Expertise Identification in Enterprise Social Media

October, 2012

Master's Program

Systems Engineering, Policy Analysis and Management
Information Architecture Track

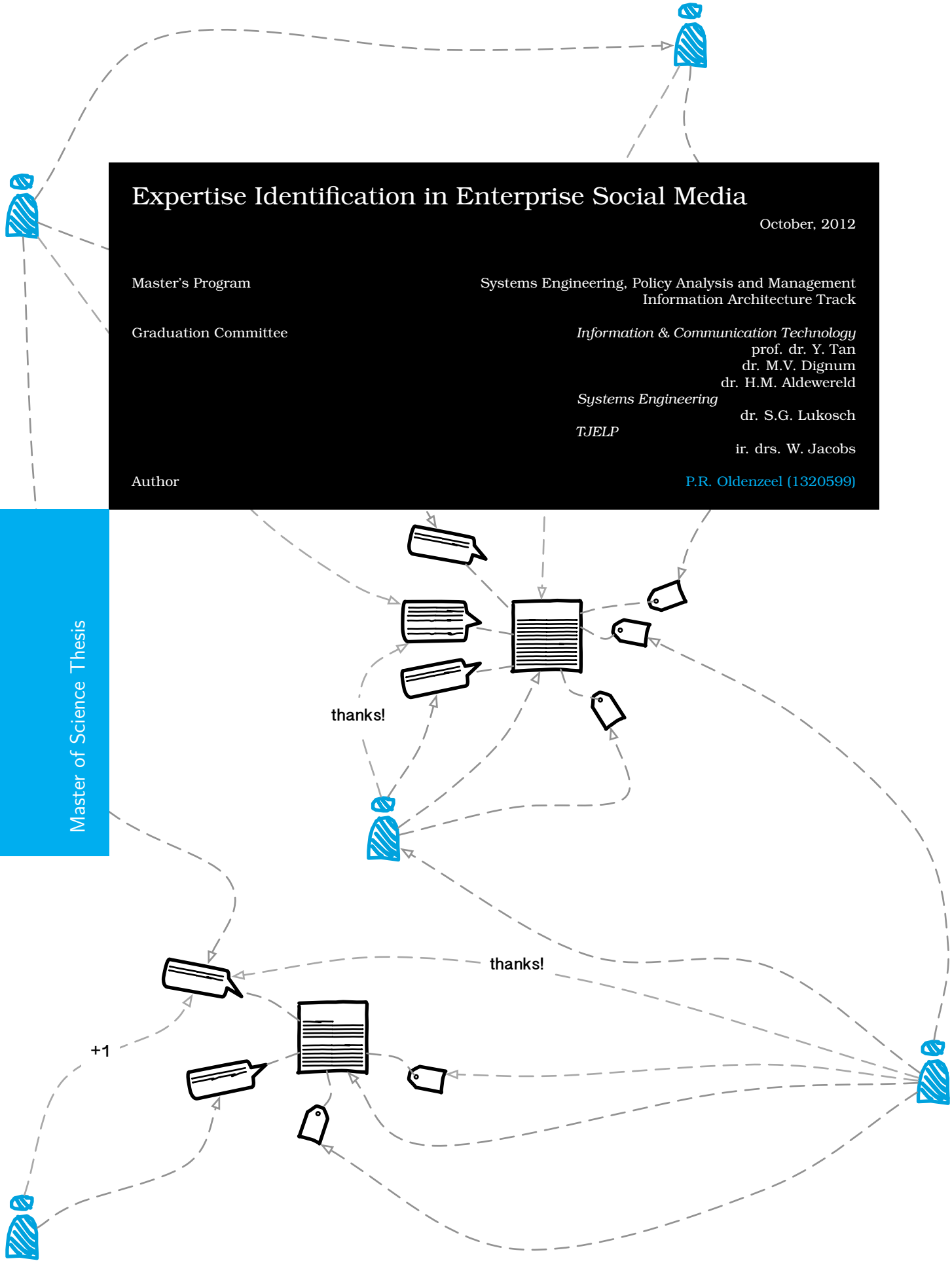
Graduation Committee

Information & Communication Technology
prof. dr. Y. Tan
dr. M.V. Dignum
dr. H.M. Aldewereld
Systems Engineering
dr. S.G. Lukosch
TJELP
ir. drs. W. Jacobs

Author

P.R. Oldenzeel (1320599)

Master of Science Thesis



TU Delft Delft University of Technology
Faculty of Technology, Policy and Management

Expertise Identification in Enterprise Social Media

Master of Science Thesis

For the degree of Master of Science in Systems Engineering, Policy Analysis and Management - Track Information Architecture at Delft University of Technology

P.R. Oldenzeel

October, 2012

Faculty of Technology, Policy and Management · Delft University of Technology

This research was conducted in cooperation with TJELP, an Amsterdam-based company working on Expertise Identification in Enterprise Social Media.

Delft University of Technology
Department of
Infrastructure Systems & Services

The signees hereby certify that they have read and recommend to the Faculty of Technology, Policy and Management, for acceptance, a thesis entitled

Expertise Identification in Enterprise Social Media

by

P.R. Oldenzeel

in partial fulfillment of the requirements for the degree of

Master of Science Systems Engineering, Policy Analysis and Management - Track Information Architecture

Dated: October, 2012

Graduation Committee:

prof. dr. Y. Tan
Infrastructure Systems & Services Department - Section ICT

dr. M.V. Dignum
Infrastructure Systems & Services Department - Section ICT

dr. H.M. Aldewereld
Infrastructure Systems & Services Department - Section ICT

dr. S.G. Lukosch
Multi Actor Systems Department - Section SE

ir. drs. W. Jacobs
TJELP

Abstract

The increasing adoption of Enterprise Social Media (ESM) systems within enterprises is driven by the need for the explicit facilitation of sharing expertise. Expertise Identification (EI) functionality can satisfy this need. The social-media-like content and Collaborative Filtering (CF) annotation data available in ESM, however, pose unique requirements on EI. In this light, we perform an elaborate study into literature and practice surrounding ESM, expertise, and EI, in order to formulate a number of design requirements and choices for EI in ESM. In our case study on E-view, a live ESM system, we design, implement, and test an EI prototype that stores all ESM relationships in a social graph and all user content into a search engine, which are then combined to estimate user expertise scores. Our results reveal that relevant content used to estimate expertise scores should be selected on the basis of both full-content *and* tags. Due to the sparsity of CF appreciation data in the dataset, EI strategies that complement content relevance scoring with appreciation scoring for the estimation of expertise scores, perform equally well as strategies based only on content relevance, in terms of ranked lists of experts. As such, we recommend that future work retests the EI strategies with the constructed prototype, using an ESM dataset that contains more CF appreciation data. We present a number of recommendations for EI in ESM and for reusing our evaluation methods in future research.

Contents

List of Figures

List of Tables

Acknowledgements

I am still getting used to the idea that my studies at the Delft University of Technology have finally come to an end. Welcoming the challenge of what lies beyond, I must admit that, without the help of a number of people, I would still be stuck in demarcating my thesis subject. As such, I want to use this opportunity to thank all the people that helped me to 'get it done'.

I would like to thank Willem, my first supervisor at TJELP, for all the contributions and time he invested in this project. Without his persistent conviction and enthusiasm to have me graduate at TJELP, I might have made the mistake of graduating somewhere else completely.

Besides Willem, I want to thank all the other TJELP people for their enthusiasm, help, brainstorms and insights. As time progressed, I was privileged to see TJELP grow in numbers, into what is now an even more accomplished company then when I arrived.

Also, I want to thank Virginia and Huib, my primary supervisors at the university, for their patience and enthusiasm, as well as their abundant contributions to my research. I still remember the endless conversations about further demarcation of my subject, something I at times mistook for forcibly down-scaling my ambitions. Without their help, I would have needed at least another 200 pages to complete this thesis.

I feel I should express my sincere gratitude toward Stephan, my secondary supervisor at the university. If he would not have joined my graduation committee when he did, I would have been stuck in university bureaucracy for the rest of the summer.

Lastly, I would like to thank all of my friends, roommates and family for bearing through this extended period of social isolation, disinterest and egocentricity. I can only imagine the irritation some of them must have felt at times.

Delft, University of Technology — P.R. Oldenzeel
October, 2012

"Knowledge must be gained by ourselves. Mankind may supply us with the facts; but the results, even if they agree with previous ones, must be the work of our mind."

— *Benjamin Disraeli*

Chapter 1

Introduction

With enterprises recognizing the potential of social media for use within their organizations, Enterprise Social Media (ESM) are becoming increasingly popular (Miles, 2011; Forrester, 2010). An ESM system is a professional digital environment within an organization's boundaries in which employees communicate and collaborate by means of social-media-like digital content, sharing knowledge within and outside the enterprise.

Research on the benefits and key drivers of ESM within organizations has pointed out that *sharing knowledge and expertise* is one of the most prominent reasons for organizations to adopt an ESM system (Richter & Riemer, 2009; Forrester, 2010; Miles, 2011). While more effectively and efficiently sharing knowledge and expertise within an organization has been one of the promises of Knowledge Management (KM) systems for approximately two decades, such systems lack the support and contribution of employees to realize this promise (Dagostino, 2004). By drawing on employees' support for social-media-like communication, ESM systems are thought to have the potential to facilitate the sharing of knowledge and expertise.

1-1 Problem

The need for sharing knowledge and expertise within an ESM system is in fact twofold: employees with an information need have to be facilitated in locating and retrieving information, and employees facing a complex problem have to be facilitated in locating and contacting other employees or people outside of the organization that possess knowledge or expertise with regard to the problem at hand (Richter & Riemer, 2009). The first need is typically serviced by a search engine that matches a query with relevant documents and other digital resources. Expertise Identification (EI), the identification and retrieval of experts with respect to a certain topic, an active field in both academic and corporate research, is used to address the latter need. While research in this field has yielded a number of methods for EI in a variety of digital environments, ESM environments have not yet been considered. This is to be expected, since the first of these environments came into existence only a few years ago.

We already briefly touched upon the social-media-like content by which users communicate and collaborate in ESM systems. Besides this social-media-like content, these systems typically also exploit Collaborative Filtering (CF) techniques to annotate that content. CF techniques enable users to annotate their and other users' content in order to improve content retrievability, discourage system abuse and reward user performance (Amatriain, Pujol, Tintarev, & Oliver, 2009; Halpin, Robu, & Shepherd, 2007; Huang & Zeng, 2011). Common examples of CF techniques are the possibility for users to *tag* their own content and *rate* others'. While there has been an amount of research on using CF data for EI, to our knowledge, there is no published research on the issue of EI in ESM systems and the utilization of these CF techniques.

1-2 Research Objective

The objective of this research is to identify viable EI methods and to select the most suitable of these methods for ESM. This EI method should take into account the unique requirements ESM pose on EI, as well as fully exploit the unique combination of social-media-like content and Collaborative Filtering data contained in ESM.

Accordingly, we formulate the following main research question:

What form of Expertise Identification is most suitable for Enterprise Social Media, accounting for its social-media-like content and Collaborative Filtering characteristics?

We split this main research question into a number of sub questions:

1. How is Expertise Identification in Enterprise Social Media systems different from that in other digital environments for knowledge sharing and which requirements do these differences pose?
    (a) How do Enterprise Social Media systems differ from other digital environments for knowledge sharing?
    (b) What requirements do these differences pose on Expertise Identification in Enterprise Social Media systems?
2. How can expertise be defined and quantified in the context of an Enterprise Social Media system?
3. Which best practices from existing Expertise Identification systems can be reused in ESM?
    (a) How do existing Expertise Identification systems estimate expertise?
    (b) How can Collaborative Filtering data contribute to Expertise Identification?
4. How can Expertise Identification be implemented in Enterprise Social Media systems?

The results of this research comprise answers to all formulated sub research questions, subsequently providing an answer to the main research question.

1-3 Demarcation

In this thesis, we discuss a number of KM, ESM and EI systems. We are only able to review a small portion of these systems, and we do not claim to give a conclusive overview and argumentation. We hope other research will extend our findings by considering other KMS, ESM and EI systems.

We will focus specifically on *Expertise Identification*, not *Expertise Explication*. *Expertise Identification* refers to identifying the experts with respect to a certain query X, whereas *Expertise Explication* refers to explicating what expertise an expert Y possesses, i.e. explaining what that user's fields of expertise are (McDonald, 2001).

Furthermore, we only take into account professional use of ESM, not personal use. In reality, some ESM use can be expected to be personal, as is the case sometimes with professional e-mail accounts.

1-4 Approach and Methods

This section describes our approach and methods in answering the posed research questions.

Literature Study

To answer research questions 1, 2 and 3, in Part 1 of this thesis, we perform an elaborate literature study. First, we discuss the background of KM systems, ESM systems and EI. Then, by considering a number of existing ESM systems as well as relevant ESM characteristics, we determine requirements they pose on EI. Next, we establish a conceptual definition of expertise, which we will later quantify for the case of EI in ESM systems. Finally, we collect best practices for EI in ESM by exploring the methods for identifying experts in existing EI systems and reviewing these methods using the collected requirements. The result of our literature study comprises a set of design requirements and design choices to fulfill these requirements.

Case Study Analysis

In Part 2, we address research question 4 by performing a case study on E-view, an ESM system focused on the corporate sharing of knowledge and expertise. In order to answer research question 4, we pose the following sub-research questions as part of this case study:

1. How can Expertise Identification be implemented in E-view?
    (a) What kind of ICT architecture is required to facilitate the identification of expertise in E-view?
2. Which form of Expertise Identification is most suitable for E-view?

By designing an EI prototype and implementing it in E-view, we test a number of EI strategies, based on our design choices from Part 1. Ultimately, we establish the most suitable form of EI for E-view as a concept for EI in ESM in general.

1-5 Outline

This thesis is divided into two parts. Part 1 comprises of chapters 2 through 6. Chapter 2 gives background information regarding the various systems and techniques mentioned in the introduction. In chapter 3, we research ESM in more detail, specifically looking into the characteristics that set apart ESM systems from other digital environments for knowledge sharing and the requirements these differences pose on EI. Chapter 4 describes contemporary definitions surrounding expertise. Moreover, in this chapter we determine a suitable mapping of expertise onto ESM attributes. Then, in chapter 5, we describe various existing EI systems and methods as well as their most important characteristics, extracting best practices for use of EI in ESM. Finally, in chapter 6, we answer research questions 1, 2 and 3.

Part 2 comprises of chapters 7 through 12. In chapter 7, we discuss the background of our case study on E-view. Chapter 8 lays out our technical design of an EI prototype for ESM and our implementation in E-view. In chapter 9, we describe our tests and explain how we derived various strategies for EI in E-view, taking into account requirements and best practices from Part 1. In chapter 10, we discuss the results of our tests on the established strategies for EI in E-view and for ESM in general. Then, in chapter 11, we answer research question 4 and our main research question. Next, we present our recommendations, the limitations of this research and future work in chapter 12. Finally, in chapter 13, we reflect on the results and process of this thesis.

Figure 1-1 visualizes the outline of this thesis. The grey middle block of the figure represents our main research question. The upper half of the figure represents Part 1 of our thesis. The lower half represents Part 2.

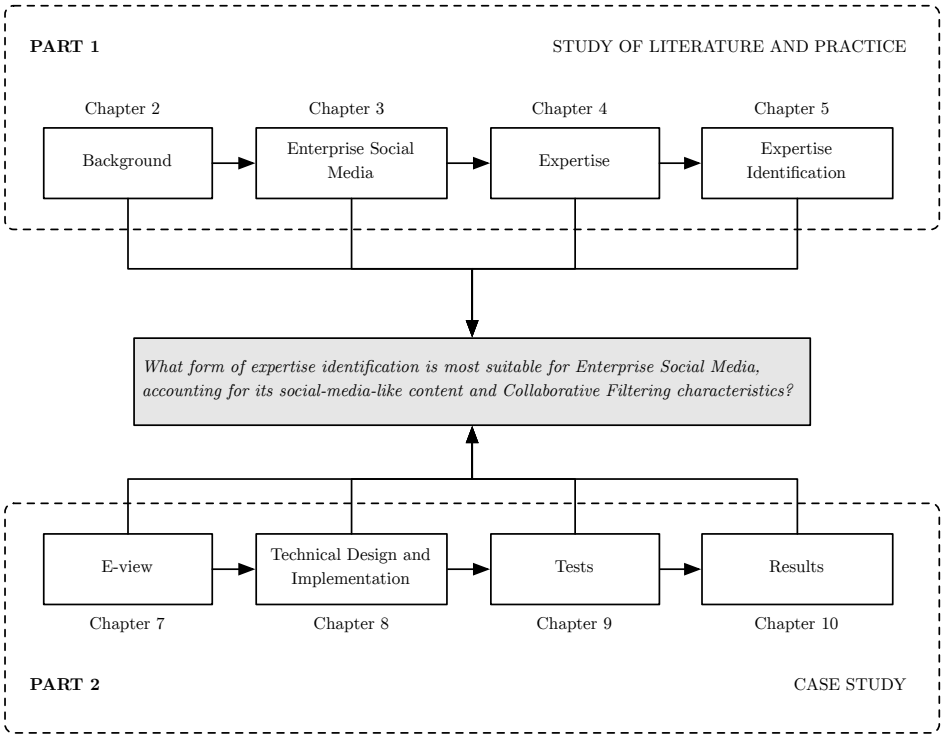


Figure 1-1: Visual outline of this thesis.

PART 1

Chapter 2

Background

In the introduction, we explained the need to be able to find relevant experts in Enterprise Social Media (ESM) systems. Furthermore, we introduced Expertise Identification (EI) as a way to fulfill that need. Before we start collecting best practices from existing ESM and EI systems in order to elicit requirements ESM pose on EI, we need to further explore the context and background of EI in digital environments for knowledge sharing. In order to do so, we first discuss the relationships between Knowledge Management (KM) systems and ESM. Next, we describe the purpose and context of EI in more detail. Finally, we elaborate on Collaborative Filtering (CF) as an alternative or complement to content as the main source for identifying expertise, as well as its relationship to Reputation systems (RS).

2-1 Enterprise Social Media and Knowledge Management Systems

KM refers to identifying individual and collective knowledge in an organization and leveraging that knowledge into a competitive edge (Alavi & Leidner, 2001; Krogh, 1998). Ever since the rise of personal computers in the workplace, enterprises have attempted to facilitate knowledge management practices by implementing KM systems, designed to ease storage and retrieval of company data, information and knowledge (Rastogi, 2000; Alavi & Leidner, 2001; Dagostino, 2004). Alavi and Leidner (1999) define KM systems as *"a line of systems which target professional and managerial activities by focusing on creating, gathering, organizing, and disseminating an organization's 'knowledge' as opposed to 'information' or 'data' "*. Within the enterprise, KM systems are often referred to as *intranets*, *extranets* or *collaborative environments*. While promising, the implementation of these systems has shown an 80 percent failure rate as enterprises' social cultures did not consolidate well with the workflow dictated by these systems (Dagostino, 2004). Knowledge management systems traditionally require employees to not only communicate about their work activities via e-mail, but also redundantly upload related documents and notes to the system for storage, causing a duplication of work (Alavi & Leidner, 2001; Becerra-Fernandez, 2006). This results in little cooperation from employees. ESM systems integrate digital social-media-like communication

with KM functionality in order to create seamless integration of corporate communication and knowledge sharing that does enjoy cooperation from employees. In contrast to traditional KM systems, ESM systems are often adopted bottom-up, by having employees starting to use the social-media-like communication functionality. Traditional KM systems are known for their top-down implementation, which, combined with the redundancy in work described earlier in this paragraph, is believed to cause the bad adoption rates. KM systems tend to suffer from the cold-start-problem caused by top-down implementation: if only few employees use the system, their efficiency gain is very small, since they will only be able to find part of the files and information they need in the KM systems and part in their mailboxes. In order for KM systems (and ESM for that matter) to function properly, all employees should ideally participate actively. Getting employees to actively participate in any ICT system hinges on many different aspects, ranging from organizational culture and structure to employees' proficiency in working within digital environments. The aim of this research, however, is not to determine such drivers and thresholds for employees to share knowledge. Our considerations stretch as far as considering the ease with which employees *could* collaborate within an ICT system, ignoring behavioral impediments to share knowledge in the first place.

Furthermore, we do not claim that ESM will overcome all of the problems of KM systems, neither do we posit that ESM will succeed where KM systems did not. We do posit, however, that by integrating social-media-like means of communication and collaboration, its adoption rate and ease of use are more promising than that of traditional KM systems.

2-2 Expertise Identification

The move from traditional KM systems and e-mail to ESM improves the retrievability of information from communication, since e-mail forces the storage of numerous copies of the same message within replies and mailboxes of recipients (Miles, 2011; Forrester, 2010). However, ESM are typically focussed on organizing unstructured communication rather than facilitating the finding and sharing of expertise. ESM systems currently on the market host the unstructured Facebook-type communication between employees, much like that on social media sites Facebook and Twitter. Consequently, employees' primary activities within ESM are founded in organizational *"culture"* and *"fun and communication"* (Raj, Dey, & Gaonkar, 2011).

For ESM to overcome the shortcomings and bad adoption rates of KM systems, social-media-like communication alone appears insufficient to fulfill the need for KM in an organization. ESM systems need to explicitly facilitate the finding and sharing of expertise. Accurately providing relevant information to users fulfills part of this requirement. An answer to a user's question, however, is not the same as a solution to a user's problem. The latter goes beyond providing the right information to a user and towards connecting that user with one or more experts that are able to help solve the problem at hand. In order to suggest subject matter experts, the system needs to be able to estimate the relevant *valid expertise* of employees in an organization.

Suggesting experts to a user based on a question or problem is the main topic of EI research, a subfield of Information Retrieval. EI systems comprise a special category of KM systems, aimed at finding and suggesting people rather than information. In literature and in practice, EI systems are also called Expertise Locator Systems (ELS) (Becerra-Fernandez, 2006),

Expertise Finders (EF) and Expert Recommender Systems (ERS) (Balog, Azzopardi, & Rijke, 2006; Huang et al., 2006). Furthermore, some research positions EI systems as a type of Recommender system (Balog et al., 2006; Hennis, Lukosch, & Veen, 2011; McDonald, 2001). Recommender systems are used to suggest relevant content or – more generally – objects to users based on some query. NASA's Expert Seeker was one of the first EI systems, in which the expertise of each employee was entered manually, enabling management to find the right expert for the job or problem at hand (Becerra-Fernandez, 2006). It is important to realize that EI systems are typically found as a component of KM systems, ESM systems or other digital environments for knowledge sharing (Balog et al., 2006). For the purpose of clarity, illustration 2-1 shows this overlap between EI systems and the types of digital environments for knowledge sharing we have mentioned up to now.

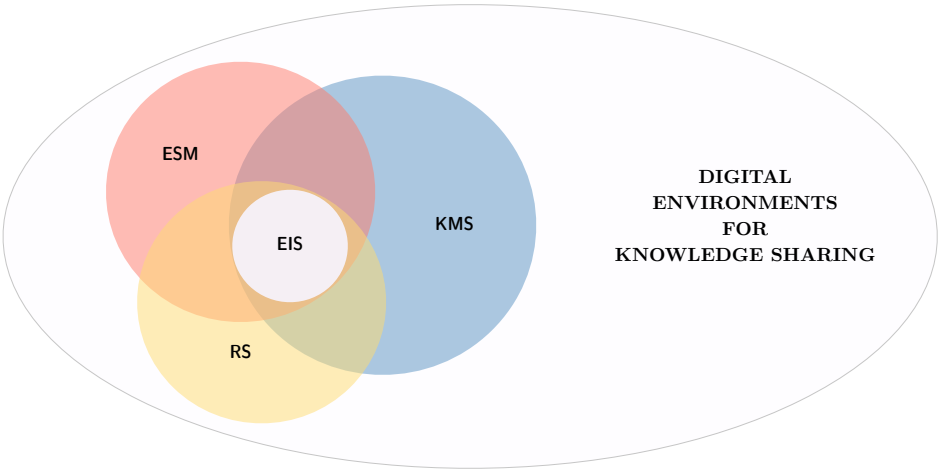


Figure 2-1: The overlap between Enterprise Social Media (ESM), Knowledge Management (KM) systems, Recommender systems (RS) and Expertise Identification (EI) systems.

In practice, it seems that only few KM systems facilitate EI (Venkateshprasanna, Gandhi, Mahesh, & Suresh, 2011). Those that do facilitate it, often require manual creation and management of User Expertise Profiles (UEPs), revealing a number of disadvantages (Yang & Huh, 2008; Becerra-Fernandez, 2006; Oosterman, 2011). Firstly, the vast number of data and users of KM systems makes manually maintaining UEPs into a very time consuming task. Secondly, this task is typically assigned to either system administrators or users themselves, all of whom can be considered biased with respect to judging their and others' expertise. Early EI systems had in common that the identification of experts was based on a homogenous set of content, i.e. the manually inputted account of each employee's knowledge and competencies along with manually provided levels of proficiency (Jansen, 2010). Examples of such systems are Hewlett Packard's CONNEX system, the KSMS system with the American National Security Agency (Becerra-Fernandez, 2006) and Microsoft's SPuD system. Another big problem with these systems was that people had little or no incentive to carefully fill out their knowledge profiles, let alone keep them up to date (Jansen, 2010; Balog et al., 2006; Becerra-Fernandez, 2006). Summarizing, a number of issues surround the manual inputting and updating of User Expertise Profiles:

- it is time-consuming for employees, and thus costly for the organization (Jansen, 2010; Balog et al., 2006);

- employees tend to either exaggerate or downplay their competencies in fear of losing their jobs or being assigned more responsibilities (Becerra-Fernandez, 2006);
- both system administrators and users cannot be expected to be objective with respect to judging users' expertise and proficiency levels;
- employees who do keep their profiles up to date, often do not directly benefit from doing so, e.g. in the form of requests to work on projects (Jansen, 2010; Perry, Candlot, & Schutte, 2009), and;
- because predefined concepts of knowledge and competence areas tend to be more generic than when freely described by employees, choosing the right concepts to represent knowledge and competencies is difficult (Becerra-Fernandez, 2006; Jansen, 2010; Balog et al., 2006).

EI functionality within ESM systems needs to counter these problems in order to gain employee cooperation and be effective. We argue they can be countered by posing a number of general requirements on EI in ESM. It should:

- be unsupervised if possible, i.e. performed largely automatically;
- incentivize employees to make use of the ESM system for their corporate activities, and;
- not predefine the categories or topics of expertise, or at least not limiting to users.

2-3 Data-mining Content

More modern EI systems estimate a user's expertise in a mostly unsupervised fashion by data-mining a selection of that user's content within the enterprise. That content may exist in the form of e-mails in a mailbox, documents in a KM system, messages within an ESM system or some other form of digital information. Table 2-1 shows a non-exhaustive overview of different types of content typically present in either KM or ESM systems, based on an inquiry of live KM and ESM systems. Just like EI systems, ESM systems are often components of digital environments for knowledge sharing, or of a companies' intranets. In table 2-1, however, we consider KM and ESM systems as distinct systems. Moreover, note that the table only reflects the *presence* of types of content in KM systems and ESM, not the functionality available to manipulate that content. We have found files, for example, to be present in both KM systems and ESM. However, in KM systems, there is usually more functionality in place to manage files (e.g. rights management and check-in-check-out functionality) than there is in ESM.

KM systems primarily facilitate document storage and often support some types of organizational communication (e.g. news messages). In the case of KM systems, e-mail still serves as the primary means of communication, whereas ESM systems integrate that communication.

The question is whether conventional forms of EI are equally suitable for the social-media-like contents and CF data in ESM.

Content Type	Content Data	KMS	ESM
Files	Documents	Yes	Yes
	Images	Yes	Yes
	Videos	No	Sometimes
	Hyperlinks	Yes	Yes
	Bookmarks	Sometimes	Yes
Communication	Chatter	No	Yes
	Q&A	No	Yes
	Ideas	Sometimes	Yes
	Polls	Sometimes	Sometimes
	News messages	Sometimes	Yes
	Discussions	No	Yes
	Meetings	No	Yes
Workflow	Tasks	Sometimes	Sometimes
	Appointments	Sometimes	Sometimes
	Contacts	Sometimes	Yes

Table 2-1: Overview of different types of content typically present in Knowledge Management and Enterprise Social Media systems.

2-4 Collaborative Filtering

The aim of digital environments for knowledge sharing is to have all employees participate actively, which generates a large number of largely unstructured data. Especially when enabling and storing social-media-like communication in ESM systems, one such system's contents can be vast, making it difficult for both the system and its users to distill relevant information from it. With the rise of interactive online communities and e-commerce over the last decade, CF (also called 'Collaborative Tagging' (Benz, Körner, Hotho, Stumme, & Strohmaier, 2011), 'Social Tagging' (Abel, Cardosodearaujo, Gao, & Houben, 2011) or 'Social Information' (Amitay, 2008)) has emerged as a popular means of annotating content. CF involves users annotating their and others' contributions, for instance by providing keywords that describe a message's contents on a blog or adding a tag *toread* to a personal message. Other examples are marking a user's answer to a question on Stackoverflow[1] as the 'best answer', up- or down-voting users' comments to express appreciation for their contribution, and rating a seller on Ebay with respect to product quality, delivery and service.

CF data can be divided into *abstraction data* (e.g. providing keywords and adding tags) and *appreciation data* (e.g. up- or down-voting users' comments or rating a seller on Ebay). Abstraction data can be used as an alternative for data-mining the full textual content of a message or document. We suggest that instead of data-mining full content, EI can be performed by mining only the CF abstraction data, assuming this data is sufficiently representative for a message's content (Heymann, Koutrika, & Garcia-Molina, 2008). Appreciation data can be used to collect value judgements about users' content. Subsequently, appreciation data can be used to collect information about a user's reputation; the expectations other users may pose on that user's expertise with respect to a certain topic (Hennis et al., 2011).

[1]Stackoverflow.com is a public online Q&A platform used by professional and private users to pose questions and answers related to IT.

Information on a user's reputation, in turn, enables other users to make a value judgement about that user's expertise (Farmer & Glass, 2010). CF appreciation data has the structure of a *reputation statement*, described in Reputation systems literature as consisting of a *reputing source* exerting a *claim* about a *reputed target* (Farmer & Glass, 2010). Figure 2-2 shows an example of a reputation statement in the case of a user adding another user's post to favorites.

Figure 2-2: A reputation statement of a user adding another user's post to favorites.

CF data, if sufficiently provided by users, enables ESM systems to suggest relevant content and people to a user based on similarities with that user's search query, profile or other characteristics (Z. K. Zhang, Zhou, & Zhang, 2011). ESM systems are typically equipped with a range of Collaborative Filtering (CF) techniques to facilitate users in providing abstraction and appreciation data. A non-exhaustive overview of these techniques is listed in table 2-2. In chapter 3, we discuss CF techniques in three existing ESM systems.

CF Type	CF Data	KMS	ESM
Abstraction data	Keywords	Yes	Yes
	Tags	No	Sometimes
Appreciation data	Likes	No	Sometimes
	Rating	No	Sometimes
	Votes	No	Sometimes
	Thanking	No	Sometimes
	Favorites	Sometimes	Sometimes
	Following	No	Sometimes

Table 2-2: Overview of different types of Collaborative Filtering data typically present in Knowledge Management and Enterprise Social Media systems.

EI might also utilize CF data, because the expertise of a user with respect to a certain topic correlates with the reputation of that user with respect to the topic. Besides the benefits to EI, reputation systems have also been found to stimulate user participation in online environments where successful knowledge sharing depends on all participants (Hennis et al., 2011). A common method to determine the expertise of a user X with respect to a topic Y based on CF data, is to calculate the term frequency - inverse document frequency (TFIDF) of some tag Y in that user's messages. Combining the frequency of a term in a document and the inverse document frequency of that term in the whole document collection, TFIDF is used to calculate *relative* term importance. It appoints high scores to terms (in this case tags) that occur a lot of times in the messages a user submitted and a small number of times in all users' messages. Equation 2-1 is used to calculate the inverse document frequency. Equation 2-2 is used to calculate TFIDF for a term t in a document d.

$$idf(t) = \log \frac{|D|}{(1 + |\{d : t \in d\}|)} \tag{2-1}$$

Where D is the total number of documents, and $(1+|\{d : t \in d\}|)$ is the number of documents term t appears in. When the term frequency equals 0, the denominator becomes 1 to avoid dividing by zero.

Then, TFIDF(t, d) is formulated as:

$$TFIDF(t, d) = tf(t, d) * idf(t) \tag{2-2}$$

If an employee, for example, submits five messages regarding HRM procedures and subsequently tags these messages with the tag *HRM*, based on TFIDF that employee will be considered to possess relatively more expertise on HRM, if few or none of other users' messages contain the tag *HRM*.

2-5 Concluding Remarks

In this section, we discuss the main conclusions from this chapter, together with a number of design requirements for EI in ESM. Table 2-3 enumerates these design requirements. In the next chapter, we extend this table with requirements imposed by ESM systems.

We have explored the theoretical and practical background of ESM as successors of traditional KM systems, with social-media-like content and CF data as its main differentiators. Additionally, we posited that ESM systems need to be able to suggest both relevant information and relevant experts to users in order to facilitate the finding and sharing of expertise. Suggesting relevant experts requires the ability to estimate *valid expertise*. Therefore, estimating valid expertise is our first design requirement. Incorporating EI into ESM, typically implemented as a system component rather than an autonomous system, can fulfill this requirement. From history, we know that having employees manually construct the User Expertise Profiles necessary for EI does not yield the desired results. Consequently, EI in ESM should base User Expertise Profiles on employees' *automatically* data-mined digital content within the enterprise. We argue that this automatic extraction of UEPs is in fact a design choice to fulfill ease of use. Consequently, we posit *easy to use* as our second design requirement and the automatic extraction of UEPs as a design choice to realize that ease of use.

	Design Requirement	**Design Choice**
1.	Estimate valid expertise.	
2.	Easy to use.	1. Automatically extract User Expertise Profiles.

Table 2-3: Overview of the design requirements and choices for Expertise Identification in Enterprise Social Media systems, formulated in chapter 2.

Users' social-media-like content in ESM, however, differs greatly from users' content in KM systems, hence we need to find out whether conventional methods for EI can also provide good results in ESM. In chapter 5, we investigate the issues and best practices from these existing systems and methods for EI.

Besides data-mining users' content for EI, CF data in ESM systems may also provide useful input. We have divided CF data into *abstraction* and *appreciation* data. Abstraction data may be used as an alternative to data-mining full content, assuming the abstraction data inputted by users is sufficiently representative for the full content. Appreciation data can be used to determine a user's reputation with respect to topics of expertise.

The content and CF data in ESM systems can serve as input for EI. The challenge lies in combining best practices from existing EI systems (based on user content) with data on the appreciation and abstraction of that content offered by CF techniques. In the following chapters, we collect these best practices from existing ESM and EI systems, in the form of design requirements and matching design choices.

