EXPLOITING ENTITY SEMANTICS FOR QUERY EXPANSION

ABSTRACT

Many user queries nowadays contain references to *named entities*, which has motivated the development of new methods that exploit entity semantics for query expansion. At the same time, Wikipedia has been widely recognized as a large network of named entities, where entity-related articles are organized into a comprehensive hierarchy of categories and present summarized information on these entities in the so-called *infoboxes*. In this paper, we present a new query expansion method that uses entity semantics derived from Wikipedia. The main appeal of our method is that, differently from previous methods in the literature, it exploits valuable human-refined information available in infoboxes to obtain candidate query expansion terms and to associate entities identified in queries with categories. Indeed, by taking advantage of the semantic structure implicitly provided by infoboxes templates, we leverage well-known term-selection functions, adapting them to deal properly with entities and ultimately improving their accuracy in selecting good query expansion terms. Experimental results show that our method presents gains of 19.05% (from 0.1370 to 0.1631) in terms of MAP and 77.99% (from 0.2381 to 0.4238) in terms of P@10. In addition, the two approaches for obtaining expansion terms based on information found in infoboxes present a better trade-off between quality of results and time required to process the expanded query.

KEYWORDS

Query Expansion, Query Reformulation, Entity, Semantic, Wikipedia, Infobox

1 INTRODUCTION

Query expansion is the process of enriching queries posed by users with additional meaningful terms, in an attempt to improve the search results obtained. It is motivated by the observation that users are frequently unable to chose the best terms to formulate their queries. In fact, it is a notorious fact that users often use only a couple of terms in their queries, what may lead to poor results. Many methods for query expansion have been proposed over the years. They are based on several different approaches, including: relying on user's relevance feedback, employing information theoretic metrics to select expansion terms from documents in the collection, and acquiring knowledge from external sources to obtain additional query terms (Baeza-Yates and Ribeiro-Neto, 2011). More recently, it has been observed that many queries posed nowadays by users contain references to *named entities*. Indeed, according to a recent analysis reported in (Guo et al., 2009), the percentage of queries mentioning entities reaches 71% already. This has motivated the development of new methods that exploit entity semantics for query expansion. In summary, these methods identify entities in queries, match entities in entities repositories, and reformulate the original queries using entity-related information taken from these repositories. The best-known methods in the literature use Wikipedia as a repository of entity-related information (Xu et al., 2008, 2009; Maskey and Dakka, 2009).

Wikipedia is widely recognized as a large network of related named entities, containing articles that cover around 1.5 million entities and 6.5 million relationships between them (Maskey and Dakka, 2009). More important, most of Wikipedia articles bring summarized information on entities, in the form of explicitly structured data represented in the so-called *infoboxes*. In addition, articles are organized into a comprehensive hierarchy of categories, covering a broad spectrum of knowledge domains. Remarkably, these information sources and semantic structures are built and maintained by a voluntary community-based process whose goals is to enforce quality and accuracy.

In this paper, we present a new query expansion method that also uses knowledge acquired from Wikipedia, exploiting the entity semantics available in its content and structure. We propose three alternative approaches for obtaining terms related to Wikipedia entities. The first, we call *content-based*, uses terms available in the textual content of articles. The other two, called *property-based* and *relationship-based*, use, respectively, attribute values and

references to other entities, both found in infoboxes. Indeed, the main appeal in our method is that, differently from methods previously presented in the literature, it extensively uses information available in infoboxes in a principled way. Such information is not only closely related to entities, but it is also continuously refined by human editors. Thus, they form valuable sources of semantic knowledge to obtain terms to be used in query expansion. In addition, our method uses infoboxes to associate entities identified in queries with categories. By doing so, we leverage previously proposed term-selection functions, adapting them to deal properly with entities, ultimately improving the accuracy of such functions in selecting the best terms to be used for query expansion.

