Bridging the Gap – Using External Knowledge Bases for Context-Aware Document Retrieval

Benjamin Köhncke[1], Patrick Siehndel[1], and Wolf-Tilo Balke[2]

[1] L3S Research Center; Hannover, Germany
[2] TU Braunschweig, Germany
{koehncke,siehndel}@L3S.de, balke@ifis.cs.tu-bs.de

Abstract. Today, a vast amount of information is made available over the Web in the form of unstructured text indexed by Web search engines. But especially for searches on abstract concepts or context terms, a simple keyword-based Web search may compromise retrieval quality, because query terms may or may not directly occur in the texts (vocabulary problem). The respective state-of-the-art solution is query expansion leading to an increase in recall, although it often also leads to a steep decrease of retrieval precision. This decrease however is a severe problem for digital library providers: in libraries it is vital to ensure high quality retrieval meeting current standards. In this paper we present an approach allowing even for abstract context searches (conceptual queries) with high retrieval quality by using Wikipedia to semantically bridge the gap between query terms and textual content. We do not expand queries, but extract the most important terms from each text document in a focused Web collection and then enrich them with features gathered from Wikipedia. These enriched terms are further used to compute the relevance of a document with respect to a conceptual query. The evaluation shows significant improvements over query expansion approaches: the overall retrieval quality is increased up to 74.5% in mean average precision.

Keywords: conceptual query, query expansion, semantic enrichment.

1 Introduction

Today's information gathering in many domains is almost entirely focused on Web searches. However, handling the growing amount of available information poses severe challenges even for focused information providers, such as digital libraries and archives. When searching for information, users usually describe their broad information needs with several keywords which are likely to be different from the words used in the actually relevant documents. As a consequence the results returned by the information provider may miss relevant documents with respect to the user's information needs. This leads to a dramatically decreased retrieval quality and thus a bad usage experience. To guarantee high quality retrieval it is therefore important to bridge the gap between the query terms and the documents' vocabulary. The challenge of expressing the user's information need by finding the right query terms is widely known as the vocabulary problem [1]. Users often try to solve this problem by

S.R. Urs, J.-C. Na, and G. Buchanan (Eds.): ICADL 2013, LNCS 8279, pp. 11–20, 2013.

refining their query, i.e. adding or changing query terms in case the retrieval results have not been satisfying [2]. However, considering scenarios where users are searching for information about abstract concepts the problem of word mismatch is even bigger: such abstract concepts or context terms hardly ever occur directly in Web documents. Imagine a user who is interested in *information retrieval*. By entering the conceptual query '*information retrieval*' he only receives documents dealing with this very general concept. Closely related and also relevant documents not containing the exact term, like, for instance, documents about *Web search*, will not be returned. This also holds for more specific conceptual queries, like, e.g., *polyomavirus infections* in the biomedical domain or searches for chemical classes, like, e.g., *alcohol*, in the domain of chemistry.

To solve this problem, in some domains documents are already pre-annotated with suitable context terms. The most prominent example is the MEDLINE corpus which is currently the largest document repository of life science and biomedical documents, containing more than 20 million publications. Each of these documents is manually annotated with several terms from the Medical Subject Heading (MeSH) ontology which offers a controlled vocabulary for indexing and retrieval purposes. However, document collections like MEDLINE are a rare case and most collections lack suitable context annotations. For most domain specific collections no suitable controlled vocabularies or even better, ontologies, are available.

The traditional way of searching for documents relevant for conceptual queries is to use query expansion. It expands the query term issued by a user with suitable related terms, called expansion terms, matching the documents' vocabulary. In general, query expansion leads to higher recall, but strongly decreases the retrieval precision. The reason is that usually the context of the query is not known and thus the expansion terms do not meet the user's search intention. More advanced retrieval models, like Latent Semantic Analysis (LSA), try to solve this by producing sets of concepts related to the documents and their contained terms. However, as we will see in our experiments, the resulting quality is still not sufficient. For digital library providers it is important to enable conceptual queries while also ensuring their high quality requirements. While the context of the query can hardly be determined, the context of each document is defined based on its contained terms. Thus, instead of expanding the query, the idea of this paper is to expand the documents with semantic annotations. To find suitable annotations for semantic enrichment external knowledge bases are necessary.

