Explicit Extraction of Topical Context

Rianne Kaptein
Archives and Information Studies, Faculty of Humanities, University of Amsterdam, Turfdraagsterpad 9, 1012 XT Amsterdam, The Netherlands, E-mail: kaptein@uva.nl

Jaap Kamps
ISLA, Informatics Institute, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. E-mail: kamps@uva.nl

This article studies one of the main bottlenecks in providing more effective information access: the poverty on the query end. We explore whether users can classify keyword queries into categories from the DMOZ directory on different levels and whether this topical context can help retrieval performance. We have conducted a user study to let participants classify queries into DMOZ categories, either by freely searching the directory or by selection from a list of suggestions. Results of the study show that DMOZ categories are suitable for topic categorization. Both free search and list selection can be used to elicit topical context. Free search leads to more specific categories than the list selections. Participants in our study show moderate agreement on the categories they select, but broad agreement on the higher levels of chosen categories. The free search categories significantly improve retrieval effectiveness. The more general list selection categories and the top-level categories do not lead to significant improvements. Combining topical context with blind relevance feedback leads to better results than applying either of them separately. We conclude that DMOZ is a suitable resource for interacting with users on topical categories applicable to their query, and can lead to better search results.

Introduction

In this article, we study one of the main bottlenecks in providing more effective information access: the poverty on the query end. With an average query length of about two terms (Jansen, Spink, & Koshman, 2007; Jansen, Spink, & Saracevic, 2000; Lau & Horvitz, 1999), users provide only a highly ambiguous statement of the, often complex, underlying information need. This significantly restricts the ability of search engines to retrieve exactly those documents that are most relevant for the user's needs. To overcome this problem, we associate the query with topical context. If query topics can successfully be associated with topic categories, this topical context can be used in different ways, i.e., to improve retrieval effectiveness, to filter out results on nonrelevant topic categories or to cluster search results. In this article, we will investigate how to get and use topical context at different levels of granularity.

We make use of a web directory to obtain a hierarchy of topically organized web sites to use as a source of topical context. Two large web directories that have organized their information into hierarchical topical categories are DMOZ[1] and Yahoo! Directory.[2] In addition, Wikipedia[3] has an extensive category hierarchy to classify its articles. In the early days of the internet, web directories were used as a starting point for most activities. Nowadays, browsing in these types of directories is largely replaced by search. Yet, in China directories are still popular (Lee, 2008). There has been a stream of articles that use some form of topic model or context that use the DMOZ directory to represent categories (see Related Work). DMOZ has a lot of attractive features. It is hierarchical, large, and it covers a wide range of topics. The sites in the DMOZ directory are of high quality and selected by human editors, thus, providing us with potentially good feedback documents. A disadvantage of using a topic directory is that there is not an applicable topic category for every query. The DMOZ directory is very general, however, and if there is no topic category that applies to the query, there is usually a higher level category under which the query can be placed. Effectively communicating the category to the user is essential, and topical feedback using DMOZ categories by design generates clear intelligible labels (in contrast with, for example, clustering techniques such as described by Hearst &

Received December 9, 2010; revised March 21, 2011; accepted March 29, 2011

 DOI: 10.1002/asi.21563

[1] http://www.dmoz.org/
[2] http://dir.yahoo.com/
[3] http://www.wikipedia.org/

Pedersen, 1996). In this article, we therefore use the DMOZ directory to represent topical categories.

Queries can be associated with a topical category by using implicit or explicit techniques. Implicit techniques unobtrusively obtain information about users by watching their natural interactions with the system (Kelly & Teevan, 2003). Topical context can be elicited implicitly by using a user profile built on previous information seeking behavior, previously issued queries, selection and reading time of documents, etc. We elicit the context explicitly as a first step, i.e., ask the user to classify a query into a topical category. Eliciting the context implicitly is another challenge, which is only useful to explore once we can ascertain that topical context can indeed be used to improve retrieval effectiveness.

The DMOZ directory consists of hundreds of thousands categories; hence, for users it might not be so easy to find the DMOZ category that applies best to their query. There is a trade off between the specificity of the user categorization and the effort that is needed to select this category. Searching or browsing the complete directory requires the most effort from the user, but can result in finding more specific categories. Another option is to aid the user by a list of suggested categories. Choosing from a list of suggested categories requires less effort from the user, but there is a risk that the best possible category is not included in the list of suggestions.

Once the queries are associated with topical context, we experiment with using this topical context to improve retrieval results. We use the topical context in a similar way as relevance feedback, that is, we expand the query with terms from documents from the associated DMOZ category. We examine whether there is also a trade off between the level of categorization, and retrieval effectiveness when the topical context is used. We expect that low level and thus specific categories will prove most beneficial for retrieval effectiveness, because for low-level categories, the specificity of the category will be more similar to the specificity of the query than for high-level categories. The closer the topic of the query is to the topic of the category, the more likely the documents in this category will contain terms relevant to the query, and thus the more likely these are beneficial query expansion terms.