Chapter 3

Enterprise Social Media

In chapter 1, we defined an Enterprise Social Media (ESM) system as *"a professional digital environment within an organization's boundaries in which employees communicate and collaborate by means of social-media-like content, sharing knowledge within and outside the enterprise"*. As this definition suggests, ESM constitute a range of systems, rather than one specific digital environment. Furthermore, in chapter 2, we described ESM systems as successive or complementary to legacy Knowledge Management (KM) systems. It is clear that ESM can have a great number of forms, both in terms of functionality and implementation.

As we want to derive requirements that ESM systems pose on Expertise Identification (EI), we first look into the ESM system type and determine its main characteristics. We demarcate what we mean by ESM. Then, we briefly describe three ESM platforms currently available: Yammer, Rypple and Icon. By doing so, we gain insight into the content typically present in ESM platforms and users' annotations of that content using available Collaborative Filtering (CF) techniques. Finally, we derive a number of requirements that ESM platforms pose on EI.

3-1 Social Media in Enterprise Knowledge Sharing

Enterprise Social Media, Enterprise Social Networking (Forrester, 2010; Raj et al., 2011), Enterprise 2.0 (McAfee, 2006; Bughin & Chui, 2010; Miles, 2011), Social Networking 2.0 (Zyl, 2009) and Social Business systems (Miles, 2011) are all terms often used to refer to a category of digital enterprise systems. The first characteristic these systems have in common is that they are typically based on the workings of highly interactive public social media. A direct consequence is that ESM typically contain a lot of *chatter* and other informal communication. The unstructured social-media-like collaboration in these systems, however, better represents the unstructured knowledge work, input and output that takes place in corporate practice (McAfee, 2006).

We prefer the term ESM, because it explicitly states using social media within the enterprise. By using social media within the enterprise we mean using a dedicated social media platform

Application	Use
Blog	Online journal where one or a number of employees can periodically post messages, which can be commented on by readers as well as shared with others.
Wiki	Website containing explicit information that can be expanded, edited and linked collaboratively by employees.
Social Network	Facebook-like enterprise social network in which employees can add others to their corporate social network and share information via loosely structured content.
Bookmarking	Website that enables employees to bookmark and rate interesting content (websites, documents etc.), be it within or outside the organization.
Q&A	Question-and-answer-website enabling employees across the organization to post questions and answers. Providing the right answer is often rewarded in some fashion.
Discussion forum	Online environment where employees can freely discuss topics of interest.

Table 3-1: Examples of Enterprise Social Media implementation.

primarily for use *within* the enterprise, not to be confused with using public social media (like Facebook and Twitter) for corporate purposes. The latter focus on an organization's marketing and public relations, rather than on its internal knowledge management.

According to market research, as well as a small amount of academic literature on ESM, it is used for a variety of corporate activities. However, the top five of these activities differs greatly between reports. Research on an Enterprise Social Network platform by Raj et al. (2011) stipulates *"fun and entertainment"* as the main category of activities, along with organizational *"culture"* and *"work & life balance"* in second and third place (Raj et al., 2011). This is not surprising, since ESM are based on the popular social-media-like communication style of public social media. Market research commissioned by Cisco (2010) points out the need to *"work better in distributed teams"* and *"reduce the amount of email"* as the most important drivers for ESM adoption (Forrester, 2010). According to market research by AIIM Market Intelligence (2011), *"finding and sharing expertise"* and the breaking down of *"departmental/geographic barriers"* are also key drivers for adoption.

Because ESM serve a wide spectrum of corporate goals and activities, its different implementations are also numerous. Table 3-1 shows a non-exhaustive list of various types of ESM systems. These types are non-exclusive, i.e. there can be overlap between two or more types depending on their implementation.

The basis for all these types of ESM systems is highly interactive online collaboration, often referred to as web 2.0 capabilities (McAfee, 2006). In many enterprises, some combination of these types of ESM systems has been implemented on top of or in parallel with existing KM systems. As the capabilities of ESM mature, it is reasonable to expect that KM systems will only serve as document storage and all other information will reside in an organization's ESM system.

3-1-1 Criteria for ESM

According to Zyl (2009), ESM (Social Networking 2.0 as she phrases it) systems should live up to three criteria. They should:

1. support social networking (must contain all three components);
   (a) build a digital expression of people's personal relationships and links;
   (b) aid in the discovery of potential ties, and;
   (c) aid in the conversion of potential ties into weak or strong ties;
2. support computer-mediated communication (must contain at least two components);
   (a) one-on-one, or;
   (b) one-to-many/one-to-few, or;
   (c) many-to-many/few-to-few;
3. and allow social feedback;
   (a) contributions by a member are rated by other users.

These three criteria strongly emphasize the social and interactive character of ESM. Criterion 1 refers directly to the social network within ESM, facilitating employees in finding each other based on some commonality. Criterion 2 posits that a combination of communication modes is present in ESM. A discussion forum, for instance, supports all three modes of communication. A blog, however, supports only one-to-many (e.g. a post by the author) and one-on-one (e.g. private feedback on an article) communication. Criterion 3 emphasizes the importance of CF techniques in enabling users to annotate each others' content.

Based on Zyl's criteria, the content types in table 2-1 and the ESM types in table 3-1, we argue that ESM systems always contain three aspects of corporate processes: *communication*, *people (users)* and *objects*. From Zyl's criteria we can conclude that ESM place heavy emphasis on people, with the requirement of explicating users' social networks and encouraging them to expand it. All of the ESM systems in table 3-1 facilitate communication, people (users) and objects. Users communicate with each other in a public social-media-like fashion, with posts and comments to those posts. Furthermore, communication always centers around objects, be it an author's blog post in a blog, a problem or question in a Q&A system or a bookmarked web resource in a bookmarking system. One can easily find other systems that also involve these three aspects, such as an e-mail client. And we do not claim that the aspects *communication*, *people* and *objects* is unique to ESM. However, it provides us with a structural approach to exploring the global entities in an ESM system.

McAfee discusses two more user-oriented ground rules for ESM (or as he calls it, Enterprise 2.0) (McAfee, 2006). An ESM system should:

1. be easy to use, with no special software or skills required, and;
2. not *"impose on users any preconceived notions about how work should proceed or how output should be categorized or structured"*.

Advantages	Disadvantages
Lower costs and energy consumption due to economies of scale and hardware virtualization.	Complete reliance on ESM developer for hardware and software, i.e. lock-in effect.
More responsive problem resolution due to easily accessible central software and hardware.	More susceptible to hardware and software issues caused by other clients, little control over issue resolution.
Pricing model based on pay-per-use.	Less predictable than traditional fixed fee.
Scalability and flexibility (no long term commitment necessary).	Privacy and security issues.

Table 3-2: Advantages and disadvantages of Software-as-a-Service (SaaS) when compared to traditional service models.

Mcafee's first ground rule is established by integrating corporate communication and knowledge sharing formerly done using KM systems and e-mail. Furthermore, most ESM systems can be operated using only an internet browser, increasing ease of use (McAfee, 2006). The second ground rule would prevent a predetermined dictionary of content categories for tagging and areas of expertise to be used in an ESM system. Organization and categorization of content should emerge from system use. In chapter 5, we further discuss the implications of this ground rule.

3-2 SaaS in the Cloud

ESM systems are typically available as Software-as-a-Service (SaaS), a form of Cloud Computing. One of the most popular hosting providers for SaaS products is *Amazon Elastic Compute Cloud (EC2*[2]*)*. As such, client organizations can acquire a turnkey ESM system, which is hosted by either the product developer or a third party Cloud provider. That way, client organizations do not have to concern themselves with a lot of the requirements of reliable and secure application hosting. Organizations' existing KM systems are traditionally installed on either dedicated servers online or offline (in the case of an internal datacenter). Well-known examples of growing ESM SaaS products are *Google Apps* by Google, a range of products by *Salesforce*, Yammer's *Enterprise Social Network* and *blogs* by Wordpress[3].

With ESM systems online in the cloud, one ESM system contains data of multiple organizations, neatly separated and shielded from each other. This way, client organizations outsource the hardware, software, network infrastructure, energy costs, maintenance and updates to the ESM provider. SaaS shows several advantages and disadvantages when compared to traditional dedicated hosting, displayed in table 3-2 (Leavitt, 2009; Armbrust et al., 2010).

We do not claim that ESM systems are never installed on organizations' internal mainframes rather than in the cloud. However, since SaaS are practically always implemented in the cloud, we argue ESM are too.

[2]Visist http://aws.amazon.com/ec2 for more information on Amazon EC2.
[3]Wordpress.com offers free blogs to users, configurable and accessible via their site.

3-3 Target User Group

Because ESM can be used for a variety of corporate purposes and is typically implemented as SaaS, we argue that any organization that is prepared to host part of their software and data in the cloud is a potential ESM customer. However, in reality most organizations that adopt ESM are geographically dispersed with a lot of mobile employees, a high degree of familiarity with public social media driving adoption, and finally a moderate to aggressive ICT innovation policy (Forrester, 2010; Miles, 2011). In section 3-2 we considered ESM usually facilitating a number of different client organizations, each operating within a distinct part of the system. Consequently, it typically contains a number of different client organizations. This setup makes for a diverse overall user base grounded in different languages and cultures, operating a range of business activities. This diverse user base makes it difficult to tailor an ESM to the wishes of individual client organizations, an inherent problem of the SaaS model (Leavitt, 2009).

3-3-1 Scalability

ESM systems, like public social media, thrive when actively used by many people. Because of this network effect, the more employees that participate in an ESM system, the more comprehensive its contents (McAfee, 2006). This is especially the case with respect to peer-based voting, rating and other peer-based appreciation (Hennis et al., 2011). With the sum of the users being greater than the parts, data use and required capacity will rapidly rise when more users participate in the system. Initially, when a system is empty, EI – and other functionality that depends on content and/or CF data – will not function properly. This is called the *cold-start problem*, and has also been recognized as a common problem of existing KM systems, as we discussed in section 2-1.

Once the system is used increasingly, data and required capacity must be expected to rise more rapidly as well. Moreover, if a lot of complex algorithms are in place to perform intelligent system behavior, for example EI, scalability is even more important. Therefore, the software and hardware of ESM should be instantly scalable. Because SaaS usually runs on virtualized hardware, hardware capacity is extremely scalable depending on the hosting agreement in place. Accordingly, methods for EI should be light-weight when performed in real-time and should be performed independent from main processes in an ESM system as much as possible.

3-4 Examples of Enterprise Social Media in Practice

Having explored the basic characteristics and surroundings of ESM systems in the previous sections, we discuss three actively used ESM – Yammer, Rypple and Icon – in order to gain better understanding of typical ESM content and CF data in practice. We will discuss these ESM by investigating each environment's implementation of the three aspects defined in section 3-1-1: communication, people and objects. Furthermore, we explore their *content* and *CF data*, in accordance with the distinction we made in chapter 2. We discuss content and CF data types based on table 2-1 and table 2-2.

3-4-1 Yammer

Microsoft's Yammer promotes itself as an Enterprise Social Network. Its user interface greatly resembles that of Facebook, offering new users who are already acquainted with Facebook familiar surroundings. In contrast with Facebook, in Yammer a user is, by default, only allowed to operate within its *company group*. If required, it is also possible to create an *external group*, through which clients and partners can be involved. Figure 3-1 shows a user's Yammer dashboard, closely resembling Facebook's *wall*.

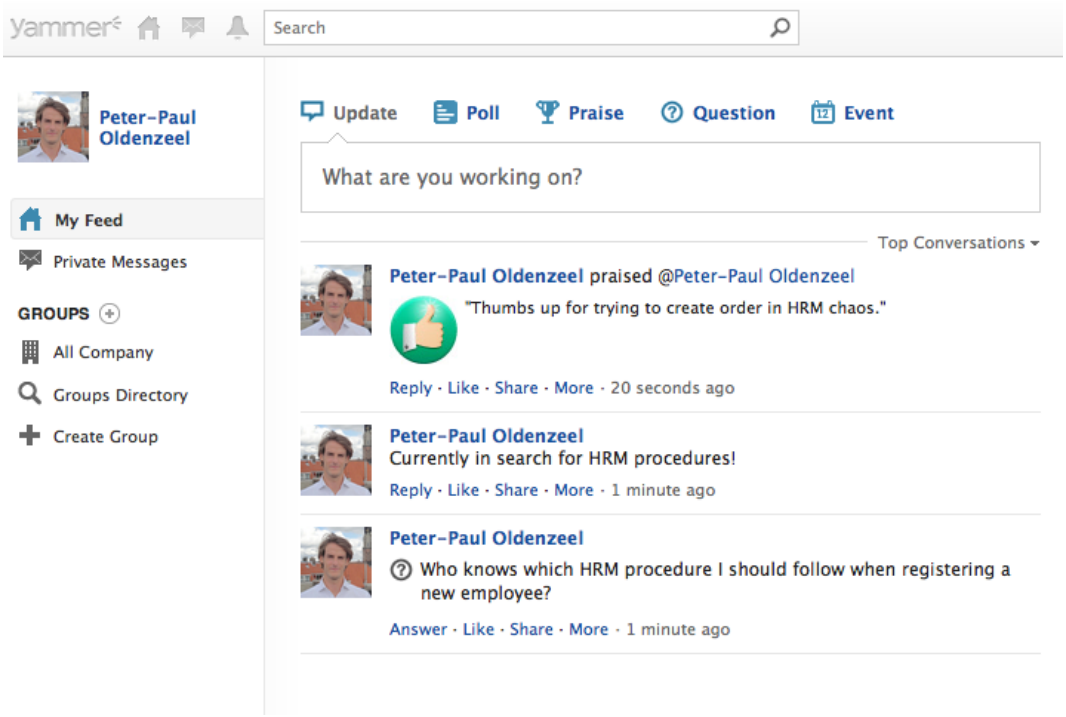


Figure 3-1: Creating a dialog in Yammer.

Content

By default, Yammer supports all *file*, *communication* and *workflow* content types from table 2-1. By purchasing one or more apps within Yammer, other content types – often mixtures of the types in table 2-1 – can be added.

Collaborative Filtering Possibilities

Yammer supports most of the CF possibilities from table 2-2: *tagging* content, *'liking'* posts, *mentioning* posts and people, *following* posts and people and *thanking* people for their contribution by *'praising'* them for it. Rating content (e.g. on a scale from -5 to 5) is not possible.

Expertise Identification

Users can purchase additional apps within Yammer to extend Yammer's functionalities. By default, Yammer does not support EI, but purchasing the Senexx[4] app for Yammer, a client can extend Yammer with EI.

3-4-2 Rypple

Salesforce's Rypple resembles Yammer with regard to user interface, but focuses on objectives, goals, tasks and achievements rather than all round conversation. Figure 3-2 illustrates the Rypple interface. Like in Yammer, Rypple users operate within the company environment. More than Yammer, Rypple places emphasis on building employee reputation. Reputation pages are accessible to all employees, containing an employees' achievements, goals and rewarded *'thank-you'* badges. Rypple does not promote extension or integration with third party functionality like that of Sennex.

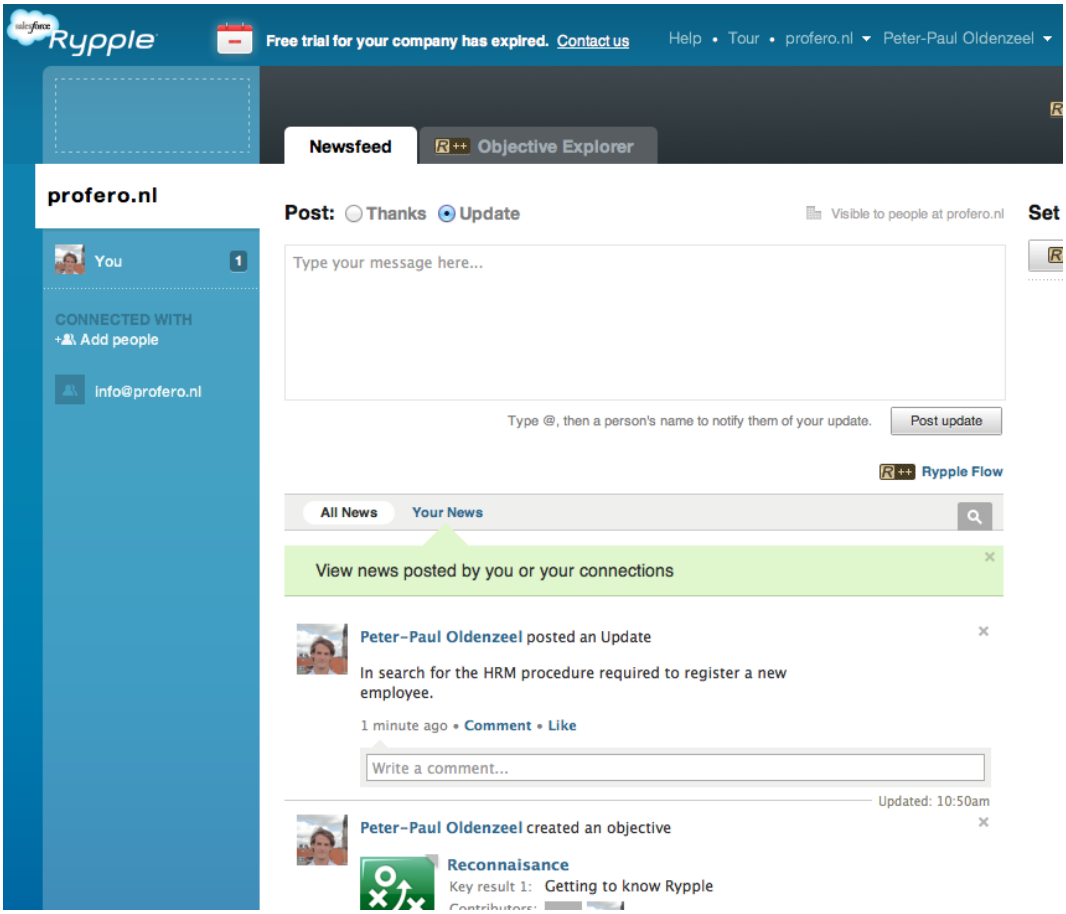


Figure 3-2: Creating a dialog in Rypple.

[4]Senexx is an EI component for enterprise systems that operates on the basis of a black box: documents and communication go in, relevant experts come out. For more information on Senexx, visit http://www.senexx.com.

Content

Rypple supports less content data types than Yammer and Icon do. From the content types in table 2-1 Rypple supports *hyperlinks*, *chatter*, *discussions* and *tasks*.

Collaborative Filtering Possibilities

From the CF techniques in table 2-2, Rypple supports thanking a user for a contribution by awarding preset and custom badges. There is no support for *abstraction data*, such as *keywords* or *tags*.

Expertise Identification

Identifying experts is not explicitly facilitated in Rypple. However, because Rypple keeps track of users' overall reputations, some form of EI is possible. This would require looking up relevant authors of messages manually, and then reviewing their reputations manually. And user reputation is estimated as an overall score, not a topic-based metric.

3-4-3 Icon

Spigit's Icon platform facilitates corporate Q&A by enabling users to post questions with an expiration date and time. Users can provide answers to questions and choose the *'best'* answer out of equally rated answers. Like Yammer and Rypple, Icon users operate within their companies' Icon environment. In contrast with Yammer and Rypple, Icon only focuses on Q&A. Figure 3-3 shows a question in Icon.

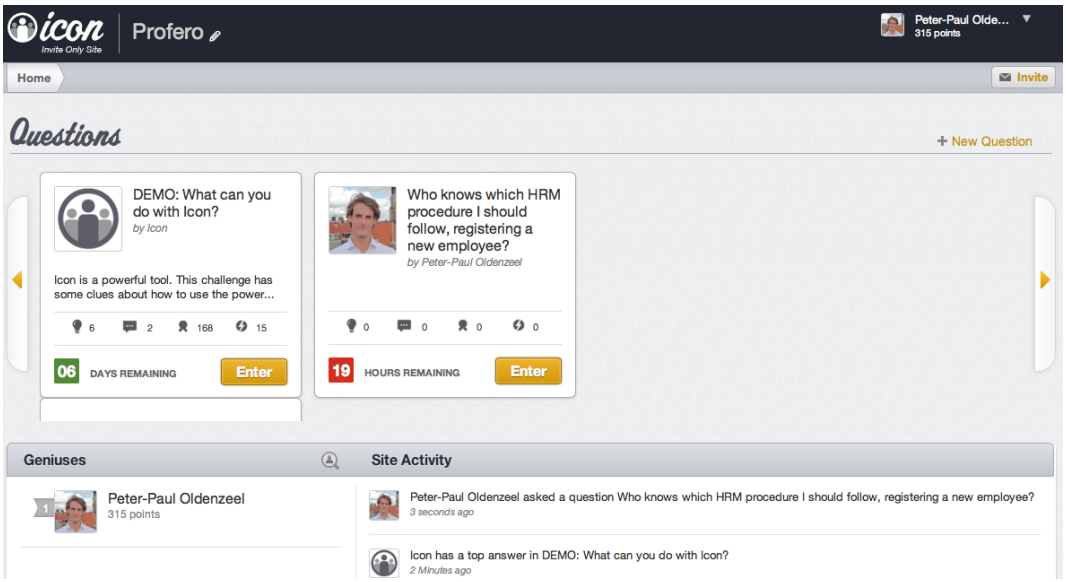


Figure 3-3: Question dashboard in Spigit.

Content

Icon does not facilitate *file* content types. It is, however, possible to link to files using hyperlinks. Users can communicate about ideas and Q&A, but there is no general wall or status board for chatter and other informal communication. Finally, Icon does not support *workflow* content.

Collaborative Filtering Possibilities

In Icon, users can send each other *'gifts'*, each representing 10 *'prize points'*. A user can collect additional prize points by actively participating. Next to prize points, Icon supports *votes* and *best answer* rewards. Besides manually awarded prize points and votes, it keeps track of the number of answers and comments a user has been given. Except for the naming of CF data in Icon, the techniques are a lot like those on the public Q&A platform Stackoverflow.

Expertise Identification

Icon keeps track of the top experts with respect to each question and bases the ranking on the votes the answers of the experts have been given. So users can lookup relevant experts by first looking up relevant questions previously posed by other users. Icon also keeps track of overall *'geniuses'*: users that have the highest scoring combination of prize points, votes, comments and answers within the organization.

3-4-4 Enterprise Social Media Differences in Practice

None of the discussed ESM systems explicitly suggests experts based on a user's search query. Icon is the only system that suggests experts. The ranking of these experts, however, is based on the individual question at hand, not on the overall topic the question pertains to. Yammer supports black-box EI, but only when third party software Senexx is purchased.

Furthermore, our observation in chapter 2 that different ESM systems support different content and CF data types appears to be correct. All three ESM systems use different types of content and CF data types, as well as different naming.

3-5 Concluding: Design Requirements

Now that we have further demarcated ESM characteristics and considered real-world examples of ESM systems, we can complement our design requirements from chapter 2 with a number of requirements ESM pose on EI. Table 3-3 enumerates the requirements and choices we point out in this section, together with the requirement of *valid expertise* we posed in chapter 2.

Throughout this chapter, we have shown that ESM cover a range of different systems, serving a variety of corporate goals and drivers for adoption. They are designed to facilitate internal knowledge sharing within organizations. Furthermore, they are based on highly interactive online collaboration and should support social networking, computer-mediated communication and social feedback. From a user-oriented point of view, ESM systems should be *"easy*

to use" and should *"not impose on users any preconceived notions about how work should proceed or how output should be categorized or structured"* (McAfee, 2006). Hence, the *free categorization and structuring of content* is our third design requirement.

ESM systems are typically available as SaaS, hosted online 'in the cloud'. An ESM system can be expected to contain the ESM environments of multiple organizations, operating in different areas of business, cultures and languages. This makes it difficult to customize the system to the needs of individual client organizations. Consequently, any organization that is prepared to host part of their software and data in the cloud is a potential ESM customer. In reality, ESM client organizations are characterized as geographically dispersed with a lot of mobile employees, a high degree of familiarity with public social media driving adoption of ESM, and finally a moderate to aggressive ICT innovation policy. Concluding, *support for domain-independent, multilingual and multicultural content and users* is our fourth design requirement.

Moreover, these systems are used for a plethora of corporate purposes, so the nature of its contents can be ambiguous. An employee might post some idea about organizing a coffee contest in the morning and post a critical note on the mistreatment of a client in the afternoon. As our exploration of real-world ESM pointed out, they facilitate numerous dialog types, from questions in Icon to objectives in Rypple to polls in Yammer. EI in ESM must take this dialog diversity into account. Consequently, *support for dialog type diversity* constitutes our fifth design requirement. We go into the implications of this ambiguous content for EI more elaborately in chapter 5.

It is important that the hardware and software of an ESM system are instantly scalable. Therefore, *instant scalability of hardware and software* is our sixth design requirement for EI in ESM. While virtualized hardware can be scaled up easily, EI algorithms must be tuned to ensure instant software scalability.

	Design Requirement		**Design Choice**
1.	Estimate valid expertise.		
2.	Easy to use.	1.	Automatically extract User Expertise Profiles.
3.	Ensure free categorization and structuring of content.		
4.	Support for domain-independent, multilingual and multicultural content and users.		
5.	Support for dialog type diversity.		
6.	Ensure instant scalability of hardware and software.		

Table 3-3: Overview of the design requirements for Expertise Identification in Enterprise Social Media systems, complemented in chapter 3.

In chapters 4 and 5, we present further design choices to fulfill the design requirements in this section.

Chapter 4

Defining Expertise in ESM

In 1976, Steven Spurrier organized a wine tasting event in Paris that would later become known as the *Judgment of Paris* (Ericsson & Cokely, 2007). Spurrier, a British wine merchant specialized in French wines and backed by numerous French wine experts, wanted to prove once and for all that French wines were superior to the increasingly popular California wines. Nine french wine experts were asked to blind-taste a selection of the best California and French wines. Surprisingly, California wines received higher scores on all accounts. Furthermore, the experts mistook a number of the French wines for California wines, and vice versa.

This example illustrates the complexity of expertise: when confronted with new wines, the expertise of the French wine experts fell short.

In this chapter, we assess the ambiguity surrounding *expertise* in order to reach a sound conceptualization of expertise in the context of ESM. As discussed in chapter 3, different Enterprise Social Media (ESM) systems estimate user expertise and reputation differently. Consequently, we need to gain an understanding of expertise in ESM in order to investigate its elicitation. Coherent with the distinction made in chapter 2, we concretize expertise into aspects found in content and Collaborative Filtering (CF) data. First, we look into dictionary definitions of *expertise* and related concepts of *knowledge* and *skills*. Second, we discuss defining and eliciting expertise from a philosophical perspective. Then, we consider knowledge by means of two dimensions, that of declarative versus procedural knowledge and that of explicit versus tacit knowledge. Finally, we describe how we define expertise in this thesis and how data available in an ESM system can be used to express it. At the end of this chapter, we present a number of additional design choices to fulfill the design requirements from chapters 2 and 3.

4-1 Dictionary

The Oxford English dictionary defines the concepts *expertise*, *expert*, *knowledge* and *skill* as follows (Oxford, 2012):

1. expertise: *"expert skill or knowledge in a particular field"*;
2. expert (adjective): *"having or involving a great deal of knowledge or skill in a particular area"*;
3. knowledge: *"facts, information, and skills acquired through experience or education; the theoretical or practical understanding of a subject"*;
4. skill: *"the ability to do something well"*, and;
5. skill (count noun): *"a particular ability"*.

Combining definitions 1 and 2, in modern English, *having expertise* means *"having or involving a great deal of knowledge or skill in a particular area"*. This definition only partly disambiguates the concept of expertise; the exact amount of knowledge required for a person to be called an expert on a particular subject area is still unclear.

On the basis of definitions 3 and 4, we argue that the distinction between skill (in the sense of a general ability; meaning 4) on the one hand and knowledge on the other is very vague. Skill appears to be practical in nature, e.g. riding a bike, whereas knowledge seems both practical and theoretical, e.g. possessing a lot of facts, information and skills with respect to HRM procedures. Accordingly, knowledge and skill can be acquired through experience and education. Knowledge seems to encapsulate skill, hence we focus on the definition and measurement of knowledge.

Following through on our line of reasoning, a more concrete definition of having expertise is *"having or involving a great deal of facts, information and/or practical abilities in a particular area, acquired through experience or education"*. The question remains what *amount* of facts, information and/or practical abilities is required for a person to be considered an expert. We believe this *amount* varies from subject area to subject area and from organization to organization. The amount of knowledge required is then relative to the amount of knowledge others possess. It is, however, difficult to measure this amount. If we assume that an organization's ESM system contains a representative reflection of the collective and individual knowledge of the organization's employees, we can use the content at hand to gauge knowledge and thus expertise. We do not ignore that employees will use numerous other modalities to convey knowledge, e.g. by mobile phone or in person. We simply assume that the knowledge that does get entered into the ESM system constitutes a representative reflection of the organization's collective and individual knowledge.

Instead of looking at the *amount* of knowledge required to be considered an expert, one can also look at the output of expertise. Expertise must *"lead to performance that is consistently superior to that of the expert's peers"*, *"produces concrete results"* and *"can be replicated and measured in the lab"* (Ericsson & Cokely, 2007). While software cannot determine such performance or distinguish between concrete and in-concrete results, users' CF appreciation of their content might be able to do just that. It can at least be used to determine which content and which authors are most appreciated in a certain subject area.

4-2 Epistemology

In epistemology, the philosophical study of knowledge, usable definitions of the concept *expertise* and surrounding concepts are subject to continuous debate. In Plato's classic definition of knowledge, a statement has to fulfill three requirements to be considered knowledge (Chappell, 2004):

1. it has to be true;
2. it has to be justified, and;
3. it has to be believed.

All of these requirements are ambiguous and subjective, further demonstrating that the classification and value of data, information, knowledge and expertise rely heavily on an individual's interpretation and perception. In IT systems, in particular, it is extremely difficult or even impossible to determine whether a statement (in the form of a post) fulfills these requirements.

Perry et al. (2009) posit that the multi-user nature of ESM-like platforms requires a shared language between experts *"to confirm relevance, authority and confidence in resources and the information therein"*.

Using these terms they define the *validity* of expertise as follows (Perry et al., 2009):

validity = relevance + authority + confidence, where;

relevance = corresponds to the recipient's interest;

authority = has been assessed by a trusted mediator; is recognized by a large community, could be assumed as proof, and;

confidence = seems interesting to the recipient; is something that is personally trusted.

This explanation of expertise validity seems much in line with Plato's original considerations. In order to gauge expertise in the case of Expertise Identification (EI) in ESM, we thus need to consider the relevance of employees' knowledge with respect to a search query, the subject authority of the employees and the confidence other employees place in their knowledge on the subject.

We argue that the relevance of an employee's knowledge with respect to a search query, can be determined by examining the content that employee has published and is involved in. Equivalently, we propose that the authority and confidence other employees place in an employee with respect to a certain subject, can be determined by analyzing the CF appreciation data related to that employee.

4-3 Declarative, Procedural, Tacit and Explicit Knowledge

Another way of defining knowledge is by considering different dimensions along which knowledge can be distinguished. One such dimension is that of *declarative* versus *procedural* knowledge (Nurius & Nicoll, 1992). Declarative knowledge is about concepts and facts, e.g. being able to explain in detail what HRM means and what it usually comprises within an organization. Procedural knowledge is about knowing how to (re)act in case of specific events, e.g. knowing how to act when firing an employee and coping with his/her reaction. We expect that it is difficult to distinguish between declarative and procedural knowledge in ESM, because of the textual nature of posts and comments. Procedural knowledge is presumably gained through experience rather than theoretical education, and posts and comments in an ESM at best only convey *information* on how to (re)act in certain situations, not *experience* with doing so.

A second dimension to consider is that of *explicit* and *tacit* knowledge (Nonaka, 1994; Polanyi, 1967). Explicit knowledge is the kind of knowledge that can be stored in IT systems and is typically *"articulated, codified, and communicated in symbolic form and/or natural language"* (Alavi & Leidner, 2001). For instance, knowledge on how to fix a specific computer problem. Tacit knowledge is less easily stored as information, highly personalized and resides in employees' minds. An example of tacit knowledge – in the same HRM context we have used so far – is the best means of firing an employee, e.g. gently and empathic or fiercely and harsh, based on that employee's characteristics.

When speaking of EI in the remainder of this thesis, we are referring to the expertise derived from the kind of explicit declarative knowledge present in ESM. We acknowledge that much of an employee's knowledge cannot be captured by ESM or other IT systems.

We do not claim that the 2 dimensions we have described are the only ones. There are countless other knowledge dimensions, such as individual versus collective, and causal versus conditional. For the purpose of illustrating which knowledge can be stored in ESM, however, we believe the described dimensions are sufficient.

Figure 4-1 illustrates the resulting knowledge space. We argue that both Knowledge Management (KM) and ESM systems contain mostly declarative and explicit knowledge, i.e. facts, information and theoretical know-how. The space representing KM systems is situated more to the right, because these are primarily used to classify and store documents, e.g. manuals, descriptions or reports. These are exemplary forms of explicit declarative and procedural knowledge. ESM is used for human communication as well, thus containing much more chatter, ideas and situational information. In contrast with documents in KM systems, the communication in ESM systems captures more of the knowledge transfer that usually takes place in real life. Hence, ESM is situated more to the lower left of the knowledge space.

4-4 Concluding Remarks

In this section, we discuss the main findings of this chapter, and present a number of design choices for EI in ESM. These design choices fulfill part of the design requirements from chapters 2 and 3. Table 4-1 enumerates our design choices alongside these requirements. In

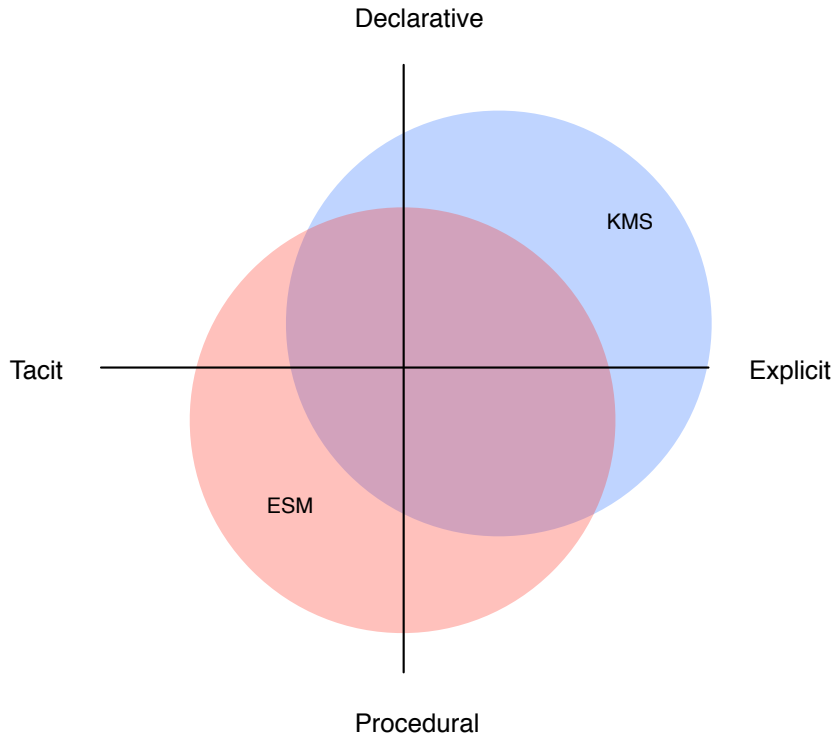


Figure 4-1: Knowledge space illustrating the types of knowledge stored in Enterprise Social Media and Knowledge Management systems.

the next chapter, we will complement table 4-1 with a final round of design choices, based on best practices from existing EI methods and systems.

