Personalized Microblog Search on Twitter

Master's Thesis, April 1st 2014

Shaoyi Duan

<<Page left blank intentionally>>

Personalized Microblog Search on Twitter

THESIS

Submitted in the partial fulfillment of
The requirements for the degree of

MASTER OF SCIENCE

in

COMPUTER SCIENCE
TRACK INFORMATION ARCHITECTURE

by

Shaoyi Duan
Born in Xi'An, China

Web Information Systems
Department of Software Technology
Faculty, EEMCS, Delft University of Technology
Delft, the Netherlands
http://wis.ewi.tudelft.nl

Personalized Microblog Search on Twitter

Author: Shaoyi Duan
Student ID: 4181115
Email: dduan@student.TUDelft.NL

Abstract

With the development of microblogging services, the information sharing process on network has been facilitated. Users can broadcast and share their emotions and opinions on such platforms. Among those microblogging service providers, Twitter is the most notable one and has been used worldwide. Due to the significant amount of information flows on Twitter, the search engine is required to help users in seeking relevant information. However, researchers noticed that people tend to issue short queries on Twitter. Consequently, a "one size fits all" search approach may fail to satisfy uses' particular information needs, since short queries may not effectively describe their information needs. To improve the retrieval effectiveness on Twitter, search results can be personalized based on users' personal interests. Thus, this thesis aims to combine microblog search with personalization techniques in order to incorporate users' specific information needs.

This thesis describes our approach to microblog search personalization. The approach utilizes implicit information about the user's interests to personalize original search results. We first present how do we model users' preferences on Twitter. Subsequently, we investigate different ways to represent search results so that they can be compared with users' preferences. In addition, we provide a set of personalized strategies and evaluate them in two experiments. Furthermore, we compare the performance of our personalized strategies and further analyze their impacts on the retrieval effectiveness. Our research suggests that our personalized search approach can enhance the retrieval performance on Twitter.

To the best of our knowledge, few works have been done in the domain of microblog search personalization. The research work of this thesis incorporated personalization techniques into microblog search and empirically showed the feasibility of search personalization on Twitter.

Keywords: User Modeling, Personalized Search, Microblog Search, Information Retrieval

Graduation Committee:

Chair: Prof. dr. ir. G.J. Houben, Faculty EEMCS[1], TU Delft

Committee Member: Dr. Guido Wachsmuth, Faculty EEMCS, TU Delft

Committee Member: Dr. Claudia Hauff, Faculty EEMCS, TU Delft

Committee Member: Ke Tao, Faculty EEMCS, TU Delft

[1] Electrical Engineering, Mathematics and Computer Science

Preface

This thesis has been produced as my final piece of work for my study at Delft University of Technology. I performed my thesis within the Web Information Systems(WIS) Group. During my time at Delft University of Technology, I had taken courses related to both Business and IT domains as an Information Architect(IA) student. Finally I decided to choose a technical thesis topic to complement my technical background. I have met many challenges during the thesis work, but in the end I found those challenges have given me valuable experiences on tackling research problems.

I would like to take this opportunity to thank my direct supervisor, Ke Tao. This result would not have been possible without his support and guidance. He is always ready to provide comments and push me into the right direction of my research. Secondly, I would like to thank my professor, Geert-Jan Houben, for allowing me to perform my thesis in the WIS group and teaching me the ways of thinking in research.

My thanks also goes to Dr. Qi Gao, who always give me advices when I encounter problems. I could not complete this thesis without him. Furthermore, I would specially thank prof. Claudia Hauff and prof. Alessandro Bozzon. They both provided me with feedback, guidance and critique. I have learnt a lot from their expertise in the field of computer science.

Shaoyi Duan

Delft, the Netherlands

March 28, 2014

Contents

List of Figures

List of Tables

1 Introduction

Microbloging is a new global phenomenon. It provides a light-weight, easy form of communication that enables users to broadcast and share their emotions and opinions [1]. In other words, users can publish brief text updates to describe their current statuses and send them to friends and other interested observers. Twitter is one of the most popular microblog service providers, attracting over 500 million registered users and publishing more than 340 million tweets per day [1]. The Twitter platform has seen a lot of growth since its launch in 2006. It actually facilitates online information sharing activities. For instance, it plays a vital role in broadcasting many breaking news and real-time events [2]. Nowadays, the number of registered users on Twitter is still increasing dramatically. Therefore, both the large scale of the social network as well as the significant amount of information flows reflect the popularity of this new form of information exchange.

Apart from sharing information on microblog platforms, people also show two types of information seeking behavior on microblogs [3]. The first is "Asking for information", while the second is "Retrieving for information". The former refers to the broadcasting of questions to their followers in hopes that people in their social network will answer them, whereas the latter refers to the conducting of searches over microblog data. For instance, Twitter provides a search interface for users to access popular or recent public tweets. This platform processed 340 million posts and 1.6 billion search queries per day[2]. However, an earlier study [4] showed that users differed significantly in how they personally judged the relevance of search results to the same query. This phenomenon can be explained by people's different information needs. Although they may express their needs through the same query, the underlying intents are actually different. Teevan et al. [5] suggest that the current web search approaches can achieve a high performance on satisfying the range of intents people have for a given query, but they have less capability to discern individuals' search goals. Thus, there is an opportunity to improve the retrieval performance by providing tailored results to users.

In addition, microblog search has unique characteristics compared with web search. Teevan et al. [2] present a systematic overview of how search behavior on Twitter differ from web search. Their work reveals that Twitter users often issue short queries to find temporally relevant information. Moreover, the length restriction of Twitter messages lead to a problem in discriminating terms within a given item [6]. Furthermore, the purpose of a tweet may not merely be restricted to share information. Tweets can be published to express personal emotions or opinions. A previous study showed that search results on Twitter include more social content and event information than web search [2]. These alternative purposes pose a challenge to microblog search, because those private contents are of less interest to users with particular information needs. Therefore, we believe there exists a gap between

[2] https://blog.twitter.com/2011/engineering-behind-twitter%E2%80%99s-new-search-experience

the intents of current search approaches and satisfying Twitter users' particular information needs.

To provide tailored search results on Twitter, this thesis aims to incorporate personalization techniques into microblog search. Since users tend to issue short queries on Twitter [2], their queries may not effectively describe one's information needs. Thus, personalized search approach can become a solution to this problem. By collecting information related to an individual's preferences, the personalized search can improve the retrieval performance. For instance, two users are interested in the search topic "2020 Olympics" and try to find recent results from Twitter search engine. However, they may care about different topics related to "2020 Olympics". More specifically, basketball commentaries on Twitter may be more interesting for a basketball fan rather than a tennis enthusiast. In this case, the search engine will not be able to capture such contextual information from the users' queries. Thus, the gathering of users' preferences can facilitate the search process. If the search engine can be aware of the particular information needs of users, it can provide tailored results and thus improve the retrieval performance on Twitter. Meanwhile, the accessibility of wealthy user-generated data enables researchers to gather their interest information implicitly on Twitter.

To the best of our knowledge, few works have been done in the domain of microblog search personalization. We address the problem that a sole query is an insufficient expression of the Twitter users' information needs. In this thesis, we plan to investigate the personalization of Twitter searches based on Twitter user modeling techniques. We propose a framework to achieve Twitter search personalization. Detailed information of our work will be presented in the rest of this thesis.

1.1 Research Questions

To provide tailored search results, the initial research problem is how can we collect users' interest information. In this subsection, we first present the question related to the gathering of users' preferences. Subsequently, the research objective is transformed to adapting results to users' preferences. Thus, the second question related to our personalized approach. Finally, we need to evaluate our approach with Twitter users. The last question was designed for the evaluation of retrieval performances.

1. How can we gather a user's preferences information on Twitter?

   To provide search results which are both topically relevant and of particular interest to users, we should first understand their individual information needs. Since a query is an insufficient expression of information needs, our approach should be able to collect users' preference information on which personalization can be based. Subsequently, we need to determine the preferences information gathering approach as well as the source of individual information.

Current personalized search systems often utilize user information (e.g. name, age or country) or usage information (e.g. browsing history or interacting activities) [7]. We considered individual Twitter activities as the source of preferences information, since users can discuss about any topics they are interested in or concerned with via microblog posts. More specifically, users can customize their information by following others, and retweet or reply to tweets that relevant to their interests. In addition, researchers find that the distributions of topic types differ between Twitter and traditional news media [8]. This finding implies that general user interests on Twitter are relatively unique. Consequently, we believe Twitter activities is the best source to model user interests, since these contents are directly related to users interests on Twitter. Thus, this type of usage information allows us to infer user interests in various domains that make it valuable to personalize results of different search topics. Furthermore, the real-time nature of microblog posts should be taken into consideration during the personalization process. Popular topics as well as user interests evolve over time [8]. Subsequently, Abel et al. reveals that [9] short-term user interests on Twitter are time-sensitive. They found that users' recently concerned concepts can reflect their current interests. Meanwhile, concepts referenced by users can be extracted from their recent Twitter activities. Consequently, individual Twitter activities are worthwhile to personalize search results related to popular search topics.

Generally speaking, usage information can be gathered in an implicit or an explicit manner [7]. However, users may consider the explicit data collecting behavior from online system as violating their privacy [10]. To avoid the privacy issues, we plan to gather the preferences information in an implicit manner. Meanwhile, Twitter provides API operations that allow us to implicitly obtain individual Twitter activities. Thus, we will investigate the modeling of user preferences based on the usage information on Twitter.

After the gathering of preferences information, a representation of individual's interests should be constructed after the gathering of such information. Many previous studies [9, 11] investigate how individual's interests on Twitter can be modeled by different user modeling strategies. These studies have shown that the performance of tweets recommendation can be improved by identifying concepts in user profiles. This thesis adopts their approaches to model Twitter users' interests. In other words, users' preferences are represented as Twitter user profiles in our thesis work.

2. How can we improve the retrieval effectiveness by providing tailored search results to Twitter users?

Proposing a personalized search approach is the core part of this thesis. Because people are not good at specifying detailed information needs, we use in-

formation about the user to infer their implicit intents. Given the user's implicit intents, the subsequent question is how can we utilize this information during the search process. Generally speaking, existing methods achieve personalization mainly by three means [12]: 1) query-adaption, 2) results re-ranking and 3) items filtering. In our thesis work, we choose the re-ranking method based on our design context.

In addition to the general method, we propose four strategies based on the available options in design space. The customized results list is provided based on the similarity between the user profile and original search results. However, there are different ways to generate representations of both the user and the tweet. We thus investigate different combinations of these two types of representations in this thesis. In addition, we plan to compare the performances among different strategies for evaluation.

3. How can we evaluate the personalized search approach on Twitter?

Apart from conventional web search, evaluation of the personalized approach poses a challenge for us, since relevance judgments can only be acquired by enquiring the users. In other words, only the users can subjectively judge whether a specific result satisfies their particular information need. Generally speaking, there are two ways to realize this. The first method is to conduct a user study in which real users are involved. However, user studies are often costly and time-consuming [13]. An alternative method is to utilize the indirect relationship between user and results based on the user's interacting behavior [13, 14].

To investigate the gathering of subjective assessments for Twitter search personalization, we plan to set up two experiments: an online experiment and an offline experiment. These two experiments utilize the explicit and implicit user feedback respectively. We will try to find out whether our personalized strategies can improve the relevance judgment performance in both experiments.

1.2 Contributions

Our work aims to apply personalization techniques in the field of Searching on Twitter. Final results of our approach have shown that it can improve the relevance judgment performance. In this thesis we make the following contributions:

First, we investigated the mining of users' search preference by implementing Twitter-based user modelling approaches. These approaches have been proposed in previous studies [9, 15] to model user interests on the Twitter platform. Our thesis work has shown that us-

er profiles constructed by those approaches can provide feasible user representation in the personalized search process.

Second, we developed a set of approaches for search personalization strategies on the Twitter platform. By exploring the semantic enrichment and concepts classification techniques, we achieved the purpose of improving preference relevance judgment for Twitter search. Our design space includes user modeling, resource profiling and similarity measurement.

Last but not least, we presented two evaluation approaches for personalized search on Twitter. The first approach seeks the collection of relevance judgments from real users for individual results. It is based on a general evaluation framework proposed by Vallet et al. [16]. The second approach utilizes users' implicit feedback. We considered users' interactions on related result as personal relevance judgments. In terms of the performance of our strategies in both experiments, we get the following results:

- Our personalization approach has a positive effect on relevance judgment performance in general.
- Semantic enrichment, which aims to identify meaningful concepts of the actual content, can increase the effectiveness of personalized microblog search.

1.3 The Approach

A brief summary of our methodology is: firstly, a literature survey has been conducted to provide further knowledge about recent researches. Then we transform the issue to developing a re-ranking approach based on twitter user modeling techniques. In terms of evaluating the performance of our approach, we first create a tweets dataset with 47 search topics along with associate search results returned by the Twitter search engine. Subsequently, we conduct online and offline evaluations with two groups of users. A user study framework is designed for the online evaluation experiment to check the performance according to feedback from users. In terms of the offline experiment, we choose users from the search topic dataset who have displayed interacting behavior with the results. Those search results are used as ground truth to evaluate the performance of the approach.

1.4 Thesis outline

The rest of this thesis is organized as follows: Chapter 2 states the existing studies about our thesis topic. In Chapter 3, we present our main design for search personalization on Twitter at both a high level and detailed level. The evaluation is performed in Chapter 4, in which we present two evaluation methods. The conclusion and future works are discussed in Chapter 5.

2 Related Work

Chapter 2 introduces works that are related to this thesis. We will first briefly introduce the research works done about Twitter. Then we summarize the existing studies related to our thesis topic. Since we try to combine microblog search and search personalization, we will introduce related works in both domains. Three main components of our related work are: information needs on Twitter, microblog search and search personalization.

2.1 Twitter

The popularity of Twitter has not only attracted attention of the general public but also that of researchers. They have started to understand users' intentions [1], closely look at how trending topics evolve [17], and found the topical differences between Twitter and traditional news media [18]. Furthermore, some researchers conduct content analysis on information needs extracted from users [19]. To summarize, recent studies examined Twitter from different perspectives and explain the underlying causes of the microblogging phenomena.

In addition, the current microblog search engine provides the user with the ability to reach a large amount of new information. Although the users can be exposed to a large amount of information on Twitter, some of that information may be redundant. To provide relevant search results to users, this thesis suggests that combining the microblog search engine with personalization service can result in a better retrieval performance.

2.2 Microblog Search

Microblog search is the first component of our related work. Several studies examined users' motivations and search behaviors. Teevan et al. [2] reveal that people's motivation to search Twitter is to find temporally relevant information and information related to celebrities. They also compare users' search behaviors on Twitter with those conducted on web search engines, and found that Twitter search focusses more on monitoring content, whereas Web search is used to get information about a topic. In addition, Twitter queries are more common, are repeated more, and change less than Web queries.

To understand and subsequently make use of Twitter (or microblogging in general), an initial question is what kind of useful information is required on Twitter by users. Previous works focus on information seeking behaviors and individual behavioral patterns of Twitter users and how their needs differ from conventional web search [2, 19]. Teevan et al. [2] report three types of information that users are willing to seek on Twitter: 1) timely information (e.g. news, trending topics, summaries of events), 2) social information (e.g. information related to other Twitter users, people's overall opinions on particular topics) and 3) topical information (public sentiment about topics of interest).

