# Mining query subtopics from social tags

Maayan Zhitomirsky-Geffet

Information Science Department, Bar-Ilan University Ramat-Gan Israel

Yossi Daya

Information Science Department, Bar-Ilan University Ramat-Gan Israel

**Introduction.** One of the principal difficulties confronted by modern search engines is the overwhelming amount of, and lack of focus in, retrieved results. The goal of this study is identification and extraction of an assortment of diverse subtopics for the query and reduction and re-ranking of search results, according to the most prominent and discriminative subtopics.

**Method.** Social tagging as a 'wisdom of crowds' source for characteristic query subtopics was employed. A new method for extraction of subtopics from social tags was devised. Three different ranking measures for query result are presented and comparatively evaluated.

**Analysis**. We propose a new automatic methodology for evaluation of the quality of the extracted subtopics. For this purpose, we assess the minimal potential contribution of our approach to enhancement of query result ranking and focusing. The proposed methodology is based on standard information retrieval measures, such as precision, recall and the mean average precision.

**Results**. The obtained results with the Google search engine and the Delicious social bookmarking site as a source of social tags indicate that the mean average precision of search results was improved after their filtering and re-ranking according to the subtopics extracted from social tagging.

**Conclusion**. Social tags can be effectively utilized for query subtopics extraction. The quality of the subtopics can be automatically evaluated by their contribution to focusing and ranking query results. Query result representation can be improved by their selection and re-ranking according to the prominent social subtopics.

**Introduction**

As the internet has developed over the past two decades, accessing information has become one of its primary functions. Search engines have become the second most useful tool of the internet, following email. A survey conducted by Pew Internet (2010) indicates that 87% of internet users in the United States reported using search engines to some extent, in contrast to 94% percent who reported using electronic mail. Despite the improvements made to search engines, results often focus on one subtopic of the query, while the subject relevant to the user appears only marginally in the search results (Zheng *et al*., 2011). Compounding the problem is the fact that users often utilise brief generalized queries, because they do not know how to precisely define their information needs. This restricts the system's ability to understand the user's true interests, i.e., what subtopics to choose from the broader topic selected by the user. For example, search results for the word 'quotes' focus on the financial aspect of this term and not on its usage in writing and literature. In this situation, the user interested in the latter meaning of the query is forced to dig through the many retrieved results until he finds the desired information. Additionally, search results for multifaceted queries such as 'Mozart' can include a number of aspects such as: Mozart's bibliography, his works, musicians that play his works and video clips about the composer.

User research (Zhang and Moffat, 2006) indicates that average users check only the first ten search results. Moreover, Jansen and Spink (2003) have shown that 26% Web search sessions are approximately 5 minutes or less. Over 90% of searchers use only noun query terms (Jansen, Spink, and Saracevic, 2000). Issues include finding appropriate query terms, retrieving too many results, not retrieving enough results, and retrieving zero results among many others. Numerous solutions have been proposed for the problem of too much information, information which needs to be re-ranked and reduced (Finkelstein *et al*., 2001; Agichtein, Brill, and Dumais, 2006). However, these solutions do not address the problem of the query's lack of focus on subtopics relevant to the user. Searchers seldom reformulate their queries, even though there may be other terms that relate directly to their information need (Bruza, McArthur, and Dennis, 2000; Jansen *et al*., 2000). On the other hand, it has been found that users perform shifts in topics of their queries during the search process (Ozmutlu and Cavdur, 2008).

These problems highlight the necessity of diversifying the top search results to ensure that they cover a variety of subtopics of the query, and, on the other hand, to focus them on the subtopics of interest to the specific user. Problematic situations like the ones described above generally occur when the original query is vague and unfocused. As proclaimed by Henninger and Belkin (1994, page 1): *Information retrieval systems must not only provide efﬁcient retrieval, but must also support the user in describing a problem that s/he does not understand well.*

According to the review provided by Jansen (2005) there has been much work developing IR systems that offer some type of automated assistance. However, many searchers have difficulty effectively utilising IR systems. On the other hand, Jansen and Spink (2003) have found that searchers implemented the system's assistance 82% of the times they viewed the assistance. This finding indicates that the users perceived the offered assistance as beneficial to the vast majority of searches.

Broder (2002) claims that over half of the searches conducted on search engines are made to receive website navigational information or transactional information (downloading a file or purchasing a product) and not informational material. Consequently, it would be appropriate to present users with categories of subjects related to their search query (in place of a superficial, lengthy list of links) so that they could focus their search, thereby reducing and pinpointing results. These categories can be defined as *query subtopics*. Many short and general queries can be decomposed into several specific subtopics which represent different aspects and facets of the queries. For example, the query 'Lake Tahoe' would lead to subtopics such as its location, characteristics of its water, and local attractions (Efrati, 2012). Thus, there is a need to *identify* or *mine* the query's subtopics in order to retrieve more specific results for a user that cover different query's subtopics. Recently, several algorithms were developed for query subtopics mining (Hu *et al*., 2012; Wang *et al*., 2013) aiming at diversifying the search results. In particular, query context, semantics and logs were employed to extract the prominent subtopics of queries. Even the most popular search engine, Google, is now working to change search results from a long, unstructured listing of links to sites into a list of subtopics. The smart search would retrieve subtopics or attributes related to the query.

In contrast to previous research, in this work we propose to learn the most popular and discriminative user-centered subtopics of the query from other users who tagged this document on Web 2.0 social tagging sites, such as social bookmarking sites like Delicious or Diigo. In this context, the social Web might contribute to semantic search, since it can be assumed that whatever people thought about the results characterize the subtopics of the query.

Social tagging platforms provide a rare opportunity to take advantage of *the wisdom of crowds*, since it enables the identification of valuable information recommended by internet users. Social tagging is used to identify Web pages on social bookmarking platforms. On these platforms, internet surfers upload webpages that interest them and label them with tags or keywords that describe them. Earlier studies indicated that this information regarding Web pages can be employed to enhance the quality and ranking of retrieval results (Heymann, Koutrika, and Garcia-Molina, 2008). The information existing on these systems regarding Web pages (the volume of people who uploaded the page and the content of the tags assigned to a page) can signify the subject of the page and its relevancy, information which did not exist before these tags were uploaded to these platforms. The primary weaknesses of these systems are the relatively small volume of pages that they cover relative to the search engines and the usage of tags with their subjective implications or vague connotations.