To evaluate our method, we performed a comprehensive set of experiments using ClueWeb09 Category B, a standard TREC collection that is regarded as being representative of the Web (Koolen and Kamps, 2010). We experiment different configurations, combining our proposed approaches for obtaining sets of entity-related terms from Wikipedia with several different strategies for selecting expansion terms from these sets. The experiments show that all of our three approaches have a positive impact on the quality of the results, leading to gains of 19.05% (from 0.1370 to 0.1631) in terms of MAP and 77.99% (from 0.2381 to 0.4238) in terms of P@10. In addition, the *property-based* and *relationship-based* approaches present a better trade-off between quality of results and time required to process the expanded query.

The remainder of this paper is organized as follows. Section 2 covers related work. Section 3 describes our proposed method in details. Section 4 reports the experiments we have carried out and analyses the obtained results. Finally, Section 5 presents our conclusions and directions for future work.

2 RELATED WORK

Query expansion has been used as an effective way to improve the quality of information retrieval systems. Several methods use relevance feedback, information theoretic approaches and knowledge acquired from external sources to automatically obtain additional query terms. The most important traditional query expansion techniques are covered in (Baeza-Yates and Ribeiro-Neto, 2011). Recently, new techniques that exploit entity semantics for query reformulation have been proposed. Basically, the techniques identify entities in queries, use them to match entities in entities repositories, and reformulate the original queries using entity-related information. Along this line, the work in (Xu et al., 2008) presents an automatic query reformulation approach that first identifies the most representative entities for a given query. Then it collects a set of terms/phrases closely related to these entities, according to Wikipedia entity pages. Finally the query is enriched by adding the most semantic related terms/phrases for new round of retrieval.

The work in (Xu et al., 2009) also exploit Wikipedia as a repository of entities and uses a relevance feedback framework for query reformulation. Similarly to (Xu et al., 2008), their method first detects the most representative entity in a query. Second, it classifies TREC topics into three categories based on Wikipedia: 1) entity queries, 2) ambiguous queries, and 3) broader queries. Then, the authors propose and study the effectiveness of three methods for expansion term selection, each modeling the Wikipedia based pseudo-relevance information. The work in (Maskey and Dakka, 2009) uses Wikipedia to build a large network of related named entities. Query expansion experiments were carried out to validate the usefulness of named entities for document retrieval tasks. The proposed query expansion approach uses all the relationships of an entity as feature to select the top k best ranked entities related to the entity. The work in (Li et al., 2007) proposes a re-rank strategy to order results based on the category assignments of Wikipedia articles, improving performance on weak queries. The work in (Milne et al., 2007) presents a thesaurus-based query expansion method using Wikipedia. The method identifies relevant topics in a query by checking consecutive sequences of words in the query against a thesaurus derived from Wikipedia, and automatically selects concepts from the thesaurus to expand the query. The work in (Elsas et al., 2008) investigates link-based query expansion using Wikipedia. They run the original query on Wikipedia and use a phrase scoring function to extract anchor texts, next using them as expansion terms.

In this work we also use knowledge acquired from Wikipedia exploiting entity semantics for query expansion. The main differences between the aforementioned works and this work is that we extensively use information available in

infoboxes for detecting entity semantics, and we adapt previously proposed term-selection functions from the literature to deal properly with entities. Three examples of how we explore the information existing in infoboxes are: (i) we classify the most representative entity in a query considering infobox templates as classes; (ii) we use values of attributes obtained from the infobox of a Wikipedia article as entity descriptors; and (iii) we exploit entity relationships extracted from anchor text of infobox links to other Wikipedia articles. They form important sources of semantic knowledge that are used to rank terms to be used as expansion terms.

3 METHOD DESCRIPTION

3.1 Modeling Entities in Wikipedia

In our method, we consider Wikipedia as a repository of entities represented by *articles*. Given an entity e, we assume there is an article A_e describing e. We consider A_e an *entity page* and its title is used to name e. We also assume that A_e contains an infobox I_e composed by a set of attribute-value pairs $\{\langle a_1, V_1\rangle,\ldots,\langle a_m, V_m\rangle\}$, where each V_i is a set of values referred to as *properties*. There are cases in which properties are more than literal values, as they may also include a reference (i.e., a hyperlink) to other entity page. In such cases, properties are referred to as *relationships*.