Previous studies investigated the differences between microblog search and web search. Zhao et al. [18] empirically compared the content of Twitter with a typical traditional news media. They found that Twitter and traditional news media cover a similar range of topical categories, but the distributions of different topical categories differ. For instance, Twitter users care more on personal life and pop culture than world events. And although Twitter users show a relatively low interest in world news, they actively help spreading news of important world events. In addition, a large-scale analysis of information needs on Twitter by Zhang et al. [19] illustrates that information needs on Twitter are likely to be socially driven rather than information driven. Their work has shown that information needs detected on Twitter have a considerable power of predicting the trends of search engine queries. The availability of large-scale user-generated content on Twitter has provided a decent platform for this type of analysis. Meanwhile, such works can facilitate several tasks such as personalization [20], query expansion [21] or advertising [22]. Furthermore, some previous works have found that the average query length on Twitter (1.64 words) is significantly shorter than those on web search (3.08 words) [2]. This fact implies that users tend to issue short queries on Twitter.

Apart from understanding information needs and search behavior, various retrieval models have been proposed to facilitate the Twitter search process. Naveed et al. [23] address two challenges for microblog search: 1) Sparsity is inherent to microblog documents, since the technical constraints on the message length. Thus, the inherent sparsity result in a problem o discriminating terms within a result, 2) some tweets aim to support social interaction or express emotions rather than communicate information. This nature makes tweets of less interest to a user with a concrete information need. Thus they propose a retrieval model which incorporates term and length features to measure the interestingness of search results on Twitter. Magnani et al. [24] proposed a user-based tree model for retrieving conversations from microblogs. A query-likelihood retrieval model can be used to identify subtopics for further browsing [25]. Lau et al. [21] propose a feature extraction algorithm to capture meaningful pattern of tweet, and also investigate the effectiveness of different features for microblog search.

In addition, different unique features of Twitter are also exploited for retrieval purposes. Efron et al. [26] have shown that hashtags can be used to improve relevance feedback via query expansion. Duan et al. [27] find that the presence of URL correlated to the relevance of a given tweet. Apart from the presence of URL, Tao et al. [28] argue semantics and topic-sensitive features also have influence on the prediction of tweet relevance. Furthermore, the statistics of tweets published, followers count and following-followers ratio can be used to estimate the authority of users to rank and improve the retrieval result [29].

Although microblog search has been studied from various dimensions, there is few works exploiting information from microblog to perform personalized search. This thesis aims to bridge the gap between satisfying users' particular information needs and current microblog search approach.

2.3 Personalized Search

The traditional method used for Information Retrieval is to build an index of the document collection and use this to look up the documents that include keywords submitted as a query [12]. However, it relies solely on the content of the documents (i.e. the keywords they contain) to determine the relevance of a page to a query. Personalized search systems address the limitation that a query may not reflect users' whole information needs. Generally speaking, it aims to build systems that provide individualized collections of pages to the user, based on some form of model representing their needs and the context of their activities. Thus, personalized search systems should be able to keep track of the information needs of their users [7] . Given a particular need, the results are tailored to the preferences, tastes, backgrounds and knowledge of the user [30].

To obtain user information, a personalized system could request that users explicitly supply this information or it could implicitly gather this information from other sources (e.g. query log, click-through analysis and desktop data) [30]. In this thesis, we focus on the use of implicit representations of a user's short-term interests. With this approach to personalization, there is no need for users to specify their interests.

2.3.1 Personalized Web Search.

Personalized web search has been studied extensively. Micarelli et al. [30] classify the current personalized search approaches into two categories: content-based and collaborative-based. Content-based approaches utilize user-generated data or documents (e.g. current working context, search history, user click history) to model user interest, whereas collaborative-based approaches employ users' social relation and based on the assumption that users with similar interest are likely to share the same information needs.

Various content-based approaches have been proposed to achieve search personalization. Raghavan et al. [31] propose an approach to provide tailored results to users by integrating a past queries database, if the similarity between a past query and a current query is significant, the past results which refer to the past query are presented to users. Tan et al. [32] propose a language model approach for query history mining. Their work has shown that the history-based language model can be used to achieve personalization over normal retrieval process. Furthermore, many researchers have investigated building user profiles based on search history for personalization. Liu et al. [33] construct user profiles by mapping users' search history to the Open Directory Project (ODP) category hierarchy. This type of profile will then be used to achieve personalization. Qiu et al. [34] incorporate the history of user click data to achieve search result personalization. Chirita et al. [35] present a personalized algorithm based on the click-through data analysis. This approach utilizes three types of information during the search process: 1) information about the user, 2) the query and 3) the visited pages in the result set. These data are represented as triples to reflect users' interest. Teevan et al. [20] also examine variability in user intent of the query by incorporating large-scale log analysis of user behavior patterns.

Apart from exploring query log and history information to derive user interest, there are studies that utilize desktop data and external sources. For example, Teevan et al. [5] introduce a rich model of user interest, which is built from not only search-related information but also documents on the user's desktop and emails the user has read. This approach modifies the query term weights to incorporate user interests as captured by their desktop indexes. Their research suggests that rich representations of the user can be used to facilitate the degree of personalization. In addition, there are approaches that utilize the current context of the user task. For example, Dou et al. [36] have shown that the variability in results that people click for a query is related to how well they can personalize results for a query.

With regard to the collaborative-based approaches, which aims to deliver relevant resources based on previous ratings by users with similar tastes and preferences [12]. Claypool at al. [37] investigate a possible combination of collaborative and content-based approaches by basing the interest prediction of a document on a weighted average adapted to the individual user. Sugiyama et al. [38] compares two search systems in different scopes. Their work has shown that the community-based system outperformed the individualized system in terms of retrieval accuracy. In addition to the web search personalization systems, collaborative-based approaches can also be found in the social search domain. These approaches focus on employing the user's social network. Previous works such as [10, 39] are typical examples.

2.3.2 Personalized Social Search.

Personalized social search is a search process over "social" data gathered from web applications [39] such as social bookmarking systems, social platforms, forums, and blogs. Recently, many personalized social approaches focus on utilizing social tagging and bookmarking systems (e.g. Flickr, Del.icio.us) also known as "folksonomy". For instance, Bao et al. [40] proposed two algorithms, SocialSimRank and SocialPageRank, both of which incorporate user generated tags and annotations to influence the results set ranking. Xu et al. [14] also focus on utilizing folksonomy for personalized social search. Heymann et al. [41] explored the feasibility of using social book-marking to facilitate web search personalization.

Achieving personalization in social search usually refers to two approaches [13]: query adaptation and result adaptation. The former approach creates a query modification phase after users submit their query, during which the original query is modified based on the user's preferences. Zhou et al. [42] propose a query expansion framework, which incorporates annotation data such as user-generated tags. Their approach constructs the user interest profile by mining the resources a user has marked and annotated. They assume that "the most appropriate expansion terms for a query are likely to be associated with, and influenced by terms extracted from the documents ranked highly for the initial query". Bender et al. [39] propose a unified graph model to represent the users, content, and tags to facilitate the query personalization on social tagging system. Zhou et al. propose a user query modification approach based on the user's tags and bookmarks. They create a statistical

model based on these bookmarks to identify topics in documents, and subsequently use this model to enrich the user query.

The alternative approach focus on the adaptation of results lists. Search personalization can be achieved with result scoring, result re-ranking, or result filtering based on this type of approach. Xu et al [14] present a re-rank approach based on ODP classification and folksonomies. The categorization, keyword, and structure property of the system are explored for the topic space estimation. The relevance of a document is determined not only by the topic relevance between the query and the document, but also by the topic similarity between the user's interests and the web page's topics. In addition, Carmel et al. [13] propose a re-rank approach based on users' social network. A document is first scored by SNaD (Social Networks and Discovery), which is an aggregation tool for information discovery and analysis over the social data, based on its non-personalized scoring mechanism, and then its score is re-ranked based on its relationship with user profile. Wang et al. [10] present a system which constructing user profiles based on their online activities on social systems. The relevance score of original search results is modified by the combination of the topic relevance score and interest similarity score.

These recent studies listed here indicate that the likelihood of a result to a user is based on context and social network of this user. However, their approaches mainly utilize the users' social network relationship and self-generating tags, while the methods proposed in this paper are focused on the constructing of user profiles on Twitter and content-based similarity of results. In this thesis work, we facilitate microblog search personalization by exploiting implicit user activities.

3 Twitter Search Personalization

In this chapter, we go through the design phase of our work. We first study the existing personalized approaches and then make the decision based on our research objectives. Subsequently, we introduce the architecture of our personalized search system. We then describe the core framework of the personalization component. In addition, different design dimensions will be elaborated upon. In the end, we will select and combine different design dimensions and alternatives in order to provide a set of personalized strategies for Twitter search.

3.1 Twinder

Twinder (**Twi**tter Fi**nder**) is a search engine for Twitter streams that aims to improve search for Twitter messages by going beyond keyword-based matching. Fundamentally, it is equipped with Twitter Analysis Language, so that an existing set of tools is available for conducting data analytic tasks with Twitter data [43].

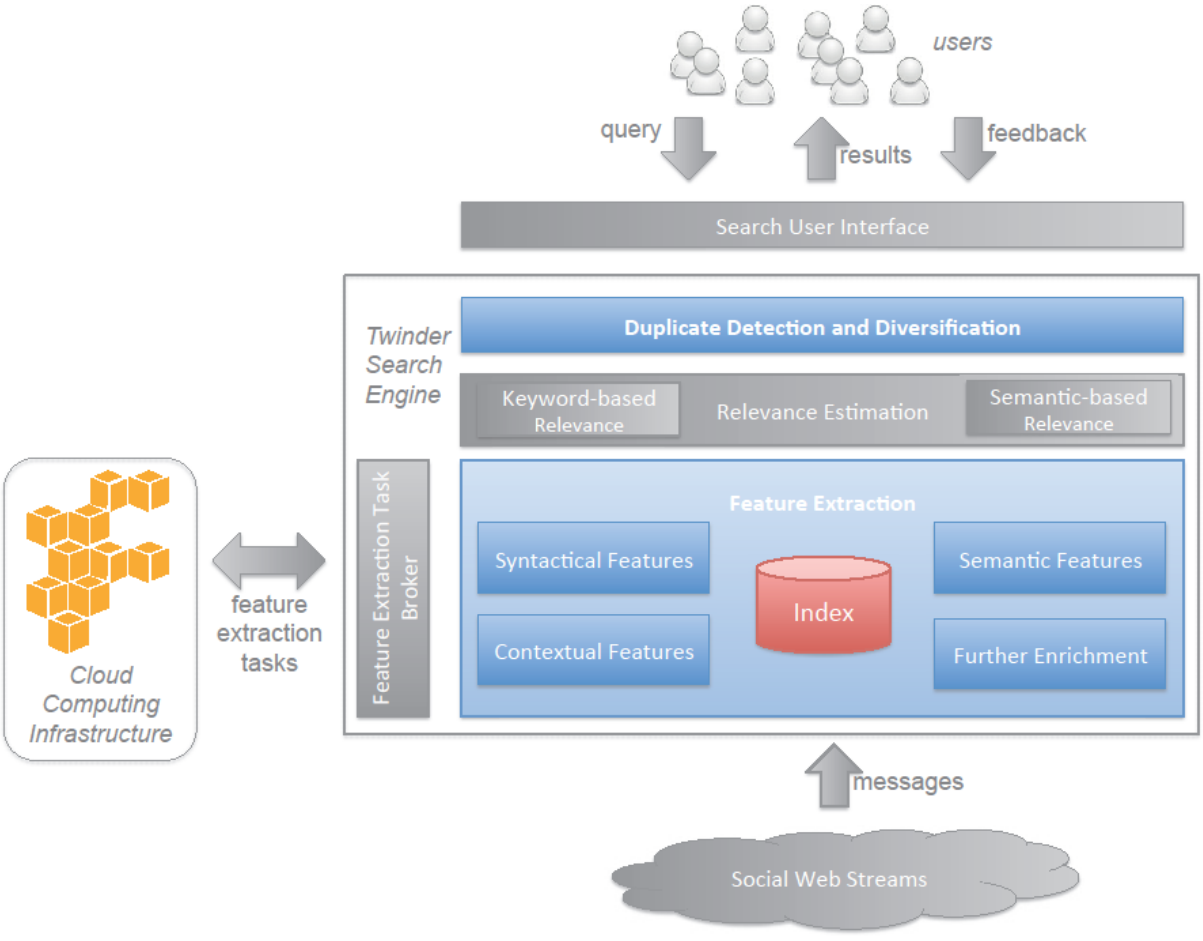

Figure 1 The core components of the Twinder architecture

In the previous version of Twinder, different types of features ranging from syntactical to contextual features are considered by Twinder in order to predict the relevance of tweets for a given search query [43]. Moreover, the duplicate contents from the search results can be detected and removed to achieve better diversity in the search results [44]. Figure 1 shows the core components of the Twinder architecture. Different components are concerned with extracting features from the incoming messages of a Twitter stream. Given the huge amount of Twitter messages that are published every day, the system is designed to be scalable. For this reason, Twinder makes use of cloud computing infrastructures for processing-intensive tasks such as feature extraction and indexing.

This thesis aims to further enhance Twinder search engine with the feature of search results personalization so that it can provide users with the microblog posts that are adapted to their personal preferences.

3.2 The Choice of Personalized Approaches

Search personalization can usually be achieved by means of three main approaches [12]:

1. *Re-ranking*: the reordering of search results to provide a tailored list.
2. *Query expansion/modification*: the augmentation of the user's keyword-based query.
3. *Filtering*: the removal of results that are determined to be irrelevant to the user.

The proposed design solution in this thesis is based on the re-ranking approach. The reasons for utilizing this approach are as follows:

First, the main objective of our thesis work is to provide tailored Twitter search results. Thus we focus on identifying the information that are personally most relevant to an individual user [45]. Apart from the web search, which aims to identify topic-relevant (typically keyword-based) results to a query, the personalized search must deal with the diversified information needs of users rather than solely rely on their queries. Also, there are several unique features (e.g. length of a tweet, presence or absence of a URL or a hashtag, etc.) which may help in determining content relevance in microblog search [27]. However, we consider the filtering approach to be inappropriate, since well-defined and personalized filtering constraints are not available. Proposing a common pattern or a set of rules to identify personally relevant tweets is infeasible for us. In addition, the filtering approach, which removes items based on patterns, has an inherent tendency to exclude groups of relevant results altogether [12].

Second, we addressed the limitation that Twitter's current search engine may not entirely guarantee that the most personally relevant tweets are presented at the top. Twitter's current search engine ranks results based on the chronological order or the popularity [29]. However, this method of ranking emphasizes time constraints or popularity rather than their potential interestingness to users [46]. Given this drawback of Twitter's search engine, we plan to investigate how to improve upon the standard ranked-list presentation of results.

Thus, we try to make a comparison among different ranking methods that have been applied on a given set of tweets. Given this research purpose, the query expansion approach is not the best design alternative. Since it can only affect the ranking by altering the query representation. In other words, the query expansion approach may generate a new ranking with dissimilar results rather than results with relevance-based sequence.

Last but not least, many researchers have shown that exploring users' posts is an efficient way to model users' preferences on Twitter [9, 15]. Their studies inspired us to create user preference representations by constructing user profiles. Meanwhile, the re-ranking approach allows us not only to employ user profiles during the personalization process, but also to apply personalized ranking strategies selectively. Furthermore, the re-ranking approach facilitates straightforward evaluation. To explore different ranking strategies, we only need to collect relevance assessments for the top-N returned results, rather than collecting evaluations for all different rankings. To summarize, the re-ranking approach is the best design alternative based on our research questions. Therefore, these are the reasons why we choose this approach rather than other alternatives in this work.

3.3 Architecture of the Personalized Twitter Search System

We now introduce the re-ranking based search process on Twitter, and explain how personalization can be achieved by incorporating user profile. This process is shown in Figure 2.