This study analyses the potential latent in social tagging sites for the identification of subtopics for user queries and the focused reduction of search results. We integrate the abilities of standard search engines with the information existing on the social tagging and bookmarking sites.

Our primary research questions are as follows:

1. Can social tagging sites serve as a quality source for mining query subtopics?
2. Can result reducing and re-ranking, according to query subtopics extracted from social tags, improve the accuracy of information retrieval?

To answer these questions, we present an algorithm for mining the most popular and discriminative query subtopics from the tags of Web pages on social bookmarking sites. We hypothesize that social tags, after filtering and re-ranking, can represent a comprehensive set of subtopics of the query. These subtopics will reflect the aggregated *wisdom of crowds* description of the Web document.

Once identified, these subtopics can be used to reformulate the query by substituting its more general and short formulation with the concrete selected subtopics. Each subtopic is assigned a weight that determines its importance for the query. Further, the obtained results have to be re-ranked according to their relevance to the new query (consisting of the original query's subtopics). To estimate the importance of a subtopic to a query and of a result to subtopic, various measures were developed and comparatively tested.

In addition, we have developed an automated method to evaluate the minimum potential of utilising subtopics generated by social tags for enhancing retrieval results. The method is based on filtering out the least relevant subtopics of the query and the results linked to them, and re-ranking the remaining results. In the future, it will be possible to develop an interactive user interface that will present query subtopics and allow users to filter out subtopics that do not interest them. We expect that user selected subtopics will produce even better results than our automated system.

The rest of this article is organized as follows. The related work is reviewed in the next section. This research methodology is described, followed by the analysis of the obtained results before the conclusion of the paper.

**Related work**

In this chapter, we will review the studies that most closely relate to the goals of our research. As we are not aware of a previous research using social tagging for query subtopic mining, two types of works will be discussed separately: 1) utilization of social tagging for enhancing result ranking; and 2) query subtopic mining for retrieval enhancement.

**Social tagging as an aid for enhancing retrieval ranking**

Social tagging is a process in which a large number of users add metadata to shared content in the form of tags (Golder and Huberman, 2006). Lately, the world of social tagging has become very popular throughout the internet. These sites fall into two categories: social bookmarking sites that allow users to upload bookmarks and tag them, and informational sites that allow users to tag the site's existing information.

There are a number of popular social tagging sites, such as Delicious, Diigo and Flickr. These sites allow users to manage personal information and share it with other users. These systems have advantages over automated cataloging and ranking systems such as spiders and crawlers, since the tag-based cataloguing is performed by humans who understand the content of the information. In addition, users are able to tag internet pages that were not yet indexed by search engines (the invisible Web).

The question is whether this information can be used to upgrade the abilities of internet information retrieval systems. A number of studies (e.g., Heymann *et al*., 2008; Bischoff *et al*., 2008; Zhou *et al*., 2008) have assessed the hidden potential in social bookmarking systems for information retrieval. Bischoff *et al*. (2008) investigate the use of different kinds of tags for improving search. To this end, they analyse and classify tags from three most prominent social tagging websites: (Del.icio.us – http://www.delicious.com), (Last.fm – http://www.last.fm), and (Flickr – http://www.flickr.com).

The authors found that more than 50% of existing tags bring new information to the resources they annotate. A large amount of tags are accurate and reliable. In music domain for example 73.01% of the tags also occur in online music reviews. In addition, most of the tags can be used for search, and in most cases tagging behavior has the same characteristics as searching behavior. Heymann *et al*. (2008) collected a dataset including 3 million unique URL addresses from the Delicious website and a number of random queries were executed on the Yahoo search engine. Their results showed that 9% of the top 100 results were covered on Delicious, while 19% of the top 10 results existed on the social bookmarking site. Consequently, they concluded that social bookmarking sites can help with search engine ranking.

A significant amount of popular search terms coincide with bookmarking tags. Therefore, it can be inferred that the information latent in the tags can significantly help with information retrieval. Most of the tags are relevant and comprehensible to users. Interestingly, about 12.5% of the pages uploaded by users are new and not yet indexed by search engines. At the time of this study, the information located on the social bookmarking sites is not adequately significant to influence information retrieval. Nonetheless, if social tagging and bookmarking platforms continue growing at the same rate that they have grown in recent years, their influence will become more significant. In (Zhou *et al*., 2008) the authors introduce a probabilistic method for employing social tagging from Delicious to improve information retrieval. They discover topics in documents and queries and enhance document and query language models by incorporating topical background models.

In recent years, a number of articles have been written suggesting ranking methods based on information existing on social bookmarking sites (e.g., Yanbe *et al*., 2007; Bao *et al*., 2007; Zhang *et al*., 2009; Choochaiwattana and Spring, 2009). Yanbe *et al*. (2007) suggested enhancing result ranking by integrating the PageRank algorithm with the tag information on social bookmarking sites. Bao *et al*. (2007) devised two algorithms for ranking according to social bookmarking: 1) the SocialSimRank algorithm which assesses the resemblance between the query and the tags; and 2) the SocialPageRank algorithm which measures the quality of the page according to its popularity. Their study indicates that these two algorithms improved significantly the quality of result ranking. A similar method was presented in an additional study (Zhang *et al*., 2009). This method ranked search results according to a query's resemblance to the tags, with the rank weight determined by the popularity of the tags. Another study conducted in 2009 (Choochaiwattana and Spring, 2009) considered the number of social tags that matched the query terms. The study authors report that the ranking method that yielded the best results ranked the document according to the number of users who tagged it on Delicious with tags that matched the terms of the search query.

Inspired by the above findings we choose to employ Delicious social tags as a source for the most significant tags (subtopics of the query) in order to pinpoint the search and reduce its results number. Note, however, that the algorithm proposed in our study does not involve the analysis of semantics or lexical resemblance of the results or the queries and the tags. In contrast, it focuses exclusively on the analysis of the potential contribution of social tagging.

**Mining search query subtopics for enhancing information retrieval**

Understanding the intent of the user's query is one of the primary goals of information retrieval systems (Wang *et al*., 2013). With the widespread usage of internet search engines, this issue has become relevant and challenging. Queries entered by users on search engines are often ambiguous or multifaceted (Clarke *et al*., 2008). This is particularly true because queries tend to be brief (Wang *et al*., 2013). A number of recent studies have suggested solving the query reformulation and diversification problem by mining subtopics and clustering search results accordingly when a query is submitted. This approach is discussed in more detail below.