In this article, we address the following main research question:

- How can we explicitly extract and exploit topical context from the DMOZ directory?

This main research question consists of two parts, the first part deals with the extraction of topical context:

1. How well can users classify queries into the DMOZ categories?

   We conduct a user study to answer our first research question. We explore whether the DMOZ categories are representative for queries, that is, whether the DMOZ directory contains categories into which queries can be classified. The DMOZ directory contains a large number of categories, 590,000 in our test collection. This equals the amount of words in the Oxford English Dictionary (2011). Although we have to keep in mind that categories can be composed of multiple words, the amount of categories in DMOZ seems to be a promising repository to classify queries. Furthermore, we compare two different forms of extracting context explicitly, i.e., free search or browsing of the categories on the DMOZ site, and evaluation of categories from a list of suggestions.

   To answer the second part of our main research question, we use the results from our user study to look at the effects of using topical context on retrieval performance:
2. How can we use topical feedback to improve retrieval results?

   We compare the performance of runs using topical feedback in addition to the query. The topical feedback consists of categories on different levels in the DMOZ directory. In our work, topical feedback is feedback in the form of a (DMOZ) category and relevance feedback is feedback in the form of a document. We focus on the use of explicit topical feedback, i.e., a user has explicitly marked a category as relevant, and implicit or blind relevance feedback, i.e., it is assumed that top-ranked documents are relevant to the query. We do not study implicit topical feedback in this article.

   A question that arises when applying feedback techniques is how they relate to blind as wells as true relevance feedback, the most common use of feedback. Our third research question therefore is:
3. Does topical feedback improve retrieval results obtained using relevance feedback?

   The remainder of this article is organized as follows. In the next section, we discuss related work. We describe the data in the third section, i.e., the queries, the test collection, and the DMOZ directory. In the section Models, we describe the language models that we are using for topic categorization and retrieval. In the next section, we discuss the user study we have conducted to categorize queries into DMOZ categories. We describe the retrieval experiments where we use the topical context elicited in our user study to improve retrieval effectiveness in the section Retrieval Using Topical Feedback. In the final section, we draw our conclusions.

Related Work

In this section, we discuss related work on relevance feedback and topical feedback, other sources of context including user profiles, cluster-based retrieval, and latent semantic analysis.

The most common form of exploiting query context is through relevance feedback. When relevance feedback is applied, documents that are considered relevant, either because the documents are top-ranked in the initial ranking, or because users marked them as relevant, are exploited in a second iteration of the retrieval process.

Relevance feedback has been around for a long time already. In the seventies, Rocchio (1971) first applied relevance feedback on a vector space retrieval model. This relevance feedback approach maximizes the difference between the average vector of the relevant documents and the average vector of the non-relevant documents by adding

DOI: 10.1002/asi

query terms and by the reweighing of query terms to reflect their utility in discriminating relevant from non-relevant documents. Some years later feedback methods based on the probabilistic feedback model were introduced. Probabilistic retrieval models rank documents in decreasing order of probabilities of relevance, where initial probabilities of relevance are estimated by a constant for the query terms for the relevant documents and by the probabilities of terms in the whole background collection for non-relevant documents. Relevance feedback is applied by substituting the initial estimated probabilities of terms by using the accumulated statistics relating to the relevance or non-relevance of previously retrieved items (Salton & Buckley, 1990).

A widely used relevance feedback model was introduced by Lavrenko and Croft (2001). This so-called relevance model provides a formal method to determine the probability $P(w|R)$ of observing a word w in the documents relevant to a particular query. They are using the top-ranked documents retrieved by the query as implicit feedback, but the same model can be used when explicit relevance judgments are available. The method is a massive query expansion technique where the original query is completely replaced with a distribution over the entire vocabulary of the feedback documents. An overview of relevance feedback techniques can be found in (Ruthven & Lalmas, 2003).

A problem with systems incorporating relevance feedback is that they generally do not give the user enough context on which to base their relevance decisions, e.g., how many documents should be marked as relevant, how relevant should a document be before being marked as relevant, and what does not relevant mean? Getting the user to provide explicit feedback is not easy, and making the process of assessing relevance more difficult may result in less interaction not more (Ruthven & Lalmas, 2003). Another factor that influences the interaction of the user with the system is the user's experience with searching in general, and the experience with the system at hand. More experienced users are more flexible and are more likely to use different search strategies according to the familiarity to the search topic (Hsieh-Yee, 1993).

Instead of using previously retrieved documents for feedback, we aim to use other sources of information that are topically related to the query. There is a range of studies that use topical context similar to our approach, i.e., by exploiting an external knowledge source to group topically related documents into categories and associate these categories with the query. Categories can be associated with queries explicitly by users, or implicitly by a query categorization method.

Wei and Croft (2007) manually assign DMOZ categories to queries according to some basic rules. A topic model is built from the documents in the selected category, and queries are smoothed with the topic model to build a modified query. A query likelihood model using this modified query does not outperform a relevance model using pseudo-relevance feedback. A combination of applying the relevance model for queries with low clarity scores, meaning clear queries, and the topic model smoothing otherwise, leads to minor improvements over the relevance model.