Figure 1 presents an excerpt of the infobox found in the entity page entitled *Alan Turing*. This infobox contains, among others, an attribute named *Nationality*, which has a single property, *British*, as its value, and an attribute named *Fields*, whose values are the relationships *Mathematics*, *Cryptanalysis* and *Computer Science*.

Fig. 1. Example of a Wikipedia infobox.

Infoboxes are usually built based on *templates* that define a set of attributes to be used in articles of a certain category C. For instance, there are infobox templates for categories such as *Scientist*, *Philosophers* and *Countries*. In our method, we only consider entity pages that follow an infobox template and thus belong to the category related with this template. The infobox illustrated in Figure 1 follows the template of category *Scientist*. Consider $e_1, \ldots, e_k$ a set of entities. Each entity e_i: (i) corresponds to an entity page A_{e_i} and (ii) has an infobox that follows a template that refers to a category C. The set of entities $e_1, \ldots, e_k$ form a *class*. Thus, we assume that any entity e_i belongs to a single category or class, and we use both terms interchangeably. This view enable us to adopt Wikipedia as a source of entity semantics for query expansion, as described next.

3.2 Method Overview

Given a query q, our method first identifies a citation to an entity e in this query.Then, the method finds in Wikipedia an entity page A_e that describes e and determines a category C for e. Next, the method obtains from A_e a set of *entity descriptors*, which are terms that represent concepts strongly related to entity. As many of such descriptors are often available, our method selects only the k most representative of them to be used as query expansion terms. We claim that, if properly selected, entity descriptors can be effectively used as query expansion terms, improving search results. Thus, we experiment several different functions to perform query term selection. Finally, we add the selected query terms to the set of terms from the original query, producing an expanded query q'.

In the following, we first describe how our method detects entities in queries. Then, we introduce our three approaches to obtain entity descriptors. Finally, we discuss how our method selects query expansion terms from entity descriptors.

3.3 Recognizing Entities in Queries

Named entity recognition in queries has become an important problem in web search. Although there are very effective methods to accomplish this task (Guo et al., 2009; Du et al., 2010), for simplicity we use a method based on (Xu et al., 2008), since this issue is not central to this work. The method detects the most representative entity in a query, rather than extracting all possible entities. It starts by selecting an n-word candidate string (n consecutive words in the original query) and tests if there is an entity named with this candidate string. If no such a candidate string is found, it tries all $(n-1)$-word candidate strings in the original query, and so on. The process stops when all query terms are covered by entities, or it reaches 1-word candidates. A term might be associated with more than one entity, what characterizes an ambiguity problem. Our method solves this problem differently from (Xu et al., 2008). Instead of choosing the first entity referred in the disambiguation page, it chooses the most cited entity, i.e., the one with the larger number of inlinks from the list of entities referred in the disambiguation page.

3.4 Obtaining Entity Descriptors

In this section we exploit the semantics implied by the structure of Wikipedia articles. We propose the following three approaches to obtain entity descriptors from page A_e related to an entity e.

1. *Content-based (CB)*: entity descriptors are obtained from the unstructured textual description available of the entity page. This approach simply extracts terms from the page content and applies standard procedures for stemming and stop-word removal.
2. *Property-based (PB)*: entity descriptors are obtained only from properties (i.e., literals in attribute values) that occur in the entity page infobox. This approach takes advantage of the richer semantics implicitly provided by: (i) human editors, which selected the more important terms about an entity to compose the infobox; and (ii) the community of users which defines, by means of the infobox template, what information is important to describe entities of a certain class.
3. *Relationship-based (RB)*: entity descriptors are obtained from relationships (i.e., references in attribute values) that occur in the entity page infobox. This approach refines the property-based approach by taking only the names used by human editors to refer to other entities. These names are obtained from the anchor text of hyperlinks to other entities pages.