Hu *et al*. (2012) suggested a method for automatic mining of subtopics from search log data as a source for the wisdom of crowds present in search engines. Researchers used the logs to amass data about searches, search results, and links clicked by users. They assumed that the user is interested in one subtopic of each query, and therefore all the URLs s/he clicks serve as an indication of that one subtopic. In addition, query clarifications by the user include keywords that serve as the subtopics of the original query. The researchers also suggest a method for improving ranking results based on subtopics and the popularity of the search results. After submitting the search query, the users are presented with a result screen with a list of subtopics from which they can choose the one that interests them. After the user makes his/her selection, the URL addresses associated with this subtopic will be moved to the top of the result list. Another study (Yin *et al*., 2009) employs Wikipedia and Google Insights in order to extract diverse query subtopics. Their approach indicates an improvement in contrast to similar algorithms of the query diversification track of TREC 2009.

Recently, a method was suggested (Wang *et al*., 2013) for mining and ranking query subtopics by extracting fragments including the query terms from document text. Text fragments were then used to divide the query into subtopics. The authors of this method measured the results of their study against the NTCIR-9 INTENT Chinese subtopic mining test collection (Song *et al*., 2011). This set of tests included a list of 100 topics in Chinese, a series of possible descriptions and a list of subtopics for each topic. In addition, the researchers evaluated their results against other studies conducted on this subject (for example, Zheng *et al*., 2011). The results indicated that their technique efficiently mined and ranked subtopics.

These studies generally evaluated the quality of the extracted subtopics by various metrics that assessed their novelty and diversity (Clarke *et al*., 2008). In contrast, we suggest using query subtopics to focus and reduce results and not to diversify or expand them. Additionally, we use a different source for obtaining the wisdom of crowds, social bookmarking. These subtopics are not employed for submitting more queries, but for filtering and re-ranking of the original query results based on the assumption that they include all relevant material. At the conclusion, we suggest a new automated technique for evaluating the potential contribution of focusing a query by reducing its subtopics.

**Method**

To evaluate the potential of social tagging as quality query subtopics, we built a collection of queries and extracted social tags for their results. Delicious was chosen as the source site for the social tags as it is the most popular amongst the general public and has been used in other studies. Google, the most popular search engine in the world, was chosen for extracting query search results.

**Constructing a Set of Queries and their Results**

A set of queries was extracted from the Google Trends site. For the benefit of the study, a set of queries was selected from the most popular search terms of 2012 in the United States. The query collection process was conducted in a number of stages:

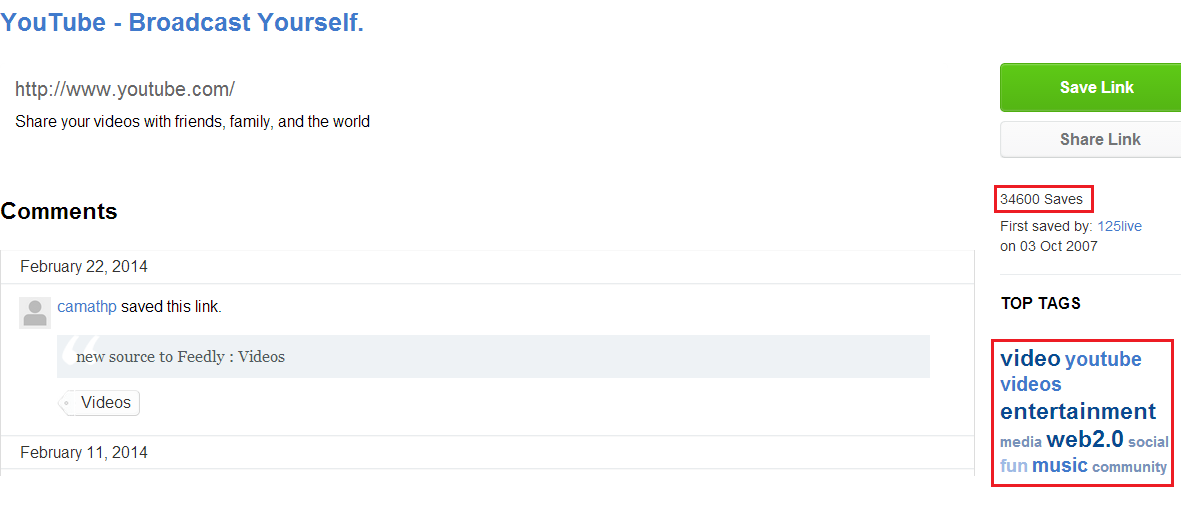
1. At the first stage, the most popular searches for 2012 in the United States were extracted from each of 20 categories;
2. Four queries were randomly selected from each category - 80 queries altogether; and
3. Out of these 80 queries, 13 queries with the highest amount (at least 75%) of the retrieved Google results tagged on Delicious were chosen.

Using these queries, we tested the potential effectiveness of the system in the ideal situation, when most of the search results retrieved for a query exists on social bookmarking sites with tag information.

At the next stage, the most popular queries that could be defined as named entities were obtained from each category. A named entity identifies an entity that can be clearly differentiated from a list of other items with similar characteristics. For example, personal names, product names, geographic locations and websites and companies (<http://searchbusinessanalytics.techtarget.com/definition/named-entity>.) 20 queries were extracted, one from each category. The rationale behind this classification was the construction of a set of clearly defined queries so that the relevance of their results could be evaluated. This classification is very important for the evaluation of the system (e.g., the query 'Youtube' is defined and better evaluated than the query 'bank').

The more popular a query is, the more it tends to be general (Wang *et al*., 2013). We therefore chose a collection of popular queries, since they would reflect the problem of generality and lack of definition in queries, and because they were the most thoroughly covered on social bookmarking sites like Delicious. We then compared the sub-collection of most popular queries to the named entity query collection, which are rarer and better defined.

We chose 100 of the first Google results for each of the above-chosen 33 queries to serve as the result collection of our experiment. This corpus of 3300 online documents was then judged for relevance according to the accepted specifications of the Text Retrieval Conference (TREC) (<http://trec.nist.gov/data/reljudge_eng.html>) by three independent human annotators. The annotators were guided as follows: *If you were writing a report on the subject of the topic and would use the information contained in the document in the report, then the document is relevant*. Only binary judgments (relevant or not relevant) were made, and a document was judged relevant if any piece of it was relevant (regardless of how small the piece may have been in relation to the rest of the document). The final judgment was designated according to the majority vote. The Kappa values for inter-annotatoragreement were good (ranging between 0.69 and 0.81 for different pairs of annotators).