In previous work we have already shown the usefulness of Wikipedia categories to summarize documents' content [3]. Therefore, in the presented approach we use external knowledge provided by Wikipedia to semantically enrich documents, bridging the gap between conceptual queries and documents' vocabulary. We extract the most important terms from each document and enrich them with several semantic features gathered from Wikipedia. The enriched terms are used to compute the relevance of a document to a conceptual query. Our experiments show that our approach outperforms traditional query expansion methods using statistical query expansion, showing an increase of more than 30% in mean average precision. We also compare against stronger baselines using LSA and Random Indexing showing an improvement of more than 15%. All results have been proven to be statistically significant. Another advantage of our approach is that it can be easily integrated in the metadata enrichment process of a digital library.

The rest of the paper is organized as follows: In section 2 we give an overview of the related work, followed by a detailed description of our approach in section 3. The evaluation is presented in section 4. Finally, we conclude and give an outlook to our future work in section 5.

2 Related Work

One major problem of current information retrieval systems is their low retrieval quality caused by the inaccuracy of the query composed of a few keywords compared to the actual user information need. In case the user enters a query containing several topic-specific keywords the system is able to return good results, but in most cases queries are rather short and since language is inherently ambiguous this leads to worse retrieval results [4]. The most critical problem influencing the retrieval quality is the term mismatch problem (also known as the vocabulary problem [1]), meaning that the query terms chosen by the user are often different from the vocabulary used in the documents. In case of conceptual queries one possibility is to let users choose from a fixed set of context terms from a controlled vocabulary, like, e.g., provided by the MeSH ontology. In our scenario a conceptual query is defined as the search for documents relevant for an abstract concept, like, for instance, *Polyomavirus Infections* in the biomedical domain. In current search interfaces this context restriction is offered using facetted browsing, see, e.g., GoPubMed. Other approaches use controlled vocabularies to suggest suitable query terms to the user to avoid the vocabulary problem. In [5] such an approach is presented showing that discipline-specific search term recommendations improve the retrieval quality significantly.

In general, one well known method to overcome the term mismatch problem is automatic query expansion. A good summary of different query expansion approaches is given in [4]. Automatic query expansion approaches can be generally categorized into global and local analysis [6]. Global analysis is usually based on statistics of co-occurrences of pairs of words, resulting in a matrix of word similarities [7]. Although these approaches are relatively robust, the computation of the corpus-wide statistics is computational intensive. In contrast, local analysis uses only a subset of the returned documents for a given query to find suitable expansion terms. This kind of local feedback has the drawback that the documents retrieved in the initial search strongly influence the retrieval quality. These methods have, for instance, been evaluated on TREC datasets, see, e.g., [8] or [9]. Since these methods need to know which documents are relevant for a given query and pseudo relevance needs a multi-phase retrieval process they cannot be applied directly to commercial search engines. Therefore, recently a number of expansion approaches have been developed using query logs. Generally, these approaches also derive expansion terms from (pseudo-) relevant documents, which are extracted from search logs by analyzing user clicks, see, e.g., [10] or [11]. This approach is further extended in [12]. The authors replaced the correlation model with a statistical translation model which is trained on pairs of user queries and the titles of clicked documents. Furthermore, they translated the word-based translation model to a concept-based model. These concepts are used to expand the original query. However, also the idea of representing queries and documents as a set of related concepts rather than a bag of individual words is used in several approaches. The first necessary step is to identify the concepts in queries and documents. Afterwards, the concepts are introduced as hidden variables in the query expansion model to capture term dependencies.