The content-based and the property-based approaches may generate a large number of entity descriptors. However, not all of them are equally relevant as query expansion terms. Next we describe how our method addresses this issue.

3.5 Selecting Expansion Terms

Consider $t_1, \ldots, t_n$ a set of entity descriptors obtained for entity e_j using one of the approaches described previously. Consider $w_{i,j}$ the weight that estimates the relevance of an entity descriptor t_i to an entity e_j. Our method uses $w_{i,j}$ to rank the entity descriptor t_i considering the category C to which e_j belongs to. Next, the top k entity descriptors are selected to be used as expansion terms.

To compute the weights, we experiment with a number of information theoretic and information retrieval measures we adapt from the literature (Baeza-Yates and Ribeiro-Neto, 2011; Carpineto et al., 2001; Croft et al., 2009). Notice that our method uses infoboxes to associate each entity e_j with a category C. We use such a valuable knowledge to adapt these well-known measures to deal properly with entities as described below. For the discussion that follows, consider the following notation:

- E is the set of all entities that are associated to a category;
- E_C is the set of entities that belong to the same category C as e_j;
- E_i is the set of entities for which t_i is an entity descriptor;

– $E_{i,C} = E_i \cap E_C$ is the set of entities that belong to category C and have t_i as entity descriptor;
– $P(t_i) = |E_i|/|E|$ is the probability of a given entity descriptor t_i in E; and
– $P(t_i|C) = |E_{i,C}|/|E_C|$ is the probability of entity descriptor t_i given C.

We now define the following score functions to compute $w_{i,j}$:

Pearson's CHI-squared (CHI2) Measures the relationship between an expected frequency in the general population and an observed frequency. In our case, the expected frequency is the term likelihood $P(t_i)$, and the observed frequency is the probability of t_i given category C, i.e., $P(t_i|C)$. It is given by $w_{i,j} = \frac{(P(t_i|C)-P(t_i))^2}{P(t_i)}$.

Dice's Coefficient (DICE) Measures the similarity between two sets. We use it to evaluate the similarity between E_i and E_C. If these two sets are similar, then t_i is considered closely related to C. We define it as $w_{i,j} = 2 \times \frac{|E_{i,C}|}{|E_i|+|E_C|}$.

Inverted Document Frequency (IDF) Commonly used in information retrieval to measure the discriminative power of a term in a set of documents. In our case, we measure how good is t_i to discriminate the entity e_j in C as $w_{i,j} = \log \frac{|E_C|}{|E_{i,C}|}$.

Kullback-Lieber Divergence (KLD) Also known as *relative entropy* or *information gain*. It estimates the difference between two probability mass functions $p(x)$ and $q(x)$, i.e., the distance between two probability distributions. In our case, $p(x)$ represents the observations of $P(t_i|C)$, and $q(x)$ is $P(t_i)$. Thus, we define it as $w_{i,j} = P(t_i|C) \times \log \frac{P(t_i|C)}{P(t_i)}$.

Mutual Information (MI) Also known as *transinformation*. Measures the mutual dependence of the two random variables X and Y, i.e., the information that X and Y share. In our case, we consider that X is E_C and Y is E_i so the greater the amount of shared information between X and Y, the more t_i is closely related to C. We calculate it as $w_{i,j} = |E_{i,C}| \times \log(\frac{|E_{i,C}|}{|E_i| \times |E_C|})$.

To rank the entity descriptors t_i base on $w_{i,j}$ we use a simple term ranking procedure that disregards entity descriptors with score equal to zero and order the others in descending order of their scores. Besides the previous five score functions, we also use two strategies to combine them, based on the ranks they generate.

Borda Count[1] *(BC)* We score each entity descriptor using its position on each ranking. For instance, considering that we have 5 different rankings, each one with k different entity descriptors. Let $p_{i,r}$ be the position of the entity descriptor t_i in the rank r, so that if t_i is not present in a ranking r, $p_{i,r} = k + 1$. We define $score(i) = \sum_{j=1}^{5} p_{i,r}$. Next, we rank entities descriptors in the ascending order of this score.