**Figure 1: A sample page from Delicious presenting user information and tags for Youtube.**

A sample page of user tags is presented in Figure 1. Here the URL http://[www.youtube.com](http://www.youtube.com) is displayed. The number of users who uploaded the address to Delicious is circled in red at the top right of the screen. The tag cloud reflecting the popular tags attributed to the URL is circled in red at the center right of the screen. A tag cloud is a visual presentation of textual information that usually contains keywords/tags. The significance of each tag is expressed in the size of its font. The Delicious tag cloud consists of four differently sized fonts.

**Query subtopic selection: methods for tag weighting, ranking and filtering**

For every query (out of 33) we collected the list of tags associated with each of its top 100 Google results that also appeared on Delicious. Then, a combined tag list representing all the results of the query was constructed. The next stage was to mine the most prominent tags for a query that would be further defined as its subtopics. This involves calculating a relative weight of each tag for the query. Two measures were devised and compared for this purpose. The first was to consider subtopics' *popularity* on Delicious, while the second was to compute the *discriminativeness* of the subtopic.

As can be viewed in Figure 1, the Delicious interface discloses the following information about a URL address:

* *R –* result popularity. The number of users who uploaded the URL address to the site.
* *T –* tag popularity. Cloud of popular tags in four differently sized fonts. The font size reflects the importance of the tag without providing a numeric weight.

We aimed to calculate a weight reflecting the relative weight of the tags for a specified query (the number of people who utilized the tag) when given the quantitative weight of the URL address (the number of people who uploaded the result to the site), in order to create a uniform scale for tags associated with different results.

Hence, we defined a baseline metric *QTW* (*Query Tag Weight*) for weighting the tag, *t*, for a specified query, *q*, such that for each results *r* in query result set *RQ* so that as *t* is linked to on Delicious, it reflects both *R* (the result popularity) and *T* (the tag popularity) as follows:

where *N* is the normalization factor in order to create a uniform scale for tags associated with different results (10 was the optimal value for *N* in our experiments).

|  |  |  |
| --- | --- | --- |
| **Number of Results** | **QTW Weight** | **Tag** |
| 65 | 13338 | Education |
| 87 | 12464 | Math |
| 50 | 10789 | Resources |
| 60 | 10528 | Mathematics |
| 52 | 9872 | Maths |
| 42 | 6563 | Games |
| 39 | 6439 | Interactive |
| 19 | 4101 | Kids |
| 11 | 3696 | Science |
| 7 | 3361 | Tools |

**Table 1: A list of top-10 tags for the query 'math' ranked according to their *QTW* values.**

Intuitively, this measure weights the subtopics by their relative *popularity* on Delicious. It assigns higher ranks to tags with a higher frequency of appearance which are linked to the documents with a higher number of uploads on Delicious. However, a sample testing of tag ranking based on the above weights indicated that such highly popular tags that reached the top of the list were too general and were often identical to the text of the query, as shown in Table 1. Consequently, these tags covered most of the results without distinguishing between their subtopics.

Therefore, an alternative measure for mining subtopics from tags was required. This measure had to promote more specific tags highly characteristic for some of the results but not for the others. Hence, to create a list of more focused and discriminative tags, we calculated their TF-IDF weight (Manning, Raghavan, and Schutze, 2009, p. 118). This measure captures the significance of a term in a document relative to the other documents in the corpus. The classic formula was adapted for our goals as TF-IQF (tag frequency – inverse query frequency). The *QTW* value of each tag was positioned in the formula as TF:

where *M* is the overall number of results for query *q* on Delicious and *|RQ|* is the number of query results to which the tag *t* is linked. This measure weights the subtopics by their *discriminativeness*. This decreases weights of tags frequently used for most of the query results and increases weights of tags frequently used for a certain result. Further, to filter out tags that are spam, too general, rare, or lacking significance from the final list of representative tags for a query, we determined that they would be included in the list of potential subtopics for this query only if:

* The tag is attributed to at least 5% of query results which exist on Delicious;
* The TF-IQF weight of the tag is equal to or higher than the average TF-IQF weight of the tags in the list.

|  |  |  |
| --- | --- | --- |
| **Number of results** | **TF-IQF** **Weight** | **Tags** |
| 7 | 3775 | tools |
| 11 | 3426 | science |
| 50 | 2908 | resources |
| 12 | 2871 | teaching |
| 19 | 2829 | kids |
| 11 | 2641 | learning |
| 5 | 2619 | research |
| 39 | 2430 | interactive |
| 7 | 2340 | fun |
| 42 | 2266 | games |

**Table 2: Top-10 tags for the query 'math', ranked according to their TF-IQF weights**.

Table 2 presents the top ten tags on the list for the query 'math', ranked according to their TF-IQF weight. It can be observed that some general tags (like education) and morphological variation of the query term (like mathematics) were excluded from the list.

|  |  |  |
| --- | --- | --- |
| **Top-10 TF-IQF tags for query 'Twitter'** | **Top-10 TF-IQF tags for query 'Pizza Hut'** | **Top-10 TF-IQF tags for query 'fashion'** |
| Facebook | dining | style |
| networking | online | clothing |
| social networking | restaurants | online |
| Web design | restaurant | photography |
| tools | shopping | design |
| web2.0 | pizza | shop |
| design | lessons | blog |
| social | incentives | clothes |
| Twitter tools | food | moda |
| social media | kids | trends |

**Table 3: Top-10 TF-IQF-weighted tags for 3 different queries in the dataset.**

Table 3 shows the top-10 tags sorted by their TF-IQF weights for a general query 'fashion' and for two named entity queries: 'Twitter' and 'Pizza Hut'.

Finally, the remaining tags can serve as meaningful and discriminative subtopics for the query. Likewise, in contrast to other methods, typically these subtopics do not include the terms of the original query. While being semantically related to the original query, they are not necessarily morphologically or lexically similar to it. This is difficult to achieve through other sources such as search log data or analysis of lexical resemblance that are based on query terms collocations (as described in Section 2). Between 40 and 100 subtopics were created for each of the 33 queries analysed in this study.