Ravindran and Gauch (2004) designed a conceptual search engine where users can input DMOZ categories as context for their search. Document scores for retrieval are a combination of the key word match and the category match. This improves the precision of the search results. Additionally, search results are pruned, i.e., documents that do not match any of the categories provided with the query are removed, leading to further significant improvements of the retrieval results.

Topical categories as a source of query context have also been used in TREC (Text REtrieval Conference) for ad hoc retrieval. The topics in TREC 1 and 2 include a topical domain in the query descriptions, which can be used as topical context. It has been shown that these topical domains can successfully be used as query context for ad hoc retrieval (Bai, Nie, Bouchard, & Cao, 2007). In this article, the automatic and the manual assignment of categories is compared. Category models are created by using the relevant documents or the top 100 documents retrieved for the in-category queries. The top terms in the category models are used to expand the query. Automatic query classification is done by calculating KL-divergence scores. Although the accuracy of the automatic query classification is low, the effectiveness of retrieval is only slightly lower than when the category is assigned manually. Both lead to significant improvements over a baseline that does not incorporate topical context.

Haveliwala (2002) considers two scenarios to assign categories to queries. In the first scenario, unigram language models are used to calculate the class probabilities given a query for each of the 16 top-level DMOZ categories. The three categories with the highest probabilities are selected to compute topic-sensitive PageRank scores. Offline a set of PageRank scores has been calculated for each page and each category. In the second scenario, context of the query is taken into account. For example, users can highlight a term in a web page, and invoke a search. The context, in this case the web page, is then used to determine the category. Instead of only the query terms, the terms of the whole page are used to rank the 16 top-level DMOZ categories. Two other sources of query context are also suggested. First, using the history of queries issued leading up to the current query. Second, if the user is browsing some sort of hierarchical directory, the current node in the directory that the user is browsing at can be used as context. Potential query-independent sources of context include browsing patterns, bookmarks, and e-mail archives.

Successful, domain-specific applications of exploiting topical context can be found in the social science and genomics domain. Meij, Trieschnigg, De Rijke, and Kraaij (2010) leverage document-level concept annotations for improving full-text retrieval using the Medical Subject Headings (MeSH) thesaurus to improve genomics information retrieval and annotations of the CLEF collections to improve results in the CLEF domain-specific track. The original query is translated into a conceptual representation by means of relevance feedback, which is subsequently used to expand

DOI: 10.1002/asi

the query. Trieschnigg, Pezik, Lee, De Jong, Kraaij, and Rebholz-Schuhmann (2009) automatically annotate queries with MeSH concepts. A K-Nearest Neighbor classifier classifies documents by looking at the manual classification of similar or neighboring documents. Combining the textual and conceptual information leads to significant improvements on the TREC Genomics test collection.

Besides topical context, other forms of context can be explored, e.g., entity-type information (Balog, Vries, Serdyukov, Thomas, & Westerveld, 2009; Demartini, Iofciu, & Vries, 2009), document-type information (Kim & Croft, 2010), genres of web pages or lexical context. Rosso (2008) explores user-based identification of web genres. He defines genre as: a document-type based on purpose, form, and context, e.g., genres can be resumes, scientific articles or tax income forms. In this study, users develop and agree upon a genre ontology or palette for the edu domain. Lexical context of query terms can, for example, be extracted from Wordnet Miller (1995), which contains all kind of lexical relations to terms such as synonyms, hyponyms, and antonyms. Voorhees (1994) finds query expansion by lexical-semantic relations provides the potential to improve short, imprecise queries, but on average little improvement is achieved.

Instead of using groups of documents that are topically related to the query as context, the context can also consist of documents that are associated with a user. In this case, a user profile independent of the query is created and used at retrieval time to personalize and improve the retrieval results. These user profiles can be built in different ways, e.g., by monitoring the user's search behavior or by asking the user for explicit feedback. When explicit feedback is requested from the user, topical categories from web directories such as DMOZ can be used to represent the user's search profile. Chirita, Nejdl, Paiu, & Kohlshuetter (2005) let users pick multiple DMOZ categories to create user profiles that fit their interests. At run-time the output of a search engine is reranked by considering the distance between a user profile and the sets of DMOZ categories covered by each URL returned in the regular web search. Trajkova and Gauch (2004) build user profiles implicitly based on the user's search history. Web pages that a user has visited for at least a minimum amount of time are classified into a category from the top three levels of the DMOZ directory by using the highest weighted 20 words are to represent the content of the web page.

Liu, Yu, and Meng (2002) combine user profiles with query specific profiles to map a user query to a set of categories. User profiles are created automatically by using the search history, which consists of the issued queries, relevant documents, and related categories. A new incoming query is mapped to a set of categories using the user profile, the query specific profile, or a combination of both. Categories from DMOZ are ranked, and the top three categories are shown to the user who can select the category that best fits his search intention. Although this work provides a promising method to determine the categories associated with a query for a specific user, no method to exploit this information to improve the search results is suggested.

