shallow features. Statistics of POS tag distribution are generated for the textual page content. They include various POS parse tree statistics as, for example, the average number of noun phrases per sentence or the token/type ratio of observed words.

Entity occurrences (6) Commonly, children's text is not only syntactically, but also semantically, simpler than general text. Children's cognitive abilities are not yet suited for understanding complex texts, as for example newspaper articles, which often contain several different entities per sentence. Feng et al. [71] see this reflected by a smaller number of entities per article and per sentence in low reading level texts. We use the LingPipe toolkit[6] to extract the entity types person, location and organization.

[5] The numbers in brackets denote the number of distinct features within each category.
[6] http://alias-i.com/lingpipe/

OOV rates (8) On child-friendly Web sites we notice not only a simpler text structure in terms of shorter sentences containing less named entities, but also the use of more basic language. To reflect this difference, we constructed 7 distinct vocabularies of the most frequent/ most basic English words. We expect the out of vocabulary (OOV) rates of adult texts for these vocabularies to be higher than those of children's texts which commonly use a smaller and simpler range of words. We use an additional vocabulary of academic terms for which the opposite tendency is expected.

Wiktionary features (4) A possible way of capturing textual complexity makes use of the Wiktionary on-line dictionary. Ambiguous words which have a number of possible meanings (dictionary definitions) are supposedly harder to understand and use than unambiguous ones. The great coverage of Wiktionary should enable us to capture vocabulary difficulty and cognitive complexity of texts in a more universal way than mere OOV rates would allow. We use statistics on, for example, the average number of definitions or average definition length of a page's textual content to represent its cognitive complexity.

Presentation

Child-friendly Web sites should be presented in a way that is appealing to children. They should make use of the types of media that children, even at younger ages, are familiar with. While longer textual resources often cause frustration, the use of videos, images and colours in general have been shown to appeal to them. Large et al. [124] concluded that a page's content can be relevant and still children will not look at it unless its presentation is also attractive. In this work we will measure child friendliness of presentation through HTML page characteristics as well as visual features.

HTML features (10) Many high quality children's pages run a great number of scripts and animations in order to offer an interesting and accessible interface for an audience with limited reading abilities. We incorporate various HTML features as for example the tag distribution or the number of scripts on a page in order to capture the page's specific presentation style. Especially scripts were expected to be a strong indicator of child-friendliness.

Visual features (8) Child-friendly Web pages often rely on great numbers of images to convey their message. Especially for age groups that have not yet developed high literacy skills, visual resources are easier to understand. To further pursue this notion, we analyse the use of visual elements in terms of the number, type and size of pictures on the page.

Navigational

While the previous two aspects of child-friendliness were concerned with the page's content, navigational aspects target the embedding of the page into its link neighbourhood. Previous research in topical classification of Web pages by Qi and Davison [174] has found that a page's neighbours are strong indicators of its topic. However the assumption that pages on a given topic contain links to or are linked from other pages on that topic does not always hold. If we however transfer the same principle onto an age-scale

it may prove even more powerful as children's pages regardless of the actual topic should not link to non-child-friendly pages [86]. Large et al. [123] found that children prefer browsing over searching and generally trust links without closer inspection of anchor text. This exploratory search behaviour of children requires that safe children's Web sites do not contain links to Web pages for adults.

Link neighbourhood features (2) We use a basic version (without neighbourhood analysis) of our classifier to analyse a Web page's outgoing links. The share of pages that were classified as for adults/ for kids with at least a given threshold confidence are incorporated as features. Link analysis approaches sometimes take into account not only the page's immediate neighbourhood but also pages with a distance of 2 or more links. Since using more than one level of neighbouring pages did not yield significantly better results for our application, we restrict the neighbourhood to Web sites directly linked on the classified page to limit computational cost. Although previous work has analysed incoming links, we exclusively consider outgoing links. The reason for this is the hypothesis on which these features are built. While a children's page should not link to an adult page, there is no such limitation for an average page for adults.

Ethical

The final dimension of child-friendliness to be considered in this work is based on ethical considerations. The particular ethical concern lies in the presenting of advertisement to children who Nielsen [162] found to be less resistant to marketing strategies than adults.

Commercial intent (1) As stated in the ODP's content guidelines child pages should not be of commercial nature even though their products (e.g., toys) may be targeted towards children. We use the Microsoft AdCenter on-line commercial intent detection [59] as an indicator of suitability. Pages with a high likelihood of commercial intent are considered not child-friendly. It is important to note that there are two possible reasons for commercial intent on Web pages. A given page can either offer products or services itself or display advertisements which lead to actual commercial pages. Both variants are considered inappropriate for children as they clearly try to exploit their inexperience.

Focus towards children

Sometimes, Web pages are easy to understand and not harmful, although they do not convey the impression of being intended for children. These pages qualify as child-friendly according to all previously discussed dimensions but they are not targeted towards a child audience. To capture this second aspect of appropriateness, we will discuss what makes a page focused on children. Often this focus is already expressed through the choice of topic, and, some topics such as *colouring books* or *The Sesame Street* are mainly interesting for children. Focus can also be visible in the way the reader is addressed on the page. Web sites focusing on child audiences often employ a distinct style of addressing them (e.g., using child talk or diminutives).

LM scores (9) Language models have been widely shown to be strong representations of topical affiliation in Web scenarios. Using a language modelling approach, we hope to

capture the language use specific to children's Web pages. Textual resources from Simple Wiktionary[7] (18,206 entries), Simple Wikipedia [8] (108 articles) and Web page content for children (A subset of DMOZ pages topically disjoint from our training and test sets) are used to build up character-n-gram, uni-gram and token-n-gram language models for this purpose. We used the simple Wikipedia and Wiktionary versions because their more basic language use is easier to understand for young readers and thus closer to the language good children's pages will display. The language model score $P_{LM}(T|cat)$ is computed as the maximum likelihood estimate of the observed text T given the category's language model.

$$P_{LM}(T|cat) = \prod_{t \in T} P_{LM}(t|cat) \; P_{LM}(t|cat) = \lambda \frac{count(t,cat)}{|cat|} + (1-\lambda) P_{backoff}(t)$$

For each term, the number of occurrences within the category $count(t,cat)$, divided by the overall number of category terms $|cat|$ is computed. An interpolated character-n-gram model $P_{backoff}(t)$ serves for smoothing purposes in Jelinek-Mercer fashion with smoothing factor λ. Each page is scored against these models and the scores are used as features.

Reference features (2) In order to find children's pages, we localized cue words (considering all affixations of the terms "child" and "kid" and manually rejecting those terms that had no relevance to the children's domain, e.g., "kidney"). We analysed text windows of variable size around these terms. N-gram counts for the windows are collected. The notion behind using this approach is that there should be a difference in the way children are referred to on general pages as opposed to on child pages. On a general page about education or childcare we expect to observe higher frequencies of strings like "your child" or "the average child". Here children are **talked about**. The reference style should be different on actual children's pages where phrases like "for kids" or "us kids" in which kids are **talked to** are assumed to be more dominant. Finally the share of about-references and to-references are reported as features.

$$p(kids|page) = \frac{1}{|M_{n,page}|} \sum_{w \in M_{n,page}} p(kids|w) \; p(kids|w) = \begin{cases} 1 & if \; c_{rel}(w) > \delta_{threshold} \\ 0 & else \end{cases}$$

$$c_{rel}(w) = \frac{count(w,kids)}{count(w)}$$

Where $M_{n,page}$ denotes the set of text windows of size n around the page's cue word occurrences. $p(kids|w)$ expresses whether the term w is a to-reference. $c_{rel}(w)$ is the ratio of n-gram w's occurrences on children's pages versus its general frequency. $\delta_{threshold}$ is the threshold value of $c_{rel}(w)$. Only terms w that reach this threshold are considered relevant. Best results could be achieved for a window size of 2 words (one before and one after the actual cue word) and a $\delta_{threshold}$ of 0.66.

[7] http://simple.wiktionary.org
[8] http://simple.wikipedia.org

URL features (5) Well-designed Web sites put considerable effort into the choice of domain name. Good examples are www.pbskids.org or www.kidsdinos.com. Previous surveys by Large et al. [124] even found that children preferred pages with funny URL names. We inspect the occurrences of child terms within the URL by considering all its sub-strings that form valid terms contained in our simple Wikipedia vocabulary. The maximum likelihood estimate of these terms according to our children's text language model is incorporated as an additional feature.

Page segmentation

Previous research by Golub and Ardö [83] has shown page segmentation to be beneficial for Web page classification. We investigated its impact by splitting the Web pages into title, headlines, anchor texts and main text. For each of these segments, the above features were extracted and used as an extended feature space. Some features have to be considered on page level rather than on segment level. Therefore the HTML, URL and visual features, a page's commercial intent and its link neighbourhood are extracted per page rather than per segment. Using page segmentation, the original set of 80 features is strongly enlarged, yielding a 242-dimensional new feature space. A complete overview of all features used in this work can be found in Appendix 1.

Feature Analysis & Selection

Previously, we introduced a number of promising features founded on very different assumptions, which we assumed will capture the essence of what makes a good children's Web site. Our aim is, however, to gain a deeper understanding of what makes a Web site a suitable and valuable resource for children. Trying to comprehend the individual contributions of each of the features in a 242-dimensional space is however hardly possible for humans. Therefore, we employ different means of feature subset selection and feature/category evaluation in order to better understand how to automatically assess suitability.

Reducing the number of features is not just helpful for human interpretation of results, but often also beneficial for classification performance. Several state-of-the-art machine learning methods tend to perform sub-optimally when facing a large number of redundant or even irrelevant features.

An accepted and well-known estimator of feature performance is given by the mutual information I between features and the class label. It measures the reduction in entropy when introducing an additional feature. To get a first notion of the importance of the various features described previously, we rank them by their mutual information. Table 4.1 shows the top 10 features according to this ranking.

The strongest overall feature according to mutual information is the share of linked pages that were classified as suitable for children. This supports Gyllstrom and Moens' finding that Web page neighbourhoods should be homogeneous for good children's pages [86]. The high ranking of the number of occurrences of the term "kid" on the page as well as the child reference ratios for main text and title follow the requirement that children's Web pages should be targeted towards them. As required good children's pages will mention kids (kid term occurrence) and will address them (high child reference ratio) rather than talk about them. Finally, among the top-rated dimensions, we observe many fea-

Table 4.1: Top 10 features by mutual information

Feature	I
Child neighbourhood rate	0.050
Occurrences of "kid" in main text	0.026
Kid's 1-gram LM on title	0.021
Kid's 3-gram LM on title	0.016
Wiki 3-gram LM on title	0.016
Wiktionary 3-gram LM on title	0.016
To-references on title	0.014
Coleman-Liau on headlines	0.013
To-references on main text	0.012
Kid's character LM on title	0.010

Table 4.2: Avg. Mutual information by categories.

Category	Average I
Neighbourhood	0.0280
LM	0.0050
URL	0.0049
Reference	0.0043
Visual	0.0034
Shallow	0.0022
POS	0.0022
HTML	0.0020
Readability	0.0020
Entity	0.0013
Wiktionary	0.0010
OOV	0.0009
Commercial	0.0007

tures from the title segment and the category of language models. This suggests that, despite its relative brevity, the title of a Web page is a strong predictor of its suitability, and that strong indicators of suitability are encoded within the page's vocabulary use.

To further inspect the predictive potential of individual Web page aspects, we compare average mutual information scores per category in order to see which categories contribute most to the reduction of entropy. The results of this comparison can be found in Table 4.2. We can observe the same tendency that already emerged in the top 10 ranking of single features. Web page neighbourhoods remain by far the strongest overall category, followed by language model scores. Entity, Wiktionary and OOV features, at this stage, hardly add information that was not already expressed by the language models. At the bottom of the ranking we find the commercial intent score. This clearly surprised us as according to the DMOZ editors the children's section should not contain commercial Web pages. We inspected the data further and found that there are no significant differ-

Table 4.3: Avg. mutual information per segment

Segment	Average IG
Title	0.0057
Body	0.0041
Headlines	0.0032
Anchor text	0.0030

ences in commercial intent of children's (intent confidence $\mu = 0.31$ and $\sigma = 0.255$) and general pages ($\mu = 0.28$ and $\sigma = 0.27$). Since Dai et al. [59] have shown commercial intent detection to yield reliable results, we have to assume, that in spite of their content guidelines, the distribution of commercial intent among pages for children and grown-ups in DMOZ is rather arbitrary. Manual assessment of Web pages showed that the majority of commercial pages from the kids class published advertisement banners rather than actively offering products or services. This carelessness with respect to advertisements, however, is a key weakness of today's Internet for children.

Finally, the same comparison was conducted on a per-segment basis. The results can be found in Table 4.3. Again, we see a high importance of the title segment confirmed. The results presented here give an early impression of the predictive strength of individual features, categories and segments. However, as we will show in the following, a high mutual information score does not guarantee that a feature will end up in the best-performing subspace. Often the relevant information lies in the interplay of several features. Single-feature mutual information is not suitable to capture such synergies.

Feature subset selection

After the previous information theoretic inspection, we will now evaluate feature subsets' actual prediction performance in order to find strong indicators of child suitability. We employed a number of different state-of-the-art classification methods and found logistic regression to be the strongest overall method for this application. The results reported in this section will therefore generally refer to a logistic regression classifier trained on the varying feature sets and evaluated using 10-fold stratified cross validation. We will report precision and recall for this task as well as the $F_{0.5}$-measure and ROC area under curve. We decided for the precision-biased F-measure, as recall is desirable but precision is the crucial aspect for child-friendly Web search that aims to promote as few as possible false positive results.

As a first step to understanding the predictive power of different feature sets we follow the natural division present in the data, namely feature categories and Web page segments. We will inspect the individual performance of each division by excluding all other features for classification. The results of this analysis are shown in Table 4.4.

The tendency that was observed for information theoretic analysis in the previous section is repeated on classification scores. Link neighbourhood is still the strongest feature aspect, closely followed by language model scores. Most categories, regardless of their singular predictive power, add a small contribution to the overall score. The relatively weak performance of reference features is due to the low share of pages actually

Table 4.4: Classification performance per category.

Category	P	R	$F_{0.5}$	ROC
All features	0.79	0.55	0.73	0.78
Neighbourhood	0.76	0.54	0.70	0.75
LM	0.69	0.67	0.69	0.76
Shallow	0.74	0.45	0.66	0.71
HTML	0.73	0.45	0.65	0.71
URL	0.72	0.47	0.65	0.72
POS	0.66	0.50	0.62	0.68
Readability	0.66	0.47	0.61	0.66
Wiktionary	0.67	0.45	0.61	0.65
OOV	0.65	0.45	0.60	0.64
Entity	0.63	0.47	0.59	0.63
Reference	0.79	0.12	0.37	0.55
Visual	0.29	0.11	0.22	0.56
Commercial	0.10	0.05	0.08	0.50

Table 4.5: Classification performance per segment

Segment	P	R	$F_{0.5}$	ROC
Full page	0.79	0.55	0.73	0.78
Title	**0.78**	**0.49**	**0.70**	**0.74**
Body	0.70	0.45	0.63	0.69
Anchor	0.65	0.50	0.61	0.66
Headlines	0.62	0.37	0.55	0.59

mentioning children. While child mentions prove to be a valuable page characteristic when they occur, for most pages this notion is not applicable. When considering these results, it should be taken into account that the feature categories differ in size. Some of them contain 10 or 12 features while the commercial intent likelihood even has its own category. If we however assume that the features presented in this work capture the majority of the information a certain feature category holds, we can treat them as atomic. Under this hypothesis we can use the insights gained to discriminate those Web page aspects that contain most information of age-suitability.

Table 4.5 shows evaluation results per page segment. As in the IG comparison using one segment at a time, the page's title proved to be the most predictive one in general. This finding is very promising with respect to the on-line scenario. A computationally light classification based on Web page titles that, in case of low confidences, backs off to using the full page content would constitute an efficient way of satisfying response time requirements.

Previously, we inspected the information theoretic importance per feature. To evaluate information theoretically motivated feature subsets we ranked features by their mutual information with the class label constructing a feature subset out of the top n fea-

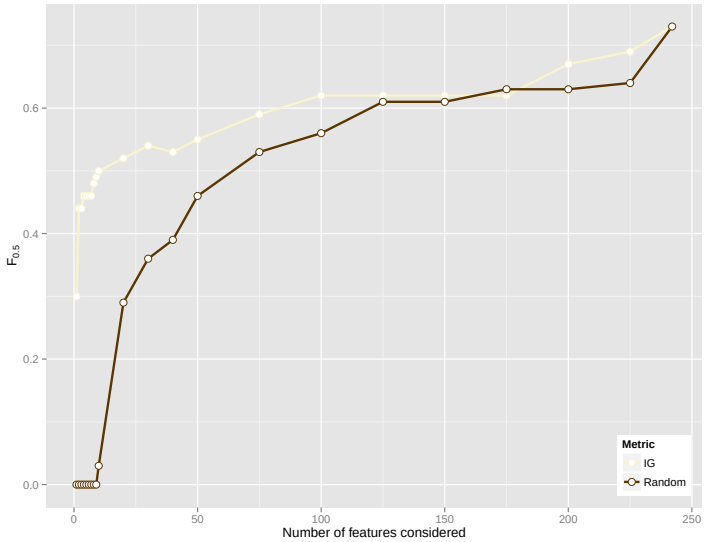

Figure 4.1: Performance of top n features ranked by mutual information for increasing n

tures for consecutively increasing values of n.

Figure 4.1 shows the performance of subsets in terms of $F_{0.5}$ scores in comparison to randomly drawn rankings of features. It can be observed that the early subsets constructed only of those features with the greatest mutual information scores significantly outperform randomly drawn subsets. As the number of features grows, however, the relative performance of the random sets approaches and even locally surpasses that of information theoretically selected features. This behaviour can be attributed to synergies between features not discovered by the current IG ordering.

Having shown that single-feature mutual information and subset selection by segment are capable of conserving most of the discriminative information while greatly reducing feature set size, we will now inspect the performance of automatically selected sets. Subset selection was done using the Weka[9] library's Genetic Search approach.

In difference to the natural subsets along the boundaries of categories or segments and those constructed on mutual information rankings, this approach yields feature spaces which can outperform the full set of features. This, again, supports the hypothesis that the suitability decision can hardly be made based on either only a single page aspect (category) or a set of strong individual features (IG). Evidently, a well-balanced combination of different features yields best results. The overall strongest set of features reached an $F_{0.5}$-score of 0.8 using cross-validation and is described in Table 4.6. This 16-dimensional feature space can be extracted and classified far more quickly than the full range of features, while additionally yielding better performance through elimination of redundant and irrelevant dimensions.

When inspecting the resulting feature set we notice that the general diversity of fea-

[9] http://www.cs.waikato.ac.nz/ml/weka/

Table 4.6: Best-performing feature subset

Kid link ratio	Number of words
Domain length	Total entities/unique entities
URL kid term score	Simple Wiktionary 1-gram LM
Script/word ratio	term freq “child” (headline)
Term frequency “kid”	number of words (title)
to-reference ratio	average word length (title)
Kid’s pages 3-gram LM	OOV Academic (title)
Average word length	kid’s 1-gram LM (title)

Table 4.7: Experiments with 10-fold cross-validation

Method	P	R	$F_{0.5}$	ROC
Random	0.50	0.50	0.50	0.50
Intuitive	0.82	0.05	0.21	0.56
LM	0.69	0.67	0.69	0.76
SVM	0.71	0.66	0.70	0.77
Classifier	0.85	0.64	0.80	0.83

tures is preserved. Most aspects of child-friendliness and focus towards children are present in form of members of the relevant feature categories. For all further experiments and evaluations in this work, we will assume the use of this feature set.

Classifier evaluation

After having selected a promising classification method and the feature space in which to represent Web pages, we will now measure our classifier’s performance in comparison with several baseline methods.

There has not been any specific research on the classification of Web pages specifically according to their suitability for different age groups. In order to evaluate our classifier’s performance this section will introduce two baseline approaches from similar Web tasks.

In the initial stage of our research we manually inspected different ways of how a human would tackle the task of finding child-friendly Web content with the common means of information retrieval at hand. An intuitive approach is to use a conventional Web search engine and expand the informational query by “kids”, “children” or similar cue words. This expansion method obviously does not work reliably all the time because Web pages for children are by no means bound to explicitly mention them. On the other hand, there will also be general pages talking about children while not being meant for a child audience. This method does however shift the focus of the results into the child-friendly domain to a certain degree. In order to exploit this intuitive notion, our first baseline method will be to interpret the presence of the terms “kid”, “child” and affixed forms thereof as an indicator of child suitability.

To construct a stronger baseline we apply an approach from text categorization as

Table 4.8: Experiments on unseen sample

Method	P	R	$F_{0.5}$	ROC
Intuitive baseline	0.80	0.04	0.17	0.53
LM baseline	0.61	0.57	0.60	0.69
SVM baseline	0.63	0.60	0.62	0.70
Classifier	0.72	0.71	0.72	0.76
Human performance	**0.76**	**0.72**	**0.75**	**0.79**

suggested by Liu et al. [137]. For this method we use unique terms as feature dimensions for an SVM classifier. Each dimension's value is the tf/idf-weighted term frequency. Stop words and terms that occur in less than 3 distinct documents within the collection are removed in order to keep the model from becoming too large. Our earlier findings on feature evaluation suggested language models to be powerful features. Because of term distribution statistics we expect this second approach to be a strong performance baseline.

Table 4.7 shows a comparison of the baseline methods and our classification approach. As expected, the fairly naive intuitive method achieves high precision (most general pages will not mention children) at low recall (Children's pages are not bound to explicitly mention them). This coverage problem results in a worse-than-random F-score. The SVM-based text classification approach performs solidly. Our classifier that combines topical aspects expressed by language modelling approaches with non-topical notions of suitability achieves best performance for cross-validation.

Evaluation on unseen data

Our test collection is a set of 1800 Web pages listed in the ODP containing 900 instances for children and 900 for adult audiences. The pages were randomly sampled from the English part of the directory (again excluding the *World* and *Regional* branches) and ensuring disjointness with the training data. Aside the ODP annotation, we had the test set additionally judged by external human annotators. The suitability decision is a highly subjective one. Using the overlap of several independent judgements will help us to reduce this degree of subjectivity. Furthermore we were able to collect information beyond the page's mere suitability, that we will use for a more fine-grained analysis. For each of the 1800 pages at least 5 independent human judgements were collected through the crowdsourcing platform CrowdFlower [178]. In Chapter 5 of this thesis, we will discuss human computation and in particular crowdsourcing, as well as some relevant considerations to take into account in greater detail. All further results reported refer to the performance on predicting the label assigned by the majority of CrowdFlower judges. 53% of the participants stated they were helping children with Web search on a regular basis. An additional 33% do so less frequently. 49.57% said to have children themselves. Based on these numbers we are confident that their judgements represent sensible notions of suitability.

Table 4.8 shows the performance of our classifier in comparison with the baseline methods as well as human judges for unseen test pages. The naive baseline approach

Table 4.9: Testing on pages exclusively for kids

Method	P	R	$F_{0.5}$	ROC
Intuitive	0.81	0.05	0.19	0.58
LM	0.66	0.58	0.63	0.71
SVM	0.68	0.61	0.66	0.73
Classifier	0.77	0.68	0.75	0.78
Human performance	**0.79**	**0.75**	**0.78**	**0.81**

that simply considers pages mentioning kids as suitable achieves high precision but due to the number of suitable pages not explicitly mentioning children its low coverage makes it the weakest overall method. The SVM classifier consistently proved to be the next better approach. Our classification method was able to perform significantly better than both baseline methods (with a relative improvement of 14% over the strongest baseline) and came close to human performance. We were able to outperform both baseline methods at $\alpha < 0.05$ significance level. (Determined using Wilcoxon Signed Rank Test.) Inspecting human judgement behaviour gives further evidence for the task's high degree of complexity. We observed an agreement ratio of 68% for the suitability decision among independent CrowdFlower workers. The decision whether a Web site deals with sports appears to be far easier to make than the one whether the page is suitable for a child of a given age. While topical classification tasks are mainly objective, the suitability decision involves subjective factors such as personal understanding of children's needs and ethical aspects.

Besides determining the actual classifier performance, we will answer the following four research questions: 1) Is it easier to distinguish pages for age groups that are divided by a broad age margin? 2) Does the performance of our method depend on the Web page's topic? 3) Does the page's quality have an impact on classification performance? 4) Can we make the suitability decision for pages independently of their language?

Age margin analysis

For our previous experiments we considered all pages that are suitable for children but at the same time might be interesting for teenagers. Now, we will alter the test set by using only those pages that ODP considered exclusively suitable for children. For this category, we observed inter-judge agreement ratios of 71%. This finding supports the hypothesis that a bigger age margin between the classes makes the decision easier for humans. Table 4.9 shows the automatic approaches' performances for distinguishing pages exclusively for children from those for adults. Notice that only the target set for evaluation is restricted, not the training process.

While this change in experiment set-up does not affect the ranking of methods for the task, it consistently raises performance of all approaches. To further increase the age margin between the classes we asked our CrowdFlower judges to additionally judge every page's suitability for very young children (aged 3-6). For this final experiment we rejected all kids' pages that were not deemed appropriate for young children.

Table 4.10 shows that the task of distinguishing pages for very young audiences from

Table 4.10: Testing on pages for young kids

Method	P	R	$F_{0.5}$	ROC
Intuitive	0.84	0.08	0.29	0.60
LM	0.70	0.63	0.68	0.75
SVM	0.73	0.67	0.72	0.77
Classifier	0.80	0.76	0.79	0.82
Human performance	**0.86**	**0.84**	**0.86**	**0.84**

general ones experiences another boost in performance. As an answer to our first research question, both experiments show how the information needs of different age groups become easier to discern for broad age group margins.

Topical analysis

An error analysis in the early stages of this work showed us not only that the suitability decision is often even difficult to make for humans, but also, that there are topics that are in general harder to classify than others. We found that our classifier had particular problems with scientific and biographical content. Pages from these areas, even if they are deemed suitable for children, often use a rather complex vocabulary and also focus more on pure information than appealing presentation. One might argue that such pages will only be relevant to a small group of children who are really interested in such topics, but the challenge of correctly dealing with these pages remains. To get a more precise notion of our observation we analysed topical difficulty for human judges and our classifier. Page topics on which the inter-annotator agreement is very low, imply a hard decision. Table 4.11 shows examples of topics that proved to be especially hard or easy. We can see that among the branches with high agreement scores we find typical children's and grown-ups' topics respectively. Examples include "English grammar" or "school". Those branches with low agreement rates often belong to complex topics. Even if they were written and presented in a child-friendly way many annotators rejected them because they doubted the topic's general suitability for young audiences. Examples are "history", "geography" or "mathematics".

Since the same notion of topical difficulty that humans face might also apply for automatic approaches, we determined the correlation between average inter-annotator agreement and our classifier's decision confidence. We found a Spearman Rank Correlation Coefficient $\rho = 0.58$ between the two dimensions. Figure 4.2 shows the distribution of classifier confidence scores and human agreement per page. To answer our second research question, we note that although it is a rather weak correlation, there is apparently an underlying topic-dependent difficulty that applies to humans as well as automatic approaches. Inspecting the figure, one can note a cluster of pages with high human agreement ratios (0.8-0.9) and at the same time low classifier confidence (0.5-0.6). Manual analysis showed that the majority of these pages relied heavily on flash objects and images. Such pages are easy to classify for humans while our classifier finds hardly any accessible content. Dealing with pages with little to no textual content is clearly one of the future challenges in this domain.

Table 4.11: Inter-annotator agreement by topic

Topic	Agreement
Kids_and_Teens/.../English/Grammar	1.00
Kids_and_Teens/Health/Safety/Fireworks	1.00
Society/Genealogy/Royalty	0.83
Kids_and_Teens/...Cartoons/Hello_Kitty	0.83
...	...
Kids_and_Teens/School_Time/.../Geography	0.59
Home/Family	0.55
Arts/Television/History	0.50
Kids_and_Teens/People/Biography	0.50
Kids_and_Teens/.../Math/Geometry	0.50

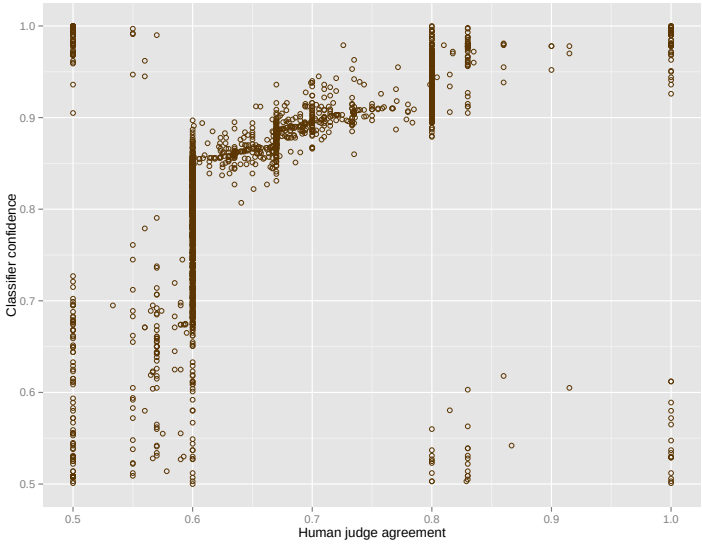

Figure 4.2: Correlation between human agreement on page suitability and classifier confidence

Table 4.12: Quality-dependent performance

		Human	Classifier			
Qual	Share	Agreement	P	R	$F_{0.5}$	ROC
1	100%	0.68	0.72	0.71	0.72	0.76
2	95%	0.68	0.77	0.73	0.76	0.80
2.25	69.3%	0.70	0.77	0.75	0.77	0.80
2.5	51.7%	0.71	0.78	0.75	0.77	0.81
2.75	28.4%	0.73	0.79	0.75	0.78	0.81
3	11.4%	0.75	0.81	0.70	0.79	0.83
3.25	0.9%	0.82	0.90	0.82	0.88	0.90

Analysis of Page Quality

Previously, we inspected the page topic's influence on automatic prediction performance and inter-judge agreement. Another interesting dimension to pursue is the notion of Web page quality. During our manual study of ODP pages we found examples whose textual content might have been generally suitable for children but whose layout and general look and feel was not appealing. Although a Web site's content might have been suitable for children many of the CrowdFlower workers rejected it stating that they disliked the confusing interface or the general impression the page made. In order to quantify this intuition, we asked the CrowdFlower workers to rate each page's quality on a scale ranging from 1 (lowest quality) to 4 (highest quality). While quality of Web pages is a highly subjective notion humans tend to have similar ideas of "good" and "bad" pages. The annotators' low standard deviation of quality judgements per page (0.79 quality grades) results in a coherent overall impression of page quality. The first issue we will address is the connection of page quality and annotator agreement for that page. Intuitively, we would assume that high-quality pages are clear-cut specimen of their specific types and should therefore be easier to assign an age group to. Table 4.12 shows the agreement scores of human judges for Web pages satisfying at least a given quality threshold. For very low quality threshold values the whole set of pages is considered. As we subsequently raise the quality threshold, we note substantial rises in agreement levels. Based on the previously shown correlation between annotator agreement and classification confidence we were interested in the influence of Web page quality on classification performance. Considering the above results with respect to agreement scores, we expected greater classification performance for higher quality pages. We can note that the classifier performance also increases for higher quality page sets. Both agreement and classifier performance rise steadily with quality. As an answer to our third research question, we can see that while our classification method is robust to the quality of a page, it performs best on high quality pages. The same tendency holds true for human judgements.

Language-independent analysis

Finally, it should be mentioned that one of the remaining limitations of our approach still is its strong focus on textual features. While these have been shown to perform well,

Table 4.13: Language-independent performance

Language	# pages	P	R	$F_{0.5}$	ROC
English	2000	0.64	0.51	0.61	0.71
Chinese	1200	0.59	0.50	0.57	0.69
Dutch	1100	0.65	0.47	0.60	0.72
French	2000	0.64	0.46	0.59	0.70
German	2000	0.65	0.48	0.61	0.70
Russian	500	0.62	0.45	0.58	0.69
Spanish	2000	0.63	0.50	0.60	0.69
All	8800	0.63	0.48	0.58	0.70

they force us to operate within the confines of a given language. Even shallow features like sentence lengths or general text lengths can easily become meaningless when crossing language boundaries between training data and on-line examples. Although we do not explicitly address this issue we were still interested in the performance level we could achieve on non-English resources with our current feature set. For assessment we relied solely on those features which are language independent, as for example the visual, commercial or HTML features. We extracted a number of non-English Web pages from the ODP and ran the reduced classifier trained on the original English set on it. Table 4.13 shows the distribution of languages amongst the Web pages and our classifier's performance. For each language 50% of the pages originate from the ODP children's set and 50% from the ODP general set. Regarding our fourth research question we conclude that even without any language-specific training we were able to reliably reach a minimum score of $F_{0.5} = 0.57$ over a set of very different languages. Further research in this direction might be dedicated to applying well-known language-specific scaling factors (e.g., a language's average word length) to use a wider range of features without having to retrain.

Conclusion

In this section, we demonstrated how relevance dimensions, can be automatically estimated from document content. In the given example, we predicted the suitability of Web pages for children based on a wide number of on-page features, reaching an accuracy comparable to that of human judges who were presented with the same task. Previous work on classification along Web taxonomies either excluded age-dependent categories or reported comparably low result scores. We argue however that with appropriate consideration classifying pages according to their suitability for different age groups is not only possible but also highly desirable. Given the ubiquity of the Internet we should investigate how to make Web search less frustrating for all users whose needs differ from the mainstream. In the next section, we will disregard document content and fully turn towards user interaction data. There, our goal will be to estimate relevance based on how people interact with and react to a document.

4.2. Interaction-based Methods

In recent years, content sharing platforms have become very popular. In particular, video sharing platforms have experienced massive growths in both, the amount of shared content as well as the number of viewers. In a recent survey, Cheng et al. [49] attributed YouTube as being solely responsible for approximately 10% of the global Internet traffic. Content sharing services typically enhance the publishing and distribution of pieces of media by social networking features such as friend relationships, messaging, collaborative tagging and commenting functionalities. In order to make content available to the users, most state-of-the-art content sharing platforms rely on tagging. The step of assigning tags, however, is often left to the user community.

While there are users who relish this task, and some platforms even integrate it into games to make it more entertaining, there are many who regard it as a rather tedious burden. Ames and Naaman [14] studied user tagging behaviour and found that a frequently expressed motivation for tagging lies in the necessity to do so in order to make the content available to the user base. Additionally, they noted a significant share of tags to be strongly dependent on the tagger's socio-context, rendering them less useful for users that do not share the same context (i.e., friends, place of residence, cultural background).

To overcome this challenge in related domains, automatic tagging mechanisms have been proposed that extract keywords from textual meta data and content. Larson et al. [125] find that in the case of shared multimedia content, however, this is often not feasible with satisfying precision, as meta data can be sparse or ambiguous and concept detection from audio-visual signals is still considered more difficult than text-based alternatives. For example, many videos on YouTube feature only a title and a brief textual description. Statistical tag prediction approaches face significant problems when operating in such resource-impoverished domains.

Commenting, on the other hand, appears to be a more natural activity for most users. We can observe extensive threads of comments related to shared media items. In this work, we propose the use of time series analyses for audio-visual content with sparse meta data. The investigation is not targeted towards the actual content and meta data but will focus exclusively on people's comments towards the content. To this end, we employ a language modelling approach to utilise the naturally created community information on content sharing platforms, to infer potential tags and indexing terms. In this way, we aim to mitigate the vocabulary gap between content and query. Previous studies, e.g., by Oghina et al. [164], often doubted the usefulness of user comments for retrieval tasks due to the high rate of noise in the chat domain. However, given the large scale at which user comments are currently available, we will show that informed means of interpreting noisy natural language communication streams as well as aggregation with orthogonal types of (social) media can help to identify valuable pieces of information in the abundant underlying noise.

While tag prediction from short, noisy user communication has not been extensively studied, there are several prominent methods for keyword extraction directly based on content. Hu et al. [98] introduced a graph-based method for discussion summarisation through sentence extraction from Weblog posts. Budura et al. [40] propagate tags along the edges of a Web page similarity graph that is built based on a range of content fea-

tures. Matsuo and Ishizuka [149] present an approach of extracting keywords from single documents without the need for a background corpus. Using intra-document term distributions, the authors report performances that approximate those of *tf/idf*-based methods. Wartena et al. [232] propose to infer keyword candidates from the semantic relationships between terms in academic abstracts and BBC news stories. Tomokiyo and Hurst [219] present a language modelling approach to keyword extraction from longer coherent news articles. Their use of the divergence between term frequency distributions is based on an intuition similar to our method. Due to the high amount of noise in user comments, additional steps are required to successfully apply their method in this domain. To this end, we apply time series analyses to identify informative comments. Amodeo et al. [16] investigated temporal relationships between time of publication of blog posts and their probability of relevance. The authors employ a notion of activity bursts similar to the one proposed in this work. However, where their approach applies time series analyses directly to documents in order to prune the list of pseudo relevant results, we aim to improve the general indexing quality by broadening the document vocabulary.

Tag prediction is most prominently used to describe pieces of textual content, as semantic concepts can be conveniently observed in the form of term occurrences. However, there are several pieces of work dedicated to predicting tags directly from multimedia content. Eck et al. [67] present an approach of predicting tags from the audio signal of music pieces. Similar approaches for other types of media include [198]'s automatic video tagging method which propagates tags across videos containing redundant or similar content, or [237]'s photo tagging scheme.

While the previously discussed publications concentrate solely on extracting tags from actual content, we can identify a body of work that makes additional use of community-created information. As an example, Mishne and Glance [153] first employed user comments to enhance Weblog retrieval. Heymann et al. [92] predict tags from a range of local Web page features enriched by information from social bookmarking services. Yee et al. [243] presented a method of improving search performance by utilising user comments by means of a *tf/idf*-based method. Most recently, Filippova and Hall [72] employ user comments to aid content classification performance. The promising results achieved by previous work support the feasibility of our goal: Describing content exclusively based on user comments. We will employ statistical language models aided by time series analyses and external Web resources such as Wikipedia, to find potential index terms and evaluate their quality in a series of TREC-style experiments.