**Measures for reducing and re-ranking query results according to subtopics**

As described above, the subtopics for each query were mined out of the most prominent tags on Delicious with high values of TF-IQF measure. Only tags that were assigned to at least 5% of query results which exist on Delicious were considered as subtopics for a query. Thus, each of the subtopics was linked to the corresponding results.

The objective of the next step is to utilize these subtopics in order to reduce and re-rank the results in a manner that will enhance retrieval accuracy. This can be done by filtering out some isoteric results with low relevance to the query and then by re-ranking the rest of the results such that the most relevant of them appear at the top. To this end, we needed to compute a relative weight for each result for a given query. In the framework of this study, we tested three methods for weighting query results. Each of these weighting methods can be used for further ranking the results by simply sorting them in descending order of their weights.

All the analysed measures took the Delicious weight into consideration. In other words, they consider the number of people who uploaded the address to the site and/or tagged it with a tag that was chosen as a query subtopic.

The following are the ranking methods that were analysed:

1. Result Weight (RW)

Upon receiving the list of subtopics (tags), we gather all the results linked to them and rank the results in descending order of *R*, the result popularity, on Delicious. This ranking method does not consider the weight of the selected subtopics, but only of the results themselves to which the tags are attributed. In this case, the subtopics are only used to filter out the results that were not linked to any subtopic on the selected list. The rationale behind this method is to assess the potential of directly using raw Delicious popularity rates of the results to improve the retrieval precision.

2. Tag Weight (TW)

Given a list of subtopics, the system gathers all of the results attributed to them. For each result, it sums up the weight of all the subtopics (*QTW*) attributed to it. It then sorts the results according to this total weight in descending order. The cumulative weight of the subtopics for a result captures the number of users who used these subtopics to tag this result. More formally, when *r* is a query result from top-100 Google results and also appearing on Delicious, and (*tl, t2,...,tn*) is the list of subtopics that users selected and attributed as tags to result *r,* then the weight of *r* for the query *q* is defined as follows:

This measure ranks higher results associated with more subtopics of the query and which were more popular on Delicious. Thus, this measure does not rely solely on the result popularity, but rather combines it with the rating of the subtopics of the query. Also, it is based on the subtopics' popularity rates rather than on their discriminativeness.

3. TF-IQF Weight (TIW)

For each result *r* from top-100 Google results for a given query which also appears on Delicious we sum up the TF-IQF weights of the subtopics (*t1, t2, …, tn*) attributed to it, the weight of *r* for this query is defined as follows:

This measure is based solely on the cumulative weights of the subtopics associated with the specified result. Thus, it shows the potential of subtopics in improving the ranking and precision of retrieval. Note that this measure considers subtopics' discriminativeness rather than their popularity rates.

Note that for all 3 measures presented in this subsection, the query results that are not linked to any of the selected query subtopics are removed from the result list for this query.

**Automatic evaluation of the quality of the extracted subtopics**

In contrast to previous studies, we assessed the quality of subtopics that were extracted from social bookmarking in an automated manner according to their potential contribution to retrieval. Therefore, the results of each query were reduced and re-ranked iteratively, at different steps. At the first step, we tested the quality of the results with 100% of the selected query subtopics according to the subtopic ranking and filtering procedure above. At the second step, we assessed the quality of the results with 90% of the subtopics (the ones ranked the highest), and so on, until only the top 10% of subtopics were left in the list. At each step, only the results attributed to the chosen subtopics remained and were re-ranked.

As long as the result was linked to a subtopic that was not filtered out from the list at this step, it remained in the result list for the query. The result was only removed once all of the tags attributed to it were filtered out. This process reduces the number of query results according to the prominent subtopics represented by the social tags.

After each step, we calculated the remaining result quality with the standard recall precision (van Rijsbergen, 1979, Evaluation section) and the Mean Average Precision (MAP) (Manning *et al*., 2009, p. 159) measures. The *recall* of the system is the proportion of relevant documents actually retrieved in answer to a search query. The *precision* of the system is the proportion of retrieved documents that is actually relevant to the query. The Mean Average Precision measure captures both precision and recall and computes the quality of result ranking for a query set for various recall levels. It first calculates an average precision for each query as a percentage of relevant documents among top-*k* retrieved documents for different values of *k*. These average precision values for individual queries are then averaged over all the queries in the set. Thus, with the more relevant documents concentrated at the top ranks in the result list, the higher is the Mean Average Precision value of the specified ranking method.

In this manner, we also determine the optimal percentage of tags to filter out. In other words, at which step does the Mean Average Precision score indicate maximal improvement. The suggested methodology also simulates user selection of query subtopics to focus and reduce results. But in our case, subtopics were chosen according to what was considered the most important and characteristic of the result collection by a substantial number of users on the Delicious website.

This type of methodology is indispensable for evaluating information retrieval systems, because user testing is complicated and subjective in addition to being expensive and time consuming. To achieve credible results, it would be necessary to recruit a large number of users who would examine various search scenarios. On the other hand, the direct evaluation of the quality of extracted subtopics (as performed in previous work such as Clarke *et al*., 2008; Wang *et al*., 2013) is based on subjective human judgment. In contrast, our methodology offers a fast, objective, automated evaluation of many scenarios and various filtering and ranking metrics. The automated evaluation that we conducted in this work assessed the bottom level of effectiveness of the suggested system, since we can assume that user selection of relevant subtopics (instead of the automated system) would result in even higher average precision.

We could not compare our method to other techniques mentioned in the literature review, since those that were based on social bookmarking did not perform our task - query subtopic mining, but only re-ranked the query results. On the other hand, works that did mine query subtopics evaluated them directly according to the TREC database. These queries were inappropriate for our study, since they were not adequately covered on Delicious. Additionally, we believe that for our goals, it is more important to evaluate the contribution of subtopic mining to retrieval enhancement than to compare their quality isolated from their practical application.