Another area of related work does not use an external knowledge source to identify groups of topically related documents. Instead, groups of topically related documents or terms to the query are identified implicitly by using search log and click data, by using the document collection at hand, the so-called cluster-based retrieval, or by latent semantic analysis.

An example of the use of search logs for topical search can be found in (Sondhi, Chandrasekar, & Rounthwaite, 2010). Contextual key words derived from topic-specific query logs are added to the initial query and submitted to a standard search engine. The altered queries help to focus the search engines results to the specific topic of interest. Cluster-based retrieval is a retrieval method inspired by the cluster hypothesis: "closely associated documents tend to be relevant to the same requests" (Van Rijsbergen, 1979). Documents are grouped into clusters, which can be used in different ways during the retrieval stage, i.e., clusters can be returned in their entirety in response to a query, or they can be used as a form of document smoothing. Document clustering can be performed online at retrieval time, depending on the query, which can be expensive, or offline and query independent, which may be based on factors irrelevant to the user information need (Liu & Croft, 2004). Effectively communicating the category to the user is essential in user interaction. In contrast with clustering techniques, our topical feedback method will by design generate clear intelligible labels, because we use the DMOZ category labels.

A more mathematical approach using topic models is latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Latent semantic indexing uses linear algebra techniques to detect conceptual relations in a document collection. An underlying or latent structure is assumed in the document-term matrix. This latent semantic structure is modeled on the basis of topics rather than individual terms. The result is a much smaller representation space, which can retrieve documents that share no words with the query. Two more latent topic models have since been developed, both applicable retrieval tasks. Hofmann (1999) introduced probabilistic latent semantic indexing, which is based on the likelihood principle and defines a generative model of the data. Each document is modelled as a mixture of topics. Latent Dirichlet allocation (Blei, Ng, Jordan, & Lafferty, 2003) is similar to probabilistic latent semantic indexing, but the topic distribution is assumed to have a Dirichlet prior. Latent Dirichlet allocation does not outperform a relevance model using pseudo-relevance feedback, but it can be calculated offline, which could be an advantage for some applications (Wei & Croft, 2006). Azzopardi, Girolami, and Rijsbergen (2004) use a document-specific term prior based on inferred topics induced from the corpus using LDA. The method achieves results comparable to the standard models, but when combined in a two-stage language model, it outperforms all other estimated models.

DOI: 10.1002/asi

<topic>
<num>Number: 701

<title>
U.S. oil industry history

<desc>Description:
Describe the history of the U.S. oil industry

<narr>Narrative:
Relevant documents will include those on historical exploration and drilling as well as history of regulatory bodies. Relevant are history of the oil industry in various states, even if drilling began in 1950 or later.

</topic>

FIG. 1. TREC ad hoc query topic 701.

Comparing our work to the related work described in this section, our contributions are:

- We conduct a user study to have participants explicitly assign DMOZ categories to queries shedding light on the (im)possibility of using topical context.
- Our approach is tested on a larger test collection with a larger number of queries than in the previous work. All previous works use either small document collections or a small number of queries created by the authors, which lead to questionable results and also avoid issues with efficiency.
- Most related work does not take into account the relation of topical feedback to relevance feedback. We take this into account and can therefore measure the additional value of topical feedback.

Data

In this article, we investigate whether we can use the DMOZ directory as a source of topical context. We use topics from the TREC 2008 Terabyte and Relevance Feedback tracks as test data (TREC, 2011). The TREC Terabyte track ran for 3 years, and provides us with 150 ad hoc topics that consist of three components, i.e., title, description, and narrative. The title field contains a key word query, similar to a query that might be entered into a web search engine. The description is a complete sentence or question describing the topic. The narrative gives a paragraph of information about which documents are considered relevant and/or irrelevant. An example query topic is shown in Figure 1. To retrieve documents, we will only use the title part of the query and not the description and the narrative. The relevance feedback track reuses topics from the terabyte track, but adds sets of known relevant and non-relevant documents to the topics that can be used for feedback.

Common query topics are on health, animals, and education. Some of the topics request information about specific U.S. or U.S. government matters. From our test collection, we remove the topics that are too specific for the U.S. and will be difficult to understand for the non-American participants in our user study. NIST (National Institute of Standards and Technology) assessors create the topics and judge the relevancy of documents in the test collection. They have good knowledge about the test collection and the US government in general.

The DMOZ directory is organized as a tree, where the topic categories are inner nodes and pages are leaf nodes. An example of a typical page in DMOZ can be found in Figure 2. As you can see the page for the category, Amsterdam contains a number of links to subcategories, as well as two links to pages about Amsterdam. Nodes cannot only have multiple child nodes, but by using symbolic links, nodes can appear to have several parent nodes as well. Since the DMOZ directory is free and open, everybody can contribute or re-use the data set, which is available in RDF. Google for example uses DMOZ as basis for its Google Directory service (Chirita et al., 2005).