Comments as Bursty Streams

Common methods for characterising individual documents d within a collection C are often based on the intuition that some terms will occur more frequently locally in d than in the collection-wide average. This notion is for example expressed in the popular tf/idf family of formulae but is also implicit in the language modelling framework (see e.g., [93]). The same method can be applied to the video retrieval setting, in which each shared video corresponds to a distinct d. We assume a unigram collection model LM_C comprised of all comments in C and dedicated document models LM_d based on the comment thread of document d. Subsequently, we assume good descriptors of d can

be determined by the term-wise KL-divergence between both models (LM_C and LM_d), identifying locally densely occurring terms w (those that display a high negative value of $KL(w)$).

$$KL(w) = P(w|d)\ log\frac{P(w|d)}{P(w|C)} \tag{4.1}$$

This method has been applied for a wide number of settings and is known for its robustness and generalizability [219]. The domain at hand, however, imposes a number of specific challenges on automatic keyword extraction. There are several sources of comment noise that require appropriate treatment. Firstly, there is a significant share of comments that are uninformative for the task of keyword extraction, either because they are off-topic (spam) or because they simply do not convey much meaning (e.g., "Cool."). In order to address this type of messages, we introduce a resource selection step that identifies informative comments based on Kleinberg's burstiness criterion [113]. When analysing the usage statistics of his personal email account, Kleinberg noticed that his incoming email was subject to sudden, typically short, peaks of activity. A first investigation in the domain of shared Web videos showed that most comment threads (98%) display the same peaking behaviour.

These so-called *bursts* can be related to external triggers such as a famous musician winning an award, causing a sudden increase of attention and commenting activity on his music videos. Often, however, the trigger is of internal nature, e.g., caused by controversial comments that spark an avid discussion. This latter class of triggers lets us assume that comments submitted within an activity burst may be more informative than regular ones. We formulate a variation of Kleinberg's original burst detection scheme to better fit the notion of threaded chat communication: We consider each coherent sequence of messages $m_i ... m_j$ with inter-comment intervals $\delta_t(i, i+1)$ shorter than a threshold value δ_t as candidate bursts. In this work, we set δ_t to be the median time between comments for each document, however, further tuning of this parameter could prove beneficial. In order to select informative bursts, we apply a burstiness function $b(i, j)$, according to which we rank all candidates. The underlying intuition is that a "good" burst should cover many comments in as little time as possible. This is represented by $length_{rel}(i, j)$, the relative share of comments contained in the burst, divided by $\delta_{rel}(i, j)$, the relative amount of time for which the burst lasted. Consequently, we pool all comments from the n highest-ranked bursts to train LM_d. This filtering step eliminates a significant proportion of unrelated "background noise" comments from the modelling step.

$$b(i, j) = \frac{length_{rel}(i, j)}{\delta_{rel}(i, j)} \tag{4.2}$$

Considering the merit of using bursty comments, and assuming them to be triggered within the comment stream, we further suspect that the event triggering the increased commenting activity may be of import as well. In order to verify this hypothesis, we use a history of h comments immediately preceeding each burst as an alternative resource. Manual qualitative investigations showed an optimum in extracted tag quality at $h = 7$ history comments preceeding each burst.

In order to harmonize the evidence from pre-burst histories and actual bursts, we turn to the simplest setting of Ogilvie's method for language model combination [165]. Instead of directly estimating the probabilities of observing given terms from the whole comment thread, we now use a weighted combination of two such models. $P_B(w|D)$ is based on the maximum likelihood estimate of term occurrence according to the comments within bursts. $P_H(w|D)$ is based on the 7-comment pre-burst history. The mixture parameter λ determines the relative importance of burst comments over history comments. Higher values of λ give more weight to comments within the bursts.

$$P_{HB}(w|D) = \lambda P_B(w|D) + (1-\lambda)P_H(w|D) \tag{4.3}$$

In order to assess tag extraction quality, we randomly sampled 50 videos from YouTube, applied our four tag prediction methods (based on the entire comment thread, on bursts, on pre-burst histories, and, on the burst/history mixture) and measured the overlap of the respective with the gold standard tags as assigned by YouTube users. Figure 4.3 shows tag prediction performance as we vary the composition of the model mixture. Best results could be achieved for settings of $\lambda = 0.65$. Language models trained on the entire comment thread resulted in an F_1 score of 0.061, significantly below any of the compared settings in Figure 4.3 (tested using Wilcoxon Signed rank test with $\alpha < 0.05$).

Wikipedia as a Surrogate for Natural Language Vocabulary

Previously, we addressed noise in the form of unrelated and uninformative comments within the thread. The second source of noise are misspellings, abbreviations, chatspeak and foreign language utterances, all of which are frequently encountered in on-line chat communication. To address this, we use the online encyclopaedia Wikipedia for regularization. We formally introduce the $\eta(w)$ criterion. Terms w that do not have a dedicated article in the English version of Wikipedia are assumed to be noise and, subsequently, rejected from the list of candidate terms. Due to Wikipedia's high coverage, the number of false positives, valid terms rejected by this filter, has been found to be negligible.

$$\eta(w) = \begin{cases} 1 & \text{if } w \text{ has an English Wikipedia article,} \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

With our noise filtering components in place, our final extraction scheme is:

1. Within a comment thread d, find all message sequences (so-called bursts) with inter-comment intervals no longer than δ_t.
2. Rank the bursts according to their burstiness $b(i,j)$ (Eq. 4.2) and keep top n.
3. Train LM_d on the previously selected most bursty comments (Eq. 4.3).
4. Rank all terms w according to (Eq. 4.1).
5. Return top k terms $w_1 \ldots w_k$, rejecting all w with $\eta(w) = 0$.

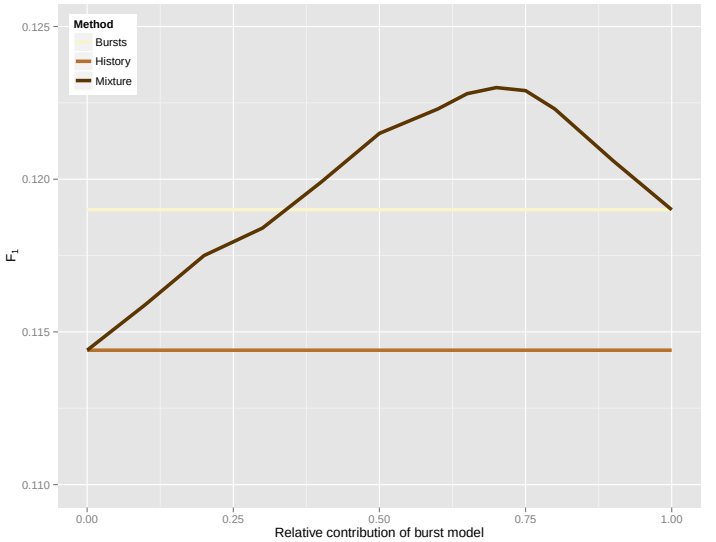

Figure 4.3: Performance of burst-history mixture models for varying weights.

Evaluation

Previously, we investigated our method's performance at replicating the gold standard labels assigned by YouTube users. Ultimately, however, we aim to improve retrieval performance of shared video content by extracting representative terms *a priori* at indexing time. In this way, we can enrich sparsely annotated content (e.g., in the audio-visual domain) by harnessing community knowledge in the form of user comments.

Our evaluation dataset is comprised of 4.7 million user comments issued towards more than 10.000 videos. It was collected between December 2009 and January 2010. The crawling process was limited to textual information, omitting the actual audio-visual content, and was started from a diverse selection of manually formulated seed queries, following the "related videos" paths. On average, every video in this collection has 360 ($\sigma = 984$) dedicated user comments and 14 tags ($\sigma = 11.8$) assigned to it. The only source of textual meta information are titles and video descriptions provided by the uploader.

To evaluate our method's merit at indexing time, we conduct a TREC-style retrieval experiment. We use the Lucene search engine library[10] and a BM25F retrieval model [182]. We manually designed a set of 40 topics that are well represented in our collection (e.g., *"Lady Gaga Music Video"* or *"Swine Flu 2009"*). Finally, we obtained binary relevance judgements for the top 10 retrieved results per query via crowdsourcing. On average, 36 results per query were evaluated. Alonso and Mizzaro [11] describe a similar setting for collecting pairwise query/document judgements, concluding that a group of untrained workers can produce relevance judgements of a quality comparable to that of a single domain expert. As a consequence, we collected 10 redundant binary judgements per unique topic/video pair and aggregate the results in a majority vote. The task was offered

[10] http://lucene.apache.org/

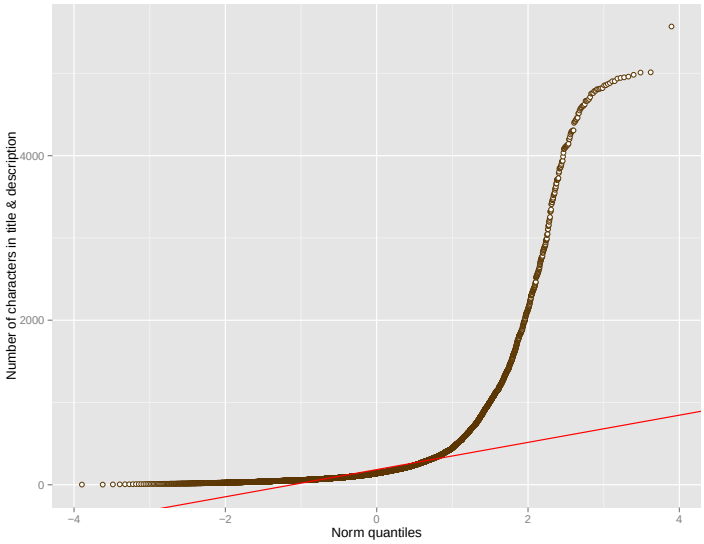

Figure 4.4: Quantile distribution of YouTube video meta data length in characters.

on the crowdsourcing platform Amazon Mechanical Turk (AMT) at a pay level of 2 cents per single judgement. In order to ensure result quality, we employ gold standard judgements as well as honey pot questions as recommended by Kazai [107]. In Chapter 5, we will give a detailed discussion of quality control in crowdsourcing efforts. Our method's parameter settings were determined on the first 5 topics of our data set by varying the number of most intense bursts, n, and the number of terms extracted per video, k. These training queries were not used further for evaluation. The best observed performance could be achieved at $n = 20, k = 15$. Table 4.14 compares the retrieval performance of various BM25F indexes, using either only original meta information, extracted terms, or combinations of both. We measure result quality in terms of *Mean Reciprocal Rank* of first relevant results (MRR), *Mean Average Precision* (MAP) as well as precision at rank 10 (P@10). In a baseline performance run, we rely exclusively on video titles and textual descriptions, each of which becomes an individual field in the retrieval model's index. This is comparable to the information based on which the standard YouTube search API operates[11]. Unless stated differently, all experiments were conducted on the full video corpus regardless of the number of comments per video. Statistically significant performance improvements over the baseline are denoted by the asterisk character (tested using a Wilcoxon signed rank test at $\alpha = 0.05$-level). In a second experiment, we exclusively use the top $k = 15$ terms extracted by our method to form the index. We can note a significant and consistent improvement over the original index's retrieval performance. When combining extracted terms and original meta data by interleaving a pool of k terms from both source selections, we experience another significant performance gain. Indexing the full comment thread alongside the original meta data introduces a high number of

[11] https://developers.google.com/youtube/2.0/developers_guide_protocol_api_query_parameters#qsp

Table 4.14: Retrieval performance on shared video content.

Index type	MRR	MAP	P@10
Title & description	0.81	0.48	0.46
k extracted terms	0.85*	0.52*	0.51*
k extracted terms (bursts only)	0.80	0.49	0.46
k extracted terms (no regularization)	0.63	0.32	0.25
k random comment terms	0.08	0.03	0.05
Title, description & extracted terms	**0.89***	**0.67***	**0.64***
Title, description & full comment thread	0.48	0.33	0.34

false positives, ultimately hurting retrieval performance. As a point of comparison, we include runs for extracted terms based solely on bursts (not using the pre-burst history), as well as those not using Wikipedia regularization. In both cases, we note performance drops as compared to the regularized mixture setting.

The domain at hand is particularly challenging, since a high percentage of videos is annotated only sparsely. Our investigation shows that both titles and descriptions contain only small amounts of text (titles have an average length of 32.8 (σ = 12.8) characters, and, descriptions average at 211 (σ = 220) characters each). Figure 4.4 shows the quantile distribution of video description lengths in our data sample. A significant percentage (58%) of videos in our corpus is described with no more than 140 characters each. This represents the same amount of information that could be conveyed in a single tweet. For video titles, we observed a similar behaviour with more than 50% of all titles being shorter than 35 characters. In combination, this lack of explicit content annotation may hinder successful retrieval. In order to confirm this assumption, we repeat the retrieval experiment and restrict the corpus to those videos that are sparsely annotated. More concretely, we index only those videos that feature either less than 35 title characters *OR* less than 140 description characters. The resulting set contains 7840 videos, an equivalent of 77% of the original collection.

Table 4.15 details the performance of the previously-introduced indexes when textual information is sparse. We can see that performance scores are consistently lower, while the performance-based ranking of approaches remains the same. However, the difference in performance between comment-based and exclusively meta data-based indexes becomes more expressed. Again, we can note a clear merit of using burst / and pre-burst information, as well as Wikipedia regularization. In conclusion, we observe significant performance improvements across all experimental settings when applying keyword extraction to user comment threads for the task of video retrieval on online content sharing platforms such as YouTube.

Discussion

The previous sections detailed concrete, task-driven performance evaluations of our method. In this section, we will dedicate some room to lessons learned and will discuss several observations that could not be confirmed to be statistically significant but yet deserve attention as they may become more salient in related applications or domains.

Table 4.15: Retrieval performance for sparsely annotated content.

Index type	MRR	MAP	P@10
Title & description	0.74	0.41	0.35
k extracted terms	0.79*	0.44*	0.39*
k extracted terms (bursts only)	0.75	0.38	0.33
k extracted terms (no regularization)	0.56	0.25	0.27
k random comment terms	0.08	0.04	0.05
Title, description & extracted terms	**0.82***	**0.63***	**0.59***
Title, description & full comment thread	0.41	0.31	0.25

In order to give qualitative insights into comment-based keyword extraction, let us visit an example that we encountered during the manual inspection of extraction results on the YouTube dataset and that is representative for a large number of cases. The video in question shows scenes from a Mafia-related computer game followed by several action film shooting scenes. While the original title ("Mafia Shootout") and description ("Mafia members in a huge shooting.") are very brief and uninformative, the results of our term extraction method show convincing tendencies. The highest-ranked term was "Mafia", which, considering that we do not peek into the actual meta information of the video, is a very good match. Subsequent ranks contained further unsurprising terms such as "shoot" or "gun". The interesting matches, however, were "Corozzo" and "Guarraci", referring to Joseph "Jo Jo" Corozzo, Sr. and Francesco "Frank" Guarraci, two infamous criminals. Additionally, the term "Mississippi" ended up on a high rank. At first we considered it a false positive, before looking more deeply into the matter and discovering the Dixie Mafia, an organization that heavily operated in the southern U.S. states in the 1970s. An additional facet of the video that our method picked up was the keyword "game". The existing meta information did not cover this essential aspect of the content. Considering this example, we can see how comment-based keyword extraction manages to discover novel aspects of a topic rather than exclusively sticking to the literal content of a video item. The general observation was that our method often picks up very specific topical aspects of a given piece of content. As a consequence of relying on locally densely occurring terms, we discover "Guarraci" rather than "criminal".

One particular application that became obvious throughout the course of our research is using term extraction from comments as a means of summarizing the discussed content. When manually inspecting the output of our methods, we arrived at the impression that the set of top-ranked keywords was sufficient to convey a reliable description of the content itself. We aim to further confirm this notion and determine the method's merit for content summarisation in a dedicated series of future experiments.

In this work, we investigated the usefulness of user comments for two tasks, (1) reproducing the user-assigned YouTube tags without using any form of video-related meta information, and, (2) Improving retrieval performance of shared videos by expanding the index by terms extracted from user comments. In the future, it would be interesting to evaluate the degree to which our findings generalize to different domains and media types.

The step towards alternative media is not assumed to introduce significant changes to the method since we did not make any assumptions on the content other than the existence of time-stamped user comments. Therefore, our method should be conveniently portable to platforms such as Flickr (images) or Last.fm (music). A more challenging but also potentially more interesting generalization step could be taken to explore novel domains besides shared and commented media content.

Conclusion

In this section, we investigated the potential use of user interaction information in order to estimate document relevance. Concretely, we used textual comments to infer the topical relevance of YouTube videos. We found that it was possible to deduce meaningful tag candidates from comment streams without using any form of direct content information such as titles or video descriptions. Results improved significantly when incorporating time series analyses to identify informative regions in the discussion. We were able to benefit from external resources such as Wikipedia by using them to reduce the background noise of the chat domain. After a series of experimental runs against a set of gold standard tags, we confirmed the usefulness of the extracted terms for retrieval purposes in a sizeable TREC-style experiment based on several million user comments. We showed, that including only a high-precision set of tags extracted from user comments achieves better retrieval performance than either ignoring comments altogether or indexing the full comment stream.

In the following, final, section to this chapter, we will join our insights about content-based and interaction-based relevance estimation and show an estimation scheme based on the combination of both sources.

4.3. Combined Methods

In this section, we will pick up and combine the challenges and methods from the two previous sections and demonstrate the merit of a multi-source approach for estimating document relevance. As a concrete task, we will stick to Section 4.1's challenge of predicting a document's suitability for children, but, will now turn towards audio-visual documents, rather than textual ones. This choice is motivated in children's substantial attraction towards video content. While textual resources often require considerable literacy skills, videos also appeal to children who can not yet read very well. The potential danger lies in the unmoderated consumption of videos. Parents or teachers who guide children's information seeking naturally filter which content to show to their wards and which to avoid. As stated in the beginning of this chapter and supported by studies such as presented by [163], parental supervision and support are not always available and automatic means of inference can make a considerable difference.

Video-sharing communities that specifically target young audiences appear to be a promising alternative. Manually selected collections of videos, as was for example offered on Totlol[12], provide high quality content for children and parents. They do however impose a high work load of manual editing and selecting on the community. Hand-picked collections typically feature a comparably low coverage rate and low agility as it

[12] http://www.totlol.com

takes time for the community to explore the available content. Since permanent assistance of a prudent adult can in reality not always be ensured and hand-selected video collections require high maintenance efforts, an automatic method for identifying suitable video resources is desirable. To this end, we propose an approach that makes use of the existing content-based and interaction-based information on video sharing platforms in order to determine suitability.

While there has been no previous work dedicated to determining the suitability of Web videos for children a range of relevant research on related topics has been conducted. In recent years, much research effort has been invested into automatic video classification. Traditional video classification approaches (see e.g., [230]) at first commonly employed audio-visual features, often using Hidden Markov Models as described by [138] and [140]. There have however been promising advances, e.g., by [211], into using more sophisticated machine learning techniques such as Support Vector Machines. The growing amount of shared and tagged video content available has given rise to approaches making use of textual features [133]. The fundamental difference to our method however lies in the objective. While topic information appears to be well-contained in tags and headlines, age appropriateness can be widely independent of the video's subject. Another active line of literature investigates the predictive and indicative potential of user comments on content-sharing platforms. Lange [122] found that user comments on YouTube often more clearly express the relationships between users than the platform's explicit friendship function. Siersdorfer et al. [199] asked the more general question of the overall usefulness of YouTube user comments. In several experiments they were able to build models to identify polarising comments as well as predicting comment ratings. De Choudhury et al. [61] ,finally, tried to identify characteristics of interesting conversations by analysing discussions in YouTube comments. These various successful exploitations of information contained in comments encourage our approach of determining suitability of video content based on user comments.

The structure of YouTube content

As a first step towards identifying suitable video content for children, we will introduce a range of potential features that may convey suitability information. We will start with a brief inspection of the information offered on a typical YouTube page. A typical example is shown in Figure 4.5. The pieces of information on the page can be grouped into 4 distinct categories:

Video information This category contains all information concerning the actual video content. Examples from this category are the video title, tags, full text descriptions, its genre or the play time. Intuitively, one would assume this category to contain the strongest indicators of suitability as it is directly dedicated to the video content. We will however show, that other sources of information can yield comparably strong indications.

Author information The second source of information available on the page is related to the video's uploader. While on the actual video page, we can only find the author's user name, retrieving information from the related user profile can offer many interesting clues such as the user's age or popularity in the community.

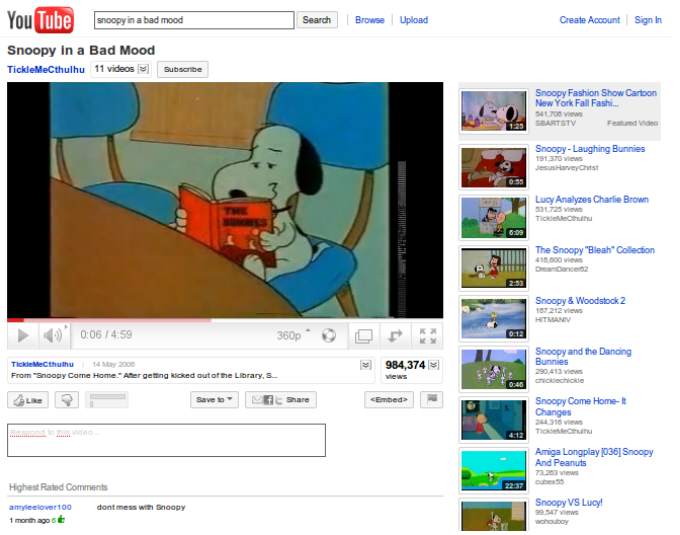

Figure 4.5: Example of a shared video on YouTube

Meta Information This category is dedicated to automatically generated meta information on the video. Examples from this group are the video's view count or the number of times it was added to a user's favourite list.

Community-created information The final category is concerned with all information that was created by the user community. Examples are user ratings, video responses or comments.

Capturing Suitability

After having introduced the range of information available on typical YouTube pages, we will now discuss their applicability for predicting suitability of video content. Most notably, we will group the available sources of information into *content-based* and *interaction-based* ones. Following intuition, video and author information will be considered content-related data, while meta and community information fall into the interaction category.

Content Information The main textual resources in this category are the video's tags and its description. The title and the category turn out to be already contained term by term in the list of tags. We propose using uni-gram tag-based language models to represent topical similarities and tag co-occurrences of child/non-child videos. Due to the relative brevity of video descriptions, we expect only limited growth in complexity when using higher order models at this point. Another feature of interest is the video's play time. Topical classification was hardly able to exploit this aspect. Anderson and Levin [17], however, report children to show a significantly lower attention span than adults. We will investigate whether this is reflected in the length of children's videos compared to general video clips.

User profiles offer a wide range of possibly interesting information. The user's age may be relevant towards the suitability decision of her or his content. It is likely that user age forms a valid prior on suitability as children's videos will in the majority of cases be watched by actual children or by parents which fall into empirically distinguishable age groups. Furthermore, the profile text (either free text or information on favourite books, films, hobbies, etc.) is expected to provide valuable information on the user's background and interests. We will use a language modelling approach to reflect these specifics for authors of children's and general content. It is however important to consider the potentially sparse nature of these texts. The user is not bound to provide and maintain a detailed profile of his likes and dislikes.

Interaction Information YouTube comments are typically short user messages, related to the published video content (although we also encountered extensive paragraphs of text). Users additionally have the possibility to rate comments. The highest-rated comments are then shown in a priority position above the main discussion. As a first step towards using comments for suitability prediction, we will build comment language models for children's videos as opposed to general videos. These models are expected to give further evidence of the video's topic and related subjects.

One of the strongest motors of commenting is controversy. People feel more far more inclined to contribute to a discussion whose general position they disagree with. This results in controversial topics being discussed more vividly with comments pouring in every few seconds [122]. Considering the nature of children's videos we expect to see far fewer heated debates on an episode of "Hello Kitty" than there might be on a news story that deals with the recent changes to the US healthcare system. As a consequence we will consider the total number of comments but also the median time between comments as features.

We will furthermore capture this notion by applying sentiment analysis to the video comments. Children's content is in general expected to cause positive affection rather than negative sentiments. The typical behaviour of antagonism that is often observed in on-line discussions (e.g., by [121]), is expected to be less frequent for children's content. Our sentiment analysis is based on the sentiment corpus by Pang and Lee [167]. The likelihood of a given comment c being well-intentioned is expressed as the average of its constituent term's likelihoods. The likelihood of each single term t being positive is finally defined as the number of times t was observed in a positive comment in relation to the total occurrences of t. The analogous score for the negative case is computed as well and both are reported as features. A complete overview of all features and their category affiliations is given in Table 4.16.

$$P(positive|c) = \frac{1}{|c|}\sum_{i=0}^{c|} p(positive|t_i)$$
$$p(positive|t) = \frac{count_{pos}(t)}{count(t)}$$

Experiments

To conduct our experiments, we use the same corpus of YouTube videos, that was previously described in Section 4.2. While the average video in our collection had a total of

Table 4.16: YouTube features

Content	Interaction
Play time	Average rating
Tag LM	# comments
Description LM	Comment LM
Presence of tag “kid”	positive sent. score
Presence of tag “child”	negative sent. score
Author age	# favourites
Author profile text LM	# views
# subscribers to channel	Median time between comments

360 comments, there are outliers which solely feature as many as 281,000 comments for famous pieces of popular culture. In order to reduce the crawling and processing time of such videos to reasonable dimensions we capped the number of comments fetched at = 4950 comments. At that point more than 96.8% of the videos do not have additional comments. Without affecting the majority of videos, the computational load could be reduced significantly. Out of the whole collection, an initial sample of 1000 videos (50% suitable for children and 50% for adult audiences) have been rated concerning their child-friendliness by a domain expert with a background in childcare and education.

Our further exploration will be guided by the following three research questions: (1) Can we automatically identify child-friendly videos using exclusively non-audio-visual information? (2) Can community expertise outperform directly video-related information at predicting child-friendliness? (3) Can video play time indicate child-friendliness of shared Web videos? To begin our inspection of the domain we split the corpus into stratified training (90%) and test (10%) sets, extracted the previously described features and trained a range of state-of-the-art machine learning classifiers. Table shows a performance comparison of the various approaches on the previously unseen test set. Performance will be captured in terms of precision and recall as well as their combination in the $F_{0.5}$-measure. As in Section 4.1, we decided for the precision-biased F score to reflect the nature of our task. Showing as few as possible unsuitable videos to children (high precision) is clearly more important than retrieving all suitable videos (high recall). The area under the ROC curve is additionally reported to give a notion of classification confidence.

We can observe that already straight forward Naïve Bayesian approaches yielded convincing performance. The overall best-performing model was an SVM classifier (We employed an SVM using a Pearson VII universal kernel function as suggested by Qifu et al. [175]. The following parameter settings were used: $\omega = 1$, $\sigma = 2$, $\epsilon = 10^{-12}$, and, $c = 1$.) Although we did not invest time into further feature engineering and parameter tuning at this point of our research, the results look promising.

We see our first research hypothesis confirmed; Suitability of shared videos can be reliably estimated using exclusively non-audio-visual features. In order to gain a deeper understanding of how the crucial information is reflected in the data, we will further analyse the individual prediction performance per feature category and per single fea-

Table 4.17: Classification performance

Classifier	P	R	$F_{0.5}$	ROC
SVM	**0.85**	**0.67**	**0.81**	**0.85**
Random Forest	0.77	0.86	0.79	0.87
Decision Table	0.75	0.83	0.76	0.79
Logistic Regression	0.72	0.63	0.7	0.74
Decision Tree	0.74	0.79	0.75	0.82
Ada Boost	0.72	0.89	0.75	0.79
MLP	0.78	0.6	0.74	0.77
Naïve Bayes	0.72	0.63	0.7	0.72

Table 4.18: Feature category performance comparison

Category	P	R	$F_{0.5}$	ROC
Content Information	0.62	0.77	0.65	0.66
Interaction Information	0.68	0.75	0.69	0.71

ture. For this purpose we again used the previously described SVM classifier which was now, however, trained on one feature category at a time. Table 4.18 shows a ranking by feature category performance. Returning to our second research question, we note that interaction information represents a more powerful predictor of suitability for children than directly content-related features. This finding is statistically significant at $\alpha < 0.05$-level (determined using Wilcoxon Signed-Rank test). We conducted the analogous experiment training the classifiers on just a single feature at a time. The results are shown in Table 4.19.

Closer examination of the results shows a surprising tendency. The strongest single feature turned out to be the number of times the video was watched. This finding is most likely due to the fact that the majority of extremely popular videos on YouTube are about current pop culture. The number of people who are actually interested in watching children's videos will be limited in comparison to the fan community of a famous musician or music style. While this finding represents an interesting foundation for further research, it is of course not prudent to deduce that every video that is not largely popular in terms of number of views could be shown to children without concern.

Reconsidering our third research question, we have to note that (at least at the moment) the mere duration of a YouTube video does not give useful clues as to its suitability for children. Despite the findings of Anderson and Levin [17] who were able to measure significant differences in the attention spans of children of different age groups, the hypothesis of children's videos being shorter did not hold true for our application. We assume that this is largely due to the fact that most video clips on YouTube are short by definition. Video sequences of no more than 10 minutes are apparently easily understandable for children so that no significant differences could be measured. This observation should be revisited in the future, if the distribution of video lengths should change due to new terms of service on the platform.

Table 4.19: Individual feature performance comparison

Feature	P	R	F0.5	ROC
# of views	0.75	0.54	0.7	0.72
Average rating	0.66	0.85	0.69	0.62
# favourites	0.72	0.54	0.68	0.7
Median inter-comment interval	0.71	0.55	0.67	0.7
Author age	0.64	0.82	0.67	0.64
Tag LM (1-gram)	0.59	0.92	0.64	0.55
Profile text LM (3-gram)	0.6	0.8	0.63	0.65
Comment LM (3-gram)	0.58	0.87	0.62	0.56
Sentiment score positive	0.57	0.91	0.62	0.52
Sentiment score negative	0.57	0.88	0.61	0.54
Description LM (3-gram)	0.56	0.91	0.61	0.53
# of subscribers	0.55	0.99	0.6	0.56
Presence of tag “kid”	0.55	1	0.6	0.54
Uncapped # of comments	0.55	0.98	0.6	0.5
Presence of tag “child”	0.55	1	0.6	0.5
Video Play time	0.54	0.48	0.53	0.51

4.4. Conclusion

In this chapter, we showed examples of three automatic means of estimating relevance. First, addressing Research Question 2.a), we used document content of Web pages in order to infer the pages' child suitability. We showed that, using modern machine learning methods, we can combine several document-level features to obtain a strong predictor of suitability that could match human performance for the task. Turning away from actual document content, in Section 4.2, we employ a term extraction scheme on textual comments submitted towards YouTube videos in order to extract indexing terms, and, eventually, estimate topical relevance without taking into account the actual document content. Finally, in Section 4.3, we combined both data sources and present a joint approach of estimating the child suitability of YouTube videos based on features related to both content and user interaction.

A particular merit of such hybrid estimation schemes, employing multiple sources of data, becomes salient when we take into account the life cycle of an information object. In the examples given in this chapter, we found interaction-based evidence to frequently outperform purely content-based efforts due to a greater volume of available information. However, such methods will face severe “cold start” problems before sufficiently many user interactions have been registered. At such early points in the life span of a document, it is more promising to mainly rely on content-based evidence. A mixture model that explicitly takes into account the document's life cycle to determine the relative contributions of individual mixture components seems appropriate.

Throughout this chapter, we designed statistical models based on prior observations and manually created ground truth labels in order to estimate the distribution of relevance for unseen future data. So far, we did not discuss this, necessary, first step of

manual labelling in much detail. Since, however, it is an integral part of many real-life development cycles in both industry and academia, we will dedicate the following chapter to the challenge of integrating human expertise in the computation loop.

References

[11] Omar Alonso and Stefano Mizzaro. “Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment”. In: *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 2009, pp. 15–16.

[14] Morgan Ames and Mor Naaman. “Why we tag: motivations for annotation in mobile and online media”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 971–980.

[16] Giuseppe Amodeo, Giambattista Amati, and Giorgio Gambosi. “On relevance, time and query expansion”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 1973–1976.

[17] Daniel R. Anderson and Stephen R. Levin. “Young Children’s Attention to" Sesame Street"”. In: *Child Development* (1976), pp. 806–811.

[29] Paul N. Bennett and Nam Nguyen. “Refined experts: improving classification in large taxonomies”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 11–18.

[40] Adriana Budura et al. “Neighborhood-based tag prediction”. In: *The semantic web: research and applications*. Springer, 2009, pp. 608–622.

[42] Jamie Callan and Maxine Eskenazi. “Combining lexical and grammatical features to improve readability measures for first and second language texts”. In: *Proceedings of NAACL HLT*. 2007, pp. 460–467.

[44] Sandra L. Calvert. “Children as consumers: Advertising and marketing”. In: *The Future of Children* 18.1 (2008), pp. 205–234.

[47] Carlos Castillo et al. “Know your neighbors: Web spam detection using the web topology”. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, pp. 423–430.

[49] Xu Cheng, Cameron Dale, and Jiangchuan Liu. “Understanding the characteristics of internet short video sharing: YouTube as a case study”. In: *arXiv preprint arXiv:0707.3670* (2007).

[52] Kevyn Collins-Thompson and Jamie Callan. “A language modeling approach to predicting reading difficulty”. In: *Proceedings of HLT/NAACL*. Vol. 4. 2004.

[59] Honghua Kathy Dai et al. “Detecting online commercial intention (OCI)”. In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 829–837.

[61] Munmun De Choudhury et al. “What makes conversations interesting?: themes, participants and consequences of conversations in online social media”. In: *Proceedings of the 18th international conference on World wide web*. ACM. 2009, pp. 331–340.

[67] Douglas Eck et al. “Automatic generation of social tags for music recommendation”. In: *Advances in neural information processing systems* 20.20 (2007), pp. 1–8.

[70] Lijun Feng. “Automatic readability assessment for people with intellectual disabilities”. In: *ACM SIGACCESS Accessibility and Computing* 93 (2009), pp. 84–91.

[71] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. “Cognitively motivated features for readability assessment”. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2009, pp. 229–237.

[72] Katja Filippova and Keith B. Hall. “Improved video categorization from text metadata and user comments”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 835–842.

[77] Evgeniy Gabrilovich and Shaul Markovitch. “Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization”. In: *Journal of Machine Learning Research* 8 (2007), pp. 2297–2345.

[83] Koraljka Golub and Anders Ardö. “Importance of HTML structural elements and metadata in automated subject classification”. In: *Research and Advanced Technology for Digital Libraries*. Springer, 2005, pp. 368–378.

[86] Karl Gyllstrom and Marie-Francine Moens. “Wisdom of the ages: toward delivering the children’s web with the link-based agerank algorithm”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM. 2010, pp. 159–168.

[92] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. “Social tag prediction”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, pp. 531–538.

[93] Djoerd Hiemstra. “A probabilistic justification for using tf× idf term weighting in information retrieval”. In: *International Journal on Digital Libraries* 3.2 (2000), pp. 131–139.

[98] Meishan Hu, Aixin Sun, and Ee-Peng Lim. “Comments-oriented blog summarization by sentence extraction”. In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 901–904.

[107] Gabriella Kazai. “In search of quality in crowdsourcing for search engine evaluation”. In: *Advances in information retrieval*. Springer, 2011, pp. 165–176.

[112] George R. Klare. “The measurement of readability: useful information for communicators”. In: *ACM Journal of Computer Documentation (JCD)* 24.3 (2000), pp. 107–121.

[113] Jon Kleinberg. “Bursty and hierarchical structure in streams”. In: *Data Mining and Knowledge Discovery* 7.4 (2003), pp. 373–397.

[115] Pranam Kolari, Tim Finin, and Anupam Joshi. “SVMs for the blogosphere: Blog identification and splog detection”. In: *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. Vol. 4. 2006, p. 1.

[121] Patricia G. Lange. “Commenting on comments: Investigating responses to antagonism on YouTube”. In: *Annual Conference of the Society for Applied Anthropology. Retrieved August*. Vol. 29. 2007, p. 2007.

[122] Patricia G. Lange. “Publicly private and privately public: Social networking on YouTube”. In: *Journal of Computer-Mediated Communication* 13.1 (2007), pp. 361–380.

[123] Andrew Large, Jamshid Beheshti, and Alain Breuleux. “Information seeking in a multimedia environment by primary school students”. In: *Library & Information Science Research* 20.4 (1998), pp. 343–376.

[124] Andrew Large, Jamshid Beheshti, and Tarjin Rahman. “Design criteria for children’s Web portals: The users speak out”. In: *Journal of the American Society for Information Science and Technology* 53.2 (2002), pp. 79–94.

[125] Martha Larson et al. “Automatic tagging and geotagging in video collections and communities”. In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM. 2011, p. 51.

[133] Wei-Hao Lin and Alexander Hauptmann. “News video classification using SVM-based multimodal classifiers and combination strategies”. In: *Proceedings of the tenth ACM international conference on Multimedia*. ACM. 2002, pp. 323–326.

[135] Bing Liu, Minqing Hu, and Junsheng Cheng. “Opinion observer: analyzing and comparing opinions on the Web”. In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 342–351.

[137] Tie-Yan Liu et al. “Support vector machines classification with a very large-scale taxonomy”. In: *ACM SIGKDD Explorations Newsletter* 7.1 (2005), pp. 36–43.

[138] Zhu Liu, Jincheng Huang, and Yao Wang. “Classification TV programs based on audio information using hidden Markov model”. In: *Multimedia Signal Processing, 1998 IEEE Second Workshop on*. IEEE. 1998, pp. 27–32.

[140] Cheng Lu, Mark S. Drew, and James Au. “Classification of summarized videos using hidden Markov models on compressed chromaticity signatures”. In: *Proceedings of the ninth ACM international conference on Multimedia*. ACM. 2001, pp. 479–482.

[149] Yutaka Matsuo and Mitsuru Ishizuka. “Keyword extraction from a single document using word co-occurrence statistical information”. In: *International Journal on Artificial Intelligence Tools* 13.01 (2004), pp. 157–169.

[152] G. Harry McLaughlin. “SMOG grading: A new readability formula”. In: *Journal of reading* 12.8 (1969), pp. 639–646.

[153] Gilad Mishne and Natalie Glance. “Leave a reply: An analysis of weblog comments”. In: *Third annual workshop on the Weblogging ecosystem*. 2006.

[159] Shiva Naidu. "Evaluating the usability of educational websites for children". In: *Usability News* 7.2 (2005).

[162] Jakob Nielsen. "Kids' corner: Website usability for children". In: *Jakob Nielsen's Alertbox* (2002).

[163] Ofcom. *UK children's media literacy: Research Document*. http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf. 2010.