**Analysis and Results**

First we compared Google's precision to that of Delicious for our 33 selected queries.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Delicious precision in %** | **Google precision in %** | **No. of results appearing in Delicious out of 100 Google results (estimated Delicious recall)** | **Query text** | **Query no.** |
| 86 | 86 | 87 | bank | 1 |
| 72 | 61 | 74 | Bible | 2 |
| 92 | 91 | 85 | college | 3 |
| 57 | 49 | 68 | Facebook | 4 |
| 73 | 68 | 80 | fashion | 5 |
| 81 | 81 | 90 | games | 6 |
| 71 | 66 | 86 | Google | 7 |
| 99 | 96 | 76 | hotels | 8 |
| 88 | 83 | 76 | Java | 9 |
| 95 | 93 | 93 | math | 10 |
| 100 | 100 | 84 | news | 11 |
| 100 | 100 | 86 | recipes | 12 |
| 95 | 95 | 80 | weather | 13 |
| 85.4 (13.04) | 82.2 (16.33) | 81.92 (6.87) | all | average (standard deviation) |
| **Table 4: Overlapping results for the top queries.** | | | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Delicious precision in %** | **Google precision in %** | **No. of results appearing in Delicious out of 100 Google results (estimated Delicious recall)** | **Query text** | **Query no.** |
| 78 | 72 | 74 | YouTube | 1 |
| 91 | 75 | 32 | USPS | 2 |
| 76 | 74 | 63 | Apple | 3 |
| 67 | 71 | 27 | Bank of America | 4 |
| 96 | 91 | 24 | Pizza Hut | 5 |
| 93 | 93 | 46 | Xbox | 6 |
| 57 | 50 | 14 | Walgreens | 7 |
| 94 | 92 | 68 | Lottery USA | 8 |
| 94 | 77 | 16 | Home Depot | 9 |
| 91 | 78 | 55 | Yahoo | 10 |
| 95 | 95 | 83 | Steve Jobs | 11 |
| 79 | 87 | 39 | CNN | 12 |
| 84 | 74 | 75 | Twitter | 13 |
| 100 | 96 | 53 | Craigslist | 14 |
| 61 | 63 | 74 | Amazon | 15 |
| 78 | 83 | 41 | ESPN | 16 |
| 80 | 78 | 60 | Las Vegas | 17 |
| 67 | 60 | 73 | Shakespeare | 18 |
| 94 | 95 | 71 | Yoga | 19 |
| 94 | 94 | 81 | American Airlines | 20 |
| 83.45 (12.46) | 79.9 (12.73) | 53.45 (21.66) | all | average (standard deviation) |
| **Table 5: Overlapping results for the named entity queries.**  For the data presented in Table 4 and Table 5 the t-test for paired samples was performed. The differences between Google and Delicious average precision were found significant with p < 0.01 for the data in Table 4. For the data in Table 5, the differences between Google and Delicious average precision were found significant with p < 0.05.  Tables 4 and 5 indicate that for both the top and named entity queries for Delicious the average precision is about 3% higher than that of Google. Note that precision values are similar for both groups of queries. However, there is a significant difference in recall on Delicious relative to Google for different types of queries. The top queries have received a relatively high recall value on Delicious (82% on average) while those for the named entity queries are much lower (53% on average). | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TIW ranking** | | **TW ranking** | | **RW ranking** | | **No. of top selected sub topics** |
| **Delta in % compared to the original Google ranking** | **Average MAP** | **Delta in % compared to the original Google ranking** | **Average MAP** | **Delta in % compared to the original Google ranking** | **Average MAP** |
| 4.68 | 0.89 | 5.32 | 0.90 | 0.77 | 0.85 | 100 |
| 5.02 | 0.89 | 5.50 | 0.90 | 0.92 | 0.85 | 90 |
| 5.52 | 0.90 | 5.96 | 0.90 | 1.52 | 0.86 | 80 |
| 5.60 | 0.90 | 5.98 | 0.90 | 1.76 | 0.86 | 70 |
| 5.93 | 0.90 | 6.54 | 0.91 | 2.34 | 0.87 | 60 |
| 5.96 | 0.90 | 6.65 | 0.91 | 2.31 | 0.87 | 50 |
| 6.36 | 0.91 | 6.91 | 0.91 | 3.11 | 0.88 | 40 |
| 6.81 | 0.91 | 7.60 | 0.92 | 3.66 | 0.88 | 30 |
| 6.73 | 0.91 | 7.25 | 0.92 | 4.47 | 0.89 | 20 |
| 8.05 | 0.93 | 9.45 | 0.94 | 7.56 | 0.92 | 10 |
| **Table 6: Summary of the average improvement in MAP scores for all the queries, according to the subtopic filtering steps.** | | | | | | |

In Table 6 **Google original Mean Average Precision scores are used as a baseline for comparison. The paired-sample t-test was performed to compute the significance level of the differences between the Mean Average Precision of the proposed measures and the baseline. The differences were found significant with p < 0.01 only for TW and TIW measures.**

We then computed and compared the Mean Average Precision scores of the original Google ranking as a baseline to the rankings according to the metrics that were presented in the previous section. The findings in Table 6 exhibit that the Mean Average Precision scores of TW and TIW metrics were progressively enhanced with a similar significant improvement over the baseline, except for RW. This performance demonstrates that filtering subtopics and results linked to them does contribute to result precision and ranking and focuses the query result list such that mostly relevant results are concentrated at the top ranks.

The TW measure was the most effective for re-ranking results and achieved up to 10% enhancement to the Mean Average Precision scores. TW measure is the only one that combines both result's popularity on Delicious with the weights of query's subtopics associated with it. Hence, the tag popularity on Delicious for the query results (which is not considered by our RW measure) is a crucial factor for identification of the high quality subtopics. On the other hand, the result's popularity, which is not considered explicitly in TIW, is important as well. Interestingly, subtopics' popularity (reflected in TW measure) was found to be a better indicator of result's relevance to a query than subtopics' discriminativeness (summed up in TIW based on TF-IQF scores).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Google baseline ranking** | **RW Ranking** | | **TW Ranking** | | **TIW Ranking** | |
|  | Average MAP (standard deviation) | Average MAP (standard deviation) | Delta % | Average MAP (standard deviation) | Delta % | Average MAP (standard deviation) | Delta % |
| **Top queries** | 0.84 (0.15) | 0.88 (0.12) | 4.76 | 0.91 (0.10) | 8.33 | 0.90 (0.12) | 7.10 |
| **Named entity queries** | 0.86 (0.09) | 0.865 (0.13) | 1.74 | 0.91 (0.11) | 5.81 | 0.91 (13.38) | 5.81 |
| **Table 7: Summary of the average enhancement (over all steps) of Mean Average Precision scores for the top and named entity queries in separate.** | | | | | | | |

**The paired-sample t-test was performed to compute the significance level of the differences between the Mean Average Precision of the proposed measures and the baseline in Table 7. The differences were found significant with p < 0.05 only for TW and TIW measures.**

**Figure 2: Distribution of the average improvement over the baseline Google Mean Average Precision scores for all steps (axis X shows the percentage of top subtopics) and various ranking methods for the top queries.**

**Figure 3: Distribution of the average improvement over the baseline Google Mean Average Precision scores for all steps (axis X shows the percentage of top subtopics) and various ranking methods for the named entity queries.**

In summary, the elicited results have exhibited that on average, for all queries and at all steps, two out of the three suggested ranking methods achieved Mean Average Precision scores higher than that of the baseline Google ranking (see Table 7 and Figures 2, 3 above). We therefore conclude that there is a positive potential for reducing, focusing and re-ranking search engine results through the employment of tags as query subtopics. It can be assumed that when users select relevant subtopics (instead of the automated system) precision will increase even more.