At the time of writing, the complete DMOZ directory contains one million categories. At the time of our data dump in the beginning of 2008, it consisted of over 590,000 categories. The number of sites included in the directory is however stable at 4.8 million sites. In our experiments we exclude categories under the "World" category, because it contains categories in languages other than English. The number of categories and sites at different levels in the DMOZ directory is given in Table 1. For levels 1–4, these numbers are calculated using our test collection, for the complete directory (row "All") the numbers are taken from the DMOZ homepage.

We use the DMOZ corpus as the background collection for our language models. It consists of the raw text of all web pages up to level 4 we were able to crawl (459,907 out of 600,774). For efficiency reasons, all words that occur only once are excluded from the background corpus. The corpus consists of a total number of 350,041,078 words.

The web collection that is used to search relevant pages for these topics is the .GOV2 collection, a collection of web data crawled from web sites in the .gov domain during early 2004. Topics are only created if the .GOV2 collection contains relevant pages for the topic. The DMOZ directory is intended to cover the whole web, thereby also including the .gov domain. In total, 5,339 sites, i.e., around 1% of the sites in our test collection consisting of levels 1–4 of the DMOZ directory is from the .gov domain. Some of the DMOZ categories hardly contain any sites from the .gov domain, e.g., games, shopping,

DOI: 10.1002/asi

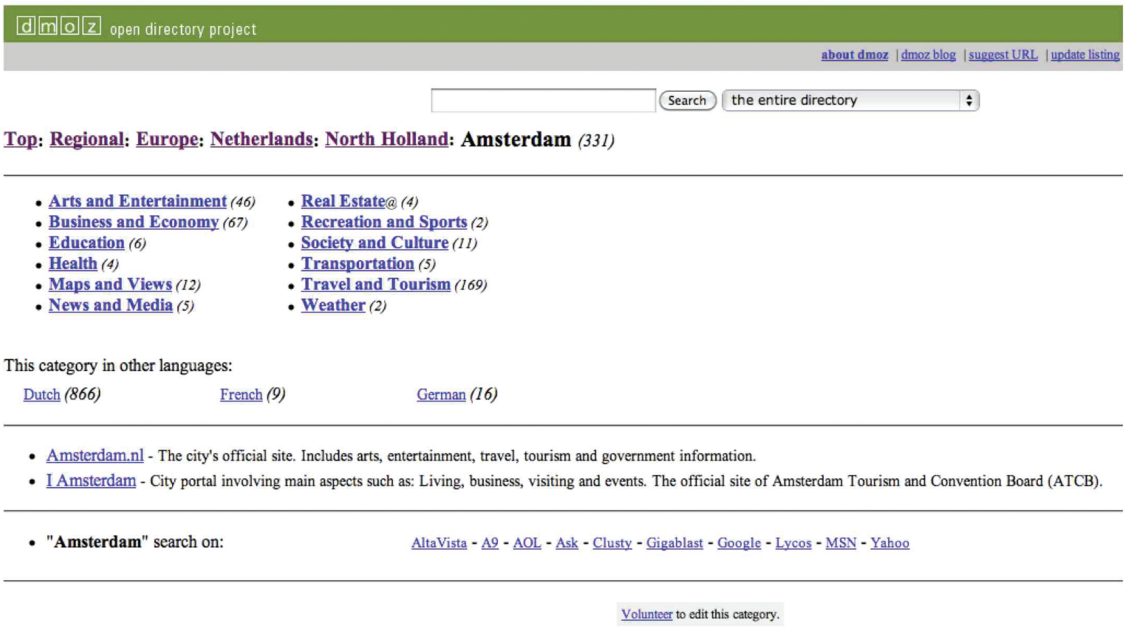

FIG. 2. Page of category 'Amsterdam' in the DMOZ directory. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 1. Size of our DMOZ test collection.

Level	# Categories	# Sites
1	15	86
2	574	6,776
3	6,501	128,379
4	29,777	379,619
All	over 590,000	4,830,584

and sports. The categories health, regional, and science contain the most sites from the .gov domain. We expect therefore that also most topics will be categorized into the categories health, regional, and science.

Models

Throughout this article, we use the language modeling approach for retrieval, feedback, query categorization, and other tasks. For an introduction of the language modelling approach, we refer to (Zhai, 2008). In this section, we explain how we use language models for query categorization to generate a list of suggested categories, and we describe the model we use to incorporate topical and relevance feedback in our retrieval model.

Query Categorization

In this section, we discuss three methods to generate a list of suggested categories for a query to display to the user. The first method we use to categorize the query is the simplest.

1. *Title match*: Match query words with the label of the DMOZ category.

When all query words are present in the category label, this category is assigned to the query. The label of the category consists of the whole path of categories in the hierarchy, e.g., "Regional: Europe: Netherlands: North Holland: Amsterdam." Not all words from this label have to be present in the query, e.g., the queries "Amsterdam" and "Amsterdam Netherlands" are matches to the given example category. When a category matches all query words, all its descendants automatically also match all query words, we then only assign the highest level matching category to the query, e.g., if the query is "Netherlands," only the category "Regional: Europe: Netherlands" is assigned to the query. Both the query words and the category labels are stemmed using a Porter stemmer (Porter, 1997).