[164] Andrei Oghina et al. "Predicting imdb movie ratings using social media". In: *Advances in Information Retrieval*. Springer, 2012, pp. 503–507.

[165] Paul Ogilvie and Jamie Callan. "Combining document representations for known-item search". In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM. 2003, pp. 143–150.

[167] Bo Pang and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts". In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2004, p. 271.

[174] Xiaoguang Qi and Brian D. Davison. "Web page classification: Features and algorithms". In: *ACM Computing Surveys (CSUR)* 41.2 (2009), p. 12.

[175] Zheng Qifu et al. "Support Vector Machine Based on Universal Kernel Function and Its Application in Quantitative Structure-Toxicity Relationship Model". In: *Information Technology and Applications, 2009. IFITA'09. International Forum on*. Vol. 3. IEEE. 2009, pp. 708–711.

[178] Clare Richards. *Crowdflower: The world is at work. Right now.* http://crowdflower.com/. 2014.

[182] Stephen E. Robertson, Hugo Zaragoza, and Michael Taylor. "Simple BM25 extension to multiple weighted fields". In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM. 2004, pp. 42–49.

[188] John Schacter, Gregory KWK Chung, and Aimée Dorr. "Children's Internet searching on complex problems: performance and process analyses". In: *Journal of the American society for Information Science* 49.9 (1998), pp. 840–849.

[194] Sarah E. Schwarm and Mari Ostendorf. "Reading level assessment using support vector machines and statistical language models". In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2005, pp. 523–530.

[198] Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. "Automatic video tagging using content redundancy". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 395–402.

[199] Stefan Siersdorfer et al. “How useful are your comments?: analyzing and predicting youtube comments and comment ratings”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 891–900.

[211] Vakkalanka Suresh et al. “Content-based video classification using support vector machines”. In: *Neural Information Processing*. Springer. 2004, pp. 726–731.

[219] Takashi Tomokiyo and Matthew Hurst. “A language model approach to keyphrase extraction”. In: *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*. Association for Computational Linguistics. 2003, pp. 33–40.

[230] Yao Wang, Zhu Liu, and Jin-Cheng Huang. “Multimedia content analysis-using both audio and visual clues”. In: *Signal Processing Magazine, IEEE* 17.6 (2000), pp. 12–36.

[231] Ellen A. Wartella, Elizabeth A. Vandewater, and Victoria J. Rideout. “Introduction: electronic media use in the lives of infants, toddlers, and preschoolers”. In: *American Behavioral Scientist* 48.5 (2005), p. 501.

[232] Christian Wartena, Rogier Brussee, and Wout Slakhorst. “Keyword extraction using word co-occurrence”. In: *Database and Expert Systems Applications (DEXA), 2010 Workshop on*. IEEE. 2010, pp. 54–58.

[237] Lei Wu et al. “Distance metric learning from uncertain side information with application to automated photo tagging”. In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 135–144.

[243] Wai Gen Yee et al. “Are web user comments useful for search”. In: *Proc. LSDS-IR* (2009), pp. 63–70.

5

Human Computation for Relevance Estimation

Alone we can do so little.
Together we can do so much.

-Helen Keller, Author

Many scientific fields, including information retrieval, artificial intelligence, machine translation or natural language processing rely heavily on large-scale corpora for system building, training and evaluation. The traditional approach to acquiring these data collections is employing human experts to annotate or create the relevant material. A well-known example from the area of information retrieval is the series of extensive corpora created in the context of the NIST's *Text REtrieval Conference* (TREC) (see, e.g., [88]). Since the manual creation of such resources typically requires substantial amounts of time and money, there have long been advances into using automatically generated or extracted resources, examples are given by [180], [131], [204], or, [15]. While there are several promising directions and methods that are reported to correlate well with human judgements, for many precision-sensitive applications, experts such as Marcus et al. [148] still find manual annotation efforts indispensable.

Especially in novel or niche areas of research for which there are few or no existing resources that could be re-used, the demand for an alternative way of data acquisition becomes apparent. With the advent of commercial crowdsourcing, a new means of satisfying this need for large-scale human annotations emerged. Starting in 2005, Ama-

zon Mechanical Turk[1] and others provide platforms on which task *requesters* can reach a large number of freelance employees, the so-called *workers*, to solve *human intelligence tasks* (HITs). The payment is typically done on micro level, e.g., a few US cents per quickly solvable HIT. This process is now widely accepted and, following the pioneering work of Ambati et al. [13] and Kittur et al. [111], represents the basis for data collection, resource annotation or validation in many recent research projects.

In the light of these recent developments, this chapter will dedicate some room to discussing human computation for relevance estimation. In particular, we will discuss how to ensure the quality, timeliness and cost-efficiency of such data collection efforts. Section 5.1 investigates current challenges in commercial crowdsourcing and presents insights into overcoming them by means of task and interface design. In Section 5.2, we will look into *gamification*, a paradigm that phrases data collection and annotation tasks in an entertaining way with the goal of eliciting greater user engagement and immersion while collecting valuable data.

5.1. Crowdsourcing

Over the years, crowdsourcing became substantially more popular among both requesters and workers. As a consequence, the composition of the relatively small initial crowd of workers that was mainly attracted by the prospect of completing odd or entertaining tasks as a diversion, changed. Nowadays, according to [22], [106], and, [183], the number of users who are mainly attracted by the monetary reward represents a significant share of the crowd's workforce. At the same time, we observe a significant share of *cheaters* entering the market. Those workers try to maximise their financial gain by submitting quick, generic, non-reflected answers that, in turn, result in weak or altogether compromised corpus quality. In response to this trend, research work based on crowdsourcing nowadays has to pay careful attention to monitoring result quality. The accepted way of addressing cheat submissions is the use of high quality gold standard data or inter-annotator agreement ratios to check on, and, if necessary, reject deceivers. Here, we present an alternative approach by designing HITs that are less attractive for cheaters. Based on the experience gained from several previous crowdsourcing tasks and a number of dedicated experiments, this work aims to quantify the share of deceivers as well as to identify criteria and methods to make tasks more robust against this new form of annotation taint.

Despite the fact that crowdsourcing is being widely used to create and aggregate data collections for scientific and industrial use, the current amount of research work dedicated to methodological evaluations of crowdsourcing is relatively limited. One of the early studies by Sorokin and Forsyth [207] investigated the possibility of using crowdsourcing for image annotation. They found an interesting non-monotonic dependency between the assigned monetary reward per HIT and the observed result quality. While very low pay resulted in sloppy work, gradually increasing the reward improved annotation quality up to a point where further increases even deteriorated performance due to attracting more cheaters. In the same year, Kittur et al. [111] published their influential overview on the importance of task formulation to obtaining good results. Their main

[1] http://mturk.com

conclusion was that a task should be given in such a way, that cheating takes approximately the same time as faithful completion. The authors additionally underline the importance of clearly verifiable questions in order to reject deceivers.

In the course of the following year, several independent research groups such as Snow et al. [203] and Hsueh et al. [97] studied the performance of experts versus non-experts for various natural language processing tasks such as paraphrasation, translation or sentiment analysis. The unanimous finding, also confirmed by Alonso and Mizzaro [11], was, that a single expert is typically more reliable than a single non-expert. However, aggregating the results of several cheap non-experts, the performance of an expensive professional can be equalled at significantly lower cost. In the same year, Little et al. released TurkIt [134], a framework for iterative programming of crowdsourcing tasks. In their evaluation, the authors mention relatively low numbers of cheaters. This finding is somewhat conflicting with most publications in the field, that report higher figures. We suspect that there is a strong connection between the type of task at hand and the type of workers attracted to it. Later on, we will carefully investigate this dependency through a series of dedicated experiments.

There is a line of work dedicated to studying HIT design in order to facilitate task understanding and worker efficiency. Examples are Khanna et al [109]'s investigation of the influence of HIT interface design on Indian workers' ability to successfully finish a HIT or Grady and Lease [84]'s study of human factors in HIT design. Our own interface-related study inspects a very different angle by using interface design as a means of making cheating less efficient and therefore less attractive.

We can conclude that there are numerous influential publications that detail tailor-made schemes to identify and reject cheaters in various crowdsourcing scenarios. Snow et al. [203] do not treat cheaters explicitly, but propose modelling systematic worker bias and subsequently correcting for it. For their sentiment analysis of political blog posts, [97] rely on a combination of gold standard labels and majority voting to ensure result quality. Soleymani et al. [205] use a two-stage process. In a first round, the authors offer a pilot HIT as recruitment and manually invite well-performing workers for the actual task. Hirth et al. [94] describe a sophisticated workflow in which one (or even potentially several) subsequent crowdsourcing step is used in order to check on the quality of previously crowdsourced results. In her recent work, Gabriella Kazai discusses how the HIT setup influences result quality, for example through pay rate, worker qualification or worker type (see [107] and [108]).

Here, we take a different approach from the state of the art by (1) Aiming at discouraging cheaters rather than detecting them. While there is extensive work on the posterior identification and rejection of cheaters, we deem these methods as sub-optimal as they bind resources such as time or money. Instead, we try to find out what makes a HIT look appealing to cheaters and subsequently aim to remedy these aspects. (2) While there are many publications *also* detailing the authors' cheater detection schemes, we are not aware of comprehensive works on cheat robustness that are applicable to a wide range of HIT types. By giving a broad overview of frequently encountered adversarial strategies as well as established countermeasures, we hope to close this gap.

How to Cheat

Before proceeding to our experimentally supported inspection of various cheat countering strategies, we will spend some thought on the nature of cheating on large-scale crowdsourcing platforms. The insights presented here are derived from related work, discussions with peers, as well as our own experience as HIT requesters. They present, what we believe is an overview of the most frequently encountered adversarial strategies in the commercial crowdsourcing field. While one could theorize about many more potential exploits, especially motivated by the information security domain (e.g., [170], [157]), we try to concentrate on giving an account of the main strategies HIT designers have to face regularly.

As we will show in more detail, the cheaters' methods are typically straightforward to spot for humans, but, given the massive HIT volume, such a careful manual inspection is not always feasible. Cheating as a holistic activity can be assumed to follow a breadth-first strategy in that the group of cheating workers will explore a wide range of naive cheats and move on to a different HIT if those prove to be futile. When dealing with cheating in crowdsourcing, it is important to take into consideration the workers' different underlying motivations for taking up HITs in the first place [102]. We believe that there are two major types of workers with fundamentally different motivations for offering their work force. Entertainment-driven workers primarily seek diversion by taking up interesting, possibly challenging, HITs. For this group, the financial reward plays a minor role. The second group are money-driven workers. These workers are mainly attracted by monetary incentives. We expect the latter group to contain more cheating individuals as an optimization of time efficiency and, subsequently, an increased financial reward, is clearly appealing given their motivation. Here, we also regard any form of automated HIT submission, i.e., bots, scripts, etc. to originate from money-driven workers. We were able to get an interesting insight into the organization of the money-driven crowdsourcing subculture when running a HIT for the child suitability corpus described in Section 4.1 that involved filling a survey with personal information. For this HIT, we received multiple submissions by seemingly unique workers that contained largely contradictory statements about fundamental credentials such as gender, age or education level. We suspect these workers to be organised in large-scale offices from where multiple individuals connect to the platform under the same worker id. While this rather anecdotal observation is not central to our work and demands further evidence in order to be quantifiable, we consider it an interesting one, that is worth sharing with the research community.

Our following overview of cheating approaches will be organised according to the types of HITs and quality control mechanisms they are aimed at.

Closed-class Questions Closed-class questions are probably the most frequently used HIT elements. They require the worker to choose from a limited, pre-defined list of options. Common examples of this category include radio buttons, multiple choice question, check boxes and sliders. There are two widely encountered cheating strategies targeting closed-class tasks: (1) Arbitrarily picked answers can often easily be rejected by using good gold standard data or by inspecting agreement with redundant submissions by multiple workers, either in terms of majority votes or more sophisticated combina-

tion schemes [60]. (2) Clever cheaters may learn from previous HITs and devise educated guesses based on the answer distribution underlying the HIT. An example could be the typically sparsely distributed topical relevance in Web search scenarios for which a clever cheater might learn that arbitrarily selecting only a very small percentage of documents closely resembles meaningful judgements. This is often addressed by introducing a number of very easy to answer gold standard awareness questions. A user that fails to answer those questions can be immediately rejected as he is clearly not trying to produce sensible results.

Open-class Questions Open questions allow workers to provide less restricted answers or perform creative tasks. The most common example of this class are text fields, but it potentially includes draw boxes, file uploads or similar. Focussing on the widely used text fields, there are three different forms of cheats: (1) Leaving the field blank can be prohibited during HIT interface design. (2) Entering generic text blocks is easily detectable if the same text is used repeatedly across questions. (3) Providing unique (sometimes even domain-specific) portions of natural language text copied from the Web is very hard to detect automatically.

Internal Quality Control Most current large-scale crowdsourcing platforms collect internal statistics of the workers' reliability in order to fend off cheaters. Reliability is, to the best of our knowledge, measured by all major platforms in terms of the worker's share of previously accepted submissions. There are two major drawbacks of this approach: (1) Previous acceptance rates fail to account for the high share of submissions that are uniformly accepted by the HIT provider and are post-processed and filtered, steps, to which the platform's reputation system has no access. (2) According to [103], previous acceptance rates are prone to gaming strategies such as rank boosting in which the worker simultaneously acts as a HIT requester. He can then artificially boost his reliability by requesting and submitting small HITs. This gaming scheme is very cheaply implementable as the cycle only loses the service fee deducted by the crowdsourcing platform.

In addition to these theoretical considerations concerning the shortcomings of current quality control mechanisms, later on, we will show an empirical evaluation backing the assumption that we need better built-in quality measures than prior acceptance rates can deliver.

External Quality Control Some very interactive HIT types may require more sophisticated technical means than offered by most crowdsourcing platforms. During one of our early experiments, we dealt with this situation by redirecting workers to an external, less restricted Web page on which they would complete the actual task and receive a confirmation code to be entered on the original crowdsourcing platform. Despite this openly announced completion check, workers tried to issue made-up confirmation codes, to resubmit previously generated codes multiple times or to submit several empty tasks and claim that they did not get a code after task completion. While such attempts are easily fended off, they offer a good display of deceiver strategies. They will commonly try out a series of naive exploits and move on to the next task if they do not succeed. Section 5.2 will describe an example of such an external HIT in more detail.

Experiments

After our discussion of adversarial approaches and common remedies, we will give a quantitative experimental overview of various cheating robustness criteria of crowdsourcing tasks. The starting point of our evaluation are two very different HITs that we originally requested throughout 2010 and that showed substantially different cheat rates. The first task is a straightforward binary relevance assessment between pairs of Web pages and queries. The second task asked the workers to judge Web pages according to their suitability for children of different age groups and to fill a brief survey on their experience in guiding children's Web search as described in Section 4.1. The interfaces of both tasks can be found in the appendix.

All experiments were run through CrowdFlower[2] in 2010 and 2011. The platform incorporates the notion of "channels" to forward HITS to third party platforms. To achieve broad coverage and results representative of the crowdsourcing market, we chose all available channels, which at that time were Amazon Mechanical Turk[3] (AMT), Gambit[4], SamaSource[5] as well as the GiveWork smartphone application jointly run by Samasource and CrowdFlower. Table 5.1 shows the overall distribution of received submissions according to the channels from which they originated. The figures are reported across all HITs collected for this study, as there were no significant differences in distribution between HIT types. The vast majority of submissions came from AMT. We are not are not aware of the reason why we did not receive any submissions from the GiveWork app. HITs were offered in units of 10 at a time with initial batch sizes of 50 HITS. Each HIT was issued to at least 5 independent workers. Unless stated differently, all HITs were offered to unrestricted audiences with the sole qualification of having achieved 95% prior HIT acceptance, the default setting on most platforms. The monetary reward per HIT was set to 2 US cents per relevance assessment and 5 US cents per filled Web page suitability survey. We did not change the reward throughout this work. Previous work, e.g., by Harris [89], has shown the influence of different financial incentive models on result quality. Statistical significance of results was determined using a Wilcoxon Signed Rank test with $\alpha < 0.05$.

A key aspect of our evaluation is identifying cheaters. There is a wide range of indicators for this purpose, including: (1) Agreement with trusted gold standard data can be used to measure the general quality of an answer. (2) Agreement with other workers enables us to identify hard tasks on which even honest workers occasionally fail. (3) HIT completion times (either compared per HIT or HIT type) give an estimate of how much effort the worker put into the task. (4) Simple task awareness questions that are correctly and unambiguously answerable for every worker can be introduced to identify distracted or cheating individuals. Mistakes on this kind of question are typically penalized heavily in cheater detection schemes.[6] The concrete scheme chosen in this work will be formally detailed in the following section.

Our analysis of methods to increase cheating robustness was conducted along four

[2] http://www.crowdflower.com
[3] http://www.mturk.com
[4] http://getgambit.com/
[5] http://samasource.org/
[6] The original suggestion of this trick resulted from personal communication with Mark Sanderson and William Webber.

Table 5.1: Submission distribution for all HITs.

Channel	Absolute	Relative share
Amazon Mechanical Turk	4285	85%
Samasource	454	9%
Gambit	303	6%
GiveWork	0	0%

research questions: 1. How does the concrete task type influence the number of observed cheaters? 2. Does interface design affect the share of cheaters? 3. Can we reduce fraudulent tendencies by explicitly filtering the crowd? 4. Is there a connection between the size of HIT batches and observed cheater rates?

Before beginning our inspection of different strategies to fend off cheaters in crowdsourcing scenarios, let us dedicate some further consideration to the definition of cheaters. Following our previous HIT experience, we can extend the worker classification scheme by Kittur et al. [111] to identify several dysfunctional worker types. *Incapable workers* do not fulfil all essential requirements needed to create high quality results. *Malicious workers* try to invalidate experiment results by submitting wrong answers on purpose. *Distracted workers* do not pay full attention to the HIT which often results in poor quality. The source of this distraction will vary across workers and may be of external or intrinsic nature. The exclusively money-driven cheater introduced earlier falls into the third category, as he would be theoretically capable of producing correct results but is distracted by the need to achieve the highest possible time efficiency. As a consequence, we postulate the following formal definition of cheaters for all subsequent experiments in this work:

Definition 1. *A cheater is a worker who fails to correctly answer simple task awareness questions or who is unable to achieve better-than-random performance given a HIT of reasonable difficulty.*

In the concrete case of our experiments, we measure agreement as a simple majority vote across a population of at least 5 workers per HIT. Disagreeing with this majority decision for at least 50% of the questions asked will flag a worker as a cheater. Additionally, we inject task awareness questions that require the worker to indicate whether the current document is written in a non-English language (each set of 10 judgements that a worker would complete at a time would contain one known non-English page). Awareness in this context represents that the worker actually visited the Web page that he is asked to judge. Failing to answer this very simple question will also result in being considered a cheater. The cheater status is computed on task level (i.e., across a set of 10 judgements in our setting) in order to result in comparable reliability. Computing cheater status on batch level or even globally would serve for very strict labels as a single missed awareness question would brand someone a cheater even if the remainder of his work in the batch or the entire collection were valuable. Our approach can be considered lenient as cheaters are "pardoned" at the end of each task. Our decision is additionally motivated by the fact that the aim of this study is to gauge the proportion of cheaters attracted by a given HIT design rather than achieving high confidence at iden-

Table 5.2: Task-dependent share of cheaters before and after using gold standard data.

Task	before gold	after gold
Suitability	2.2%	1.6%
Relevance	37.3%	28.4%

tifying individuals to be rejected in further iterations. At the same time, we are confident that a decision based on 10 binary awareness questions and the averaged agreement across 10 relevance judgements produces reliable results that are hard to bypass for actual cheaters.

This approach appears reasonable as also it focuses on workers that are distracted to the point of dysfunctionality. In order to not be flagged as a cheater, a worker has to produce at least mediocre judgements and not fail on any of the awareness questions. Given the relative simplicity of our experiments we do not expect incapability to be a major hindrance for well-meaning workers. Truly malicious workers, finally, can be seen as a rather theoretical class given the large scale of popular crowdsourcing platforms on the Web. This worker type is expected to be more predominant in small-scale environments where their activity has higher detrimental impact. We believe that our definition is applicable in a wide range of crowdsourcing scenarios due to its generality and flexibility. The concrete threshold value of agreement to be reached as well as an appropriate type of awareness question should be selected depending on the task at hand.

5.1.1. Task-dependent Evaluation

As a first step into understanding the dynamics of cheating on crowdsourcing platforms, we compare the baseline cheater rate for the two previously introduced HIT types. The main differences between the two tasks are task novelty and complexity. Plain relevance judgements are frequently encountered on crowdsourcing platforms and can be assumed to be well-known to a great number of workers. Our suitability survey, on the other hand, is a novel task that requires significantly more consideration. Directly comparing absolute result quality across tasks would not be meaningful due to the very different task-inherent difficulty. Table 5.2 shows the observed share of cheaters for both tasks before and after comparing answers to gold standard data. We can find a substantially higher cheater rate for the straightforward relevance assessment. The use of gold standard data reduced cheater rates for both tasks by a comparable degree (27.3% relative reduction for the suitability HIT and 24% for the relevance assessments). With respect to our first research question, we note that more complex tasks that require creativity and abstract thinking attract a significantly lower percentage of cheaters. We assume this observation to be explained by the interplay of two processes: (1) Money-driven workers prefer simple tasks, that can be easily automated, over creative ones. (2) Entertainment-seekers can be assumed to be more attracted towards novel, enjoyable and challenging tasks. For all further experiments in this work, we will exclusively inspect the relevance assessment task as it has a higher overall cheater rate that is assumed to more clearly illustrate the impact of the various evaluated factors.

We have shown that innovative tasks draw a higher share of faithful workers that are

Table 5.3: Interface-dependent percentage of cheaters for variable queries, variable documents and fully variable pairs.

Interface type	Observed cheater rate
Variable queries	28.4%
Variable documents	21.9%
Both variable	18.5%

assumed to be primarily interested in diversion. However, in the daily crowdsourcing routine, many tasks are of rather straightforward nature. Now, we will evaluate how interface design can influence the observed cheater rate even for well-known and repetitive tasks such as image tagging or relevance assessments. According to Shneiderman [196], traditional interface design commonly tries to not distract the user from the task at hand. As a consequence, the number of context changes is kept as low as possible to allow focused and efficient working. While this approach is widely accepted in environments with trusted users, crowdsourcing may require a different treatment. A smooth interaction with a low number of context changes makes a HIT prone to automation, either directly by a money-driven worker or by scripts and bots. We investigate this connection at the example of our relevance assessment task.

Table 5.3 shows the results of the comparison. In the first step, we present the workers with batches of 10 Web page/query pairs using gold standard data. In order to keep the number of context changes to a minimum, we asked the workers to visit a single Web page[7] and create relevance judgements for that page given 10 different queries. The resulting share of cheaters turns out to be substantial (28.4%). Now we increase the amount of interaction in the HIT by requiring the worker to create 10 judgements for query/document pairs in which we keep the query constant and require visiting 10 unique Web pages. Under this new setting the worker is required to make significantly more context changes between different Web pages. While in a controlled environment with trusted annotators this step would be counterproductive, we see a significant decline by 23%, giving a total proportion of 21.9% cheaters. In a final step, we issue batches of 10 randomly drawn query/document pairs. As a result, the proportion of cheaters decreases by another 15% to 18.5%. The general HIT interface remains unchanged from the original example shown in Figure 5.1, only the combinations of query/document pairs vary. With respect to our second research question, we find that greater variability and more context changes discourage deceivers as the task appears less susceptible to automation or cheating, and therefore less profitable.

Crowd Filtering At this point, we will address our third research question by inspecting a number of commonly used filtering strategies to narrow down the pool of eligible workers that can take up a HIT. In order to make for a fair comparison, we will regard two settings as the basis of our juxtaposition: (1) The initial cheat-prone relevance assessment setup with 10 queries and 1 document, using gold standard verification questions.

[7] The pages used in this study originate from the ClueWeb'09 collection[8] and the queries and gold standard judgements for topics 51-57 from NIST's TREC 2010 Web track adhoc task, described by Clarke [51].

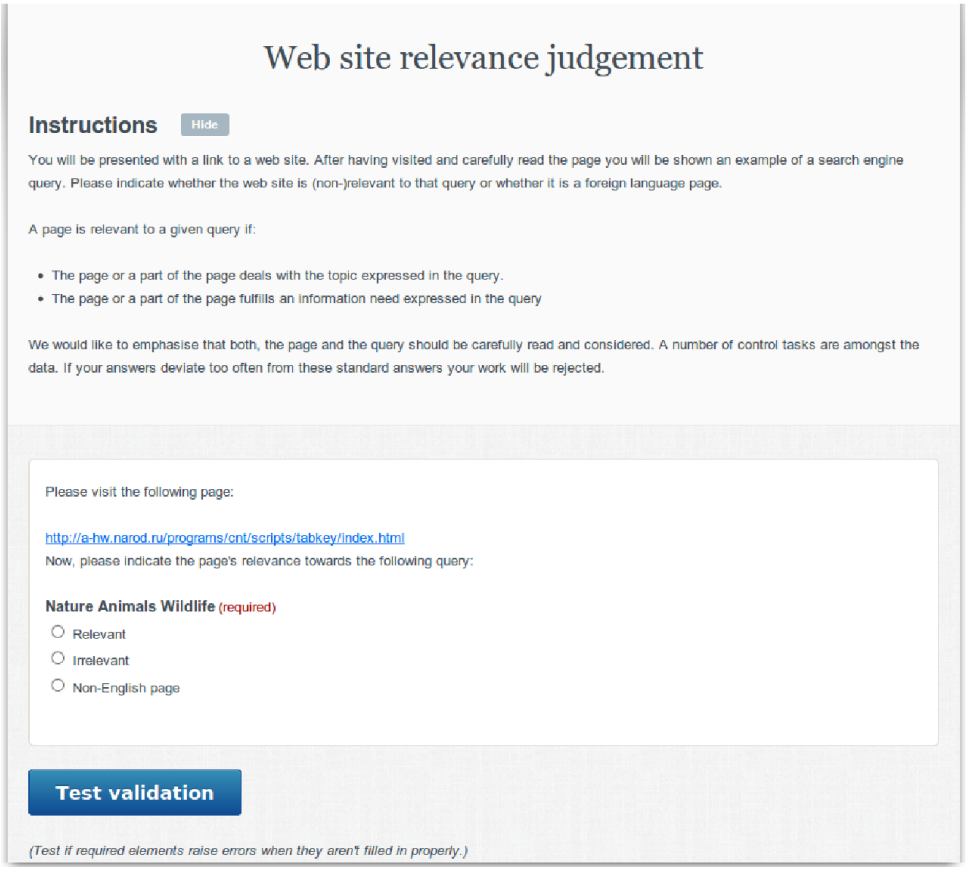

Figure 5.1: Relevance judgement HIT

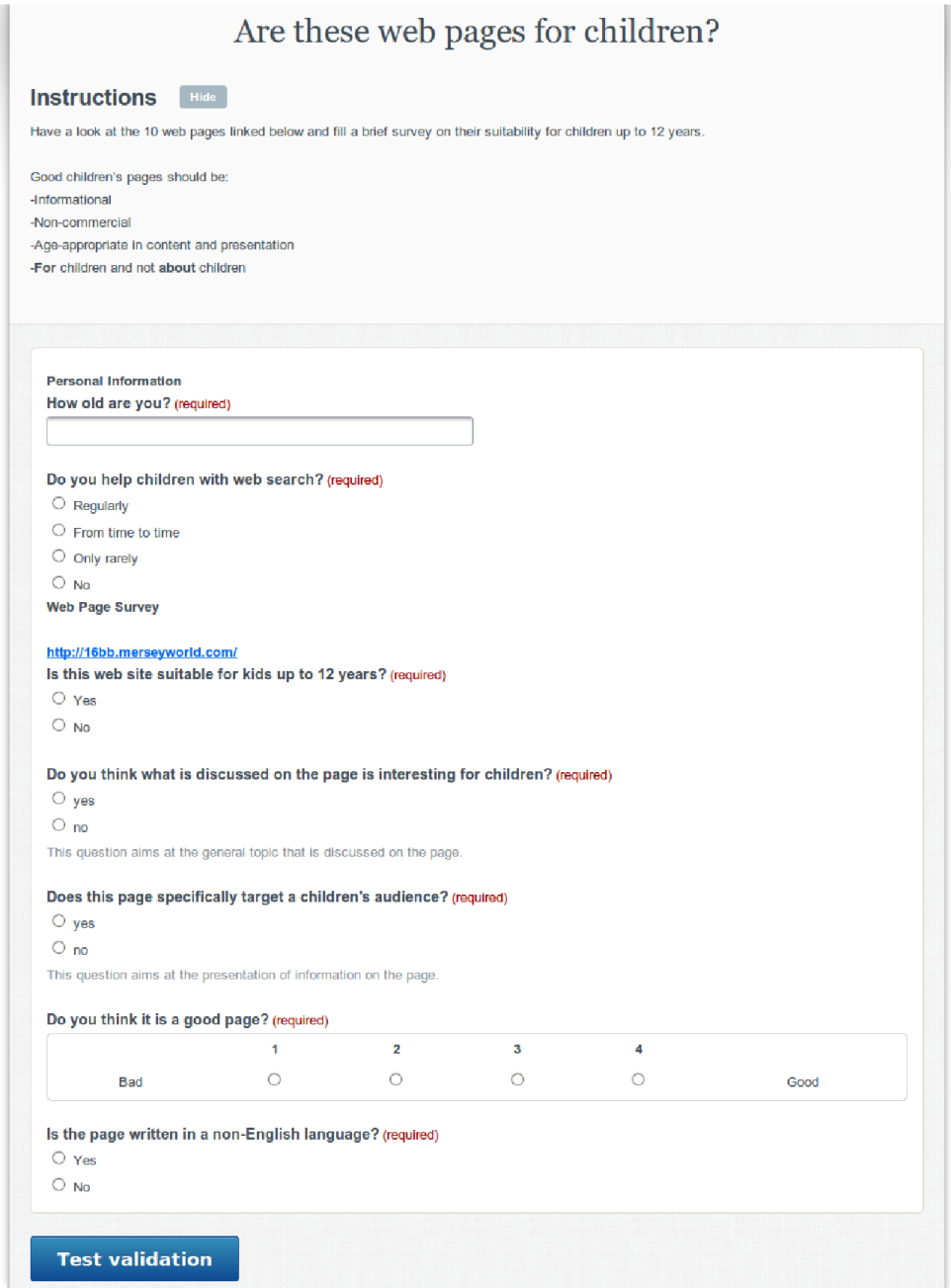

Figure 5.2: Web page Suitability HIT

(2) The previous best performance that was achieved using gold standard verification sets as well as randomly drawn query/document pairs as described earlier. Please note that the crowd filtering experiments were exclusively run on AMT, since not all previously used platforms offer the same filtering functionalities. The HITs created for this experiment are not shown in Table 5.1, as they would artificially boost AMT's prominence even further. We will investigate three filtering instruments:

Prior Acceptance Rate Earlier on, we argued that the widely used prior acceptance rates are not an optimal means of assessing worker reliability. In order to evaluate the viability of our hypothesis, we increase the required threshold accuracy to 99% (The default setting is 95%). Given a robust reliability measure, this, at least seemingly, very strict standard should result in highest result quality.

Country of Origin Offering the HIT exclusively to inhabitants of certain countries or regions is a further commonly-encountered strategy for fending off cheaters and sloppy workers. Following our model of money-driven and entertainment-driven workers, we assume that offering our HITs to inhabitants of developed countries should result in lower cheater rates. In order to evaluate this assumption, we repeat the identical HIT that was previously offered unrestrictedly, on a crowd restricted to US workers.

Previous Recruitment Recruitment (sometimes also called qualification) HITs are a further means of *a priori* narrowing down the group of workers. In a multi-step process, workers are presented with preparatory HITs. Those workers that achieve a desired level of result quality are eligible to take up the actual HIT. In our case, we presented workers with the identical type of HIT that was evaluated later and accepted every worker that did not qualify as a cheater (according to Definition 1) for the final experiment.

The first two columns of Table 5.4 show an overview of the three evaluated filtering dimensions. Raising the threshold of prior acceptance from the 95% default to 99% only gradually affected the observed cheater rate. Filtering depending on worker origin was able to cut cheater rates down to less than a third of the originally observed 28.4%. However, this substantial reduction comes at a cost. The run time of the whole batch increased from 5 hours to slightly under one week, as we limit the crowd size. Providers of time-sensitive or very large HIT batches may have to consider this trade-off carefully. The introduction of a recruitment step prior to the actual HIT was able to reduce the cheater rate, however, ratio of cheat reduction vs. increase in completion time is worse than for origin-based filtering. To further confirm and understand these trends, columns 3 and 4 of the same table display the same statistics for the varied HIT setting in which we assigned random query/document pairs. In general, the effect of filtering turned out to be largely independent of the previously applied interface changes. The relative merit of the applied methods was found to be comparable for both the initial and the high-variance interface.

The conclusion towards our third research question is twofold: (1) We have seen how prior crowd filtering can greatly reduce the proportion of cheaters. This narrowing down of the workforce may however result in longer completion times. (2) Additionally, we

Table 5.4: Effect of crowd filtering on cheater rate and batch processing time.

	initial		varied	
Filtering method	Cheaters	Time	Cheaters	Time
Baseline	28.4%	3.2hr	18.5%	5.2hr
99% prior acc.	26.2%	3.8hr	17.7%	7.6hr
US only	8.4%	140hr	5.4%	160hr
Recruitment	19%	145hr	12.2%	144hr

could confirm the assumption that a worker's previous task acceptance rate cannot be seen as a robust stand-alone predictor of his reliability.

The Influence of Batch Sizes Crowdsourced HITs are typically issued at large scale in order to collect significant amounts of data. Currently, HIT batch sizes are often adjusted according to practical or organizational needs but with little heed to result quality. Wang et al. [228] give a first intuition of an instrumental use of batch sizes by showing that small batches typically have longer per-HIT completion times than large ones. We assume that this tendency is explained by large HIT batches being more attractive for workers interested time-efficiency. A batch of only 2 HITs has a relatively large overhead of reading and understanding the task instructions before completing actual work. For large batches, workers have a significantly higher reuse potential. The same holds true for cheating. Investing time into finding a way to game a 5-HIT batch is far less attractive than doing the same for a batch of 100 HITs. As a consequence, we expect large HIT batches to attract relatively higher cheater rates than small batches. Previously, all HITs were offered in batches of 50. In order to evaluate our hypothesis, we issued several batches of relevance assessment HITs (see Figure 5.1) and compared the observed cheater rates conditioned on batch sizes. For each setting, we collected judgements for 100 query/document pairs. Except for the batch size, all experiment parameters were kept at the settings described in Section 5.1. Batches were requested one at a time. Only after a batch's completion would we publish the following one. In this way, we aim to avoid giving the impression that there was a large amount of similar HITs available to be preyed on by cheaters. As a consequence, we do not expect major external effects caused by the resulting higher number of batches offered as they are never available at the same time. Figure 5.3 shows the result of this comparison. The figure represents the mean observed cheater rate for each batch size s across a population of $n = \frac{100}{s}$ batches. We can note a statistically significant (using Wilcoxon signed Rank test with $\alpha < 0.05$) increase in cheating activity for batch sizes of at least 10 HITs. As a consequence of determining cheater status at task level, we do not expect any influence of the batch size on the confidence of our cheater detection since the number of HITs per task remained unchanged across batch size settings.

As a further piece of evidence, let us investigate how the cheater rate develops within a batch as HIT submissions arrive. The previously made observations would imply that, as the batch approaches completion, the arrival of new cheaters should become less frequent as the batch of available HITs shrinks in size. To pursue this intuition, workers

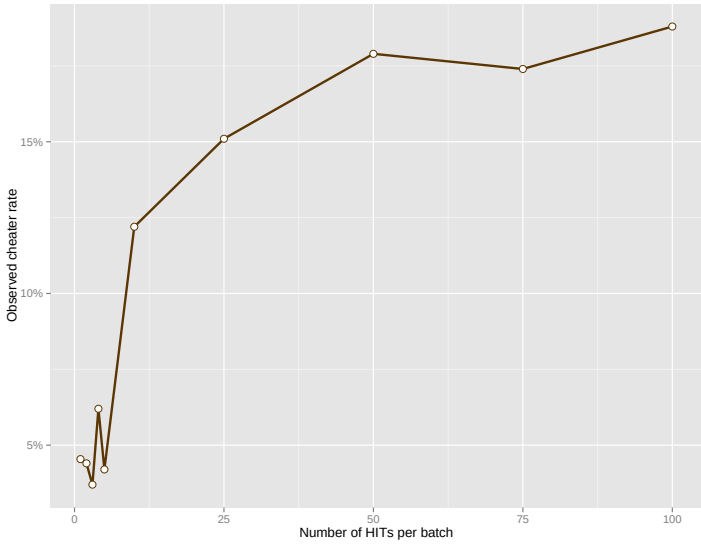

Figure 5.3: Observed percentage of cheaters for HIT batches of variable size.

were ordered according to their time of first submission to the batch. Subsequently, we determined the maximum likelihood estimate of encountering a new cheater, $p(c)$ given an original batch size and a degree of completion as:

$$p(c|s,\omega) = \frac{|C_{s,\omega}|}{|W_{s,\omega}|}$$

Where $|C_{s,\omega}|$ is the number of new cheaters observed for size s at degree of completion ω, and $|W_{s,\omega}|$ is the overall number of new workers arriving at that time. Figure 5.4 shows the resulting distributions. For sufficiently large s, we can clearly see our intuition confirmed. As the number of remaining HITs declines, new cheaters are observed less and less frequently. For settings of $s < 25$ the distributions are near uniform and we could not determine significant changes over time.

With respect to our fourth research question, we conclude that large batches indeed attract more cheaters, as they offer greater potential of automation or repetition. This finding holds interesting implications for HIT designers, who may consider splitting up large batches into multiple smaller ones.

Conclusion

In this section, we investigated various ways of making crowdsourcing HITs more robust against cheat submissions. Many state-of-the-art approaches to deal with cheating rely on posterior result filtering. We choose a different focus by trying to design and formulate HITs in such a way that they are less attractive for cheaters. The factors evaluated in this article are: (1) The HIT type. (2) The HIT interface. (3) The composition of the worker crowd. (4) The size of HIT batches.