Also methods like Latent Semantic Analysis (LSA) [13] can be used in the proposed scenario. LSA analyzes a set of documents to produce a set of concepts which are related to the documents and terms. The main idea behind this approach is that words which are used in the same context may have a similar meaning. These concepts are generated by applying singular value decomposition to a generated word-document matrix.

Other approaches try to use the knowledge of external information sources to enrich the user query [14], [15]. An interesting approach dealing with effective query formulation is presented in [16]. The authors present a unified framework trying to automatically optimize the use of different information sources for formulating the user query. The experimental results show better results than state-of-the-art baseline approaches performing concept weighting, query expansion or both.

In contrast to these approaches, in our approach we use external knowledge bases to enrich the documents with semantic annotations. In previous work we already analyzed the usefulness of Wikipedia as external source for semantic enrichment [3].

3 Semantic Enrichment Using Wikipedia

3.1 Architecture Overview

In this section we describe the architecture of our approach. The goal is to semantically enrich documents enabling conceptual queries. Our basic idea is to extract important terms from documents and use community maintained knowledge bases to compute the semantic similarity between these terms. Previous work used Wikipedia to help users finding relevant query terms and interactively guide them on their search [17]. In [3] we used Wikipedia categories to describe the content of chemical documents. Experiments showed that also for specialized domains like chemistry, knowledge gathered from Wikipedia is more useful for domain experts than a domain specific ontology. Since Wikipedia uses the wisdom of the crowds, which has been proven to provide tremendous quality [18], the contained knowledge is growing fast and updated regularly. In our approach we further exploit the provided knowledge by creating semantic features based on the Wikipedia categories and the link structure. Fig. 1 gives an overview of our approach.

The architecture is composed of two basic components. The term extractor is responsible for annotating and extracting important terms from the documents. For annotating the documents we use the Wikipedia Miner toolkit [19]. The main purpose of the Wikipedia Miner is to annotate a given fulltext in the same way a human would annotate a Wikipedia article. The methods are based on a machine learning approach which is used to identify relevant terms and links them to Wikipedia. The approach is two folded in the way that the first task is to disambiguate the terms which occur in a given text, and the second task is to check whether the detected terms are useful links to Wikipedia articles.

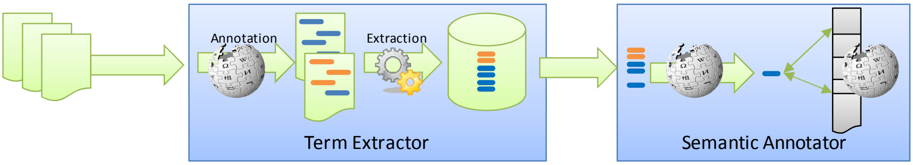

Fig. 1. Architecture Overview

The extracted terms are further processed by the semantic annotator. For each term its associated Wikipedia categories, and its in- and out-links are extracted. These features are used for computing the semantic similarity between different terms. The measures used for calculating the feature similarities are based upon the Jaccard coefficient and are described in detail in [20].

3.2 Retrieval Workflow

The user enters a query and submits it to our system. For retrieving a ranked list of relevant documents the system is composed of two components: the semantic annotator and the ranking engine. Fig. 2 gives an overview of the retrieval workflow.