Considering the ambiguity of knowledge and the fact that only explicit, externalized knowledge can truly be stored in the form of posts in ESM, we argue that most of the knowledge present in an organization cannot be stored in an ESM system at all. It can be expected to capture more of the tacit and procedural knowledge that resides in employees' minds than legacy KM systems do, but in order to share truly tacit knowledge, employees will still have to consult each other in real life. Nevertheless, employees with great explicit declarative knowledge in a particular field are likely to possess tacit procedural knowledge in that field as well, gained through experience and education. Because of this, we believe that we can point out *likely* subject matter experts based on data available in ESM. Therefore, EI in ESM should *suggest* likely experts rather than *claim* or *state* the absolute experts on a particular subject. Accordingly, this is design choice 1 in table 4-1.

The validity of an employee's expertise comprises of the *relevance* of, and the *confidence* and *authority* placed in that employee's knowledge with regard to the subject at hand, constituting design choice 2. Content in ESM can be used to estimate the relevance of an employee's knowledge with respect to a subject, whereas CF appreciation data can be used to gauge the confidence and authority other employees place in it. Consequently, the use of ESM content in estimating knowledge relevance is design choice 3, and estimating confidence and authority users place in knowledge by means of CF appreciation data is design choice 4. It is clear that an employee's knowledge and expertise are strongly subjective to other employees' perceptions of that knowledge and expertise.

Consequently, we define an organization's experts as follows:

Experts are those employees that are, within the boundaries of the organization, confided in and authorized by others to possess the greatest amount of relevant facts, information and/or practical abilities with respect to a particular subject.

As in KM systems, we believe that the goal of ESM should not be to store all knowledge in an organization, but to *facilitate* employees in finding the knowledge and the expertise they seek (Alavi & Leidner, 2001).

	Design Requirement		**Design Choice**
1.	Estimate valid expertise.	1.	Suggest likely experts rather than *claim* or *state* the absolute experts on a particular subject.
		2.	The validity of an employee's expertise comprises of the *relevance*, *confidence* and *authority* placed in that employee's knowledge with respect to a subject.
		3.	Knowledge relevance can be estimated by means of content in ESM.
		4.	Confidence and authority other employees place in knowledge can be estimated by means of CF appreciation data in ESM.
2.	Easy to use.	5.	Automatically extract User Expertise Profiles.
3.	Ensure free categorization and structuring of content.		
4.	Support for domain-independent, multilingual and multicultural content and users.		
5.	Support for dialog type diversity.		
6.	Ensure instant scalability of hardware and software.		

Table 4-1: Overview of the design choices for Expertise Identification in Enterprise Social Media systems, complemented in chapter 4.

Chapter 5

Expertise Identification

In chapter 2, we concluded that Expertise Identification (EI) in Enterprise Social Media (ESM) should base User Expertise Profiles (UEPs) on employees' automatically mined digital content within the enterprise. In chapter 3, we determined that EI in ESM should be based on a combination of both the social-media-like content and Collaborative Filtering (CF) data contained in ESM, with CF abstraction data possibly used as an alternative to full-content. Furthermore, we determined that EI in ESM should account for dialog type diversity, multilingual and multicultural content, and scalability.

Then, in chapter 4, we concluded that content in ESM can be used to estimate the relevance of an employee's knowledge with respect to a subject, while CF appreciation data can be used to gauge the confidence and authority other employees place in that knowledge. The relevance, confidence and authority of an employee's knowledge with respect to a subject constitute the employee's *valid* expertise on that subject.

In this chapter, we elicit best practices from existing methods and systems for EI. We describe the main types of systems and methods, along with their main characteristics. We also describe a number of common issues and trade-offs with respect to EI methods throughout this chapter. Finally, we present a number of design choices, based on the requirements we established in previous chapters and the considerations of chapter 5.

5-1 Existing Expertise Identification Systems and Research

As we discussed in section 1-3, this thesis is about Expertise *Identification*, not Expertise Explication. The first refers to identifying experts with respect to a certain query X, whereas the latter refers to explicating what expertise an expert Y possesses, i.e. explaining his or her fields of expertise are (McDonald, 2001). We argue that Expertise Explication (also called *Expertise Selection* in literature) follows logically from EI; once we are able to identify likely experts based on content and CF data, we are only one step away from explicating an employee's expertise with respect to a certain subject.

In section 2-2, we argued that more modern EI systems make use of data-mining techniques in order to derive expertise from content. Reality is more nuanced, as different EI systems use different:

- content sources (e-mail, unstructured web content, structured scientific papers, document and candidate priors and so on);
- processing steps to get from sources to a ranked list of experts;
  - the use of User Expertise Profiles;
  - taking into account meta-data (e.g. freshness and decay of content);
  - exploitation of the social graph in an ESM, and;
- ranking logic.

In order to better understand the differences between EI systems, we first discuss a number of successful methods for EI in recent years. Then, we go into the main characteristics that are different among EI systems.

5-2 Text REtrieval Conference

From 2005 to 2008, the annually held international Text REtrieval Conference (TREC)[5] included an EI task in their Enterprise Track. The objective of this task was to return a ranked list of experts for a list of given topics (Craswell, Vries, & Soboroff, 2005). In 2009, this task was replaced by an Entity Retrieval task, which was ultimately removed from TREC in 2012 (*TREC Track Overview*, 2012).

In this section, we investigate the best performing EI methods from TREC participants. We leave out results from 2005 because of a reported lack of representativity of the dataset, severely influencing the results (Craswell et al., 2005). Additionally, 2005 was the pilot year for the EI task.

In 2006, 23 groups participated in the EI task, submitting a total of 91 systems. A dataset of the World Wide Web Consortium (W3C) was used as a corpus. The *kmiZHU* group from the Open University achieved first place. The kmiZHU system used a two-stage language modeling[6] approach. They combined a document relevance model with a window-based co-occurrence model (Soboroff, Vries, & Craswell, 2006). Using traditional Information Retrieval tools (BM25[7] and TFIDF[8]), their document relevance model estimates query-document association. Documents that are cited or referenced to more often, are assumed to be more authoritative and thus more relevant (PageRank[9]). Their co-occurrence model measures the query-candidate association by matching candidate names in documents. In doing so, the

[5] Apart from TREC, there are several other international conferences on Information Retrieval. The most prominent are the Special Interest Group on Information Retrieval (SIGIR) and the European Conference on Information Retrieval (ECIR). Please visit http://sigir.org and http://ecir.org for more information.

[6] For an explanation of *language modelling*, visit http://en.wikipedia.org/wiki/Language_model.

[7] Visit http://en.wikipedia.org/wiki/Okapi_BM25 for an explanation of the BM25 ranking function.

[8] See section 2-4 for an explanation of TFIDF.

[9] Visit http://nl.wikipedia.org/wiki/PageRank for more information on the PageRank algorithm.

model takes into account that names matched in different places in the document (e.g. the title versus the body) should be weighted differently when calculating relevance. Moreover, the system takes into account various content window-sizes, with smaller window-sizes receiving higher weights than larger window-sizes. An example is matching a query subject in a document sentence (small window) versus matching that subject in a document section. Finally, the window-based relevance scores are aggregated.

TREC 2007 offered a completely different corpus for EI: the CERC corpus (CSIRO Enterprise Research Collection, http://es.csiro.au/cerc/), which represents the public side of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO). Most importantly, this corpus did not contain a predetermined list of candidate experts. Instead, these were contained within the corpus documents (authors of CSIRO's public web pages). Participants submitted 55 systems, of which 45 automated (Bailey, 2007). The winning team from Tsinghua University constructs a Person Description Document (PDD) (a form of User Expertise Profiles) for candidate experts found in the corpus documents. For around 15% of all the candidate experts they found, they were able to add contextual information from their corresponding personal websites to the PDDs.

In 2008, the same dataset was used: the CERC corpus. A total of 42 systems was submitted, of which 32 automated. The best team was that of the University of Amsterdam. They use candidate *proximity* in their expertise estimation. Additionally, like the winning team of 2007, they import external evidence from the web to enhance User Expertise Profiles. Lastly, they apply UEP-based query expansion, meaning that a topic query is expanded to include additional information in order to enhance query precision (Balog et al., 2008).

5-3 Candidate and Document Models

Forrester (2010) states that *"at the core of social networks are profiles to allow users and the broad community to establish expertise information about an individual"*, pointing out that a knowledge sharing system should construct an expertise profile for each user. In the previous chapters, we discussed User Expertise Profiles and the way these are constructed either manually by users or automatically by the system. EI systems that construct UEPs and use them to estimate user expertise with respect to a queried subject are called *candidate-model* systems (Balog et al., 2006). In contrast with candidate-model EI systems, *document-model* systems estimate user expertise with respect to a query by analyzing the authors of content relevant to the subject (Balog et al., 2006). The difference seems subtle, but is very important. In candidate-model EI systems, UEPs need to be created and then updated continuously, causing information redundancy: each time a user updates content, the system has to update both the content and involved users' UEPs. In contrast, in document-model EI systems, the system has to free up resources to perform content updates *and* at the same time work through all employees associated with the search results to derive likely experts.

These models aim to determine the probability of a candidate expert given a search query, or $p(ca|q)$. Candidates are then ranked according to this probability in order to obtain a ranked list of the top-N experts. By applying Bayes' Theorem, we obtain:

$$p(ca|q) \equiv \frac{p(q|ca)p(ca)}{p(q)}, \tag{5-1}$$

where $p(q|ca)$ represents the *likelihood that the candidate ca publishes content regarding the query q*. Because the *likelihood of a candidate*, $p(ca)$, and the *likelihood of the query*, $p(q)$, can be considered constant for all candidates, we need only focus on $p(q|ca)$.

Following the candidate-model approach, $p(q|ca)$ is calculated by determining the relevance of a candidate's User Expertise Profile with respect to the query. Hence, we speak of a query-profile association and a profile-document association. A UEP can be as simple as containing all content contributed by a user. The query-profile association and consequently $p(q|ca)$ is calculated by matching the query with the content in a user's UEP and weighting these occurrences.

In contrast, the document-model approach requires iterating over all documents relevant to the query and aggregating the relevance scores of documents the candidate is involved in. This approach is depicted as:

$$p(q|ca) \equiv \sum_{d} (p(q|d) * p(d|ca)), \tag{5-2}$$

where $p(q|d)$ represents the *probability of a query given a document*, i.e. the relevance of a document d with respect to a query q. This document-relevance is often calculated using conventional search engine algorithms (Smirnova & Balog, 2011). $p(d|ca)$ is the *probability of a document given a candidate*. In contrast with the candidate-model, the candidate-document association is determined directly from search results, not by constructing a User Expertise Profile. If possible candidates are known, this candidate-document association can be determined by matching candidate names in the document. Iterating over all relevant documents and aggregating document relevance scores for each associated candidate returns the estimated expertise of those candidates.

Both candidate- and document-model systems take content relevance as the basis for estimating the expertise of associated content authors. While this approach fulfills the *relevance* component of valid expertise we discussed in section 4-2, it foregoes the requirements of *authority* and *confidence*. A candidate-model approach misses out on collecting CF appreciation data, because this data is usually not stored in User Expertise Profiles. A document-model approach is more technically complex because all document-candidate associations and aggregation of relevance scores need to be performed in real-time. Furthermore, we argue that a big difference between these two approaches is that candidate-models focus on candidate information, whereas document-models focus on document information. Since we have to expand the content *relevance* with information about the *authority and confidence* other users place in the candidate experts with respect to a queried subject, we believe that EI in ESM systems should combine the models into a hybrid model.

5-4 Supervised versus Unsupervised Expertise Identification

In chapter 2, we argued that EI in ESM should be *unsupervised, i.e. performed largely automatically*. In reality, only very few systems for EI can run completely unsupervised. Some tasks, e.g. solving problems with the software and combatting spammers, will always require human intervention. However, a lot of the tasks that used to be done manually, such as

composing queries from query topics, manual term expansion, manual relevance feedback and manual combination of results, do not deliver performance superior to that of their automated equivalents (Bailey, 2007). While administrative intervention seems largely unnecessary and unwanted, user intervention may be crucial in the evolution of an EI system. Users may provide useful feedback on the precision of ranked lists of experts and may point the system in the right direction in case of errors. Lastly, having users provide an ESM system with CF appreciation data is an inherently manual task.

Concluding, EI in ESM systems should be performed unsupervised and automated, with the exception of:

- administrative maintenance and development;
- user feedback to correct errors and improve system behaviour, and;
- CF data provided by users.

5-5 Categorization: Ontologies and Taxonomies versus Folksonomies

Many ESM systems operate on the basis of taxonomies: up-front categorization schemes developed by either the client organization or the system developers (McAfee, 2006; Jansen, 2010). Taxonomies typically allow (or even *require*) users to choose between categories when creating content (e.g. a post).

A *folksonomy* refers to emergent bottom-up categorization of content by means of users' annotations. Folksonomies emerge when users *collaboratively tag* content, providing both an abstraction of content meaning as well as personal categorization to improve content retrievability (Hotho, Jischke, Schmilz, & Stumme, 2006; Milicevic, Nanopoulos, & Ivanovic, 2010; Yao, Cui, Han, Zhang, & Zhou, 2011; B. Zhang, Zhang, & Gao, 2011; Cai & Li, 2010). Folksonomies are usually flat (no hierarchical distinction) and redundant (synonymous terminology). Their main advantage is that they reflect the information structures and relationships that people actually use, instead of the ones that were planned for them in advance (McAfee, 2006). They also have the advantage of allowing for non-skill-related terms such as project names or highly specific technical terms (Becerra-Fernandez, 2006). Hence, folksonomies can be expected to provide more accurate content abstraction data than taxonomies can.

Ontologies are usually created and maintained by online communities. Together with Natural Language Processing (NLP) techniques – statistical and semantic techniques to process natural language into machine-compatible meaning – they are often used for entity extraction in EI systems. Entity extraction constitutes the extraction of relevant entities, such as names, locations or in our case topics of expertise, from unstructured natural language content. Ontologies resemble taxonomies, with the exception that ontologies are typically used for automated content *classification*, not content *categorization* by users. They work best with domain-specific content, tend to be language-dependent, and need to be created and maintained manually (Abel, Henze, Kawase, Krause, & Siehndel, 2010; Yang & Huh, 2008; Solskinnsbakk & Gulla, 2008). It is typically difficult to obtain a list of knowledge areas or categories from automatically data-mined content (Balog, 2008, 2007; Venkateshprasanna

et al., 2011). Moreover, both taxonomies and ontologies may not (yet) contain category-names that are preferred by users (Cattuto, Benz, Hotho, & Stumme, 2008). An example of a domain-specific ontology is an ontology for *computers*, containing all computer names, measurements, parts and other characteristics. This ontology would, for instance, link a *Dell notebook* and an *Apple Macbook* to one superseding node *Notebook*.

There are a number of semi-domain-independent ontologies, supporting a great number of languages. Examples of such ontologies are DBPedia[10] and WordNet[11]. The DBPedia ontology describes *things* from Wikipedia, while the WordNet ontology describes English words and the way they relate. Although the DBPedia ontology describes *things* in 111 different languages, the English language is overly represented. And WordNet is in fact English-only, while its sister-ontologies (e.g. EuroWordNet) contain other languages.

Many EI systems use static knowledge areas as categories or topics of expertise. Raj et al. (2011), for example, use predetermined categories of questions on Q&A platforms. Datta et al. (2011) derive academics' expertise by ranking research papers on their linkage with other papers (based on citations and authorship). Subsequently, they use a static range of 15 different topics of expertise (e.g. *'Game Theory'* and *'Water Quality'*).

In section 3-1-1, we discussed the requirement of ESM systems *to not impose on users any preconceived notions on the way work should proceed or how output should be categorized or structured*. Following this requirement, EI in ESM should advocate the use of folksonomies. Ontologies and taxonomies may, however, still be used to provide synonyms (e.g. using WordNet), suggest alternative naming and provide meaningful context or background (e.g. using DBPedia).

5-6 Full-content versus Abstraction Data

Besides using CF abstraction data to build up a folksonomy and to improve content retrievability, it can be used as an alternative to full-content in EI. Specifically in determining the relevance of content with respect to a queried subject ($p(q|d)$), or the relevance of a candidate expert UEP with respect to that subject ($p(q|ca)$). The assumption then, is that the CF abstraction data – often *tags* in the case of ESM, as discussed in chapter 4 – are representative[12] for the content they annotate.

By using CF abstraction data for EI, data-mining full-content can potentially be avoided. In that case, the user is doing part of the system's work voluntarily. This can improve scalability both EI and ESM systems. However, since users are not required to provide CF data, the question arises whether they will provide sufficient and high-qualitative tags to support EI. And if users do provide such tags, how can we prevent tags from being abused as a means to be judged an expert falsely?

In order to acquire a deeper understanding of the trade-offs between using full-content or CF abstraction data for EI, we will discuss both sources in the remainder of this section.

[10]DBPedia (http://dbpedia.org) contains extracts of Wikipedia and describes 10.3 million unique *things* in up to 111 different languages (DBPedia, 2012).

[11]WordNet (http://wordnet.princeton.edu) is a lexical database that contains a great number of English words and their relationships.

[12]See also appendix A for an overview of important assumptions.

5-6-1 Content

As argued in section 3-5, content in ESM is diverse, ranging from professional questions to chatter between two employees in informal dialogs. Attached files (documents, images et cetera) can be even more unpredictable with respect to content.

In existing research on EI, the source often consists of structured documents or e-mails. Campbell et al. (2003) estimate expertise by examining an e-mail corpus, using senders and recipients to construct an expertise graph. Because of the structured header information of e-mail messages, candidate experts are extracted fairly easily. Datta et al. (2011) use an academic dataset for EI, containing not only authors of papers, but also co-authors, citation information (links to other papers) and keywords. Compared to ESM content, these sources for EI are quite predictable in terms of format and structure. Content diversity can make data-mining very difficult. One way to increase content predictability would be to have users categorize or annotate their content, in effect providing CF abstraction data.

Although automatically determining the structure and meaning of an ESM system's content may be difficult, determining its authors and recipients is easily accomplished, because each user in an ESM system is typically logged into a user account. Consequently, all possible candidate experts are known. This does not hold for files, which may have been created by someone else or by a group of people other than the user that uploaded the file. Existing research typically determines document-candidate association by matching candidate names in the file (Oosterman, 2011; Balog et al., 2006). We argue that this method is not very reliable, if only because documents tend to mention the author's name as well as a great number of other names, or no names at all (Hertzum & Pejtersen, 2000; Huang et al., 2006). Let alone images, which do not contain author information by default. Because taking into account files in EI constitutes a distinct research task, we leave files – attached/uploaded to ESM, including documents and images – out of scope.

5-6-2 Tags

Just like full-content, tags can be ambiguous. As we have argued in section 5-6, user incentives to provide tags typically lie in improved retrievability of content. At the same time, users will also benefit from each others' tags. Sometimes, users provide tags that do not make sense to other users, e.g. *'#book-ll#'* as a tag to depict that the content it represents is made up of a book, or simply not useful to other users, e.g. *'toread'* to describe the user still has to read the content. Research by Suchanek et al. (2008) found that collections of user-generated tags feature more noise than the terms from either page content or search queries. This noise can diminish overall tag value and even prevent users from providing tags in the first place. Consequently, in this section, we explore ways of coping with that noise and related issues surrounding tags.

Disambiguation

Using ontologies or dimension reduction techniques, tags can be categorized into clusters automatically. That way, the system can recommend tags to users *and* recommend consolidating tags that suffer from:

- *synonymy* (multiple tags with a single meaning, e.g. *'Macintosh'* and *'iPhone'* are both Apple products) versus *'polysemy'* (one tag with multiple meanings, e.g. *'apple'*) (Kawase, Papadakis, & Abel, 2011; Kim, Roczniak, Lévy, & El Saddik, 2012);
- *singularity* versus *plurality* (*'cat'* versus *'cats'*) (Z. K. Zhang et al., 2011);
- different formatting conventions, e.g. the tags *'employee-meeting'* and *'emplMeeting'* (Z. K. Zhang et al., 2011), and;
- different levels of tag aggregation: some users may prefer to tag content about trees with the tag *'tree'*, while others may use *'Red Oak tree'* because the content specifically concerns Red Oak trees.

Abel et al. (2010) use the DBPedia ontology to automatically enhance tags in their *TagMe* system with context. For example, if a user tags a piece of content with *'Amsterdam Central Station'*, TagMe automatically derives that it concerns a building, and more specifically, a train station located in the Netherlands.

The problem of word sense disambiguation is even more apparent in multilingual systems. The Italian word *'penna'*, for example, translates into English as *'feather'*, *'pen'* or *'author'*, depending on its context (Navigli, 2009).

Automatically categorizing or classifying tags seems like a useful way of countering tag ambiguity. However, the tools available for accomplishing tag sense disambiguation are knowledge intensive and require either ontologies, taxonomies or trained machine learning (Navigli, 2009). In chapter 2, we determined that EI systems in ESM need to be *unsupervised if possible*, and that *categories or subjects of expertise must not be predefined, or at least not limiting to users*. The methods available to perform automatic categorization or classification of tags are far from unsupervised, as they require a lot of manual training, updating and other manual supervision. Moreover, these methods are not language-independent and require a lot of computer resources (Navigli, 2009). Therefore, we believe that EI in ESM should allow for an emergent categorization scheme, as we discussed at the start of section 5-5. Until convergence of the tag collection sets in, tags can be expected to be messy and ambiguous. We ignore this initial ambiguity by assuming[13] clean and well formatted tags. Future research should point out the best way of coping with tag ambiguity before convergence, for the case of EI in ESM. Ontologies and taxonomies may still be used to do non-committal user recommendations. We discuss such use in the next section.

Suggestions and Convergence

With users being highly susceptible to tag suggestions, automated tag suggestions can be used to speed up tag convergence (Halpin et al., 2007; W. T. Fu, Kannampallil, Kang, & He, 2010; Suchanek et al., 2008). If users are suggested tags used by others, they tend to tag more accurately. These suggestions may, for example, be provided on the basis of popularity or co-occurrence with other tags. A prerequisite for tag convergence is a large number of users and tags, so that the tag vocabulary eventually stabilizes to an agreed upon folksonomy with a relatively small number of popular and meaningful tags and a long-tail containing a large

[13]See appendix A for an overview of this and other assumptions.

number of idiosyncratic tags (Halpin et al., 2007). As we discussed in section 5-5, ontologies may be used for tag suggestions in ESM. However, as tag collections tend to converge as they become larger in terms of size and use, we argue that simple co-occurrence-based suggestions are preferable. Especially considering the multilingual, multicultural and domain-independent content and tags in ESM systems, as we discussed in section 3-5.

5-6-3 Top Contributors

With both content and tags, it is common for a small number of enthusiastic users to contribute the most (sometimes over 40%) (Holley, 2010; Halpin et al., 2007). If expertise is estimated on the basis of some term frequency measure, it is likely that these top contributors will be rated as experts more quickly than other users. An EI system should account for this system behavior, for instance by taking into account freshness and decay of content and tags.

Moreover, the number of times a tag (and a term in general) is used in content does not necessarily imply the degree of the author's expertise (Balog, 2008; Yeung, Noll, Gibbins, Meinel, & Shadbolt, 2011). Some topics, for instance, engender more opinion than facts. Others are simply more commonly subjects of discussion and work flow. Moreover, an ESM system can be expected to lack information about employee performance and expertise development in the past (Balog, 2008).

As we discussed earlier, we assume the knowledge in an ESM system is representative for the knowledge in the client organization, see also appendix A-1. Our reasoning is, that it is in fact useful to have users, that participate more fully in knowledge sharing, displayed more prominently. In section 2-2 we discussed that in order to *incentivize employees to make use of the ESM for their corporate activities*, they need to experience direct return on their invested time. The fact that users will benefit more from EI in their ESM environment as they participate more actively, may serve as such direct feedback and may incentivize users to participate.

5-6-4 Freshness and Decay

Users that have recently contributed content with respect to a subject of expertise, can be expected to be more knowledgeable on that subject than users that have been active in the past but not recently (Raj et al., 2011). This issue of freshness and decay of expertise is common in EI. Although the issue is obvious, the *rate* of decay is unclear. Besides decaying, expertise may also become *outdated* with time. An example is expertise with respect to building car engines: in 30 years, if all cars run on non-fossil fuel engines, that expertise can be expected to be largely outdated. Expertise on computer software, for example, may become outdated even faster.

Although we recognize this issue, we will not investigate the *rate* of decay, because of the scope and time constraints of this thesis. Instead, we use an expertise half-life of *5 years*, leaving the matter of more dynamically determining and accounting for expertise decay in ESM to future research.

5-6-5 Cold-Start Problem

If the system only contains little content and a small number of tags, the identification of experts is likely to be very ineffective due to the so-called cold-start problem (Z. K. Zhang, Liu, Zhang, & Zhou, 2010; Milicevic et al., 2010; Z. K. Zhang et al., 2011; Abel et al., 2011). And even as the system is gradually filled with content, expertise from the past may still remain absent. Past work on old company projects, for example, may not find its way into the ESM system.

A way to counter this problem is by importing existing information on users and the organization into the system. Importing users' LinkedIN profiles, for instance, can provide some preliminary insight into current and past user expertise. Importing e-mail messages can help to include knowledge that was exchanged in the past. In the case of tags, some research reports that importing user tags from public social media can help overcome the cold-start problem. Abel et al. (2011) show promising results by using tags from public social media for tag suggestions in a new environment. Moreover, as we discussed in section 5-2, expertise identification can also be enhanced by importing external web data on candidate experts.

Importing such external data is often a time-consuming task, because it may not be compatible with the data in the ESM system, and it may be created using a different vocabulary dictated by the external system or by the type of activities in that system (Abel et al., 2011). For these reasons, we will leave the enrichment of User Expertise Profiles with external web data out of scope in the remainder of this thesis.

5-7 Exploiting the Social Graph

So far, we have discussed full-content and CF abstraction data as sources for EI. Research on EI recognizes the exploitation of the social graph in a digital environment to be a good source for EI as well. Using the social graph of a client organization – the network of employees and their relationships to each other – user-oriented EI is possible, accounting for the differences between users and each user's personal view on useful recommendation of experts. Figure 5-1 shows a social graph depicting an individual's linkedIN network. Every dot represents a person in the user's linkedIN network. People are grouped according to affiliation, employment and other commonalities.

Smirnova et al. (2011) argue that the ranking of experts must take into account the time needed for a user to contact each expert, based on social distance, their position in the firm and the people they know. It is also known that people are more comfortable approaching physically similar experts, based on race, gender, sex and age (Becerra-Fernandez, 2006). Moreover, Fu et al. (2007) argue that expertise propagation should be taken into account, appointing extra relevance to experts that are closely connected to other experts.

While we expect the personalization of expert recommendation to be of great importance for the future development of EI, in this thesis we focus on determining the *basic* characteristics of EI in ESM systems. We consider user personalization to be one of the next steps. Therefore, we leave user personalization to future research. Furthermore, we consider expertise propagation as an important factor in EI in ESM. Implementing such propagation, however, requires additional research on the requirements the organizational nature of ESM poses on

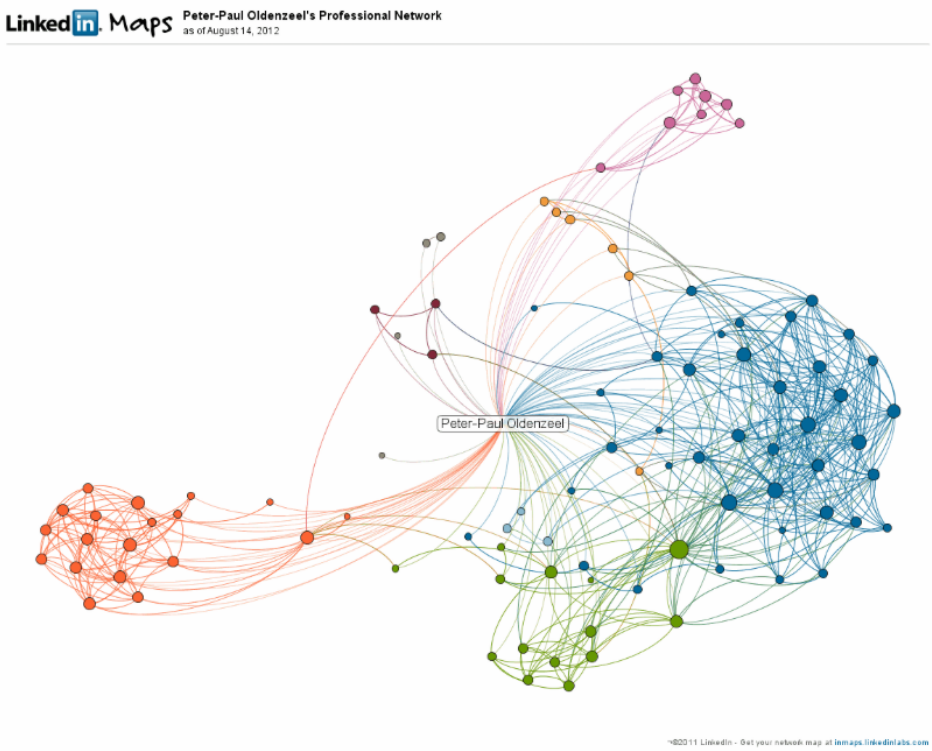


Figure 5-1: A social graph of a user's linkedIN network.

it. How should we, for instance, measure the distance between an employee in one department and an employee in another? Future research should extract lessons on expertise propagation from public social media and fit these onto ESM in order to come to solid metrics.

5-8 Collaborative Filtering Appreciation Data

In order to determine *valid* expertise as explained in section 4-2, EI also has to determine the *confidence and authority* other employees place in an employee's knowledge. We suggest that CF appreciation data can be used to do so. However, it is fairly easy for users to abuse the CF instruments in place. As we have seen in chapter 3, ESM environments typically support a combination of *thanking*, *voting*, *following* and *liking*. If a group of employees decides they want to promote a group member's expertise with respect to a topic X, they can systematically thank, up-vote, follow and like that group member's content in order to boost expertise scores. By taking into account the social distance between employees, an EI system can demote appreciation data that is cast between employees in close organizational proximity (Yeung et al., 2011). This demotion can be set to increase with the frequency of voting, thanking or other appreciative actions. Equivalently, appreciation data that is cast between employees with great social distance could be promoted.

Although countering abuse of CF appreciation techniques seems very important for a production system, we leave it out of scope in the remainder of this thesis. Research by Yeung et al. (2011) describes state of the art methods to prevent abuse, which can be implemented

in any EI system.

We know of only few methods for EI that exploit appreciation data. One example is taken from a paper by Raj et al. (2011) on EI in a Q&A environment. Expertise $E_{i,j}$ of a user i in a set of predetermined question categories j is denoted as the sum of that user's *"subject matter expertise"* ($U_{i,j}$) and *"propensity to answer"* ($H_{i,j}$) (Raj et al., 2011):

$$E_{i,j} \equiv \frac{U_{i,j} + H_{i,j}}{2}. \tag{5-3}$$

A user's subject matter expertise is estimated by the votes and best-answer-markings of his answers and questions. Propensity to answer is estimated by the answers the user has provided to questions.

We have described several instruments that can be used to prevent abuse of CF appreciation techniques by malicious users, and we conclude that there exist few methods for EI that utilize CF appreciation data. This is probably because of the absence of CF appreciation techniques in many existing digital environments for knowledge sharing. Furthermore, it can be expected that part of industries' methods for involving CF appreciation data in EI is kept secret (Becerra-Fernandez, 2006).

5-9 Concluding: Design Choices

In this chapter, we have discussed a number of characteristics and issues of existing methods and systems for EI. Now, we will aggregate and consolidate our findings from previous chapters and this chapter to complement our design requirements and choices for EI in ESM from chapters 2, 3 and 4, with a number of design choices. Table 5-1 at the end of this section, provides a complete overview of all design requirements and choices established in Part 1 of this thesis. In Part 2, we apply our design choices to the technical design of our EI prototype.

5-9-1 Hybrid Approach

By exploring the winning EI methods from TREC 2006, 2007 and 2008, we conclude that best practices for EI are often combinations of candidate- and document-model approaches. Combining the models enables us to find a balance between the real-time capacity load of the document-model and the data redundancy of the candidate-model. We determined that both models conventionally revolve around content relevance. As we discussed in chapter 4, EI in ESM systems should estimate *valid* expertise by estimating content *relevance*, as well as the *confidence* and *authority* others place in that content. Consequently, only taking into account content relevance is not enough. Therefore, we argue that EI in ESM should be based on a hybrid approach, combining the candidate-model, document-model, and CF appreciation data. This is design choice 5. That way, all three components of valid expertise can be estimated. Figure 5-2 illustrates this hybrid approach, ranging from data sources (in the lower part of the figure) to components of valid expertise (in the upper part of the figure).

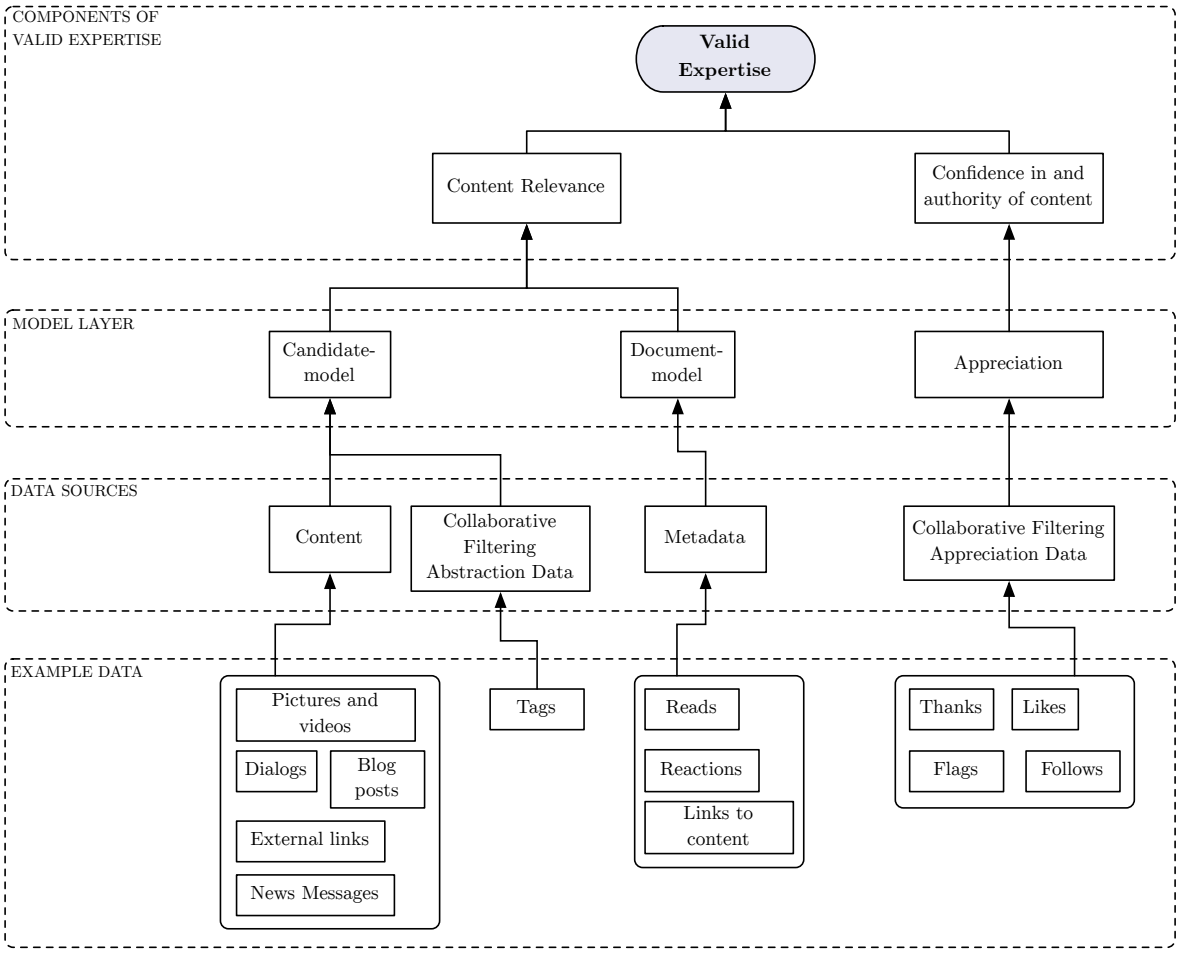


Figure 5-2: Illustration of our proposed hybrid approach to Expertise Identification in Enterprise Social Media, combining the candidate-model and document-model approaches with Collaborative Filtering appreciation data.