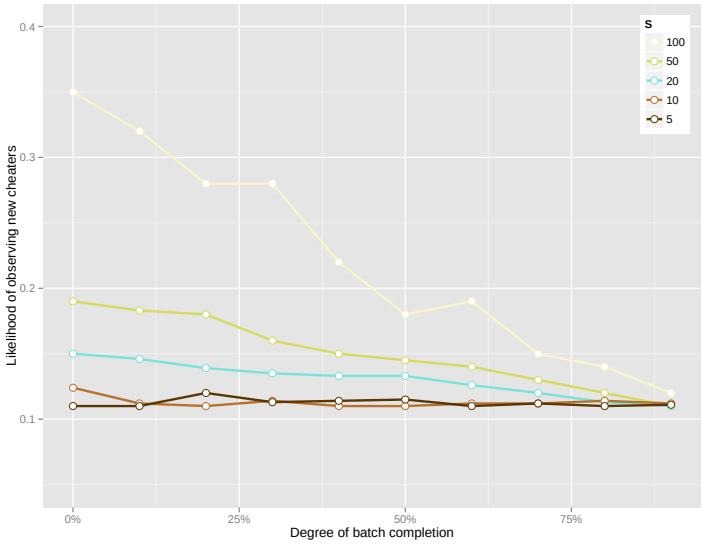

Figure 5.4: Likelihood of observing new cheaters as batches approach completion.

Based on a range of experiments, we note that cheaters are less frequently encountered in novel tasks that involve creativity and abstract thinking. Even for straightforward tasks we could achieve significant reductions in cheater rates by phrasing the HIT in a non-repetitive way that discourages automation. Crowd filtering could be shown to have significant impact on the observed cheater rates, while filtering by origin or by means of a recruitment step were able to greatly reduce the amount of cheating, the batch processing times multiplied. We are convinced that implicit crowd filtering through task design is a superior means to cheat control than excluding more than 80% of the available workers from accessing the HIT. An investigation of batch sizes further supported the hypothesis of fundamentally different worker motivations, as the observed cheater rates for large batches that offer a high reuse potential, were significantly higher than those for small batches.

Finally, our experiments confirmed that prior acceptance rates, although widely used, cannot be seen as a robust measure of worker reliability. Recently, we have seen a change in paradigms in this respect. In June 2011, Amazon introduced a novel service on AMT that allows to issue HITs to an selected crowd of trusted workers, so-called *Masters* (at higher fees). Master status is granted per task type and has to be maintained over time. For example, as a worker reliably completes a high number of image tagging HITs, he will be granted Master status for this particular task type. Currently, the available Master categories are “Photo Moderation” and “Categorization”. Due to the recency of this development, we were not able to set up a dedicated study of the performance-cost trade-off of Master crowds versus regular ones.

Novel features like this raise an important general question to be addressed by future work: Temporal instability is a major source of uncertainty in current crowdsourcing re-

search results. The crowdsourcing market is developing and changing at a high pace and is connected to the economical situation outside the cloud. Therefore, it is not obvious whether this year's findings about worker behaviour, the general composition of the crowd or HIT design would still hold two years from now. Besides empirical studies, we see a clear need for explicit models of the crowd. If we could build a formal representation of the global (or, depending on the application, local) crowd, including incentives and external influences, we would have a reliable predictor of result quality, process costs and required time at our fingertips, where currently the process is trial-and-error-based.

One particularly interesting aspect of such a model of crowdsourcing lies in a better understanding of worker motivation. On the basis of activity logs and usage histories, we can solicit more sophisticated worker reliability models. To this end, the following section will discuss the use of games in commercial crowdsourcing tasks. Following previous experience, we hypothesise that workers who are mainly entertainment-driven and for whom the financial reward only plays a subordinate role, are less likely to cheat during the task. Using games, such workers can be rewarded more appropriately by representing HITs in engaging and entertaining ways.

5.2. Gamification

As we saw in the previous section, in many time-insensitive applications, HIT providers restrict the crowd of workers to certain nationalities (in the earlier example, we included only US workers) who they trust will provide higher quality results. Although the approach is widely accepted and has been shown to significantly reduce the number of spam submissions, we believe that this uptake may be treating symptoms and not the actual underlying cause. Rather than attributing the confirmed performance differences between the inhabitants of different countries to their nationality, we hypothesize that there are 2 major types of workers with fundamentally different motivations for offering their workforce on a crowdsourcing platform: (1) *Money-driven* workers are motivated by the financial reward that the HIT promises. (2) *Entertainment-driven* workers primarily seek diversion but readily accept the financial incentives as an additional stimulus. We are convinced that the affiliation (or proximity) to one of those fundamental worker types can have a significant impact on the amount of attention paid to the task at hand, and, subsequently, on the resulting annotation quality. We realize that money-driven workers are by no means bound to deliver bad quality; however, they appear to be frequently tempted into sloppiness by the prospect of a higher time efficiency and therefore stronger satisfaction of their main motivation. Entertainment-driven workers, on the other hand, appear to work a HIT more faithfully and thoroughly and regard the financial reward as a welcome bonus. They typically do not indulge in simple, repetitive or boring tasks. We propose to more strongly focus on entertainment-driven workers by phrasing crowdsourcing problems in an entertaining and engaging way: As games. Csikszentmihalyi [57]'s theory of *flow*, a state of maximal immersion and concentration at which optimal intrinsic motivation, enjoyment and high task performance are achieved, further encouraged our design.

We intend to increase the degree of satisfaction entertainment-driven workers experience. This can lead to (a) higher result quality, (b) quicker batch processing rates, (c)

lower overall cheater rates, (d) better cost efficiency. An additional incentive for delivering high-quality results in a game scenario would be the element of competition and social standing among players. Taking into account recent behavioural analyses of online communities and games [120], entertainment seekers can be expected to put considerable dedication into producing high-quality results to earn more points in a game to progress into higher difficulty levels or a rank on the high score leader board.

In this section, we will describe a game-based approach to collecting document relevance assessments in both theory and design. Based on NIST-created TREC data, we conduct a large-scale comparative evaluation to determine the merit of the proposed method over state-of-the-art relevance assessment crowdsourcing paradigms. Venturing beyond "hard" quality indicators such as precision, cost-efficiency or annotation speed, we discuss a wide range of socio-economical factors such as demographics and alternative incentives to enhance a fundamental understanding of worker motivation. In a separate study, we demonstrate the generalizability of the proposed game to other tasks on the example of noisy image classification.

Document relevance assessments have been playing a central role in IR system design and evaluation since the early Cranfield experiments [225]. Explicit judgements of (the degree of) relevance between a document and a given topic are used as a proxy of user satisfaction. Based on such test collections, the results of retrieval systems can be compared without undergoing numerous iterations of user studies. Voorhees and Harman [226] look back on two decades of assessing document relevance in IR evaluation and benchmarking for NIST's Text Retrieval Conference (TREC). In order to be suitable for the evaluation of state-of-the-art Web-scale systems, the requirements in terms of size, topical coverage, diversity and recency that the research community imposes on evaluation corpora have been steadily rising. As a consequence, the creation and curation of such resources becomes more expensive.

The majority of crowdsourced tasks are plain surveys, relevance assessments or data collection assignments that require human intelligence but very little creativity or skill. An advance into bringing together the communities of online games and crowdsourcing is being made by the platform Gambit[9], that lets players complete HITs in exchange for virtual currency in their online gaming world. This combination, however, does not change the nature of the actual HIT carried out beyond the fact that the plain HIT form is embedded into a game environment. Instead, we propose using an actual game to leverage worker judgements.

A number of techniques have been designed to make participation in human computation efforts as engaging as possible. The perhaps most effective technique among these is a genre of serious games called *games with a purpose* (GWAP) [4] which have been developed with the focus of efficient and entertaining transformation of research data collection into game mechanics. By equating player success in the game with providing quality inputs, the idea is to extract higher-quality data than is currently done with dull repetitive tasks such as surveys. More than half a billion people worldwide play online games for at least an hour a day – and 183 million in the US alone [151]. Richards [179] finds the average American, for example, to have played 10,000 hours of video games by the age of 21. Channeling some of this human effort to gather data

[9] http://getgambit.com/

has shown considerable promise. People engage in these GWAPs for the enjoyment factor, not with the objective of performing work. Successful GWAPs include the ESP Game [5], which solicits meaningful, accurate image labels as the underlying objective; Peekaboom [7], which locates objects within images; Phetch [8], which annotates images with descriptive paragraphs; and Verbosity [6], which collects common-sense facts in order to train reasoning algorithms. The typical research objective of these GWAPs is to have two randomly-selected players individually assign mutually-agreed document labels, with the game mechanics designed to reward uncommon labels. In contrast, the game mechanics of our proposed game is to encourage and reward consensus labeling. Ma et al. [142]'s Pagehunt game presented players with a Web page and asked them to formulate a query that would retrieve the given page in the top ranks on a popular search engine to investigate the findability of Web pages. While task-specific games have been shown to be engaging means of harnessing the players' intelligence for a certain research goal, there has not been a formal investigation of the merits of game-based HITs over conventional ones. Additionally, current GWAPs are typically highly tailored towards a certain (often niche) problem at hand and do not lend themselves for application across domains. We will demonstrate the generalizability of our approach in Section 5.2.

Methodology

At this point, we will introduce the annotation game as well as the necessary pre- and post-processing steps in order to acquire standard topic/document relevance assessments. Careful attention will be paid to highlighting motivational aspects that aim to replace HIT payment by entertainment as a central incentive.

The central concept of our proposed game is to require players to relate items to each other. In order to preserve its general applicability, we tried to make as few as possible assumptions about the nature of those items. The game shows $n = 4$ concept buckets $b_1 \dots b_n$ at the bottom of the screen. From the top, a single item i slides to the bottom, and has to be directed into one of the buckets by the player. Doing so expresses a relationship between i and b_j. Additional information about i can be found in an info box in the top left corner.

In the case of document relevance assessments, the concept buckets b_j display topic titles, item i is a keyword from a document and the info box displays the context in which the keyword appears in that document. Figure 5.5 shows a screenshot of our game. A live version can be found online[10]. For each assigned relation between i and b_j the player is awarded a number of points. The score in points is based on the degree of agreement with other players. In addition to this scheme, the game shows a tree that grows a leaf for every judgement that consents with the majority decision. A full tree awards bonus points. In this way, we reward continuous attention to the game and the task. As a final element, the game is divided into rounds of 10 judgements each. After each round, the speed with which item i moves is increased, making the task more challenging to create additional motivation for paying close attention.

After up to 5 rounds (the player can leave the game at any point in time before that) the game ends and the achieved points as well as the player's position in our high-score

[10] http://www.geann.org

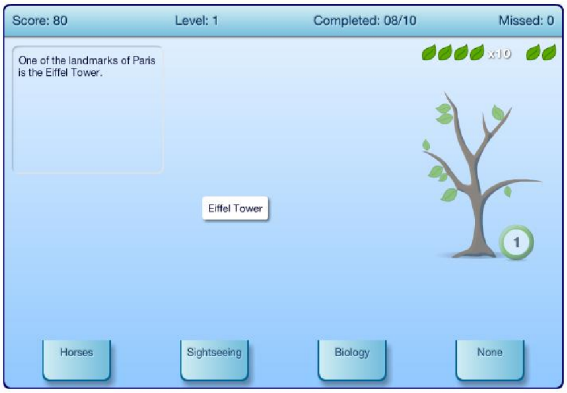

Figure 5.5: A screenshot of the annotation game in which the keyword "Eiffel Tower" has to be related to one of a number of concepts.

leader board are shown. Together with the log-in concept, this aims to encourage replaying as people want to advance into the higher ranks of the leader board.

In order to use the envisioned game with TREC data for relevance assessments, a number of pre-processing steps are required. All relevant game information is stored in a relational database from which items, concepts and additional information are drawn and into which user judgements are stored, subsequently. We assume a list of pairs p consisting of a query q and a document d for which we will collect user judgements.

For every d, we extract the textual Web page content and break it up into a set S of sentences $s_{d,1}...s_{d,|S|}$ using the LingPipe library[11]. In the following step, S is reordered by a ranking function $r(s)$, based on decreasing mean inverse document frequency (idf) of sentences s with a number of $|s|$ constituent terms t_n. $|C|$ denotes the number of documents in the collection and $df(t)$ is the number of documents containing term t. In this way, we promote the most salient and informative sentences in the document. The underlying $idf(t)$ statistics for this step are computed collection-wide.

$$idf(t) = \frac{|C|}{df(t)+1}$$

$$r(s) = \frac{1}{|s|}\sum_{n=1}^{|s|} idf(t_n)$$

Finally, the k highest-ranking sentences (those with the highest scores of $r(s)$) are selected and used in the game. The concrete setting of k depends on the size of document d and was set to $0.1|S|$ in order to account for different document lengths. Higher settings of k result in a higher judgement density per document. For each of the selected sentences, we extract the single highest-idf term t_n as sliding keyword i, and the full sentence as context information to be shown in the top left corner. The concept buckets b include the original query q, two randomly selected topics from the database, as well

[11] http://http://alias-i.com/lingpipe/

as an "other" option to account for none of the offered concepts being related to i. The buckets are shown in random order to prevent selection biases.

As a final step, we have to transform the players' conceptual associations into document-wide relevance judgements. Each player annotation a can be understood as a quintuple $a = (p_a, i_a, s_a, c_a, r_a)$ in which player p associated item i occurring in context sentence s with concept c in round r of the game. In a first step, we map all associations a to relevance votes. We interpret associations of any $s \in d$ to the concept of the original query q as a player's binary relevance vote $v_{p,s,q,r}$ between sentence s and query q as described in Equation 5.1.

$$v_{p,s,q,r} = \begin{cases} 1 & \text{if } c_a = q \\ 0 & \text{else} \end{cases} \tag{5.1}$$

In order to account for wrong associations and diversity in personal preference, we aggregate a global sentence-level vote $v_{s,q}$ across all players p. As the game speeds up in higher rounds, players have less time available for making the relevance decision. In a preliminary inspection of annotation results, we noticed significant drops of accuracy across subsequent rounds of the game. In order to account for this effect, we introduce a weighting parameter λ_r representing the confidence that we put into judgements originating from round r of the game being correct. For simplicity's sake, we reduce the confidence by 0.05 per round after the first one. Alternative strategies could for example include learning this parameter as a maximum likelihood estimate across previous observations. Equation 5.2 details the aggregation to a global sentence-level vote $v_{s,q}$ across the set of players $P_{s,q}$ that had encountered the combination of sentence s and query q.

$$v_{s,q} = \frac{1}{|P_{s,q}|} \sum_{p_i \in P_{s,q}} \lambda_r \, v_{p_i,s,q,r} \tag{5.2}$$

Finally, we aggregate across all sentence-level votes $v_{s,q}$ of a document d in order to get one global page-wide judgement that is comparable to well-known (e.g., NIST-created) annotations. Equation 5.3 outlines this process formally. It should be noted, that omission of this third step may, given the application at hand, be beneficial for the evaluation of tasks such as passage-level retrieval or automatic document summarization.

$$v_{d,q} = \frac{1}{|D|} \sum_{s_i \in D} v_{s_i,q} \tag{5.3}$$

Experimentation

At this point, we describe the setup and results of a large-scale experiment conducted on several major commercial crowdsourcing platforms. Our performance comparison of traditional and game-based HITs will be guided by the following 7 fundamental directions:

Quality. How does the result quality of game-based crowdsourcing HITs relate to that of traditional ones given the same underlying crowd of workers and comparable

financial means? We evaluate result quality in terms of agreement with gold standard NIST labels as well as with consensus labels across all participanting groups of the TREC 2011 Crowdsourcing Track [127].

Efficiency. Are game-based HITs more popular, resulting in quicker HIT uptake and completion than conventional ones? We investigate time to completion (for a given batch size) as well as the duration per document and per single vote.

Incentives. How much is fun worth? We investigate the influence of incentives such as fun & social prestige vs. monetary rewards on the uptake rate of HITs. Do users prefer entertaining HIT versions even though they pay less?

Consistency. Does our game encourage a stronger task focus, resulting in better within-annotator consistency? We investigate this dimension using a fixed set of workers (of varying reliability and quality levels) who are exposed to re-occurring HIT questions in a game-based or conventional setting to measure their self-agreement as an estimate of consistency and alertness.

Robustness. Does the share of (alleged) cheaters attracted to our game / attracted to conventional HITs differ? Independent of the overall result quality, the observed cheater rate is a surrogate of how reliable results are and how much sophistication a HIT should dedicate to fraud protection.

Population. Does the use of games lead to a different crowd composition? Offering game-based and conventional HITs, we collect surveys to investigate whether game-based HITs attract different kinds of workers.

Location. State-of-the-art crowdsourcing approaches frequently filter their crowd by nationality in order to improve result quality. We investigate whether there are indeed geographical preferences for game-based or conventional HITs and whether those can be related to the crowd composition in those areas.

In this comparative study, we replicate the setting that was proposed by [127] in the TREC 2011 Crowdsourcing Track assessment task. A total of 3200 topic/document pairs (30 distinct topics, 3195 unique documents) were judged for relevance. The documents are part of the ClueWeb09 collection [43], and the topics originate from the TREC 2009 Million Query Track [46]. A comprehensive list of all topics and document identifiers are available from the TREC 2011 Crowdsourcing Track home page[12]. We contrast the performance and characteristics of our proposed gamified HIT with those of a traditional one. To attribute for assessment mistakes and personal preference, we collected judgements from at least 3 individual workers per topic/document pair in both settings. All HITs were run in temporal isolation (No more than 1 batch at any given time) to limit mutual effects between the tasks. In the following, we describe the respective task designs in detail.

[12] https://sites.google.com/site/treccrowd2011/

Traditional HIT As a performance baseline, we designed a state-of-the-art relevance assessment HIT. Its design follows accepted insights from previous work as detailed in the following. In order to limit the number of context changes, the document is shown in-line on the platform's HIT form as proposed by Kazai [107]. In this way, no distracting opening and closing of windows or browser tabs is required. To further enhance the task, we highlight every occurrence of query terms in the document. This technique was reported to be beneficial by several previous approaches, e.g., by [222]. Finally, in order to deal with malicious submissions, we measure agreement with NIST gold standard pairs of known relevance. Workers who disagree on more than 50% of the gold labels are rejected from the judgement pool. In the HIT instructions, we briefly introduce the available relevance categories. The definition of relevance was introduced according to the TREC guidelines [127]. The HIT form contains 2 questions:

"Please indicate the relevance of the shown document towards the topic <T> ".

"Do you have any remarks, ideas or general feedback regarding this HIT that you would like to share?"

For each HIT, the place holder <T> is replaced by the current topic. Offering the possibility for worker feedback has been frequently reported (see, e.g., [10]) to improve task quality and track down bugs or design flaws quickly. The HIT was offered at a pay rate of 2 US cents per topic/document pair assessments; a reward level previously found adequate by [9].

Gamified HIT The central piece of our proposed gamified version of the relevance assessment HIT is the annotation game that was described in Section 5.2. Instead of having the workers complete tasks locally on the crowdsourcing platform, the technical requirements of our game demanded running it off-site on a dedicated server. In order to verify task completion, workers are provided with a confirmation token after playing one round of the game (10 term associations). Back on the crowdsourcing platform, they enter this token in order to get paid. As a consequence, the actual HIT contained only a brief instruction to the off-site process and two input fields:

"Please enter the confirmation token you obtained after completing one round of the game."

"Do you have any remarks, ideas or general feedback regarding this HIT that you would like to share?"

Again, we solicit worker feedback. The HIT was offered at a pay rate of 2 US cents for one round (10 term associations) of the game. All experiments described in this section were conducted between December 2011 and February 2012 on two crowdsourcing platforms: Amazon Mechanical Turk[13] as well as all available channels on CrowdFlower[14].

[13] http://mturk.com
[14] http://crowdflower.com/

Table 5.5: Annotation quality.

HIT type	Accuracy (NIST)	Accuracy (TREC-CS)
Conventional	0.73	0.74
TREC-CS	0.79	1.0
Game (plain)	0.65	0.75
Game (sent)	0.77	0.87
Game (doc)	**0.82**	**0.93**

Initial evaluation did not show any significant differences in the work delivered by workers from different platforms. We will therefore not split the pool of submissions along this dimension.

In total, 795 unique workers created 105,221 relevance judgements via our game. Additionally, 3000 traditional relevance judgements were collected for comparison. In total, we invested $90 to collect a volume of 108,221 annotations across the two compared experimental conditions. Together with the TREC 2011 Crowdsourcing Track annotations and the original NIST labels, this makes the T11Crowd subset of ClueWeb09 one of the most densely-annotated Web resources known to us. To enable reproducibility of our insights and to further general crowdsourcing research, the complete set of our judgements and the game itself are available to the research community[15].

As a starting point to our performance evaluation of game-based crowdsourcing of relevance assessments, we investigate the quality of the collected labels. Table 5.5 details the performance of our game in terms of overlap with gold standard NIST labels as well as the global consensus across all TREC 2011 Crowdsourcing Track participants (TREC-CS). We can note that already the conventional HIT delivers high result quality. Ratios between 65% and 75% are often considered good rules-of-thumb for the expected agreement of faithful human judges given a relevance assessment task [233]. TREC consensus labels show a high overlap with NIST annotator decisions. The third row in Table 5.5 shows the performance of direct unaggregated sentence-level votes from our game as described in Equation 5.1. While agreement with the TREC crowd is already substantial, the overlap with high-quality NIST labels lags behind. As we aggregate across multiple workers' annotations of the same sentence (Equation 5.2) and, finally, across all sentences extracted from the same document (Equation 5.3), the performance rises significantly, outperforming all compared methods. We used a Wilcoxon signed rank test at $\alpha < 0.05$-level to test significance of results.

In order to confirm the usefulness of our assumption from Section 5.2 concerning the decline of label quality as the game speeds up and the player has less time to make decisions, we evaluated annotation performance of raw labels according to the round in which they were issued. Table 5.6 shows a near-linear decline in agreement of plain game scores with TREC consensus as the game progresses. Agreement with NIST scores also consistently shrinks from round to round.

Finally, previous work on prediction in crowdsourcing systems (e.g., [212] and [169]) demonstrates that reliability of the average predicted scores by the crowd improves as

[15] http://sourceforge.net/projects/geann/

Table 5.6: Annotation quality as a function of the game round in which judgements were issued.

Round	Accuracy (NIST)	Accuracy (TREC-CS)
1	0.72	0.81
2	0.67	0.77
3	0.62	0.73
4	0.60	0.69
5	0.54	0.65

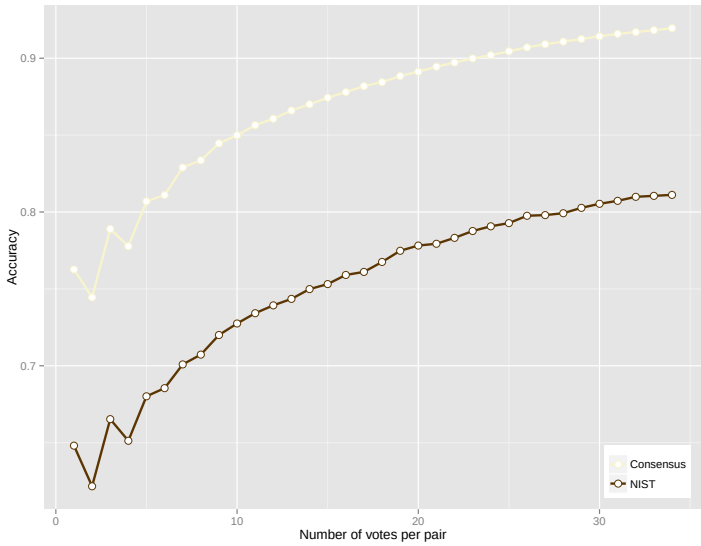

Figure 5.6: Quality as a function of votes per pair.

the size of the crowd increases. The benefit of our game-based HIT is its popularity that allows us to collect more judgements per topic / document pair than traditional HITs. On average, each topic / document pair in the collection received 32 unique user judgements at the sentence level (some of which may originate from the same user as she or he rates different passages of the same document). Figure 5.6 shows how annotation quality develops as we add more judgements per pair. After initial fluctuation, as single votes have great influence on the overall decision, accuracy consistently improves as we add more votes per pair. The effect levels out as we approach the upper performance limit.

The second official performance indicator besides label quality in the TREC 2011 Crowdsourcing Track was the time necessary to collect the required judgements. For many use cases in human computation, low latencies are essential. The particular nature of the game imposed a time limit on players within which they had to make their decisions. As we detailed in the previous section, aggregation across users, weighting votes according to the difficulty level under which they were created, ensured competi-

Table 5.7: Annotation efficiency.

	Conventional	Game-based
t per vote	40.1 sec	5.2 sec
t per doc	40.1 sec	27.8 sec
Uptake (votes per hour)	95.2	352.1

tive result quality. At the same time, however, a sequence of, individually quick, concept matchings enables workers to be more efficient and motivated than in conventional settings. Table 5.7 shows how conventional HITs take slightly longer to judge documents even when aggregating the duration of all passage-level votes in the game-based setting. Taking into account the significantly higher uptake rate (number of judgements issued per hour) of the game HITs, this serves for a considerably more efficient batch processing.

Especially in conjunction with the previous section's findings of high degrees of redundancy serving for better result quality, high uptake rates become crucial as they allow for timely, yet accurate decisions.

The third and final evaluation criterion employed for TREC 2011 was the cost involved in the collection of relevance labels. With our game-based approach, we aim to, at least partially, replace the financial reward of the HIT with entertainment as an alternative motivation. In this section, we will investigate to which degree this change in incentives can be observed in worker behaviour.

In order to be paid via the crowdsourcing platform, workers had to complete at least one round (10 concept matchings) of our game. At that point the required confirmation token was displayed to them and they could return to the platform in order to claim their payment. However, the game offered an additional 4 levels to be played. From a purely monetary-driven perspective there would be no reason for continuing to play at that point. As we can see in Table 5.8, however, over 70% of games are played beyond the first round. This essentially results in crowdsourcing workers creating judgements free of charge because they enjoy the game experience. Additionally, we can observe players to return to the game after a number of hours to play again and improve their score and their resulting position on the leader board. Subsequent visits often happen directly to the game page, without being redirected from (and paid through) the crowdsourcing platform. Almost 80% of all players (633 out of 795) return after their first round played, with an average time gap of 7.4 hours between games. For regular HITs, we observed a return rate of only 23%.

When inspecting the concrete distribution of judgements across workers, as shown in Figure 5.7, we see this trend continued. Crowdsourcing tasks often tend to exhibit Power-law distributions of work over unique workers with some strong performers and a long tail of casual workers who only submit single HITs. Here, however, we notice a strong center group of medium-frequency players. We hypothesise that replacing the workers' extrinsic motivation ("*do the HIT to earn money*") by an intrinsic one ("*let's have some fun*"), causes these tendencies.

This has a number of noteworthy consequences: (1) We can attract workers to a HIT at a comparatively low pay rate. Even without playing beyond the first round, 2 US cents

Table 5.8: Game-based assessment behaviour.

Criterion	Observation
Games with 2+ rounds	70.9%
Rounds per game	3.5
Players with 2+ games	79.5%
Games per player	4.36
Time between games	7.4 hrs

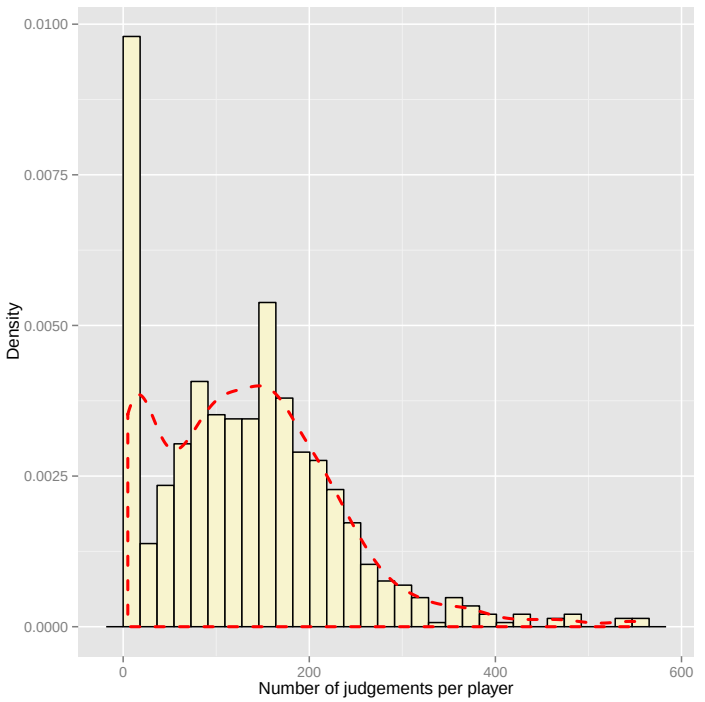

Figure 5.7: Distribution of judgements across users.

Table 5.9: Effective annotation cost.

	Conventional	Game-based
Cost per doc	$0.06	$0.0004
Whole corpus	$192	$1.28
Effective hourly rate	$1.80	$0.18

for 10 concept associations would roughly result in a prospective hourly pay of $1.20. (2) Furthermore, as most workers continue playing, additional annotations are created with no expectation of financial compensation. (3) Drawn by the competitive aspect of the game, workers return and create even more unpaid assessments. As a consequence, the overall amount of money invested into the game-based collection of more than 100,000 sentence-level relevance judgements was $27.74. This includes all administrative fees charged by the crowdsourcing platforms. In comparison, the participants to the TREC 2011 Crowdsourcing Track reported overall costs of $50 – $100 for the collection of significantly fewer labels.

Table 5.9 shows a final cost comparison of the conventional and game-based versions of the inspected relevance assessment HIT. While all indicators of cost efficiency from a worker's perspective clearly speak for choosing the conventional, better-paying HIT, the previously described figures of HIT uptake rates as well as the high number of alternative HITs available at all times on large-scale platforms such as AMT, indicate, that we reach workers who consider the entertainment potential of a HIT before choosing it. If we consider all judgements made in rounds after the first one and all judgements from revisits that were not paid for on the crowdsourcing platform as free-of-charge judgements, we arrive at a share of 83.7% of all labels having been created free of charge. Additionally, a number of players (39 out of 795) accessed the game without being prompted (and paid) by a crowdsourcing platform. These players were recruited from the authors' professional and private networks or word of mouth of other players. We could not find significant differences in the judgement quality or volume created by this group of players. The invested amount of money can be seen as advertisement costs rather than actual payments. In a traditional setting, collecting the same annotation density would have cost $2104.

Following [57]'s theory of Flow, a state of deep immersion is a good foundation for high performance independent of the concrete task at hand. With our game-based HIT, we aimed to exploit this observation in order to create greater task focus than workers typically achieve on conventional HIT types. The previously shown result quality figures support this hypothesis. As an additional performance indicator, we will measure the workers judgement consistency. Faced with the same passage of text and choice of concepts multiple times, a situation-aware worker is expected to display a high degree of intra-annotator agreement. In the course of our judgement collection, we showed identical assignments to workers 837 times and observed an intra-annotator agreement of 69.8%. We set up a dedicated crowdsourcing experiment in which a portion of the offered topic / document pairs re-occurred. The HIT was set up in the "traditional" fashion described earlier. Across 500 redundantly issued assignments, we observed an intra-annotator agreement of only 61.3%, a significantly lower ratio (determined using

Wilcoxon signed rank test at $\alpha < 0.05$) than in the game-based setting. While the game setting resulted in higher consistency than usual crowdsourcing schemes, we could not match the consistency Scholer et al. [193] report for professional assessors as for example employed by NIST.

Cheating, spamming and low-quality submissions are well-known and frequently-observed incidents on commercial crowdsourcing platforms. Previously, we demonstrated convincing result quality of gamified document relevance assessments when labels are aggregated across a sufficiently large number of workers. Since our approach appeals more to the entertainment-seeking rather than money-driven workers, we did not include a dedicated cheat detection scheme as would often be considered necessary in state-of-the-art HITs. However, we realise that the observed cheat rate in an assignment can serve as a surrogate for the confidence and reliability of the overall results. To this end, we measure the observed proportion of cheat submissions to our game as well as to the conventional HIT version. Following Definition 1, we categorizing workers who disagree with the majority decision in more than half of all cases as cheaters. In order to deliver a conservative estimate of the rate of cheat submissions, we tighten their definition and consider a worker as cheating if at least 67% of their submissions disagree with the majority vote. This scheme was applied to both, the conventional HIT as well as the gamified version. In the game-based case, we additionally flagged all submissions as cheat that tried using forged confirmation tokens.

Overall, this resulted in a share of 13.5% of the conventional HIT's judgements being considered cheated. For the game-based version, the percentage was a significantly lower 2.3%. This finding conforms with our earlier observation from Section 5.1, where we observed innovative, creative tasks being less likely to be cheated on.

In our game-based approach, we did not make use of any form of *a priori* filtering the pool of workers eligible to access our HITs. We hypothesise, however, that HIT type, financial reward and task phrasing influence the underlying crowd that decides to work on a given assignment. To better understand the composition of the group of commercial crowdsourcing workers that are interested in games, we accompanied parts of our HITs by surveys in which we asked for high-level participant demographics and their preference for either the conventional or the game-based HIT. Table 5.10 shows an overview of several salient outcomes of the survey. The split in decisions was roughly equal, with 24% of workers not indicating clear preferences. The entertainment-seeking worker is on average several years younger, more likely to hold a university degree and will typically earn a higher salary. Finally, women were found to be significantly less interested in games than their male co-workers. This conforms with general observations about gender differences made for example by Ko et al. [114]. A worker's language background did not influence his or her likelihood to prefer games.

Many commercial crowdsourcing schemes report performance gains when filtering the crowd by nationality. We noted similar tendencies in the previous section, ourselves. Due to different distributions of education, language skills or other cultural properties, such steps can influence result quality. As a final dimension of our investigation of games for use on commercial crowdsourcing platforms, we will inspect whether worker origin has an influence on result quality. From our survey, we found Indian workers, with a share of 60%, to be the dominant group in both settings. US workers were consistently

Table 5.10: Composition of the crowd

	Conventional	Game-based
Preference	39%	37%
Female	47%	35%
Age	34	27
Univ. degree	46%	62%
Income	$20k	$45k
English Native Speaker	24%	25%

the runners-up with a proportion of approximately 25%. There was no significant difference in the likelihood to prefer games over conventional HITs between countries.

Finally, when inspecting result quality from our game, again, no difference in performance or likelihood to cheat could be found. This suggests that filtering workers by nationality may not be ideal. In fact, the underlying worker motivation and HIT type preference can be assumed to have a far greater impact on observed uptake, performance and trustworthiness.

Image classification

Previously, we described and evaluated the performance of the proposed crowdsourcing-powered annotation game for the task of TREC-style document relevance assessments. To demonstrate the generalization potential of the described concept-matching method, we applied the same game in an image classification pilot.

In the course of the Fish4Knowledge project[16], several underwater cameras have been placed in selected locations in south-east Asian coral reefs. The continuous recordings are supposed to further knowledge about behaviour, frequency and migration patterns of the resident tropical fish species. A key step to coping with the large amounts of image data produced by these cameras is a reliable automatic species classification. In order to train such systems, numerous training examples are required. While the project employs a team of marine biologists, their greater expertise is costly. Using our annotation game, we crowdsource the task of classifying the encountered species. Instead of relating keywords to TREC topics, the objective is now to match a shot of the underwater camera (often low quality) to high-quality examples of resident fish species. By initializing the underlying database with images rather than textual items, no changes to the actual game were necessary. Figure 5.8 shows a screenshot of this alternative game setting.

Our feasibility study encompassed 190 unique underwater camera shots for which known gold standard labels created by the marine biologists existed. Each biologist had classified all images, allowing us to contrast crowd agreement with expert agreement. The HIT was offered in January and February 2012 at the same pay rate (2 US cents per round of 10 associations) as the text-based version. Table 5.11 shows the results of the experiment in which the degree of agreement with the majority of experts as well as the crowd's inter-annotator agreement are detailed. We can see high agreement with expert

[16] http://www.fish4knowledge.eu/

Figure 5.8: GeAnn applied for image grouping.

Table 5.11: Image classification performance

	Agreement (exp.)	Inter-annot. Agreement
Experts	0.82	-
Game	0.75	0.68

labels as well as substantial agreement among workers. The popularity (qualitatively perceived through worker feedback) and uptake rate of this HIT even slightly exceeded those of the game-based one for document relevance assessments. Several workers had mentioned difficulties reading the moving text fragments in the short time available. With images, this does not appear to be an issue.

Methods like this could play an essential role either in the creation of training and evaluation data necessary for the assessment of automatic classification quality, or as part of a hybrid human-computer classification in which automatic methods narrow down the range of potential species before human annotators select the most likely species from the pre-selection. It should be noted, however, that the domain experts are by no means obsolete in this setting. While they annotated fish images simply based on their knowledge of the resident species, players of our game only had to select one out of a range of 4 species by similarity.

Discussion

We found convincing results across all inspected performance dimensions, supporting the benefit of offering alternative incentives besides the pure financial reward. To conclude, we will now discuss a number of observations and insights that were not yet fully covered by the evaluation.

Firstly, considering the fact that the round concept of the game appears to invite workers to create assessments without payment (by playing on after having received the confirmation token), it is not obvious why we should limit the game to a fixed number of rounds. In the present setting, a game inevitably ends after the fifth round. One might argue that a higher number of rounds or even an open-ended concept would result in

even greater cost efficiency. In fact, the opposite seems to be the case. In an initial version of the game, there was no upper limit to the number of rounds per game. As a consequence, some players were frustrated, as the only way to "finish" the game would be to either lose or give up. This resulted in fewer returning players. Additionally, the quality of annotations resulting from higher rounds was highly arguable as the objective of the game became mainly surviving through as many rounds of fast-dropping items as possible, rather than making sensible assessments. In the new, limited, setting, the clear objective is to do as well as possible in 5 rounds. Players who want to improve their score beyond this point have to return and start a new game.

A second key observation to be made is the fact that while we evaluate against the performance of NIST assessors and TREC participants, the tasks our workers face is a significantly different one. In the game, no worker gets to see full textual documents or is even told that the objective is to determine topical relevance of Web resources towards given information needs. We deliberately aimed for such a loose coupling between game and task as we wanted to keep the game experience entertaining without leaving the "aftertaste" of working. It is interesting that mere conceptual matching correlates well with actual relevance assessments. Also, in the data pre-processing phase, we do not extract sentences based on query terms but rather focus on pure *idf* statistics. In this way, we manage to capture the general gist of a document without artificially biasing it towards the topic.