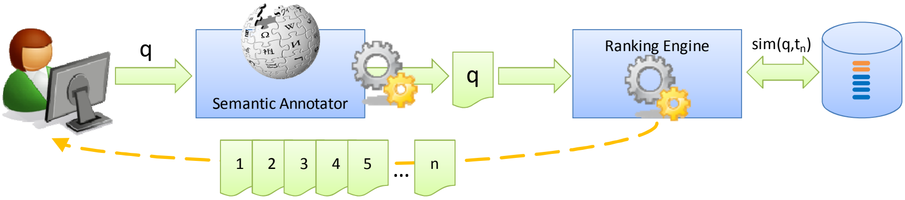

Fig. 2. Retrieval Workflow

The query term *q* is analyzed by the semantic annotator which enriches it with the different similarity features extracted from Wikipedia. The ranking engine receives the enriched query term and creates a ranked list containing all other terms. In case *q* is already known in our system the semantic similarity ranking is directly received from the relational database. Otherwise, it is necessary to compute the similarity to all terms known to the system. For our repository, containing 34324 different terms, the similarity computation for an unknown term took less than three seconds. Finally, the documents are ranked according to the similarity values of their contained terms. The relevance of a document *d* to a query *q* is computed as follows:

$$rel(\mathrm{q}, d) = \frac{\left(\sum_{\mathrm{t} \in \mathrm{q}} \sum_{t_x \in T_d} \frac{sim(t_x,\mathrm{t}) * \tau_{\mathrm{t_x}}}{|T_d|} * \frac{\omega_{\mathrm{t_x}}}{\Omega_{\mathrm{t_x}}}\right)}{|q|} \quad (1)$$

where T_d is the set of all terms included in *d* and τ is a boosting factor to give terms occurring in the document's title a higher weight. Each query *q* can consist of several terms *t*. Furthermore, ω denotes the number of times a term occurs in a document. This value is normalized by the number of times the term occurs in the whole collection, denoted by Ω. Finally, the score is normalized by the number of terms in *q*.

4 Evaluation

As document repository we use 122640 documents from the PUBMED Central corpus which is part of the MEDLINE repository. Each document in this set is manually annotated with several terms from the MeSH ontology which offers a controlled vocabulary for indexing and retrieval purposes. These terms are abstract concepts describing the general context of the respective document. Therefore, we also use MeSH terms as query terms in our experiments. To find a set of suitable query terms we analyzed the distribution of the MeSH terms in our document collection. As possible query terms we considered all terms occurring in less than 1000 but more than 10 documents. From this set we randomly choose 80 query terms which also occur in Wikipedia. As document set for the experiments we used all documents that have been annotated with at least one of these query terms. The MeSH annotation is done manually by domain experts resulting in high quality. Therefore, for our evaluations we considered all documents annotated with the respective MeSH term as relevant hits. In total our set contains 10791 documents.

4.1 Lucene Index, Statistical Query Expansion and Latent Semantic Analysis

In this experiment, we searched for all query terms in the documents' fulltext. Therefore, we created a Lucene fulltext index including all documents from our subset. To analyze the retrieval quality we considered all documents annotated with the respective MeSH term as relevant hits. The documents have been ranked according to the BM25 ranking model using standard parameters. As evaluation measure we computed the mean average precision (MAP) and the average recall over all queries. Our experiment results in a MAP of 31.53% and an average recall of 37%.

To enhance the MAP and the recall we also used a statistical query expansion method. We computed the term-to-term co-occurrence matrix based on the documents of our subset. The position of the term in the document is also taken into account, meaning two terms that are close together will get a higher score. Furthermore, we used popularity thresholds defining a required minimum and maximum popularity. Terms not fulfilling these thresholds are also not used as expansion terms. We used the following retrieval model: Let q be the query term and $C=\{c_1,c_2,\ldots,c_n\}$ the set of all expansion terms. For the expanded query the queries are formulated as q *OR* c_1 *OR* c_2 *OR* … *OR* c_n, meaning all documents are returned containing the query term or at least one expansion term. Finally, the query is expanded with the top-k co-occurring terms. Fig. 3a shows the results for the top-k expansion terms.

The best MAP of 40.28% is reached for the top-21 expansion terms. As expected, the more terms are added to the query the higher is the recall. The maximum recall of 81.91% is reached for the top-58 terms.