The next two categorization methods use topic models of the DMOZ categories to generate a list of suggested categories. Categories are assigned to each query by using either the query title, or the top 10 retrieved documents. We first create topic models of the DMOZ categories. We start by crawling the sites from each category and of all its available direct sub categories. All HTML markup is stripped from the sites, since we are only interested in the textual content. Stopwords are removed according to a standard stopword list. Stemming is not applied. If at least 10 sites are found, a parsimonious language model of the category is created. A parsimonious language model concentrates the probability mass on fewer terms than a standard language model. Instead of blindly modeling language use in a (relevant) document, it models what language use distinguishes a document from other documents (Hiemstra, Robertson, & Zaragoza, 2004). For the parsimonious model, we have to set the parameters α and the threshold parameter. We set the threshold parameter at 0.0001, i.e., words that occur with a probability less than 0.0001 are removed from the index. We set $\alpha = 0.1$ for the parsimonious model, based on initial experiments with a part of the test collection.

DOI: 10.1002/asi

We create a topic model for a category from the concatenation of all textual content of the websites belonging to the category. The web sites used to create the topic model include the sites of the category as well as the sites in all its subcategories. To produce the list of suggestions, we focus on a part of the DMOZ directory in order to reduce complexity. That is, we use the categories from the first four levels of DMOZ, which comprise around 30,000 categories. Since we have crawled only the upper four levels of the DMOZ directory, we can create topic models up until the third level of the hierarchy using also the subcategories. The topic models on the fourth level are created using only the links on that level.

After the creation of the topic models for the categories, we can start assigning categories to queries as follows. Our second method for query categorization is based on classifying documents.

2. *Top ranking documents similarity*: We use the top 10 results of a baseline model run, and select categories whose topic model is most similar to these documents.

The documents are classified into a category as follows. First, the documents are scored on DMOZ top-level categories by scoring each of the top-level topic models on the documents:

$$P(TM|D_{top}) = \sum_{d \in D_{top}} \prod_{t \in d} ((1-\lambda) P(t|TM) + \lambda P(t|C))$$

where TM is a topic model, d is a document, D_{top} is the set of top retrieved documents, t is a term, and C is the background collection. The topic models are ranked by their probabilities and saved. The documents are then classified into the second-level categories. Similarly, the documents are classified into the third- and fourth-level categories, but for computational efficiency here only subcategories from the 20 highest ranked categories are used. When the topic models up to the fourth level have been estimated, all topic models are ranked according to their probabilities, where the highest ranked topic model is the most probable category associated with the query.

Our last method directly classifies the query.

3. *Query similarity*: We classify the query, that is the short topic statement in the title field Q, by selecting categories whose topic model is most similar to the query.

In this case, the top-level topic models are scored on the query.

$$P(TM|Q) = \prod_{t \in Q} ((1-\lambda) P(t|TM) + \lambda P(t|C))$$

Again the topic models are ranked by their probabilities, and the process continues down the category hierarchy in the same way as the top 10 result classification.

To produce a list of suggestions for a topic, we merge the top 10 ranked categories from the three categorization methods. The list of suggestions is shorter than 30 categories, because some of the categories will be in the top 10 of more than one query categorization method, and the title match is not likely to generate more than one matching category.

Retrieval

For retrieval, we use the language modeling approach. We extend a baseline retrieval model to incorporate topical as well as relevance feedback.

Baseline retrieval model. Our baseline retrieval model is a standard language model. For retrieval we make use of Indri (Strohman, Metzler, Turtle, & Croft, 2005), an open-source search engine, which incorporates the language modeling approach. The baseline model uses Jelinek-Mercer smoothing to smooth the probability of a query term occurring in a document with the probability of the query term occurring in the background corpus as follows:

$$P(Q|D) = \prod_{t \in Q} (1-\lambda) P(t|D) + \lambda P(t|C)$$

where Q is the query, D the document, and C the background collection.

The standard value of the smoothing parameter λ in the language model is 0.85. From the TREC Terabyte tracks, however, it is known that the .GOV2 collection requires little smoothing, i.e., a value of 0.1 for λ gives the best results (Kamps, 2006).

Topical feedback. To retrieve documents using topical feedback, the input is not only a query Q, but also a topic model TM of a category assigned to the query. The topic model for a category is created as described in the section Query Categorization. To produce a ranking, a mixture of the query model and the topic model is calculated as follows:

$$P(Q, TM|D) = (1-\beta) P(Q|D) + \beta P(TM|D)$$

β determines the weight of the topic model. $P(TM|D)$ is estimated similar to $P(Q|D)$ as described before:

$$P(TM|D) = \prod_{t \in TM} ((1-\lambda) P(t|D) + \lambda P(t|C))$$

For efficiency reasons, we rerank the top 1,000 results retrieved by the baseline retrieval model. To estimate $P(t|D)$, we use a parsimonious model with the same parameter settings as used for the query categorization in the previous section.