Finally, the key insight gained from this work was the substantial benefit achieved by offering an alternative incentive to workers. Most of the interesting properties observed in the gamified system, such as workers producing free labels, would not have happened otherwise. This is, however, not necessarily limited to gamified tasks. In this paper we used games as one possible means of showing how a particular incentive (money) can be replaced with another one (entertainment). By doing so, we focus on a certain type of worker, entertainment-seekers, the existence of which we hypothesised based on previous experience with crowdsourcing. We are convinced that a better understanding of worker types and their specific intrinsic motivations is essential in driving the boundaries of current crowdsourcing quality. Kazai et al. [108] propose an interesting classification of workers into several categories. In their work, a number of performance-based worker types, including e.g., spammers, sloppy and competent workers are described. We believe, that more general worker models which also encompass aspects such as worker motivation capability and interest in certain HIT types, etc. can be of significant benefit for the field. Very similar to the task of advertisement placement, a worker whose motivations we understand, can be targeted with better-suited precisely-tailored HIT types.

The common example of worker filtering by nationality illustrates the practical need for a better understanding of worker motivation. This practice is not only of dubious ethical value, it may additionally address symptoms rather than causes. The original objective of identifying and rejecting such workers that try to game the task and get paid without actually working is often hard to fulfil. Filtering by nationality is straightforward to achieve, but also (at best) only correlated with reliability. This bears the significant risk of artificially thinning the pool of available workers. Our work (e.g., Tables 5.8 and 5.10), demonstrates that in an entirely unfiltered environment no significant national differ-

ences in quality, cheat rates, etc. can be found when focussing on the desired worker type. In this way, we retain a large work force but, by task design, discourage undesired worker types from taking up our work in the first place.

Looking out towards future changes to the game based on lessons learned in this work, we aim for including yet another incentive besides entertainment. The leader board concept of the current game tries to spark competition between players and has a moderate success at doing so. However, the workers do not know each other. In a reputation-aware environment, such as a social network, this effect can be expected to have a far greater impact. Having the ability to compare scores and to compete in a direct multi-player game with their friends will create much more compelling incentives for (a) performing well in each game, (b) continuing to play, (c) returning for subsequent matches and (d) recommending the game to their circle of friends. We believe that exploiting these aspects by integrating social reputation into crowdsourcing will create many interesting applications.

5.3. Conclusion

In this chapter, we discussed two alternative ways of integrating humans into the computation loop for estimating document relevance. Human expertise is considered an indispensable component in the design and evaluation cycles of many industrial and academic applications. In Section 5.1, we introduced commercial crowdsourcing as a means of reaching out to massive work forces on platforms such as Amazon's Mechanical Turk. In order to fully use the wisdom of the crowd, it is crucial to address noise in the form of inaccurate or deceiving submissions. While the state of the art focussed on posterior recognition and filtering of undesired submissions, we showed that creative tasks attract less cheat submissions than straight-forward, repetitive ones. Even for standard tasks, interface design considerations can significantly reduce the amount of cheat submissions.

Addressing Research Question 2.d), we took our theory one step further and devised a game with a purpose that allowed workers to create topical relevance judgements in a playful and entertaining way. In this way, we were able to negate the supposed effect of worker nationality by phrasing HITs in such a way that they attract more reliable workers.

References

[4] Luis von Ahn and Laura Dabbish. "Designing games with a purpose". In: *Communications of the ACM* 51.8 (2008), pp. 58–67.

[5] Luis von Ahn and Laura Dabbish. "Labeling images with a computer game". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2004, pp. 319–326.

[6] Luis von Ahn, Mihir Kedia, and Manuel Blum. "Verbosity: a game for collecting common-sense facts". In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 75–78.

[7] Luis von Ahn, Ruoran Liu, and Manuel Blum. “Peekaboom: a game for locating objects in images”. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 55–64.

[8] Luis von Ahn et al. “Improving image search with phetch”. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE. 2007, pp. IV–1209.

[9] Omar Alonso and Ricardo Baeza-Yates. “Design and implementation of relevance assessments using crowdsourcing”. In: *Advances in information retrieval*. Springer, 2011, pp. 153–164.

[10] Omar Alonso and Matthew Lease. “Crowdsourcing 101: putting the WSDM of crowds to work for you”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 1–2.

[11] Omar Alonso and Stefano Mizzaro. “Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment”. In: *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 2009, pp. 15–16.

[13] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. “Active learning and crowdsourcing for machine translation”. In: *Language Resources and Evaluation (LREC)* 7 (2010), pp. 2169–2174.

[15] Einat Amitay et al. “Scaling IR-system evaluation using term relevance sets”. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2004, pp. 10–17.

[22] Andy Baio. *The Faces of Mechanical Turk*. http://waxy.org/2008/11/the_faces_of_mechanical_turk/. 2008.

[43] Jamie Callan et al. “Clueweb09 data set”. In: *Retrieved* 12.23 (2009), p. 2010.

[46] Ben Carterette et al. “Million query track 2009 overview”. In: *Proceedings of TREC*. Vol. 9. 2009.

[51] Charles L. Clarke. *Overview of the TREC 2009 Web track*. Tech. rep. Waterloo University, 2009.

[57] Mihaly Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.

[60] A. Philip Dawid and Allan M. Skene. “Maximum likelihood estimation of observer error-rates using the EM algorithm”. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 20–28. ISSN: 0035-9254.

[84] Catherine Grady and Matthew Lease. “Crowdsourcing document relevance assessment with Mechanical Turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*. Association for Computational Linguistics. 2010, pp. 172–179.

[88] Donna Harman. “Overview of the first TREC conference”. In: *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1993, pp. 36–47.

[89] Christopher G. Harris. "You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks". In: *Crowdsourcing for Search and Data Mining (CSDM 2011)* (2011), p. 15.

[94] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. *Cheat-Detection Mechanisms for Crowdsourcing*. Tech. rep. University of Würzburg, 2010.

[97] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. "Data quality from crowdsourcing: a study of annotation selection criteria". In: *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*. Association for Computational Linguistics. 2009, pp. 27–35.

[102] Panagiotis G. Ipeirotis. "Analyzing the amazon mechanical turk marketplace". In: *XRDS: Crossroads, The ACM Magazine for Students* 17.2 (2010), pp. 16–21.

[103] Panagiotis G. Ipeirotis. *Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes*. http://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html. 2010.

[106] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk". In: *Proceedings of the Seventeenth Americas Conference on Information Systems*. 2011, pp. 1–11.

[107] Gabriella Kazai. "In search of quality in crowdsourcing for search engine evaluation". In: *Advances in information retrieval*. Springer, 2011, pp. 165–176.

[108] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. "Worker Types and Personality Traits in Crowdsourcing Relevance Labels". In: *Proceedings of 20th International Conference on Information and Knowledge Management (CIKM)*. ACM. 2011.

[109] Shashank Khanna et al. "Evaluating and improving the usability of Mechanical Turk for low-income workers in India". In: *Proceedings of the First ACM Symposium on Computing for Development*. ACM. 2010, p. 12.

[111] Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2008, pp. 453–456.

[114] Chih-Hung Ko et al. "Gender differences and related factors affecting online gaming addiction among Taiwanese adolescents". In: *The Journal of nervous and mental disease* 193.4 (2005), pp. 273–277.

[120] Joseph Lampel and Ajay Bhalla. "The role of status seeking in online communities: Giving the gift of experience". In: *Journal of Computer-Mediated Communication* 12.2 (2007), pp. 434–455.

[127] Matthew Lease and Gabriella Kazai. "Overview of the TREC 2011 crowdsourcing track (conference notebook)". In: *Text Retrieval Conference Notebook*. 2011.

[131] Gregory W. Lesher and Christian Sanelli. "A web-based system for autonomous text corpus generation". In: *Proceedings of ISSAAC* (2000).

[134] Greg Little et al. “Turkit: Tools for iterative tasks on mechanical turk”. In: *Proceedings of the ACM SIGKDD workshop on human computation*. ACM. 2009, pp. 29–30.

[142] Hao Ma et al. “Improving search engines using human computation games”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 275–284.

[148] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a large annotated corpus of English: The Penn Treebank”. In: *Computational linguistics* 19.2 (1993), pp. 313–330.

[151] Jane McGonigal. *Reality is broken: Why games make us better and how they can change the world*. Penguin books, 2011.

[157] Andrew P. Moore, Robert J. Ellison, and Richard C. Linger. *Attack modeling for information security and survivability*. Tech. rep. Carnegie-Mellon University, Software Engineering Institute, 2001.

[169] David Pennock. *The Wisdom of the Probability Sports Crowd*. http://blog.oddhead.com/2007/01/04/the-wisdom-of-the-probabilitysports-crowd/. 2007.

[170] Charles P. Pfleeger and Shari L. Pfleeger. *Security in computing*. Prentice Hall PTR, 2006.

[179] Clare Richards. *Teach the world to twitch: An interview with Marc Prensky, CEO and founder Games2train. com. Futurelab*. 2003.

[180] Ellen Riloff. “Automatically generating extraction patterns from untagged text”. In: *Proceedings of the national conference on artificial intelligence*. 1996, pp. 1044–1049.

[183] Joel Ross et al. “Who are the crowdworkers?: shifting demographics in mechanical turk”. In: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*. ACM. 2010, pp. 2863–2872.

[193] Falk Scholer, Andrew Turpin, and Mark Sanderson. “Quantifying test collection quality based on the consistency of relevance judgements”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 1063–1072.

[196] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1997. ISBN: 0201694972.

[203] Rion Snow et al. “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2008, pp. 254–263.

[204] Ian Soboroff, Charles Nicholas, and Patrick Cahan. “Ranking retrieval systems without relevance judgments”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 66–73.

[205] Mohammed Soleymani and Martha Larson. "Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus". In: *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*. 2010, pp. 4–8.

[207] Alexander Sorokin and David Forsyth. "Utility data annotation with amazon mechanical turk". In: *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*. IEEE. 2008, pp. 1–8.

[212] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[222] Julián Urbano et al. "The University Carlos III of Madrid at TREC 2011 Crowdsourcing Track". In: *Text REtrieval Conference*. 2011.

[225] Ellen M. Voorhees. "The philosophy of information retrieval evaluation". In: *Evaluation of cross-language information retrieval systems*. Springer. 2002, pp. 355–370.

[226] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge, 2005.

[228] Jing Wang, Siamak Faridani, and Panagiotis G. Ipeirotis. "Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models". In: *Crowdsourcing for Search and Data Mining (CSDM 2011)* (2011), p. 31.

[233] Peter Welinder et al. "The multidimensional wisdom of crowds". In: *Advances in Neural Information Processing Systems* 23 (2010), pp. 2424–2432.

III

Multivariate Relevance

6

Probabilistic Multivariate Relevance Modelling

Combine the extremes, and you will have the true center.

-Karl Wilhelm Friedrich Schlegel, Poet

In the previous part of this work, we discussed two alternative paradigms of estimating relevance for information retrieval. First, in Chapter 4, we investigated automatic means of estimating relevance based on document-level properties, either in terms of actual document content or based on the observed interactions between content and users. In Chapter 5, we turned towards human computation, discussing commercial crowdsourcing as well as means of gamification in order to elicit valuable information in an entertaining and enticing way. Many real-life scenarios will make use of some form of combination of the two fundamental approaches, a typically observed strategy is to harness human computation for the creation of high-quality training and evaluation sets on the basis of which automated approaches can be built to make real-time decisions at massive scale.

Whether in this way, or by means of another concrete implementation, at this point, we assume that there are reliable ways of inferring individual relevance dimensions from the available sources of data. The final challenge, however, becomes to combine all these cues of relevance into a formal model, resulting in an estimate of the overall probability of relevance.

Beyond the value of individual relevance factors, there can be complex, non-linear *dependencies* between relevance factors. For example, relevance dimensions such as topicality and credibility might appear independent for some document subsets, but extreme values in one dimension may influence the other in a way that is not easily captured by state-of-the-art approaches. As a concrete example, take TREC 2010's faceted blog distillation task (see [144]), that aims at retrieving topically relevant non-factual blog feeds. Here, the relevance space has two dimensions: topicality and subjectivity. Figure 6.1 shows the distribution of relevance scores for Topic 1171, "*mysql*", across these two relevance dimensions. We can note an apparent correlation in the lower left part of the graph that weakens as scores increase. To underline this, we computed Pearson's ρ between the two dimensions for the lower score third ($\rho = 0.37$), the upper region ($\rho = -0.4$), as well as the overall distribution ($\rho = 0.18$). Apparently, the dependency structure of the joint distribution of relevance, in this case, is not easily described by a linear model. Consequently, we can expect dissatisfying performance of linear combination models. And, indeed, when inspecting the performance of a linear combination model with empirically learned mixture parameters λ, Topic 1171 receives an average precision of only 0.14, well below the method's average across all topics of 0.25. In the course of this chapter, we will discuss practical means of addressing cases like the present one and will finally revisit this example to demonstrate the effect of our proposed method.

While the machine learning, information retrieval, data mining and natural language processing communities have significant expertise in estimating topical document relevance and additional dimensions in isolation, the commonly applied combination schemes have tended to be *ad hoc* and ignore the problem of modelling complex, multi-dimension dependencies. In practice, they follow statically weighted linear combinations with empirically determined mixture parameters (e.g., [182]) or deploy sophisticated learning to rank techniques that tend to offer only limited insight to humans about why they were weighted highly for relevance. Ideally, we would demand realistic, yet formally-grounded combination schemes that can lead to results that are both effective and with human-interpretable justification.

In a different context, the field of quantitative risk management has devised *copulas*, a flexible, varied class of probability density functions that are designed to capture rich, non-linear dependencies efficiently in multi-dimensional distributions. Copulas work by decoupling the marginal distributions of the data from the underlying dependency structure of the joint distribution. In particular, copulas can account for so-called tail dependencies, i.e., dependencies that play up at the extreme values of the interacting distributions. As an example, let us consider two commodities traded on the stock market, such as rare earth metals and pork bellies. The two commodities are sufficiently different to make the related market segments quasi-independent. However, [39] note that extreme market situations can cause investor panics that reach across otherwise independent segments and cause previously unseen interrelationships.

In the following, we give a detailed introduction to the formal framework of copulas and describe how to estimate them from empirical data. Based on a number of sizeable standard data sets such as the Blogs08 collection described by Macdonald et al. [144], we demonstrate the merit of using copulas for multivariate relevance estimation. In a

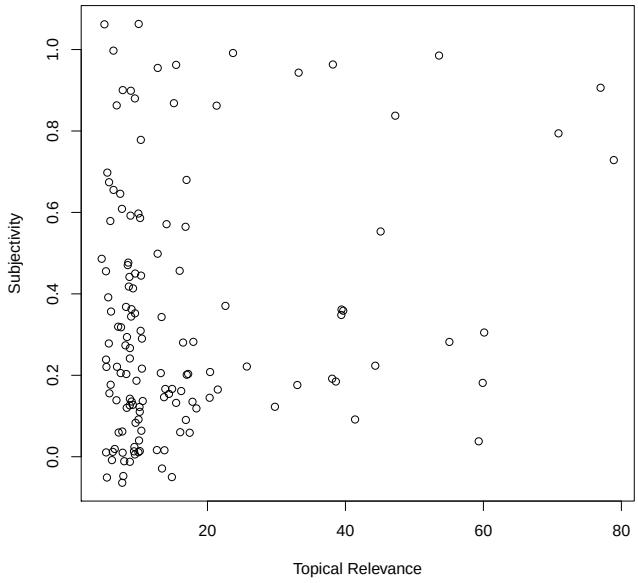

Figure 6.1: Distribution of bivariate relevance scores for TREC 2010 Blog Track Topic 1171, “mysql”.

related effort, we address the task of score fusion based on historic submissions to the TREC *ad hoc* task.

While the formal combination of several individual relevance dimensions in one model has not been extensively studied, there has been an interesting thread of research on score fusion. The task is to combine the result rankings of multiple independent retrieval systems in order to compensate for local inaccuracies of single engines. Early approaches, such as [73], to the task were based on evidence aggregation in the form of products and sums of scores across individual systems. The fused ranking is based on the absolute value of the cross-system aggregates. Vogt and Cottrell [224] first introduced linear interpolation of multiple model rankings for system fusion. Aslam and Montague [19] proposed a probabilistic rank-based method for direct combination of multiple engines. Later on, they devised a similar method based on a majority voting scheme between various retrieval systems presented in [155]. They further proposed a score normalization scheme that is more robust to outliers in the distribution of relevance than the previously used min/max technique [156]. Manmatha et al. [145] estimated a search engine’s score distribution as a mixture of normal and exponential distributions, for relevant and non-relevant documents respectively. They used the resulting distributions for score fusion across multiple engines, but did not attempt to model dependencies in the joint score distribution, instead treating the scores as independent and averaging probabilities, or discarding ‘bad’ engines altogether.

Wu and Crestani [239] introduced the first of what would become a group of fusion

approaches that define an explicit weighting scheme under which the original result lists are combined. Bordogna and Pasi [35] as well as Costa-Pereira et al. [54] employ various quality notions such as the degree to which a document satisfies a given relevance criterion to dynamically adapt the weighting scheme to the underlying distribution of relevance. Craswell et al. [55] investigated relevance model combination by linearly combining constituent scores in the log domain. Tsikrika and Lalmas [220] applied Dempster-Shafer theory for the aggregation of independent relevance dimensions in Web retrieval in the form of belief functions. Gerani et al. [80] propose non-linear score transformations prior to the standard weighted linear combination step. Their solid results demonstrate the need for models whose capabilities go beyond linear dependency structures between relevance dimensions.

In recent years, the variety of IR applications has become significantly more diverse. As a consequence, universal relevance models have become less viable in many areas. Tasks such as legal IR, expert finding, opinion detection or the retrieval of very short documents (e.g., tweets) have brought forward strongly customised relevance models tailored towards satisfying a given task (e.g., [99] or [23]). Especially for the retrieval of structured (XML) documents, score combination schemes are of central importance to combine evidence across multiple structural fields within a document. Despite the numerous potential issues pointed out by Robertson et al. [182], most state-of-the-art approaches to XML retrieval (e.g., [141]) rely on linear models. An advance towards the formal combination of several independent relevance dimensions in the form of prior probabilities for language models has been made by Kraaij et al. [117] for the task of entry page search. To date, however, most universally applicable relevance models still rely on pure linear combinations of relevance dimensions that disregard the underlying data distribution or potential dependencies between the considered dimensions.

Learning to rank (L2R) has been established as an alternative approach for signal combination. The aim is to apply machine learning methods to either directly infer a document ranking or a ranking function from a wide range of features, potentially including the previously-discussed dimensions. Examples of such methods include [41], [176], and, [136]. The downside of this approach is that the resulting models tend to yield only limited insight for humans. The classic approach of developing a unifying formal retrieval model would in our view provide better means to increase not just overall performance, but also our qualitative understanding of the problem domain.

By introducing copulas for information retrieval, this work proposes a way for closing the gap between linear combinations (that break with the probabilistic framework in which the constituent scores were estimated) and non-linear machine-learned models (that offer only limited insight to scientists and users).

Copulas have been traditionally applied for risk analyses in portfolio management [69] as well as derivatives pricing in quantitative finance [39]. Recently, however, there are several successful examples from unrelated disciplines. Renard and Lang [177] estimate water flow behaviour based on Gaussian copulas. Onken et al. [166] apply copulas for spike count analysis in neuroscience. In meteorology, Schoelzel and Friedrichs [192] use copulas to combine very high-dimensional observations for the task of climate process modelling. To the best of our knowledge, there has been no prior application of the copula framework to information retrieval problems.

6.1. Copulas

At this point, we will give a brief introduction of the general theoretical framework of copulas, before applying them to various IR tasks in subsequent sections. For a more comprehensive overview, please refer to [191] for more detail and pointers to further reading.

The term copula was first introduced by Sklar [201] to describe multivariate *cumulative distribution functions* (*cdfs*) that allow for a formal decoupling of observations from dependency structures. Formally, given

$$X = (x_1, x_2, \ldots, x_k)$$

a k-dimensional random vector with continuous margins

$$F_k(x) = \mathbb{P}[X_k \leq x]$$

we apply the probability integral transformation to obtain uniform marginals U

$$U = (u_1, u_2, \ldots u_k) = (F_1(x_1), F_2(x_2), \ldots F_k(x_k)).$$

This is where our copulas come into play. A k-dimensional copula C describes the joint cumulative distribution function of random vector U with uniform margins.

$$C : [0,1]^k \rightarrow [0,1]$$

This approach has two obvious practical benefits: (1) Separating marginals and dependency structure allows for more straightforward estimation or approximation of each component in isolation. (2) An explicit model of dependency is scale-invariant. The copula describes a reference case of dependency on the unit cube $[0,1]^k$ that can be applied to arbitrary random vectors without further adjustment.

A number of key properties make copulas an appealing theoretical framework for a wide number of applications, so we summarize those now.

- Like all cdfs, a copula $C(u_1, u_2, \ldots, u_k)$ is increasing in each component u_i
- A marginal component u_i can be isolated by setting all other components to 1:

$$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$$

- If a single component u_i in U is zero, the entire copula is zero:

$$C(u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_k) = 0$$

- Most importantly, we can assume general applicability of the copula framework, since, as a consequence of Sklar's Theorem, for each k-dimensional cdf F and all x_i in $[-\infty, \infty]$ and $1 \leq i \leq k$, there exists a copula C with

$$F(x_1, \ldots, x_k) = C(F_1(x_1), \ldots, F_k(x_k))$$

Extreme conditions

Before applying the copula framework to problems in information retrieval, let us visit a number of extreme conditions of dependency that frequently occur in IR scenarios. **(1) Independence** of observations is a frequently assumed simplification in IR theory that leads to convenient (if naïve) probabilistic models. In the copula framework, independence of events can be captured by the so-called *independence copula* C_{indep}:

$$C_{indep}(U) = \exp(-\sum_{i=1}^{k} -\log u_i)$$

which is equivalent to the product across all constituent probabilities in U. **(2) Co-monotonicity** describes the case of perfect positive correlation between observations u:

$$C_{coMono}(U) = \min\{u_1, \ldots, u_k\}$$

(3) counter-monotonicity of observations is given in the opposite case of perfect negative correlation:

$$C_{counterMono}(U) = \max\{\sum_{i=1}^{k} u_i + 1 - k, 0\}$$

Consequently, each copula lies within the so-called Fréchet-Höffding bounds [95]:

$$C_{counterMono}(U) \leq C(U) \leq C_{coMono}(U)$$

Copula families

After having covered the foundations of copula theory let us inspect some concrete examples of copulas that will be used in the course of this work. Three general families of standard copulas have been proposed in the literature, whose corresponding equations are given right after their introduction in this paragraph: **(1) Elliptical copulas** are directly derived from known distributions and are based on standard distribution functions such as the Gaussian distribution or Student's t distribution. Equation 6.1 shows the Gaussian copula that requires the observed covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$ as a parameter. Φ denotes the cdf of a standard normal distribution and Φ^{-1} its inverse. **(2) Archimedean copulas** are popular as they can be explicitly stated (note that due to their distribution dependency that is not the case for elliptical copulas) and typically depend on only a single degree of freedom. The parameter θ expresses the strength of dependency in the model. Equation 6.2 shows the Clayton copula whose θ-range is $[-1,\infty)\backslash\{0\}$. $\theta = -1$ represents counter-monotonicity, $\theta \to 0$ gives the independence copula and $\theta \to \infty$ approaches co-monotonicity. Finally, **(3) Extreme value copulas** are robust in cases of extreme observations. The Gumbel copula (Equation 6.3) has a parameter space of θ in $[1,\infty)$. For $\theta = 1$ we obtain the independence copula, and, for $\theta \to \infty$ we approach co-monotonicity.

$$C_{Gaussian}(U) = \Phi_{\Sigma}(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_k)) \tag{6.1}$$

$$C_{Clayton}(U) = (1 + \theta(\sum_{i=1}^{k} \frac{1}{\theta}(u_i^{-\theta} - 1)))^{\frac{-1}{\theta}} \tag{6.2}$$

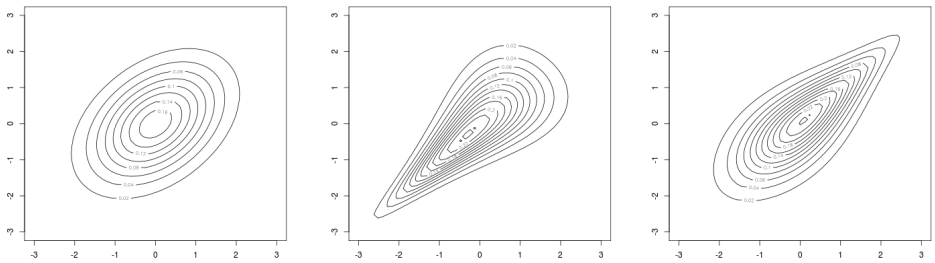

Figure 6.2: Examples of bivariate copula contour plots. (a) Gaussian copula, (b) Clayton copula with $\theta = 2.0$, (c) Gumbel copula with $\theta = 2.0$.

$$C_{Gumbel}(U) = \exp(-(\sum_{i=1}^{k}(-\log(u_i))^{\theta})^{\frac{1}{\theta}}) \tag{6.3}$$

The final probability density $c(U)$ of a copula can be obtained via partial differentiation. As we will see in the further course of this chapter, the copula density is the central component of our application of copulas for ranking purposes. While this chapter exclusively deals with fully differentiable copulas, there are other cases (e.g., empirical copulas given by a step function) in which we have to resolve to the use of *Markov Chain Monte Carlo* (MCMC) methods such as random walks in order to estimate the density function. Especially for higher choices of k, these techniques become less accurate.

$$c(U) = \frac{\partial^k}{\partial_1 \dots \partial_k} C(U) \tag{6.4}$$

Figure 6.2 shows contour plots of a number of bivariate standard copulas. Schmidt [191] reports the concrete choice of copula family and instantiation to fundamentally depend on the application domain. If no prior knowledge about the dependency structure, e.g., prevalence of asymptotic or tail dependencies, is available, practitioners often resort to goodness-of-fit tests or measures of tail dependency in order to choose an appropriate model. We will describe the use of these techniques in the subsequent sections when applying copulas for information retrieval problems.

Fitting copulas to observations

In the case of elliptical copulas, the fitting process is limited to calculating means and covariance matrices from the available observations. Here, the only degree of freedom is the concrete choice of distribution function (e.g., Gaussian vs. Student) that best approximates the original distribution that generated the observations. In the non-elliptical case, the task is to determine optimal settings of θ. Commonly, this is achieved by means of maximum likelihood estimates based on the available observations. This is also the approach chosen in this work. It should be noted that there are methods for direct empirical estimations of entire copula functions. The interested reader can find a good overview by Charpentier et al. [48] as a starting point for this line of research, the inclu-

sion of which would however go beyond the scope of this initial exploration of copulas for information retrieval.

6.2. Relevance Estimation

In the previous section, we described the theoretical foundations of copulas including concrete ways of computing $c(U)$ from multivariate observations U. We now detail their application for relevance estimation in information retrieval. First, we separately estimate the probability of relevance $P^{(k)}_{rel}(d)$ and non-relevance $P^{(k)}_{non}(d)$ for a document d, under each of the k dimensions – for example, topicality, recency, readability, etc. Next, we assume random observations U_{rel} and U_{non} to derive from these distributions and fit two distinct copulas, C_{rel} and C_{non}.

Recall that these copulas should capture the dependencies between relevance dimensions, in either the relevant (C_{rel}) or the non-relevant (C_{non}) documents retrieved. Since it is difficult to predict where these dependencies have the most effect, it is natural to consider three different general approaches of combining multivariate observation scores U into a single probability of relevance that can be used for resource ranking. **(1)** $CPOS(U)$ multiplies the independent likelihood of observing U with the relevance copula $c_{rel}(U)$, capturing only dependencies between the likelihoods of relevance. **(2)** $CNEG(U)$ normalizes the probability of relevance by the non-relevance copula $c_{non}(U)$, capturing only the dependencies between the likelihoods of non-relevance. **(3)** $CODDS(U)$, finally, multiplies the probability of relevance by the ratio of the two copulas, modelling simultaneously the dependencies between both previous notions.

$$CPOS(U) = c_{rel}(U) \prod_{i=1}^{k} u_i$$

$$CNEG(U) = \frac{\prod_{i=1}^{k} u_i}{c_{non}(U)}$$

$$CODDS(U) = \frac{c_{rel}(U)}{c_{non}(U)} \prod_{i=1}^{k} u_i$$

As performance baselines, we will compare to three popular combination methods from the literature: **(1)** $SUM(U)$ sums up the relevance scores across all dimensions k and uses the sum as the final ranking criterion. **(2)** $PROD(U)$ builds the product across all constituents. Probabilistically, this combination scheme assumes independence across all dimensions and can be expected to be too naïve in some settings where dependence is given. **(3)** Weighted linear combinations $LIN_\Lambda(U)$ build a weighted sum of constituents u_i with mixture parameters λ_i optimized by means of a parameter sweep with step size 0.1. It should be noted that all optimizations and parameter estimations, both for the baselines as well as for the copula models are conducted on designated training sets that do not overlap with the final test sets. We relied on the original training portion of the respective corpora. In the case that the original corpus did not specify a dedicated training set, we used a stratified 90%/10% split.

$$SUM(U) = \sum_{i=1}^{k} u_i$$

Table 6.1: Overview of experimental corpora.

ID	# docs	# topics	# labels	year
Blogs08	1.3M	100	38.2k	2008
Delicious	339k	180	3.8k	2012
ODP	22k	30	1k	2009

$$PROD(U) = \prod_{i=1}^{k} u_i$$

$$LIN_\Lambda(U) = \sum_{i=1}^{k} \lambda_i u_i$$

Based on three different standard datasets and tasks, we will highlight the merit of using copulas over the traditional approaches. Each of the settings specifies 2 individual relevance dimensions ($k = 2$) which are crucial for user satisfaction given the retrieval task. Table 6.1 gives a high-level overview of the relevant corpora that we used. Each of them will be described in more detail in the three following sections. Depending on the strength of tail dependency in the data, we will see varying improvements for the three inspected settings. Comparable as the scenarios appear, there seem to be significant underlying differences in the distribution of relevant documents that influence the benefit from the use of copulas. In Section 6.4, we will dedicate some room to a detailed investigation of when the use of copula-based retrieval models is most promising.

Opinionated blogs

When conducting marketing analyses for businesses, researching customer reviews of products or gauging political trends based on voter opinions, it can be desirable to focus the search process on subjective, non-factual documents. The Text REtrieval Conference (TREC) accounted for this task within the confines of their Blog Track between the years 2006 and 2010 [144]. The aim of the task is to retrieve blog feeds that are both topically relevant and opinionated. Our experimental corpus for this task is the Blogs08 collection specifically created for the venue. The dataset consists of 1.3 million blog feeds and is annotated by more than 38k manually created labels contributed by NIST assessors.

Each document is represented as a two-component vector $U_{rel}^{(2)}$. The first component refers to the document's topical relevance given the query and the second represents its degree of opinionatedness. In order for a document to be considered relevant according to the judges' assessments, it has to satisfy both conditions. Topical relevance was estimated by a standard BM25 model and opinionatedness was determined using the output of a state-of-the-art open source classifier[1]. After an initial evaluation of the domain, we chose Clayton copulas (Equation 6.2) to represent the joint distribution of topicality and opinionatedness. Table 2 shows a juxtaposition of performance scores for the baselines as well as the various copula methods. The highest observed performance per metric is highlighted by the use of bold typeface, statistically significant improvements (measured by means of a Wilcoxon signed-rank test at $\alpha = 0.05$-level) over each of the

[1] http://alias-i.com/lingpipe/

Table 6.2: Copula-based relevance estimation performance for opinionated blogs ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	**0.413**	0.360	0.181	0.289	**0.692**	0.275
SUM	0.400	0.333	0.154	0.255	0.689	0.238
LIN	0.387	0.333	0.162	0.262	0.689	0.245
CPOS	**0.413**	**0.400***	**0.182**	**0.306***	**0.692**	**0.287***
CNEG	0.373	0.373	0.181	0.290	0.545	0.245
CODDS	0.373	0.360	**0.182**	0.283	0.544	0.242

competing approaches are denoted by an asterisk. Of the baseline methods, the score product PROD performs best. However, introducing the use of copulas, we observe that the highest performance was achieved using the *CPOS* copula, which gave statistically significant gains in MAP, Bpref and precision at rank 10 over all the baseline methods.

At this point, we revisit the example query (Topic 1171) that was discussed in the introduction and depicted in Figure 6.1. For this topic, we observed a clear non-linear dependency structure alongside a lower-than-average linear combination performance of $AP = 0.14$. When applying CPOS to the topic, however, we obtain $AP = 0.22$, an improvement of over 50%.

Personalized bookmarks

Finding and re-finding resources on the Internet are frequently accompanied and aided by bookmarking. What started as a local in-browser navigation aid, has in recent years become an active pillar of the social Web society. Collaborative bookmarking platforms such as *Delicious*, *Furl*, or *Simpy* allow users to maintain an online profile along with bookmarks that can be shared among friends and collaboratively annotated by the user community. Research into tagging behaviour found that a significant amount of the tags assigned to shared media items and bookmarks are of subjective nature and do not necessarily serve as objective topical descriptors of the content [14]. This finding suggests that bookmarking has a strong personal aspect which we will cater for in our experiment. Vallet et al. [223] compiled a collection of more than 300k Delicious bookmarks and several million tags to describe them. For a share of 3.8k bookmarks and 180 topics, the authors collected manual relevance assessments along two dimensions, topical relevance of the bookmark given the topic and personal relevance of the bookmark for the user. This dataset is one of the very few corpora whose personalized relevance judgements were made by the actual users being profiled. We conduct a retrieval experiment in which we estimate topical and personal relevance for each document and use Gumbel copula models to model the joint distribution of dimensions. The set of relevant documents comprises only those bookmarks that satisfy both criteria and were judged relevant in terms of topicality and personal relevance. Table 6.3 shows an overview of the resulting retrieval performances. *CNEG* stands out as the strongest copula-based model but the overall ranking of systems depends on the concrete metrics evaluated. For some metrics such as precision at rank 10 and MRR, the linear combination baseline prevails, BPREF and precision at 5 documents favour CNEG.

Table 6.3: Copula-based relevance estimation performance for personalized bookmarks ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	0.084	0.079	**0.011**	0.051	0.192	0.043
SUM	0.095	0.095	**0.011**	0.071	0.192	0.055
LIN	0.126	**0.100***	**0.011**	0.077	**0.219***	0.063
CPOS	0.105	0.068	0.01	0.056	0.190	0.047
CNEG	**0.137***	0.090	0.010	**0.079***	0.184	**0.065**
CODDS	0.116	0.074	0.01	0.066	0.202	0.058

Table 6.4: Copula-based relevance estimation performance for child-friendly Websites ($k = 2$).

Method	P@5	P@10	p@100	BPREF	MRR	MAP
PROD	0.240	0.143	0.051	0.221	0.349	0.196
SUM	0.246	0.157	0.052	0.213	0.340	0.200
LIN	**0.320***	**0.187***	**0.071***	**0.275***	**0.357**	**0.235***
CPOS	0.238	0.140	0.053	0.215	0.351	0.200
CNEG	0.242	0.140	0.048	0.223	0.349	0.194
CODDS	0.241	0.143	0.052	0.220	0.349	0.196

Child-friendly Websites

The third application domain that we will inspect is concerned with the retrieval of child-friendly Websites. Children, especially at a young age, are an audience with specific needs that deviate significantly from those of standard Web users. As shown by Collins-Thompson et al. [53], even for adult users focussing the retrieval process on material of appropriate reading level can benefit user satisfaction. In the case of children, this tendency can be expected to be even more pronounced since young users show very different modes of interaction with search engines that reflect their specific cognitive and motor capabilities. Consequently, dedicated Web search engines for children should focus their result sets on topically relevant, yet age-appropriate documents. In Chapter 4, we constructed a corpus of 22k Web pages, 1,000 of which were manually annotated in terms of topical relevance towards a query as well as the document's likelihood of suitability for children. The class of suitable documents encompasses those pages that were topically relevant for children, presented in a fun and engaging way and textually not too complex to be understood. In our retrieval experiment, we account for both dimensions and require documents to be both on topic as well as suitable for children in order to be considered relevant. Table 6.4 gives an overview of the resulting retrieval performance. In this setting, the various copula models show comparable result quality as the non parametric baselines. Linear combinations with empirically learned weights, however, were consistently the strongest method. However we note that the distribution of child-suitable ratings has a very large mode at zero, with only a small number of non-zero scores taking a limited number of possible discrete values – limiting the amount of useful dependency information available that copulas could exploit.

6.3. Score Fusion

Previously, we investigated the usefulness of copulas for modelling multivariate document relevance scores based on a number of (largely) orthogonal dimensions of document quality. Now, we will address a different, closely related problem: *score fusion* (also known as an instance of data fusion). In this setting, rather than estimating document quality from the documents, we attempt to combine the output of several independent retrieval systems into one holistic ranking. This challenge is often encountered in the domains of metasearch or search engine fusion. To evaluate the score fusion performance of copula-based methods, we use historic submissions to the TREC Adhoc and Web tracks. We investigate 6 years of TREC (1995 - 2000) and fuse the document relevance scores produced by several of the original participating systems. Intuitively, this task closely resembles the previously addressed relevance estimation based on individual document properties. In practice, as we will show, the scenario differs from direct relevance estimation in that retrieval systems rely on overlapping notions of document quality (e.g., a variant of *tf/idf* scoring) and are therefore assumed to show stronger inter-dimension dependencies than individual dimensions of document quality might. Systematically, however, we address a set of document-level scores $U^{(k)}$, originating from k retrieval systems, exactly in the same way as we did document quality criteria in the previous section.

As performance baselines, we will rely on two popular score fusion schemes presented in [73], *CombSUM* and *CombMNZ*. CombSUM adds up the scores of all k constituent retrieval models and uses the resulting sum as a new document score. CombMNZ tries to account for score outliers by multiplying the cross-system sum by $NZ(U)$, the number of non-zero constituent scores.

$$CombSUM(U) = \sum_{i=1}^{k} u_i$$

$$CombMNZ(U) = NZ(U) \sum_{i=1}^{k} u_i$$

We introduce statistically principled, copula-based extensions of these established baseline methods: corresponding to CombSUM and CombMNZ, we define *CopSUM* and *CopMNZ* that normalize the respective baseline methods by the non-relevance copula.

$$CopSUM(U) = \frac{\sum_{i=1}^{k} u_i}{c_{non}(U)}$$

$$CopMNZ(U) = \frac{NZ(U) \sum_{i=1}^{k} u_i}{c_{non}(U)}$$

Due to the close relationship to the baseline methods, the effect of introducing copulas is easily measurable. Based on empirical evidence, we employ Clayton copulas to estimate $c_{non}(U)$.