**Improving Google ranking by subtopic-based result reduction**

In the previous section we observed that a large number of queries had a high percentage of relevant results that appeared on both Google and Delicious. Therefore, we conducted an additional experiment. In this experiment, we employed social tags as subtopics exclusively for Google result reduction. The remaining results were ranked according to the order of their original ranking on Google. The quality of this ranking was compared to the baseline Google ranking over all the results and to the best Delicious-based ranking (TW). For this test, we collected the results that existed on both Google and Delicious for each query.

**Figure 4: Distribution of average improvement of the *reduced* Google ranking.**

In Figure 4 we present the distribution of average improvement of the reduced Google ranking by applying subtopic-based result reduction, over the baseline Google Mean Average Precision scores for all queries at all steps (axis X shows the percentage of top subtopics), in comparison to the best ranking of Delicious, TW. The t-test for paired samples indicates that the improvement over the baseline Google rankings was significant for both, the reduced Google ranking and the TW Delicious ranking, with p < 0.01.

As can be observed in Figure 4, reducing the results increased the average Mean Average Precision scores for all of the steps for the Google ranking method. Interestingly, Google ranking of the remaining results were quite similar to the best Delicious-based ranking with no significant difference between the two rankings. This leads to the conclusion that employing query subtopics, extracted from social tags, for reducing Google results does enhance the quality of the result list. This finding indicates that Delicious-based subtopics help sift out many irrelevant results.

**Conclusions and future work**

This study presented a number of metrics for query subtopic mining by filtering, weighting and ranking social tags from social bookmarking sites. Our further goal was to utilise these subtopics to pinpoint the query and to reduce the amount of the results presented to the user. Our tests indicate that social tags can serve as query subtopics and can contribute to effective reducing and re-ranking of search engine results by increasing the accuracy and quality of the top ranked results. We also found that subtopics' popularity is more effective for improving result ranking than subtopics' discriminativeness. However, discriminativeness is a better criterion than popularity for subtopic mining from social tags. As the social tagging realm expands, it will cover more of the internet. Consequently, systems that integrate search engines with social tagging sites will improve the quality of information retrieval.

However, as we have seen throughout the work, the primary limitation of our approach is the insufficient level of websites covered on Delicious. We saw that the level of overlapping between the search engine and the social bookmarking system influenced the performance of the system: performance improvement for the top queries was superior to its performance for the named entity queries. In other words, there is a need to expand social tagging coverage of the Web. These limitations might decrease the performance, especially recall and the coverage, of the results obtained by our method. For more specific queries there might be a too small amount of results tagged with social tags on Delicious. This can be accomplished by integrating additional sources that provide social tagging data such as Diigo (http://www.diigo.com) and Twitter (http://www.twitter.com).

In the future, we plan to develop a system for information retrieval with an interactive user interface. This interface will present subtopics when a query is submitted, replacing lengthy, exhausting lists of results. These subtopics will be calculated from social tags originated in various sites (including social networks like Twitter). Based on this interface, we will be able to check the actual level of query result enhancement with users. The users will select the subtopics related to their information needs and will judge the relevance of the results based on the meaning of the focused query.

**Acknowledgement**

We are grateful to Dan Buchnick from the Information Science Department in Bar-Ilan University for fruitful discussions of the ideas and useful comments for the article.

**About the Authors**

Dr. Zhitomirsky-Geffet is an Assistant Professor in the Department of Information Science in Bar-Ilan University. Her main research fields include Internet research, semantic Web, social Web, and Web-based information retrieval. She can be contacted at: [Maayan.Zhitomirsky-Geffet@biu.ac.il](mailto:Maayan.Zhitomirsky-Geffet@biu.ac.il).

Yossi Daya received his M.A. degree at the Department of Information Science of Bar-Ilan University in Israel. His main areas of interest include: Internet research, social Web and information retrieval on the Web. He can be contacted at: [Yossi.Daya@gmail.com](mailto:Yossi.Daya@gmail.com).

**References**

Agichtein, E., Brill, E. & Dumais, S. T. (2006). [Improving Web search ranking by incorporating user behavior information](http://www.mathcs.emory.edu/~eugene/papers/sigir2006ranking.pdf). In S. Dumais, D. Hawking, & K. Jarvelin (Eds.), *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval (pp. 19-26).* Seattle, Washington, USA, August 2006.

Bao, S. G., Xue, X. Wu, Y. Yu, B. Fei & Z. Su. (2007). Optimizing Web search using social annotations. In P. Patel-Schnider, P. Shenoy, C.Williamson, & M. Zurko (Eds.), *WWW '07: Proceedings of the 16th International Conference on World Wide Web* (pp. 501-510). New York, NY, 2007.

Bischoff K., Firan C. S., Nejdl W., & Paiu R. (2008). Can all tags be used for search? In S. Amer–Yahia, K. Choi, A. Chowduri, D. Evans, A. Kolcz, I. Manolescu, J. Shanahan, Y. Zhang (Eds.), *CIKM '08 Proceedings of the 17th ACM Conference on Information and Knowledge Management*, (pp. 193-202). New York, NY, 2008. ACM.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum*, *36*(2), 3-10.

Choochaiwattana, W. & Spring, M. B. (2009). Applying social annotations to retrieve and re-rank Web resources. In Mahadevan V., Yi Xie (Eds.), *ICIME '09 : Proceeding of 2009 International Conference On Information Management and Engineering,(pp.* 215-219). Kuala Lumpur, Malaysia, 3-5 April 2009. IEEE, Los-Alamitos, CA,.