When a user conducts a search behavior, the Twitter search engine first returns a set of tweets containing the search query term. This original ranking list is then stored for the second step. During the subsequent step, the re-ranker will find out what those tweets are about, and generate a representation for each item based on their textual content. Before the user submits a query, her user representation (i.e. the user's search preferences) has already been generated. After the profiling tasks on both sides have been accomplished, the personalization module re-ranks the original ranking list. Results are adapted to match the user representation. Finally, the personalization module outputs the new list as a set of personalized results

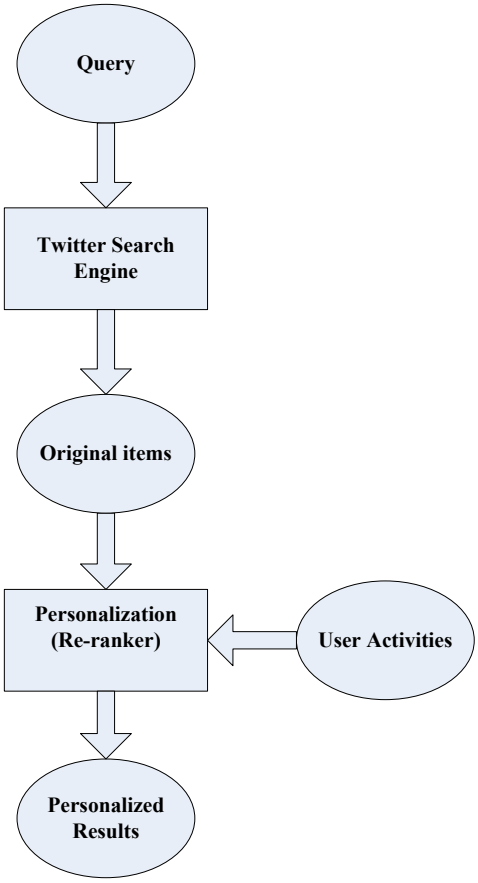

Figure 2 The general personalization process

We now briefly summarize our personalized Twitter search system based on several features. These features are proposed to describe and distinguish a personalized search system by Micarelli et al. [12]. As has been mentioned, our search system is constructed to support the personalized process shown in Figure 2, and can be summarized as follows:

- *The personalization scope*: Ghorab et al. [7] classify the scope on which personalization is performed into three categories: individualized, community-based, and aggregate-level. Our system is a typical individualized one. To identify particular information needs of a specific user, we will gather information related to their individual preference.
- *User data collection method*: As has been mentioned in Chapter 2, there are usually two ways of collecting users' preferences [30]. The first approach is to explicitly collect profile data by asking users directly, whereas the alternative method is to infer their preference implicitly based on their interacting activities. Although collecting explicitly means that we would receive direct feedback from users, some evidence suggests that users generally dislike having to spend time and effort submitting data to any system, especially when they do not benefit [12]. Our system collects user preference data implicitly by exploring their Twitter posts. We take advantage of Twitter API to crawl profiling data. The assumption lies behind it is

that users are interested in the contents that they have. Details of our user modeling process can be found in Section 3.1.1.

- *Profile construction and storage*: Our system constructs and updates user profiles offline in advance of users' search behavior. Profiles are processed and stored on the server-side. This method of construction and storage is in order to reduce the time consumption of the search process.
- *Personalization approach*: As we have explained above, in our system search personalization is achieved by the re-ranking approach. Given a list of (non-personalized) results retrieved for the user's query, search results are re-ranked by considering their relationship strength with user profiles. In Section 3.2, we will present the elaboration of the re-ranking framework.
- *The personalization algorithm*: Our personalized re-ranking algorithm is based on the vector space model [47] . A collection of vectors is generated based on the given set of profiling tweets and items, and form the representation of users' preferences and search results. Finally, similarity scores between the user profile and the items are calculated in order to determine the new sequence of personalized ranking. A further description of our algorithm and associated similarity measures is presented in Section 3.3.3.

3.4 The Framework of Personalization Re-ranker

In this section we present our core framework of the personalization re-ranker. As we explained in Section 3.1, in order to personalize Twitter search results, we plan to calculate the similarity of search items to a given user. Consequently, we propose using a re-ranker to achieve the personalization task. Three important components of our re-rankers are: 1) user representation; 2) item representation and 3) similarity measurement. Further descriptions of these components are given below.

User Representation

Search personalization requires the capability to model users' preferences and interests. To represent a user's search preferences, we employ user-modeling techniques in our thesis work. In other words, users' interactions on Twitter are structured into a user profile that can be utilized during search personalization.

We are inspired from previous studies [9, 11, 15] to develop our user modeling approaches. Users' activities on the Twitter platform are considered as the source of data needed for user profiling. By exploring their activity data, the search system is able to generate a collection of elements extracted from the contents. Subsequently, these elements are weighted by a specific weighting function. The final representation of the user profile is in a vector space model.

Item Representation

Items are tweets posted on Twitter. These posts are a special type of user generated content due to the length constrains. Their characteristics (e.g. short, ungrammatical, and noisy) pose challenges to their representation in search systems [48]. Current microblog search relies heavily on term-based approaches, such as the bag of words model [21]. Each item is a set of pre-processed terms with weight scores. However, the term-based approach is shown to be very sensitive to noise [6]. In this thesis, we compare different ways of representing items in the Twitter search system.

Given a set of original search results, the re-ranker will process those items in order to generate their representations. In these representations, the given item is represented via content features or Twitter specific features. Features such as the concepts and URLs shared in tweets are used to describe given search results.

The pre-processing phase of items is named as resource profiling. It consists of two steps: the first step is to understand the meaning of items by mining their textual contents. Meanwhile, similar vectors are created for each result using concepts from their contents. Subsequently, the weighting function is identical to the one used in the user modeling phase. The final representation of each item is a vector space model as well.

Personalization Algorithm

To realize commonalities between search results and the user's interests, we create this component in our framework in order to measure the similarity between items and a specific user profile. Given tweet representations from the top-n items, items are re-ranked based on content similarity scored by our personalized algorithm. The re-ranker then outputs this new sequence of items as personalized results.

Figure 3 shows the framework of the personalized re-ranker. Each component has different design alternatives. In terms of user modeling, we can construct different types of profiles. Likewise, a given item can be represented by concepts with different structures. In our thesis work, we explore several different methods for the three design dimensions explained before. Detailed descriptions of these design dimensions are presented in the sections that follow below.

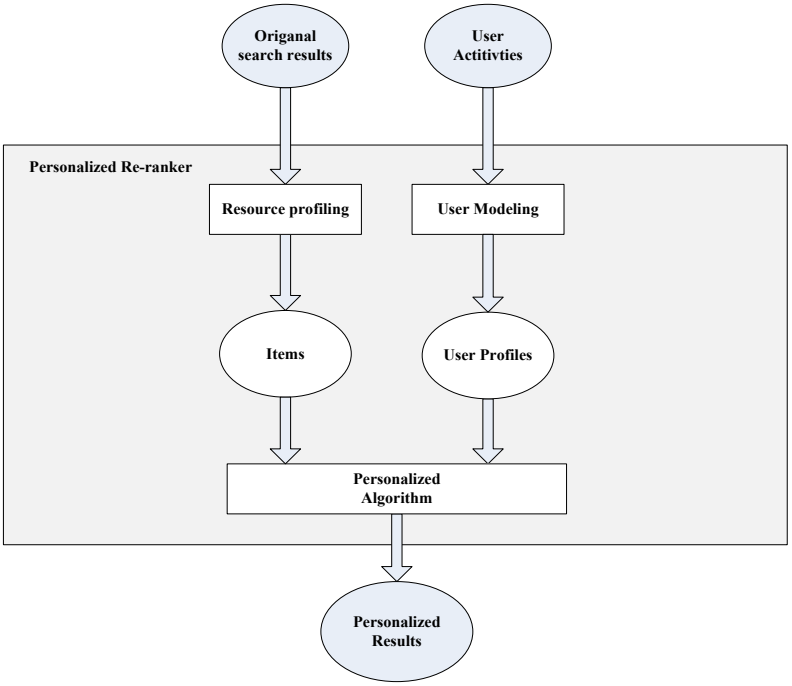

Figure 3 The framework of personalized re-ranker

3.5 Design Dimension

3.5.1 User Modeling

Personalized web search takes the query from the user as a natural but limited expression of their information need. Thus, user preferences are considered as additional sources for deriving their information need. We utilize users' interacting activities on Twitter to model their search preferences. Our assumption is that the user is likely to be interested in the search topic related to the concepts mentioned in her tweets. We therefore employ Twitter user modeling techniques to model users' search preferences.

After the profiling data have been collected, the user profile is generated in three main steps: (1) Concepts Extraction (2) Concepts Classification (3) Concepts Weighting. We will explain all these steps in the rest of this section.

In terms of the user profile definition, our model is based on a generic user-profiling model on Twitter, which is proposed by Abel et al. [15]. The definition of our user profile is given below.

Definition 1

Generic Model of User Profile (without classification): The user profile of a given user $u \in \mathbf{U}$ is a set of weighted concepts represented by $P(u)$. The weight of a specific concept $c \in \mathbf{C}$ is calculated by a certain weighting function ω. Here, $\mathbf{U}$ and $\mathbf{C}$ denote the set of users and concepts respectively.

$$P(u) = \{(c, \omega(u, c)) | c \in C\}$$

Definition 2

Generic Model of User Profile (with classification): The user profile of a given user u ϵ **U** is a set of weighted concepts. The weight of a concept cϵ **C** is calculated by a certain weighting function ω. Here, **U** and **C** denote the set of users and concepts respectively. For a given user u, $L_u(c)$ denote the set of labels in the category of concept c in this user profile.

$$P(u) = \{(c, \omega(u, c), L_u(c)) | c \in C\}$$

Concepts Extraction

This step is designed in order to identify concepts in microblog posts. Subsequently, the concepts from the given set of posts are organized to represent the user's preference. In many microblog search scenarios, identifying concepts, such as products, brands, or persons [49], plays an important role in understanding what people are expressing. In other words, the concepts in microblog post are able to determine what these posts are about. Thus, this step aims to link user preference to a set of concepts extracted from Twitter posts.

To extract meaningful concepts from individual microblog posts, items can be modeled based on semantics they carried [11]. Consequently, we utilize Web services to extract entities such as people, organizations or events from the given raw content. Services provided by OpenCalais[3] and Textwise[4] are chosen to process Twitter posts in our thesis work. To better understand this step, an example is presented below.

Example

> Tweet: "*#GameofThrones creator George R.R. Martin talked about killing off key characters last night on #Conan: http://t.co/qR4zJ… #HBO*"

Semantic meanings we can get from this item are shown in the following table:

Table 1 Semantics in the tweet

Entity	Frequency
GameofThrones	1
George R.R. Martin	1
Conan	1
HBO	1

From the examples above, we can figure out that this tweet is about an interview with the author of "Game of Thrones". By reducing noisy concepts, the search system can further

[3] http://www.opencalais.com/
[4] http://www.textwise.com/

provide a semantic representation of given item. Subsequently, this item is likely to be relevant to users who interested in concepts such as "*GameofThrones*" or "*George R.R. Martin*". To summarize, this way of providing semantics to digital items aims to generate item representation with meaningful concepts. In addition, the generation of the semantic user representation is based on entities and their frequency.

Apart from building a semantic user representation, we still have an alternative method of organizing the raw user contents: Representing individual microblog posts in a term-based way (bag-of-words). For instance, Table 2 describes the bag-of-words format of the previous example.

Table 2 Terms in the tweet

Term	Frequency	Term	Frequency
GameofThrones	1	key	1
creator	1	character	1
George R.R. Martin	1	last	1
talk	1	night	1
about	1	on	1
kill	1	conan	1
off	1	HBO	1

The term-based representation is a straightforward way of identifying concepts. Each tweet is considered as a collection of words. Subsequently, we remove all the stop words in that collection, and take the rest words along with their frequency into consideration.

To summarize, this process determines what kind of concepts should be included in the user representation. These concepts should be able to describe the users' preferences accurately.

Concepts Classification

In addition to the concepts extraction, we further take the relationships among concepts into consideration. For instance, a user would like to get tweets related to "Game of Thrones". However, searching for tweets including the keyword "Game of Thrones" may be not enough to satisfy the user's particular information needs. In this case, if some characters names are of high interest to a user, but not often mentioned in her published tweets. In this case, the search system may not identify the most personally relevant result of the user. To summarize, this step is designed in order to provide results with group of topically related concepts.

In this step, we map concepts onto domain ontology to gain their semantic relationships and derive a further understanding of users' preferences. Domain ontology is the formal representation of a set of concepts within the search results, and the relationship between the concepts [50]. Therefore, the concepts classification phase aims to identify semantic re-

lationships among entities, and then organize concepts into a hierarchical tree representation based on their relationships. In my thesis, the classification task is applied solely on entity-based approaches, because the term-based tweet representation has meaningless concepts that apparently provide no related classification information.

We now describe the general model of hierarchical structure as well as some associated concepts. Given C as a collection of elements, H_C denotes the rooted tree representation of this given collection C. Here, H_C is actually an induced tree of this given collection that describes the hierarchical domain structure of elements or concepts in C.

Given a rooted tree H, we define all the sets of categories and concepts in H as L_H, and denote the set of concepts that are leaves of the tree as LL_H. For each leaf node of H, there is a related ancestral path from the root node. The path of a given leaf node, which is a subgraph of H, describes the category of a specific concept. Furthermore, the depth of a node in the hierarchy is the number of edges on the path from the root of the tree to that node. Given any two leaves l_1 and l_2 in H, LCA(l_1, l_2) denotes the *Lowest Common Ancestor* which is the ancestor node of both l_1 and l_2 with the greatest depth.

For instance, given a Collection A= $\{GameofThrones, George\ R.R.Martin\}$, the induced tree H_A is shown in Figure 4. The sets of categories and concepts in H_A is represented by $L_H = \{Arts, Sports, Literature, Television, a, b\}$, whereas the set of concepts is represented by $LL_H = \{a, b\}$. The *Lowest Common Ancestor* of elements "$GameofThrones$" and "$George\ R.R.Martin$" which is denoted by LCA(a, b)= {Arts}.

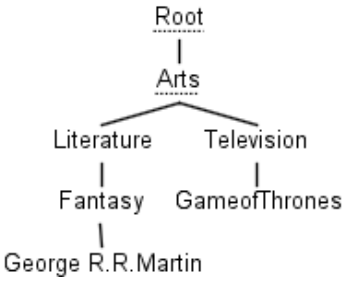

Figure 4 The induced tree for Collection A

Given a set of concepts extracted in the first step, we made use of the Textwise ODP classification service that can generate reliable results from the collection [51]. This service returns up to three possible ODP classifications of a given concept, ranked by the degree of confidence in the classification. The confidence is a score in the range of [0, 1]. For instance, the category path of an entity "NBA" is "Sports/Basketball/Professional" based on this service.

This service is based on the Web topic ontology named Open Directory Project (ODP) taxonomy[5], which is the largest and most widely used ontology [52]. In addition, it can provide

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

long category labels for concepts. Consequently, long category labels allow us to build the hierarchical tree as item representation. Other categorization service such as Open Calais can only provide short category labels (e.g. person, position and location) and are not the best alternative. Moreover, it is widely used as the basis for various research projects in the area of Web personalization [50-52]. Thus, these for are reasons for selecting Textwise classification service in our method.