Table 6.5 compares the baselines and copula methods in terms of MAP gain over the best, worst and median historic system run that were fused. Each performance score is averaged over 200 repetitions of randomly selecting k individual runs with k ranging

from 2 to 10 for each year of TREC. Statistically significant improvements over the respective baseline method, i.e., of CopSUM over CombSUM and CopMNZ over CombMNZ, are determined by a Wilcoxon signed-rank test at $\alpha = 0.05$ level and are denoted by an asterisk.

Regarding the baseline methods, CombSUM and CombMNZ perform equally well on average, but with a clear dataset bias. On TREC 4, 8 and 9, CombSUM performs consistently better than CombMNZ. For TREC 5, 6 and 7, the inverse is true. With the exception of TREC 4, the fused rankings do not match the performance of the single strongest run that contributed to the fusion.

Introducing the copula methods led to consistent improvements over their non-copula baseline counterparts. In 104 out of 168 cases, the copula-based fusion methods gave statistically significant gains, with only 14 out 168 performing worse than the corresponding baseline method. The copula-based methods achieved, on average, 7% gains over the corresponding baseline when comparing to the strongest fused system, 4% gain on median systems and 2% gain on the weakest systems.

Table 6.5: Score fusion performance based on historic TREC submissions. Evaluated in percentages of MAP improvements over the best, median, and worst original systems that were fused.

TREC 4	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.8	-	118	-4.2	20	1128	0.0	33.5	1709	3.0	39.6	2344	3.9	48.5	3116
CopSUM	-9.6*	-	116	-4.2	20.5*	1136	0.0	33.8*	1721	3.2*	40.0*	2350	4.0	49.2*	3125*
CombMNZ	-9.5	-	116	-5.4	18.3	1071	-1.1	31.6	1675	2.1	38.3	2310	3.6	48.0	3106
CopMNZ	-9.5	-	115	-5.5	18.2	1080	-1.0	31.9*	1689*	1.8	38.6*	2318*	3.8*	48.0	3117*

TREC 5	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-5.6	-	268	-10.6	12.5	614	-6.9	26.5	955	-5.3	34.3	1031	-5.6	40.1	1479
CopSUM	-5.2*	-	274*	-9.9*	13.0*	613	-6.7*	28.0*	972*	-4.9*	35.0*	1050*	-5.2*	43.3*	1503*
CombMNZ	-4.6	-	269	-6.7	17.4	652	-3.5	30.9	986	-2.5	38.2	1074	-3.3	43.5	1526
CopMNZ	-4.5	-	274*	-6.5	17.8*	667*	-3.1*	32.2*	991	-2.4	38.7*	1092	-3.0*	46.0*	1554*

TREC 6	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-18.5	-	486	-24.6	7.8	2235	-24.0	29.6	3950	-22.8	44.9	5585	-22.1	56.9	7685
CopSUM	-17.7*	-	471	-23.1*	9.1*	2279*	-22.9*	32.1*	4075*	-21.2*	48.3*	5699*	-20.8*	58.2*	7702
CombMNZ	-17.0	-	491	-18.6	15.5	2537	-18.1	38.8	4386	-16.7	55.0	6111	-17.3	65.0	8117
CopMNZ	-16.3*	-	490	-17.2*	17.4*	2601*	-17.9	40.5*	4458*	-16.7	59.6*	6202*	-16.4*	66.8*	8170

TREC 7	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.3	-	132	-16.2	6.2	303	-11.7	25.9	504	-12.8	30.0	708	-14.5	36.3	863
CopSUM	-9.4	-	145*	-15.8*	6.5*	321*	-11.1*	27.2*	538*	-12.3*	34.1*	734*	-13.8*	39.1*	877
CombMNZ	-8.8	-	130	-13.7	9.4	347	-10.1	28.1	538	-10.9	32.8	745	-13.1	38.5	891
CopMNZ	-8.8	-	139*	-13.3*	10.1*	363*	-10.2	30.5*	565*	-10.7	34.7*	786*	-12.4*	40.4*	922

TREC 8	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-15.9	-	475	-11.6	8.1	1188	-11.5	16.9	3194	-7.7	21.8	2739	-5.4	21.8	3372
CopSUM	-16.1	-	488	-10.1*	8.3	1201	-10.9*	16.7	3195	-7.3*	22.3*	2755	-4.3*	22.4	3397
CombMNZ	-17.2	-	421	-11.8	7.6	1273	-12.9	15.1	3209	-9.8	18.6	2660	-7.2	19.2	3266
CopMNZ	-17.3	-	447*	-11.2	7.9*	1292	-12.8	14.9	3216	-9.2*	19.7*	2685	-6.7*	20.5*	3301*

TREC 9	2 runs			4 runs			6 runs			8 runs			10 runs		
	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst	Best	Med.	Worst
CombSUM	-9.0	-	173	-14.9	20.4	473	-15.6	17.4	178	-21.3	18.9	202	-27.9	12.6	204
CopSUM	-8.5*	-	188*	-13.7*	21.2*	499*	-15.3	17.9	182	-20.9	19.2*	207	-26.6*	13.1*	206
CombMNZ	-11.0	-	155	-19.0	14.5	435	-17.4	14.4	172	-25.3	12.6	186	-32.7	4.7	184
CopMNZ	-10.7	-	167	-17.9*	16.0*	432	-17.1	14.7	176	-24.8*	13.0*	190	-30.4*	5.1*	187

Fusion robustness

There are significant differences in fusion effectiveness between individual editions of TREC. Comparing TREC 4 and TREC 6, for example, we observe that TREC 6 fusion results typically showcase performance losses in comparison to the best original run and very high gains for the weakest systems. We seek an explanation in the imbalance in performance of the original systems. Very weak systems have the potential of decreasing the overall quality of the fused result list by boosting the scores of non-relevant documents. As the number of very weak systems increases, so does the chance for performance losses introduced by fusion. When inspecting the number weak submissions (defined as having an MAP score that is at least 2 standard deviations lower than the average score across all participants) included in our fusion experiments, we find that, indeed, our TREC 6 sample includes ~27% more weak systems than that of TREC 4.

In order to further investigate the influence of weak runs on overall fusion performance and to measure the proposed methods' robustness against this effect, we turn to the 10-system fusion scenario and inject more and more weak systems among the regular ones. Figure 6.3 shows how the fusion improvement over the single strongest system of TREC 4 is affected as the number of weak submissions ranges from 0 to 9 out of 10. As before, each data point is an average across 200 fusions of randomly drawn runs. In the ideal setting, in which there are no weak systems, we note higher performance gains than in the uncontrolled scenario that was shown in Table 6.5. As the number of weak systems injected into the fusion increases, performance scores quickly drop. As noted earlier, CombSUM performs slightly better on TREC 4 than CombMNZ. This difference, however, is not further influenced by the number of weak systems. The copula-based fusion methods are more resistant to the influence of weak systems. We note the divide between copula-methods and baseline approaches growing as the number of weak systems increases. Each baseline system score is well-separated from the respective copula-based variant. Error bars in Figure 6.3 were omitted to prevent clutter.

6.4. When to use copulas?

In Section 6.2, we investigated three different domains in which we apply copulas to model the joint distribution of multivariate relevance scores. For each of these settings, we could observe varying degrees of usefulness of the proposed copula scheme. While for child-friendly Web search, the linear baseline performed best, we achieved significant improvements in the opinionated blog retrieval setting. At this point, we investigate the reason for this seeming imbalance in performance gains in order to find a way of deciding for which problem domains the application of copulas is most promising.

One of the key properties of copulas is their ability to account for tail dependencies. Formally, tail dependence describes the likelihood that component $u_{rel,i}$ within the observation vector $U_{rel}^{(k)}$ will take on extremely high or low values, as another component $u_{rel,j}$ with $i \neq j$ also takes an extreme value. The strength of this correlation in extreme regions is expressed by the *tail dependency indices* I_U and I_L for upper and lower tail dependency, respectively. Higher values of I signal stronger dependencies in the respective tail regions of the distribution.

$$I_U = P\{X_1 > F_i^{-1}(u_i) | X_2 > F_j^{-1}(u_j)\}$$

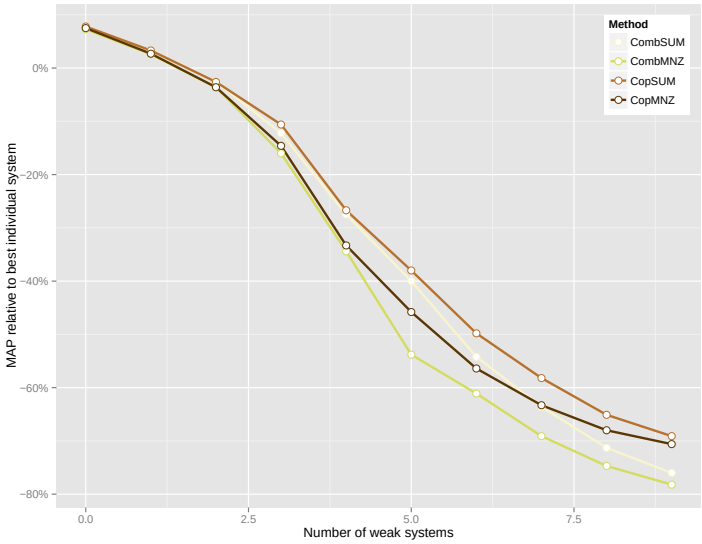

Figure 6.3: Performance in terms of MAP when 0…9 out of 10 fused original systems are weak.

$$I_L = P\{X_1 \leq F_i^{-1}(u_i) | X_2 \leq F_j^{-1}(u_j)\}$$

The literature has brought forward a number of estimators of the tail indices. We use the R implementation of [75]'s method. Tail index estimates serve as good tools for separating domains where we are likely to observe performance gains (blog and bookmark retrieval) and those that do not match linear combination performance (child-friendly search). Based on the respective copula models that we fit to our observations, the blog retrieval ($I_L = 0.07$) and personalized bookmarking ($I_U = 0.49$) show moderate tail dependencies while the child-friendly Web search task has no recognizable dependency among extrema ($I_L = I_U = 0$). Since the comparison of absolute tail index scores across observations is not meaningful, we are interested in a method to further narrow down the expected performance. To this end, we took a closer look at the actual data distribution, and investigated goodness-of-fit tests that are used to determine how well an assumed theoretical distribution fits the empirical observations. The higher the likelihood of our observations to have been generated by the copula models that we estimated, the higher resulting performance we can expect. We apply a standard Anderson-Darling test [18] to determine how well the observations are represented by the copula models. In the personalized bookmarking setting, we obtain $p = 0.47$ and for the blog data $p = 0.67$ for the null hypothesis of the observations originating from the present copula model. As we suspected based on the tail dependency strength, the child-friendly Web search data only achieved a probability of fit of $p = 0.046$.

To summarize, in this section, we have shown how a combination of tail dependence indices and goodness-of-fit tests can be used to help differentiate between domains that may benefit from copula-based retrieval models and those that may not.

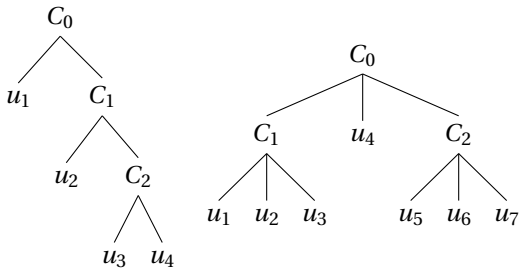

Figure 6.4: Examples of fully nested (left) and partially nested (right) copulas.

6.5. Nested copulas

In the previous sections, we demonstrated the core functionality of copulas. For simplicity, we chose several low-dimensional examples that allowed for a detailed performance analysis. The true strength of the copula model family, however, lies in their ability to capture complex dependency structures in high-dimensional relevance spaces. To this end, we will now turn to high-dimensional relevance settings with more than one hundred individual dimensions. Additionally, we further refine the previous approach by introducing nested copulas, an expansion to the basic copula framework that aims especially at high-dimensional settings.

Earlier in this chapter, we modelled low-dimensional document relevance spaces with a single copula with k components, where k equalled the cardinality of the entire relevance space. While this is possible in high-dimensional spaces as well, alternative options offer additional degrees of freedom. The use of so-called *nested* copulas is one such method. Instead of combining all dimensions in a single step as described earlier, they allow for a nested hierarchy of copulas that estimate joint distributions for sub sets of the full relevance space and subsequently combine them until one global model is obtained. Formally, fully nested copulas with k dimensions are given by

$$C(U) = C_0(u_1, C_1(u_2, C_2(\ldots, C_{k-2}(u_{k-1}, u_k))))$$

By means of the structure of the nesting "tree", nested copulas can explicitly model which dimensions depend on each other directly. The respective θ_i parameters determine the strengths of these (per dimension) dependencies. This mechanism gives nested copulas a theoretical advantage in flexibility over their non-nested counterparts. As an alternative approach to full nesting, partially nested copulas hierarchically combine subsets of dimensions. Figure 6.4 shows a fully nested copula with $k-1$ copula modelling steps (left) and a conceptual example of a partially nested copula (right).

6.5.1. Data Set

In order to evaluate the use of copulas for high-dimensional relevance spaces, we use the *MSLR-WEB10K* and *WEB30K* datasets, two publicly available collections of 10,000 (30,000, respectively) real Web search queries and an annotated sample of hundreds of thousands of related impressions. For each query-url pair, a set of 136 features is available. The majority of the feature space considers dimensions related to topicality such as

tf/idf scores, query term frequency in different sections of the page. There are, however, several features that capture alternative relevance criteria such as general page authority, quality or textual complexity. For an overview of the full list of features, please consult the data set Web page [2]. The corpus is pre-partitioned into 5 folds to allow for cross validation in a 3-1-1 split of training, validation and test sets.

6.5.2. Experiments

This section discusses our experimental set-up and findings. All experimental results were obtained by means of 5-fold cross validation on the *MSLR-WEB10K* and *MSLR-WEB10K* datasets. In order to set the performance of the various copula models into perspective, we include a common weighted linear combination scheme l as a baseline. Concrete settings of the mixture parameters λ_i are determined based on a greedy parameter sweep (ranging from $0 \ldots 1$ in steps of 0.005) on the training set of each CV fold.

$$l(U) = \sum_{i=1}^{k} \lambda_i u_i$$

Additionally, we compare to LambdaMART [238], a competitive learning-to-rank baseline. The relevant model parameters are tuned on the validation set. We rely on the implementation of the GBM package for R[3].

We investigate three types of copula models. The "flattest" nesting hierarchy is given by copulas without any sub-nesting. All 136 dimensions are included in a single model, describing all inter-dimensional dependencies by a single parameter θ. This strategy is equivalent to the method presented in Section 6.1. To study some simple, yet indicative examples of nested copulas, we include a fully nested approach in which the nesting order is determined randomly and the average results across 50 randomizations are reported. Note that the concrete nesting structure is an additional degree of freedom that holds significant modelling power. Finally, as an example of partially nested copulas, we rely on the existing semantic grouping of dimensions in the dataset (e.g., all *tf/idf* features, or all *query term coverage* features) and estimate individual copulas C_{d_i} for each group d_i. All of these group copulas are then combined in an overall copula $C_{partial}(U)$ in arbitrary order (again randomized 50 times). Groups that comprise only a single dimension are directly included into $C_{partial}(U)$. This will become especially relevant for the experiments presented later in Figure 6.5. For the extreme case of exclusively single-dimensional groups, this model becomes equivalent to non-nested copulas. All copula experiments presented in this paper are based on the publicly available implementation for R [241].

For model comparison, we use two well known metrics: Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (MAP). Table 6.6 shows the resulting cross-validation performance of the respective methods on the full 136-dimensional datasets. Statistically significant performance improvements with respect to the linear combination baseline are denoted by the $*$ character. We used a Wilcoxon signed rank test with $\alpha < 0.05$ confidence level.

[2] http://research.microsoft.com/en-us/projects/mslr/feature.aspx
[3] http://cran.r-project.org/web/packages/gbm/

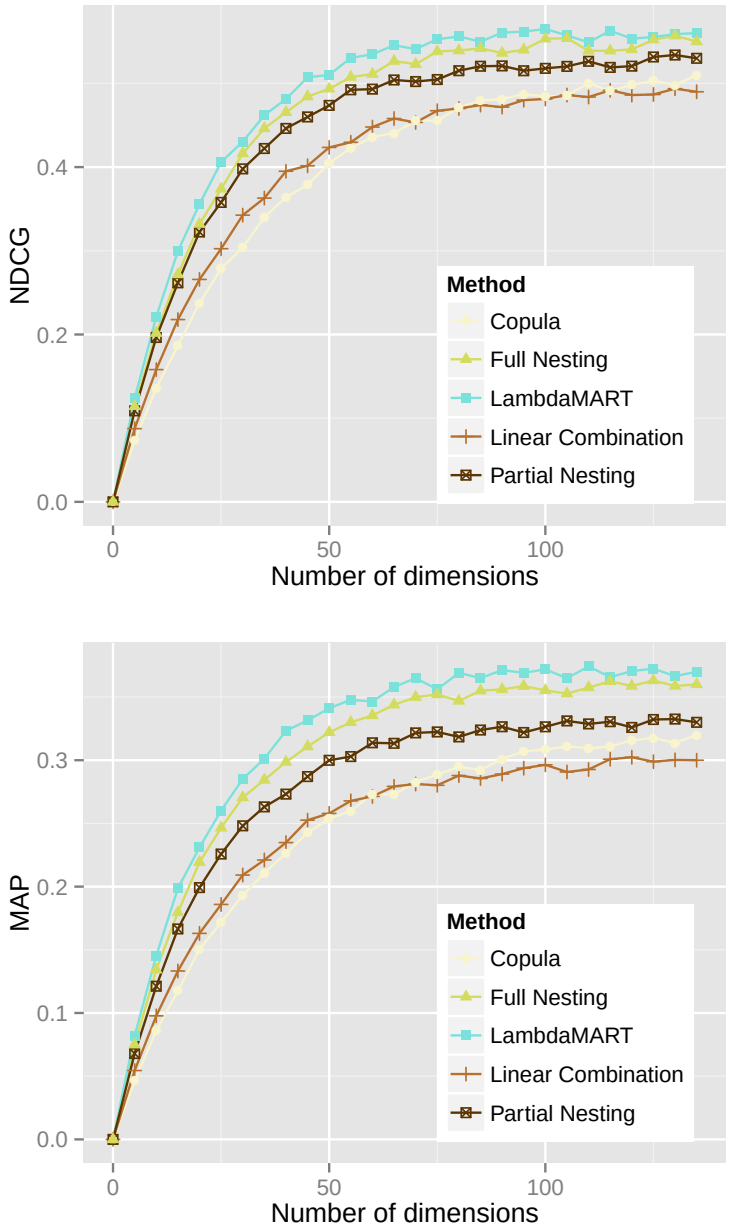

Figure 6.5: NDCG and MAP as functions of the dimensionality of the underlying *MSLR-WEB10K* relevance space.

Table 6.6: Performance comparison of copula and L2R models.

Method	$NDCG_{10K}$	$NDCG_{30K}$	MAP_{10K}	MAP_{30K}
Linear Comb.	0.49	0.46	0.30	0.28
LambdaMart	0.56*	0.55*	0.37*	0.37*
Copula	0.51	0.50*	0.32	0.32*
Nested	0.53*	0.54*	0.33*	0.33*
Fully Nested	0.54*	0.54*	0.36*	0.35*

We can note a clear ordering of approaches in which linear combinations achieve the lowest overall performance and the LambdaMART method delivers the best results. The various copula-based models lie between these extremes. Global copulas show slightly better performance than a linear feature combination, these differences were, however, not found to be significant. For both forms of nested copulas, we can note significantly higher scores in terms of NDCG and MAP. With respect to our research questions, we note that copula-based models, especially nested ones show strong ranking performance for high-dimensional settings. Fully nested copulas, especially, approximate the performance of the learning-to-rank model to the degree, that we could not note any statistically significant differences between the two methods.

In order to further investigate the individual performances of the various methods as the dimensionality of the relevance space increases, we modify the setting by varying the number of dimensions k between 1 and 136. Figure 6.5 shows the results of this experiment in terms of NDCG and MAP on the *MSLR-WEB10K* dataset. For each choice of k, we randomly sample $n = 100$ feature subsets, train the respective models on each set and average the resulting retrieval performance. For all methods, we note steep performance gains with each dimensions that is added early on. These improvements slowly level out and reach a largely stable performance for relevance spaces of size $75 \leq k \leq 136$. An especially noteworthy observation can be made in the comparison of global copulas and linear combination models. While early on, linear models show higher scores in both metrics, this tendency reverses for higher dimensional spaces ($60 \leq k \leq 80$). In the previous sections, we noted competitive ranking performance of linear combination models for most of our experimental corpora. As we can see from the current example, even in domains that are well represented by linear models, copulas can achieve performance gains as the problem scales up in dimensionality.

6.6. Conclusion

In this chapter, we introduced the use of *copulas*, a powerful statistical framework for modelling complex dependencies, for information retrieval tasks. We demonstrated the effectiveness of copula-based approaches in improving performance on several standard IR challenges. First, we applied copulas to the task of multivariate document relevance estimation, where each document is described by several potentially correlated relevance dimensions. We learned and evaluated copula models for three different IR tasks, using large-scale standard corpora: (1) opinionated blog retrieval; (2) personalized social bookmarking; and (3) child-friendly Web search, obtaining significant im-

provements on the first two of these tasks. Second, following up on Research Question 3.b), we investigated the performance differences of copula models between different domains, and proposed the use of tail dependency indices and goodness-of-fit tests to understand the likely effect of using copulas for a given scenario. Addressing Research Question 3.c), we introduced copula-based versions of two existing score fusion methods, COMB-SUM and COMB-MNZ, and showed that these improve the performance of score fusion on historic TREC submissions, in terms of both effectiveness and robustness, compared to their non-copula counterparts. Finally, we demonstrated the use of nested copulas, an expansion that equips the basic model with a potentially high number of additional degrees of freedom and that is especially suited for high-dimensional relevance spaces.

References

[14] Morgan Ames and Mor Naaman. "Why we tag: motivations for annotation in mobile and online media". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 971–980.

[18] Theodore W. Anderson and Donald A. Darling. "A test of goodness of fit". In: *Journal of the American Statistical Association* 49.268 (1954), pp. 765–769.

[19] Javed A. Aslam and Mark Montague. "Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems (poster session)". In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2000, pp. 379–381.

[23] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. "Formal models for expert finding in enterprise corpora". In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2006, pp. 43–50.

[35] Gloria Bordogna and Gabriella Pasi. "A model for a soft fusion of information accesses on the web". In: *Fuzzy sets and systems* 148.1 (2004), pp. 105–118.

[39] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[41] Chris Burges et al. "Learning to rank using gradient descent". In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 89–96.

[48] Arthur Charpentier, Jean-David Fermanian, and Olivier Scaillet. "The estimation of copulas: Theory and practice". In: *Copulas: From theory to application in finance* (2007), pp. 35–60.

[53] Kevyn Collins-Thompson et al. "Personalizing web search results by reading level". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 403–412.

[54] Célia da Costa Pereira, Mauro Dragoni, and Gabriella Pasi. "Multidimensional relevance: A new aggregation criterion". In: *Advances in information retrieval*. Springer, 2009, pp. 264–275.

[55] Nick Craswell et al. “Relevance weighting for query independent evidence”. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 416–423.

[69] Paul Embrechts, Filip Lindskog, and Alexander McNeil. “Modelling dependence with copulas and applications to risk management”. In: *Handbook of heavy tailed distributions in finance* 8.1 (2003), pp. 329–384.

[73] Edward Fox and Joseph Shaw. “Combination of multiple searches”. In: *NIST SPECIAL PUBLICATION SP* (1994), pp. 243–243.

[75] Edward W. Frees and Emiliano A. Valdez. “Understanding relationships using copulas”. In: *North American actuarial journal* 2.1 (1998), pp. 1–25.

[80] Shima Gerani, ChengXiang Zhai, and Fabio Crestani. “Score transformation in linear combination for multi-criteria relevance ranking”. In: *Advances in Information Retrieval*. Springer, 2012, pp. 256–267.

[95] W. Höffding. “Scale-invariant correlation theory”. In: *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin* 5.3 (1940), pp. 181–233.

[99] Xuanjing Huang and W. Bruce Croft. “A unified relevance model for opinion retrieval”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 947–956.

[117] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. “The importance of prior probabilities for entry page search”. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, pp. 27–34.

[136] Tie-Yan Liu. “Learning to rank for information retrieval”. In: *Foundations and Trends in Information Retrieval* 3.3 (2009), pp. 225–331.

[141] Wei Lu, Stephen Robertson, and Andrew MacFarlane. “Field-weighted XML retrieval based on BM25”. In: *Advances in XML Information Retrieval and Evaluation*. Springer, 2006, pp. 161–171.

[144] Craig Macdonald et al. “Blog track research at TREC”. In: *ACM SIGIR Forum*. Vol. 44. ACM. 2010, pp. 58–75.

[145] Raghavan Manmatha, Toni M. Rath, and Fangfang Feng. “Modeling score distributions for combining the outputs of search engines”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 267–275.

[155] Mark Montague and Javed A. Aslam. “Condorcet fusion for improved retrieval”. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM. 2002, pp. 538–548.

[156] Mark Montague and Javed A. Aslam. “Relevance score normalization for metasearch”. In: *Proceedings of the tenth international conference on Information and knowledge management*. ACM. 2001, pp. 427–433.

[166] Arno Onken et al. “Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation”. In: *PLoS computational biology* 5.11 (2009), e1000577.

[176] Filip Radlinski and Thorsten Joachims. “Query chains: learning to rank from implicit feedback”. In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 239–248.

[177] Benjamin Renard and Michel Lang. “Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology”. In: *Advances in Water Resources* 30.4 (2007), pp. 897–912.

[182] Stephen E. Robertson, Hugo Zaragoza, and Michael Taylor. “Simple BM25 extension to multiple weighted fields”. In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM. 2004, pp. 42–49.

[191] Thorsten Schmidt. “Coping with copulas”. In: *Chapter forthcoming in Risk Books: Copulasfrom theory to applications in finance* (2006).

[192] Christian Schoelzel and Petra Friederichs. “Multivariate non-normally distributed random variables in climate research–introduction to the copula approach”. In: *Nonlin. Processes Geophys.* 15.5 (2008), pp. 761–772.

[201] Abe Sklar. “Fonctions de répartition à n dimensions et leurs marges”. In: *Publ. Inst. Statist. Univ. Paris* 8.1 (1959), p. 11.

[220] Theodora Tsikrika and Mounia Lalmas. “Combining evidence for relevance criteria: a framework and experiments in web retrieval”. In: *Advances in Information Retrieval*. Springer, 2007, pp. 481–493.

[223] David Vallet and Pablo Castells. “Personalized diversification of search results”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 841–850.

[224] Christopher C. Vogt and Garrison W. Cottrell. “Fusion via a linear combination of scores”. In: *Information Retrieval* 1.3 (1999), pp. 151–173.

[238] Qiang Wu et al. “Ranking, boosting, and model adaptation”. In: *Technical Report, MSR-TR-2008-109* (2008).

[239] Shengli Wu and Fabio Crestani. “Data fusion with estimated weights”. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM. 2002, pp. 648–651.

[241] Jun Yan. “Enjoy the joy of copulas: with a package copula”. In: *Journal of Statistical Software* 21.4 (2007), pp. 1–21.

7

Conclusion

In this thesis, we investigated several challenges in delivering contextual, multidimensional relevance models for a range of information retrieval applications. Part 1 of the thesis started out with a motivation of our problem by giving examples of two types of search context (person-specific and situation-specific) that can significantly influence the retrieval process. In the second part, we investigated several alternative ways of estimating individual relevance dimensions. Concretely, we employed content-based and interaction-based automatic classification methods as well as human computation in order to judge the relevance of textual and multimedia documents. Finally, in Part 3, we investigated the use of the copula framework to formally model document relevance based on a number of previously estimated relevance dimensions.

Let us now return to the three fundamental research questions that inspired the work presented in this dissertation and that were stated in Chapter 1:

Q1: How is relevance influenced by subjective and situational factors?

How does search behaviour differ between individual user groups? Our dedicated user studies in Chapter 2 showed that children require a very different mix of relevance dimensions than adult users. In all studies, we observed children struggling with formulating queries, scanning result lists, understanding result pages and distinguishing content from sponsored material and advertisements. Similar tendencies are reported by others, e.g., [66]. We accounted for these different needs by designing a child-friendly search system with query assistance functionality, comprehension aids and material from trusted sources. In our user study, we saw a clear tendency of children preferring the aided search system over the generic Web search engine for adults. Children need different documents than adults simply by virtue of being children. Let this serve as a concrete example for the subjective nature of relevance that shows clear differences between individual users and groups of users.

How to predict search success based on search behaviour? We observed significant differences in the behaviour of young and adult searchers. In a dedicated effort, we in-

vestigated whether search behaviour is an indicator of expected search success. Users that used short keyword queries and scanned SERPs as well as clicked pages in an efficient manner are much more likely to succeed in their searches than those that formulate verbose natural language queries while performing excessive mouse movements on the page. This observation, again, suggests that the fundamental differences between users, and user groups, here manifested in search behaviour, require different tools, means of support and relevance criteria in order to achieve search success.

Are there situational influences on search behaviour and relevance? Having established the importance of the searcher's personal context, we investigated the way and extent to which situational factors can influence document relevance and search behaviour. In Chapter 2, we observed evidence of searchers straying from their regular behaviour in order to satisfy novel, often externally motivated, information needs. While such *atypical searches* occur only rarely to each individual searcher, in sum, they represent a significant volume of search sessions and affect the majority of the user base of modern search engines. This observation highlights the importance of considering signals beyond the general topical preferences and needs of the user by including the situational context in which the search is motivated and conducted.

How can we personalize search result lists when they diverge from the user's general profile? Finally, we investigated in which way existing user profiles can be used for personalizing atypical search sessions. Comparing different strategies, including regular personalization, no personalization at all, and several hybrid forms, we concluded that atypical sessions can benefit from personalization. In such cases, one should not rely on the full global history, but should use the immediate (situational) session context instead.

Q2: How to estimate individual relevance dimensions?

Relevance estimation based on document content. In Chapter 4, we investigated the use of automatic regression methods for estimating the child-suitability dimension of Web pages based on document content. Using a wide range of on-page features as well as network topology properties, we were able to achieve near-human accuracy. We noted that certain topics, especially those belonging to technical or scientific domains, had strongly varying inter-annotator agreement ratios due to variance in education styles and personal definitions of suitability.

Estimation based on user interaction. As an alternative to content-based relevance estimation, we determined topical document relevance on the basis of user comments in which YouTube videos are discussed. The greatest challenge towards this goal was to appropriately deal with the significant amounts of noise inherent to the medium. In Chapter 4, time series analyses and distributional semantics techniques allowed us to achieve significant gains in retrieval performance by enriching state-of-the-art indexing strategies by extracted terms.

How to use crowdsourcing for relevance estimation? Despite the considerable power of automatic learning methods, there are many inference scenarios in which humans clearly outperform machines. Crowdsourcing is an efficient way of harnessing considerable work forces for such human intelligence tasks. In Chapter 5, we demonstrated how to phrase standard tasks such as the estimation of topical document relevance as crowdsourcing jobs. A special focus was robustness to sloppy workers. To this end, we found that tasks should be designed with variability and ingenuity in mind in order to discourage workers from trying to "automate" their work either by means of scripts and robots or by submitting arbitrary judgements.

How to cast the relevance estimation process as a game? Based on the assumption that entertainment-seeking users are less likely to submit low-quality judgements, we phrased the topical relevance annotation task in the form of a game. As a result, we noted significantly higher agreement ratios and much fewer low-quality submissions at much lower pay rates. Additionally, we found evidence to question the common practice of restricting crowdsourcing jobs to First-World audiences. When using a gamified approach that explicitly addresses entertainment seekers, we could not note any significant correlations between country of origin and worker reliability.

Q3: How to obtain a single probability of relevance?

How do copulas compare to established multidimensional relevance frameworks? In Chapter 6, we demonstrated the use of copula-family models for relevance estimation. In three different scenarios, we compared their performance of combining individual relevance dimensions with standard baselines such as weighted averages or score multiplications. The resulting estimates of overall document relevance were found to be of greatly varying accuracy depending on the domain at hand, ranging from mediocre quality to clearly superior results.

Is there a way of predicting the merit of using copulas for a given domain? Due to these volatile results, we tried to establish a forecast of how well copulas will work for a given relevance space and distribution. The copula family allows for easy modelling of co-movements in extreme regions across relevance dimensions, so-called tail dependencies. We found that indicators of tail dependency strength are good indicators of the merit for applying copula-based models. The stronger the tail dependencies, the more promising is the use of copulas. Furthermore, well-known goodness-of-fit tests were able to predict how accurately the model represented the true distribution of relevance.

How robust are copula-based relevance models to outliers? We studied the quality and robustness of fusing multiple original result rankings of historic TREC submissions. Many state-of-the-art fusion approaches suffered significant performance losses when bad individual rankings were added to the fusion. Some modern competitors pre-scan for such low-quality contributions and assign mixtures weights such, that unreliable contributions receive only mild consideration for the overall score. The copula framework allowed us to achieve superior robustness against low-quality outliers without having to make any modifications to their formal framework.

Summary

Information retrieval systems centrally build upon the concept of *relevance* in order to rank documents in response to a user's query. Assessing relevance is a non-trivial operation that can be influenced by a multitude of factors that go beyond mere topical overlap with the query. This thesis argues that relevance depends on personal (Chapter 2) and situational (Chapter 3) context. In many use cases, there is no single interpretation of the concept that would optimally satisfy all users in all possible situations.

We postulate that relevance should be explicitly modelled as a composite notion comprised of individual relevance dimensions. To this end, we show how automatic inference schemes based on document content and user activity can be used in order to estimate such constituents of relevance (Chapter 4). Alternatively, we can employ human expertise, harnessed, for example, via commercial crowdsourcing or serious games to judge the degree to which a document satisfies a given set of relevance dimensions (Chapter 5).

Finally, we need a model that allows us to estimate the joint distribution of relevance across all previously obtained dimensions. In this thesis, we propose using *copulas*, a model family originating from the field of quantitative finances that decouples observations and dependency structure and which can account for complex non-linear dependencies among relevance dimensions (Chapter 6).

Samenvatting

Information retrieval systemen zijn gebaseerd op het concept van *relevantie* om documenten te rangschikken voor de zoekopdracht van een gebruiker. Het bepalen van relevantie is een niet-triviale stap die kan worden beïnvloed door een groot aantal factoren. Het is daarom niet voldoende om relevantie als pure overlap tussen trefwoorden en documenten te meten. Dit proefschrift stelt dat relevantie afhankelijk is van persoonlijke (Hoofdstuk 2) en situationele (Hoofdstuk 3) context. In veel gevallen bestaat er geen enkele interpretatie van het concept dat optimaal zou voldoen aan de eisen van alle gebruikers in alle mogelijke situaties.

We postuleren dat relevantie expliciet moet worden gemodelleerd als een combinatie van onafhankelijke dimensies. Met deze doelstelling tonen wij hoe automatische methodes die gebaseerd zijn op documentinhoud en gebruikersactiviteit kunnen worden gebruikt om de bestanddelen van relevantie te schatten (Hoofdstuk 4). Als alternatief voor deze automatische benaderingen kunnen we gebruik maken van menselijke ervaring, bijvoorbeeld via commerciële *crowdsourcing* of *serious games*, om te beoordelen in welke mate een document voldoet aan een bepaalde set van dimensies van relevantie (Hoofdstuk 5).

Tot slot hebben we een formeel model nodig dat ons een schatting van de gezamenlijke distributie van relevantie in alle eerder verkregen dimensies geeft. In dit proefschrift gebruiken wij *copula's*, een model afkomstig uit het vakgebied van de kwantitatieve financiële economie, die marginale verdelingen ontkoppelen van hun afhankelijkheidsstructuur en die bijzonder geschikt zijn om complexe niet-lineaire afhankelijkheden tussen relevantie dimensies af te beelden (Hoofdstuk 6).

Acknowledgements

- Special thanks go out to Esther for her unconditional love, for cheering me up and for supporting me to the point of camping with me in the office during paper-writing night shifts before SIGIR deadlines.
- I would like to thank my family for always encouraging my curiosity and supporting me all these years in all matters small and big.
- A big thanks goes to my promotor, Arjen de Vries. Over the past years he taught me much about IR but also research and life in general and has become a dear friend.
- Many thanks to my friends for being great guys and being there whenever you are needed. Come wind, come rain, come Orc hordes with spiky things on sticks, you are the best.
- I would like to thank the staff and patients at Emma Kinderziekenhuis in Amsterdam for their valuable input that led to the development and refinement of the EmSe system described in Chapter 2.
- The research described in this thesis was in part supported by the PuppyIR project[1]. It was funded by the European Community's Seventh Framework Programme FP7/ 2007-2013 under grant agreement no. 231507.
- Finally, I would like to thank the ACM special interest group on information retrieval (SIGIR) for having supported my conference travels throughout the years.