Beside query expansion we also evaluated how an LSA approach would perform in this scenario. To analyze the performance of an LSA based approach we used the Semantic Vectors[1] toolkit which is build upon Apache Lucene. We used LSA and

[1] https://code.google.com/p/semanticvectors

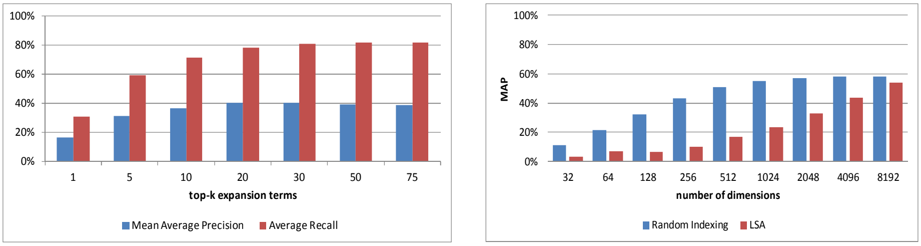

Fig. 3. MAP and average recall for the top-k expansion terms (a)MAP for Random Indexing and LSA (b)

Random Indexing for building the vectors for our corpus. Random Indexing is an alternative approach to standard word space models, which is efficient and scalable [21]. For both methods we used the standard parameters and varied the number of dimensions used for the vectors. We started with 32 dimensions and went up to 8192 dimensions. The resulting MAP of booth methods was continuously growing with an increase of the number of dimensions. We did not use a higher number of dimensions because of the runtime complexity and memory requirements for the resulting model. The results are shown in Fig. 3b. We see that the MAP based on Random Indexing is higher in all cases, reaching up to a maximum MAP of 58.2%. Using a very high number of dimensions we archive quite similar results using LSA (54.1%).

4.2 Semantic Enrichment Using Wikipedia

In this experiment we evaluate the usefulness of our approach for conceptual queries. For each document and each annotated term a confidence value has been computed describing the reliance of the assignment between Wikipedia article and term. We did two main experiments analyzing the influence of the confidence value. In the first one we computed the MAP using different confidence thresholds. In this experiment, for computing the relevance of a document to a query term only the assigned terms having a higher confidence value than the threshold are used. In the second main experiment we ordered the assigned terms for each document by their confidence values. For the relevance computation only the top-k terms for each document are used. Furthermore, in both experiments we also analyzed the influence of giving terms occurring in the document's title a higher weight. In addition, we also considered the number of times the term appears in the document in the ranking function. To do not prefer frequently used terms that are not descriptive for the respective document, we normalized this value by the number of times the term occurs in the whole collection. Since our method computes the relevance of a query to all documents in our set, the recall is always 100% and therefore not meaningful at all. To evaluate the different rankings and compare them to the baseline approaches we compute MAP.

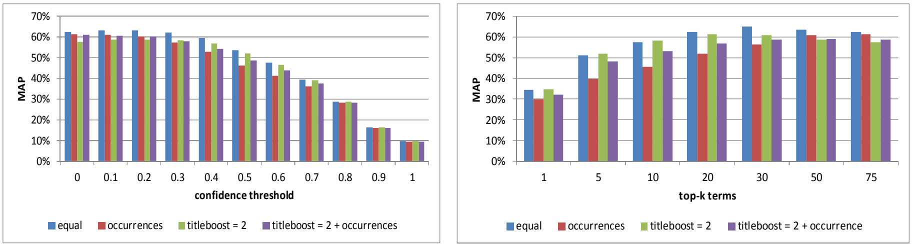

Fig. 4. MAP for varying confidence thresholds (a)MAP for top-k terms (b)

Fig. 4a shows the results for the confidence threshold experiment. A confidence threshold of 0 means that all terms have been used for the relevance computation. The results show that giving the terms occurring in the documents' title a higher score leads to a decrease of the MAP. We only show the results for a title boost factor of 2, meaning the title terms are twice as important as other terms. In our experiments we varied the boosting factor from 1 to 15. But, the higher the boosting factor the worse the results. Also the number of occurrences of a term does not lead to better overall results. The combination of title boost an occurrences leads to better results for smaller thresholds than using the features alone, but the overall best results are achieved if all terms are considered as equally important. The best MAP of 63.14% is reached for a confidence threshold of 0.1. The higher the threshold the fewer is the number of assigned terms for each document.