However, another challenge of mapping concepts to ontology is that some entities may belong to multiple categories in the ontology. To disambiguate these categories, we employ a disambiguation algorithm [52] for the hierarchical tree representation. If a given concept belongs to multiple categories, the category with the largest probability will be regarded as the final category based on this algorithm. Given a rooted tree H, the probability of a concept that $c \in LL_H$ belongs to a category at level 1, is calculated as the global occurrence of the concepts under this category multiplying the confidence probability of the category.

To better understand this mechanism, we provide an example to show how we disambiguate multiple categories of concepts.

Example

Tweet: *"I am watching Game of Thrones, The Rains of Castamere (S03E09). http://t.co/I4kXIGj9pS #GameOfThrones"*

Given this tweet, Figure 5 shows the induced tree of related concepts. These concepts are extracted from the tweet text and associate external resource. As we can see from Figure 5, the concept "*jack gleeson*" has two categories.

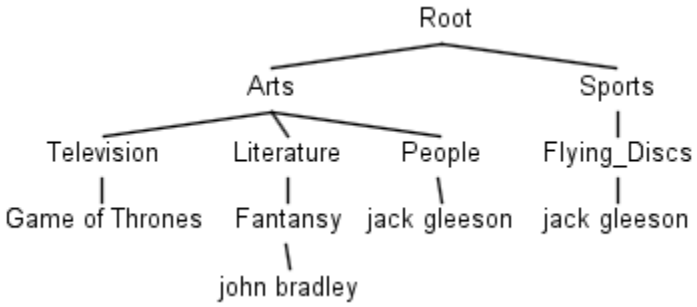

Figure 5 The induced tree for multiple category concepts

As we can see from the graph above, there are 3 leaf nodes under "*Arts",* and total number of leaf nodes is 4. The numbers of leaf nodes providing a current context related to "*Arts*". Given the confidence probability of path "*Arts –People*" is 0.18 returned by the Web service, the probability of concept c "*jack gleeson*" belongs to "*Arts*" is calculated as $P_{gc}(c \in \text{Arts}) = \frac{3}{4} \times 0.18$. Description of the algorithm [52] is shown in Figure 6.

For $c \in LL_H$ **do**

For $node \in child(root(Tree))$ **do**

$$P_{gc}(node) = \frac{number\_of\_categories\_bypass\_node}{total\_number\_of\_leaf\_nodes}$$

$$Node\_score = P_{gc}(node) \times P_{cp}(node)$$

End for

$Node^* = arg\, Max_{node}(Node\_score)$

$Tree = Subtree(Tree, Node^*)$

End for

Figure 6 Algorithm for concepts disambiguation

Concepts Weighting

Definition 3

Weighting Function. Concepts in both user representation and Twitter post representation are weighted based on their term frequency. For a given concept $c \in C$ of a specific user $u \in U$, its weight is assigned by the weighting function $\omega(u, c)$. Here, U and C denote the set of users and related concepts respectively. The equation is given as below.

$$\omega(u, c) = \frac{n}{N}$$

In the equation, lower case n represents the number of a specific concept occurrence, while N represents the number of all concepts occurrence. The implicit assumption of this weighting schema is that the more concepts of interests are mentioned by users' tweets, the more relevant these concepts are to this user.

After assigning the weight, the user profiles will be represented by vector space model with the purpose of calculating the similarity between the user profiles and candidate items.

3.5.2 Resource Profiling

To generate item representations for personalization, we need to preprocess original search results. This resource profiling process is performed through four main steps: (1) external contents extraction, (2) concepts extraction, (3) concepts classification and (4) concepts weighting. This process takes original items as input data, and outputs a set of item representations.

Previous studies have shown that some Twitter posts are short, noisy, or full of ungrammatical text [49, 53], thus making these types of posts provide a limited context for internal words, and consequently little meaningful concepts can be extracted from candidate items. Apart from generating user preferences representation based on a group of posts, constructing representation for items can rely solely on the content of the item itself. Therefore, it makes tackling the data sparsity problem a challenge.

To tackle this potential problem in Twitter posts, this process begins with a step named external contents extraction. To provide a semantically enriched representation, we apply a semantic enrichment approach proposed by Abel et al. [11], whose approach depends on matching tweets to external news articles via URLs in those tweets, followed by semantic enrichment based on external textual content. If a given item contains an external URL, we consider it a pointer to the external content, and we then utilize BoilerPipe [17], a library based on linguistic rules, to extract the main contents of the external resource. Finally, the textual contents from the external page are considered supplemental profile data, and outputted along with original item content for further concepts extraction tasks.

The ways concepts are extracted and classified applied in search items are identical to how we process user profile data. A detailed description of these two steps can be found in section 3.2.1. Weighting schema of concepts in Twitter post representation is identical to one previous one. For a given item $t \in T$, its weight is defined by $\omega(t, c)$, where T denotes the set of items.

3.5.3 Personalization Algorithm

The personalized algorithm takes the weighted user profiles and a set of candidate items as inputs. Given the required inputs, it then re-ranks the candidate items based on their similarity to the user profile. The assumption behind the algorithm is that, for each candidate item on the original list, the more similar it is to the user profile, the higher the probability it has to become a relevant item. Therefore, the final sequence of the new ranking is determined by the similarity between the user profile and items. Meanwhile, this ranking is regarded a personalized list, because user preference information is incorporated in the search process. Given a user profile $\overrightarrow{P}(u)$ and a set of candidate items T, our general personalized algorithm is as shown in Figure 7.

For $t_i \in T$ **do**
    Caculate $Similarity\left(\overrightarrow{P}(u), \overrightarrow{P}(t_i)\right)$
End for
 # highest to lowest stable sort
Sort *items* **BY Similarity**
Return *items*

Figure 7 The personalized algorithm

Similarity Metric

Various approaches can be employed to measure the similarity between two vectors. In this project we choose two metrics: Cosine-Similarity and Generalized Cosine-Similarity Measure. The detailed definitions of these metrics are given below.

Cosine-Similarity: Given a user profile $\overrightarrow{P}(u)$ and a set of candidate items T , both represented in the vector space model using the same vector representations. In the following equation, $\overrightarrow{P}(u)$ and $\overrightarrow{P}(t)$ ($t \in T$) are the vector representations for user profiles and a candidate item respectively, the similarity between two vector representations is defined as below:

$$sim_{cosine}\left(\overrightarrow{P}(u), \overrightarrow{P}(t)\right) = \frac{\overrightarrow{P}(u) \cdot \overrightarrow{P}(t)}{\left\|\overrightarrow{P}(u)\right\| \cdot \left\|\overrightarrow{P}(t)\right\|}$$

Generalized Cosine-Similarity Measure (GCSM): GCSM [54] is an expansion of the vector space model. It takes the domain relationships among concepts into consideration, and exploits a hierarchical domain structure in computing similarity. In this algorithm, both the user profile and items are represented in the expanded vector space model using the same vector representations along with their related hierarchical structures.

Suppose the user profile is denoted as $\overrightarrow{P}(u)$, and an item in the original result list T is denoted as $\overrightarrow{P}(t)$ ($t \in T$), GCSM defines $\overrightarrow{P}(u) \cdot \overrightarrow{P}(t)$ as:

$$\overrightarrow{P}(u) \cdot \overrightarrow{P}(t) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \cdot b_j \cdot \theta(P_i(u), P_j(t))$$

Where a_i and b_i are the term frequency for two concepts $P_i(u)$ and $P_j(t)$ respectively, the usage of weight here is identical to the standard vector space model. In the above equation, $\theta(P_i(u), P_j(t))$ is a similarity measure, which is developed to describe the similarity between two nodes in the induced hierarchical tree, for any two elements l_1 and l_2. The similarity is defined as below:

$$\theta(l_1, l_2) = \frac{2 \times depth(LCA_H(l_1, l_2))}{depth(l_1) + depth(l_2)}$$

Where H denotes the rooted hierarchical tree, l_1 and l_2 are two nodes of a tree, $LCA_U(l_1, l_2)$ is the lowest common ancestor of l_1 and l_2, $depth(l_1)$ and $depth(l_2)$ are the depth (from root) of these two nodes in the tree respectively.

Finally, the normalized GCSM similarity of these two vectors is given as:

$$sim_{GCSM}\left(\overrightarrow{P}(u), \overrightarrow{P}(t)\right) = \frac{\overrightarrow{P}(u) \cdot \overrightarrow{P}(t)}{\sqrt{\overrightarrow{P}(u) \cdot \overrightarrow{P}(u)} \cdot \sqrt{\overrightarrow{P}(t) \cdot \overrightarrow{P}(t)}}$$

This metric is proposed to facilitate the vector space model by adding a hierarchy, and the hierarchy aims to describe the relationships among domain elements or concepts. The semantic relationships in the hierarchy help us identify objects sharing common characteristics, leading to improved measures of similarity.

3.5.4 Summary

In this section, we summarize our design dimensions. There are three design dimensions of search personalization on Twitter: 1) User modeling, 2) Resource Profiling and 3) Similarity Measurement. The first dimension aims to extract users' preferences and generate associate user representation. While the second dimension determines how items are represented in the search system. Finally, the last dimension determines how we measure the relevance of items to the user. Furthermore, our personalized strategies are based on different design alternatives explained in this section. Table 3 describes the design space for the personalized strategies.

Table 3 Design Space

Design dimension	Design choice	
User modeling	Term-based	
	Entity-based	1) without classification 2) with classification
Resource profiling	Term-based	
	Entity-based	1) with URL enrichment 2) without URL enrichment
		1) with classification 2)without classification
Similarity Measurement	1) Cosine-Similarity 2) GCSM	

3.6 Personalized Strategies

In this section, we introduce our four personalized strategies: (1) Term-based (2) Entity-based (3) URL-based and (4) Hierarchy-based. For each personalized strategy, we explain how we represent users' preferences and items based on our previous design dimensions. In addition, we describe how to determine the preference relevance of a given result to the user.

3.6.1 Term-based Strategy (Bag-of-Words)

Many twitter search approaches rely on the bag-of-words (BOW) model [6, 53]. Each tweet is considered a collection of pre-processed (e.g. normalized, stemmed) terms with weight scores (e.g. TF-IDF) assigned. We adopt this model in our first strategy. A given user $u \in \mathbf{U}$ is represented by one's term-based profile (see definition 1). Concepts in this user representation are terms that are extracted from user's published tweets. Thus, the term-based profile of user u is a vector space model $\overrightarrow{P}(u) = (\beta_1, \beta_2 .. \beta_n)$, where β_i is the weight of a word i in $\overrightarrow{P}(u)$ and n donates the given number of concepts. Likewise, a given Twitter post $t \in T$ is

represented by a vector space model $\overrightarrow{P}(t) = (\alpha_1, \alpha_2 .. \alpha_n)$. Their similarity is measured by cosine similarity $sim_{cosine}(\overrightarrow{P}(t), \overrightarrow{P}(u))$.

The term–based strategy compares the tweet representation and the user representation based on their cosine similarity (see section 3.3.3). Personalized results are provided after the similarity calculation phase. This strategy is a straightforward way to model items and user preferences. In addition, this strategy treats all words in a post equally important after removing stop words.

However, characteristics of tweets pose challenge to this way of item representing. Twitter posts often has unusual spelling or emoticons [48], thus some tweets may be too noisy to provide concise information for similarity calculation, and some various pointless terms may flood those semantically concepts. Therefore, in the following strategies, we investigate how to provide a more concise representation by reducing the amount of noisy concepts.

3.6.2 Entity-based Strategy

Naveed et al. [6] mention that microblog messages contain few terms in general and very rarely contain a term more than once. They point out that this term sparsity will have an impact in a retrieval setting. Given the restriction of microblog messages to contain very few terms or meaningless concepts, we plan to explore how to construct item representation with meaningful concepts.

In this strategy, to provide concise representations for user preferences, we model Twitter posts by exploring their semantic meanings. User activities on social platforms are often triggered by specific topics and related entities (e.g. sports events, celebrations, crises, news articles, persons, locations) [48], and can be utilized by an entity-based approach. Thus, tweets are enriched with entity extraction services such as OpenCalais and TextWise (see Section 3.1.2) in our thesis work.

As has been explained, users' activities on the Twitter platform are considered an additional resource for the modeling of their preferences. For each user $u \in \mathbf{U}$, we construct the user's entity-based profile (see definition 1). Concepts in this type of profile are named entities identified from the user's published posts. Therefore, users' preferences are represented by a vector space model $\overrightarrow{P}(u)$. Similarly, a given Twitter post $t \in T$ is represented by a tor $\overrightarrow{P}(t) = (e_1, e_2 .. e_n)$, where e_i is the weight of a specific entity in $\overrightarrow{P}(t)$, and n denotes the given number of entities. The entity-based personalized strategy ranks original search results by calculating their cosine similarity $sim_{cosine}(\overrightarrow{P}(t), \overrightarrow{P}(u))$.

Apart from the term-based strategy, we filter out some of the meaningless or noisy concepts by employing semantic enrichment techniques, and then generate semantically representation for users and items. For a given ranking returned from the Twitter search engine, we compare the similarity of semantically representation of a given result to the user.

3.6.3 URL-based Strategy

In this strategy, we address the problem that short tweets provide insufficient basis for generating semantic item representation. The limited size of tweets poses a challenge to understanding items by identifying sufficient entities. In this strategy, to provide enriched representations for items, we try to understand tweets by linking them to external articles. If a post already contains URLs, we assume that contents from external source are related to the semantic meaning of this post. Consequently, we enrich the semantic of tweets by exploring external resources.

The URL-based strategy relates the Twitter post t with an external article a, entities identified from the article and the tweet are merged to form a concepts collection E. The post $t \in T$ is represented by a vector $\overrightarrow{P}(t) = (e_1, e_2 .. e_n)$, where e_i denotes the frequency of a specific entity in E and n is denotes the given number of distinct entities in E. User representation of URL-based strategy is identical to the entity-based strategy, both two strategies present users by their entity-based profile (see definition 1).

To better understand this strategy, we provide an example on how to enrich a tweet.

Example

Tweet: "*Game 3 preview: Heat vs. Spurs - ESPN http://t.co/0wiR2agced*"

From the examples above, we can only get few semantics such as "Heat", "Spurs" and "ESPN". However, if we relate this tweet to the news article published by ESPN, we can know that this tweet is about an upcoming basketball final in NBA. Furthermore, the tweet representation can be enriched by additional entities (e.g. Lebron James, Tim Duncan, and NBA) from the news article. If a fan of "Lebron James" is searching for some tweets related to "Miami Heat", then this tweet is likely to be relevant to his or her search preference.

Given a set of items which are considered relevant to the search query, this strategy compares the semantic user representation and the enriched tweet representation based on their cosine similarity. Finally, we present personally relevant items on top by re-ordering their sequence.

3.6.4 Hierarchy-based Strategy

Previous strategies identify the personally relevant results by strictly matching items to user profiles. In other words, the new relevance score is determined by the number of identical concepts in the intersection between two representations. However, concepts in the user profile may not fully reflect one's search preference. For instance, if a user searches a query "Game of Thrones", this user may also be interested in the characters in this TV series. If these types of character names are mentioned in some items but not included in one's content-based profile, the personalized search engine is likely to neglect these potential candidate items.

Consequently, we propose this hierarchy-based strategy to provide personalized results by exploring relations among concepts. Our assumption is: If a concept is highly topically related to those which are relevant to a user's preference, then this user is likely to be interested in this concept as well.

A Twitter posts $t \in T$ are represented by $P(t) = \{(c, \omega(t, c), L_t(c)) | c \in C\}$, where C is a set of concepts identified in t. The weight of a concept $c \in C$ is calculated by a certain weighting function ω. For a given tweet t, $L_t(c)$ denote the set of labels in the category of concept c in this item. Users are represented by entity-based user profile with classification $P(u)$ (see definition 2). We present how to mapping concepts in a given representation to a hierarchical ontology in Section 3.3.1.