Weighted topic query expansion. A general problem of feedback approaches is that they work very well for some queries, and that they degrade the results for other queries. In our experiments, we analyze the performance of all approaches on individual queries. To tackle this problem, we experiment with an alternative query expansion method, we call weighted topic query expansion. This method reweighs the original query terms according to the inverse fraction of query terms

DOI: 10.1002/asi

that occur in the category title. If the query terms are equal to the category title, this topic model is a good match for the query, hence, the weight of the topic model terms can be high. On the other hand, if none of the query terms occur in the category title, it is unlikely that the topical feedback will contribute to retrieval performance, hence, the weight of the topical feedback is lowered. The original weights of the query words are $1/|Q|$, the adjusted weights of the query terms are $1/(|Q|{*}\textit{fraction of query terms in category title})$. A fraction of 1/5 is used when none of the query terms occur in the category title. Since we do not want to divide by zero, and the large majority of queries consists of less than five query terms, this is an approximate lower bound on the range of fractions.

Relevance feedback. Besides topical feedback, we also apply the more standard relevance feedback, instead of a topic model of a category, a model of (pseudo)relevant documents to the query is used. Relevance feedback is applied using an adaptation of the relevance model of Lavrenko and Croft (2001). Their relevance model provides a formal method to determine the probability $P(w|R)$ of observing a word w in the documents relevant to a particular query. The method is a query expansion technique where the original query is completely replaced with a distribution over the entire vocabulary of the relevant feedback documents. Instead of completely replacing the original query, we include the original query with a weight W_{orig} in the expanded query. Our relevance feedback approach only uses positive relevance feedback. The approach is similar to the implementation of pseudo-relevance feedback in Indri, and takes the following steps:

1. $P(t|R)$ is estimated using the given relevant documents either using maximum likelihood estimation, or using a parsimonious model.
2. Terms $P(t|R)$ are sorted. All terms in the parsimonious model are kept, but in case of MLE only the 50 top-ranked terms are kept.
3. In the original baseline query Q_{orig}, each query term gets an equal weight of $1/|Q|$. The relevance feedback part, Q_R, of the expanded query is constructed as:

$$\#weight(P(t_i|R)\ t_i \ldots P(t_n|R)\ t_n)$$

4. The fully expanded Indri query is now constructed as:

$$\#weight(W_{orig}\ Q_{orig}\ (1 - W_{orig})\ Q_R)$$

5. Documents are retrieved based on the expanded query.

Adjusting the query is a simple and efficient way to implement parsimonious relevance feedback. When MLE is used to estimate $P(t|R)$, our feedback approach is equal to the feedback approach implemented in Indri. When pseudo-relevance feedback, also known as blind relevance feedback, is applied, we use the top 10 documents of the initial ranking for feedback.

Categorizing Queries

In this section, we describe the user study we conducted to let the participants assign topic categories to query topics.

User Study Setup

The user study is designed as follows. Participants first read an instruction, and do a training task. Before starting the actual tasks, participants fill out a pre-experiment questionnaire that consists of some demographic questions. The main part of the study consists of 15 tasks. Each task corresponds to one query like the example query shown in Figure 1. No specific knowledge is needed to understand the queries.

The queries in the user study are taken from the three TREC Terabyte tracks 2004, 2005, and 2006 (.GOV2 collection of 25 M documents) (TREC, 2011). Queries from topics 801–850 are categorized and evaluated by two to four participants, all other queries are covered by one participant. In total 135 out of the 150 Terabyte queries are covered. The order and the selection of queries are randomized.

At the beginning of each task the query, consisting of query title, description, and narrative, is given. Each task is then divided into four subtasks:

1. Pre-task questions
2. The evaluation of a list of suggested categories.

   In subtask 2 the participant evaluates a list of suggested categories. The list of suggestions is composed of the categories resulting from the three query categorization methods described in the section Query Categorization. For each suggestion the participant evaluates how relevant the category is to the query by answering the question: "For each suggested category evaluate how relevant it is to the query". The four options are: "Not at all", "Relevant, but too broad", "Relevant, but too specific", and "Excellent". We do not define the define the concepts of relevancy, but leave the interpretation to the user.
3. Search or browse on the DMOZ site to find the best category.

   In subtask 3 the participant is free to select a category from the DMOZ site that he or she thinks applies best to the query. Categories can be found by browsing the DMOZ site or by using the search function on the DMOZ site. Besides the category label the participants can use the information available on the DMOZ pages to determine the relevancy of the category such as a description of the category, the sites belonging to the category, related categories, and subcategories. If the participant finds more than one category that applies best to the query, there is a possibility to add a second DMOZ category. Also in this subtask the participant evaluates the relevance of the selected category to the query.
4. Post-task questions.

In the second and third task, some questions are also asked on how easy the task was, and how confident the participants are about their categorization. After the 15 tasks each participant fills out a post-experiment questionnaire that consists of questions on how they experienced and liked the different

TABLE 2. Coverage of queries.