[1] http://www.puppyir.eu

List of Publications

Part I - Motivation

1. **"Designing Human-Readable User Profiles for Search Evaluation"**. C. Eickhoff, K. Collins-Thompson, P. Bennett and S. Dumais. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR), Moscow, Russia*, 2013

2. **"Personalizing Atypical Web Search Sessions"**. C. Eickhoff, K. Collins-Thompson, P. Bennett and S. Dumais. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM), Rome, Italy*, 2013

3. **"Supporting Children's Web Search in School Environments"**. C. Eickhoff, P. Dekker, and A. P. de Vries. In *Proceedings of the 4th Conference on Information Interaction in Context (IIiX), Nijmegen, The Netherlands*, 2012

4. **"EmSe: Initial Evaluation of a Child-friendly Medical Search System"**. C. Eickhoff, L. Azzopardi, D. Hiemstra, F. de Jong and A. P. de Vries. In *Proceedings of the 4th Conference on Information Interaction in Context (IIiX), Nijmegen, The Netherlands*, 2012

5. **"EmSe: Supporting Children's Information Finding Needs within a Hospital Environment"**. L. Azzopardi, R. Glassey, K. Gyllstrom, F. van der Sluis, C. Eickhoff and S. Duarte. In *Proceedings of the 34th European Conference on Information Retrieval (ECIR), Barcelona, Spain*, 2012

6. **"Web Search Query Assistance Functionality for Young Audiences"**. C. Eickhoff, T. Polajnar, K. Gyllstrom, S. Duarte Torres and R. Glassey. In *Proceedings of the 33rd European Conference on Information Retrieval (ECIR), Dublin, Ireland*, 2011

Part II - Relevance Measures

1. **"Exploiting User Comments for Audio-visual Content Indexing and Retrieval"**. C. Eickhoff, W. Li and A. P. de Vries. In *Proceedings of the 35th European Conference on Information Retrieval (ECIR), Moscow, Russia*, 2013

2. **"Increasing Cheat Robustness of Crowdsourcing Tasks"**. C. Eickhoff and A. P. de Vries. In *Information Retrieval, Springer*, 2012

3. **"Quality through Flow and Immersion: Gamifying Crowdsourced Relevance Assessments"**. C. Eickhoff, C. G. Harris, A. P. de Vries and P. Srinivasan. In *Proceedings of the 35th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Portland, Oregon, USA*, 2012

4. **"GEAnn - Games for Engaging Annotations"**. C. Eickhoff, C. G. Harris, P. Srinivasan and A. P. de Vries. In *Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR), Beijing, China*, 2011

5. **"How Crowdsourcable is Your Task?"**. C. Eickhoff and A. P. de Vries. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM), Hong Kong, China*, 2011

6. **"A Combined Topical/Non-topical Approach to Identifying Web Sites for Children"**. C. Eickhoff, P. Serdyukov and A. P. de Vries. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining (WSDM), Hong Kong, China*, 2011

7. **"GeAnn at TREC 2011"**. C. Eickhoff, C. G. Harris, P. Srinivasan and A. P. de Vries. In *The 20th Text REtrieval Conference (TREC 2011) Notebook*, 2011

8. **"Identifying Suitable YouTube Videos for Children"**. C. Eickhoff and A. P. de Vries. In *Proceedings of the 3rd Networked and Electronic Media Summit (NEM), Barcelona, Spain*, 2010

9. **"Web Page Classification on Child Suitability"**. C. Eickhoff, P. Serdyukov and A. P. de Vries. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM), Toronto, Canada*, 2010

Part III - Multivariate Relevance

1. **"Modelling Complex Relevance Spaces with Copulas"**. C. Eickhoff and A. P. de Vries. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management (CIKM), Shanghai, China*, 2014

2. **"Copulas for Information Retrieval"**. C. Eickhoff, A. P. de Vries and K. Collins-Thompson. In *Proceedings of the 36th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Dublin, Ireland*, 2013

3. **"Relevance as a Subjective and Situational Multidimensional Concept"**. C. Eickhoff. In *Proceedings of the 35th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Portland, Oregon, USA*, 2012

Other Publications

1. **"Interactive Summarization of Social Media"**. W. Li, C. Eickhoff, and A. P. de Vries. In *Proceedings of the 6th Conference on Information Interaction in Context (IIiX), Regensburg, Germany*, 2014

2. **"Crowd-Powered Experts"**.
C. Eickhoff. In *Proceedings of the ECIR Workshop on Gamification for Information Retrieval (GamifIR), Amsterdam, The Netherlands*, 2014

3. **"Lessons from the Journey: A Query Log Analysis of Within-Session Learning"**. C. Eickhoff, J. Teevan, R. White and S. Dumais. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining (WSDM), New York, USA*, 2014

4. **"Geo-Spatial Domain Expertise in Microblogs"**. W. Li, C. Eickhoff and A. P. de Vries. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR), Amsterdam, The Netherlands*, 2014

5. **"Proceedings of the 13th Dutch-Belgian Workshop on Information Retrieval"**. C. Eickhoff and A. P. de Vries (Eds.). In *CEUR Workshop Proceedings*, 2013

6. **"The Downside of Markup: Examining the Harmful Effects of CSS and Javascript on Indexing Today's Web"**. K. Gyllstrom, C. Eickhoff, A. P. de Vries and M. F. Moens. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), Maui, USA*, 2012

7. **"BooksOnline'12: 5th Workshop on Online Books, Complementary Social Media, and Crowdsourcing"**. G. Kazai, M. Landoni, C. Eickhoff, and P. Brusilovsky. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management (CIKM), Maui, USA*, 2012

8. **"Towards Role Detection in Virtual Worlds"**. C. Eickhoff and V. P. Lavrenko. In *Computers in Entertainment (CIE), ACM*, 2012

9. **"Want a Coffee? Predicting Users' Trails"**. W. Li, C. Eickhoff and A. P. de Vries. In *Proceedings of the 35th Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR), Portland, Oregon, USA*, 2012

10. **"The BladeMistress Corpus: From Talk to Action in Virtual Worlds"**. A. Leuski, C. Eickhoff, J. Ganis and V. Lavrenko. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC), Istanbul, Turkey*, 2012

11. **"BooksOnline'11: 4th Workshop on Online Books, Complementary Social Media, and Crowdsourcing"**. G. Kazai, C. Eickhoff, and P. Brusilovsky. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM), Glasgow, UK*, 2011

12. **"The Where in the Tweet"**. W. Li, M. Larson, C. Eickhoff, A. P. de Vries and P. Serdyukov. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM), Glasgow, UK*, 2011

13. **"How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy"**. J. Vuurens, A. P. de Vries and C. Eickhoff. In *Proceedings of the SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval (CIR), Beijing, China*, 2011

14. **"Investigating Factors Influencing Crowdsourcing Tasks with High Imaginative Load"**. R. Vliegendhart, M. Larson, C. Kofler, C. Eickhoff and J. Pouwelse. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM), Hong Kong, China*, 2011

15. **"DMIR on Microblog Track 2011"**. W. Li, A. P. de Vries and C. Eickhoff. In *The 20th Text REtrieval Conference (TREC 2011) Notebook*, 2011

16. **"Managing the Quality of Large-Scale Crowdsourcing"**. J. Vuurens, A. P. de Vries and C. Eickhoff. In *The 20th Text REtrieval Conference (TREC 2011) Notebook*, 2011

17. **"Role Detection in Virtual Worlds"**. C. Eickhoff. The University of Edinburgh, School of Informatics, *M.Sc. Thesis, Edinburgh, Scotland*, 2009

18. **"Abbilden der Geschäftsregeln der VHV mit einem Business Rule Framework"**. C. Eickhoff. FHDW Hannover, *Diplomarbeit (M.Sc. Thesis), Hannover, Germany*, 2008

Curriculum Vitæ

Carsten Eickhoff

23-02-1985 Born in Twistringen, Germany.

Education

1997–2004 Grammar School
Gymnasium Sulingen

2005–2008 M.Sc. in Computer Science
FHDW Hannover

2008–2009 M.Sc. in Artificial Intelligence
The University of Edinburgh

2009 – 2014 Ph.D. in Computer Science
Technische Universiteit Delft
Thesis: *"Contextual Multidimensional Relevance Models"*
Promotor: Prof. dr. ir. A. P. de Vries

Awards

2011 Microsoft Bing Most Innovative Paper Award (ACM WSDM/CSDM)

2014 ECIR Best Reviewer Award

Bibliography

[1] Mikhail Ageev et al. “Find it if you can: a game for modeling different types of web search success using interaction data”. In: *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011)*. 2011.

[2] Eugene Agichtein et al. “Finding high-quality content in social media”. In: *Proceedings of the international conference on Web search and web data mining*. ACM. 2008, pp. 183–194.

[3] Denise E. Agosto and Sandra Hughes-Hassell. “Toward a model of the everyday life information needs of urban teenagers, part 1: Theoretical model”. In: *Journal of the American Society for Information Science and Technology* 57.10 (2006), pp. 1394–1403.

[4] Luis von Ahn and Laura Dabbish. “Designing games with a purpose”. In: *Communications of the ACM* 51.8 (2008), pp. 58–67.

[5] Luis von Ahn and Laura Dabbish. “Labeling images with a computer game”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2004, pp. 319–326.

[6] Luis von Ahn, Mihir Kedia, and Manuel Blum. “Verbosity: a game for collecting common-sense facts”. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 75–78.

[7] Luis von Ahn, Ruoran Liu, and Manuel Blum. “Peekaboom: a game for locating objects in images”. In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM. 2006, pp. 55–64.

[8] Luis von Ahn et al. “Improving image search with phetch”. In: *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. Vol. 4. IEEE. 2007, pp. IV–1209.

[9] Omar Alonso and Ricardo Baeza-Yates. “Design and implementation of relevance assessments using crowdsourcing”. In: *Advances in information retrieval*. Springer, 2011, pp. 153–164.

[10] Omar Alonso and Matthew Lease. “Crowdsourcing 101: putting the WSDM of crowds to work for you”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 1–2.

[11] Omar Alonso and Stefano Mizzaro. “Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment”. In: *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 2009, pp. 15–16.

[12] Giuseppe Amato and Umberto Straccia. “User profile modeling and applications to digital libraries”. In: *Research and Advanced Technology for Digital Libraries*. Springer, 1999, pp. 184–197.

[13] Vamshi Ambati, Stephan Vogel, and Jaime Carbonell. “Active learning and crowd-sourcing for machine translation”. In: *Language Resources and Evaluation (LREC)* 7 (2010), pp. 2169–2174.

[14] Morgan Ames and Mor Naaman. “Why we tag: motivations for annotation in mobile and online media”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2007, pp. 971–980.

[15] Einat Amitay et al. “Scaling IR-system evaluation using term relevance sets”. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2004, pp. 10–17.

[16] Giuseppe Amodeo, Giambattista Amati, and Giorgio Gambosi. “On relevance, time and query expansion”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 1973–1976.

[17] Daniel R. Anderson and Stephen R. Levin. “Young Children’s Attention to" Sesame Street"”. In: *Child Development* (1976), pp. 806–811.

[18] Theodore W. Anderson and Donald A. Darling. “A test of goodness of fit”. In: *Journal of the American Statistical Association* 49.268 (1954), pp. 765–769.

[19] Javed A. Aslam and Mark Montague. “Bayes optimal metasearch: a probabilistic model for combining the results of multiple retrieval systems (poster session)”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2000, pp. 379–381.

[20] Richard Atterer, Monika Wnuk, and Albrecht Schmidt. “Knowing the user’s every move: user activity tracking for website usability evaluation and implicit interaction”. In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 203–212.

[21] Anne Aula, Rehan M. Khan, and Zhiwei Guan. “How does search behavior change as search becomes more difficult?” In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 35–44.

[22] Andy Baio. *The Faces of Mechanical Turk*. http://waxy.org/2008/11/the_faces_of_mechanical_turk/. 2008.

[23] Krisztian Balog, Leif Azzopardi, and Maarten de Rijke. “Formal models for expert finding in enterprise corpora”. In: *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2006, pp. 43–50.

[24] Carol L. Barry. “User-defined relevance criteria: an exploratory study”. In: *Journal of the American Society for Information Science* 45.3 (1994), pp. 149–159.

[25] Marcia J. Bates. “Information search tactics”. In: *Journal of the American Society for information Science* 30.4 (1979), pp. 205–214.

[26] Nicholas J. Belkin. “Anomalous states of knowledge as a basis for information retrieval”. In: *Canadian journal of information science* 5.1 (1980), pp. 133–143.

[27] Nicholas J. Belkin and W. Bruce Croft. “Information filtering and information retrieval: two sides of the same coin?” In: *Communications of the ACM* 35.12 (1992).

[28] Michael Bendersky and W. Bruce Croft. “Analysis of long queries in a large scale search log”. In: *Proceedings of the 2009 workshop on Web Search Click Data*. ACM. 2009, pp. 8–14.

[29] Paul N. Bennett and Nam Nguyen. “Refined experts: improving classification in large taxonomies”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 11–18.

[30] Paul N. Bennett, Krysta Svore, and Susan T. Dumais. “Classification-enhanced ranking”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 111–120.

[31] Paul N. Bennett et al. “Modeling the impact of short-and long-term behavior on search personalization”. In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 185–194.

[32] Dania Bilal. “Children’s use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks”. In: *Journal of the American society for information science* 51.7 (2000), pp. 646–665.

[33] Dania Bilal and Joe Kirby. “Differences and similarities in information seeking: children and adults as Web users”. In: *Information processing & management* 38.5 (2002).

[34] Dania Bilal, Sonia Sarangthem, and Imad Bachir. “Toward a model of children’s information seeking behavior in using digital libraries”. In: *Proceedings of the second international symposium on Information interaction in context*. ACM. 2008, pp. 145–151.

[35] Gloria Bordogna and Gabriella Pasi. “A model for a soft fusion of information accesses on the web”. In: *Fuzzy sets and systems* 148.1 (2004), pp. 105–118.

[36] Christine L. Borgman et al. “Children’s searching behavior on browsing and keyword online catalogs: the Science Library Catalog project”. In: *Journal of the American Society for Information Science* 46.9 (1995).

[37] Pia Borlund. “The concept of relevance in IR”. In: *Journal of the American Society for information Science and Technology* 54.10 (2003), pp. 913–925.

[38] Pia Borlund and Peter Ingwersen. “Measures of relative relevance and ranked half-life: performance indicators for interactive IR”. In: *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1998, pp. 324–331.

[39] Jean-Philippe Bouchaud and Marc Potters. *Theory of financial risk and derivative pricing: from statistical physics to risk management*. Cambridge University Press, 2003.

[40] Adriana Budura et al. “Neighborhood-based tag prediction”. In: *The semantic web: research and applications*. Springer, 2009, pp. 608–622.

[41] Chris Burges et al. “Learning to rank using gradient descent”. In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 89–96.

[42] Jamie Callan and Maxine Eskenazi. “Combining lexical and grammatical features to improve readability measures for first and second language texts”. In: *Proceedings of NAACL HLT*. 2007, pp. 460–467.

[43] Jamie Callan et al. “Clueweb09 data set”. In: *Retrieved* 12.23 (2009), p. 2010.

[44] Sandra L. Calvert. “Children as consumers: Advertising and marketing”. In: *The Future of Children* 18.1 (2008), pp. 205–234.

[45] Robert Capra. “HCI browser: A tool for studying web search behavior”. In: *Proceedings of the American Society for Information Science and Technology* 47.1 (2010), pp. 1–2.

[46] Ben Carterette et al. “Million query track 2009 overview”. In: *Proceedings of TREC*. Vol. 9. 2009.

[47] Carlos Castillo et al. “Know your neighbors: Web spam detection using the web topology”. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, pp. 423–430.

[48] Arthur Charpentier, Jean-David Fermanian, and Olivier Scaillet. “The estimation of copulas: Theory and practice”. In: *Copulas: From theory to application in finance* (2007), pp. 35–60.

[49] Xu Cheng, Cameron Dale, and Jiangchuan Liu. “Understanding the characteristics of internet short video sharing: YouTube as a case study”. In: *arXiv preprint arXiv:0707.3670* (2007).

[50] Chun Wei Choo, Brian Detlor, and Don Turnbull. “Information seeking on the Web: An integrated model of browsing and searching”. In: *First Monday* 5.2 (2000).

[51] Charles L. Clarke. *Overview of the TREC 2009 Web track*. Tech. rep. Waterloo University, 2009.

[52] Kevyn Collins-Thompson and Jamie Callan. “A language modeling approach to predicting reading difficulty”. In: *Proceedings of HLT/NAACL*. Vol. 4. 2004.

[53] Kevyn Collins-Thompson et al. “Personalizing web search results by reading level”. In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 403–412.

[54] Célia da Costa Pereira, Mauro Dragoni, and Gabriella Pasi. “Multidimensional relevance: A new aggregation criterion”. In: *Advances in information retrieval*. Springer, 2009, pp. 264–275.

[55] Nick Craswell et al. “Relevance weighting for query independent evidence”. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 416–423.

[56] Fabio Crestani et al. ""Is this document relevant?... probably": a survey of probabilistic models in information retrieval". In: *ACM Computing Surveys (CSUR)* 30.4 (1998), pp. 528–552.

[57] Mihaly Csikszentmihalyi. *Flow: The psychology of optimal experience*. Harper Perennial, 1991.

[58] Sebastian Czajka and Sabine Mohr. "Internetnutzung in privaten Haushalten in Deutschland". In: *Ergebnisse der Erhebung* (2008).

[59] Honghua Kathy Dai et al. "Detecting online commercial intention (OCI)". In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 829–837.

[60] A. Philip Dawid and Allan M. Skene. "Maximum likelihood estimation of observer error-rates using the EM algorithm". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28.1 (1979), pp. 20–28. ISSN: 0035-9254.

[61] Munmun De Choudhury et al. "What makes conversations interesting?: themes, participants and consequences of conversations in online social media". In: *Proceedings of the 18th international conference on World wide web*. ACM. 2009, pp. 331–340.

[62] Pieter Dekker. "Children's roles in web search". In: *Master Thesis, Delft University of Technology* (2011).

[63] Allison Druin et al. "Children's roles using keyword search interfaces at home". In: *Proceedings of the 28th international conference on Human factors in computing systems*. ACM. 2010, pp. 413–422.

[64] Allison Druin et al. "How children search the internet with keyword interfaces". In: *Proceedings of the 8th International Conference on Interaction Design and Children*. ACM. 2009.

[65] Sergio Duarte Torres and Ingmar Weber. "What and how children search on the web". In: *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM. 2011, pp. 393–402.

[66] Sergio Raúl Duarte Torres. "Information retrieval for children: search behavior and solutions". In: (2014).

[67] Douglas Eck et al. "Automatic generation of social tags for music recommendation". In: *Advances in neural information processing systems* 20.20 (2007), pp. 1–8.

[68] Carsten Eickhoff, Pavel Serdyukov, and Arjen P. de Vries. "A combined topical/non-topical approach to identifying web sites for children". In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 505–514.

[69] Paul Embrechts, Filip Lindskog, and Alexander McNeil. "Modelling dependence with copulas and applications to risk management". In: *Handbook of heavy tailed distributions in finance* 8.1 (2003), pp. 329–384.

[70] Lijun Feng. "Automatic readability assessment for people with intellectual disabilities". In: *ACM SIGACCESS Accessibility and Computing* 93 (2009), pp. 84–91.

[71] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. "Cognitively motivated features for readability assessment". In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2009, pp. 229–237.

[72] Katja Filippova and Keith B. Hall. "Improved video categorization from text metadata and user comments". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 835–842.

[73] Edward Fox and Joseph Shaw. "Combination of multiple searches". In: *NIST SPECIAL PUBLICATION SP* (1994), pp. 243–243.

[74] Steve Fox et al. "Evaluating implicit measures to improve web search". In: *ACM Transactions on Information Systems (TOIS)* 23.2 (2005), pp. 147–168.

[75] Edward W. Frees and Emiliano A. Valdez. "Understanding relationships using copulas". In: *North American actuarial journal* 2.1 (1998), pp. 1–25.

[76] Thomas J. Froehlich. "Relevance Reconsidered - Towards an Agenda for the 21st Century: Introduction to Special Topic Issue on Relevance Research". In: *Journal of the American Society for Information Science* 45.3 (1994), pp. 124–134.

[77] Evgeniy Gabrilovich and Shaul Markovitch. "Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization". In: *Journal of Machine Learning Research* 8 (2007), pp. 2297–2345.

[78] Jianfeng Gao et al. "Smoothing clickthrough data for web search ranking". In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 355–362.

[79] Susan Gauch, Jason Chaffee, and Alexander Pretschner. "Ontology-based personalized search and browsing". In: *Web Intelligence and Agent Systems* 1.3 (2003), pp. 219–234.

[80] Shima Gerani, ChengXiang Zhai, and Fabio Crestani. "Score transformation in linear combination for multi-criteria relevance ranking". In: *Advances in Information Retrieval*. Springer, 2012, pp. 256–267.

[81] Richard Glassey, Tamara Polajnar, and Leif Azzopardi. "PuppyIR Unleashed: A Framework for Building Child-Oriented Information Services". In: *In Proc. of the 11th Dutch-Belgian IR Workshop*. 2011.

[82] Sharad Goel et al. "Anatomy of the long tail: ordinary people with extraordinary tastes". In: *Proceedings of the third ACM international conference on Web search and data mining*. ACM. 2010, pp. 201–210.

[83] Koraljka Golub and Anders Ardö. "Importance of HTML structural elements and metadata in automated subject classification". In: *Research and Advanced Technology for Digital Libraries*. Springer, 2005, pp. 368–378.

[84] Catherine Grady and Matthew Lease. “Crowdsourcing document relevance assessment with Mechanical Turk”. In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics. 2010, pp. 172–179.

[85] Karl Gyllstrom and Marie-Francine Moens. “Clash of the Typings”. In: *Advances in Information Retrieval*. Springer, 2011, pp. 80–91.

[86] Karl Gyllstrom and Marie-Francine Moens. “Wisdom of the ages: toward delivering the children's web with the link-based agerank algorithm”. In: *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM. 2010, pp. 159–168.

[87] Mark Hall et al. “The WEKA data mining software: an update”. In: *ACM SIGKDD Explorations Newsletter* 11.1 (2009).

[88] Donna Harman. “Overview of the first TREC conference”. In: *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1993, pp. 36–47.

[89] Christopher G. Harris. “You're Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks”. In: *Crowdsourcing for Search and Data Mining (CSDM 2011)* (2011), p. 15.

[90] Stephen P. Harter. “Psychological relevance and information science”. In: *Journal of the American Society for Information Science* 43.9 (1992), pp. 602–615.

[91] Taher H. Haveliwala. “Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search”. In: *Knowledge and Data Engineering, IEEE Transactions on* 15.4 (2003), pp. 784–796.

[92] Paul Heymann, Daniel Ramage, and Hector Garcia-Molina. “Social tag prediction”. In: *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2008, pp. 531–538.

[93] Djoerd Hiemstra. “A probabilistic justification for using tf× idf term weighting in information retrieval”. In: *International Journal on Digital Libraries* 3.2 (2000), pp. 131–139.

[94] Matthias Hirth, Tobias Hoßfeld, and Phuoc Tran-Gia. *Cheat-Detection Mechanisms for Crowdsourcing*. Tech. rep. University of Würzburg, 2010.

[95] W. Höffding. “Scale-invariant correlation theory”. In: *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin* 5.3 (1940), pp. 181–233.

[96] Christoph Hölscher and Gerhard Strube. “Web search behavior of Internet experts and newbies”. In: *Computer networks* 33.1 (2000), pp. 337–346.

[97] Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. “Data quality from crowdsourcing: a study of annotation selection criteria”. In: *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*. Association for Computational Linguistics. 2009, pp. 27–35.

[98] Meishan Hu, Aixin Sun, and Ee-Peng Lim. "Comments-oriented blog summarization by sentence extraction". In: *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM. 2007, pp. 901–904.

[99] Xuanjing Huang and W. Bruce Croft. "A unified relevance model for opinion retrieval". In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 947–956.

[100] Hilary Hutchinson et al. "How do I find blue books about dogs? The errors and frustrations of young digital library users". In: *Proceedings of HCII 2005* (2005).

[101] Peter Ingwersen and Kalervo Järvelin. *The turn: Integration of information seeking and retrieval in context*. Vol. 18. Kluwer Academic Pub, 2005.

[102] Panagiotis G. Ipeirotis. "Analyzing the amazon mechanical turk marketplace". In: *XRDS: Crossroads, The ACM Magazine for Students* 17.2 (2010), pp. 16–21.

[103] Panagiotis G. Ipeirotis. *Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes*. http://behind-the-enemy-lines.blogspot.com/2010/10/be-top-mechanical-turk-worker-you-need.html. 2010.

[104] Hannah Jochmann-Mannak. *Websites for Children: Search strategies and interface design*. Twente University, 2014.

[105] Rosie Jones and Kristina Lisa Klinkner. "Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs". In: *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM. 2008, pp. 699–708.

[106] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. "More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk". In: *Proceedings of the Seventeenth Americas Conference on Information Systems*. 2011, pp. 1–11.

[107] Gabriella Kazai. "In search of quality in crowdsourcing for search engine evaluation". In: *Advances in information retrieval*. Springer, 2011, pp. 165–176.

[108] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. "Worker Types and Personality Traits in Crowdsourcing Relevance Labels". In: *Proceedings of 20th International Conference on Information and Knowledge Management (CIKM)*. ACM. 2011.

[109] Shashank Khanna et al. "Evaluating and improving the usability of Mechanical Turk for low-income workers in India". In: *Proceedings of the First ACM Symposium on Computing for Development*. ACM. 2010, p. 12.

[110] Jin Young Kim et al. "Characterizing web content, user interests, and search behavior by reading level and topic". In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 213–222.

[111] Aniket Kittur, Ed H. Chi, and Bongwon Suh. "Crowdsourcing user studies with Mechanical Turk". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2008, pp. 453–456.

[112] George R. Klare. “The measurement of readability: useful information for communicators”. In: *ACM Journal of Computer Documentation (JCD)* 24.3 (2000), pp. 107–121.

[113] Jon Kleinberg. “Bursty and hierarchical structure in streams”. In: *Data Mining and Knowledge Discovery* 7.4 (2003), pp. 373–397.

[114] Chih-Hung Ko et al. “Gender differences and related factors affecting online gaming addiction among Taiwanese adolescents”. In: *The Journal of nervous and mental disease* 193.4 (2005), pp. 273–277.

[115] Pranam Kolari, Tim Finin, and Anupam Joshi. “SVMs for the blogosphere: Blog identification and splog detection”. In: *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*. Vol. 4. 2006, p. 1.

[116] Alexander Kotov et al. “Modeling and analysis of cross-session search tasks”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 5–14.

[117] Wessel Kraaij, Thijs Westerveld, and Djoerd Hiemstra. “The importance of prior probabilities for entry page search”. In: *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2002, pp. 27–34.

[118] Carol C. Kuhlthau. “Inside the search process: Information seeking from the user’s perspective”. In: *Journal of the American Society for information Science* 42.5 (1991).

[119] Giridhar Kumaran and Vitor R. Carvalho. “Reducing long queries using query quality predictors”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 564–571.

[120] Joseph Lampel and Ajay Bhalla. “The role of status seeking in online communities: Giving the gift of experience”. In: *Journal of Computer-Mediated Communication* 12.2 (2007), pp. 434–455.

[121] Patricia G. Lange. “Commenting on comments: Investigating responses to antagonism on YouTube”. In: *Annual Conference of the Society for Applied Anthropology. Retrieved August*. Vol. 29. 2007, p. 2007.

[122] Patricia G. Lange. “Publicly private and privately public: Social networking on YouTube”. In: *Journal of Computer-Mediated Communication* 13.1 (2007), pp. 361–380.

[123] Andrew Large, Jamshid Beheshti, and Alain Breuleux. “Information seeking in a multimedia environment by primary school students”. In: *Library & Information Science Research* 20.4 (1998), pp. 343–376.

[124] Andrew Large, Jamshid Beheshti, and Tarjin Rahman. “Design criteria for children’s Web portals: The users speak out”. In: *Journal of the American Society for Information Science and Technology* 53.2 (2002), pp. 79–94.

[125] Martha Larson et al. “Automatic tagging and geotagging in video collections and communities”. In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM. 2011, p. 51.

[126] Victor Lavrenko and W. Bruce Croft. “Relevance based language models”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 120–127.

[127] Matthew Lease and Gabriella Kazai. “Overview of the TREC 2011 crowdsourcing track (conference notebook)”. In: *Text Retrieval Conference Notebook*. 2011.

[128] Dutch Legislation. *Wet bescherming persoonsgegevens*. http://wetten.overheid.nl/BWBR0011468. 2000.

[129] Dutch Legislation. *Wet medisch-wetenschappelijk onderzoek met mensen*. http://wetten.overheid.nl/BWBR0009408. 1998.

[130] European Legislation. *European directive on privacy and electronic communications*. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:201:0037:0047:EN:PDF. 2002.

[131] Gregory W. Lesher and Christian Sanelli. “A web-based system for autonomous text corpus generation”. In: *Proceedings of ISSAAC* (2000).

[132] Lin Li et al. “Dynamic adaptation strategies for long-term and short-term user profile to personalize search”. In: *Advances in Data and Web Management* (2007), pp. 228–240.

[133] Wei-Hao Lin and Alexander Hauptmann. “News video classification using SVM-based multimodal classifiers and combination strategies”. In: *Proceedings of the tenth ACM international conference on Multimedia*. ACM. 2002, pp. 323–326.

[134] Greg Little et al. “Turkit: Tools for iterative tasks on mechanical turk”. In: *Proceedings of the ACM SIGKDD workshop on human computation*. ACM. 2009, pp. 29–30.

[135] Bing Liu, Minqing Hu, and Junsheng Cheng. “Opinion observer: analyzing and comparing opinions on the Web”. In: *Proceedings of the 14th international conference on World Wide Web*. ACM. 2005, pp. 342–351.

[136] Tie-Yan Liu. “Learning to rank for information retrieval”. In: *Foundations and Trends in Information Retrieval* 3.3 (2009), pp. 225–331.

[137] Tie-Yan Liu et al. “Support vector machines classification with a very large-scale taxonomy”. In: *ACM SIGKDD Explorations Newsletter* 7.1 (2005), pp. 36–43.

[138] Zhu Liu, Jincheng Huang, and Yao Wang. “Classification TV programs based on audio information using hidden Markov model”. In: *Multimedia Signal Processing, 1998 IEEE Second Workshop on*. IEEE. 1998, pp. 27–32.

[139] Sonia Livingstone and Leslie Haddon. “EU Kids Online: Final Report”. In: *LSE, London: EU Kinds Online (EC Safer Internet Plus Programme Deliverable D6. 5)* (2009).

[140] Cheng Lu, Mark S. Drew, and James Au. “Classification of summarized videos using hidden Markov models on compressed chromaticity signatures”. In: *Proceedings of the ninth ACM international conference on Multimedia*. ACM. 2001, pp. 479–482.

[141] Wei Lu, Stephen Robertson, and Andrew MacFarlane. “Field-weighted XML retrieval based on BM25”. In: *Advances in XML Information Retrieval and Evaluation*. Springer, 2006, pp. 161–171.

[142] Hao Ma et al. “Improving search engines using human computation games”. In: *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM. 2009, pp. 275–284.

[143] Zhongming Ma, Gautam Pant, and Olivia R. Liu Sheng. “Interest-based personalized search”. In: *ACM Transactions on Information Systems (TOIS)* 25.1 (2007), p. 5.

[144] Craig Macdonald et al. “Blog track research at TREC”. In: *ACM SIGIR Forum*. Vol. 44. ACM. 2010, pp. 58–75.

[145] Raghavan Manmatha, Toni M. Rath, and Fangfang Feng. “Modeling score distributions for combining the outputs of search engines”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 267–275.

[146] Gary M. Marchionini. “Exploratory search: from finding to understanding”. In: *Communications of the ACM* 49.4 (2006).

[147] Gary M. Marchionini. *Information seeking in electronic environments*. 9. Cambridge Univ Pr, 1997.

[148] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. “Building a large annotated corpus of English: The Penn Treebank”. In: *Computational linguistics* 19.2 (1993), pp. 313–330.

[149] Yutaka Matsuo and Mitsuru Ishizuka. “Keyword extraction from a single document using word co-occurrence statistical information”. In: *International Journal on Artificial Intelligence Tools* 13.01 (2004), pp. 157–169.

[150] Nicolaas Matthijs and Filip Radlinski. “Personalizing web search using long term browsing history”. In: *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM. 2011, pp. 25–34.

[151] Jane McGonigal. *Reality is broken: Why games make us better and how they can change the world*. Penguin books, 2011.

[152] G. Harry McLaughlin. “SMOG grading: A new readability formula”. In: *Journal of reading* 12.8 (1969), pp. 639–646.

[153] Gilad Mishne and Natalie Glance. “Leave a reply: An analysis of weblog comments”. In: *Third annual workshop on the Weblogging ecosystem*. 2006.

[154] Stefano Mizzaro. “Relevance: The whole history”. In: *Journal of the American Society for Information Science* 48.9 (1997), pp. 810–832.

[155] Mark Montague and Javed A. Aslam. “Condorcet fusion for improved retrieval”. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM. 2002, pp. 538–548.

[156] Mark Montague and Javed A. Aslam. “Relevance score normalization for metasearch”. In: *Proceedings of the tenth international conference on Information and knowledge management*. ACM. 2001, pp. 427–433.

[157] Andrew P. Moore, Robert J. Ellison, and Richard C. Linger. *Attack modeling for information security and survivability*. Tech. rep. Carnegie-Mellon University, Software Engineering Institute, 2001.

[158] Penelope A. Moore and Alison St George. “Children as Information Seekers: The Cognitive Demands of Books and Library Systems”. In: *School Library Media Quarterly* 19.3 (1991), pp. 161–68.

[159] Shiva Naidu. “Evaluating the usability of educational websites for children”. In: *Usability News* 7.2 (2005).

[160] Ramesh Nallapati. “Discriminative models for information retrieval”. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2004, pp. 64–71.

[161] Nikolaos Nanas, Victoria Uren, and Anne De Roeck. “Building and applying a concept hierarchy representation of a user profile”. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM. 2003, pp. 198–204.

[162] Jakob Nielsen. “Kids’ corner: Website usability for children”. In: *Jakob Nielsen’s Alertbox* (2002).

[163] Ofcom. *UK children’s media literacy: Research Document*. http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf. 2010.

[164] Andrei Oghina et al. “Predicting imdb movie ratings using social media”. In: *Advances in Information Retrieval*. Springer, 2012, pp. 503–507.

[165] Paul Ogilvie and Jamie Callan. “Combining document representations for known-item search”. In: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM. 2003, pp. 143–150.

[166] Arno Onken et al. “Analyzing short-term noise dependencies of spike-counts in macaque prefrontal cortex using copulas and the flashlight transformation”. In: *PLoS computational biology* 5.11 (2009), e1000577.

[167] Bo Pang and Lillian Lee. “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”. In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2004, p. 271.

[168] Taemin Kim Park. “The nature of relevance in information retrieval: An empirical study”. In: *The library quarterly* (1993), pp. 318–351.

[169] David Pennock. *The Wisdom of the Probability Sports Crowd*. http://blog.oddhead.com/2007/01/04/the-wisdom-of-the-probabilitysports-crowd/. 2007.

[170] Charles P. Pfleeger and Shari L. Pfleeger. *Security in computing*. Prentice Hall PTR, 2006.

[171] Peter Pirolli and Stuart Card. "Information foraging in information access environments". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM Press/Addison-Wesley Publishing Co. 1995, pp. 51–58.

[172] Pitkow, James and Schütze, Hinrich. "Personalized search". In: *Communications of the ACM* 9.45 (2002), pp. 50–55.

[173] Maja Pusara and Carla E. Brodley. "User re-authentication via mouse movements". In: *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*. ACM. 2004.

[174] Xiaoguang Qi and Brian D. Davison. "Web page classification: Features and algorithms". In: *ACM Computing Surveys (CSUR)* 41.2 (2009), p. 12.

[175] Zheng Qifu et al. "Support Vector Machine Based on Universal Kernel Function and Its Application in Quantitative Structure-Toxicity Relationship Model". In: *Information Technology and Applications, 2009. IFITA'09. International Forum on*. Vol. 3. IEEE. 2009, pp. 708–711.

[176] Filip Radlinski and Thorsten Joachims. "Query chains: learning to rank from implicit feedback". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 239–248.

[177] Benjamin Renard and Michel Lang. "Use of a Gaussian copula for multivariate extreme value analysis: Some case studies in hydrology". In: *Advances in Water Resources* 30.4 (2007), pp. 897–912.

[178] Clare Richards. *Crowdflower: The world is at work. Right now.* http://crowdflower.com/. 2014.

[179] Clare Richards. *Teach the world to twitch: An interview with Marc Prensky, CEO and founder Games2train. com. Futurelab.* 2003.

[180] Ellen Riloff. "Automatically generating extraction patterns from untagged text". In: *Proceedings of the national conference on artificial intelligence*. 1996, pp. 1044–1049.

[181] Stephen E. Robertson. "The probability ranking principle in IR". In: *Journal of documentation* 33.4 (1977), pp. 294–304.

[182] Stephen E. Robertson, Hugo Zaragoza, and Michael Taylor. "Simple BM25 extension to multiple weighted fields". In: *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. ACM. 2004, pp. 42–49.

[183] Joel Ross et al. "Who are the crowdworkers?: shifting demographics in mechanical turk". In: *Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems*. ACM. 2010, pp. 2863–2872.

[184] Tefko Saracevic. "Relevance: A review of and a framework for the thinking on the notion in information science". In: *Journal of the American Society for Information Science* 26.6 (1975), pp. 321–343.

[185] Tefko Saracevic. “Relevance reconsidered”. In: *Proceedings of the second conference on conceptions of library and information science (CoLIS 2)*. 1996, pp. 201–218.

[186] Tefko Saracevic and Paul Kantor. “A study of information seeking and retrieving. III. Searchers, searches, and overlap”. In: *Journal of the American Society for information Science* 39.3 (1988), pp. 197–216.

[187] Reijo Savolainen and Jarkko Kari. “User-defined relevance criteria in web searching”. In: *Journal of Documentation* 62.6 (2006), pp. 685–707.

[188] John Schacter, Gregory KWK Chung, and Aimée Dorr. “Children’s Internet searching on complex problems: performance and process analyses”. In: *Journal of the American society for Information Science* 49.9 (1998), pp. 840–849.

[189] Linda Schamber and Judy Bateman. “User Criteria in Relevance Evaluation: Toward Development of a Measurement Scale.” In: *Proceedings of the ASIS Annual Meeting*. Vol. 33. ERIC. 1996, pp. 218–25.

[190] Linda Schamber, Michael B. Eisenberg, and Michael S. Nilan. “A re-examination of relevance: toward a dynamic, situational definition”. In: *Information processing & management* 26.6 (1990), pp. 755–776.

[191] Thorsten Schmidt. “Coping with copulas”. In: *Chapter forthcoming in Risk Books: Copulasfrom theory to applications in finance* (2006).

[192] Christian Schoelzel and Petra Friederichs. “Multivariate non-normally distributed random variables in climate research–introduction to the copula approach”. In: *Nonlin. Processes Geophys.* 15.5 (2008), pp. 761–772.

[193] Falk Scholer, Andrew Turpin, and Mark Sanderson. “Quantifying test collection quality based on the consistency of relevance judgements”. In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 2011, pp. 1063–1072.

[194] Sarah E. Schwarm and Mari Ostendorf. “Reading level assessment using support vector machines and statistical language models”. In: *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics. 2005, pp. 523–530.