5-9-2 Using Collaborative Filtering to Estimate Confidence and Authority

With only few existing methods for EI utilizing CF appreciation techniques, use of these techniques for estimating *confidence* and *authority* of a user's knowledge needs to be trialled for the case of ESM systems.

5-9-3 Full-content or Content Abstraction Data

In section 5-6, we elaborately considered full-content and abstraction data as sources for EI in ESM systems.

Data-mining full content can be difficult in ESM because of the diverse, multilingual, multicultural and domain-independent content. When using tags for EI, users are voluntarily performing part of the abstraction task. However, tag collections are far more ambiguous than content.

Methods for tag disambiguation are language-dependent, require a lot of computer resources, and require manual supervision. Convergence of the resulting tag collection can be achieved through a large number of active users and tags, as well as helpful tag suggestions. Until convergence of the tag collection sets in, tags can be expected to be messy and ambiguous. We ignore this initial ambiguity by assuming[14] clean and well formatted tags. Future research should point out the best way of coping with tag ambiguity before convergence of the tag collection, for the case of EI in ESM.

The sources for content relevance estimation in EI – full-content and tags – have to be trialled for the case of ESM in order to determine which single source or combination of sources can best be used. Therefore, we establish design choice 2. As discussed in section 2-2, EI systems are typically found as components of ESM systems. Consequently, we argue that the choice to base an EI system on either content, tags, or some combination of the two, depends on the quality and usefulness of these sources in determining content relevance. If the EI system needs to be integrated into an ESM system in which tags are used sparsely, for example, content relevance estimation should be based on content alone. If an ESM system tags messages automatically, with few adjustments by users, using tags as a source for EI might even be redundant, not adding any value.

5-9-4 Conventional Content Relevance

The existing methods for EI discussed in this chapter all use conventional methods from the field of Information Retrieval for determining the relevance of content to a query. This is not surprising, since Information Retrieval techniques for determining query-content similarity have matured through the extensive use and development of search engines. We argue that EI in ESM should also reuse conventional techniques for estimating content relevance of a source. Therefore, we adopt the use of conventional relevance scoring methods as design choice 6.

[14] See appendix A for an overview of this and other assumptions.

5-9-5 Largely Unsupervised

We discussed that EI in ESM systems should take place largely unsupervised, with the exception of *administrative maintenance and development*, *user feedback* and *CF data provided by users*. Consequently, this is reflected in design choice 10.

5-9-6 Folksonomies Over Ontologies and Taxonomies

We considered the trade-offs between ontologies and taxonomies on the one hand, and folksonomies on the other, concluding that EI in ESM systems should advocate folksonomies, enabling users to freely categorize content. Moreover, folksonomies can be expected to provide more accurate content abstraction data than taxonomies can, because they reflect the information structures and relationships that people actually use, instead of ones that were planned for them in advance. Therefore, we adopt the use of folksonomies over ontologies and taxonomies as design choice 10. Ontologies and taxonomies may, however, still be used to provide synonyms (e.g. using WordNet), suggest alternative naming and provide meaningful context or background (e.g. using DBPedia).

5-9-7 Expertise Decay and Degradation

In subsection 5-6-4, we discussed the freshness and decay of expertise in ESM systems. Determining how to account for freshness and decay of expertise requires much additional research, so we use an expertise half-life period of 5 years, as reflected by design choice 7.

5-9-8 Out of Scope

In this section, we describe a number of aspects of EI we have decided to leave out of scope in this thesis, and consider in future work.

Files

In section 5-6-1, we decided to leave files out of scope. It is often very difficult to determine the authors and abstracted content of files, e.g. documents and images, which makes accounting for files in EI into a big task. Therefore, we solely focus on content other than files. Examples of such content are dialogs, news messages, blog posts et cetera. Authors of such content in ESM are easily determined because users have to be logged into the system in order to publish any content. Future research should investigate whether files can offer added value in performing EI in ESM.

User Personalization

While recognizing the potential benefits of user personalization to EI systems, we also leave user personalization out of scope. User personalization poses a dinstinct research task, touching upon aspects from research fields like User Modeling and Behavioral studies, which should be performed by future research.

Expertise Propagation

Furthermore, we consider expertise propagation as an important factor in EI in ESM. Implementing this propagation, however, requires additional research into the requirements the organizational nature of ESM poses on it. Future research should extract lessons on expertise propagation from public social media and fit these onto ESM in order to come to solid metrics.

Preventing Abuse of Collaborative Filtering Data

Although countering abuse of CF appreciation techniques seems very important for a production system, we leave it out of scope in the remainder of this thesis. Research by Yeung et al. (2011) describes state of the art methods to prevent such abuse, which can be implemented in any EI system. Future research should investigate the use of these methods for EI in ESM systems.

Tag Disambiguation

By ignoring initial tag ambiguity before convergence of the tag collection, we leave tag disambiguation out of scope. As discussed in subsection 5-9-3, disambiguation of tags is language-dependent, and requires a lot of computer resources, as well as manual supervision. One way to cope with this initial tag ambiguity is not using tags for ESM functionality other than tag suggestions. Our tests in Part 2 of this thesis investigate the use of full-content and tags as means of estimating content relevance. If it is desirable to use both tags *and* full-content as sources for estimating content relevance with respect to a query, the system could also be configured to only start using tags for EI after a while. Future research should point out the period of time after which tag convergence occurs, as well as the best way to disambiguate tags before convergence.

Enriching User Expertise Profiles with External Web Data

In section 5-6-5, we discussed that the initial absence of content and tags in an ESM system, also called the cold-start problem, can be countered by importing existing information about users and the organization into the system. Moreover, as discussed in section 5-2, two out of three winning TREC EI methods use additional techniques for the inclusion of external web data on participants, further enhancing precision and accuracy. Compatibility issues between the external data and the data in ESM systems, however, make enriching User Expertise Profiles with such external web data into a time-consuming task. For the sake of time constraints, we leave it out of scope in the remainder of this thesis. Future research on EI in ESM should point out whether this enrichment step indeed offers enhanced precision, accuracy, and possibly a remedy for the cold-start problem.

Design Choices

Here, in table 5-1, we present the aggregated design requirements and choices from our literature study in Part 1 of this thesis.

	Design Requirement		Design Choice
1.	Estimate valid expertise.	1.	Suggest likely experts rather than *claim* or *state* the absolute experts on a particular subject.
		2.	The validity of an employee's expertise comprises of the *relevance*, *confidence* and *authority* placed in that employee's knowledge regarding a subject.
		3.	Knowledge relevance can be estimated by means of a selection of full-content and CF abstraction data in ESM (see subsection 5-9-3).
		4.	Confidence and authority other employees place in knowledge can be estimated by means of CF appreciation data in ESM.
		5.	Hybrid approach (see subsection 5-9-1).
		6.	Conventional content relevance estimation methods (see subsection 5-9-4).
		7.	Expertise decay (see subsection 5-9-7).
2.	Easy to use.	8.	Automatically extract User Expertise Profiles.
3.	Ensure free categorization and structuring of content.		
4.	Support for domain-independent, multilingual and multicultural content and users.	9.	Folksonomies over ontologies and taxonomies (see subsection 5-9-6).
5.	Support for dialog type diversity.		
6.	Ensure instant scalability of hardware and software.	10.	Largely unsupervised EI (see subsection 5-9-5).

Table 5-1: Overview of the design choices for Expertise Identification in Enterprise Social Media systems, completed in chapter 5.

Chapter 6

Conclusion Part 1

In the introduction of this thesis, we set out to answer the following main research question:

What form of Expertise Identification is most suitable for Enterprise Social Media, accounting for its social-media-like content and Collaborative Filtering characteristics?

Now Part 1 is complete, we can answer sub-research questions 1, 2 and 3.

6-1 Answering Research Question 1

1. *How is Expertise Identification in Enterprise Social Media systems different from that in other digital environments for knowledge sharing and which requirements do these differences pose?*
   (a) *How do Enterprise Social Media systems differ from other digital environments for knowledge sharing?*
   (b) *What requirements do these differences pose on Expertise Identification in Enterprise Social Media systems?*

In contrast with other, more conventional, environments for knowledge sharing, we found that the main differentiators of Enterprise Social Media (ESM) systems are their integrated social-media-like content and the Collaborative Filtering (CF) techniques available for users to enrich their and others' content with abstraction and appreciation data. ESM systems are typically focussed on unstructured communication rather than facilitating the finding and sharing of expertise, while the latter is one of the main reasons for companies to adopt ESM systems.

In chapter 2, we concluded that conventional Knowledge Management (KM) systems primarily facilitate document storage and often support several types of organizational communication (e.g. news messages). Conventional digital environments for knowledge sharing assume corporate communication takes place via e-mail and require redundant knowledge sharing

via the environment. ESM systems integrate that corporate communication by means of social-media-like communication. As they thrive by this low-threshold communication, it is important that Expertise Identification (EI) is performed *largely unsupervised*, that it *incentivizes employees to make use of the system for their corporate activities* and that *the categories of expertise are not predefined or at least not limiting to users.*

ESM systems are typically available as SaaS products in the cloud, containing data of multiple client organizations. Because ESM client organizations are geographically dispersed with a lot of mobile employees, those data are typically multilingual, multicultural and domain-independent. The SaaS delivery model makes it difficult to tailor ESM to the wishes and specifics of a client organization, hence ESM systems should take into account client and content diversity.

ESM systems require many users and a lot of system activity in order to build up comprehensive content. Consequently, ESM systems require instantly scalable hardware and software.

From a user-oriented perspective, ESM systems should be *easy to use* and should *not impose on users any preconceived notions about how work should proceed or how output should be categorized or structured.* Moreover, as we discuss in chapter 3, ESM systems typically support *social networking*, *computer-mediated communication*, and *allow social feedback between users.*

ESM systems are typically equipped with a range of CF techniques to facilitate users in providing abstraction and appreciation data to content. In chapter 4, we emphasize the necessity of CF data in estimating the *confidence and authority* others place in an employee's expertise.

As discussed in chapter 2, EI systems are typically found as components of a variety of digital environments for knowledge sharing. Consequently, when implemented in an ESM system, an EI system needs to facilitate all of the aforementioned requirements on ESM systems.

An overview of the resulting design requirements ESM systems pose on EI can be found in table 3-3.

6-2 Answering Research Question 2

2. *How can expertise be defined and quantified in the context of an Enterprise Social Media system?*

As we discussed in chapter 4, we posit that the *validity* of an employee's expertise with respect to a subject comprises of the *relevance* of his or her knowledge with respect to the subject, and the *confidence and authority* that other employees place in that knowledge.

Consequently, we define organization's experts as follows:

Experts are those employees that are, within the boundaries of the organization, confided in and authorized by others to possess the greatest amount of relevant facts, information and/or practical abilities with respect to a particular subject.

Considering the ambiguity of knowledge and the fact that only explicit, externalized knowledge can truly be stored in the form of posts and comments in ESM systems, we argue that most of the knowledge present in an organization cannot be stored in an ESM environment.

The latter can be expected to capture more of the tacit and procedural knowledge that resides in employees' minds than legacy KM systems do, but in order to share truly tacit knowledge, employees will still have to consult each other in real life.

Nevertheless, employees with great explicit declarative knowledge in a particular field are likely to possess tacit procedural knowledge in that field as well, gained through experience and education. Because of this, we believe that we can point out *likely* subject matter experts based on data available in ESM. Therefore, EI in ESM should *suggest* likely experts rather than *claim* or *state* the absolute experts on a particular subject.

We believe that the validity of an employee's expertise comprises of the *relevance*, *confidence* and *authority* placed in that employee's knowledge with respect to the subject at hand. Content in ESM can be used to estimate the relevance of an employee's knowledge with respect to a topic, whereas CF appreciation data can be used to gauge the confidence and authority other employees place in that knowledge. An employee's knowledge and expertise are clearly subjective to other employees' perceptions of that knowledge and expertise.

So far, we have conceptualized how *valid* expertise in ESM should be *estimated* by analyzing users' content and CF data. In our case study in Part 2 of this thesis, we will further quantify this approach. Therefore, in chapter 11, we complete the answer to research question 2.

Table 4-1 lists the design requirements and design choices resulting from our answers to research questions 1 and 2.

6-3 Answering Research Question 3

3. *Which best practices from existing Expertise Identification systems can be reused in ESM?*

   (a) *How do existing Expertise Identification systems estimate expertise?*

   (b) *How can Collaborative Filtering data contribute to Expertise Identification?*

Existing methods and systems for EI, as we discussed in chapter 5, estimate a user's expertise by estimating the relevance of his or her *content* with respect to the subject at hand. This is not surprising, since conventional digital environments for knowledge sharing primarily store users' content. However, this conventional approach ignores the *confidence* and *authority* components of *valid* expertise. CF Appreciation data should be used to estimate *confidence* and *authority* other employees place in an author's expertise regarding a subject.

EI in ESM systems should take place largely unsupervised, with the exception of *administrative maintenance and development*, *user feedback* and *CF data provided by users*.

We considered the trade-offs between ontologies and taxonomies versus folksonomies, concluding that EI in ESM systems should use folksonomies, enabling users to freely categorize content. Moreover, folksonomies can be expected to provide more accurate content abstraction data than taxonomies can, because they reflect the information structures and relationships that people actually use, instead of the ones that were planned for them in advance. Ontologies and taxonomies may, however, still be used to provide synonyms (e.g. using WordNet),

and suggest alternative naming and provide meaningful context or background (e.g. using DBPedia).

When using tags for EI, users are voluntarily performing part of the abstraction task. However, tag collections are far more ambiguous than content. Methods for tag disambiguation are language-dependent, require a lot of computer resources, as well as manual supervision. Consequently, we argue that EI in ESM should allow for an emergent categorization scheme: a folksonomy. The convergence of tag collections requires a large number of users and tags, as well as helpful tag suggestions. Until convergence of the tag collection, tags can be expected to be messy and ambiguous. We ignore this initial ambiguity by assuming clean and well formatted tags. Future research should point out the best way of coping with tag ambiguity before convergence, for the case of EI in ESM.

With users being highly susceptible to tag suggestions, automated tag suggestions can be used to stimulate tag convergence. A prerequisite for tag convergence is a large number of users and tags, so that the tag vocabulary eventually stabilizes to an agreed upon folksonomy. Accordingly, we argue that simple co-occurrence-based suggestions are preferable.

Furthermore, we argue that the top contributors of content and tags in ESM systems benefit more from EI than less active users do, causing direct return for user participation in EI. Additionally, the initial absence of content and tags in an ESM system, also called the cold-start problem, may be countered by importing existing information on users and their organization into the system.

An overview of all the design requirements and design choices resulting from research questions 1, 2 and 3, can be found in table 5-1.

6-4 Case Study

Based on the answers to research questions 1, 2 and 3, discussed in Part 1, Part 2 of this thesis contains a case study on E-view, an ESM system. Based on this case study, we will complete our answer to research question 2 and answer research question 4, displayed below.

4. *How can Expertise Identification be implemented in Enterprise Social Media systems?*
    (a) *How can Expertise Identification be implemented in E-view?*
        i. *What kind of ICT architecture is required to facilitate the identification of expertise in E-view?*
    (b) *Which form of Expertise Identification is most suitable for E-view?*

First, chapter 7 describes the background of our case study. In chapter 8, we present the technical design of our EI prototype, as well as its implementation in E-view. Then, in chapters 9 and 10, we describe the setup and results of the tests we performed in order to gain preliminary insights into performance of the prototype. Chapters 11 through 13 contain conclusions, recommendations, limitations, future work and our reflection on the process and product of this thesis.

PART 2

Chapter 7

Background Case Study

In this chapter, we describe the background of our case study on Expertise Identification (EI) in E-view, a live Enterprise Social Media (ESM) system. First, we discuss E-view's origins and its congruence with ESM systems we have discussed in part 1 of this thesis. Then, we explore the content and Collaborative Filtering (CF) techniques present in E-view. Next, we discuss a number of typical workflows in E-view that benefit from EI. Finally, we reveal design constraints that E-view imposes on EI.

7-1 TJELP and E-view

In the summer of 2011, TJELP, a consultancy and IT start-up located in Amsterdam, set out to develop *Expert View*, commonly abbreviated to E-view[15]. With E-view, TJELP aims to fill the gap between organizations' demands on ESM systems and existing ESM systems, as well as other digital environments for knowledge sharing. E-view's main purpose is to support transparent and intuitive communication, based on the success of social-media-like dialogs. In contrast with several other ESM systems, these dialogs are not designed to support a single organizational task such as *idea management* or *Q&A*. Instead, E-view is designed to support all-round organizational communication. As such, it is difficult to categorize E-view as one of the ESM implementation types from table 3-1. We argue, however, that in its current form, E-view best fits in the *Enterprise Social Networking* category.

By the time this thesis is completed at the end of September, 2012, E-view is in beta phase. By September, 2012, it supported all basic functionalities, including, among others, creating a dialog and adding participants, files and tags. However, as discussed, the success of ESM depends on the explicit facilitation of sharing knowledge and expertise. Moreover, TJELP's change management consultancy vision strongly advocates the philosophy of explicitly connecting employees that posses knowledge relevant to an inquiry. Consequently, E-view should be equipped with EI functionality.

[15]For more detailed insight into E-view, we suggest visiting their website: http://e-view.com. In order to experience E-view, interested parties are free to sign up for a free account at the same address.

By September, 2012, E-view was able to calculate the top three experts given a query, e.g. "HRM registration", by counting the number of times users occur in individual search results returned by that query. This approach is in fact a simplistic implementation of the document-model approach explained in chapter 5. The resulting top three of experts may be invalid. For instance, when an overly active user comments on all posts in E-view with "Good one!", without actually contributing to the content matter.

7-2 Content

Analogous to findings in chapter 3 on the content of ESM systems, the content in E-view comprises of social-media-like dialogs between users or groups of users. Figure 7-1 illustrates a dialog in E-view. The contents of a dialog are made up of the *title*, the *body* of the main post, comments, and comments on comments. File attachments can also be considered content. However, as discussed in chapter 5, we leave files out of scope.

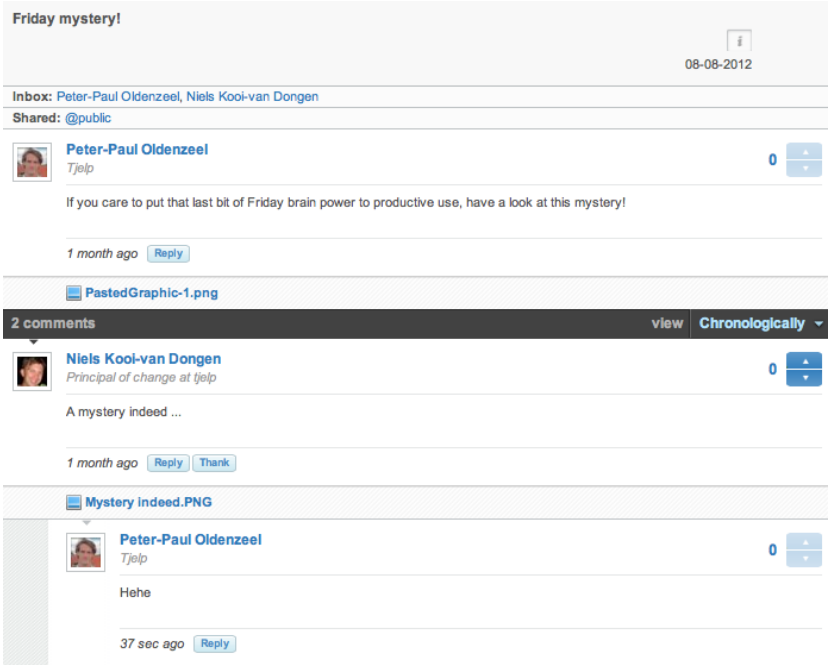


Figure 7-1: A dialog in E-view.

7-3 Collaborative Filtering Techniques

E-view supports a number of CF techniques for both the abstraction and appreciation of content. Table 7-1 depicts which CF data users can provide for different types of posts, tags and users. EI should take into account the availability of CF appreciation and abstraction data for each technique.

Collaborative Filtering Data	Main Post	Comment	Tag	User
Up- and down-vote	✓	✓	✗	✗
Thank You	✗	✓	✓	✗
Following	✗	✗	✓	✓
Favorite	✓	✗	✗	✗
Tag	✓	✓	✗	✗

Table 7-1: Overview of the Collaborative Filtering data users can provide for different types of content, users and tags in E-view.

7-3-1 Abstraction Techniques

In order to get users to provide E-view with CF abstraction data, tagging is supported. By September, 2012, tagging was not mandatory. There was no taxonomy or ontology dictating choice, categorization or classification of tags. E-view suggests auto-completed tags to users, but does not provide tag suggestions based on tag co-occurrence. When E-view is released by the beginning of October, 2012, however, it will contain more advanced tag suggestion techniques that are expected to help users converge on tag use. Figure 7-2 shows the act of adding a tag in E-view and receiving auto-completed suggestions.

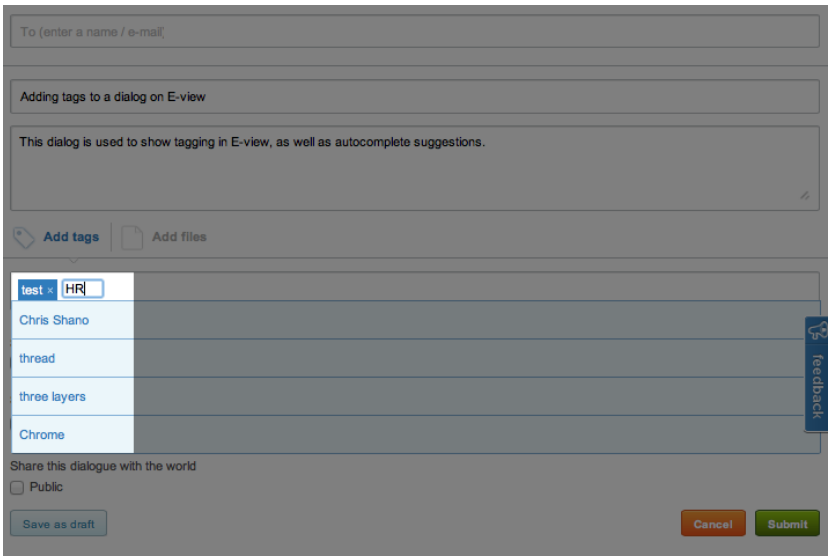


Figure 7-2: Tags and autocompleted suggestions in E-view.

7-3-2 Appreciation techniques

E-view supports a number of techniques to collect CF appreciation data:

- *up- and down-voting* of other users' posts (dialogs and comments);
- flagging dialogs as *favorite*;

- *thanking* other users for their comments, and;
- *following* tags or users.

Figure 7-3 illustrates these techniques within a dialog in E-view.

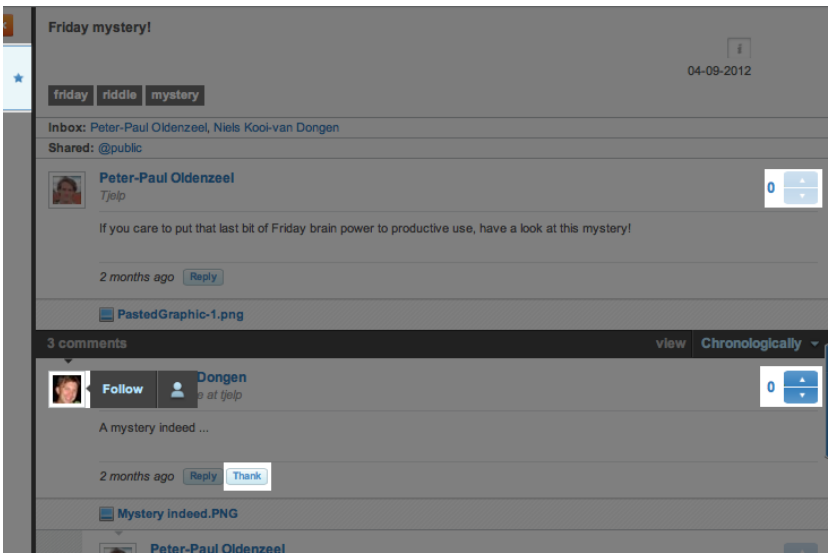


Figure 7-3: Techniques for collecting Collaborative Filtering appreciation data in E-view.

7-3-3 Metadata

E-view stores a number of metadata that can prove trivial in the identification of users' subject matter expertise. These metadata are listed in table 7-2.

	Type of Metadata	Description
1.	Timestamps	The date and time content or CF data was created and/or edited.
2.	@Mentions	Using the prefix '@', while writing a main post or comment, enables users to link to other users.
3.	Reads	E-view stores whether a user has read a post. At the time of this research, however, E-view did not track the total number of reads from the perspective of a main post or comment.

Table 7-2: Metadata available in E-view.

Read counts can contain information about the popularity of certain content: if a piece of content has been read a lot more times than other content, it can be expected to be more valuable as a resource. This is coherent with the PageRank methods from TREC, discussed in section 5-2. Because of the perspective of the read counts and the @mentions pertaining only to users, not content, EI in E-view can currently take into account only timestamp metadata.

7-4 Expertise Identification Workflows

Because E-view supports all-round organizational communication, typical workflow in E-view is unclear. To gain insight into the possibilities of EI functionality in E-view workflows, we describe two workflow scenarios wherein EI can be expected to fulfill a vital role. These workflow scenarios have been constructed in consultation with TJELP employees.

7-4-1 Multidisciplinary Team Composition

An employee may need to compose a multidisciplinary team to take on a new project. EI may provide the employee with leads on which employees to review for open team positions, based on the queried team's purpose and the resulting list of suggested experts. The composing employee can then review the experts' track records within the company to decide on team composition.

7-4-2 Finding a Knowledgeable Colleague

If an employee is in search of information to solve an urgent problem, the right information may not always be present as (part of) content in E-view. Searching for *'how to register an employee'*, for instance, may not return any sensible results. Moreover, in large organizations, different people, teams and departments may choose to store certain information in E-view, while leaving out other information. It goes without saying that some information can be expected to be absent altogether because there have been no dialogs or questions on that subject. Lastly, a portion of company knowledge may be stored as tacit knowledge only, in the minds of its employees.

In these cases, the employee in search of information will not be able to find any. If no search results are returned when searching for the direct query, it is sensible to try a more indirect query. The employee might search for *'HRM'* or *'new employee'*. These queries will probably return search results pertaining to situations different from registering a new employee. Typically, the employee could qualitatively select the best search results and contact their respective authors.

Using EI, E-view can facilitate employees in contacting each other more easily, by suggesting experts to contact in case a query does not return any results. If the query *does* return a set of results, EI can serve to extrapolate the most likely experts, so that the employee does not have to plough through the results one by one to determine who to contact.

7-5 E-view-specific Design Constraints

The characteristics of E-view we discussed in this chapter impose a number of additional constraints on the implementation of EI. Table 7-3 lists these constraints. In the next chapter, we present our technical design of an EI prototype for ESM. Moreover, we explain how we implemented the system in E-view, accounting for the design constraints in table 7-3.

	Design Constraint	Description
1.	Limited CF data.	E-view currently offers limited means of providing CF data (see section 7-3).
2.	Metadata.	Only timestamps can be used for EI (see subsection 7-3-3).

Table 7-3: Overview of additional E-view-specific design constraints on Expertise Identification.

Chapter 8

Technical Design and Implementation

In this chapter, we discuss the technical design of our Expertise Identification (EI) prototype for Enterprise Social Media (ESM). Starting from September, 2012, around the time this thesis research was completed, a version of the EI prototype we present in this chapter was implemented in E-view 1.0 (production version). Our technical design is applicable to both E-view and ESM systems in general, although in order to implement the design in E-view we had to account for a number of additional design constraints laid out in the previous chapter.

Throughout this chapter, we explain our implementation of the design choices for EI in ESM, as established in Part 1 of this thesis. The result is the technical design of our prototype and an implementation in E-view. First, we discuss the hybrid-approach to EI we use, based on the candidate- and document-model approaches we discussed in chapter 5. Moreover, we further quantify expertise in the context of an ESM system. Subsequently, we present the EI process within an ESM system. Next, we discuss the data-structure and software we (re)used and developed to implement our prototype in E-view, as well as the way our implementation handles the design constraints posed in the previous chapter. Finally, we discuss future expansion and development of the prototype.

8-1 Hybrid Approach

Our research into existing EI methods and practice points out that, in the case of ESM systems, the candidate- and document-model approaches to EI are best combined into a hybrid approach. Accordingly, our technical design combines the candidate- and document-model into a hybrid approach. The resulting approach enables us to estimate *valid expertise* by estimating its components: *relevance*, *confidence* and *authority*, as posited in chapter 4.

The candidate-model component of our EI prototype comprises of User Expertise Profile construction, based on all content and tags a user publishes in the ESM system. The novelty of our approach lies in the document-model component of our EI system. We estimate *confidence* and *authority* of users in content, by complementing content relevance scores with Collaborative Filtering (CF) appreciation scores. Our prototype estimates this *confidence* and

authority by evaluating the appreciation data for every piece of content in a User Expertise Profile that is relevant to the query.

Besides content-specific appreciation data, we believe that each user's *overall reputation* should also contribute to the estimation of a user's confidence and authority. Although many ESM and public Social Media already calculate overall user reputation by accumulating user popularity, contributions, attrition rate and other activity measures, using that reputation in conjunction with estimated user expertise with respect to a query is more difficult. Possible issues are:

- distinguishing between users that are specialized in one or few subjects, and generalist users that contribute to many subjects, and;
- accounting for possible overlap/interaction between content-specific confidence and authority and the author's overall reputation: it can be expected that these are based at least partially on the same CF appreciation data, causing for overlap and feedback loops that distort their combined value.

For these reasons, determining overall reputation comprises a distinct research task. Consequently, we leave it out of scope in the remainder of this thesis. Future research should experiment with overall confidence and authority other users place in a user and that user's reputation to determine its value to EI.

We continuously collect all CF appreciation data in a social graph – a data-structure typically used to store relational information – from which we retrieve appreciation data associated with the content matched by a query in User Expertise Profiles. We go into the details of this social graph in section 8-1-2. In subsection 8-1-1 and 8-1-2, we elaborate on the way our prototype estimates content relevance, confidence and authority.

8-1-1 Estimating Content Relevance

Selecting relevant posts and estimating query-post relevance is an Information Retrieval task. Highly-developed techniques for doing so can be found in numerous search engines. Reusing these techniques, our prototype employs the open-source search engine Apache Solr[16], an expansion of the well-known open-source engine Apache Lucene[17], to select posts and estimate their query-content relevance $R_u(c_i)$, where R represents the relevance score of each selected post c_i for a user u.

We use Solr to index and store all of a user's posts into a User Expertise Profile. The resulting UEP index is stored as a number of segmented documents that reside on a physical disk. In case the index becomes very large, part of the index files (the most recent and most updated files) can be stored in memory instead, making searching against the index even faster. Solr indexes all posts with their content and tags into UEPs, performing stemming (e.g. the word 'computers' is indexed as 'computer' and 'computers') and stop-word removal (e.g. words like 'or' and 'I' are removed) in the process. This way, we can perform partial and wildcard searches on the resulting UEP index. Solr also supports the use of ontologies to disambiguate

[16]For more information on Apache Solr, visit http://apache.org/solr.
[17]For more information on Apache Lucene, visit http://apache.org/lucene.

the indexed content, i.e. find synonyms, typos, polysemous terms, et cetera. However, because we want to ensure our prototype is largely unsupervised and runs fully automatically, we have disabled these features. As explained in chapter 5, we assume that the folksonomy of tags in an ESM system disambiguates full-content, once the tag collection has converged into a stable vocabulary. This leaves the period of use before convergence. During this period, the system will provide tag suggestions based on existing tags and possibly domain dependent ontologies. Future research should examine the use of suggestions based on ontologies, in overcoming the cold-start problem of our prototype as explained in subsection 5-6-5.

While Solr supports a range of similarity measures to calculate query-content relevance, we have chosen to use the standard Solr relevance scoring. This standard scoring metric is based on Lucene's combination of the Boolean and Vector Space models from Information Retrieval (*Lucene Relevancy Scoring*, 2012). The Boolean model is used to narrow down the content in UEPs that needs to be scored given a query, using Boolean logic. Then, the Vector Space model is used to determine scores for the selected content, using Term Frequency - Inversed Profile Frequency (TFIPF)[18]. Accordingly, the following factors are taken into account in determining query-content relevance of pieces of content in a UEP (*SOLR Relevancy Scoring*, 2012):

- term frequency (TF) - the more times a search term appears in a User Expertise Profile, the higher the score;
- inverse profile frequency (IPF) - matches on rarer terms in UEPs score higher than matches on more common terms;
- coordination factor - if there are multiple terms in a query, the more terms that match, the higher the score;
- lengthNorm[19] - matches on a smaller field[20] score higher than matches on a larger field;
- boosting - a query may explicitly boost the contribution of content in one field in Solr over that in another.

Because a match between a query and a post in a User Expertise Profile can occur in different types of content in that post, e.g. the *title* of a dialog or its *body*, it is important that these matches can be weighted separately. As discussed in chapter 5, the winning EI system of the Text REtrieval Conference in 2006 pointed out that such window-based document relevance contributes significantly to system performance. Our prototype indexes different content types in UEPs into different Solr fields. The field *dialog_title*, for example, contains the titles of dialogs the owner of the UEP has created in E-view. Because of this setup, we are able to weight matches between the query and specific content types individually. This can prove to be extremely helpful in determining proper weights for parts of the posts. In order to, for

[18]See chapter 2 for an explanation of TFIDF, the equivalent of TFIPF in the case of documents instead of User Expertise Profiles.