Ranking Frequency (RF) We score each entity descriptor that appears in any ranking using the ranking frequency, i.e., the number of different rankings in which the entity descriptor occurred. Next, we rank entity descriptors in the descending order of its scores.

[1] Borda count is a consensus-based electoral system where candidates are ranked in order of preference (Saari, 1999).

4 EXPERIMENTS

For all the experiments here reported, we used the English Wikipedia dump of 09/16/2010, downloaded from `http://download.wikipedia.org`. Our method only considers what we have defined as entity pages, i.e., pages containing an infobox that follow a template. Table 1 presents some descriptive statistics on the Wikipedia corpus we used.

Table 1. Wikipedia corpus we used

Wikipedia Dump File:	2010.09.16
# of Entities:	7,452,529
# of Infobox Templates:	3,115
# of Entities with Infobox:	1,436,054 (19.27%)

As the target to queries, we used the ClueWeb09 Category B standard TREC collection, a representative web text collection with 50,220,423 of documents which has been recently used to evaluate information retrieval methods. For the experiments, we generated queries from the title field of the TREC topics. However, we only found entity pages with an infobox for entities mentioned in 21 out of the 50 queries available. Thus, unless when explicitly noticed, we executed all experiments using this 21 queries. For each query submitted we retrieved 1000 documents. We choose a ClueWeb09 collection because it is recent (2009), much larger than previous collections used at TREC 8 and 9 web tracks, and it was crawled in a way similar to commercial search engines, what means it is more similar to real web collections (Koolen and Kamps, 2010).

We implemented our method on top of the Terrier IR Platform (Terrier IR Platform, 2010), which Provides support for indexing and querying large datasets. We used Terrier version 3.0, running on Java 1.6.0. The preprocessing of documents and queries included stemming with the Porter stemmer and stopwords removing.

4.1 Term Selection Results

We first evaluate the effectiveness of the term selection functions presented in Section 3.5. For this, we measure the quality of the answers obtained when using top-k terms selected by each term selection function, with k varying from 0 to 100. This procedure was executed for the content-based and the property-based approaches. Recall that the relationship-based approach generates only a few entity descriptors, so that there is no need to use term selection functions with it. The results in terms of MAP and P@10 are presented in Figure 2.

These graphs clearly show that the property-based approach leads to better entity descriptors than the content-based approach. Indeed, with both MAP and P@10, all term selection functions have a similar behavior when used with the property-based approach, with a slight advantage to DICE in all cases. In contrast, only some of the term selection functions achieve good results with the content-based approach, e.g., IDF, CHI-2, and MI. This occurs because the property-based approach successfully leverages the semantics of entities implicitly encoded in Wikipedia.

More importantly, in all cases the results with the property-based approach reach their best values with about top-40 terms and stay stable from this point. On the other hand, the best results with the content-based approach are reached, in this experiment, only with top-100 terms. This means that shorter queries generated with the property-based approach are as good as longer queries generated with the content-based approach. This is an important issue, since shorter queries are preferable in terms of performance.

4.2 Query Expansion Results

We now compare the quality of the search results achieved by our three proposed approaches for obtaining entity descriptors. Table 2 presents a direct comparison between the following configurations: (i) BM-25, that is, no query expansion; (ii) CB-40, the content-based approach with 40 query expansion terms selected with the MI function; (iii) CB-100, the content-based approach

Table 2. Comparison of Results

Method	**MAP**	**P@10**
BM-25	0.1370	0.2381
CB-40	0.1317 (-3.87%)	0.3524 (+48.00%)
CB-100	0.1631 (+19.05%)	0.4190 (+75.98%)
PB-40	0.1556 (+13.58%)	0.4238 (+77.99%)
RB	0.1571 (+14.67%)	0.3600 (+51.20%)