The similarity measure of hierarchy-based strategy is the *Generalized Cosine-Similarity Measure (GCSM)* [54], which is used to calculate similarity between concepts in the hierarchical structure. The detailed description of this metric is presented in Section 3.3.1. In the hierarchy-based strategy, search context and the concepts relations are taken into consideration in order to calculate the semantic similarity between the user profile and candidate items.

3.7 Conclusion

In this chapter, we present the personalized re-ranking approach of microblog search. We now present the answers to the first and second research questions.

1. How can we gather a user's preferences information on Twitter?

To represent a user's search preferences, we utilize their usage information on Twitter. In other words, users' interactions on Twitter are considered as the source of preferences information. By exploring their activity data, user profiles are constructed as the user representation. We are inspired from previous studies [9, 11, 15] to develop our user modeling approaches. Two approaches are applied to construct Twitter user profiles. The first one is term-based and the second is entity-based.

2. How can we improve the retrieval effectiveness by providing tailored search results to Twitter users?

To provide tailored search results, we propose the personalization framework. We focus on three design dimensions of this framework. Our design dimensions are as follows.

- User modeling
- Resource profiling
- Similarity measurement

Each design dimension has several design alternatives. Based on these alternatives, we gave a set of personalized strategies. The term-based strategy is a straight forward method

based on the bag-of-words (BOW) model. The entity-based strategy aims to reduce noisy concepts in both user and item representations based on semantic enrichment techniques. In addition to the entity-based strategy, we propose the URL-based strategy to provide rich semantic representations for users and items. This strategy can thus identify relevant items with additional concepts. Moreover, we propose the hierarchy-based strategy based on the categories of concepts. According to this strategy, the relevance of an item to a user depends on the number of topically related concepts as well as their semantic relationships.

In the next chapter, we will evaluate the performance of our personalized strategies. In addition, we will also attempt to understand and explain how it is achieved.

4 Evaluation

In this chapter, we evaluate our personalization search approaches. Two experiments are designed to answer the research questions introduced in Section 1. The first experiment is an online user study, while the second one is an offline evaluation based on users' implicit feedback. In this section, we first detail our experimental setup and dataset descriptions. Subsequently, results of the evaluation including overall performance and individual analysis will be described in detail.

4.1 Experimental Setup

The evaluations of general information retrieval usually refer to user-independent evaluation tasks. The relevance assessments collected from experts can thus be used to evaluate any search approach independently. However, user-independent evaluation approaches are not feasible for personalized search, since personalized results can only be assessed by the users themselves. Current evaluation approaches for personalized search are often based on a user study, where users subjectively judge whether a specific result satisfies their personal needs. In this thesis, we will first conduct a user study to evaluate our approach. This online experiment will be presented in Section 4.1.1.

In addition to the user study, we investigated the alternative method in our offline evaluation. The offline experiment utilized users' feedback on results. For instance, some personalized search approaches on social tagging system utilize user-generated tags to indicate their objective preference on items [13, 14]. Apart from personalized approaches for social tagging system, Twitter users don't have such direct ways as they might have on social tagging system to indicate their personal preference. Alternatively, we utilized users' implicit interacting behaviors on Twitter such as retweet or reply. A detailed description of our offline experiment will be given in Section 4.1.2

In this thesis, we build our own search topic set along with related original results. This set of search results will be used during both the online and offline experiment. We collected results of 48 search topics returned by the Twitter search engine. Details of our dataset will be presented in Section 4.2. Further, we set the dimensions of user profiles and item vector models to 20000 in both two experiments.

4.1.1 The Online Experiment

To evaluate our personalization techniques with real users and potentially real Twitter search topics, we propose an evaluation framework based on Vallet et al. [16] to conduct a user study. In this framework, users need to perform relevance assessments of individual search results. We applied pooling on the top 10 re-ranked items for each strategy and evaluated the items in the pool for relevance.

To implement the framework, we developed a web application to collect relevance judgments from users. To assess the performance of our personalized approaches, we needed subjective relevance judgments. In addition to subjective relevance, we additionally ask users to judge the topic relevance of a given item. The purpose of this extension is to distinguish between technically relevant (i.e. about the topic, containing the query terms) and personally relevant (whether a given item is relevant to users' personal search preference). As we will see later, users assessment behaviors towards these two types of relevance are different. Although a given search result can be highly relevant to the search topic, it may not satisfy a specific information need of a given user.

In addition, we didn't specify evaluation topics for a given user, but users were asked to choose two queries based on their interests. Previous works have shown that users have more positive responses to those evaluation topics which are known by the user [51]. If we specify evaluation topics for users, they may not be familiar with some of them and cannot be expected to assess those results based on their preference.

The Experimental Procedure

The complete result generation and assessment process is presented in Figure 8.

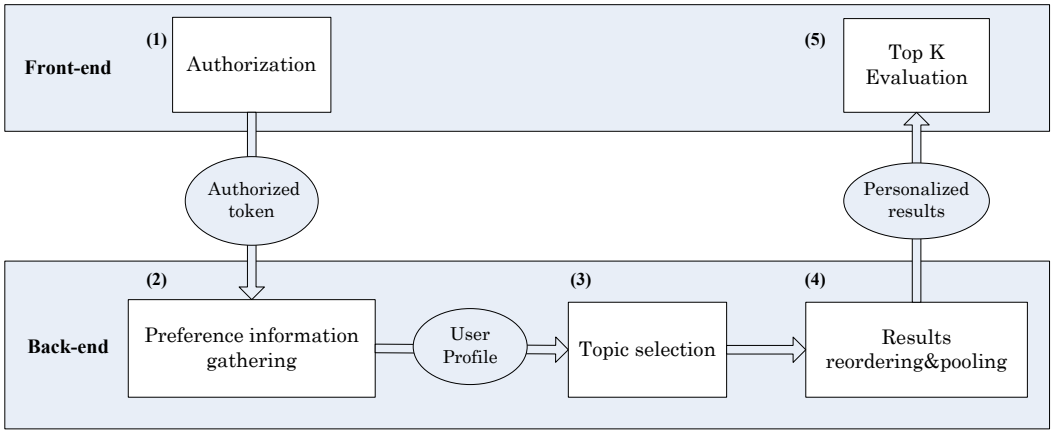

Figure 8 The process of result generation and evaluation

Each step of this process is explained as follows:

1) The user enters her/his Twitter account name and password. The OAuth standard is applied for authorization.

2) After receiving the authorized token, we collect the user's published tweets. Concepts from these published tweets are extracted in order to construct the user's profile. Details of our user modeling approach are presented in Section 3.3.

3) Next, search topics from different domains are presented to the user. He or she should select two of them as familiar topics. To avoid the potential data sparsity problem in the user profile, we provided some pre-defined concepts for users to select. Concepts of a specific

search topic were extracted from the associate search results dataset. Once the evaluation topics have been determined, the user should select several additional concepts as supplements of his or her user profile.

4) For each search topic, we use four different personalized strategies to reorder the original results. In this thesis work, we re-rank the Top-500 original search results based on the user's profile. The re-ranking process and related algorithms are described in Section 3.2. Finally, the top 10 results from each evaluated strategy are aggregated into a single pool of items.

5) The aggregated output of different strategies is presented as a search result list to the user in random order. The user should evaluate each result individually by answering two questions:

- Q1 (user): a 3-grade scale assessment on how relevant the result is to the evaluated topic.
- Q2 (topic): a 3-grade scale assessment on how relevant the result is to the user's interests.

Q1 provides feedback for measuring the accuracy of the evaluated approaches with respect to the overall search topic. Q2 is used to evaluate how relevant the result is to the overall search topic. Figure 9 shows the interface of this evaluation step.

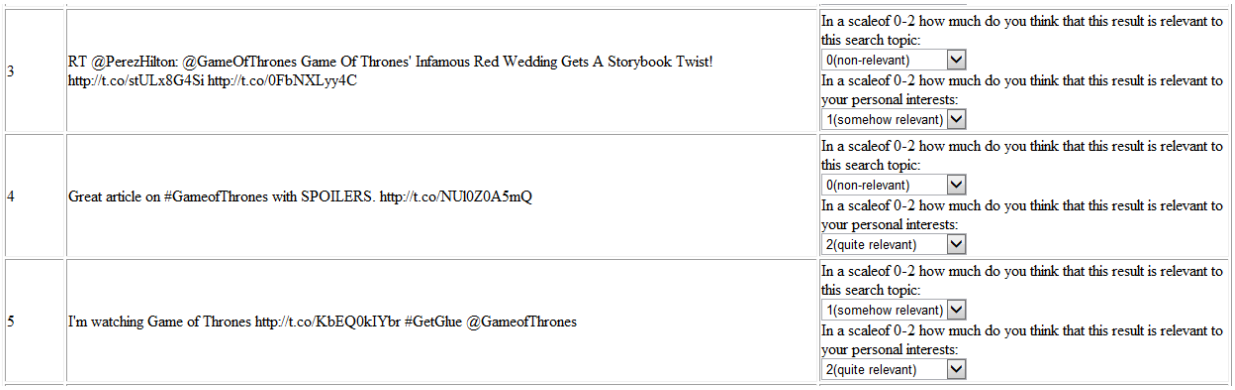

Figure 9 A snapshot of the user questionnaire

<u>The Participants</u>

The assessment collection process spanned a period of six weeks from August 15th, 2013 to September 30th, 2013. During this period, we were able to collect information from fourteen students from TUD between the ages of 21 to 28, eight of which are male. They provided assessments on 18 search topics along with 1230 individual judgments.

4.1.2 The Offline Experiment

Generally speaking, methods of evaluating personalized search approaches often include user study or user interviews [5]. However, conducting a user study with a large number of participants is oftentimes expensive and time consuming [6]. To test our personalized approach with a relatively large user set, we employ an offline evaluation approach in lieu of a user study. In addition, we construct a user sample for our offline experiment.

To simulate users' search behavior, we use the tweets dataset described in Section 4.1 as candidate sets. These items are regarded as original search results returned from the search engine during a certain period of time. To determine which item is relevant to a given user's preference, we utilize his or her interacting behaviors on Twitter. Thus, if a given item i related to search topic t is retweeted or replied to by a specific user u, we consider this item t as ground truth of the user's search preference under the search topic t. The relationship of users and search topics are established based on the number of items which are both relevant to user preference and belong to a specific search topic. Apart from the online evaluation, we didn't distinguish the topic relevance from preference relevance in here, due to the fact that we have no access to offline users' subjective preference information.

Given the ground truth, we select users who meet our criteria from the candidate set. We then exploit tweets published before the ground truth to construct their Twitter profile. Given a pair of user and search topic, the four personalized strategies are applied to generate new personalized rankings based on users' profiles. The performance of relevance judgment was measured by metric described in Section 4.3.

The User Sample

The user sample for offline experiment contains 142 users. These users are selected from the dataset of candidate items, and based on two criteria: 1) they have personally relevant tweets on a specific search topic; 2) they had published sufficient tweets before July 6th, 2013. In terms of this sample, the numbers of profiling tweets and personally relevant tweets are listed in Table 4.

Table 4 The numbers of profiling tweets and relevant items

	Profiling tweets	Relevant items
The average number	185.4	4.1
The maximum number	200	10
The minimum number	37	2

Timestamps of candidate items ranged from June 6th, 2013 to July 30th, 2013. Thus, we assume that these users conducted search behaviors during this period of time. Meanwhile, we used their published tweets in advance to their search behaviors (before June 6th, 2013) to construct their user profiles. Finally, the performances of four personalized strategies (see Section 3.4) are evaluated based on the existing ground truth.

4.2 Dataset Descriptions

This subsection introduces how we select the search topics and the associate original search results. To investigate the effectiveness of our personalized search system, we create an evaluation collection. We first determine a set of search topics to simulate the search behaviors of Twitter users. In Section 4.2.1, we will explain how we select these search topics and present details of these. Subsequently, we collected related results returned from the Twitter search engine for each topic. This set of results is considered to contain candidate items for the future evaluation tasks. Details of candidate items in this collection will be presented in Section 4.2.2. To empirically analyze and compare the impact of different personalized strategies on retrieval performance, this collection had been used as the original results dataset for both the online and offline experiment in this thesis work.

4.2.1 Evaluation Topics

Twitter search queries in this collection were manually selected from Twitter trending topics. Trending topics are the various popular and most often mentioned phrases, words, and hashtags on Twitter [17], and are shown on the left sidebar of users' homepage by default. Furthermore, trending topics could also affect users' search behaviors on Twitter.

We now explain the reason for utilizing trending topics to simulate search queries. Teevan et al. [2] illustrate that Twitter queries were significantly more likely to be celebrity names or event names. In addition, Zhao et al. [8] show that Twitter covers more entity-oriented topics on celebrities and brands which may not be covered in traditional media [8]. Since trending topics often related to popular concepts on Twitter, we consider these to be an alternative to simulate popular Twitter search queries. To summarize, we believe that our queries can simulate the searches that the participants conducted in the real world.

Table 5 Query terms

Query Type	Query Terms
Arts	Game of Throne, The Hangover 3, George R.R. Martin, Harry Potter, Dom Brown, Man of Steel, The Great Gatsby, The Incredible, Toy Story 3, Iron Man 3, J.K. Rowling
Sports	2020 Olympics, Robben, C. Ronaldo, Lebran James, Miami Heat, Ray Allen, Rooney, Tiger Woods, Tim Duncan, Tony Parker, Tracy McGrady
Products	Bitcoin, HTC One, Mac Pro, Nexus 4,Xbox One, IOS 7, IPhone 5
Politics	Ben Bernake, President Obama, Paul Krugman, Milton Freedman
Entertainment	Jennifer Lopez, Beyoncé, Adele, American Idol, Jeremy Kyle, Jon Stewart, Justin Bibber, Taylor Swift

Search topics from different domains constitute the query set. For instance, we collected results related to several famous sports players, new products, movie stars and politicians.

All of the query terms are shown in Table 5. In our thesis work, most of the search topics were selected from May to June 2013. The associate search results were crawled after a certain period of time. We didn't immediately collect results as long as a specific concept becomes a trending topic. The reason is that several popular tweets may flood the dataset during the early phase.

We hypothesize that users with information needs on certain topics can benefit from search personalization. For instance, if a football fan conduct a search on topic "2020 Olympics", then the results about some football stars should be presented on the top of the list. In addition, if a Japanese user searches for this topic, news related to "Tokyo becomes the host city of 2020 Olympics" might be more relevant to this user. Further, we guess some search on movie topics may be beneficial from personalization. For instance, if a user issues a query "Game of Thrones", the current search engine will return a set of results based on time or popularity. However, if a character "John Snow" is of more interest to the given user, then results about "John Snow" could be more relevant to his or her preference.

4.2.2 The Dataset of Candidate Items

In this subsection, we will introduce our candidate items dataset. As has been mentioned, each search topic has a group of original items which are topically related to it. Thus, we collected results on Twitter for the selected queries using Twitter's Search API[6]. The set of candidate items was crawled from June to July 2013. We sample 500 tweets for each query from its original tweets collection. Given a search query, the corresponding 500 tweets are considered baseline search results

The candidate dataset contains 48 queries and 24000 related tweets. The OpenCalais and TextWise Web Services have been used to preprocess candidate items. Details of these preprocessing tasks can be found in Section 3.3. The period of this data collection is from June 6th, 2013 to July 30th, 2013. The statistics of the dataset is detailed in Table 6.