[19]The LengthNorm factor used by Solr resembles the *window-based* occurrence factor used by the winning TREC team in 2006, as described in section 5-2.

[20]A *field* in Solr contains one or several content types from E-view; the field *'dialog_title'*, for instance, contains dialog titles.

instance, weight comments twice as important as dialogs, the weights for the corresponding fields can simply be adjusted.

Figure 8-1 shows an indexed User Expertise Profile, visualized using BaseX[21], a popular XML database manager. The indexed Solr fields are displayed using red headers. Each block inside a field represents an instance of the field type. So each block in the field *dialog_title* represents one title of one dialog.

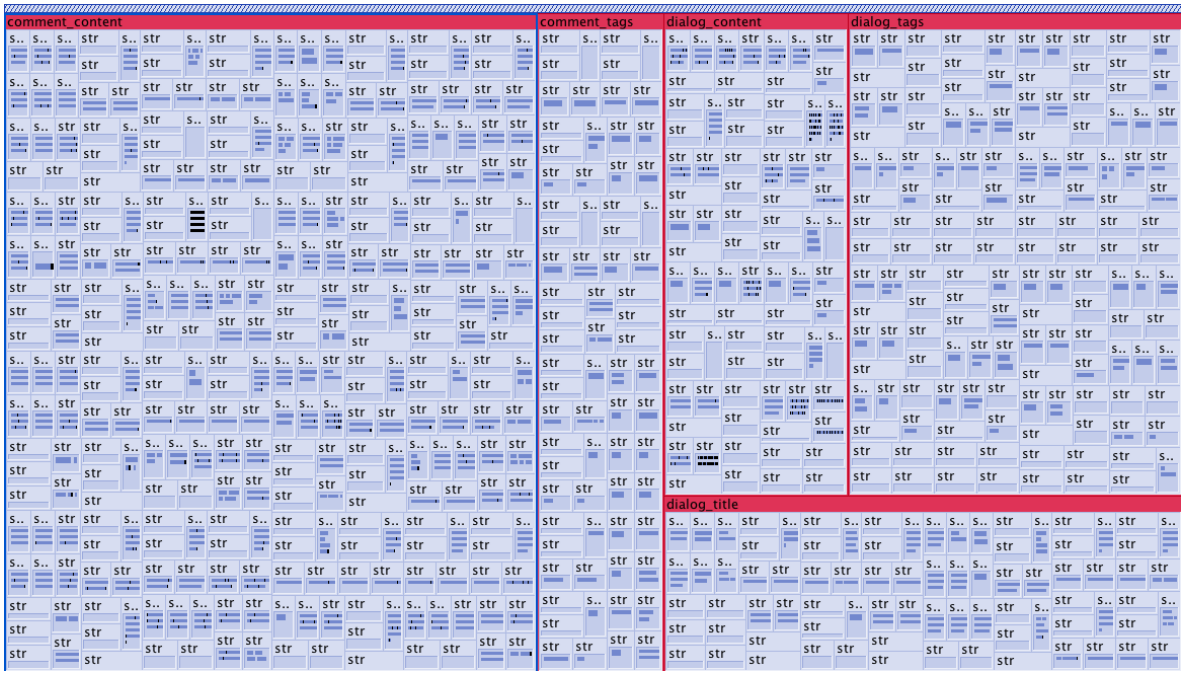


Figure 8-1: An indexed User Expertise Profile in our Expertise Identification system, visualized using BaseX.

As we discussed in chapter 6, it is unknown whether EI in ESM systems should be based on content, tags (abstraction data), or a combination of the two. Because of the modular structure of our EI system, it has the ability to base EI on content, part of that content, tags, or on some combination. In chapter 9, we will test a number of EI strategies in E-view, using the prototype we present here. There, we will test strategies that base EI on full-content, tags, or both.

Concluding Remarks

Summarizing, the candidate-model component of our EI prototype indexes and stores all of a user's posts in a User Expertise Profile index in Solr, enabling the ESM system to select posts from UEPs that are relevant to a query subject. A version of TFIDF is then used to determine relevance scores for the selected posts, by examining the different parts of a post. Using Solr fields, our prototype enables weighting those parts individually. Our prototype only uses simple ontology-free techniques to disambiguate full-content and tags, counting on a converged folksonomy of tags to fulfill this role once the system is actively used. Future

[21] For more information on BaseX, visit http://basex.org.

research should look into the possibilities of using Solr, in combination with ontologies, to provide tag suggestions before the tag collection has converged.

8-1-2 Estimating Content-Specific Confidence and Authority

In order to complement the content relevance of a post with the confidence and authority other users place in that post, we need to estimate appreciation scores for these posts. We estimate the confidence and authority of each relevant post in a User Expertise Profile by performing a real-time look-up of the associated appreciation data in a social graph. As explained in section 5-3, this document-model look-up step of our EI prototype poses a possible bottleneck for the speed at which a ranked list of experts can be provided in response to a query. In order to minimize the capacity and time needed to look up the required appreciation data, we have chosen to store the social graph into a graph-database. Graph-databases are, in contrast with conventional relational databases such as SQL databases, much faster in traversing *'deep'*[22] relationships. Because of the characteristics of ESM discussed in chapter 3, i.e. a high degree of interaction between users and content, ESM systems typically contain a large number of deep relationships. Hence, graph-databases are very suitable for the storage of relationships in ESM. Public social media have in fact driven the increased use of graph-databases worldwide (Weikum, 2007). Graph-databases are often seen as a first step in integrating Database Management and Information Retrieval, because they are more suitable for the execution of Information Retrieval algorithms, which typically require traversing deep relationships.

The graph-database we have constructed consists of a number of vertices (entities) and edges (relationships). Figure 8-2 shows a schema of these vertices, edges, and their respective properties. Entities have only few properties, because all content, names, mail addresses et cetera are stored in the SQL database. The graph-database only contains a mapping of all relationships between users, main posts, comments and tags. Users' *uids*, dialogs' *dids*, comments' *cids* and tags' *tids* are used to reference between the SQL database and the graph-database. By default, all edges and vertices contain a timestamp and a name. We omitted the edge properties from figure 8-2 for the sake of figure readability. Appendix B gives a more elaborate description and visualization of an actual social graph filled with data from E-view, from the perspective of a tag.

Equation 8-1 depicts the look-up of appreciation data and metadata from the social graph, associated with a relevant post in a UEP.

$$A_u(c_i) = \frac{\sum_{j=1}^{n} a_j(c_i)}{max_{k=1}^{n}(a_k(c_i)) - min_{l=1}^{n}(a_l(c_i))}, \tag{8-1}$$

where the *normalized confidence and authority score* $A_u(c_i)$ of a single relevant post c_i of a user u, equals the sum of the j = 1,...,n appreciation data $a_j(c_i)$, where $a_j(c_i) = +1$ for positive appreciation and $a_j(c_i) = -1$ for negative appreciation, divided by the range of appreciation for all users. So, in a setting with 3 users, if a post has received two thank you's (positive appreciation) and 1 down-vote (negative appreciation), $\sum_{j=1}^{n} a_j(c_i)$ equals 1. Assuming that

[22] In this context, *deep* relationships comprise data that would normally require numerous table joins in SQL databases, in order to be extracted.

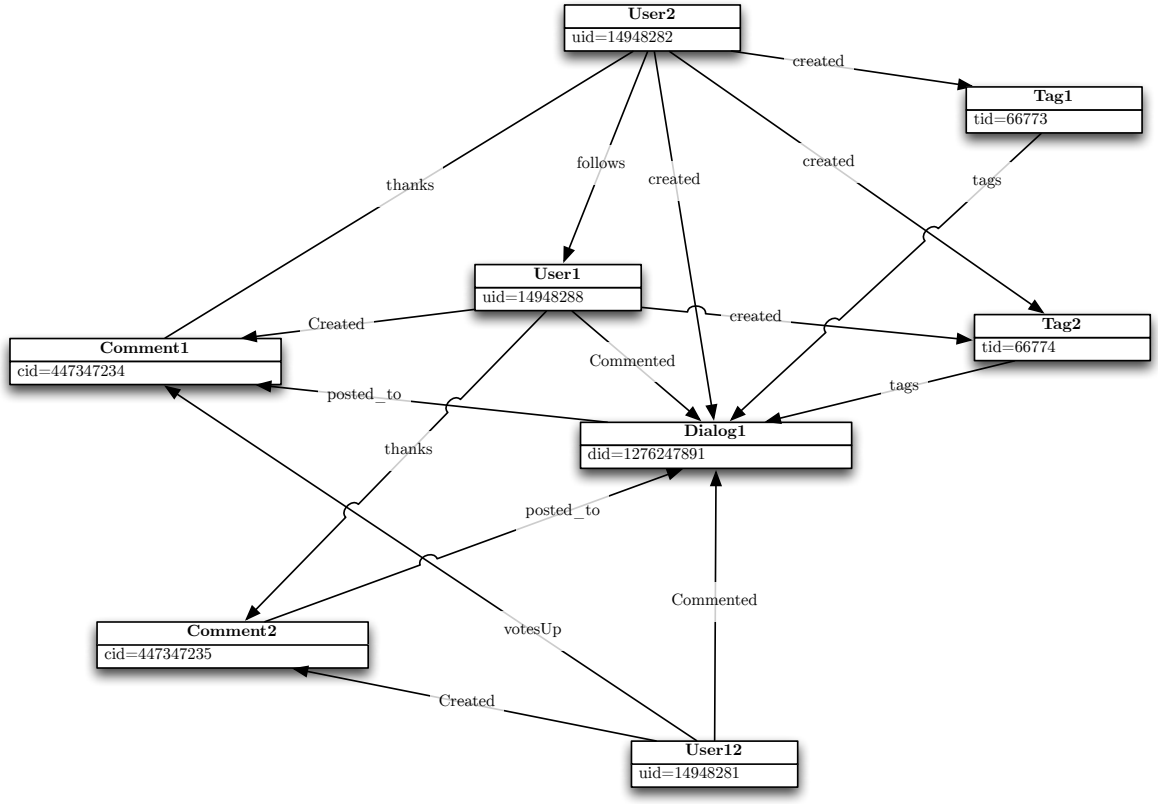


Figure 8-2: Schema of the social graph constructed for our Expertise Identification prototype.

the maximum for all relevant posts of all users equals 6 and the minimum equals 3, then $A_u(c_i)$ is $\frac{1}{3}$.

Equation 8-1 results in a *normalized content-specific confidence and authority factor* ranging between 0 and 1.

8-1-3 Popularity and Decay Factors

From the metadata we collect in the social graph, read counts, share counts and timestamps, we formulate two additional factors: a *popularity* factor and a *decay* factor. The popularity factor is a simple implementation of the PageRank method from TREC, as discussed in section 5-2. It is based on the *read* and *share* counts. The more times a post is read or shared, the higher the popularity factor. Equation 8-2 depicts how this works.

$$P_u(c_i) = 1 + \frac{1}{10} * \frac{r_{c_i} + s_{c_i}}{max_{k=1}^{n}(r_k + s_k) - min_{l=1}^{n}(r_l + s_l)}, \tag{8-2}$$

where $P_u(c_i)$ represents the normalized popularity factor of post c_i by user u, r_{c_i} the read count of a post c_i and s_{c_i} the share count of a post c_i. The resulting factor has a range between 1.0 and 1.1, enabling it to adjust the expertise score by a maximum of 10%. So the most popular post retrieved in response to a query will receive an expertise boost of 10%.

The decay factor implements the expertise decay rate we established in subsection 5-6-4. It uses the timestamp metadata from the social graph, containing the creation dates of posts. Although this gives a fairly decent representation of the age of a message, ideally edit dates should also be taken into account. Because of time constraints on this research, we have not implemented these edit dates as an attribute of the social graph, so we leave them out of scope. Future research should investigate how these edit dates are best taken into account when constructing a decay factor. Equation 8-3 shows the mathematical representation of the factor.

$$D_u(c_i) = (1 - \frac{t(c_i)}{10}), \tag{8-3}$$

where $D_u(c_i)$ represents the decay factor of a post c_i by a user u, and t represents the creation date stored in the timestamp of a piece of content c_i. The resulting decay factor has a range between 0 and 1. Five years after the creation of a post, the decay factor will be 0.5, conforming to an expertise half-life of five years.

8-1-4 Aggregating Content Relevance, Confidence and Authority

Finally, we aggregate the content relevance of a relevant post in a UEP, the confidence and authority placed in that content, and the popularity and decay factors, resulting in an estimate of the expertise of the candidate expert with respect to the query subject. Equation 8-4 illustrates this aggregative step. For each post, we multiply the sum of the *content relevance* and *confidence and authority* factors, with the *decay* and *popularity* factors acquired from metadata read counts, share counts and timestamps.

$$X(u) = \sum_{i=1}^{n} ((R_u(c_i) + A_u(c_i)) * P_u(c_i) * D_u(c_i)), \tag{8-4}$$

where a user's expertise score X equals the sum of that user's content relevance scores $R_u(c_i)$ and the appreciation score $A_u(c_i)$, multiplied with the popularity and decay factors, for each piece of relevant content c_i the user u authored with respect to the query subject.

8-1-5 Expertise Identification Process

As we discussed in chapter 2, EI systems are typically found as components of ESM systems. Accordingly, we have developed our EI prototype to work in parallel with existing ESM systems. While this approach may cause for more data redundancy and be less lean than more integrated approaches, we believe it is essential to have an EI system working in parallel. Especially during the first period of use, when the weights of the prototype need to be tuned a lot, it should be able to operate the EI system separately from the ESM system. This way, the EI configuration may be altered without users experiencing any downtime.

The candidate-model component of our prototype (estimating content relevance) can retrieve content directly from an existing SQL database of an existing ESM system. Assuming the content in the ESM system changes continuously (new content may be added, existing content may be altered), Solr is configured to re-index individual pieces of content the moment they are altered. New items are also detected and indexed automatically.

The document-model component of the prototype looks up CF data in a social graph, associated with pieces of relevant content. We fill this graph with all relationships in the existing ESM, and like indexing in Solr, any change in CF relationships in the ESM system triggers an update in the graph as well. Figure 8-3 illustrates the described process. Except for periodically adjusting the operational weights and having users provide CF data as well as queries, the prototype is fully unsupervised.

8-2 Developed and Re-used Software

We already discussed our use of Apache Solr in extracting, indexing and searching against UEPs containing users' posts in the ESM system. Moreover, we use the JDBC SQL connector to enable Solr to connect to SQL. JDBC offers numerous other database connectors, enabling Solr to connect to databases other than SQL in case an ESM system uses a different relational database.

For the construction of the social graph, we used OrientDB, an *"open-source NoSQL Graph-Document DBMS"* (*What is OrientDB?*, 2012). Since OrientDB and Solr both provide JAVA APIs, we have written our prototype as a JAVA program. By default, our prototype does not have a front-end. For the sake of the tests dicussed in chapter 9, we configured the prototype to output a CSV file containing results and peripherals in response to each query subject. For the implementation of our prototype in E-view, we created a front-end in E-view in cooperation with TJELP developers. The front-end presents ranked lists of *likely* experts, together with

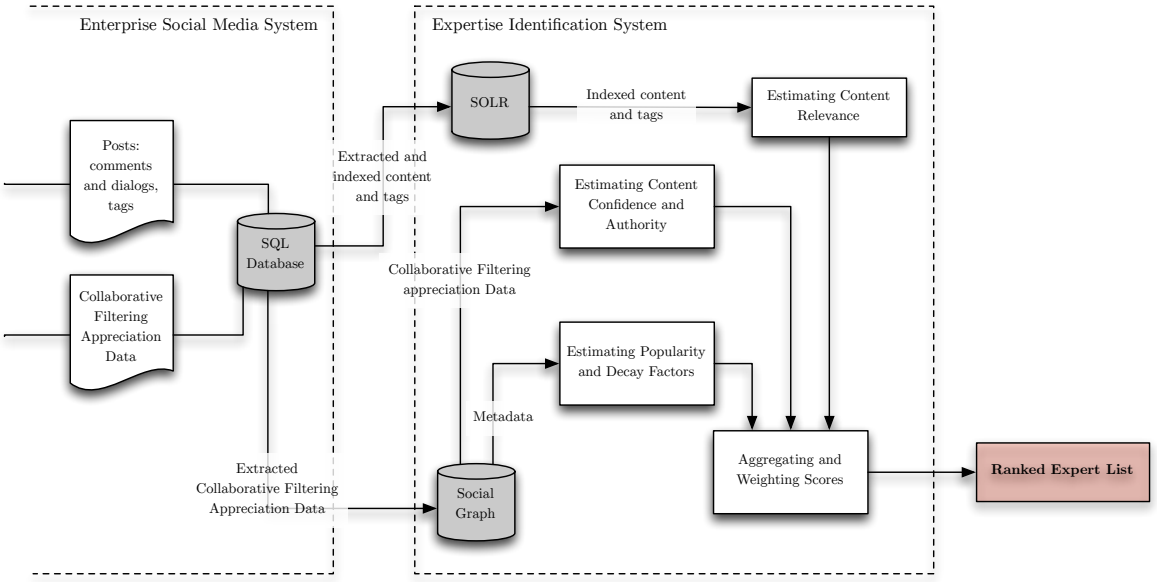


Figure 8-3: ICT process of our Expertise Identification system.

expertise scores. Moreover, the back-end of E-view's production version seamlessly integrates both OrientDB and our code, as it is constructed using the Play![23] framework.

8-3 Implementation in E-view

We have implemented our EI system, as presented in this chapter, in E-view. In doing so, we have had to account for the design constraints laid out in table 7-3 in chapter 7.

We have used the CF appreciation data that E-view provides to estimate the confidence and authority users place in a post. E-view contains *votes up*, *votes down*, *thank you's* and flags. Accordingly, a_j in equation 8-1 is made up of the sum of votes up, votes down, thank you's and flags. We do not know whether thank you's represent a more valuable judgment than votes, or whether votes up and votes down should be quantified with equal weights. We leave these weight issues to future research, which should investigate whether these types of CF data should be weighted differently in ESM systems. In this thesis, we assume that a vote represents either $+1$ (up) or -1 (down) and that a thank you or a flag represents $+1$.

Furthermore, E-view only supports timestamp metadata, not read and share counts. So we left out the popularity component discussed in subsection 8-1-3. We recommend that in the future, E-view should store *read* and *share* counts, so that the popularity factor can also be included in the calculation of expertise scores. Equation 8-5 shows the mathematical representation of estimated user expertise scores as depicted earlier in equation 8-4, adjusted for E-view.

$$X(u) = \sum_{i=1}^{n} ((R_u(c_i) + A_u(c_i)) * D_u(c_i)) \tag{8-5}$$

[23] Visit http://www.playframework.org for more information on the Play! framework.

Although E-view does support the timestamp metadata necessary to obtain the expertise decay factor $D_u(c_i)$, the dataset used, ranging from February 1st to September 5th, 2012, does not contain sufficient history for $D_u(c_i)$ to have a notable impact on system performance. As a consequence of $D_u(c_i)$, the oldest post in E-view receives only a 5.8% penalty on account of its age. Future research, using a dataset with a longer history, should investigate the impact of $D_u(c_i)$ on system performance.

As we explained in subsection 8-1-5, our EI prototype operates in parallel with E-view, merely causing some extra load on E-view's SQL database server.

8-4 Future Expansion and Development

In the future, our EI prototype can be expanded with a range of additional functionalities. One of the most important next steps is to realize Expertise Explication, i.e. explicating the stated expertise of each expert in our outputted ranked list of experts. Another future development could be to enable users to interact with their User Expertise Profiles. While in our prototype, these profiles are invisible to users, in the future they could serve as professional identities within and outside the organization.

8-4-1 From Expertise Identification to Expertise Selection

In chapter 1, we demarcated our research to focus on *Expertise Identification*, not *Expertise Explication*. *Expertise Identification* refers to identifying the experts with respect to a certain query X, whereas *Expertise Explication* refers to explicating what expertise an expert Y possesses, i.e. explaining what that user's fields of expertise are (McDonald, 2001).

Expertise explication can have many forms, and manifests through the information next to experts shown to users in the ESM. Examples are to display scores next to the experts, provide links to the most influential dialogs and comments a score was based on, or even provide a complete break down of all content and appreciation data underlying a score.

Inspiration for expertise explication can be drawn from successful existing score explication initiatives, such as users' profiles at Stackoverflow.com[24], scientists' research profiles at Researchgate.com[25], and public influence explained by Klout.com[26]. The difference between these examples and the case of extending EI with Expertise Explication in ESM, however, is that these examples show *overall* user profiles, not just the information underlying the expertise score of a user with respect to the query at hand. Moreover, Stackoverflow.com profiles are fully public.

[24] A good example of a filled Stackoverflow.com user profile is that of Jon Skeet, the author of *C# in Depth*. It can be found at http://stackoverflow.com/users/22656/jon-skeet?tab=summary.

[25] Researchgate.com is a closed community, so in order to review expertise explication you have to create an account at http://researchgate.com.

[26] Klout.com takes all your public social media activity and measures the activity and reach of your messages, comments, followers, followings et cetera. For more information on Klout, you can create an account at http://klout.com.

8-4-2 Validated Public Professional Identity

Our prototype hides users' Expertise Profiles. Instead of only using these profiles to provide the user with a ranked list of experts with respect to a query, they could also serve as a professional identity, summarizing all of a users' activity within the ESM system. Just like profiles on Stackoverflow.com, Researchgate.com and Klout.com, the user profiles in an ESM system can be extended to include information on users' activities, appreciation data, most active topics, and more.

Stackoverflow.com already offers users the possibility to automatically mine user profiles to build a professional identity. On Stackoverflow.com, this novel functionality is called *Careers 2.0*[27]. Here, users' Q&A activity from Stackoverflow.com is used to automatically explicate their professional identity as software developers. Companies looking for developers with a certain locale and skill-set can subscribe to Careers 2.0 to contact developers for job offers. They can review their detailed public activities on Stackoverflow.com in order to assess their skills. This way, companies avoid the costs and time normally needed to assess developers themselves.

The difference between Stackoverflow.com and ESM, however, is the fact that on Stackoverflow.com, all user content is *public*. In ESM systems, user content can be private or sensitive, making it difficult to use it in explicating users' professional identities. Users would be required to mark which content is sensitive and which is public. Another option would be to somehow anonymize user content automatically. Future research should investigate ways to validate professional identity using ESM content.

8-4-3 User Interaction with Expertise Profile

Once users' Expertise Profiles in an ESM system have been made visible to users, interaction between a user and his or her expertise profile could also enhance the precision of the profile. The question is, however, which aspects of the profile users should be able to edit.

8-5 Implementation of Design Choices

Now we have presented the technical design of our EI prototype for ESM and our implementation of the prototype in E-view, we summarize the implementation of the design choices laid out in chapter 5. Table 8-1 contains the design choices from chapter 5, together with their actual implementation. In chapters 9 and 10, we will evaluate design implementations 3, 4, 6, 7 and 9.

[27] Stackoverflow Careers 2.0 can be found on http://careers.stackoverflow.com.

	Design Choice	Implementation
1.	Suggest likely experts rather than *claim* or *state* the absolute experts on a particular subject.	We only *suggest likely experts* using our prototype (see section 8-2).
2.	The validity of an employee's expertise comprises of the *relevance*, *confidence* and *authority* placed in that employee's knowledge regarding a subject.	Combine estimated content relevance with estimated confidence and authority (see section 8-1).
3.	Knowledge relevance can be estimated by means of a selection of full-content and CF abstraction data in ESM.	We can use a selection of full-content and tags with individual weights (see subsection 8-1-1).
4.	Confidence and authority other employees place in knowledge can be estimated by means of CF appreciation data in ESM.	Estimate confidence and authority users place in content using CF appreciation data, stored in social graph (see section 8-1 and subsection 8-1-2).
5.	Hybrid approach.	Candidate-model in the sense of indexing all user content into User Expertise Profiles in Solr. Document-model in the sense of complementing content relevance score with appreciation data score (see section 8-1).
6.	Conventional content relevance estimation methods.	Solr uses conventional search engine similarity measures (see subsection 8-1-1).
7.	Expertise decay.	Estimated expertise scores decay with a half life of 5 years (see subsection 8-1-4).
8.	Automatically extract User Expertise Profiles.	Solr indexing of user content into a User Expertise Profile is done automatically (see section 8-1-5).
9.	Folksonomies over ontologies and taxonomies.	No use of ontologies and taxonomies. Full reliance on folksonomy of tags to abstract data (see subsection 8-1-1).
10.	Largely unsupervised EI.	Fully unsupervised, except for adjusting weights for the impact of relevance and appreciation, and the provision of CF data by users (see section 8-1-5).

Table 8-1: Overview of our implementation of design choices for Expertise Identification in Enterprise Social Media systems.

Chapter 9

Tests

In order to test the implemented Expertise Identification (EI) prototype described in the previous chapter, we executed a number of tests. These tests provide preliminary insights into the performance of the prototype, rather than a statistically significant foundation for formally validating or benchmarking it. We have chosen to do these tests over a complete validation, because of the limited number of data present in E-view at the time of this thesis. We discuss these data limitations more elaborately in section 9-3. First, we describe our approach to evaluating a number of EI strategies in E-view, using our prototype as described in chapter 8. In section 9-3, we discuss the used dataset as well as its representativity. Chapter 10 discusses the test results.

9-1 Approach

The EI prototype implemented in E-view, as discussed in section 8-3, enables us to vary the basis for the selection and scoring of relevant posts by Solr. In order to evaluate design choice 3 listed in table 8-1 – the estimation of post relevance by means of a selection of full-content and tags – we need to find out which selection method performs the best in E-view. We test three selection methods:

1. tags only: the relevant posts are selected from Solr by matching the query subject with the tags of a post;
2. full-content only: relevant posts are selected by matching the query subject with the full-content of a post, excluding the tags in that post, and;
3. full-content and tags: the combination of 1 and 2.

With respect to selection method 1, we decided to consider all tags in a thread for each dialog's main text and comments. So if 3 tags were added during the creation of a dialog,

and 4 tags were added by comments, we consider all 7 tags as part of both the comments and the dialog's main post.

Moreover, the prototype enables us to vary the way expertise scores are calculated, by varying the weights of the two score components *relevance* and *confidence and authority*. In reality, the prototype supports varying a number of other weights as well, enabling administrators to adjust the importance of both the relevance and appreciation data of comments versus dialogs, post components such as the dialog title versus the dialog body, and positive versus negative appreciation data. However, for the purpose of these tests, we are primarily interested in the combination of content relevance and appreciation data. We will gain first insights into the added value of appreciation data in estimating expertise scores in an Enterprise Social Media (ESM) system, enabling us to evaluate design choice 4 – estimating confidence and authority using Collaborative Filtering (CF) appreciation data – by comparing EI with and without the use of appreciation data. We assume[28] the weights of various types of appreciation data to be representative when equalized to 1. Lastly, varying the scoring methods enables us to test design choice 6 – using conventional Solr search engine similarity measures to determine content relevance. We test two scoring methods:

1. content relevance: only the Solr content relevance scores of posts are taken into account, and;
2. content relevance and appreciation: both content relevance and appreciation are taken into account, with equal weights.

9-1-1 Expertise Identification Strategies

Combining all possible selection and scoring methods results in a total of 9 available EI strategies, and the Zero Option. Figure 9-1 shows these strategies.

The *Zero Option* strategy represents the current method for EI in E-view, which estimates a top three of user expertise ranks by examining user frequency in dialogs returned in response to a query. Coherent with the selection and scoring methods chosen earlier, we will test six EI strategies. Table 9-1 provides an overview of these strategies.

#	Name	Scoring method	Selection method
7	CR_T	content relevance	tags
8	CR_{FC}	content relevance	full-content
9	CR_{FC+T}	content relevance	full-content and tags
1	$(CR+A)_T$	content relevance and appreciation data	tags
2	$(CR+A)_{FC}$	content relevance and appreciation data	full-content
3	$(CR+A)_{FC+T}$	content relevance and appreciation data	full-content and tags

Table 9-1: Tested strategies for Expertise Identification in E-view, using our EI prototype.

In order to measure the precision of the ranked lists of likely experts that our prototype generates for each strategy, we need a ground truth to compare them with, as well as a

[28] See appendix A for more information on this assumption, as well as other assumptions.

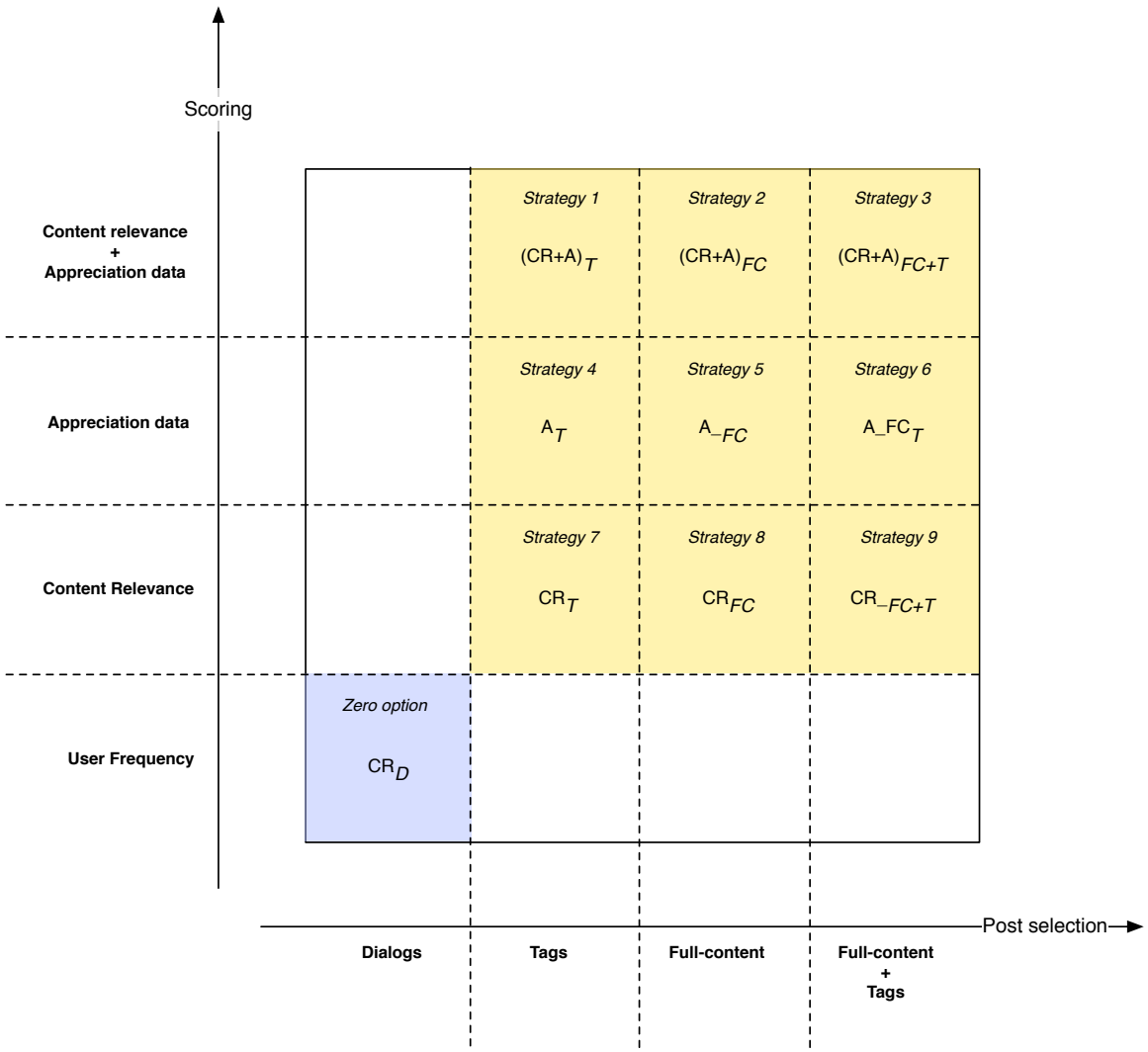


Figure 9-1: The main strategies for identifying experts as available in E-view using our Expertise Identification prototype.

number of query subjects to test. Let us first devise query subjects, and then establish a ground truth.

9-1-2 Query Subjects

In order to test the strategies in table 9-1, we need a set of subjects that are sufficiently present in E-view. Furthermore, it is important that the employees working on those subjects can provide a ground truth for the ranked lists of likely experts. Most of the content in E-view is in Dutch. Where helpful, we provide translations into English. In cooperation with TJELP employees, we have established the following set of query subjects:

1. *Enexis*, one of TJELP's clients;
2. *Weteringschans*, the informal name of TJELP's new office in Amsterdam;
3. *bug*, issues with E-view, and;
4. *Drupal*, the web development framework E-view's beta version was built in.

9-1-3 Ground Truth

We established the ground truth for each of the query subjects, by having each TJELP employee rank all employees (including him- or herself) on their expertise with respect to the subjects we proposed in the previous subsection. The ranking was done individually, so that employees could not knowingly influence each other knowingly in the process. Averaging all employees' ranked lists for each query subject results in the average and rounded rankings in table 9-2. The ranks range from 1(highest) to 7(lowest). From the average ranks, we can see that some users lie closer to the corresponding ranks than others. The 1st rank (with average rank = 1.00) with respect to the subject 'Weteringschans', for instance, was appointed to Rempko unanimously. Justus's 1st rank (with average rank = 1.83) with respect to the query subject 'Bug', however, can be disputed, as it lies very close to Rempko's average rank.

Rank	*avg. rank*	**Enexis**	*avg. rank*	**Weteringschans**	*avg. rank*	**bug**	*avg. rank*	**Drupal**
1	1.17	Henno	1.00	Rempko	1.83	Justus	1.50	Ruud
2	2.17	Niels	2.67	Willem	2.00	Rempko	1.67	Justus
3	2.67	Willem	3.50	Henno	3.00	Ruud	3.33	Rempko
4	4.17	Rempko	4.33	Justus	3.83	Henno	3.83	Willem
5	5.50	Freek	5.00	Ruud	5.17	Willem	5.33	Niels
6	5.83	Justus	5.17	Niels	5.33	Niels	6.00	Henno
7	6.50	Ruud	6.33	Freek	6.83	Freek	6.33	Freek

Table 9-2: Ground truth rankings for the query subjects in our tests, including the average ranks.

For client organizations of E-view with a lot of employees, say 100 or more, it quickly becomes impossible to acquire a ground truth like the one in table 9-2. Not only because it is difficult to have 100 employees fill out ranked lists of likely experts, but more so because the employees

cannot be expected to be apprised of the subject matter expertise of all the other employees. This is exactly where E-view can fulfill a vital role, as we explained in subsection 7-4-2. In order to tune the EI system, however, feedback on EI performance can be extremely helpful. We imagine that in order to collect feedback on EI performance in E-view within large organizations, enabling users to freely rate the ranks and likelihood of experts in expert lists might provide administrators with good feedback. Another way to collect it could be to confront a user with an expert, rank and query subject combination just before logging out, asking the user to rate the expert. With time, this method can provide a fine-grained and continuous ground truth for EI, representative for the client organization.