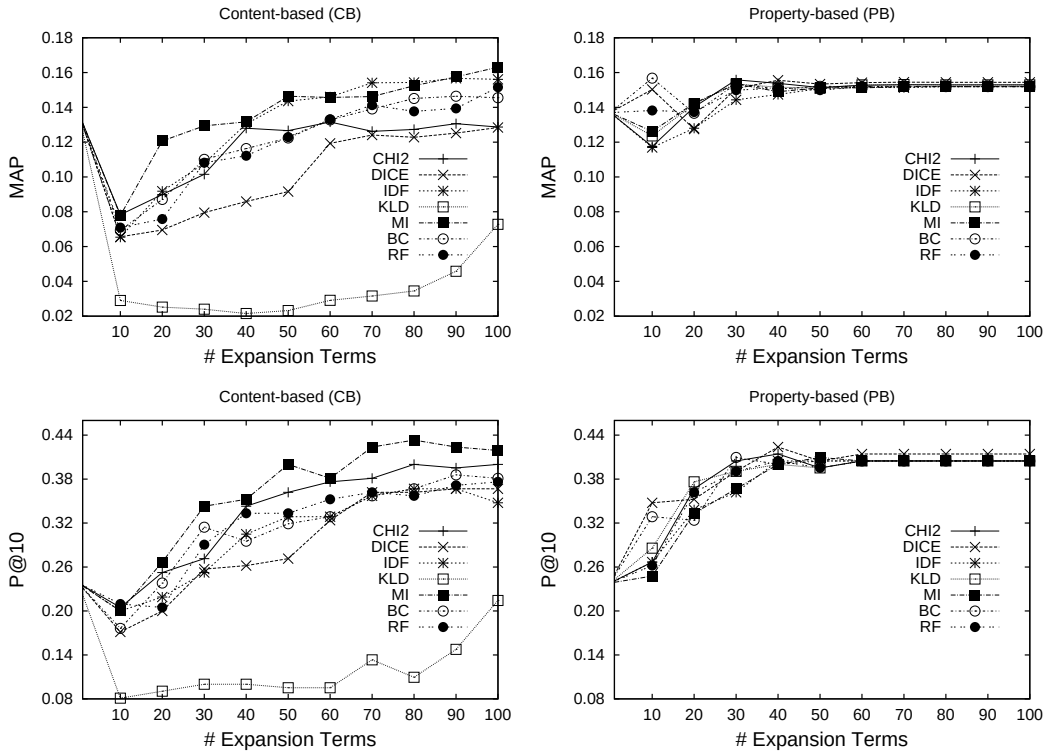

Fig. 2. Effectiveness of term selection functions.

with 100 query expansion terms, also selected with the MI function; (iv) PB-40, the property-based approach with 40 query expansion terms selected with the DICE function; (v) RB, the relationship-based approach with all entity descriptors being used as query expansion terms.

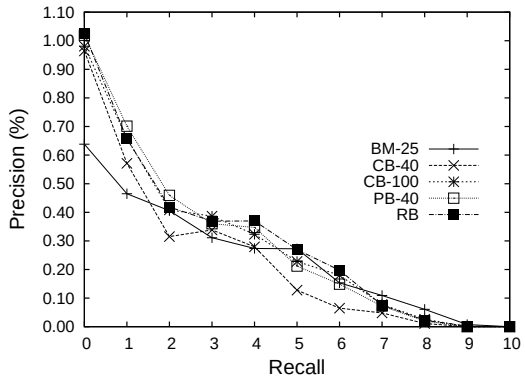

Fig. 3. Interpolated precision at 11 recall levels.

Figure 3 show precision/recall curves plotted with search results achieved when using these configurations. Notice that CB-100 and PB-40 are respectively the best configurations for the content-based and the property-based approaches in the results presented in Section 4.1. CB-40 configurations were included to allow comparing the content-based approach and the property-based approach when the same amount of expansion terms is used. As it can be noticed, all approaches yielded a clear improvement over BM-25. The RB approach achieved the best results, while CB-100 and PB-40 had a very similar performance. As expected, they were better than CB-40. It is worth mentioning that RB uses only about 18 expansion terms per query.