Table 6 Dataset Statistics

Total tweets	24000
HasURL	25.1%
Hasconcept	75.6%
Average number of concepts	1.153

In addition, we give our further analysis of candidate items. As mentioned in Section 3.3.2, item representation will be enriched via textual contents from related external articles. Figure 10 compares the total frequency of entities in two types of item representation grouped by different search topics. Generally, it can be seen that the number of entities in item representations with URL enrichment per each search topic is higher than the number of those without URL enrichment. In terms of search personalization, users' specific infor-

[6] https://dev.twitter.com/docs/api/1.1/get/search

mation need should be satisfied via information or concepts in items. If user profiles cannot fully reflect one's whole information needs, a non-informative tweet may also have less possibility to be relevant to one's preference due to the lack of semantics.

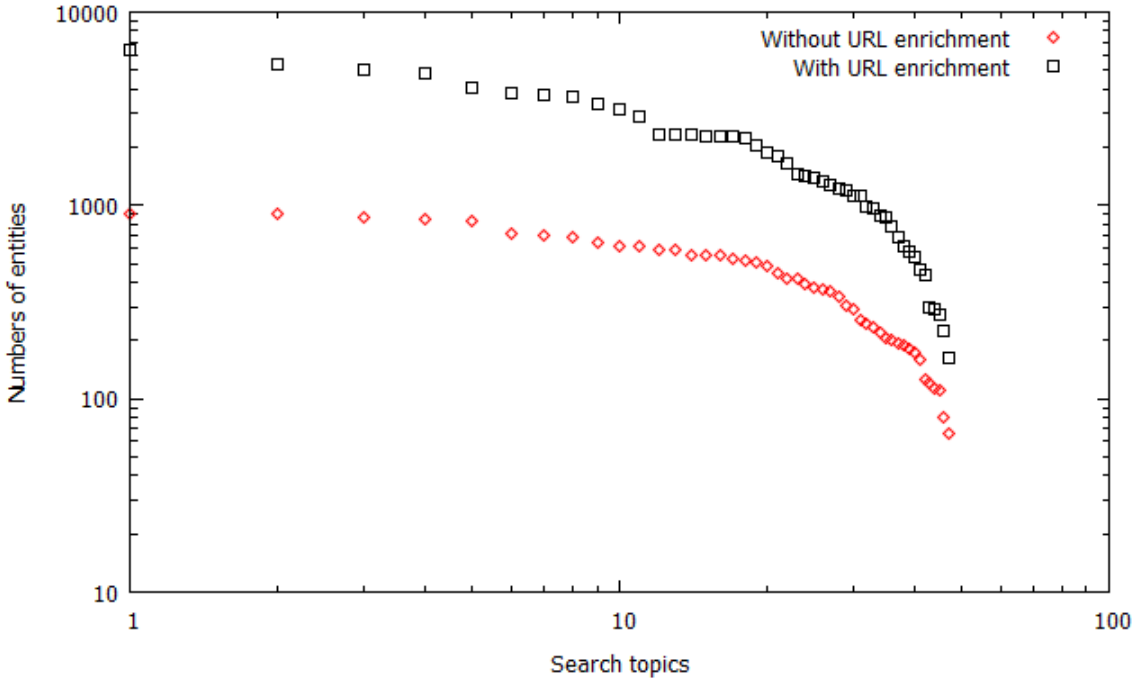

Figure 10 The number of entities per search topic

4.3 Evaluation Measures

To measure the ranking quality and the performance of personalization, we use the following metrics to evaluate re-ordering search results in online evaluation. However, we employ the following metrics except the nDCG@10 in offline evaluation, since we are not able to collect 3-grade scale assessment from offline users.

S@k

Success at rank k is the ratio of times where at least one relevant item in the first k was returned. Success at rank k is regarded the probability of finding a good descriptive item among the top k recommendation items. In this thesis, we use the success at 5(S@5) and success at 10(S@10) to measure our performance of recommendation systems.

MRR

In addition to S@k, we also select Mean reciprocal rank (MRR) to quantify the ability of ranking relevant items. MRR measures at which rank the first item relevant to the user occurs on average. This measure provides the ability of recommendation systems to provide the relevant item at the top of the ranking. If the first correct recommendation result is

ranked as the 3^{rd}, then the reciprocal rank (RR) is 1/3. Mean reciprocal rank is defined as the average reciprocal rank of results for a sample of queries Q. The equation is given as below.

nDCG@n

To measure the effectiveness of our personalized search approach, we use the Normalized Discounted Cumulative Gain (nDCG) [55]. DCG is a measure that gives more weight to highly ranked documents and allows us to incorporate different relevance levels (*quite relevant*, *somehow relevant*, and *non-relevant*) by assigning these different gain values. The equation is shown as below.

$$DCG(i) = \begin{cases} G(1), & if\ i = 1 \\ DCG(i-1) + \frac{G(i)}{\log(i)}, & otherwise \end{cases}$$

In the online experiment, we used G(i) = 1 for somehow relevant results, and G(i)=2 for quite relevant results, reflecting their relative importance. Because queries associated with higher numbers of relevant items will have a higher DCG, the DCG was normalized to a value range from 0 (the worst possible DCG given the ratings) to 1 (the best possible DCG given the ratings). In the online evaluation, we use the nDCG@n measure as an evaluation metric, where scores of questions QR1 (To what extent the given item is relevant to a user's preference) are mapped to relevance weights assigned to each candidate item.

4.4 Results

4.4.1 User Study Analysis

To investigate how users' information needs can be satisfied, we take a deeper look at their behaviors during the relevance judgment. In our user study, we distinguish the topic relevance from the preference relevance. We consider that merely relying on the topic relevance may lead to a neglect of users' implicit information needs. Consequently, the search approach may not be aware of users' implicit search intents and result in a poor retrieval performance. For instance, a top ranked result can be technically relevant to a given search topic, but it may be of less interest to a specific user. In this case, the actual retrieval performance decreased. To summarize, we believe that search approaches should be able to incorporate both two types of relevance in order to improve retrieval effectiveness. Thus, we use results of the user study to test our thoughts.

Relevance Judgment Distribution

Table 7 shows the distribution of relevance assessment of the user study. With regard to the preference relevance, of the 1230 online relevance judgments collected, 28.53% were Quite Relevant, 33.82% were Somehow Relevant and 36.74% were Non-Relevant. Although only 13.74% of the total results were non-relevant to associated search topics, 36.74% of the

total results were non-relevant to users' preferences. This group of results indicates that users' objective assessments towards these two types of relevance are not identical. Thus, we believe that users' judgment on topic relevance may not fully reflect the satisfaction of their information needs. It is necessary to consider preference relevance in the search system rather than merely considering the topic relevance.

Table 7 Relevance Judgment Distribution

Relevance	Non Relevant	Relevant	Highly Relevant
Preference relevance	463(36.74%)	416(33.82%)	351(28.53%)
Topic relevance	169(13.74%)	663(53.9%)	398(32.35%)

We now give some examples related to the topic and preference relevance. The first example is both technically relevant and peronally relevant to users who selected the topic "Game of Thrones".

Example

Tweet: "*Great article on #GameofThrones with SPOILERS. http://t.co/NUl0Z0A5mQ*"

This tweet conctains a external article related to the search topic. This article has mentioned some concepts (e.g. Eddard Stark, Red Wedding) which are relevant to the given user's preference. In addition, this result is of high interest to users who interested in "SPOILERS".

The following example is topically relevant but not personally relevant to some users.

Example

Tweet: "*@GoogleFacts: Game of Thrones has a 6 million dollar budget. Per episode. @aadityadamani*"

This tweet provides information in financial aspects. However, it can hardly attract users with particular information needs on characters or spoilers.

In addition, we also observed some results that are personally relevant but not topically relevant. To better understand this type of results, we give the following example.

Example

Tweet: "*RT @therealcabbie: Ray Lewis. Barack Obama. Game of Thrones. http://t.co/dFNYXdjIIr*"

This tweet is actually a joke upon some celebrities. It didn't convey any information about the TV series. However, it was labelled as personally relevant by some users. These users may understand this joke and consider this tweet as an interesting item.

In addition, we believe that introducing such a cascaded grading can be used for checking the efficiency of our strategies in a finer granularity. For instance, if a less relevant item is placed higher in the ranking, users still need to spend time on seeking the most relevant result. Thus, we believe that the percentage of highly relevant results in the ranking can reflect the effectiveness of our personalized strategies. Subsequently, we present the the distribution of preference relevance assessment in Table 8.

As shown in Table 8, results are organized based on different strategies. For each pair of user and search topic, a strategy will provide top 10 results for evaluation. Given 28 user-topic pairs, each strategy provided 280 items in total for users to assess. Table 8 indicates what percentage of items are labeled as non-relevant, somehow relevant and quite relevant to the users' preferences. We can see that the percentage of highly relevant results derived by the URL-based strategy is higher than the rest strategies. Meanwhile, the baseline group has the highest percentage of non-relevant results. In addition, we also notice that the term-based and the entity-based strategy have similar trend with regard to the relevance judgment distribution.

Table 8 Preference Relevance Judgment Distributions

Strategy	Non Relevant	Relevant	Highly Relevant
Baseline	158(56.4%)	74(26.5%)	48(17.1%)
Term-based	97(34.6%)	103(36.78%)	80 (28.57%)
Entity-based	91(32.5%)	119(42.5%)	70(25%)
URL-based	58(20.7%)	90(32.1%)	132(47.1%)
Hierarchy-based	93(33.2%)	93(33.2%)	94(33.5%)

Subjective Assessment Analysis

To further understand users' subjective preference, we made a comparison among items with different relevance level based on their features. Many researchers have investigated features such as length of a tweet, presence or absence of a URL or a hashtag and the number of entities mentioned in a message [27, 28]. Their works have shown that these features have influence on the relevance assessment or interestingness measurement that associates with users' subjective preference. Naveed et al. [23] indicate that incorporating features during the search process can improve retrieval performance in the sense of providing more relevant and generally interesting messages in the search results.

As has been mentioned, a given item can be labeled non-relevant, somehow relevant and quite relevant to the user's preference. Here, we focus on their preference relevance. Thus, items with the same relevance level constitute the related item set. Subsequently, we analyze each set of tweets based on the initial two features: 1) the presence of a URL of the tweet and 2) the length of the tweet. Table 9 shows the result of this analysis.

From the analysis of these two features, we could see that 69.2% of the quite relevant tweets have external links. The percentage of tweets which include URLs in the quite rele-

vant group is higher than that of the other groups. Thus, we thought the reason for this might be that contents from external links can provide more information or concepts. These additional concepts may increase the probability to satisfy users' specific information need of a search result.

Table 9 The comparison of features among different relevance levels

Relevance level	HasURL	Length (in characters)
Non-relevant	30.6%	98.45
Somehow relevant	55.5%	106.59
Quite relevant	69.2%	106.88

To validate our thoughts, we further analyze items that have an external link in these three groups with the third feature: the number of entities mentioned by a tweet. Of all the tweets that have URL in each group, we calculated the average number of entities mentioned in the external page referred by the item.

Table 10 Average numbers of entities in external page

Relevance level	Average number of entities
Non-relevant	9.16
Somehow relevant	16.19
Quite relevant	21.61

Table 10 shows the results for the three groups. This group of results reveals that quite relevant items tend to have more concepts than items in the other groups. Therefore, we think there may be a positive correlation between the number of concepts and the ability to satisfy a user's personal information needs of a search result.

4.4.2 Results of the Online Experiment

Overall Performance

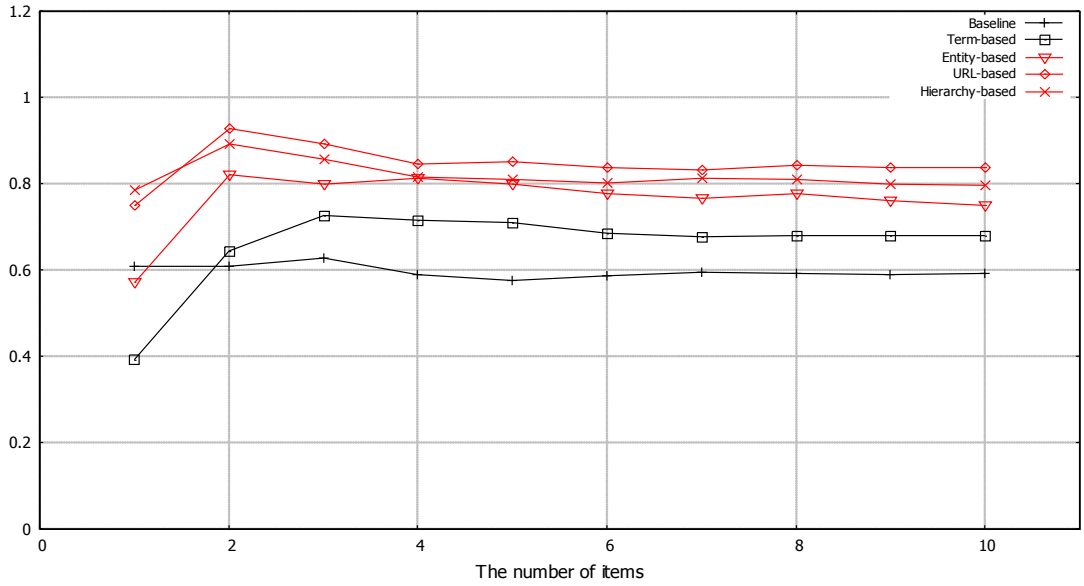

Figure 11 Results of different strategies measured by nDCG@n in online experiment

Figure 11 illustrates results of the ranking effectiveness measured by nDCG@n. All personalized strategies outperform the baseline with regards to present most personally relevant tweets on top. In addition, URL-based and Hierarchy-based strategies have better performance than the other in terms of subjective judgment. These two strategies utilize the URL enriched item representations that include more semantics than the others.

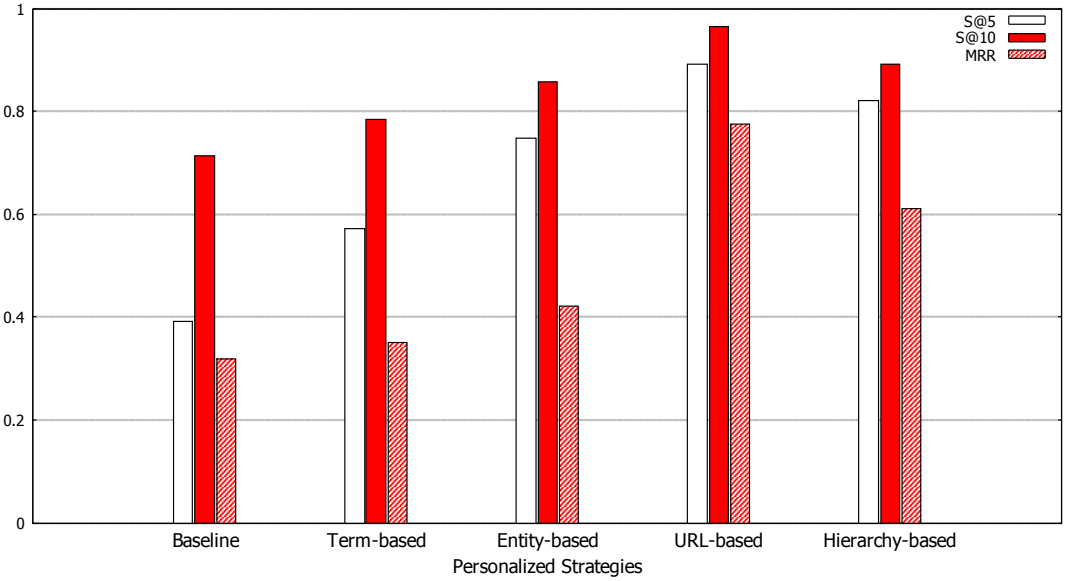

Figure 12 Results of different strategies measured by S@k and MRR in online experiment

Figure 12 shows the results of ranking performance measured by S@k and MRR. Here we observe the same trend as above. The URL-based strategy outperforms others in all three measures. Whereas the hierarch-based strategy has a slight increase compared with the entity-base strategy in terms of S@k.