9-2 Mean Absolute Error

To compare the ranked expert lists our prototype generated for each strategy with each other, the ground truth and the Zero Option, we compare the Mean Absolute Error (MAE) of all expertise scores. This metric is commonly used in the field of Information Retrieval to measure the performance of predictions versus actual values (Su & Khoshgoftaar, 2009). The MAE computes the average of the absolute difference between the estimated expertise scores and the ground truth from subsection 9-1-3. The lower the MAE, the better the prediction. We calculate the MAE for an estimated expert list using equation 9-1.

$$MAE \equiv \frac{\sum_{(i,j)} |p_{i,j} - r_j|}{n}, \tag{9-1}$$

where $p_{i,j}$ represents the estimated expert rank for a strategy i and a subject j, r_j represents the corresponding ground truth expert rank and n the total number of expertise scores. Additionally, the performance of a strategy as a whole (for all query subjects) can be calculated by averaging the MAE values of all query subjects.

It is important to note that because we compare the estimated and predicted *ranks*, and not the *expertise scores*, a lot of information is lost. The expertise scores generated by our EI prototype contain a lot more information than just the rank of a likely expert. However, because of the scale and course of these expertise scores (as a result of the TFIDF scores they are based on), we are unable to acquire a ground truth for the expertise scores rather than for the ranks.

9-3 Data

The implementation of the EI prototype in E-view, as well as the tests we ran, use data in E-view from February 1st, 2012 – the day that TJELP first started to use E-view – up to September 5th, 2012. In this section, we first consider overall characteristics of the data in E-view. Then, we investigate the data of all seven TJELP employees: dialogs and comments, tags, and finally appreciation data they provided. Because the dataset covers only 7 months of use, it is difficult to test the effect of design implementation 7 from table 8-1 – an expertise decay half-life of 5 years. This should be examined by future research using a larger dataset that covers a longer period of ESM use.

Figures 9-2 and 9-3 show the development of the number of dialogs, comments, tags, users, thank you's, votes and flags in E-view during 7 months of use. The sudden increase in almost all data in the beginning of June was caused by E-view's beta-launch on June 6th. Tags are most often added to dialogs, not comments. Comments are rarely tagged.

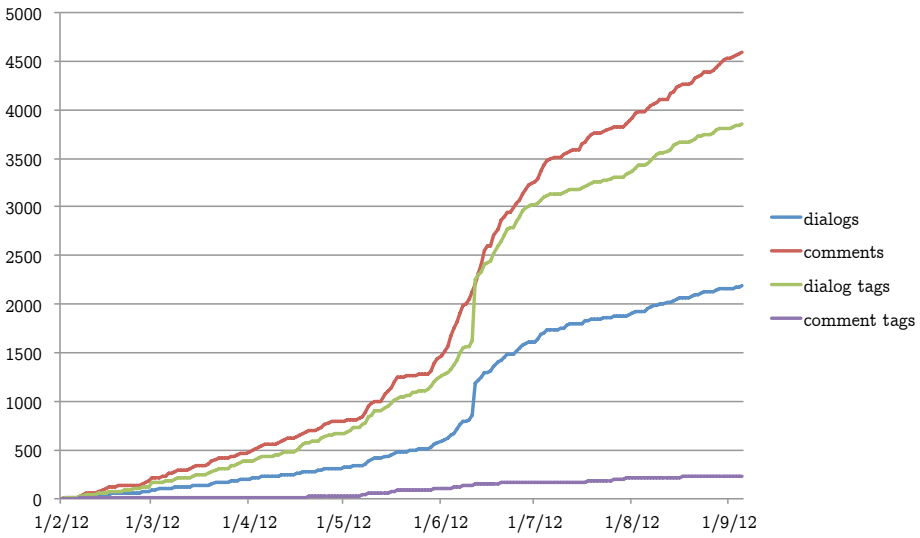


Figure 9-2: Dialogs, comments and tags in E-view over time.

The number of users present in the system (including but not limited to TJELP employees) grew fast after the beta-launch in June, and started to subside again by the beginning of July. As we will see in the next subsection, most of these users contributed barely or not at all to the data present in E-view by the beginning of September. The TJELP team authored most of the content.

From figure 9-3, up-voting dialogs and comments appears to be the most popular means of expressing appreciation. However, as we discussed in chapter 7, we should bear in mind that in E-view, users can vote on both dialogs and comments, whereas thank you's and flags are restricted to comments and dialogs respectively. Voting-down is barely used, possibly because people are more comfortable with rewarding good content than punishing bad content.

Now we have explored the overall use statistics of E-view in terms of its total number of users, posts and appreciation data, we continue to investigate the use statistics of TJELP employees.

9-3-1 Dialogs and Comments

Figure 9-4 shows the number of dialogs and comments published by TJELP employees. In total, TJELP employees published 1693 out of 2189 posts and 3731 out of 4591 comments in E-view. Considering that several hundred comments and posts comprise of automatically generated system messages, only a very small portion of the posts in E-view was not published by the TJELP team. This is not surprising, since client organizations taking part in E-view's pilot period will only start using E-view by the beginning of October. Furthermore, these posts are primarily written in Dutch. These data characteristics limit the representativity

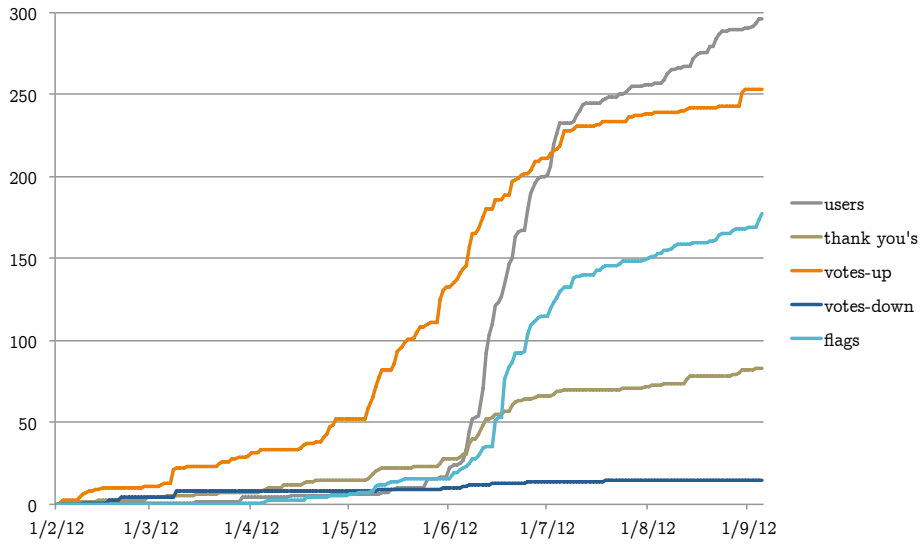


Figure 9-3: Users, thank you's, votes and flags in E-view over time.

of the dataset in terms of domain, language and nationality and may not be representative for the domain-independent, multilingual and multicultural contents of E-view in a future production state when it hosts many autonomous client organizations. Subsequently, the dataset is almost certainly not representative for middle- and large-sized companies with a lot of mobile and geographically dispersed employees. On the other hand, the way the TJELP team used E-view during the 7-month period of the dataset is representative for a startup company collaborating in an ESM system and working toward launching its next product.

Henno, Niels and Willem comment on most of the dialogs in E-view. Furthermore, with more than 500 published dialogs, Henno started most new dialogs. All 7 users have published at least around one hundred dialogs and comments.

9-3-2 Tags

It is important to examine employees' tag use in E-view, as our strategies aim to point out whether EI in ESM should base the selection of relevant posts from User Expertise Profiles on full-content, tags or both. See chapter 8 and table 8-1 for more details on this design consideration.

In section 5-6-1, we explained that a folksonomy of tags in an ESM system eventually converges as users start to agree more and more on the meaning and formulation of tags. In figure 9-5 we can see what appears to be an initial sign of such convergence. It shows that 30% of the unique tags in E-view account for more than half of *all* tag use. As such, a great number of tags is in fact being reused. Therefore, design choice 9 from table 8-1 – the use of folksonomies over ontologies and taxonomies to facilitate users' provision of abstraction data – is largely confirmed. We should emphasize, however, that in this research, we assume a fairly neat folksonomy of tags, as is the case in E-view. Moreover, a tag folksonomy still leaves the issue of the period of use before tag collection convergence: in this first period of use, EI based on tags can be very imprecise. The tests in the next chapter should point out whether

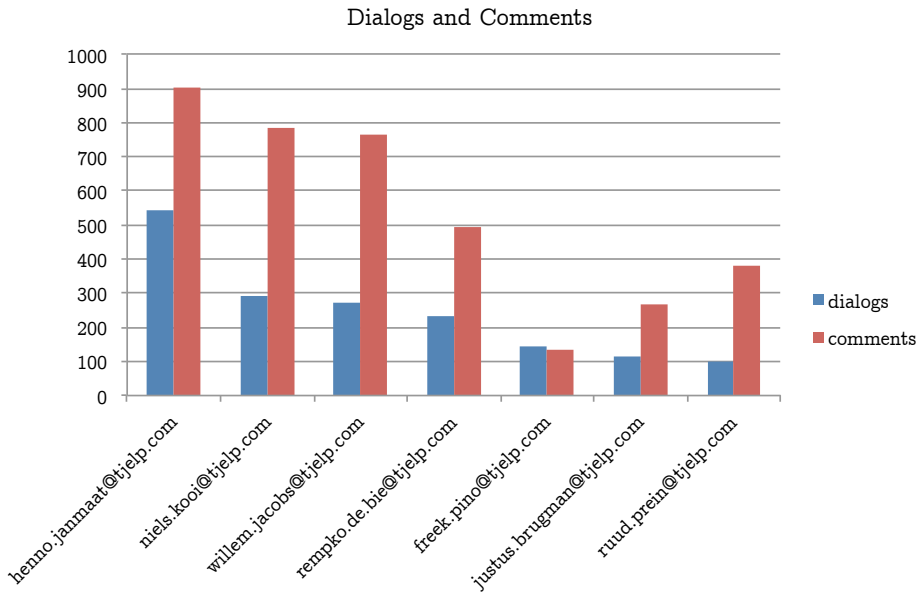


Figure 9-4: Dialogs and comments posted by TJELP employees.

estimating content relevance based on full-content can bridge this imprecise contribution of tags during the first period of use.

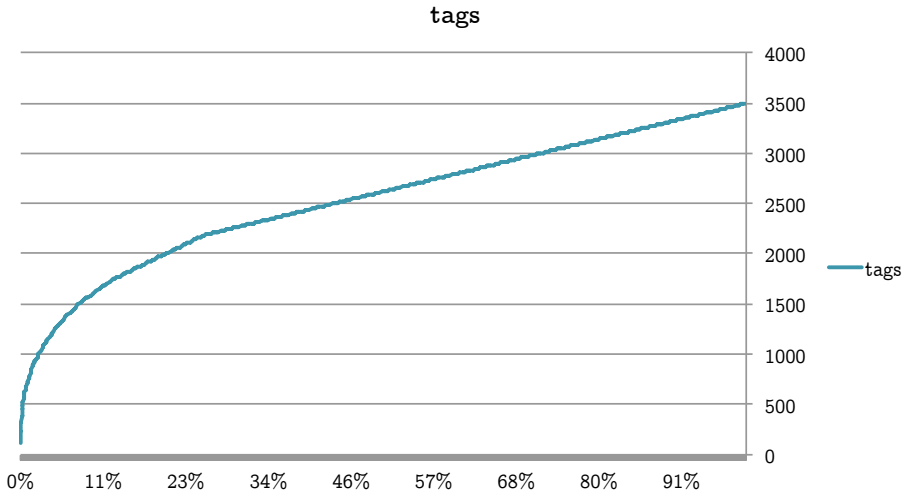


Figure 9-5: Accumulated use of unique tags in E-view.

The histogram in figure 9-6 shows the total number of tags provided by TJELP employees for posts in E-view. Willem, Niels and Henno provided most of the tags.

Figure 9-7 shows the average tag use in dialogs and comments. Overall, taking into account all posts in E-view, with or without tags, the average number of tags used approaches 0 for the case of comments, and is around 2 for the case of dialogs. Taking into account only those dialogs and comments with one or more tags, comments receive an average of 1.5 tags,

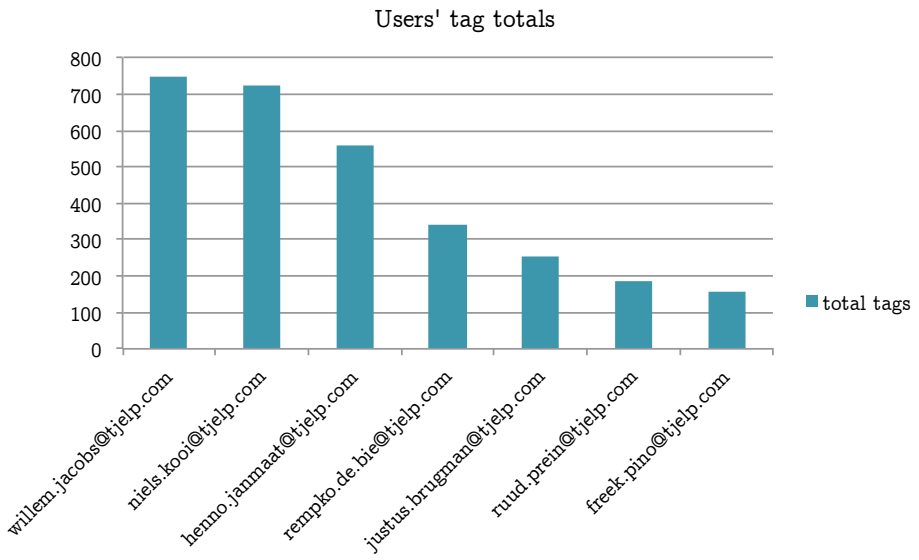


Figure 9-6: Total tag use of each TJELP employee.

whereas dialogs receive 2 tags. Freek and Niels provide the highest average number of tags for their comments. Justus and Willem, in turn, provide the highest average number of tags for their dialogs. There seem to be sufficient tags present in E-view for our EI strategies to test the selection of relevant posts on the basis of tags.

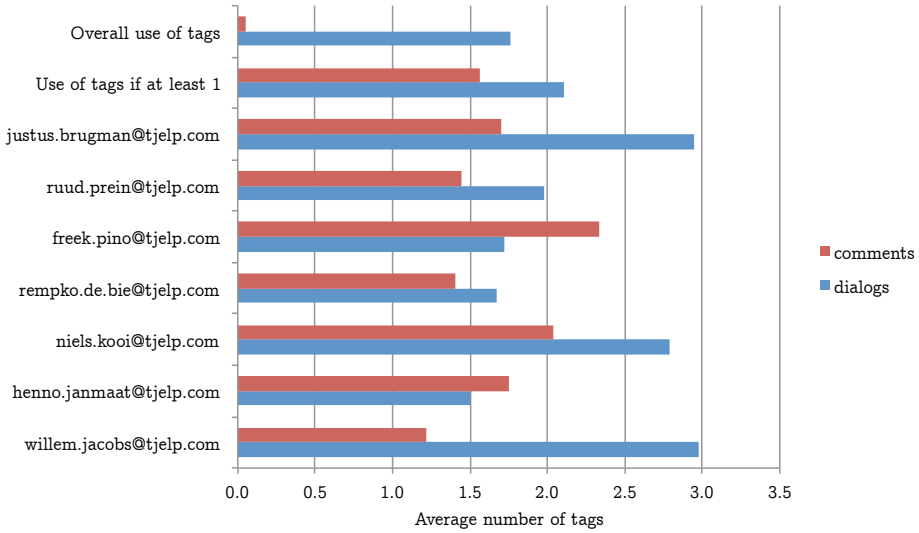


Figure 9-7: Average tag use of TJELP employees.

Finally, figure 9-8 shows the ten most popular tags in E-view. It is interesting to see that these most popular tags in fact represent the most actual topics TJELP has been involved in during the seven-month use period: from the development of *e-view*, to *testing bugs*, acquiring customers, *sales* et cetera.

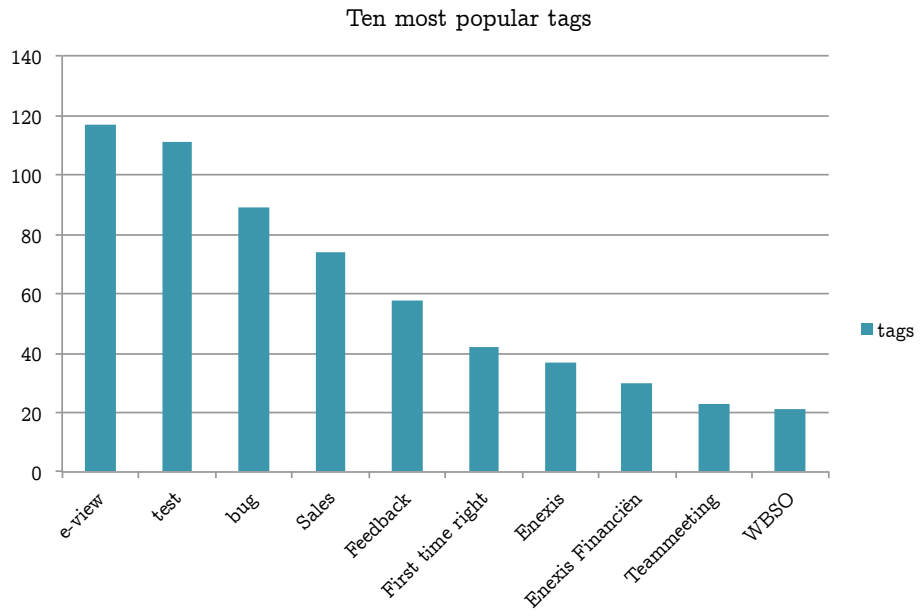


Figure 9-8: Ten most popular tags in E-view.

9-3-3 Collaborative Filtering Appreciation Data

As laid out in the beginning of this section, employees have provided only few appreciation data throughout the seven month use period. Figure 9-9 shows the number of votes, thank you's and flags provided by TJELP employees. With as little as 177 *flags*, 83 *thank you's*, 253 *votes-up* and 15 *votes-down*, TJELP employees expressed their appreciation for a total of 528 posts. With 1693 published dialogs and 3731 comments, employees provided appreciation data for less than 10% of all posts. Compared to, for instance, a Q&A system like Stackoverflow.com, of which it is known that around 60% of all answers are marked as accepted (*best answer* status or *voted up*), this seems very low (Raj et al., 2011). Apparently, users do not receive sufficient direct benefit from providing appreciation data. With tags, this direct benefit surfaces more easily: tags can be used to retrieve posts by either querying for the tag directly or drilling down search results using tags as facets.

This limited number of CF appreciation data makes evaluating the results of several of the EI strategies formulated in section 9-1-1 problematic. The EI prototype estimates expertise scores by accumulating all scores of relevant posts from User Expertise Profiles, as we explained in subsection 8-1-4. These individual scores are made up of a maximum relevance score of 1.0 and a maximum appreciation score of 1.0. With this little appreciation data present in E-view, the latter component will be 0.0 for almost all posts relevant to a query subject. Consequently, EI strategies that base expertise scores on both *content relevance* and *appreciation data* can be expected to give approximately the same results as strategies based only on content relevance.

This, however, does not prohibit us from gaining preliminary insights into the performance of the EI strategies. The small amount of appreciation data present in E-view makes for a *sparsity problem* often encountered in CF systems: what to do if there is only very little CF

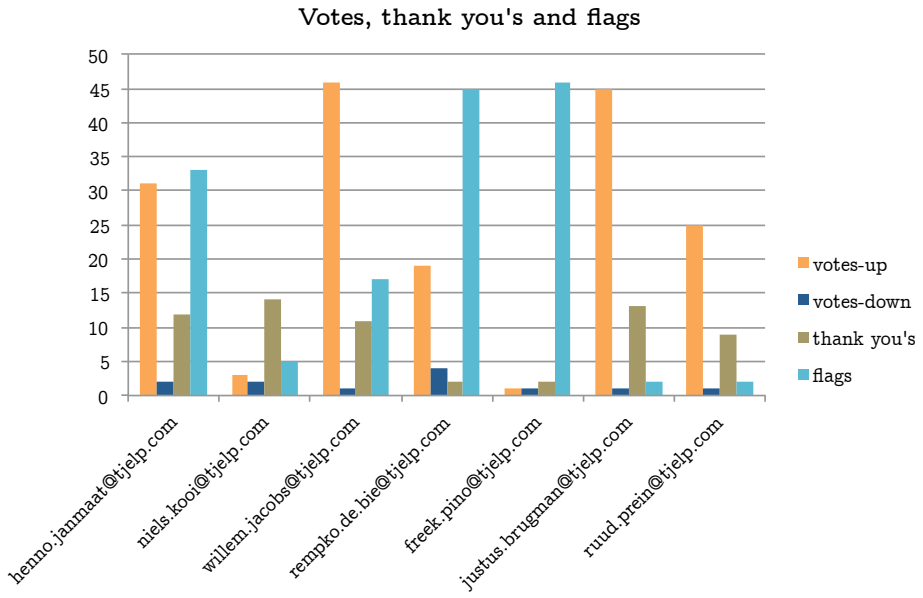


Figure 9-9: Thank you's, votes and flags appreciation data in E-view, provided by TJELP employees.

data to determine appreciation scores? One solution is to *impute* predicted or average ratings of users for all unrated posts (Su & Khoshgoftaar, 2009). While there exist many methods for doing so, calculating an average score for all posts without CF data still requires that users *have* rated a reasonable number of other posts, representative for the unrated posts. The better solution seems to create stronger incentives for users to provide appreciation data in the first place. We argue that strengthening these incentives can be accomplished through either *expertise explication* or some form of *gamification*. Expertise explication concerns enabling users to traverse a likely expert's posts and User Expertise Profile in order to examine why that candidate expert is knowledgeable on the queried subject. Not contributing any appreciation data will lead to a less impressive UEP, which can be viewed by other users through expertise explication. Gamification concerns the introduction of game-like objectives and rewards for users' provision of appreciation data. The Rypple ESM system we discussed in section 3-4 implements such gamification by rewarding user contribution with virtual badges.

9-4 Concluding Remarks

In this chapter, we introduced nine main EI strategies that are supported by our EI prototype from chapter 8. We test six of these strategies, combining different scoring and selection methods. These tests enable us to evaluate design implementations 2 (*relevance*, *confidence* and *authority as components of* valid expertise), 3 (using a selection of tags and full-content to acquire relevant posts from Solr), 4 (estimating confidence and authority of other users in a post by means of appreciation data), 6 (using conventional similarity measures from Solr to determine content relevance) and 9 (full reliance on folksonomy of tags to abstract data).

In consultation with TJELP employees, we established a ground truth for four query subjects

in E-view: *Enexis*, *Weteringschans*, *bug* and *Drupal*. In the future, when E-view hosts multiple client organizations operating in different domains, cultures and languages, determining a ground truth for EI may be done by having users rate expertise scores of likely experts, or by asking employees to rate a random EI suggestion when they log out of E-view.

In the next chapter, we compare the outcome of our EI prototype for the six strategies with each other, the ground truth and the Zero Option, based on the mean absolute errors of their *rankings* of likely experts. By doing so, part of the information contained in the actual expertise scores outputted by our prototype is lost.

We test the EI strategies using E-view data generated between February 1st and September 5th, 2012. As a result, it is difficult to test the effect of design implementation 7 from table 8-1 – an expertise half-life of 5 years. This should be examined by future research using a larger dataset that covers a longer period of ESM use. Furthermore, as most of this data was produced by TJELP employees, it may not be representative for the domain-independent, multilingual and multicultural contents of E-view in a future production state and ESM systems in general. It is, however, representative for a startup company collaborating in an ESM system and working toward launching its next product.

With 30% of all unique tags in E-view accounting for over half of the total tag use, we argue that the 7-month old folksonomy of tags in E-view is showing the first signs of convergence into an agreed upon vocabulary. Furthermore, comments are almost never tagged, while dialogs are on average provided with 2 tags. As such, design choice 9 from table 8-1 – the use of folksonomies over ontologies and taxonomies to facilitate users' provision of abstraction data – is confirmed. Assuming neat tags as explained in appendix A, this still leaves the first period of use, in which tags can be imprecise. The test results in the next chapter should point out whether leaving out tags in EI during this first period of use, as first described in section 5-9-8, is an option.

The most important issue with the dataset we used to execute our tests, is that it contains very little appreciation data. Less than 10% of all posts published by TJELP employees contains appreciation data. Because of this, the $(CR+A)_T$, $(CR+A)_{FC}$ and $(CR+A)_{FC+T}$ strategies are likely to generate approximately the same results as their equivalents based on content relevance only. We argue that users provide little appreciation data because they experience too little direct benefit from doing so. Therefore, we opt that E-view should strengthen users' incentives to provide appreciation data, by either implementing Expertise Explication or by gamifying CF techniques.

In the next chapter, we present the results of our EI strategies, comparing them with each other, the ground truth and the Zero Option. We also reflect on design implementations 2, 3, 4, 6 and 9, and propose revisions of our technical design accordingly.

Chapter 10

Results

In the previous chapter, we established a ground truth with respect to TJELP employees' expertise on a number of query subjects. We executed six Expertise Identification (EI) strategies in E-view in order to gain preliminary insights into the performance of our EI prototype with respect to that ground truth. In this chapter, we explore and compare the results of the strategies laid out in table 9-1 with each other, the ground truth and with the Zero Option.

Appendix C contains all raw output generated by our EI prototype in response to the 6 strategies and 4 query subjects. Table C-1 shows the settings we used to execute each strategy, and table C-2 provides an overview of where the output for each strategy can be found. Throughout this chapter, we will refer to the appendix and provide aggregated results when necessary.

First, we briefly discuss the general characteristics of the results. Next, we compare the generated ranked lists of likely experts with the ground truth we established, on the basis of mean absolute errors, as explained in the previous chapter. Then, we reflect on our implementation of design choices 3, 4, 6 and 9 from table 8-1. We propose a number of revisions to our design in order to improve on its shortcomings in E-view and Enterprise Social Media (ESM) systems in general. Finally, we discuss the future validation of our findings.

10-1 General Characteristics

Table 10-1 lists overall strategy performance and descriptives. We can see that strategies that select relevant posts on the basis of tags, on average return more than twice as many posts than strategies that select posts on the basis of full-content only. This means, that when a user provides an abstraction of his or her post through tags, these tags are most of the time not repeated in the full-content (title or body text). Therefore, it is not advisable to base EI solely on posts selected on the basis of full-content only. Additionally, this also means that it is not possible to leave out tags in EI during the first period of use in order to bridge the time until the tag collection has converged. Other means of tag disambiguation have to be sought.

Furthermore, and coherent with our findings on the sparsity of appreciation data in E-view in the previous chapter, we see that on average only a small portion of the posts contain appreciation data. For strategy CR_T, as few as 13 out of 369 retrieved posts for the query subject *Enexis* contained any appreciation data: less than 4%[29]. This may be explained by the nature of the posts in E-view pertaining to *Enexis*. As TJELP has been working on several projects for *Enexis*, a great number of related notices and reports was published on E-view. These news-like posts are, in contrast with for instance Q&A-like posts, often not annotated with appreciation data. In order to make EI more precise, E-view could either automatically classify post types with a higher degree of granularity, or enable users to do so. The latter is also implemented by other ESM systems as discussed in chapter 3, which enable users to create a *question*, *news update* or *event* rather than creating a *dialog* for all of these purposes.

Strategy CR_T also returned the highest portion of posts containing appreciation data. For the query subject *Drupal*, 22% of the selected posts contain appreciation data.

The average maximum appreciation data provided for a single post in the sets of relevant posts retrieved by each strategy for each query subject, is low. The CR_{FC+T} and $(CR + A)_{FC+T}$ scenarios show the highest average maximum: 2.5. In subsection 8-3, we chose values to measure the impact of both positive and negative appreciative gestures, $+1$ and -1 respectively. Moreover, in subsection 9-3-3, we revealed that TJELP employees have provided negative appreciation data (votes-down) on as few as 15 occasions. Therefore, the average maximum appreciation data of 2.5 can be expected to consist of primarily positive appreciation. This means that the posts in UEPs retrieved by these strategies, contained an average *maximum* number of appreciation data of 2.5. So within the sets of relevant posts to each query, there are no posts that have been voted on a large number of times; by all 7 employees, for example.

Non-surprising, it appears that the CR_{FC+T} and $(CR+A)_{FC+T}$ scenarios return the highest average number of relevant posts: 259.

10-2 Comparing Strategies, Ground Truth and Zero Option

In this section, we compare the results of the executed Expertise identification strategies with the ground truth, each other and with the Zero Option. We use table 10-1 to explain relative performance based on mean absolute errors.

10-2-1 Ground Truth and Strategies

Table 10-1 lists the mean absolute errors of strategies for each user (averaged over all query subjects), as well as for each overall strategy (averaged over all query subjects and users). The higher the MAE, the greater the error of the estimated ranked lists of likely experts generated by the strategy at hand. So if a strategy generates lists of ranked experts that are, on average, exactly the same as those of the ground truth discussed in subsection 9-1-3, that strategy would have a mean absolute error of 0. Because we compare the *ranks* of users in

[29]See table C-3 for the context of these numbers.

	Strategies					
	CR_T	CR_{FC}	CR_{FC+T}	$(CR+A)_T$	$(CR+A)_{FC}$	$(CR+A)_{FC+T}$
Characteristic						
Average posts selected by Solr.	200.25	93.75	259.00	200.25	93.75	259.00
Average posts containing appreciation data.	14.50	10.50	21.00	14.50	10.50	21.00
Average maximum appreciation for a single post.	2.25	1.75	2.50	2.25	1.75	2.50
User	Mean Absolute Errors					
freek.pino@tjelp.com	0.42	1.33	0.50	0.42	1.08	0.50
henno.janmaat@tjelp.com	1.79	1.54	1.04	1.79	1.29	1.04
justus.brugman@tjelp.com	0.84	2.09	0.84	0.84	1.92	1.17
niels.kooi@tjelp.com	1.50	1.59	1.42	1.50	2.00	1.50
rempko.de.bie@tjelp.com	0.71	0.88	0.38	0.71	0.88	0.38
ruud.prein@tjelp.com	1.25	1.00	1.00	1.00	1.00	0.75
willem.jacobs@tjelp.com	1.17	1.00	1.00	1.34	1.25	1.00
MAE	1.10	1.35	0.88	1.08	1.35	0.91

Table 10-1: Strategy performance with respect to the ground truth.

lists, the MAE for a strategy-user combination in fact represents the average offset of that user's estimated rank with respect to his or her average ground truth rank. From table 10-1, we can see that Rempko and Freek have relatively low MAE values. Although these users are not the top contributors in E-view, as we explored in section 9-3, the EI prototype is able to estimate their ranks more precisely. As Rempko and Freek do not seem to share any unique characteristic in terms of the descriptives in the previous chapter, we argue that this enhanced performance may be caused by the way they *compose* their content. If so, than the system's current settings are better adjusted to estimating expertise scores for their content than for other users' content. It is presumable that each user would ideally require different settings and weights.

In the previous chapter, we discussed that the small number of appreciation data in E-view may cause strategies based on both content relevance and appreciation to return nearly equal results. From table 10-1, we can see that this is indeed the case: strategies CR_T, CR_{FC} and CR_{FC+T} yield average MAE values almost equal to those of their appreciation counterparts $(CR + A)_T$, $(CR + A)_{FC}$ and $(CR + A)_{FC+T}$. The minor differences between these respective strategies with and without taking into account appreciation data are in fact negligible: they produce almost identical ranked lists of likely experts. The dataset simply contains too little appreciation data to judge its added value for the identification of expertise in E-view. We suggest repetition of this ground-truth test within, for example, 6 months. If users are, by then, given stronger incentives to provide Collaborative Filtering (CF) appreciation data to posts, we are hopeful that E-view will contain appreciation data for a greater portion of all posts by the next time the performance of EI strategies is evaluated.

Before the implementation of our EI prototype in E-view is validated using a richer dataset, it may also be validated using datasets from other ESM systems. However, because of the closed nature of ESM in comparison with public social media, such datasets are difficult to find. Moreover, publicly available datasets from public social media – like Stackoverflow.com – are not representative for the type of content in ESM systems, as described in chapters 2 and 3.

Nevertheless, it is clear that the CR_{FC+T} and $(CR + A)_{FC+T}$ strategies on average have the smallest mean absolute errors, i.e. perform the best with respect to the ground truth. Because we are comparing *rounded ranks*, the difference in average MAE between these two strategies (0.03) is negligible. It only affects rankings in two cases: Justus and Rempko in strategy $(CR + A)_{FC}$ and Justus and Ruud in strategy $(CR + A)_{FC+T}$, both in the ranked list for query subject *Drupal*. And in both cases, the absolute content relevance of the two users whose ranks where affected because of appreciation data was very low. See tables C-6 and C-8 for the concerning lists and scores. More appreciation data is required to test its impact on expertise scores.

Table 10-2 shows the same results, but from the perspective of the query subjects we used instead of the users. From this perspective, there are large differences in strategy performance. Ranked lists of likely experts generated for the subjects *Enexis* and *Drupal* are much more precise in terms of MAE than lists for the subjects *Weteringschans* and *bug*. From the data, we are unable to explain these differences. They might be explained by differences in the actual construction of the posts by users. However, because our EI prototype does not distinguish between dialog types, we can only speculate. Therefore, a higher granularity of posts in E-view appears necessary to further investigate the performance of our prototype with respect to different query subjects. Furthermore, we only tested our strategies using four query subjects. For the sake of statistically significant results, future research should test the strategies using at least 30 different queries.

10-2-2 Zero Option

As the Zero Option EI method in E-view only returns a top three of experts given a query subject, we constructed a top three variant of table 10-1. In calculating these MAE, we ignored all results beyond rank 3. Consequently, these results are not comparable with the results we discussed previously. The top three results, comparing the average MAE of all strategies and the Zero Option given the query subjects, can be found in table 10-3. The CR_T, $(CR + A)_T$ and $(CR + A)_{FC}$ strategies slightly under-perform compared to the Zero Option. This underperformance may be the cause of the relatively small number of posts selected by these strategies, only basing post selection on either tags or full-content. The CR_{FC+T} and $(CR + A)_{FC+T}$ scenarios outperform the Zero Option.

From these findings, it is again clear that EI in E-view should select relevant posts from User Expertise Profiles on the basis of both tags *and* full-content. Moreover, we conclude that our prototype outperforms the Zero Option method for EI, with respect to the ground truth.