Fig. 4b shows the evaluation results for using the documents' top-k terms. We analyzed the distribution of assigned terms for the documents in our collection. Around 10% of the documents in our collection have more than 75 terms assigned. Therefore, we computed the MAP for up to the top-75 terms. Please note, we always used all documents and only limited the number of assigned terms. As in the confidence threshold experiment the best results are achieved if all terms are considered as equally important. Using a title boost factor or taking the number of occurrences into account does not lead to better retrieval results. The best MAP of 65.14% is reached for using the top-31 terms of each document. The MAP is slightly higher as for the confidence thresholds.

As last experiment we analyzed the different combinations of the features used in our similarity measure. Fig. 5 shows the results for the different combinations. This experiment shows that the categories are performing worst with a best MAP of 59% for the top-31 terms. The overall best MAP of 74.5% is reached for the top-61 terms using only the in-links feature.

Overall we showed that for conceptual queries the proposed method leads to better results than state-of-the-art retrieval models. The best baseline approach was Random Indexing achieving an MAP of 58.2%. Our approach significantly outperforms (p-value of 0.03 using a two-tailed t-test) the best baseline by achieving an MAP of 74.5%.

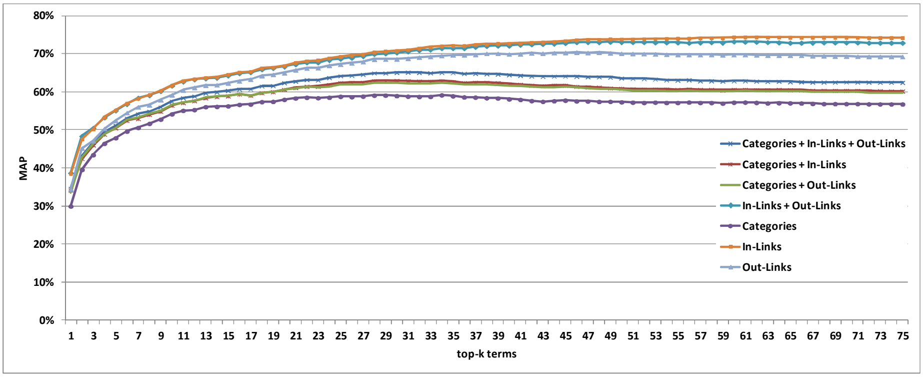

Fig. 5. Comparing MAP of different features

5 Summary and Outlook

One major problem digital library providers have to solve is the well-known vocabulary problem. Users often search for information using query terms not directly occurring in the documents. Considering conceptual queries this problem is even more complicated. To allow for suitable document retrieval meeting the high quality standards of a digital library it is important to bridge the gap between the user’s query and the documents’ fulltext. State-of-the-art solutions suggest using statistical query expansion. More advanced approaches are based on LSA or LSI models. However, especially for conceptual queries their retrieval quality is often still insufficient.

In this paper we presented an approach allowing for conceptual queries using external knowledge as provided by Wikipedia. We took a document collection from the PubMed Central repository and extracted the most important terms from each document. These terms are semantically enriched with features gathered from Wikipedia. Finally, the relevance of a document to a conceptual query is computed resulting in a ranked retrieval list. Our evaluation has shown that our approach outperforms state-of-the-art query expansion and LSA approaches resulting in an increase of the mean average precision of 58.2% for LSA to 74.5% for our approach. All results have been proven to be statistically significant (p-value of 0.03 using a two-tailed t-test). The proposed method bridges the gap between user queries and documents' fulltext by using external knowledge sources for semantic enrichment. To summarize, our results show that even without manual annotating the retrieval quality can be improved meeting the high quality standards of a digital library.