An explanation of the relative low performance of term-based and entity-based strategies is: The top ranking results derived by these two strategies contain only the query term or concepts from user profiles in a repetitive way. In addition, semantic item representations without external enrichment tend to have a smaller number of entities as we presented in Section 4.2.2. Although these results may have higher relevant scores, they are not likely to convey much information that is relevant to users' preferences. Thus, users awarded lower points to these types of top ranking results. A further analysis of subjective user assessment will be presented in the following subsections.

We also observed that the URL-based strategy has better performance than the hierarchy-based strategy. The hierarchy-based strategy assumes that users are likely to be interested in concepts that are highly topically related to the given personally relevant concepts. This strategy may put some personally relevant but less topically relevant tweets at the top of the result lists. However, some of the participants considered these types of concepts to be non-relevant information both to the topic and individual preference, whereas some other users were willing to give higher grade to some less topically relevant tweets derived by this strategy. Thus, this strategy may have both positive and negative effects on improve ranking quality in subjective assessment.

Analysis of Individual Performance

Although some users can benefit from search personalization, we notice that the retrieval performance decreased on some users. To further explain the performance of our personalized strategies, we analyze results of different strategies at an individual level by nDCG@10. We find that the performance of our approach may be good at certain topics, while being bad at others. Given 28 pairs of user and topic, the term-based and the entity-based strategies outperformed the baseline on 19 (around 67.8%) pairs of user-topic. The URL-based and the hierarchy-based strategy outperformed the baseline on 22 (78.5%) user-topic pairs. This indicates that for some users and topics, it is beneficial to personalize their search results. However, we also see that the retrieval effectiveness can decrease after personalization.

To analyze the positive and negative effects of our personalized approach, we compare the performance of the URL-based strategy to baseline in the individual level. Figure 13 presents the results of nDCG@10 for each user-topic pair. To explain why the personalized approach can increase or decrease the retrieval performance, we further look at the users and topics with the largest improvement and deterioration in the following.

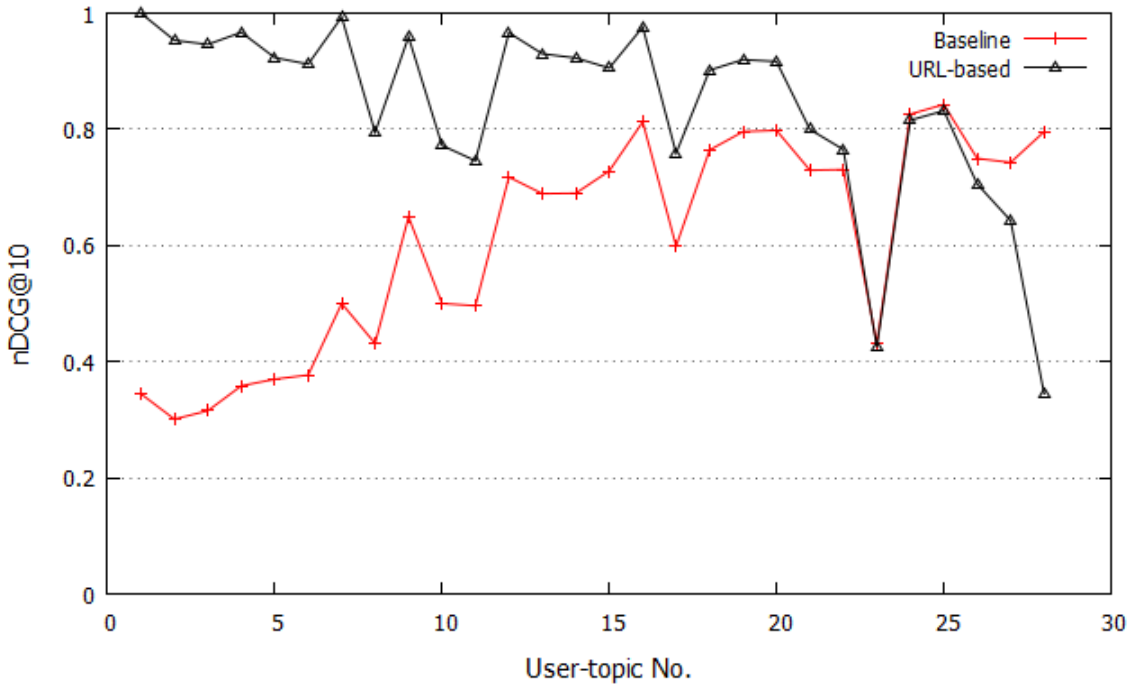

Figure 13 The difference between URL-based strategy and baseline in nDCG@10

We first present the top 5 users-topic pairs that have largest improvement based on the URL strategy. It can be observed that most of search topics are movie titles from Table 11. This type of search topics tends to have many sub-topics. For instance, there are various characters as well as actors or actress related to the movie. Meanwhile, users may have their own preferences on those characters. Thus, a user with a particular preference on some sub-topics may benefit a lot from search personalization.

Table 11 Top 5 pairs of user and topic with improvement by the URL-based strategy

No.	Search Topic	User No.
1	Harry Potter	13
2	Robben	3
3	Game of Thrones	1
4	The Great Gatsby	12
5	Game of Thrones	14

In addition, we notice that the URL-based strategy is more likely to improve the retrieval effectiveness than the rest of our strategies. As it has been mentioned, the URL-based strategy tends to provide rich semantic representations of items. Thus, we guess that results with additional concepts may increase the probability to satisfy users' specific information need.

Table 12 shows 6 pairs of user and topic that have least improvement by the URL-based strategy. Most of search topics are celebrity names in this case. We notice that results re-

lated to these topics are restricted to a specific domain. For instance, most of tweets related to "Milton Freedman" merely discuss some social science contents. In this case, users may be interested in the same contents related to this certain topic. Subsequently, this type of topic (e.g. celebrities) may not benefit from search personalization.

Table 12 Pairs of user and topic with deterioration by the URL-based strategy

No.	Search Topic	User No.
23	Milton Freedman	1
24	Harry potter	7
25	Taylor Swift	6
26	Game of thrones	5
27	Lebron James	4
28	Jon Stewart	10

The Impact of User Profile

Furthermore, we notice that the quality of user profiles may have impact on the personalization. For instance, a user may tweet something about movies but rarely talk about technical or political topics on Twitter. In this case, information from the user's posts may be good at personalizing results related to a specific movie but poor at adopting technical or political results.

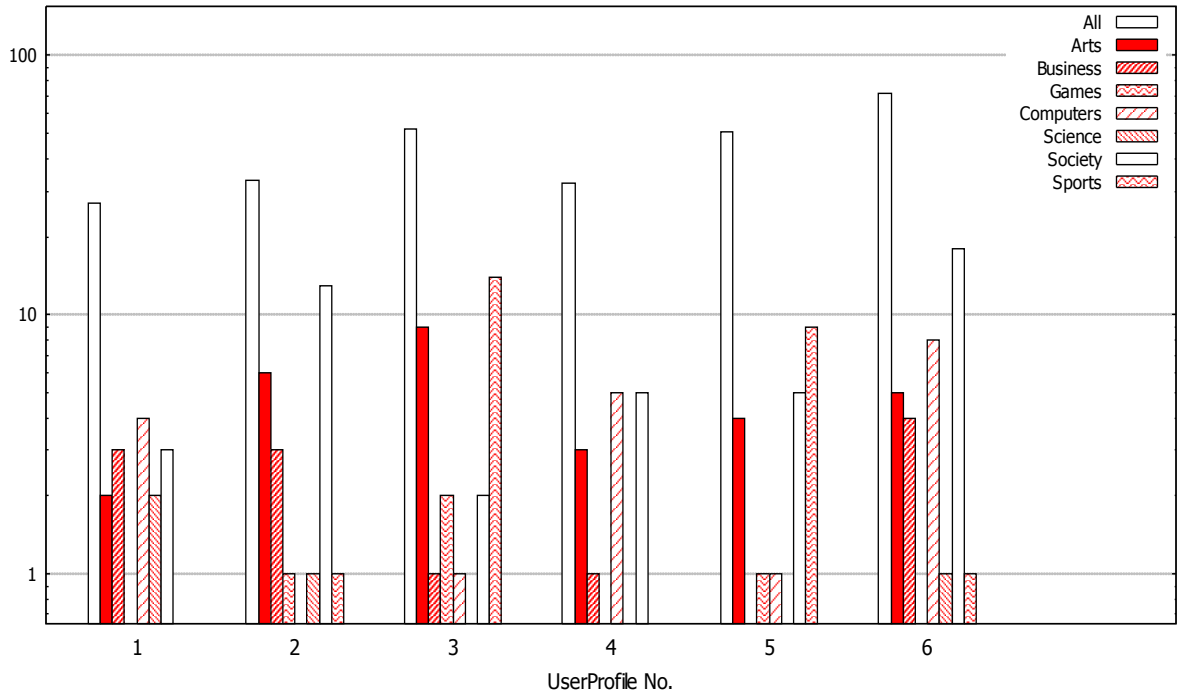

Figure 14 Numbers of concept in user profiles for users in Table 12

To validate our thoughts, we analyze the profiles of users mentioned in Table 12. The number and category of concepts are taken into consideration. Figure 14 presents the numbers of concepts classified by different domains in their profiles. As we can see from Figure 14, there are sparsity problem in their profile with regards to some specific domains. For instance, the first user profile only contains 3 concepts in business domain and 2 concepts in science domain. If this user conducts a search on an economist "Milton Freedman", then the profile of this user may not help the search approach to present the most personally relevant results on top.

The Impact of Item Representation

Apart from the user representation, we also find that the item representation may have influence on the retrieval performance. For instance, if a strategy tends to provide less informative tweets for users, then the retrieval effectiveness of this strategy may decrease.

We first notice that the hierarchy-based strategy may present results with some noisy concepts on the top. To better understand this reason, an example is presented as below.

Example

Tweet: *"@cperryy Ray Allen, Mike Miller, Chris Bosh, Dwyane Wade, Birdman, Mario Chalmers, Norris Cole, Shane Battier, isn't help?"*

This item mentioned several basketball stars. In addition, concepts in this tweet belong to a same label "Sports/Basketball", and result in a high weight score according to the hierarchy-based strategy. However, this given tweet just listed a set of names without expressing much meaningful information. Thus, it is of less interest for the user.

In addition, the shortage of concepts in item representation might also be a reason for the decrease in retrieval effectiveness. We first give an example to show how the entity-based strategy can provide non-personally relevant results.

Example

Tweet: *"taylor swift's cat http://t.co/DcGjoc1B5I #taylorswift"*

This item has a high weight score according to the entity-based strategy, since the concept "Taylor Swift" has appeared twice. However, this tweet is labeled as non-relevant by a male user. In fact, this given tweet doesn't convey any other information except the photo of a cat. Further, we guess that the concept "cat" is of less interest for the given male user.

In addition, we compare numbers of entities in the top ranked results derived by the entity-based and URL-based strategies. As has been mentioned, each strategy provide top 10 results for a given user–topic pair. Table 13 shows the average numwber of concepts in tweets

from the top 10 list. This table illustrates that top ranked items derived by the entity-based strategy have much less concepts after URL enrichment compared with others.

Table 13 Average numbers of entities in tweets from top 10 lists of each user-topic pair

Strategy	Without URL enrichment	With URL enrichment
Entity-based	3.76	7.775
URL-based	1.34	43.61
Hierarchy-based	1.29	23.86

4.4.3 Results of the Offline Experiment

Overall Performance

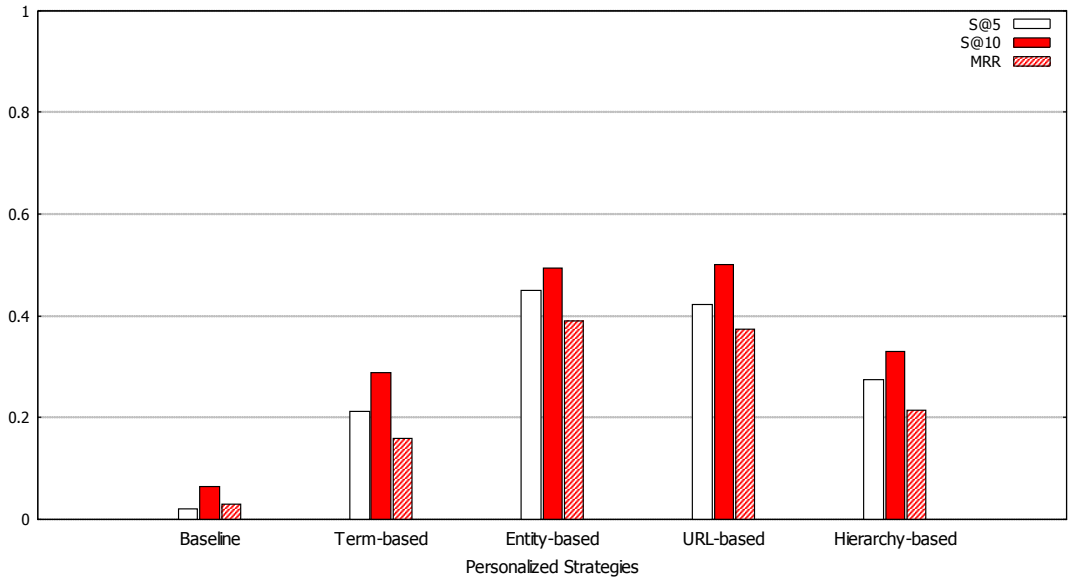

Figure 15 Results of different strategies measured by S@k and MRR in offline experiment

Figure 15 summarizes the results of ranking performance in the offline experiment. The entity-based and URL-based strategies that rely on semantic enrichment have better performances than the others. Compared to the term-based strategy, the S@5 improved from 0.211 to 0.275 by the hierarchy-based strategy, whereas the MRR improved from 0.16 to 0.214.

Apart from the online evaluation, we do not have abundant user assessments on personalized results. In online evaluation, each user can provide different levels of grades on an average of 43 items of a given search topic. However, the average number of personally relevant items is only 4.1 in this offline experiment. Thus, due to this problem, some differences among these strategies may not be fully reflected by the offline results. An item can have a higher personalized retrieval score than the given personally relevant item, but whether it is relevant to the user's preferences is not clear since we don't have further assessment behaviors from the user. Based on the existing results, we can conclude that all other strategies can improve the personalization performance compared with the term-based strategy.

Analysis of Individual Performance

We now provide the analysis of online experimental results. Since we were not able to collect nDCG@10 results in offline experiment, we analyze the MRR results of different strategies in individual level. Table 14 shows the percentage of users that have improvement by our personalized strategies. As we can see from Table 14, our personalized strategies can improve the retrieval effectiveness on some offline users, while decrease on others.

Table 14 The percentage of users with improvement in offline experiment

Strategy	Percentage of users with improvement
The term-based	71.8%(102)
The entity-based	77.4%(110)
The URL-based	79.5%(113)
The hierarchy-based	73.2%(104)

In the online evaluation, we further compare the individual performance of the URL-based strategy to baseline. Likewise, we conduct the same analysis on offline experimental results. In the offline evaluation, there are 17 users who have decrease in retrieval effectiveness based on the URL-based strategy. Figure 16 presents the number of entities in their profiles. As we can see from Figure 16, some users have data sparsity problem in their profiles. In addition, we observed that most of the associate search topics are related to technology (e.g. Windows 8.1, Mac pro) or person (e.g. Milton Freedman). Thus, we believe the quality of user profile and the type of search topic may have negative impact on the retrieval effectiveness.