These results corroborates our conclusions that the property-based and relationship-based approaches, which take advantage of the implicit semantics of en-

tities provided by infoboxes, offer the best balance between quality of the results and amount of terms used for query expansion, since both PB-40 and RB achieved results comparable to or better than CB-100, but require much less additional terms.

5 CONCLUSIONS

We presented a new query expansion method that uses knowledge acquired from Wikipedia, exploiting the entity semantics available in its content and structure. We proposed three alternative approaches for obtaining terms related to Wikipedia entities. The first, we call content-based, uses terms available in the textual content of articles. The other two, called property-based and relationship-based, use, respectively, values and references to other entities, both found in infoboxes. Differently from methods previously presented in the literature, our method extensively uses information available in infoboxes. Specifically: (1) It uses the contents of infoboxes as a source for candidate query expansion terms. These contents are not only closely related to entities, but it is also continuously refined by human editors; (2) It takes advantage of the semantic structure implicitly provided by infoboxes templates to associate entities identified in queries with categories. This allowed adapting term-selection functions previously proposed, to deal properly with entities and improve their accuracy in selecting the best terms to be used for query expansion.

Experiments we carried out using ClueWeb09 Category B standard TREC collection have shown that all of our three approaches have a positive impact on the quality of the results, leading to gains of 19.05% (from 0.1370 to 0.1631) in terms of MAP and 77.99% (from 0.2381 to 0.4238) in terms of P@10. We have also found out that the *property-based* and *relationship-based* approaches present a better trade-off between quality of results and time required to process the expanded query.

REFERENCES

Baeza-Yates, R. and B. Ribeiro-Neto, 2011: *Modern Information Retrieval*. Addison-Wesley, 2nd edition.

Carpineto, C., R. De Mori, G. Romano, and B. Bigi, 2001: An Information-Theoretic Approach to Automatic Query Expansion. *ACM Transactions on Information Systems*, **19**(1), 1–27.

Croft, B., D. Metzeler, and T. Strohman, 2009: *Search Engines: Information Retrieval in Practice*. Addison-Wesley Publishing Company, USA, 552, 1st edition.

Du, J., Z. Zhang, J. Yan, Y. Cui, and Z. Chen, 2010: Using Search Session Context for Named Entity Recognition in Query. In *SIGIR'10*, 765–766.

Elsas, J. L., J. Arguello, J. Callan, and J. G. Carbonell, 2008: Retrieval and Feedback Models for Blog Feed Search. In *SIGIR'08*, 347–354.

Guo, J., G. Xu, X. Cheng, and H. Li, 2009: Named Entity Recognition in Query. In *SIGIR'09*, 267–274.

Koolen, M. and J. Kamps, 2010: The Importance of Anchor Text for Ad Hoc Search Revisited. In *SIGIR'10*, 122–129.

Li, Y., R. Luk, E. Ho, and F. Chung, 2007: Improving Weak Ad-Hoc Queries Using Wikipedia as External Corpus. In *SIGIR'07*, 797–798.

Maskey, S. and W. Dakka, 2009: Named Entity Network based on Wikipedia. In *INTERSPEECH'09*, 1515–1518.

Milne, D., I. H. Witten, and D. M. Nichols, 2007: A Knowledge-Based Search Engine Powered by Wikipedia. In *CIKM'07*, 445–454.

Saari, D. G., 1999: Explaining All Three-Alternative Voting Outcomes. *Journal of the American Society for Information Science*, **87**(2), 313–355.

Terrier IR Platform, 2010: . `http://terrier.org/`.

Xu, Y., F. Ding, and B. Wang, 2008: Entity-Based Query Reformulation Using Wikipedia. In *CIKM'08*, 1441–1442.

Xu, Y., G. J. Jones, and B. Wang, 2009: Query Dependent Pseudo-Relevance Feedback based on Wikipedia. In *SIGIR'09*, 59–66.