	Query Subjects			
	Enexis	*Weteringschans*	*bug*	*Drupal*
Characteristic				
Average posts selected by Solr.	326.33	55.33	266.33	89.33
Average posts containing appreciation data.	14	4	26.33	17
Average maximum appreciation for a single post.	2	1.67	2.33	2.67
Strategy	Mean Absolute Errors			
CR_T	0.72	1.00	1.52	1.14
CR_{FC}	0.67	1.76	2.00	0.95
CR_{FC+T}	0.52	0.91	1.48	0.62
$(CR+A)_T$	0.72	1.00	1.48	1.14
$(CR+A)_{FC}$	0.67	1.76	2.00	0.95
$(CR+A)_{FC+T}$	0.52	0.95	1.48	0.67
Average MAE	0.64	1.23	1.66	0.91

Table 10-2: Strategy performance from the perspective of the query subjects.

Rank	**Mean Absolute Errors**						
	Zero Option	CR_T	CR_{FC}	CR_{FC+T}	$(CR+A)_T$	$(CR+A)_{FC}$	$(CR+A)_{FC+T}$
1	1.54	1.59	2.00	1.54	1.59	1.59	1.59
2	1.46	0.67	0.54	0.54	0.67	1.62	0.58
3	1.17	2.12	1.58	0.87	2.12	1.25	0.87
MAE	1.39	1.46	1.37	0.99	1.46	1.49	1.01

Table 10-3: Strategy performance with respect to the top three likely experts including the Zero Option.

10-3 Concluding: Revised Design Choices

Having explored the results of 6 EI strategies for 4 query subjects, we have gained a number of preliminary insights into the performance of our EI implementation in E-view.

The CR_{FC+T} and $(CR + A)_{FC+T}$ strategies utilize the most relevant posts and outperform all other strategies with respect to both the ground truth and the Zero Option. Because tags *are* used to abstract the contents of posts, the terms used for tagging are often *not* used in the full-content. Thus, basing EI on either tags *or* full-content leaves out a large number of relevant posts. Consequently, we revise design choice 3 in table 8-1 – the estimation of post relevance by means of a selection of full-content and tags – to estimate post relevance by means of the combination of *tags and full-content*. This also means that the impreciseness of tags during the first use period, as described in the previous chapter, cannot be compensated by only taking into account full-content during that first use period. Hence, we argue some form of tag disambiguation – other than standard stop-word removal and stemming performed by Solr – is required to be able to handle the messy tag folksonomy during the first use period. Therefore, we recommend that either users are aided in tag selection by ontology-based tag suggestions during this period, or they are given incentives to perform tag disambiguation themselves. The latter can, for instance, be accomplished through allowing active users to suggest cases wherein tags are better converged, edited, or deleted. The implementation of design choice 9 – the use of folksonomies over ontologies and taxonomies to facilitate users' provision of abstraction data – should be nuanced. If users can be stimulated to perform disambiguation of the tag collection before convergence, ontologies might not be necessary to handle the messiness of tags. If users are not willing to take care of tag disambiguation, ontologies can help to suggest tags and propose manual cases of tag convergence. If ontologies are used, it is important to make sure that the design requirements in table 5-1 are still met, meaning that ontologies should only be used to *suggest* tags, not classify or enforce them.

As the CR_{FC+T} strategy outperforms the other strategies as well as the Zero Option, we argue that conventional methods from Information Retrieval (Solr in our implementation) perform well in estimating content relevance. As such, we confirm design choice 6, as well as our implementation of that design choice in Solr.

Design choice 4 – estimating confidence and authority using CF appreciation data – cannot be confirmed nor disconfirmed by our results, due to the small amount of appreciation data available in the dataset. As we suggested in section 10-2-1, the strategies are best tested again several months from now, relying on E-view to offer users stronger incentives to provide CF appreciation data for posts. We believe that users in E-view receive too little direct benefit from providing appreciation data. Two ways to create stronger incentives are to realize Expertise Explication – the challenge of maintaining a positive expertise profile which other users can see and traverse – and to gamify CF techniques, offering users game-like objectives and rewards. Another option is to test our EI prototype using another ESM dataset. However, such datasets are not easy to find, because of the closed character of ESM systems.

From the perspective of the query subjects, there are big differences in strategy performance. We cannot explain these differences with the data we have. We speculate that the 'fitness' of a query subject for EI may depend on the type of posts related with the subject (e.g. news-like messages versus Q&A threads) or the actual content terminology used in the posts. Either way, further investigation necessitates a higher level of post granularity as well as a

larger number of tested query subjects. Higher post granularity may be achieved through the automatic classification of posts into different types of posts, or by having users provide information on the type of post using CF techniques.

Table 10-4 lists the revised set of design choices and their implementation. We boldfaced the revisions.

	Design Choice	Implementation
1.	Suggest likely experts rather than *claim* or *state* the absolute experts on a particular subject.	We only *suggest likely experts* using our prototype (see section 8-2).
2.	The validity of an employee's expertise comprises of the *relevance*, *confidence* and *authority* placed in that employee's knowledge regarding a subject.	Combine estimated content relevance with estimated confidence and authority (see section 8-1).
3.	Knowledge relevance can be estimated by means of a selection of full-content and CF abstraction data in ESM.	**We use full-content and tags to select and score posts.**
4.	Confidence and authority other employees place in knowledge can be estimated by means of CF appreciation data in ESM.	Estimate confidence and authority users place in content using CF appreciation data, stored in social graph (see section 8-1 and subsection 8-1-2).
5.	Hybrid approach.	Candidate-model in the sense of indexing all user content into User Expertise Profiles in Solr. Document-model in the sense of complementing content relevance score with appreciation data score (see section 8-1).
6.	Conventional content relevance estimation methods.	**Solr uses conventional search engine similarity measures to select and score relevant posts.**
7.	Expertise decay.	Estimated expertise scores decay with a half life of 5 years (see subsection 8-1-4).
8.	Automatically extract User Expertise Profiles.	Solr indexing of user content into a User Expertise Profile is done automatically (see section 8-1-5).
9.	Folksonomies over ontologies and taxonomies.	**Full reliance on folksonomy of tags to abstract data. Manual tag convergence by users during the first period of use, if necessary with suggestions from ontologies.**
10.	Largely unsupervised EI.	Fully unsupervised, except for adjusting weights for the impact of relevance and appreciation, and the provision of CF data by users (see section 8-1-5).

Table 10-4: Overview of our revised implementation of design choices for Expertise Identification in Enterprise Social Media systems.

Chapter 11

Conclusion Part 2

In Part 2 of this thesis, we conducted a case study into Expertise Identification (EI) in E-view. During this case study, we explored the additional design constraints E-view imposes on EI, presented the technical design of our EI prototype and its implementation in E-view, and we ran a number of tests to gain preliminary insights into its performance and shortcomings.

In this chapter, we will first complete the answer to research question 2, which went partially unanswered in chapter 6. Then, we answer research question 4. Finally, we provide an answer to our main research question as posed in the introduction of this thesis.

11-1 Answering Research Question 2

2. *How can expertise be defined and quantified in the context of an Enterprise Social Media system?*

In chapter 6, we conceptualized how *valid* expertise in ESM should be *estimated* by analyzing users' content and Collaborative Filtering (CF) data. In Part 2, we quantified this conceptualization. Our EI prototype selects and scores relevant posts using conventional methods from Information Retrieval. Using Apache Solr, we select relevant posts and calculate relevance scores $R_u(c_i)$ for each individual post. Equation 8-1 shows the computation of an aggregate *appreciation* score $A_u(c_i)$ for each individual post c_i. In computing $A_u(c_i)$, we assume equally weighted values for appreciation data: +1 for positive and -1 for negative. By multiplying the sum of the relevance score $R_u(c_i)$ and appreciation score $A_u(c_i)$ with a popularity score $P_u(c_i)$ and decay score $D_u(c_i)$, we obtain the expertise score of a user's single post. Summing all of a user's relevant scores, we obtain equation 8-4, depicting the expertise score of a user u with respect to the query at hand.

$$X(u) = \sum_{i=1}^{n} ((R_u(c_i) + A_u(c_i)) * P_u(c_i) * D_u(c_i)) \tag{8-4}$$

The implementation of our prototype in E-view uses an adjusted version of equation 8-4, as E-view does not store all of the metadata necessary to compute the popularity factor $P_u(c_i)$. Users' expertise scores are instead obtained using equation 8-5.

$$X(u) = \sum_{i=1}^{n} ((R_u(c_i) + A_u(c_i)) * D_u(c_i)) \tag{8-5}$$

We recommend that in the future, E-view should ideally store *read* and *share* counts, so that the popularity factor can also be included in the calculation of expertise scores. Although E-view *does* support the timestamp metadata necessary to obtain the expertise decay factor $D_u(c_i)$, the dataset does not contain sufficient history for $D_u(c_i)$ to have a notable impact on system performance. Future research, using a dataset with a longer history as well as read and share counts, should investigate the impact of both $P_u(c_i)$ and $D_u(c_i)$ on system performance.

The definition, conceptualization and quantification of expertise in the context of an Expertise Social Media system, as established in chapters 4 through 8, serves as the basis for the computation of expertise scores by our EI prototype.

11-2 Answering Research Question 4

4. *How can Expertise Identification be implemented in Enterprise Social Media systems?*

    (a) *How can Expertise Identification be implemented in E-view?*

        i. *What kind of ICT architecture is required to facilitate the identification of expertise in E-view?*

In Part 1, we specified a set of design requirements and design choices, by exploring best practices from theory and practice, as laid out in table 5-1. In chapter 7, we expanded these requirements and choices with additional E-view-specific design-constraints, which can be found in table 7-3. Throughout chapter 8, we presented our implementation of these design choices, accounting for the design constraints.

We have implemented a hybrid approach to EI that exploits both the candidate- and document-model approaches to EI. Coherent with the candidate-model approach, all user posts are indexed and stored into User Expertise Profiles. Solr enables us to use conventional methods from Information Retrieval to select relevant posts from UEPs and calculate their relevance scores. Moreover, our prototype is able to weight different parts of posts individually. While our implementation only uses ontology-free methods for disambiguating full-content and tags, Solr offers integration with ontologies which may be useful in providing tag suggestions in an Enterprise Social Media (ESM) system before convergence of the tag folksonomy.

Our prototype extracts all *relationships* between users, tags and full-content, into a social graph. Conform with document-model approaches, we traverse that social graph to determine *appreciation* scores of relevant posts. We then use these appreciation scores as an estimate of the *confidence and authority* other users place in the posts.

As an EI system can be expected to be implemented as a component of an ESM system, we have created our EI prototype to work in parallel with ESM systems. It automatically retrieves and indexes any change in ESM posts and relationships. Figure 8-3 shows the ICT process of the prototype.

Implementing the prototype in E-view required adjusting it to work with the CF data provided by E-view. Moreover, E-view stores limited metadata, forcing us to forego the *popularity* and *decay* factors in computing user expertise scores with respect to a query. Apart from these adjustments, however, our prototype integrates perfectly with E-view as it operates in full parallel. In fact, by the time this thesis is completed, a version of the prototype has been implemented in E-view's live environment, and is being used to suggest likely experts to pilot users with respect to every search query they run.

Although our prototype is able to suggest the likely experts with respect to a query subject, future steps are necessary to more fully exploit the possibilities of EI. First and foremost, users should be able to traverse expertise scores and the underlying User Expertise Profiles. E-view should facilitate breaking down a user's expertise score and finding relevant information in the process. Second, User Expertise scores and the underlying UEPs can be used to *validate public professional identity*. This is already being implemented by several big *public* social media systems, but is not yet used within the professional sector. This is not surprising, since a number of major challenges with respect to the closed nature of ESM systems and their content need to be overcome.

Table 8-1 lists the design choices from Part 1, together with our implementation of these choices in Part 2.

(b) *Which form of Expertise Identification is most suitable for E-view?*

In order to gain preliminary insights into the performance of different EI strategies using our prototype, we formulated six strategies in chapter 9, visualized in figure 9-1. The way these strategies vary available selection and scoring methods is laid out in table 9-1.

We determined a number of ground truth query subjects for which to execute the strategies, and had TJELP employees provide the necessary ground truth rankings to compare our prototype's output with. When E-view contains multiple large organizations, testing system performance using such a ground truth is not practical, hence users should be asked and otherwise incentivized to *rate* expertise ranks. That way, E-view can continue to collect ground truth rankings in the future.

We compared the ranked lists of likely experts outputted for each strategy with each other, the ground truth and the Zero Option, on the basis of mean absolute errors.

We found the dataset, consisting of 7 months of E-view use, to be unrepresentative for the domain-independent, multilingual and multicultural contents of an ESM system hosting many different client organizations, as discussed in chapter 3. In the dataset, TJELP employees have created almost all of the posts and appreciation data. Moreover, less than 10% of all posts contain appreciation data, causing the strategies based on *content relevance and appreciation data* to return almost exactly the same ranked lists as strategies based on *content relevance* only. As these limitations to the representativity of the dataset undermine our ability to

perform a complete validation of our EI prototype in E-view, we conducted a number of tests instead.

We recommend strengthening user incentives to provide CF appreciation data in E-view. This can be accomplished through Expertise Explication – challenging users to maintain their expertise status – and through gamification of the CF techniques.

The results of our tests, as presented in chapter 10, provide us with a number of preliminary insights into EI in E-view and in ESM systems in general. Because the terms used in tags and full-content appear to be mutually exclusive, we conclude that EI should select and score posts on the basis of both tags *and* full-content. Strategies based on both tags and full-content outperform other strategies and the zero option, with respect to the ground truth.

To disambiguate messy tags in the period of use before convergence of the tag folksonomy, E-view has to either incentivize users to disambiguate tags themselves, or aid users and administrators in disambiguating tags by providing suggestions for manual tag convergence. Furthermore, our test results point out that conventional methods from Information Retrieval perform well in estimating the content relevance of posts.

Due to the small amount of appreciation data available in the dataset, we were unable to evaluate the estimation of *valid* expertise through content relevance, as well as confidence and authority other users place in that content. The validity of estimating that confidence and authority on the basis of appreciation data in ESM systems also remains unconfirmed. The strategies are best retested in several months, relying on E-view to offer users stronger incentives to provide CF appreciation data for posts.

From the perspective of the query subjects, there are large differences in strategy performance. We cannot explain these differences with the data we have. Further investigation necessitates a higher level of post granularity as well as a larger number of tested query subjects. Higher post granularity may be achieved through the automatic classification of posts into different types of posts, or by having users provide information on the post types using CF techniques.

Although we were able to positively evaluate a number of design choices and implementations in 8-1, we were unable to report on the added value of appreciation data for EI in ESM.

Baring in mind the limitations of our test results, we argue that EI in E-view, and consequently in ESM systems in general, should be based on either strategy CR_{FC+T} or $(CR + A)_{FC+T}$ from figure 9-1. Future repetition of our tests using a richer E-view dataset should *validate* the use of the combination of *content relevance and appreciation data* over *content relevance only*.

11-3 Answering the Main Research Question

Having answered research questions 1 through 4, we can finally answer the main research question:

What form of Expertise Identification is most suitable for ESM, accounting for its social-media-like content and Collaborative Filtering characteristics?

Through the exploration of ESM, the definition and conceptualization of expertise, and the extraction of EI best practices from both literature and practice, we have established a number

of design requirements and choices for EI in ESM. These choices, as listed in table 5-1, formed the basis for our technical design of an EI system and its implementation in E-view. By testing our EI prototype using a number of strategies and query subjects, we gained preliminary insights into its performance in a live ESM system. On the basis of this performance, we revised our design implementations, resulting in table 10-4. EI in ESM should adhere to the design implementations in this table

Unfortunately, we were unable to report on the added value of appreciation data for EI in ESM systems. Baring in mind this limitation, however, EI based on content relevance *and* appreciation data was not proven to perform worse either. Concluding, EI in ESM should be based on either content relevance scores only, or the combination of content relevance scores and appreciation scores. In either case, the selection of relevant posts should be based on the *combination* of tags *and* full-content. Moreover, the tests we ran have to be repeated frequently in order to safeguard system performance and adjust weights. If users are, by the time our EI strategies are retested, more strongly incentivized to provide CF appreciation data to posts, we are hopeful that E-view will contain enough appreciation data to validate its added value

In this thesis, we have left a number of issues and research tasks out of scope, and we adopted assumptions for others. In the next chapter, we give an overview of the resulting recommendations, limitations and future work.

Chapter 12

Recommendations, Limitations and Future Work

This thesis establishes a number of recommendations for Expertise Identification (EI) in Enterprise Social Media (ESM) and specifically in E-view. Due to time constraints, we left a number of issues and research tasks out of scope that need to be addressed by future research. Moreover, our work is based on a number of assumptions, listed in appendix A, which we take into account as limitations of our research.

12-1 Recommendations

During the course of our research, we have expressed a number of recommendations for EI in ESM systems, and specifically in E-view. We discuss these recommendations one by one.

Except for design choices 4 and 7, we have been able to confirm all design choices and their implementations in table 10-4. As such, they provide a strong basis for both practice and research on EI in ESM systems.

Furthermore, we recommend that EI in ESM is based on either content relevance scores only, or the combination of content relevance scores and appreciation scores. In either case, the selection of relevant posts should be based on the *combination* of tags *and* full-content. EI performance should be tested regularly, using the approach set out in chapter 9. If users are, by the time our EI strategies are retested, more strongly incentivized to provide CF appreciation data to posts, we are hopeful that E-view will contain enough appreciation data to validate its added value

The strategies are best retested in 6 months time, relying on E-view to offer users stronger incentives to provide Collaborative Filtering (CF) appreciation data for posts.

The impreciseness of tags during the first use period, as described in the previous chapter, cannot be compensated by only taking into account full-content during that first use period.

Hence, we argue some form of tag disambiguation – other than standard stop-word removal and stemming performed by Solr – is required to be able to handle the messy tag folksonomy during the first use period. Therefore, we recommend that either users are aided in tag selection by ontology-based tag suggestions during this period, or they are given incentives to perform tag disambiguation themselves. If users can be stimulated to perform disambiguation of the tag collection before convergence, ontologies might not be necessary to handle the messiness of tags. If users are not willing to take care of tag disambiguation, ontologies can help to suggest tags and propose manual cases of tag convergence. If ontologies are used, it is important to make sure that the design requirements in table 5-1 are still met, meaning that ontologies should only be used to *suggest* tags, not classify or enforce them.

12-1-1 E-view-specific Recommendations

In subsection 9-3-3, we argue that users do not receive sufficient direct benefit from providing appreciation data in E-view. We opt the solution is to create stronger incentives for users to provide appreciation data in the first place. Strengthening these incentives can be accomplished through either *Expertise Explication* or *gamification* of CF techniques.

Furthermore, we cannot (yet) account for the differences in strategy performance and appreciation data with respect to the different query subjects. Future research into these issues necessitates a higher level of post granularity as well as a larger number of tested query subjects. We recommend either automatic classification of posts into different types of posts (based on ontology-free logic), or having users provide information on the type of post using CF techniques.

Besides timestamps, EI in E-view should also store the total read count from the perspective of main posts and comments in the social graph. That way, the popularity factor presented in subsection 8-1-3 can be included in estimating user expertise scores. As described later in this chapter, we encountered a number of issues with respect to data representativity: our E-view dataset was neither representative for that of ESM we described in chapter 3, nor did it cover sufficient use history or did it contain sufficient appreciation data for posts to report on the added value of appreciation scores in the identification of user expertise. Consequently, we recommend running the tests from chapter 9 regularly, so that the impact of increased provision of appreciation data on EI performance can be monitored, and weights can be adjusted if needed.

12-2 Limitations

First of all, we have formulated and tested six EI strategies using our EI prototype in E-view. Obviously, many more strategies can be composed. For instance, by shifting the weight settings of the EI prototype, as listed in table C-1.

As we noted in subsection 9-1-3, we compare the estimated and predicted *ranks*, not the *expertise scores*. This causes a loss of information. The expertise scores generated by our EI prototype contain more information than just the rank of a likely expert. However, because of the scale and course of these expertise scores (as a result of the TFIDF scores they are based on), we were unable to acquire a ground truth for the expertise scores rather than the ranks.

12-2-1 E-view-specific Limitations

In section 8-3, we discussed that E-view only supports timestamp metadata, not read and share counts. As a result, we had to leave out the popularity component presented in subsection 8-1-3.

Although E-view does support the timestamp metadata necessary to obtain the expertise decay factor D presented in subsection 8-1-3, our dataset ranging from February 1st to September 5th, 2012, does not contain sufficient history for $D_u(c_i)$ to have a notable impact on system performance. Future research, using a dataset with a longer history, should investigate the impact of $D_u(c_i)$ on system performance.

Since most of the dataset used to perform the tests was produced by TJELP employees, it may not be representative for the domain-independent, multilingual and multicultural contents of E-view in a future production state and ESM systems in general. It is, however, representative for a startup company collaborating in an ESM system and working towards launching its next product.

The most important issue with the dataset is that it contains very little appreciation data. Less than 10% of all posts published by TJELP employees contain appreciation data. Because of this, the $(CR + A)_T$, $(CR + A)_{FC}$ and $(CR + A)_{FC+T}$ strategies generate results that are almost identical to those of their equivalents based on content relevance only. We argue that users provide little appreciation data because they experience too little direct benefit from it.

Finally, we operated on the basis of a number of assumptions, which influenced the obtained results. These assumptions are discussed in appendix A.

12-3 Future Work

In section 5-6-1, we decided to leave files out of scope. It is often very difficult to determine the authors and abstracted content of files, e.g. documents and images, which makes accounting for files in EI into a distinct task. Future research should investigate whether files can offer added value in performing EI in ESM.

While recognizing the potential benefits of user personalization to EI Systems, we also left user personalization out of scope. User personalization poses a separate research task, touching upon aspects from research fields like User Modeling and Behavioral studies, which should be performed by future research.

Furthermore, future research should extract lessons on expertise propagation from public social media and fit these onto ESM.

Research by Yeung et al. (2011) describes state of the art methods to prevent abuse of CF techniques, which can be implemented in any EI system. Future research should investigate the use of these methods for EI in ESM systems.

In section 5-6-5, we discussed that the initial absence of content and tags in an ESM system, also called the cold-start problem, can possibly be countered by importing existing information about users and their organization into the system. Compatibility issues between the external data and the data in ESM systems, however, make enriching User Expertise Profiles with such

external web data into a time-consuming task. For the sake of time constraints, we left it out of scope. Future research should point out whether this enrichment step indeed offers a remedy for the cold-start problem.

As explained in subsection 8-1-3, we store timestamp metadata in a social graph, containing the creation dates of posts. Although this gives a fairly decent representation of the age of a message, ideally edit dates should also be taken into account. Because of time constraints on this research, we have not implemented these edit dates as an attribute of the social graph. Future research should investigate how these edit dates are best taken into account when constructing a decay factor.

In subsection 8-1-1, we explained that our prototype only uses simple ontology-free techniques to disambiguate full-content and tags, counting on a converged folksonomy of tags to fulfill this role once the system is actively in use. Future research should look into the possibilities of using Solr in combination with ontologies to provide tag suggestions in the use period before the tag collection has converged.

Subsection 8-1-1 describes how our prototype indexes different content types in UEPs into different Solr fields, enabling us to weight different post components individually. In this thesis, however, we have assumed equal weights for all components. The same is true for different appreciation data, as described in subsection 8-1-2. Future tests and research should try different weight configurations for both components of content relevance and appreciation scores. It is realistic to assume that not all types of comments, for example, are in fact of equal value. Equivalently, a comment might contribute more value to a user's expertise than a dialog.

Besides content-specific appreciation, we believe that each user's *overall reputation* also contributes to the estimation of a user's confidence and authority. Although many ESM and public Social Media already calculate overall user reputation by accumulating user popularity, contributions, attrition rate and other activity measures, using that reputation in conjunction with estimated user expertise with respect to a query is more difficult. Future research should experiment with the *overall* confidence and authority other users place in a user and determine its value to EI.

Finally, future research should investigate expansion of our EI prototype to facilitate Expertise Explication, user interaction with UEPs, and validation of professional identity using ESM content. We shortly touched upon these future expansions in section 8-4. Especially validating one's public professional identity using ESM content comprises a very challenging task. Because of the closed characteristics of both ESM systems and users' content in these systems, many technical and behavioral issues need to be addressed.

Chapter 13

Reflection

This chapter contains my personal reflection on the process and results of the 8-month graduation project behind this thesis. I will start by explaining how my Master's program at the Delft University of Technology has led to this particular graduation project. Then, I go into a number of practical dos and don'ts which I realized in hindsight.

13-1 Background of This Project

After my Bachelor's program *Technology, Policy and Management*, I started a hybrid Master's program: *Systems Engineering, Policy Analysis and Management - Information Architecture Track* (SEPAM-IA). This program, with a strong focus on Information Architecture, is being taught partially at the Faculty of Electrical Engineering, Mathematics and Computer Science. Here, I learned to progress beyond well performed and well written problem analysis and toward working with real-life demos and state-of-the-art linked-open-data and data-mining techniques. At the same time, the SEPAM component of my Master's program showed me how to design complex ICT systems, taking into account best practices and limitations from both theory and practice. In my graduation project, I was eager to combine both my SEPAM and IA backgrounds, so I sought out a project that would give me an opportunity to do exactly that.

Several months before I started my graduation project, I grew increasingly worried about my professional future, as many people consider a graduation project to be the first step toward a future occupation or at the very least one's business card in approaching future employers. I decided I wanted to do my graduation project at a company, so that I would be able to explore professional life in the process.

Eventually, in an effort to consolidate all of these demands and wishes, I got in contact with TJELP, which has proven to be the ideal company both to complete my graduation project and to broaden my horizons regarding my professional future. Moreover, I found a graduation committee willing to indulge my ambitions of a both theoretical and practical graduation project. And after more than two months of conceptual swing tides, we had laid

the foundations for this thesis. With hindsight, I must admit that my ambitions were too elaborate. Instead of exploring best practices, designing a prototype, implementing it *and* testing its performance, I should have focused on a smaller task. That way, I might have been less naive with respect to, for instance, the dataset, which appeared to be unrepresentative and too sparse to actually *validate* our technical design of an Expertise Identification (EI) system.

13-2 Practical Dos and Don'ts

I have never before undertaken any project like this graduation project. And non-surprising, by the end of this project, I had quite a list of dos and don'ts for both future graduate students and myself. Here, I will discuss the most important of these take-aways.

During my graduation project, I had two major issues with planning: *the summer* and *developing the EI prototype*. I always said I would not be so ignorant as to plan part of my graduation project during the summer[30], as many supervisors and company employees can be expected to go on vacation and overall productivity is usually lower due to the weather. In the end, I ended up doing exactly that. Only to discover that all my fears were correct: the project slowed down considerably, mostly due to my own lowered productivity. I *strongly* recommend to *not* plan (part of) a graduation project in the summer. Furthermore, I had issues with consolidating the planning and execution of my literature research (Part 1 of this thesis) with the continuous development of the EI prototype system. I did not really *plan* the latter, since the actual software I had to develop would not be a part of the thesis contents. With hindsight, I should have reserved one month near the end of the project, to create the whole prototype, instead of working on it for a day or so every week.

Although I now believe that I should have planned the development of the EI prototype more carefully, I do not regret being hands-on with data from day one. Continuously playing around with the data enabled me to try different ideas and strategies for EI in an ad-hoc fashion, and made the graduation project a lot more interesting.

I am really grateful for all of the help, insights and moral support TJELP employees have given me during this project. I believe that their contributions, together with a great external office location to do a lot of my work, greatly increased the quality of my work. And during busy times at their office, I was allowed to work on other – more quiet – locations just the same. There is one nuance to my graduating at TJELP, which is consolidating the wishes, ideas and enthusiasm of TJELP employees with the down-to-earth, realistic, scientific perspective of my supervisors at the university. At times, it was difficult to get carried away in the enthusiastic brainstorming sessions at TJELP, and then having to leave a great deal of TJELP ideas out of scope due to time constraints and the requirements on academic research.

If I had to choose all over again, I would choose to do my graduation project with TJELP again.

[30] *"The Prime Directive is not just a set of rules. It is a philosophy, and a very correct one. History has proven again and again that whenever mankind interferes with a less developed civilization, no matter how well intentioned that interference may be, the results are invariably disastrous."* –Jean-Luc Picard, Symbiosis.

Writing my thesis using Latex[31], and more specifically the *Latexian* editor, has proven to be both extremely useful and powerful. Besides the professional layout obtained using Latex, I especially found the automatic labeling and referencing to come in handy.

Lastly, I used Wunderkit[32] to manage my thesis todos. Doing so enabled me to quickly aggregate all of the different comments and suggestions made by my supervisors. Moreover, it forced me to carefully plan deadlines for (sets of) improvements to my work.

[31]I actually used a TU Delft template available for Latex. If you are interested in this template, send me an e-mail at P.R.Oldenzeel@student.tudelft.nl

[32]Visit http://wunderkit.com for more information on the Wunderkit project- and task management software.

Appendix A

Assumptions

This appendix lists the most important assumptions in this thesis, in order to maintain a clear overview.

A-1 Representative Knowledge in Enterprise Social Media

One could argue that if an organization's Enterprise Social Media (ESM) system does not contain *all* knowledge present in that organization, it is impossible to reliably measure employees' subject area knowledge in any way. We argue, however, that following this line of reasoning, ESM systems (or any other digital environment) will never contain a fully representative reflection of the collective and individual knowledge in an organization. Consequently, we assume that an organization's ESM system contains a representative reflection of that collective and individual knowledge. In reality, many other modes of communication will be used as well. Moreover, during the adoption of an ESM system, only some of an organization's employees can be expected to participate actively in the system.

A-2 Tags Representative for Annotated Full-Content

We assume that tags are representative for the annotated full-content. So the tags present in a post provide an *abstraction* of the information in that post. In reality, tags are often used for personal categorization, which means that not all tags add to this abstraction of full-content.

A-3 Tags Formatted Neatly

We assume that tags used as content abstraction data have been formatted neatly, as is the case in E-view. However, in real-life, especially when using a folksonomy that allows users

to freely organize their tags, synonyms, typos and other inherent problems are very common. And although we expect the tag collection and users' tag vocabularies to converge into a fairly neat folksonomy after some time and extensive tag use, tags can be expected to be messy during early adoption of an ESM system.

A-4 Expertise Decay

In chapter 5, we discussed the decay of expertise. The rate of this decay is strongly dependent on the knowledge domain. Expertise on computer software, for instance, may become outdated more quickly than expertise on automobiles. Because determining this expertise decay rate requires more research into the variables at play, in this thesis, we choose a fixed expertise half-life of 5 years. This means that the expertise score of a piece of content decays with approximately 13% every year.

A-5 Equal weights

The constructed Expertise Identification (EI) prototype enables separate weights for (components of) individual pieces of content as well as appreciation data. In our technical design in chapter 8, as well as our tests in chapter 9, we assumed equal weights for all these components. This means, for instance, that relevance and appreciation scores of comments are weighted equally to relevance and appreciation scores of the main text in a dialog. Votes-up have equal weights as votes-down. Votes are equally important as thank you's and flags. Finally, different dialog types are also weighted equally. A Q&A post, for instance, is given the same weights as a news message.

Appendix B

E-view's Social Graph

As we explained in chapter 8, our Expertise Identification system extracts all relationships between users and content in E-view into a social graph in OrientDB. In chapter 8, we displayed a figure showing the schema of that social graph. In this appendix, we describe a visualization of the actual graph, from the perspective of a tag in E-view.

Figure B-1 shows all the relationships in E-view with the tag 'bug' (the center object). Edges have different colors, accentuating different relationships. As can be expected, the tag has relationships with users, comments and dialogs. Users tagged comments and dialogs with the tag a great number of times. The figure shows the traversal paradigm of a graph database like OrientDB: vertices and edges can be traversed from the perspective of *one* object, be it a vertex, such as a tag, or an edge, such as a 'created' relationship.

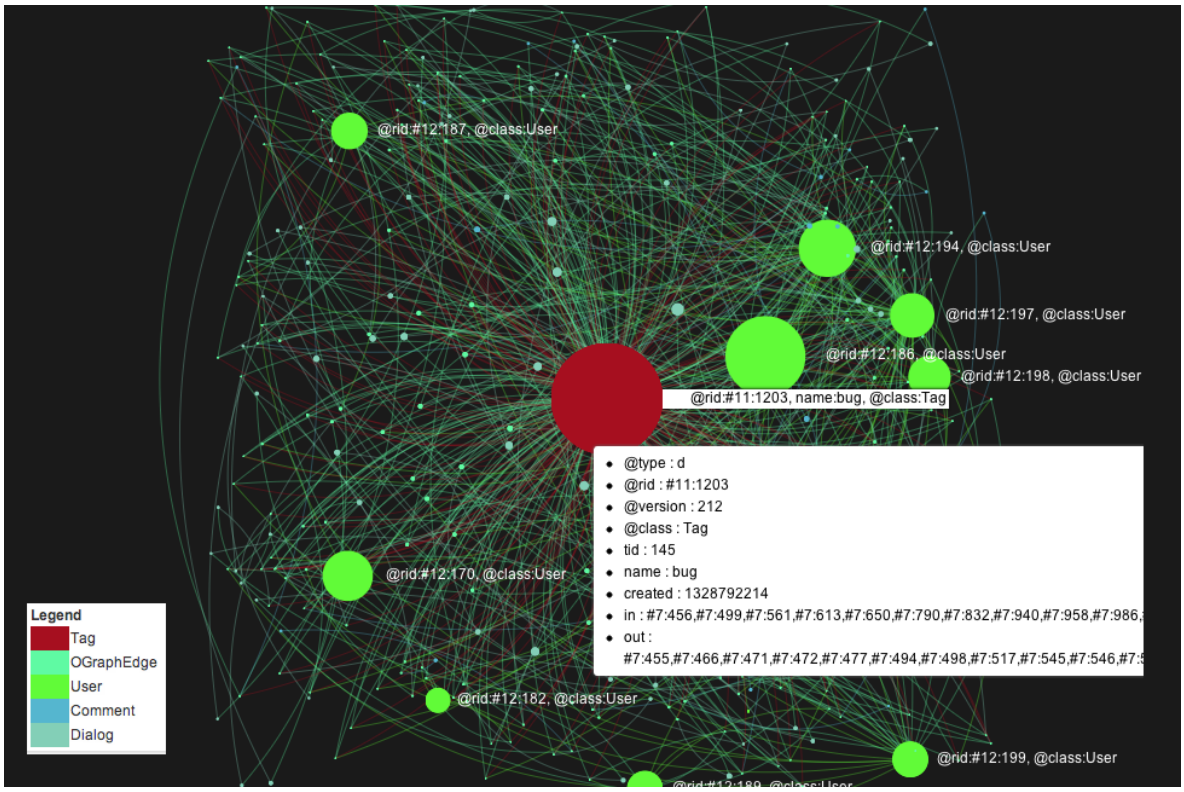


Figure B-1: Visualized social graph in OrientDB from the perspective of the tag 'bug'.

Appendix C

Strategies

In this appendix, we present and explain the full output of the Expertise Identification (EI) strategies we formulated in chapter 9 as generated by our EI prototype. For each strategy, we constructed a table containing the *descriptive output* and the *expertise output*.

C-1 Descriptive Output

The *descriptive output* is displayed for each table component and describes the *total number of posts* retrieved from Solr, as well as the portion of those posts for which users provided *appreciation data* and the *maximum appreciation data score* rewarded to a single post. In subsection 8-1-2, we explained that the appreciation score of a post is calculated by accumulating all votes, thank you's and flags related to that post. The *maximum appreciation data score* represents the highest of these scores for a single post.