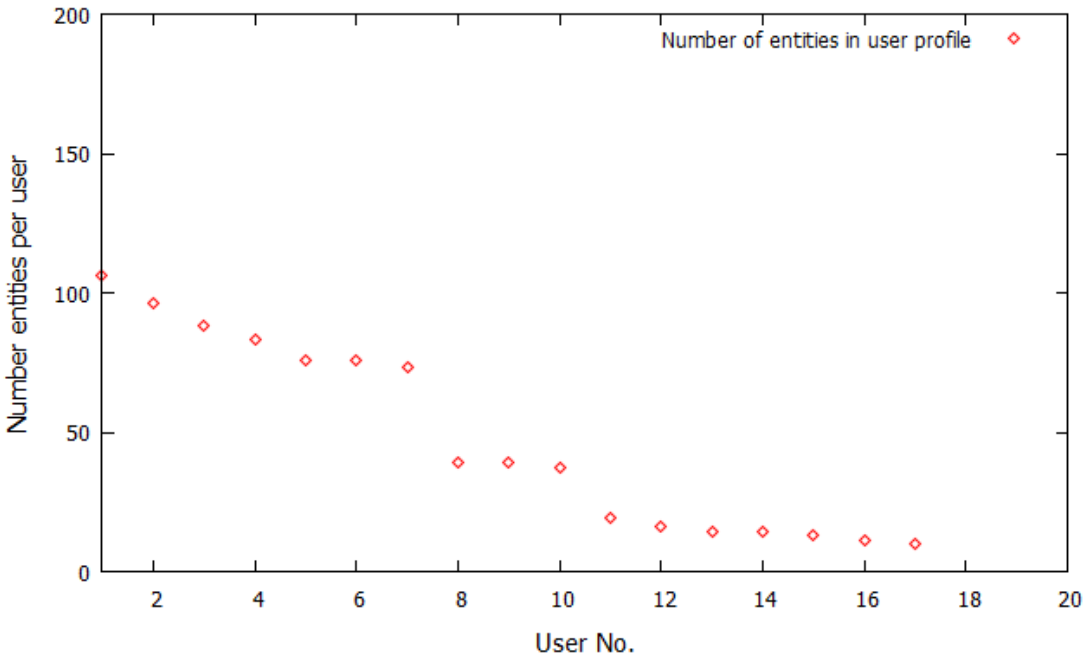

Figure 16 Numbers of concepts in user profiles for the 17 users

5 Conclusions and Future Work

5.1 Conclusion

In this thesis, we have investigated the feasibility of search personalization on Twitter by using content-based user profiles as user preference representation in our re-ranking approach. We looked into the particularities of search personalization on microblogs: modeling the users' preferences, and representing the original search results. Our thesis empirically showed that our personalized strategies can improve the retrieval performance. Answers to research questions proposed in Chapter 1 are presented as follows.

First, we use user modeling techniques to extract users' preferences on Twitter. Two approaches are applied to construct Twitter user profiles. The first one is term-based and the second is entity-based.

Second, we present a re-ranking approach to achieve personalization for Twitter search. Based on this approach, we decompose the personalization system design into three dimensions: user representation, item representation and personalized algorithm (see Section 3.2). To tackle the sparsity problem in item representation, we inferred from previous studies that semantic enrichment and hierarchy-based classification can be applied to generate informative item representation. We finally found that concepts extraction and enrichment in search results can improve the satisfaction of users' particular information needs. Based on our results (see Section 4), we could see different impacts of our strategies on the retrieval performance. Our thesis suggests that the re-ranking approach can improve the ability to satisfy particular information needs of Twitter users.

Finally, we present two evaluation methods for personalization of the Twitter search system. Our evaluation frameworks allow for the testing of different personalized approaches by including the subjective relevance assessments of Twitter users. Furthermore, we conduct a data analysis on results with different preference levels based on different features.

This thesis provides following contributions for researchers and developers:

First, we provide a set of personalized strategies for microblog search. Our personalized approach can be implemented by any microblog platforms in order to deploy a personalized social search engine.

Second, we present a framework of search personalization along with three design dimensions. Researchers can investigate how the retrieval effectiveness can be improved by other design alternatives. For instance, the performance of a new personalization algorithm can be studied based on this framework.

Last but not least, we present two evaluation approaches applicable to microblog search personalization. Our evaluation framework allows researchers to evaluate their new per-

sonalized approaches and compare different personalization strategies. In addition, we provide a method to collect subjective user assessments. Our online relevance assessment interface allows for a more realistic evaluation of different personalization approaches.

5.2 Future work

We have investigated the feasibility of Twitter search personalization by using user profiles. In this subsection, we present some future research directions of our work.

The strategies explored in our work represent only a small subset of the space of personalization. In terms of modeling user interest, we used term-based and entity-based user profiles. Several other viable types of profiles remain which can be tried in future. For instance, hashtag-based and topic-based could also prove to be additional design alternatives. Likewise, the representation of search results on Twitter can be further explored. As has been mentioned, characteristics of tweets result in a shortage of semantics in search results. Subsequently, items with few semantics may not satisfy users' particular information needs. Thus, different enrichment approaches can be applied to process search results. For instance, Meij [49] maps tweets to Wikipedia articles to facilitate concept mining on a semantic level. Benson et al. [56] try to match tweets to artist-venue pairs which can be obtained from sources such as music guides. Their studies have shown that many alternative methods can be applied to improve the concepts mining on tweets.

In terms of the re-ranking approach, it presents the potential problem of needing good quality results to re-rank [12]. A further improvement of our current re-ranking approach is that we could add a pre-filtering schema to select meaningful results in advance of the re-ranking process. Various features to describe the topic relevance of tweets have been studied in previous works [27, 28]. Naveed et al. [6] use several term and length features to measure interestingness. A possible future direction is the application of feature-based measurement to filter out meaningless items in the original results list.

In addition to the content-based approach, the alternative collaborative–based approaches can be investigated in Twitter search. Personalization can be achieved by incorporating users' social networks. Carmel et al. [13] investigate personalized search approaches based on user' social relations on social tagging system. Their work implies that the social relations derived from the user's social network can be reliable in predicting user interests and preferences. This direction is worthy of investigation on Twitter.

Finally, crowdsourcing techniques can be applied to the evaluation of personalized search approaches on Twitter. The evaluation framework should be able to integrate the Twitter API with crowdsourcing platforms. Two initial research questions are: How can we obtain Twitter user profiles of workers and how we can control their assessment quality. The possibility of building crowdsourcing-based evaluation framework for personalized microblog search can be studied in the future.

References

1. Java, A., et al., *Why we twitter: understanding microblogging usage and communities*, in *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007, ACM: San Jose, California. p. 56-65.
2. Teevan, J., D. Ramage, and M.R. Morris, *#TwitterSearch: a comparison of microblog search and web search*, in *Proceedings of the fourth ACM international conference on Web search and data mining*. 2011, ACM: Hong Kong, China. p. 35-44.
3. Efron, M., *Information search and retrieval in microblogs*. Journal of the American Society for Information Science and Technology, 2011. **62**(6): p. 996-1008.
4. Teevan, J., S.T. Dumais, and E. Horvitz, *Potential for personalization.* ACM Transactions on Computer-Human Interaction (TOCHI), 2010. **17**(1): p. 4.
5. Teevan, J., S.T. Dumais, and E. Horvitz. *Personalizing search via automated analysis of interests and activities*. in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 2005. ACM.
6. Naveed, N., et al., *Searching microblogs: coping with sparsity and document quality*, in *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011, ACM: Glasgow, Scotland, UK. p. 183-188.
7. Ghorab, M.R., et al., *Personalised information retrieval: survey and classification.* User Modeling and User-Adapted Interaction, 2013. **23**(4): p. 381-443.
8. Zhao, W.X., et al., *Comparing twitter and traditional media using topic models*, in *Advances in Information Retrieval*. 2011, Springer. p. 338-349.
9. Abel, F., et al., *Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web.* Proceedings of ACM WebSci, 2011. **11**.
10. Wang, Q. and H. Jin. *Exploring online social activities for adaptive search personalization*. in *Proceedings of the 19th ACM international conference on Information and knowledge management*. 2010. ACM.
11. Abel, F., et al., *Semantic enrichment of twitter posts for user profile construction on the social web*, in *The Semanic Web: Research and Applications*. 2011, Springer. p. 375-389.
12. Keenoy, K. and M. Levene, *Personalisation of web search*, in *Intelligent techniques for web personalization*. 2005, Springer. p. 201-228.
13. Carmel, D., et al. *Personalized social search based on the user's social network*. in *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009. ACM.
14. Xu, S., et al. *Exploring folksonomy for personalized search*. in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008. ACM.
15. Abel, F., et al., *Analyzing user modeling on twitter for personalized news recommendations*, in *User Modeling, Adaption and Personalization*. 2011, Springer. p. 1-12.

16. Vallet, D. *Crowdsourced Evaluation of Personalization and Diversification Techniques in Web Search*. in *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*. 2011.
17. Kwak, H., et al. *What is Twitter, a social network or a news media?* in *Proceedings of the 19th international conference on World wide web*. 2010. ACM.
18. Zhao, X. and J. Jiang, *An empirical comparison of topics in twitter and traditional media.* Singapore Management University School of Information Systems Technical paper series. Retrieved November, 2011. **10**: p. 2011.
19. Zhao, Z. and Q. Mei, *Questions about questions: an empirical analysis of information needs on Twitter*, in *Proceedings of the 22nd international conference on World Wide Web*. 2013, International World Wide Web Conferences Steering Committee: Rio de Janeiro, Brazil. p. 1545-1556.
20. Teevan, J., S.T. Dumais, and D.J. Liebling. *To personalize or not to personalize: modeling queries with variation in user intent*. in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. 2008. ACM.
21. Lau, C.H., Y. Li, and D. Tjondronegoro. *Microblog Retrieval Using Topical Features and Query Expansion*. in *TREC*. 2011.
22. Broder, A., et al. *Online expansion of rare queries for sponsored search*. in *Proceedings of the 18th international conference on World wide web*. 2009. ACM.
23. Naveed, N., et al. *Searching microblogs: coping with sparsity and document quality*. in *Proceedings of the 20th ACM international conference on Information and knowledge management*. 2011. ACM.
24. Magnani, M., et al., *Conversation retrieval from twitter*, in *Advances in Information Retrieval*. 2011, Springer. p. 780-783.
25. O'Connor, B., M. Krieger, and D. Ahn. *TweetMotif: Exploratory Search and Topic Summarization for Twitter*. in *ICWSM*. 2010.
26. Efron, M. *Hashtag retrieval in a microblogging environment*. in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010. ACM.
27. Duan, Y., et al., *An empirical study on learning to rank of tweets*, in *Proceedings of the 23rd International Conference on Computational Linguistics*. 2010, Association for Computational Linguistics: Beijing, China. p. 295-303.
28. Tao, K., et al., *What makes a tweet relevant for a topic?* Making Sense of Microposts (# MSM2012), 2012: p. 9.
29. Nagmoti, R., A. Teredesai, and M. De Cock. *Ranking approaches for microblog search*. in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*. 2010. IEEE.
30. Micarelli, A., et al., *Personalized search on the world wide web*, in *The Adaptive Web*. 2007, Springer. p. 195-230.
31. Raghavan, V.V. and H. Sever. *On the reuse of past optimal queries*. in *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. 1995. ACM.
32. Tan, B., X. Shen, and C. Zhai. *Mining long-term search history to improve search accuracy*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. ACM.
33. Liu, F., C. Yu, and W. Meng, *Personalized web search for improving retrieval effectiveness.* Knowledge and Data Engineering, IEEE transactions on, 2004. **16**(1): p. 28-40.

34. Qiu, F. and J. Cho. *Automatic identification of user interest for personalized search*. in *Proceedings of the 15th international conference on World Wide Web*. 2006. ACM.
35. Chirita, P.-A., C.S. Firan, and W. Nejdl. *Personalized query expansion for the web*. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.
36. Dou, Z., R. Song, and J.-R. Wen. *A large-scale evaluation and analysis of personalized search strategies*. in *Proceedings of the 16th international conference on World Wide Web*. 2007. ACM.
37. Claypool, M., et al. *Combining content-based and collaborative filters in an online newspaper*. in *Proceedings of ACM SIGIR workshop on recommender systems*. 1999. Citeseer.
38. Sugiyama, K., K. Hatano, and M. Yoshikawa. *Adaptive web search based on user profile constructed without any effort from users*. in *Proceedings of the 13th international conference on World Wide Web*. 2004. ACM.
39. Bender, M., et al. *Exploiting social relations for query expansion and result ranking*. in *Data engineering workshop, 2008. ICDEW 2008. IEEE 24th International Conference on*. 2008. IEEE.
40. Bao, S., et al. *Optimizing web search using social annotations*. in *Proceedings of the 16th international conference on World Wide Web*. 2007. ACM.
41. Heymann, P., G. Koutrika, and H. Garcia-Molina. *Can social bookmarking improve web search?* in *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 2008. ACM.
42. Zhou, D., S. Lawless, and V. Wade, *Improving search via personalized query expansion using social media.* Information retrieval, 2012. **15**(3-4): p. 218-242.
43. Tao, K., et al., *Twinder: a search engine for twitter streams*, in *Web Engineering*. 2012, Springer. p. 153-168.
44. Tao, K., et al. *Groundhog day: near-duplicate detection on twitter*. in *Proceedings of the 22nd international conference on World Wide Web*. 2013. International World Wide Web Conferences Steering Committee.
45. Steichen, B., H. Ashman, and V. Wade, *A comparative survey of Personalised Information Retrieval and Adaptive Hypermedia techniques.* Information Processing & Management, 2012. **48**(4): p. 698-724.
46. Feng, W. and J. Wang. *Retweet or not?: personalized tweet re-ranking*. in *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013. ACM.
47. Salton, G., A. Wong, and C.-S. Yang, *A vector space model for automatic indexing*. Communications of the ACM, 1975. **18**(11): p. 613-620.
48. Bontcheva, K. and D. Rout, *Making sense of social media streams through semantics: a survey.* Semantic Web, 2012.
49. Meij, E., W. Weerkamp, and M. de Rijke. *Adding semantics to microblog posts*. in *Proceedings of the fifth ACM international conference on Web search and data mining*. 2012. ACM.
50. Leung, K.W.-T., et al., *A framework for personalizing web search with concept-based user profiles.* ACM Transactions on Internet Technology (TOIT), 2012. **11**(4): p. 17.
51. Vallet, D. and P. Castells. *Personalized diversification of search results*. in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. 2012. ACM.

52. Han, X., et al., *Folksonomy-based ontological user interest profile modeling and its application in personalized search*, in *Active Media Technology*. 2010, Springer. p. 34-46.
53. Massoudi, K., et al., *Incorporating query expansion and quality indicators in searching microblog posts*, in *Advances in Information Retrieval*. 2011, Springer. p. 362-367.
54. Ganesan, P., H. Garcia-Molina, and J. Widom, *Exploiting hierarchical domain structure to compute similarity.* ACM Transactions on Information Systems (TOIS), 2003. **21**(1): p. 64-93.
55. Wang, Y., et al., *A Theoretical Analysis of NDCG Ranking Measures.* 2013.
56. Benson, E., A. Haghighi, and R. Barzilay. *Event discovery in social media feeds*. in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. 2011. Association for Computational Linguistics.