Bruza, P., McArthur, R., & Dennis, S. (2000). Interactive internet search: keyword, directory and query reformulation mechanisms compared. In N. Belkin, P. Ingwersen, M. Leong, & E. Yannakoudakis (Eds.), *SIGIR '00 Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information* *Retrieval* (pp. 280–287). ACM, New York, NY, 2000.

Clarke, C. L. , M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Buttcher, & I. MacKinnon. (2008). Novelty and diversity in information retrieval evaluation. In T. Choua, M. Leong, S. Myaeng, D. Oard, F. Sebastiani (Eds.), *SIGIR '08 Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 659–666). ACM, New York, NY, 2008.

Efrati, A. (March, 2012). Google gives search a refresh. *Wall Street Journal*, retrieved from <http://online.wsj.com/news/articles/SB10001424052702304459804577281842851136290>  [(Archived by WebCite® at http://www.webcitation.org/6WCnmui0Z)](http://www.webcitation.org/6WCnmui0Z)

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z.,Wolfman, G., & Ruppin, E. (2001). Placing search in context: the concept revisited. In W. B. Croft & G. Marchionini (Eds.), *Proceedings of the ACM Transactions on Information System, 20*(1), 116–131. ACM.

Golder, S. & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198-208.

Henniger, S. & Belkin, N. (1994). Interfaces issues and interaction strategies for information retrieval systems. In I. Katz, R. Mack, L. Marks, J. Miller (Eds.), *CHI '95 Conference On Human Factors In Computing Systems* (pp. 401-402), ACM, New York, NY 1995.

Heymann, P., Koutrika, G., & Garcia-Molina, H. (2008). Can social bookmarking improve Web search? In A. Broder, S. Chakrabarti, M. Najork (Eds.), *WSDM '08: Proceeding of the 2008 International Conference On Web Search and Data Mining (*pp. 195-206), New York, NY 2008.

Hu, Y., Qian, Y., Li, H., Jiang, D., Pei, J., & Zheng, Q. (2012). Mining query subtopics from search log data. In J. Callan, W. Hersh, Y. Maarek, & M. Sanderson (Eds.), *SIGIR '12 : Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, (pp. 305-314). ACM, New York, NY, USA.

Jansen B. J., Spink A., & Saracevic T. (2000). [Real life, real users, and real needs: a study and analysis of user queries on the Web. ‏](http://scholar.google.com/citations?view_op=view_citation&hl=iw&user=gyhAqMQAAAAJ&citation_for_view=gyhAqMQAAAAJ:UeHWp8X0CEIC) *Information Processing & Management, 36*(2), 207-22.

Jansen, B. J. & Spink, A. (2003). An analysis of Web information seeking and use: documents retrieved versus documents viewed. In R. Lienhart, Y. Liu, S. Satoh, X. Yu, Z. Zha (Eds.), *Proceedings of the 4th International Conference on Internet Computing* (pp. 65-69). Las Vegas, NV, 23–26 June.

Jansen, B. J. (2005). Seeking and implementing automated assistance during the search process. *Information Processing & Management, 41*(4), 909-928.

Manning, C., Raghavan, P., & H. Schütze. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Ozmutlu, S., H.C. Ozmutlu, & A. Spink (2008). Automatic new topic identification in search engine transaction logs using multiple linear regression. In Ralph H. Sprague Jr. (Eds.), *Proceedings of the 41st Hawaii International Conference on System Sciences* (pp. 14), Waikoloa, HI, 7-10 Jan, 2008.

Pew Internet. (May, 2010). Search engine use. Retrieved on July 16, 2011 from http://www.pewinternet.org/Static-Pages/Trend-Data/Online-Activites-Total.aspx. [(Archived by WebCite® at http://www.webcitation.org/6WDktfiB0)](http://www.webcitation.org/6WDktfiB0)

Song, R., Zhang, M., Sakai, T., Kato, M. P., Liu, Y., Sugimoto, M., Wang, Q., & Orii, N. (2011). Overview of the NTCIR-9 Intent task.In N. Kando, D. Ishikawa, M. Sugimoto (Eds.), *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies,* (pp. 82-105) .

Text Retrieval Conference (TREC). English relevance judgments. Retrieved June 28 2013 from <http://trec.nist.gov/data/reljudge_eng.html>.  [(Archived by WebCite® at http://www.webcitation.org/6WDldBg9K)](http://www.webcitation.org/6WDldBg9K)

Van Rijsbergen C. J. (1979). *Information Retrieval.* 2nd edition, Butterworth-Heinemann Newton, MA. Retrieved from: <http://www.dcs.gla.ac.uk/Keith/Preface.html> [(Archived by WebCite® at http://www.webcitation.org/6WDnLgP4d)](http://www.webcitation.org/6WDnLgP4d)

Wang, Q., Qian, Y., Song, R., Dou, Z., Zhang, F., Sakai, T., & Zheng, Q. (2013). Mining subtopics from text fragments for a Web query. *Information Retrieval, 16*(4), 484-503.

Yanbe, Y. A. Jatowt, S. Nakamura, & K. Tanaka. (2007). Can social bookmarking enhance search in the Web? In R. Larson, E. Rasmussen, S. Sugimoto, & E. Toms (Eds.), *JCDL '07 Proceedings of the 2007 Conference on Digital Libraries*, (pp. 107–116), New York, NY.

**Yin, D.,** Z. Xue, X. Qi & Davison, B.D.. (2009). **Diversifying search results with popular subtopics. In**E. M. Voorhees and Lori P. Buckland (Eds.),*TREC '09 Proceedings of* the 18th Text Retrieval Conference (pp. 5-14). Gaithersburg, MD: NIST.

Zhang, Y., & Moffat, A. (2006). Some observations on user search behaviour. *Australian Journal of Intelligent Information Processing Systems*, *9*(2), 1-8.‏

Zheng, W., Wang, X., Fang, H., & Cheng, H. (2011). An exploration of pattern-based subtopic modeling for search result diversification. In L. Cassel, G. Newton, & M. Wright (Eds.), *JCDL '11 Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 387-388), New York, NY.

Zhou D., Bian J., Zheng S., Zha H., & Giles, C. L. (2008). **Exploring social annotations for information retrieval**. In R. Chen, H. Hon, J. Huai, & Y. Liu (Eds.), *WWW '08 Proceedings of the 17th International World Wide Web 2008 Conference* (pp. 715-724), Beijing China.