C-2 Expertise Output

The *expertise output* for each strategy shows the *number of dialogs* and *the number of comments* individual users posted. Furthermore, it shows the *sum of the normalized relevance scores* and the *sum of the normalized appreciation data scores* of each user's posts. Finally, the outer right column combines these results into an *expertise score*. Users with an expertise score of 0, meaning they created *zero* posts relevant to the query subject, are not displayed in the table.

C-3 Settings

The prototype *settings* of the various strategies – weights and ratios – are shown in table C-1, and contain the:

1. *overall content relevance weight* and *overall appreciation data weight*, which together form the ratio between the importance of content relevance versus appreciation data in estimating expertise;
2. Solr field weights used in selecting E-view posts relevant to the query subject;
3. appreciation data weights to configure the importance of appreciation data in *dialogs* versus *comments*, as well as the importance of *positive* versus *negative* appreciation data.

These settings are mostly homogenous throughout all strategies, because of our assumption of equal weights, explained in section 9-1 and appendix A.

	Expertise Identification Strategies					
Setting	CR_T	CR_{FC}	CR_{FC+T}	$(CR+A)_T$	$(CR+A)_{FC}$	$(CR+A)_{FC+T}$
Overall content relevance weight	1	1	1	1	1	1
Overall appreciation data weight	0	0	0	1	1	1
Solr 'dialog title' weight	1	1	1	1	1	1
Solr 'dialog content' weight	1	1	1	1	1	1
Solr 'dialog tags' weight	1	1	1	1	1	1
Solr 'comment content' weight	1	1	1	1	1	1
Solr 'comment tags' weight	1	1	1	1	1	1
Appreciation data 'dialog' weight	1	1	1	1	1	1
Appreciation data 'comment' weight	1	1	1	1	1	1
Appreciation data 'positive' weight	1	1	1	1	1	1
Appreciation data 'negative' weight	1	1	1	1	1	1

Table C-1: Settings for the tested Expertise Identification strategies.

The various strategies are explained more elaborately in subsection 9-1-1. Figure C-2 shows all possible strategies for EI in E-view using our EI prototype, and table C-1 shows the strategies we tested to gain insight into EI performance. Both the figure and the table are copied from chapter 9. The table has been expanded with references to the tables with raw output for each strategy shown in the remainder of this chapter.

#	Name	Scoring method	Selection method	Raw output table
7	CR_T	content relevance	tags	C-3
8	CR_{FC}	content relevance	full-content	C-4
9	CR_{FC+T}	content relevance	full-content and tags	C-5
1	$(CR+A)_T$	content relevance and appreciation data	tags	C-6
2	$(CR+A)_{FC}$	content relevance and appreciation data	full-content	C-7
3	$(CR+A)_{FC+T}$	content relevance and appreciation data	full-content and tags	C-8

Table C-2: Tested strategies for Expertise Identification in E-view, using our Expertise Identification prototype.

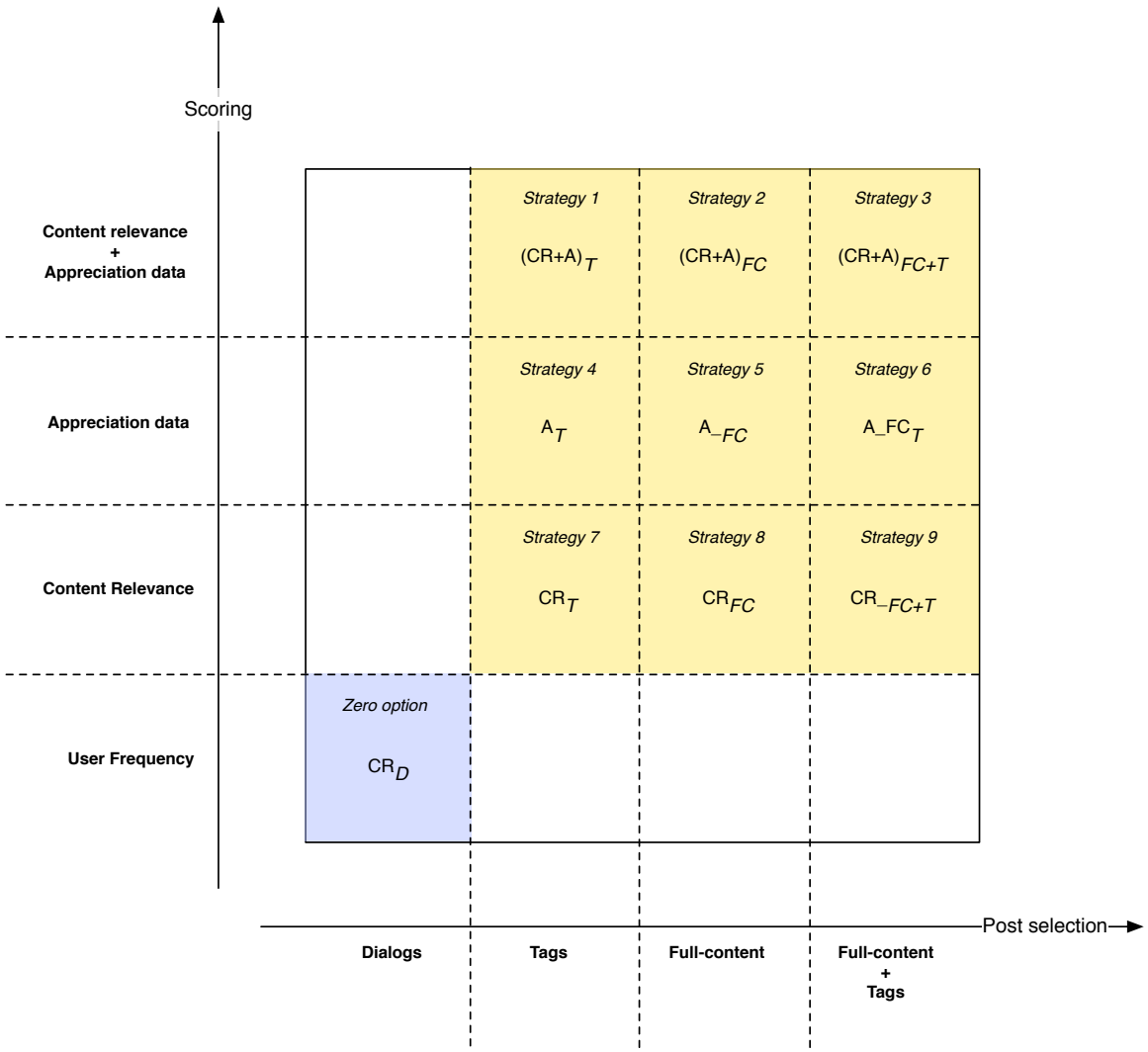


Figure C-1: The main strategies for identifying experts as available in E-view using our Expertise Identification prototype.

Enexis			369 posts, 13 with appreciation data, max appreciation 2.		
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	38	96	46.04	2.00	46.04
willem.jacobs@tjelp.com	31	73	33.69	2.00	33.69
henno.janmaat@tjelp.com	17	55	28.28	1.00	28.28
rempko.de.bie@tjelp.com	0	19	5.45	1.00	5.45
freek.pino@tjelp.com	0	4	1.14	0.50	1.14
justus.brugman@tjelp.com	0	4	1.07	0.50	1.07
ruud.prein@tjelp.com	0	2	0.69	0.00	0.69
Weteringschans			53 posts, 4 with appreciation data, max appreciation 2.		
willem.jacobs@tjelp.com	4	10	5.61	0.00	5.61
henno.janmaat@tjelp.com	1	3	2.40	0.50	2.40
rempko.de.bie@tjelp.com	1	3	1.99	0.00	1.99
justus.brugman@tjelp.com	0	4	1.67	0.00	1.67
ruud.prein@tjelp.com	0	4	1.67	0.00	1.67
niels.kooi@tjelp.com	2	1	1.35	0.00	1.35
Bug			316 posts, 27 with appreciation data, max appreciation 3.		
henno.janmaat@tjelp.com	36	36	49.45	1.33	49.45
rempko.de.bie@tjelp.com	6	44	24.94	0.00	24.94
niels.kooi@tjelp.com	17	28	23.08	1.33	23.08
justus.brugman@tjelp.com	4	37	20.79	1.33	20.79
willem.jacobs@tjelp.com	13	17	16.31	-1.67	16.31
ruud.prein@tjelp.com	3	32	16.22	1.67	16.22
freek.pino@tjelp.com	2	7	4.86	0.00	4.86
Drupal			63 posts, 14 with appreciation data, max appreciation 2.		
justus.brugman@tjelp.com	3	5	4.15	1.00	4.15
ruud.prein@tjelp.com	1	7	3.99	1.00	3.99
niels.kooi@tjelp.com	2	4	2.75	0.00	2.75
rempko.de.bie@tjelp.com	2	1	2.04	0.00	2.04
henno.janmaat@tjelp.com	1	1	1.22	0.00	1.22
willem.jacobs@tjelp.com	0	3	1.22	0.00	1.22
freek.pino@tjelp.com	0	1	0.51	0.00	0.51

Table C-3: Output for strategy CR_T.

Enexis	148 posts, 9 with appreciation data, max appreciation 2.				
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	32	29	21.92	2.00	21.92
henno.janmaat@tjelp.com	29	12	17.98	2.50	17.98
willem.jacobs@tjelp.com	10	10	5.25	0.50	5.25
rempko.de.bie@tjelp.com	2	4	1.73	0.00	1.73
freek.pino@tjelp.com	2	3	0.73	0.00	0.73
ruud.prein@tjelp.com	0	1	0.24	0.00	0.24
justus.brugman@tjelp.com	1	0	0.19	0.00	0.19
Weteringschans	33 posts, 2 with appreciation data, max appreciation 1.				
willem.jacobs@tjelp.com	7	6	4.55	0.00	4.55
rempko.de.bie@tjelp.com	6	2	1.95	1.00	1.95
niels.kooi@tjelp.com	2	2	0.82	1.00	0.82
freek.pino@tjelp.com	0	2	0.46	0.00	0.46
henno.janmaat@tjelp.com	0	2	0.39	0.00	0.39
ruud.prein@tjelp.com	0	1	0.26	0.00	0.26
justus.brugman@tjelp.com	1	0	0.18	0.00	0.18
Bug	113 posts, 17 with appreciation data, max appreciation 1.				
henno.janmaat@tjelp.com	42	8	15.34	6.00	15.34
rempko.de.bie@tjelp.com	7	8	3.99	2.00	3.99
niels.kooi@tjelp.com	5	10	3.90	3.00	3.90
ruud.prein@tjelp.com	5	10	3.71	2.00	3.71
freek.pino@tjelp.com	1	2	1.20	0.00	1.20
justus.brugman@tjelp.com	3	3	1.06	1.00	1.06
willem.jacobs@tjelp.com	2	1	0.59	0.00	0.59
Drupal	81 posts, 14 with appreciation data, max appreciation 3.				
rempko.de.bie@tjelp.com	8	6	2.87	1.00	2.87
justus.brugman@tjelp.com	9	3	2.74	1.67	2.74
ruud.prein@tjelp.com	3	9	2.51	0.33	2.51
willem.jacobs@tjelp.com	0	8	1.05	0.00	1.05
henno.janmaat@tjelp.com	2	1	0.93	0.00	0.93
niels.kooi@tjelp.com	2	4	0.78	0.33	0.78
freek.pino@tjelp.com	0	3	0.72	0.00	0.72

Table C-4: Output for strategy CR_{FC}.

Enexis	462 posts, 20 with appreciation data, max appreciation 2.				
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	48	118	47.70	3.00	47.70
henno.janmaat@tjelp.com	36	65	32.48	3.50	32.48
willem.jacobs@tjelp.com	36	79	31.06	2.50	31.06
rempko.de.bie@tjelp.com	2	20	5.45	1.00	5.45
freek.pino@tjelp.com	2	7	1.44	0.50	1.44
justus.brugman@tjelp.com	1	4	1.04	0.50	1.04
ruud.prein@tjelp.com	0	3	0.75	0.00	0.75
Weteringschans	80 posts, 6 with appreciation data, max appreciation 2.				
willem.jacobs@tjelp.com	8	16	7.42	0.00	7.42
rempko.de.bie@tjelp.com	6	5	2.71	0.50	2.71
henno.janmaat@tjelp.com	1	5	2.63	0.50	2.63
ruud.prein@tjelp.com	0	5	1.81	0.00	1.81
justus.brugman@tjelp.com	1	4	1.77	0.00	1.77
niels.kooi@tjelp.com	2	3	1.61	0.50	1.61
freek.pino@tjelp.com	0	2	0.27	0.00	0.27
Bug	370 posts, 35 with appreciation data, max appreciation 3.				
henno.janmaat@tjelp.com	44	44	42.05	2.33	42.05
rempko.de.bie@tjelp.com	12	51	21.50	0.67	21.50
niels.kooi@tjelp.com	18	34	18.93	1.67	18.93
justus.brugman@tjelp.com	5	38	15.50	1.33	15.50
ruud.prein@tjelp.com	6	36	13.72	2.00	13.72
willem.jacobs@tjelp.com	14	18	12.26	-1.67	12.26
freek.pino@tjelp.com	2	9	4.07	0.00	4.07
Drupal	124 posts, 23 with appreciation data, max appreciation 3.				
ruud.prein@tjelp.com	4	14	5.06	1.00	5.06
justus.brugman@tjelp.com	10	7	4.92	1.67	4.92
rempko.de.bie@tjelp.com	8	7	3.62	1.00	3.62
niels.kooi@tjelp.com	3	8	2.80	0.33	2.80
willem.jacobs@tjelp.com	0	11	1.83	0.00	1.83
henno.janmaat@tjelp.com	2	2	1.37	0.00	1.37
freek.pino@tjelp.com	0	3	0.73	0.00	0.73

Table C-5: Output for strategy CR_{FC+T}.

Enexis			369 posts, 13 with appreciation data, max appreciation 2.		
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	38	96	46.04	2.00	48.04
willem.jacobs@tjelp.com	31	73	33.69	2.00	35.69
henno.janmaat@tjelp.com	17	55	28.28	1.00	29.28
rempko.de.bie@tjelp.com	0	19	5.45	1.00	6.45
freek.pino@tjelp.com	0	4	1.14	0.50	1.64
justus.brugman@tjelp.com	0	4	1.07	0.50	1.57
ruud.prein@tjelp.com	0	2	0.69	0.00	0.69
Weteringschans			53 posts, 4 with appreciation data, max appreciation 2.		
willem.jacobs@tjelp.com	4	10	5.61	0.00	5.61
henno.janmaat@tjelp.com	1	3	2.40	0.50	2.90
rempko.de.bie@tjelp.com	1	3	1.99	0.00	1.99
justus.brugman@tjelp.com	0	4	1.67	0.00	1.67
ruud.prein@tjelp.com	0	4	1.67	0.00	1.67
niels.kooi@tjelp.com	2	1	1.35	0.00	1.35
Bug			316 posts, 27 with appreciation data, max appreciation 3.		
henno.janmaat@tjelp.com	36	36	49.45	1.33	50.79
rempko.de.bie@tjelp.com	6	44	24.94	0.00	24.94
niels.kooi@tjelp.com	17	28	23.08	1.33	24.42
justus.brugman@tjelp.com	4	37	20.79	1.33	22.12
ruud.prein@tjelp.com	3	32	16.22	1.67	17.89
willem.jacobs@tjelp.com	13	17	16.31	-1.67	14.65
freek.pino@tjelp.com	2	7	4.86	0.00	4.86
Drupal			63 posts, 14 with appreciation data, max appreciation 2.		
justus.brugman@tjelp.com	3	5	4.15	1.00	5.15
ruud.prein@tjelp.com	1	7	3.99	1.00	4.99
niels.kooi@tjelp.com	2	4	2.75	0.00	2.75
rempko.de.bie@tjelp.com	2	1	2.04	0.00	2.04
henno.janmaat@tjelp.com	1	1	1.22	0.00	1.22
willem.jacobs@tjelp.com	0	3	1.22	0.00	1.22
freek.pino@tjelp.com	0	1	0.51	0.00	0.51

Table C-6: Output for strategy $(CR + A)_T$.

Enexis	148 posts, 9 with appreciation data, max appreciation 2.				
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	32	29	21.92	2.00	23.92
henno.janmaat@tjelp.com	29	12	17.98	2.50	20.48
willem.jacobs@tjelp.com	10	10	5.25	0.50	5.75
rempko.de.bie@tjelp.com	2	4	1.73	0.00	1.73
freek.pino@tjelp.com	2	3	0.73	0.00	0.73
ruud.prein@tjelp.com	0	1	0.24	0.00	0.24
justus.brugman@tjelp.com	1	0	0.19	0.00	0.19
Weteringschans	33 posts, 2 with appreciation data, max appreciation 1.				
willem.jacobs@tjelp.com	7	6	4.55	0.00	4.55
rempko.de.bie@tjelp.com	6	2	1.95	1.00	2.95
niels.kooi@tjelp.com	2	2	0.82	1.00	1.82
freek.pino@tjelp.com	0	2	0.46	0.00	0.46
henno.janmaat@tjelp.com	0	2	0.39	0.00	0.39
ruud.prein@tjelp.com	0	1	0.26	0.00	0.26
justus.brugman@tjelp.com	1	0	0.18	0.00	0.18
Bug	113 posts, 17 with appreciation data, max appreciation 1.				
henno.janmaat@tjelp.com	42	8	15.34	6.00	21.34
niels.kooi@tjelp.com	5	10	3.90	3.00	6.90
rempko.de.bie@tjelp.com	7	8	3.99	2.00	5.99
ruud.prein@tjelp.com	5	10	3.71	2.00	5.71
justus.brugman@tjelp.com	3	3	1.06	1.00	2.06
freek.pino@tjelp.com	1	2	1.20	0.00	1.20
willem.jacobs@tjelp.com	2	1	0.59	0.00	0.59
Drupal	81 posts, 14 with appreciation data, max appreciation 3.				
justus.brugman@tjelp.com	9	3	2.74	1.67	4.41
rempko.de.bie@tjelp.com	8	6	2.87	1.00	3.87
ruud.prein@tjelp.com	3	9	2.51	0.33	2.84
niels.kooi@tjelp.com	2	4	0.78	0.33	1.11
willem.jacobs@tjelp.com	0	8	1.05	0.00	1.05
henno.janmaat@tjelp.com	2	1	0.93	0.00	0.93
freek.pino@tjelp.com	0	3	0.72	0.00	0.72

Table C-7: Output for strategy $(CR + A)_{FC}$.

Enexis	462 posts, 20 with appreciation data, max appreciation 2.				
Author	**Dialogs posted**	**Comments posted**	**Sum normalized relevance scores**	**Sum normalized appreciation data scores**	**Expertise score**
niels.kooi@tjelp.com	48	118	47.70	3.00	50.70
henno.janmaat@tjelp.com	36	65	32.48	3.50	35.98
willem.jacobs@tjelp.com	36	79	31.06	2.50	33.56
rempko.de.bie@tjelp.com	2	20	5.45	1.00	6.45
freek.pino@tjelp.com	2	7	1.44	0.50	1.94
justus.brugman@tjelp.com	1	4	1.04	0.50	1.54
ruud.prein@tjelp.com	0	3	0.75	0.00	0.75
Weteringschans	80 posts, 6 with appreciation data, max appreciation 2.				
willem.jacobs@tjelp.com	8	16	7.42	0.00	7.42
rempko.de.bie@tjelp.com	6	5	2.71	0.50	3.21
henno.janmaat@tjelp.com	1	5	2.63	0.50	3.13
niels.kooi@tjelp.com	2	3	1.61	0.50	2.11
ruud.prein@tjelp.com	0	5	1.81	0.00	1.81
justus.brugman@tjelp.com	1	4	1.77	0.00	1.77
freek.pino@tjelp.com	0	2	0.27	0.00	0.27
Bug	370 posts, 35 with appreciation data, max appreciation 3.				
henno.janmaat@tjelp.com	44	44	42.05	2.33	44.38
rempko.de.bie@tjelp.com	12	51	21.50	0.67	22.16
niels.kooi@tjelp.com	18	34	18.93	1.67	20.60
justus.brugman@tjelp.com	5	38	15.50	1.33	16.84
ruud.prein@tjelp.com	6	36	13.72	2.00	15.72
willem.jacobs@tjelp.com	14	18	12.26	-1.67	10.59
freek.pino@tjelp.com	2	9	4.07	0.00	4.07
Drupal	124 posts, 23 with appreciation data, max appreciation 3.				
justus.brugman@tjelp.com	10	7	4.92	1.67	6.58
ruud.prein@tjelp.com	4	14	5.06	1.00	6.06
rempko.de.bie@tjelp.com	8	7	3.62	1.00	4.62
niels.kooi@tjelp.com	3	8	2.80	0.33	3.13
willem.jacobs@tjelp.com	0	11	1.83	0.00	1.83
henno.janmaat@tjelp.com	2	2	1.37	0.00	1.37
freek.pino@tjelp.com	0	3	0.73	0.00	0.73

Table C-8: Output for strategy $(CR + A)_{FC+T}$.

Bibliography

Abel, F., Cardosodearaujo, S., Gao, Q., & Houben, G. (2011). Analyzing cross-system user modeling on the social web. In *Eleventh international conference on web engineering (icwe)*.

Abel, F., Henze, N., Kawase, R., Krause, D., & Siehndel, P. (2010). Tagme!: Enhancing social tagging with spatial context. In J. Filipe & J. Cordeiro (Eds.), *Webist (selected papers)* (Vol. 75, p. 114-128). Springer. Available from http://dblp.uni-trier.de/db/conf/webist/webist2010sp.html#AbelHKKS10

Alavi, M., & Leidner, D. E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly: Management Information Systems*, *25*(1), 107-136.

Amatriain, X., Pujol, J. M., Tintarev, N., & Oliver, N. (2009). Rate it again: Increasing recommendation accuracy by user re-rating. In (p. 173-180).

Amitay, C. D. G. N. H. N. O.-K. S. Y. S., E. (2008, July). Finding people and documents, using web 2.0 data. In *Future challenges in expertise retrieval.*

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., et al. (2010, April). A view of cloud computing. *Commun. ACM*, *53*(4), 50–58. Available from http://doi.acm.org/10.1145/1721654.1721672

Bailey, C. N. d. V. A. S. I., P. (2007). Overview of the trec 2007 enterprise track.

Balog, K. (2007). People search in the enterprise. In *Sigir '07: Proceedings of the 30th annual international acm sigir conference on research and development in information retrieval* (pp. 916–916). New York, NY, USA: ACM. Available from http://portal.acm.org/citation.cfm?id=1277741.1277985&coll=Portal&dl=GUIDE&CFID=36414553&CFTOKEN=91582420

Balog, K. (2008). The sigir 2008 workshop on future challenges in expertise retrieval (fcher). *SIGIR Forum*, *42*(2), 46-52. Available from http://dblp.uni-trier.de/db/journals/sigir/sigir42.html#Balog08

Balog, K., Azzopardi, L., & Rijke, M. de. (2006). Formal models for expert finding in enterprise corpora. In *Sigir '06: Proceedings of the 29th annual international acm sigir conference on research and development in information retrieval* (pp. 43–50). New

York, NY, USA: ACM. Available from http://portal.acm.org/citation.cfm?id=1148181

Balog, K., Soboroff, I., Thomas, P., Bailey, P., Craswell, N., & Vries, A. de. (2008). Overview of the trec 2008 enterprise track.

Becerra-Fernandez, I. (2006). Searching for experts on the web: A review of contemporary expertise locator systems. *ACM Transactions on Internet Technology*, *6*(4), 333-355.

Benz, D., Körner, C., Hotho, A., Stumme, G., & Strohmaier, M. (2011). One tag to bind them all: Measuring term abstractness in social metadata. In (Vol. 6643 LNCS, p. 360-374). Heraklion, Crete.

Bughin, J., & Chui, M. (2010). *The rise of the networked enterprise: Web 2.0 finds its payday* (Tech. Rep.). McKinsey Quartely. Available from http://download.mckinseyquarterly.com/the_rise_of_the_networked_enterprise.pdf

Cai, Y., & Li, Q. (2010). Personalized search by tag-based user profile and resource profile in collaborative tagging systems. In (p. 969-978).

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. In O. Frieder, J. Hammer, S. Qureshi, & L. Seligman (Eds.), (p. 528-531).

Cattuto, C., Benz, D., Hotho, A., & Stumme, G. (2008). Semantic grounding of tag relatedness in social bookmarking systems. In A. P. Sheth et al. (Eds.), *The semantic web – iswc 2008: 7th international semantic web conference, karlsruhe, germany* (Vol. 5318, p. 615-631). Berlin: Springer.

Chappell, T. (2004). *Reading plato's theaetetus*. Hackett Publishing Company. Available from http://books.google.nl/books?id=w-Ku6zwIyJMC

Craswell, N., Vries, A. P. de, & Soboroff, I. (2005). Overview of the trec 2005 enterprise track. In E. M. Voorhees & L. P. Buckland (Eds.), *Trec* (Vol. Special Publication 500-266). National Institute of Standards and Technology (NIST). Available from http://dblp.uni-trier.de/db/conf/trec/trec2005.html#CraswellVS05

Dagostino, D. (2004). *Expertise management: New myths and old realities.* Interview. Available from http://km.brint.com/Expertise_Management.html

Datta, A., Tan Teck Yong, J., & Ventresque, A. (2011). T-recs: Team recommendation system through expertise and cohesiveness. In (p. 201-204).

DBPedia. (2012, May). *The dbpedia data set.* Retrieved 21-05-2012, from http://wiki.dbpedia.org/Datasets

Ericsson, P. M., K., & Cokely, E. (2007). The making of an expert. *Harvard Business Review*, *85*, 114-121.

Farmer, F. R., & Glass, B. (2010). *Building web reputation systems - ratings, reviews and karma to keep your community healthy.* O'Reilly.

Forrester. (2010). *Social networking in the enterprise: Benefits and inhibitors* (Tech. Rep.). Forrester Consulting.

Fu, W. T., Kannampallil, T., Kang, R., & He, J. (2010). Semantic imitation in social tagging. *ACM Transactions on Computer-Human Interaction*, *17*(3).

Fu, Y., Xiang, R., Liu, Y., Zhang, M., & Ma, S. (2007). Finding experts using social network analysis. In (p. 77-80).

Halpin, H., Robu, V., & Shepherd, H. (2007). The complex dynamics of collaborative tagging. In (p. 211-220).

Hennis, T., Lukosch, S., & Veen, W. (2011, December). Reputation in peer-based learning

environments. In O. Santos & J. Boticario (Eds.), *Educational recommender systems and technologies.* Hershey, PA, USA: IGI Global.

Hertzum, M., & Pejtersen, A. M. (2000). Information-seeking practices of engineers: Searching for documents as well as for people. *Information Processing and Management*, *36*(5), 761-778.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve web search? In (p. 195-205).

Holley, R. (2010). Tagging full text searchable articles: An overview of social tagging activity in historic australian newspapers august 2008 — august 2009. *D-Lib Magazine*, *16*(1/2). Available from http://dlib.org/dlib/january10/holley/01holley.html

Hotho, A., Jischke, R., Schmilz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In (Vol. 4011 LNCS, p. 411-426). Budva.

Huang, Z., Chen, H., Guo, F., Xu, J. J., Wu, S., & Chen, W. H. (2006). Expertise visualization: An implementation and study based on cognitive fit theory. *Decision Support Systems*, *42*(3), 1539-1557.

Huang, Z., & Zeng, D. D. (2011). Why does collaborative filtering work? transaction-based recommendation model validation and selection by analyzing bipartite random graphs. *INFORMS Journal on Computing*, *23*(1), 138-152.

Jansen, E. (2010). *A semantic web based approach to expertise finding at kpmg.* Unpublished master's thesis, University of Technology Delft - Computer Sciences.

Kawase, R., Papadakis, G., & Abel, F. (2011). Generating resource profiles by exploiting the context of social annotations. In L. Aroyo et al. (Eds.), *International semantic web conference (1)* (Vol. 7031, p. 289-304). Springer. Available from http://dblp.uni-trier.de/db/conf/semweb/iswc2011-1.html#KawasePA11

Kim, H. N., Roczniak, A., Lévy, P., & El Saddik, A. (2012). Social media filtering based on collaborative tagging in semantic space. *Multimedia Tools and Applications*, *56*(1), 63-89.

Krogh, G. von. (1998). Care in knowledge creation. *California Management Review*, *40*(3), 133-153.

Leavitt, N. (2009). Is cloud computing really ready for prime time? *IEEE Computer*, *42*(1), 15-20. Available from http://dblp.uni-trier.de/db/journals/computer/computer42.html#Leavitt09

Lucene relevancy scoring. (2012, August). Retrieved 01-08-2012, from http://lucene.apache.org/core/old_versioned_docs/versions/3_5_0/scoring.html

McAfee, A. P. (2006). Enterprise 2.0: The dawn of emergent collaboration. *MITSloan Management Review*, *47*(3), 21-28. Available from http://sloanreview.mit.edu/the-magazine/articles/2006/spring/47306/enterprise-the-dawn-of-emergent-collaboration/

McDonald, D. W. (2001). Evaluating expertise recommendations. In *International conference on supporting group work.*

Miles, D. (2011). *Social business systems; success factors for enterprise 2.0 applications* (Tech. Rep.). AIIM Market Intelligence.

Milicevic, A. K., Nanopoulos, A., & Ivanovic, M. (2010). Social tagging in recommender systems: A survey of the state-of-the-art and possible extensions. *Artificial Intelligence Review*, *33*(3), 187-209.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys*, *41*(2), 10:1–10:69.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science*, *5*(1), 14-37.

Nurius, P. S., & Nicoll, A. E. (1992). Capturing clinical expertise: An analysis of knowledge 'mining' through expert system development. *Clinical Psychology Review*, *12*(7), 705-717.

Oosterman, J. (2011). *Finding experts in a corporate environment.* Unpublished master's thesis, TU Delft.

Oxford. (2012, May). *Dictionary pages for knowledge, expertise and skill(s).* Retrieved 01-05-2012, from http://oxforddictionaries.com

Perry, N., Candlot, A., & Schutte, C. (2009). Collaborative knowledge networks emergence for innovation: Factors of success analysis and comparison. *Journal of Decision Systems*, *19*(1), 75-91.

Polanyi, M. (1967). *The tacit dimension.* Garden City, NY: Doubleday.

Raj, N., Dey, L., & Gaonkar, B. (2011). Expertise prediction for social network platforms to encourage knowledge sharing. In (Vol. 1, p. 380-383).

Rastogi, P. N. (2000). Knowledge management and intellectual capital - the new virtuous reality of competitiveness. *Human Systems Management*, *19*(1), 39-48.

Richter, A., & Riemer, K. (2009, December). Corporate social networking sites – modes of use and appropriation through co-evolution. In *Proceedings of the 20th australasian conference on information systems.* Melbourne.

Smirnova, E., & Balog, K. (2011). A user-oriented model for expert finding. In P. Clough et al. (Eds.), *Ecir* (Vol. 6611, p. 580-592). Springer. Available from http://dblp.uni-trier.de/db/conf/ecir/ecir2011.html#SmirnovaB11

Soboroff, I., Vries, A. P. de, & Craswell, N. (2006). Overview of the trec 2006 enterprise track. In E. M. Voorhees & L. P. Buckland (Eds.), *Trec* (Vol. Special Publication 500-272). National Institute of Standards and Technology (NIST). Available from http://dblp.uni-trier.de/db/conf/trec/trec2006.html#SoboroffVC06

Solr relevancy scoring. (2012, August). Retrieved 01-08-2012, from http://wiki.apache.org/solr/SolrRelevancyFAQ

Solskinnsbakk, G., & Gulla, J. A. (2008). Ontological profiles in enterprise search. In A. Gangemi & J. Euzenat (Eds.), *Ekaw* (Vol. 5268, p. 302-317). Springer. Available from http://dblp.uni-trier.de/db/conf/ekaw/ekaw2008.html#SolskinnsbakkG08

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*, Article ID 421425.

Suchanek, F. M., Vojnovic, M., & Gunawardena, D. (2008). Social tags: meaning and suggestions. In *Cikm '08: Proceeding of the 17th acm conference on information and knowledge mining* (pp. 223–232). New York, NY, USA: ACM. Available from http://portal.acm.org/citation.cfm?id=1458082.1458114&coll=ACM&dl=ACM&type=series&idx=SERIES772&part=series&WantType=Proceedings&title=CIKM

Trec track overview. (2012, 08). Retrieved 02-08-2012, from http://trec.nist.gov/tracks.html

Venkateshprasanna, H. M., Gandhi, R. D., Mahesh, K., & Suresh, J. K. (2011). Enterprise search through automatic synthesis of tag clouds.

Weikum, G. (2007). Db&ir: both sides now. In *Sigmod '07: Proceedings of the 2007 acm sigmod international conference on management of data* (pp. 25–30). New York, NY, USA: ACM Press. Available from http://dx.doi.org/10.1145/1247480.1247484

What is orientdb? (2012, August). Retrieved 10-08-2012, from http://code.google.com/p/orient/

Yang, K. W., & Huh, S. Y. (2008). Automatic expert identification using a text categorization technique in knowledge management systems. *Expert Systems With Applications*, *34*(2), 1445-1455.

Yao, J., Cui, B., Han, Q., Zhang, C., & Zhou, Y. (2011). Modeling user expertise in folksonomies by fusing multi-type features. In (Vol. 6587 LNCS, p. 53-67). Hong Kong.

Yeung, C. M. A., Noll, M. G., Gibbins, N., Meinel, C., & Shadbolt, N. (2011). Spear: Spamming-resistant expertise analysis and ranking in collaborative tagging systems. *Computational Intelligence*, *27*(3), 458-488.

Zhang, B., Zhang, Y., & Gao, K. N. (2011). Modeling consensus semantics in social tagging systems. *Journal of Computer Science and Technology*, *26*(5), 806-815.

Zhang, Z. K., Liu, C., Zhang, Y. C., & Zhou, T. (2010). Solving the cold-start problem in recommender systems with social tags. *EPL*, *92*(2).

Zhang, Z. K., Zhou, T., & Zhang, Y. C. (2011). Tag-aware recommender systems: A state-of-the-art survey. *Journal of Computer Science and Technology*, *26*(5), 767-777.

Zyl, A. S. van. (2009). The impact of social networking 2.0 on organisations. *The Electronic Library*, *27*(6), 906-918. Available from http://dblp.uni-trier.de/db/journals/el/el27.html#Zyl09

Glossary

List of Acronyms

ESM	Enterprise Social Media
KM	Knowledge Management
UEPs	User Expertise Profiles
CF	Collaborative Filtering
W3C	World Wide Web Consortium
ELS	Expertise Locator Systems
EF	Expertise Finders
ERS	Expert Recommender Systems
EI	Expertise Identification
TFIDF	term frequency - inverse document frequency
SaaS	Software-as-a-Service
NLP	Natural Language Processing
PDD	Person Description Document
MAE	Mean Absolute Error