[195] Andrew K. Shenton and Pat Dixon. “Models of young people’s information seeking”. In: *Journal of Librarianship and Information Science* 35.1 (2003).

[196] Ben Shneiderman. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1997. ISBN: 0201694972.

[197] Ben Shneiderman and Catherine Plaisant. *Designing the user interface 4th edition*. 2005.

[198] Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. “Automatic video tagging using content redundancy”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 395–402.

[199] Stefan Siersdorfer et al. “How useful are your comments?: analyzing and predicting youtube comments and comment ratings”. In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 891–900.

[200] Ilmério Silva et al. “Link-based and content-based evidential information in a belief network model”. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2000, pp. 96–103.

[201] Abe Sklar. “Fonctions de répartition à n dimensions et leurs marges”. In: *Publ. Inst. Statist. Univ. Paris* 8.1 (1959), p. 11.

[202] Frans Van der Sluis et al. “Visual exploration of health information for children”. In: *Advances in Information Retrieval*. Springer, 2011, pp. 788–792.

[203] Rion Snow et al. “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks”. In: *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics. 2008, pp. 254–263.

[204] Ian Soboroff, Charles Nicholas, and Patrick Cahan. “Ranking retrieval systems without relevance judgments”. In: *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2001, pp. 66–73.

[205] Mohammed Soleymani and Martha Larson. “Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus”. In: *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*. 2010, pp. 4–8.

[206] Parikshit Sondhi, Vinod Vydiswaran, and ChengXiang Zhai. “Reliability Prediction of Webpages in the Medical Domain”. In: *Advances in Information Retrieval* (2012).

[207] Alexander Sorokin and David Forsyth. “Utility data annotation with amazon mechanical turk”. In: *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW’08. IEEE Computer Society Conference on*. IEEE. 2008, pp. 1–8.

[208] Micro Speretta and Susan Gauch. “Personalized search based on user search histories”. In: *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*. IEEE. 2005, pp. 622–628.

[209] Aideen J. Stronge, Wendy A. Rogers, and Arthur D. Fisk. “Web-based information search and retrieval: Effects of strategy use and age on search success”. In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 48.3 (2006).

[210] Kazunari Sugiyama, Kenji Hatano, and Masatoshi Yoshikawa. “Adaptive web search based on user profile constructed without any effort from users”. In: *Proceedings of the 13th international conference on World Wide Web*. ACM. 2004, pp. 675–684.

[211] Vakkalanka Suresh et al. “Content-based video classification using support vector machines”. In: *Neural Information Processing*. Springer. 2004, pp. 726–731.

[212] James Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[213] Bin Tan, Xuehua Shen, and ChengXiang Zhai. "Mining long-term search history to improve search accuracy". In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2006, pp. 718–723.

[214] Chenhao Tan, Evgeniy Gabrilovich, and Bo Pang. "To each his own: personalized content selection based on text comprehensibility". In: *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM. 2012, pp. 233–242.

[215] Jaime Teevan, Susan T Dumais, and Eric Horvitz. "Personalizing search via automated analysis of interests and activities". In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2005, pp. 449–456.

[216] Jaime Teevan, Susan T. Dumais, and Eric Horvitz. "Potential for personalization". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 17.1 (2010), p. 4.

[217] Andrew Thatcher. "Web search strategies: The influence of Web experience and task type". In: *Information Processing & Management* 44.3 (2008), pp. 1308–1329.

[218] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. "How users assess web pages for information seeking". In: *Journal of the American society for Information Science and Technology* 56.4 (2005), pp. 327–344.

[219] Takashi Tomokiyo and Matthew Hurst. "A language model approach to keyphrase extraction". In: *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*. Association for Computational Linguistics. 2003, pp. 33–40.

[220] Theodora Tsikrika and Mounia Lalmas. "Combining evidence for relevance criteria: a framework and experiments in web retrieval". In: *Advances in Information Retrieval*. Springer, 2007, pp. 481–493.

[221] Howard Turtle and W. Bruce Croft. "Evaluation of an inference network-based retrieval model". In: *ACM Transactions on Information Systems (TOIS)* 9.3 (1991), pp. 187–222.

[222] Julián Urbano et al. "The University Carlos III of Madrid at TREC 2011 Crowdsourcing Track". In: *Text REtrieval Conference*. 2011.

[223] David Vallet and Pablo Castells. "Personalized diversification of search results". In: *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2012, pp. 841–850.

[224] Christopher C. Vogt and Garrison W. Cottrell. "Fusion via a linear combination of scores". In: *Information Retrieval* 1.3 (1999), pp. 151–173.

[225] Ellen M. Voorhees. "The philosophy of information retrieval evaluation". In: *Evaluation of cross-language information retrieval systems*. Springer. 2002, pp. 355–370.

[226] Ellen M. Voorhees and Donna K. Harman. *TREC: Experiment and evaluation in information retrieval*. Vol. 63. MIT press Cambridge, 2005.

[227] Raven M. C. Wallace et al. “Science on the Web: Students online in a sixth-grade classroom”. In: *The Journal of the Learning Sciences* 9.1 (2000), pp. 75–104.

[228] Jing Wang, Siamak Faridani, and Panagiotis G. Ipeirotis. “Estimating the Completion Time of Crowdsourced Tasks Using Survival Analysis Models”. In: *Crowdsourcing for Search and Data Mining (CSDM 2011)* (2011), p. 31.

[229] Peiling Wang and Marilyn Domas White. “A cognitive model of document use during a research project. Study II. Decisions at the reading and citing stages”. In: *Journal of the American Society for Information Science* 50.2 (1999), pp. 98–114.

[230] Yao Wang, Zhu Liu, and Jin-Cheng Huang. “Multimedia content analysis-using both audio and visual clues”. In: *Signal Processing Magazine, IEEE* 17.6 (2000), pp. 12–36.

[231] Ellen A. Wartella, Elizabeth A. Vandewater, and Victoria J. Rideout. “Introduction: electronic media use in the lives of infants, toddlers, and preschoolers”. In: *American Behavioral Scientist* 48.5 (2005), p. 501.

[232] Christian Wartena, Rogier Brussee, and Wout Slakhorst. “Keyword extraction using word co-occurrence”. In: *Database and Expert Systems Applications (DEXA), 2010 Workshop on*. IEEE. 2010, pp. 54–58.

[233] Peter Welinder et al. “The multidimensional wisdom of crowds”. In: *Advances in Neural Information Processing Systems* 23 (2010), pp. 2424–2432.

[234] Ryen W. White, Peter Bailey, and Liwei Chen. “Predicting user interests from contextual information”. In: *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2009, pp. 363–370.

[235] Ryen W. White, Susan T. Dumais, and Jaime Teevan. “Characterizing the influence of domain expertise on web search behavior”. In: *Proceedings of the Second ACM International Conference on Web Search and Data Mining*. ACM. 2009, pp. 132–141.

[236] Ryen W. White and Dan Morris. “Investigating the querying and browsing behavior of advanced search engine users”. In: *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 2007, pp. 255–262.

[237] Lei Wu et al. “Distance metric learning from uncertain side information with application to automated photo tagging”. In: *Proceedings of the 17th ACM international conference on Multimedia*. ACM. 2009, pp. 135–144.

[238] Qiang Wu et al. “Ranking, boosting, and model adaptation”. In: *Technical Report, MSR-TR-2008-109* (2008).

[239] Shengli Wu and Fabio Crestani. “Data fusion with estimated weights”. In: *Proceedings of the eleventh international conference on Information and knowledge management*. ACM. 2002, pp. 648–651.

[240] Yunjie Calvin Xu and Zhiwei Chen. “Relevance judgment: What do information users consider beyond topicality?” In: *Journal of the American Society for Information Science and Technology* 57.7 (2006), pp. 961–973.

[241] Jun Yan. “Enjoy the joy of copulas: with a package copula”. In: *Journal of Statistical Software* 21.4 (2007), pp. 1–21.

[242] Ka-Ping Yee et al. “Faceted metadata for image search and browsing”. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2003, pp. 401–408.

[243] Wai Gen Yee et al. “Are web user comments useful for search”. In: *Proc. LSDS-IR* (2009), pp. 63–70.

[244] Yusrita M. Yusoff, Ian Ruthven, and Monica Landoni. “The fun semantic differential scales”. In: *Proceedings of the 10th International Conference on Interaction Design and Children*. ACM. 2011.

SIKS Dissertation Series

1998-01 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects

1998-02 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information

1998-03 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective

1998-04 Dennis Breuker (UM)
Memory versus Search in Games

1998-05 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting

1999-01 Mark Sloof (VU)
Physiology of Quality Change Modelling; Automated modelling of Quality Change of Agricultural Products

1999-02 Rob Potharst (EUR)
Classification using decision trees and neural nets

1999-03 Don Beal (UM)
The Nature of Minimax Search

1999-04 Jacques Penders (UM)
The practical Art of Moving Physical Objects

1999-05 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems

1999-06 Niek J.E. Wijngaards (VU)
Re-design of compositional systems

1999-07 David Spelt (UT)
Verification support for object database design

1999-08 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation

2000-01 Frank Niessink (VU)
Perspectives on Improving Software Maintenance

2000-02 Koen Holtman (TUE)
Prototyping of CMS Storage Management

2000-03 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie; een procesbenadering en actorperspectief

2000-04 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design

2000-05 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval

2000-06 Rogier van Eijk (UU)
Programming Languages for Agent Communication

2000-07 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management

2000-08 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks

2000-09 Florian Waas (CWI)
Principles of Probabilistic Query Optimization

2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations, Algorithms and Architecture

2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management

2001-01 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks

2001-02 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models

2001-03 Maarten van Someren (UvA)
Learning as problem solving

2001-04 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with Instance-Based Boundary Sets

2001-05 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style

2001-06 Martijn van Welie (VU)
Task-based User Interface Design

2001-07 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization

2001-08 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics

2001-09 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models, Views of Packages as Classes

2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice BRAHMS: a multiagent modeling and simulation language for work practice analysis and design

2001-11 Tom M. van Engers (VUA)
Knowledge Management: The Role of Mental Models in Business Systems Design

2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis

2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections

2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval

2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining

2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments inhabited by Privacy-concerned Agents

2002-06 Laurens Mommers (UL)
Applied legal epistemology; Building a knowledge-based ontology of the legal domain

2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications

2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative E-Commerce Ideas

2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy Systems

2002-10 Brian Sheppard (UM)
Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity

2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques

2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes

2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise

2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning

2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects

2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies

2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents

2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play

2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium

2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining

2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning

2004-17 Mark Winands (UM)
Informed Search in Complex Games

2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models

2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval

2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams

2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications

2005-02 Erik van der Werf (UM))
AI techniques for the game of Go

2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language

2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data

2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing

2005-06 Pieter Spronck (UM)
Adaptive Game AI

2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems

2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications

2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages

2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments

2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search

2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry

2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen

2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes

2005-16 Joris Graaumans (UU)
Usability of XML Query Languages

2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components

2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks

2005-19 Michel van Dartel (UM)
Situated Representation

2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives

2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by Exploiting Application Semantics

2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting

2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations

2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems

2006-04 Marta Sabou (VU)
Building Web Service Ontologies

2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines

2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools for Graphical Service Modeling

2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering

2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web

2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion

2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems

2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types

2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts

2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents

2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change

2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain

2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks

2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device

2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing

2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach

2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining

2006-21 Bas van Gils (RUN)
Aptness on the Web

2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation

2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources

2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC

2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval

2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories

2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval

2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures

2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach

2007-03 Peter Mika (VU)
Social Networks and the Semantic Web

2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach

2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy: a Legislative Framework for Agent-enabled Surveillance

2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs

2007-07 Natasa Jovanovic' (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings

2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations

2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation

2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols

2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a General-Purpose Adaptive Hypermedia System

2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to Dynamic Decision-Making under Uncertainty

2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology

2007-14 Niek Bergboer (UM)
Context-Based Image Analysis

2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model

2007-16 Davide Grossi (UU)
Designing Invisible Handcuffs. Formal investigations in Institutions and Organizations for Multi-agent Systems

2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice

2007-18 Bart Orriens (UvT)
On the development an management of adaptive business collaborations

2007-19 David Levy (UM)
Intimate relationships with artificial partners

2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network

2007-21 Karianne Vermaas (UU)
Fast diffusion and broadening use: A research on residential adoption and usage of broadband internet in the Netherlands between 2001 and 2005

2007-22 Zlatko Zlatev (UT)
Goal-oriented design of value and process models from patterns

2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems

2007-24 Georgina Ramírez Camps (CWI)
Structural Features in XML Retrieval

2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement

2008-01 Katalin Boer-Sorbán (EUR)
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach

2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations

2008-03 Vera Hollink (UVA)
Optimizing hierarchical menus: a usage-based approach

2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards unattended integration

2008-05 Bela Mutschler (UT)
Modeling and simulating causal dependencies on process-aware information systems from a cost perspective

2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective

2008-07 Peter van Rosmalen (OU)
Supporting the tutor in the design and support of adaptive e-learning

2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference

2008-09 Christof van Nimwegen (UU)
The paradox of the guided user: assistance can be counter-effective

2008-10 Wauter Bosma (UT)
Discourse oriented summarization

2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach

2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation

2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks

2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort

2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for the Markov Decision Process Framework in First-Order Domains

2008-16 Henriette van Vugt (VU)
Embodied agents from a user's perspective

2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises

2008-18 Guido de Croon (UM)
Adaptive Active Vision

2008-19 Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search

2008-20 Rex Arendsen (UVA)
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie van elektronisch berichtenverkeer met de overheid op de administratieve lasten van bedrijven

2008-21 Krisztian Balog (UVA)
People Search in the Enterprise

2008-22 Henk Koning (UU)
Communication of IT-Architecture

2008-23 Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia

2008-24 Zharko Aleksovski (VU)
Using background knowledge in ontology matching

2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair using Spender-signed Currency

2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled

2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design

2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks

2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans

2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content

2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval

2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes

2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; theoretical and practical issues

2008-34 Jeroen de Knijf (UU)
Studies in Frequent Tree Mining

2008-35 Ben Torben Nielsen (UvT)
Dendritic morphologies: function shapes structure

2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models

2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques

2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT

2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering

2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality

2009-06 Muhammad Subianto (UU)
Understanding Classification

2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion

2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments

2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems

2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications

2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law the Semantic Web

2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
perating Guidelines for Services

2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems

2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)

2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense

2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess

2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data

2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System

2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets

2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making

2009-21 Stijn Vanderlooy (UM)
Ranking and Reliable Classification

2009-22 Pavel Serdyukov (UT)
Search For Expertise: Going beyond direct evidence

2009-23 Peter Hofgesang (VU)
Modelling Web Usage in a Changing Environment

2009-24 Annerieke Heuvelink (VUA)
Cognitive Models for Training Simulations

2009-25 Alex van Ballegooij (CWI)
"RAM: Array Database Management through Relational Mapping"

2009-26 Fernando Koch (UU)
An Agent-Based Model for the Development of Intelligent Mobile Services

2009-27 Christian Glahn (OU)
Contextual Support of social Engagement and Reflection on the Web

2009-28 Sander Evers (UT)
Sensor Data Management with Probabilistic Models

2009-29 Stanislav Pokraev (UT)
Model-Driven Semantic Integration of Service-Oriented Applications

2009-30 Marcin Zukowski (CWI)
Balancing vectorized query execution with bandwidth-optimized storage

2009-31 Sofiya Katrenko (UVA)
A Closer Look at Learning Relations from Text

2009-32 Rik Farenhorst (VU) and Remco de Boer (VU)
Architectural Knowledge Management: Supporting Architects and Auditors

2009-33 Khiet Truong (UT)
How Does Real Affect Affect Affect Recognition In Speech?

2009-34 Inge van de Weerd (UU)
Advancing in Software Product Management: An Incremental Method Engineering Approach

2009-35 Wouter Koelewijn (UL)
Privacy en Politiegegevens; Over geautomatiseerde normatieve informatie-uitwisseling

2009-36 Marco Kalz (OUN)
Placement Support for Learners in Learning Networks

2009-37 Hendrik Drachsler (OUN)
Navigation Support for Learners in Informal Learning Networks

2009-38 Riina Vuorikari (OU)
Tags and self-organisation: a metadata ecology for learning resources in a multilingual context

2009-39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin)
Service Substitution – A Behavioral Approach Based on Petri Nets

2009-40 Stephan Raaijmakers (UvT)
Multinomial Language Learning: Investigations into the Geometry of Language

2009-41 Igor Berezhnyy (UvT)
Digital Analysis of Paintings

2009-42 Toine Bogers
Recommender Systems for Social Bookmarking

2009-43 Virginia Nunes Leal Franqueira (UT)
Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients

2009-44 Roberto Santana Tapia (UT)
Assessing Business-IT Alignment in Networked Organizations

2009-45 Jilles Vreeken (UU)
Making Pattern Mining Useful

2009-46 Loredana Afanasiev (UvA)
Querying XML: Benchmarks and Recursion

2010-01 Matthijs van Leeuwen (UU)
Patterns that Matter

2010-02 Ingo Wassink (UT)
Work flows in Life Science

2010-03 Joost Geurts (CWI)
A Document Engineering Model and Processing Framework for Multimedia documents

2010-04 Olga Kulyk (UT)
Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments

2010-05 Claudia Hauff (UT)
Predicting the Effectiveness of Queries and Retrieval Systems

2010-06 Sander Bakkes (UvT)
Rapid Adaptation of Video Game AI

2010-07 Wim Fikkert (UT)
Gesture interaction at a Distance

2010-08 Krzysztof Siewicz (UL)
Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments

2010-09 Hugo Kielman (UL)
A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging

2010-10 Rebecca Ong (UL)
Mobile Communication and Protection of Children

2010-11 Adriaan Ter Mors (TUD)
The world according to MARP: Multi-Agent Route Planning

2010-12 Susan van den Braak (UU)
Sensemaking software for crime analysis

2010-13 Gianluigi Folino (RUN)
High Performance Data Mining using Bio-inspired techniques

2010-14 Sander van Splunter (VU)
Automated Web Service Reconfiguration

2010-15 Lianne Bodenstaff (UT)
Managing Dependency Relations in Inter-Organizational Models

2010-16 Sicco Verwer (TUD)
Efficient Identification of Timed Automata, theory and practice

2010-17 Spyros Kotoulas (VU)
Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications

2010-18 Charlotte Gerritsen (VU)
Caught in the Act: Investigating Crime by Agent-Based Simulation

2010-19 Henriette Cramer (UvA)
People's Responses to Autonomous and Adaptive Systems

2010-20 Ivo Swartjes (UT)
Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative

2010-21 Harold van Heerde (UT)
Privacy-aware data management by means of data degradation

2010-22 Michiel Hildebrand (CWI)
End-user Support for Access to Heterogeneous Linked Data

2010-23 Bas Steunebrink (UU)
The Logical Structure of Emotions

2010-24 Dmytro Tykhonov
Designing Generic and Efficient Negotiation Strategies

2010-25 Zulfiqar Ali Memon (VU)
Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective

2010-26 Ying Zhang (CWI)
XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines

2010-27 Marten Voulon (UL)
Automatisch contracteren

2010-28 Arne Koopman (UU)
Characteristic Relational Patterns

2010-29 Stratos Idreos(CWI)
Database Cracking: Towards Auto-tuning Database Kernels

2010-30 Marieke van Erp (UvT)
Accessing Natural History - Discoveries in data cleaning, structuring, and retrieval

2010-31 Victor de Boer (UVA)
Ontology Enrichment from Heterogeneous Sources on the Web

2010-32 Marcel Hiel (UvT)
An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems

2010-33 Robin Aly (UT)
Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval

2010-34 Teduh Dirgahayu (UT)
Interaction Design in Service Compositions

2010-35 Dolf Trieschnigg (UT)
Proof of Concept: Concept-based Biomedical Information Retrieval

2010-36 Jose Janssen (OU)
Paving the Way for Lifelong Learning; Facilitating competence development through a learning path specification

2010-37 Niels Lohmann (TUE)
Correctness of services and their composition

2010-38 Dirk Fahland (TUE)
From Scenarios to components

2010-39 Ghazanfar Farooq Siddiqui (VU)
Integrative modeling of emotions in virtual agents

2010-40 Mark van Assem (VU)
Converting and Integrating Vocabularies for the Semantic Web

2010-41 Guillaume Chaslot (UM)
Monte-Carlo Tree Search

2010-42 Sybren de Kinderen (VU)
Needs-driven service bundling in a multi-supplier setting - the computational e3-service approach

2010-43 Peter van Kranenburg (UU)
A Computational Approach to Content-Based Retrieval of Folk Song Melodies

2010-44 Pieter Bellekens (TUE)
An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain

2010-45 Vasilios Andrikopoulos (UvT)
A theory and model for the evolution of software services

2010-46 Vincent Pijpers (VU)
e3alignment: Exploring Inter-Organizational Business-ICT Alignment

2010-47 Chen Li (UT)
Mining Process Model Variants: Challenges, Techniques, Examples

2010-48 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2010-49 Jahn-Takeshi Saito (UM)
Solving difficult game positions

2010-50 Bouke Huurnink (UVA)
Search in Audiovisual Broadcast Archives

2010-51 Alia Khairia Amin (CWI)
Understanding and supporting information seeking tasks in multiple sources

2010-52 Peter-Paul van Maanen (VU)
Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention

2010-53 Edgar Meij (UVA)
Combining Concepts and Language Models for Information Access

2011-01 Botond Cseke (RUN)
Variational Algorithms for Bayesian Inference in Latent Gaussian Models

2011-02 Nick Tinnemeier(UU)
Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language

2011-03 Jan Martijn van der Werf (TUE)
Compositional Design and Verification of Component-Based Information Systems

2011-04 Hado van Hasselt (UU)
Insights in Reinforcement Learning; Formal analysis and empirical evaluation of temporal-difference learning algorithms

2011-05 Base van der Raadt (VU)
Enterprise Architecture Coming of Age - Increasing the Performance of an Emerging Discipline

2011-06 Yiwen Wang (TUE)
Semantically-Enhanced Recommendations in Cultural Heritage

2011-07 Yujia Cao (UT)
Multimodal Information Presentation for High Load Human Computer Interaction

2011-08 Nieske Vergunst (UU)
BDI-based Generation of Robust Task-Oriented Dialogues

2011-09 Tim de Jong (OU)
Contextualised Mobile Media for Learning

2011-10 Bart Bogaert (UvT)
Cloud Content Contention

2011-11 Dhaval Vyas (UT)
Designing for Awareness: An Experience-focused HCI Perspective

2011-12 Carmen Bratosin (TUE)
Grid Architecture for Distributed Process Mining

2011-13 Xiaoyu Mao (UvT)
Airport under Control. Multiagent Scheduling for Airport Ground Handling

2011-14 Milan Lovric (EUR)
Behavioral Finance and Agent-Based Artificial Markets

2011-15 Marijn Koolen (UvA)
The Meaning of Structure: the Value of Link Evidence for Information Retrieval

2011-16 Maarten Schadd (UM)
Selective Search in Games of Different Complexity

2011-17 Jiyin He (UVA)
Exploring Topic Structure: Coherence, Diversity and Relatedness

2011-18 Mark Ponsen (UM)
Strategic Decision-Making in complex games

2011-19 Ellen Rusman (OU)
The Mind ' s Eye on Personal Profiles

2011-20 Qing Gu (VU)
Guiding service-oriented software engineering - A view-based approach

2011-21 Linda Terlouw (TUD)
Modularization and Specification of Service-Oriented Systems

2011-22 Junte Zhang (UVA)
System Evaluation of Archival Description and Access

2011-23 Wouter Weerkamp (UVA)
Finding People and their Utterances in Social Media

2011-24 Herwin van Welbergen (UT)
Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior

2011-25 Syed Waqar ul Qounain Jaffry (VU)
Analysis and Validation of Models for Trust Dynamics

2011-26 Matthijs Aart Pontier (VU)
Virtual Agents for Human Communication - Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots

2011-27 Aniel Bhulai (VU)
Dynamic website optimization through autonomous management of design patterns

2011-28 Rianne Kaptein(UVA)
Effective Focused Retrieval by Exploiting Query Context and Document Structure

2011-29 Faisal Kamiran (TUE)
Discrimination-aware Classification

2011-30 Egon van den Broek (UT)
Affective Signal Processing (ASP): Unraveling the mystery of emotions

2011-31 Ludo Waltman (EUR)
Computational and Game-Theoretic Approaches for Modeling Bounded Rationality

2011-32 Nees-Jan van Eck (EUR)
Methodological Advances in Bibliometric Mapping of Science

2011-33 Tom van der Weide (UU)
Arguing to Motivate Decisions

2011-34 Paolo Turrini (UU)
Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations

2011-35 Maaike Harbers (UU)
Explaining Agent Behavior in Virtual Training

2011-36 Erik van der Spek (UU)
Experiments in serious game design: a cognitive approach

2011-37 Adriana Burlutiu (RUN)
Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference

2011-38 Nyree Lemmens (UM)
Bee-inspired Distributed Optimization

2011-39 Joost Westra (UU)
Organizing Adaptation using Agents in Serious Games

2011-40 Viktor Clerc (VU)
Architectural Knowledge Management in Global Software Development

2011-41 Luan Ibraimi (UT)
Cryptographically Enforced Distributed Data Access Control

2011-42 Michal Sindlar (UU)
Explaining Behavior through Mental State Attribution

2011-43 Henk van der Schuur (UU)
Process Improvement through Software Operation Knowledge

2011-44 Boris Reuderink (UT)
Robust Brain-Computer Interfaces

2011-45 Herman Stehouwer (UvT)
Statistical Language Models for Alternative Sequence Selection

2011-46 Beibei Hu (TUD)
Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work

2011-47 Azizi Bin Ab Aziz(VU)
Exploring Computational Models for Intelligent Support of Persons with Depression

2011-48 Mark Ter Maat (UT)
Response Selection and Turn-taking for a Sensitive Artificial Listening Agent

2011-49 Andreea Niculescu (UT)
Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality

2012-01 Terry Kakeeto (UvT)
Relationship Marketing for SMEs in Uganda

2012-02 Muhammad Umair(VU)
Adaptivity, emotion, and Rationality in Human and Ambient Agent Models

2012-03 Adam Vanya (VU)
Supporting Architecture Evolution by Mining Software Repositories

2012-04 Jurriaan Souer (UU)
Development of Content Management System-based Web Applications

2012-05 Marijn Plomp (UU)
Maturing Interorganisational Information Systems

2012-06 Wolfgang Reinhardt (OU)
Awareness Support for Knowledge Workers in Research Networks

2012-07 Rianne van Lambalgen (VU)
When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions

2012-08 Gerben de Vries (UVA)
Kernel Methods for Vessel Trajectories

2012-09 Ricardo Neisse (UT)
Trust and Privacy Management Support for Context-Aware Service Platforms

2012-10 David Smits (TUE)
Towards a Generic Distributed Adaptive Hypermedia Environment

2012-11 J.C.B. Rantham Prabhakara (TUE) Process Mining in the Large: Preprocessing, Discovery, and Diagnostics

2012-12 Kees van der Sluijs (TUE)
Model Driven Design and Data Integration in Semantic Web Information Systems

2012-13 Suleman Shahid (UvT)] Fun and Face: Exploring non-verbal expressions of emotion during playful interactions

2012-14 Evgeny Knutov(TUE)
Generic Adaptation Framework for Unifying Adaptive Web-based Systems

2012-15 Natalie van der Wal (VU)
Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes

2012-16 Fiemke Both (VU)
Helping people by understanding them - Ambient Agents supporting task execution and depression treatment

2012-17 Amal Elgammal (UvT)
Towards a Comprehensive Framework for Business Process Compliance

2012-18 Eltjo Poort (VU)
Improving Solution Architecting Practices

2012-19 Helen Schonenberg (TUE)
What's Next? Operational Support for Business Process Execution

2012-20 Ali Bahramisharif (RUN)
Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing

2012-21 Roberto Cornacchia (TUD)
Querying Sparse Matrices for Information Retrieval

2012-22 Thijs Vis (UvT)
Intelligence, politie en veiligheidsdienst: verenigbare grootheden?

2012-23 Christian Muehl (UT)
Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction

2012-24 Laurens van der Werff (UT)
Evaluation of Noisy Transcripts for Spoken Document Retrieval

2012-25 Silja Eckartz (UT)
Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application

2012-26 Emile de Maat (UVA)
Making Sense of Legal Text

2012-27 Hayrettin Gurkok (UT)
Mind the Sheep! User Experience Evaluation Brain-Computer Interface Games

2012-28 Nancy Pascall (UvT)
Engendering Technology Empowering Women

2012-29 Almer Tigelaar (UT)
Peer-to-Peer Information Retrieval

2012-30 Alina Pommeranz (TUD)
Designing Human-Centered Systems for Reflective Decision Making

2012-31 Emily Bagarukayo (RUN)
A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure

2012-32 Wietske Visser (TUD)
Qualitative multi-criteria preference representation and reasoning

2012-33 Rory Sie (OUN)
Coalitions in Cooperation Networks (COCOON)

2012-34 Pavol Jancura (RUN)
Evolutionary analysis in PPI networks and applications

2012-35 Evert Haasdijk (VU)
Never Too Old To Learn – On-line Evolution of Controllers in Swarm- and Modular Robotics

2012-36 Denis Ssebugwawo (RUN)
Analysis and Evaluation of Collaborative Modeling Processes

2012-37 Agnes Nakakawa (RUN)
A Collaboration Process for Enterprise Architecture Creation

2012-38 Selmar Smit (VU)
Parameter Tuning and Scientific Testing in Evolutionary Algorithms

2012-39 Hassan Fatemi (UT)
Risk-aware design of value and coordination networks

2012-40 Agus Gunawan (UvT)
Information Access for SMEs in Indonesia

2012-41 Sebastian Kelle (OU)
Game Design Patterns for Learning

2012-42 Dominique Verpoorten (OU)
Reflection Amplifiers in self-regulated Learning

2012-43 Withdrawn

2012-44 Anna Tordai (VU)
On Combining Alignment Techniques

2012-45 Benedikt Kratz (UvT)
A Model and Language for Business-aware Transactions

2012-46 Simon Carter (UVA)
Exploration and Exploitation of Multilingual Data for Statistical Machine Translation

2012-47 Manos Tsagkias (UVA)
Mining Social Media: Tracking Content and Predicting Behavior

2012-48 Jorn Bakker (TUE)
Handling Abrupt Changes in Evolving Time-series Data

2012-49 Michael Kaisers (UM)
Learning against Learning - Evolutionary dynamics of reinforcement learning algorithms in strategic interactions

2012-50 Steven van Kervel (TUD)
Ontologogy driven Enterprise Information Systems Engineering

2012-51 Jeroen de Jong (TUD)
Heuristics in Dynamic Sceduling; a practical framework with a case study in elevator dispatching

2013-01 Viorel Milea (EUR)
News Analytics for Financial Decision Support

2013-02 Erietta Liarou (CWI)
MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing

2013-03 Szymon Klarman (VU)
Reasoning with Contexts in Description Logics

2013-04 Chetan Yadati(TUD)
Coordinating autonomous planning and scheduling

2013-05 Dulce Pumareja (UT)
Groupware Requirements Evolutions Patterns

2013-06 Romulo Goncalves(CWI)
The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience

2013-07 Giel van Lankveld (UvT)
Quantifying Individual Player Differences

2013-08 Robbert-Jan Merk(VU)
Making enemies: cognitive modeling for opponent agents in fighter pilot simulators

2013-09 Fabio Gori (RUN)
Metagenomic Data Analysis: Computational Methods and Applications

2013-10 Jeewanie Jayasinghe Arachchige(UvT)
A Unified Modeling Framework for Service Design

2013-11 Evangelos Pournaras(TUD)
Multi-level Reconfigurable Self-organization in Overlay Services

2013-12 Marian Razavian(VU)
Knowledge-driven Migration to Services

2013-13 Mohammad Safiri(UT)
Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly

2013-14 Jafar Tanha (UVA)
Ensemble Approaches to Semi-Supervised Learning Learning

2013-15 Daniel Hennes (UM)
Multiagent Learning - Dynamic Games and Applications

2013-16 Eric Kok (UU)
Exploring the practical benefits of argumentation in multi-agent deliberation

2013-17 Koen Kok (VU)
The PowerMatcher: Smart Coordination for the Smart Electricity Grid

2013-18 Jeroen Janssens (UvT)
Outlier Selection and One-Class Classification

2013-19 Renze Steenhuizen (TUD)
Coordinated Multi-Agent Planning and Scheduling

2013-20 Katja Hofmann (UvA)
Fast and Reliable Online Learning to Rank for Information Retrieval

2013-21 Sander Wubben (UvT)
Text-to-text generation by monolingual machine translation

2013-22 Tom Claassen (RUN)
Causal Discovery and Logic

2013-23 Patricio de Alencar Silva(UvT)
Value Activity Monitoring

2013-24 Haitham Bou Ammar (UM)
Automated Transfer in Reinforcement Learning

2013-25 Agnieszka Anna Latoszek-Berendsen (UM)
Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System

2013-26 Alireza Zarghami (UT)
Architectural Support for Dynamic Homecare Service Provisioning

2013-27 Mohammad Huq (UT)
Inference-based Framework Managing Data Provenance

2013-28 Frans van der Sluis (UT)
When Complexity becomes Interesting: An Inquiry into the Information eXperience

2013-29 Iwan de Kok (UT)
Listening Heads

2013-30 Joyce Nakatumba (TUE)
Resource-Aware Business Process Management: Analysis and Support

2013-31 Dinh Khoa Nguyen (UvT)
Blueprint Model and Language for Engineering Cloud Applications

2013-32 Kamakshi Rajagopal (OUN)
Networking For Learning; The role of Networking in a Lifelong Learner's Professional Development

2013-33 Qi Gao (TUD)
User Modeling and Personalization in the Microblogging Sphere

2013-34 Kien Tjin-Kam-Jet (UT)
Distributed Deep Web Search

2013-35 Abdallah El Ali (UvA)
Minimal Mobile Human Computer Interaction

2013-36 Than Lam Hoang (TUe)
Pattern Mining in Data Streams

2013-37 Dirk B¨rner (OUN)
Ambient Learning Displays

2013-38 Eelco den Heijer (VU)
Autonomous Evolutionary Art

2013-39 Joop de Jong (TUD)
A Method for Enterprise Ontology based Design of Enterprise Information Systems

2013-40 Pim Nijssen (UM)
Monte-Carlo Tree Search for Multi-Player Games

2013-41 Jochem Liem (UVA)
Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning

2013-42 Leon Planken (TUD)
Algorithms for Simple Temporal Reasoning

2013-43 Marc Bron (UVA)
Exploration and Contextualization through Interaction and Concepts

2014-01 Nicola Barile (UU)
Studies in Learning Monotone Models from Data

2014-02 Fiona Tuliyano (RUN)
Combining System Dynamics with a Domain Modeling Method

2014-03 Sergio Raul Duarte Torres (UT)
Information Retrieval for Children: Search Behavior and Solutions

2014-04 Hanna Jochmann-Mannak (UT)
Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation

2014-05 Jurriaan van Reijsen (UU)
Knowledge Perspectives on Advancing Dynamic Capability

2014-06 Damian Tamburri (VU)
Supporting Networked Software Development

2014-07 Arya Adriansyah (TUE)
Aligning Observed and Modeled Behavior

2014-08 Samur Araujo (TUD)
Data Integration over Distributed and Heterogeneous Data Endpoints

2014-09 Philip Jackson (UvT)
Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language

2014-10 Ivan Salvador Razo Zapata (VU)
Service Value Networks

2014-11 Janneke van der Zwaan (TUD)
An Empathic Virtual Buddy for Social Support

2014-12 Willem van Willigen (VU)
Look Ma, No Hands: Aspects of Autonomous Vehicle Control

2014-13 Arlette van Wissen (VU)
Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains

2014-14 Yangyang Shi (TUD)
Language Models With Meta-information

2014-15 Natalya Mogles (VU)
Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare

2014-16 Krystyna Milian (VU)
Supporting trial recruitment and design by automatically interpreting eligibility criteria

2014-17 Kathrin Dentler (VU)
Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability

2014-18 Mattijs Ghijsen (VU)
Methods and Models for the Design and Study of Dynamic Agent Organizations

2014-19 Vincius Ramos (TUE)
Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support

2014-20 Mena Habib (UT)
Named Entity Extraction and Disambiguation for Informal Text: The Missing Link

2014-21 Kassidy Clark (TUD)
Negotiation and Monitoring in Open Environments

2014-22 Marieke Peeters (UU)
Personalized Educational Games - Developing agent-supported scenario-based training

2014-23 Eleftherios Sidirourgos (UvA/CWI)
Space Efficient Indexes for the Big Data Era

2014-24 Davide Ceolin (VU)
Trusting Semi-structured Web Data

2014-25 Martijn Lappenschaar (RUN)
New network models for the analysis of disease interaction

2014-26 Tim Baarslag (TUD)
What to Bid and When to Stop

2014-27 Rui Jorge Almeida (EUR)
Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty

2014-28 Anna Chmielowiec (VU)
Decentralized k-Clique Matching

2014-29 Jaap Kabbedijk (UU)
Variability in Multi-Tenant Enterprise Software

2014-30 Peter de Kock Berenschot (UvT)
Anticipating Criminal Behaviour

2014-31 Leo van Moergestel (UU)
Agent Technology in Agile Multiparallel Manufacturing and Product Support

2014-32 Naser Ayat (UVA)
On Entity Resolution in Probabilistic Data

2014-33 Tesfa Tegegne Asfaw (RUN)
Service Discovery in eHealth

2014-34 Christina Manteli (VU)
The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems

2014-35 Joost van Oijen (UU)
Cognitive Agents in Virtual Worlds: A Middleware Design Approach

2014-36 Joos Buijs (TUE)
Flexible Evolutionary Algorithms for Mining Structured Process Models

2014-37 Maral Dadvar (UT)
Experts and Machines United Against Cyberbullying

2014-38 Danny Plass-Oude Bos (UT)
Making brain-computer interfaces better: improving usability through post-processing

2014-39 Jasmina Maric (UvT)
Web Communities, Immigration and Social Capital

2014-40 Walter Oboma (RUN)
A Framework for Knowledge Management Using ICT in Higher Education

2014-41 Frederik Hogenboom (EUR)
Automated Detection of Financial Events in News Text

2014-42 Carsten Eickhoff (TUD)
Contextual Multidimensional Relevance Models