For future work we plan to also consider other knowledge bases instead of Wikipedia to bridge the gap between conceptual queries and documents. Furthermore, we plan to extend our approach with a personalization component to learn the best similarity measure dependant on the individual user.

References

1. Furnas, G.W., et al.: The vocabulary problem in human-system communication. Communications of the ACM 30(11), 964–971 (1987)
2. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proc. of Int. Conf. on World Wide Web, WWW (2004)
3. Köhncke, B., Balke, W.-T.: Using Wikipedia categories for compact representations of chemical documents. In: Proc. of Int. Conf. on Information and Knowledge Management, CIKM (2010)
4. Carpineto, C., Romano, G.: A Survey of Automatic Query Expansion in Information Retrieval. ACM Computing Surveys 44(1), 1–50 (2012)
5. Lüke, T., Schaer, P., Mayr, P.: Improving Retrieval Results with discipline-specific Query Expansion. In: Zaphiris, P., Buchanan, G., Rasmussen, E., Loizides, F. (eds.) TPDL 2012. LNCS, vol. 7489, pp. 408–413. Springer, Heidelberg (2012)
6. Xu, J., Croft, W.: Query expansion using local and global document analysis. In: Proc. of Int. Conf. on Research and Development in Information Retrieval, SIGIR (1996)
7. Jing, Y., Croft, W.: An association thesaurus for information retrieval. In: Proceedings of RIAO, pp. 1–15 (1994)
8. Cao, G., et al.: Selecting good expansion terms for pseudo-relevance feedback. In: Proc. of Int. Conf. on Research and Development in Information Retrieval, SIGIR (2008)
9. Metzler, D., Croft, W.B.: Latent concept expansion using markov random fields. In: Proc. of Int. Conf. on Research and Development in Information Retrieval, SIGIR (2007)
10. Cui, H., Wen, J., Nie, J., Ma, W.: Query expansion by mining user logs. IEEE Transactions on Knowledge and Data Engineering 15(4), 829–839 (2003)
11. Wang, X., Zhai, C.X.: Mining term association patterns from search logs for effective query reformulation. In: Proc. of Int. Conf. on Information and Knowledge Management, CIKM (2008)
12. Gao, J., Nie, J.: Towards Concept-Based Translation Models Using Search Logs for Query Expansion. In: Proc. of Int. Conf. on Inf. and Knowledge Management, CIKM (2012)
13. Deerwester, S., et al.: Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science 41(6), 391–407 (1998)
14. Hu, J., Wang, G., Lochovsky, F., Sun, J., Chen, Z.: Understanding user's query intent with wikipedia. In: Proc. of Int. Conf. on World Wide Web, WWW (2009)
15. Xu, Y., et al.: Query dependent pseudo-relevance feedback based on wikipedia. In: Proc. of Int. Conf. on Research and Development in Information Retrieval, SIGIR (2009)
16. Bendersky, M., et al.: Effective query formulation with multiple information sources. In: Proc. of Int. Conf. on Web Search and Data Mining, WSDM (2012)
17. Milne, D., et al.: A knowledge-based search engine powered by wikipedia. In: Proc. of Int. Conf. on Information and Knowledge Management, CIKM (2007)
18. Surowiecki, J.: The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business. Economies, Societies and Nations (2004)
19. Milne, D., Witten, I.H.: An open-source toolkit for mining Wikipedia. Artificial Intelligence 194, 222–239 (2012)
20. Köhncke, B., Balke, W.-T.: Context-Sensitive Ranking Using Cross-Domain Knowledge for Chemical Digital Libraries. In: Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J. (eds.) TPDL 2013. LNCS, vol. 8092, pp. 285–296. Springer, Heidelberg (2013)
21. Sahlgren, M.: An Introduction to Random Indexing. In: Proc. of the Methods and Applications of Semantic Indexing Workshop (2005)