	Not available	Not relevant (%)	Too broad (%)	Excellent (%)	Too specific (%)
Free search	—	1.5	9.0	54.1	35.3
Categorization Method					
Title match	89.6%	0.0	0.0	8.9	1.5
Top docs sim.	0.0%	11.1	60.7	12.6	15.6
Query sim.	0.0%	14.1	45.2	25.2	15.6
All suggestions	0.0%	1.5	45.2	35.6	17.8

TABLE 3. Evaluations of list of suggested categories.

Categorization method	Not relevant (%)	Too broad (%)	Too specific (%)	Excellent (%)
Title match	17.9	17.9	21.4	42.9
Top docs sim.	77.2	19.8	1.9	1.1
Query sim.	78.7	15.8	3.6	2.0
All suggestions	80.1	15.8	2.6	1.6

tasks. At each stage of the user study, there are open questions for comments of any kind.

We do not rotate subtask 2 and 3 because our goal is to obtain good human feedback. Seeing the list of suggestions first means that there is a learning effect which can improve the quality of the categories selected in the free search.

The online user study records all answers, and also the time it takes for participants to do the different tasks. The open text answers, i.e., copying the URL from the DMOZ site, are manually preprocessed before the analysis to ensure that they are all in the same format.

User Study Results

In this section, we discuss and analyze the results of the user study.

Demographics. The user study has been filled out by 14 participants, of which nine are male and five female. Two participants participated twice in the user study; hence, they did 30 instead of 15 queries. The majority of the participants are studying or working within the field of information retrieval. The average age is 31 years. Half of them are familiar with the DMOZ directory, and three quarters of them are familiar with the subject of topic categorization. All of them are near-native speakers of English.

Query categorization statistics. We first look at the question: does an appropriate DMOZ category exist for the queries?

In Table 2, we present the coverage of the queries. To determine the coverage of a query for the query categorization methods, we take only the best evaluation per query, e.g., if one category from the list of suggested categories is evaluated as "Excellent" by a participant in the study, the query is counted as an excellent match. This percentage is therefore an upper bound on the coverage of the queries. When free search is used, only for 1.5% of the queries is no relevant category found. For more than half of the queries (54.1%), an excellent matching category is found. In the retrieval experiments described in the next section, we check whether the categories perceived as excellent by the participants are also excellent in terms of system performance.

When the list of suggestions is used, only for 1.5% of the queries is no relevant DMOZ category found. When the category is relevant, it is usually too broad (45.2% of the topics). Still, for 35.6% of the queries an excellent matching category is found. The query similarity categorizations provide better suggestions than the categorizations based on top-ranking documents similarity. Using the query leads to more focused categorizations, whereas using the top-ranking documents results in some topic drift leading to more "Too broad" evaluations. Using the title match method does not lead to any suggested categories for 110 out of the 135 queries (81.5%), but when a category is found, this is an excellent category in the majority of the cases.

Besides looking at the best evaluation per query, we look at all evaluations of suggested categories in Table 3. In this table, we take into account each evaluation from all participants in the user study. Keep in mind that the title match categorization method only provides a small number of suggested categories. We see here that the large majority (80%) of categories on the list of suggested categories are not relevant. Only 1.6% of all suggested categories is evaluated as excellent. Fortunately, these excellent categories are spread over a large number of queries, that is, we saw in Table 2 that an excellent category is found for 35.6% of the queries.

Next, we look at the question: what is the level in the DMOZ hierarchy where the most suitable DMOZ categories reside? With free search the participants can select a category on any level of the DMOZ directory. Figure 3 shows the distribution of categories over the level of the DMOZ hierarchy. We see that the deepest level chosen is 11, the median level is 5. Levels 1 and 2, which are often used in systems to reduce

DOI: 10.1002/asi

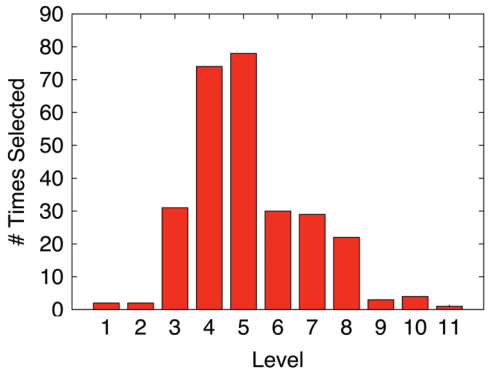

FIG. 3. Levels of DMOZ categories selected by free search. [Color figure can be viewed in the online issue, which is available at wileyonline library.com.]

TABLE 4. Free search versus suggestions list results.

	Free Search		Suggestions	
	Avg.	Post exp.	Avg.	Post exp.
Time in min.	2.0		1.3	
Speed		3.5		3.5
Confident	3.5	3.4	3.5	3.4
Easy	3.0	3.2	3.2	3.5

the complexity, are hardly ever selected. Our query categorization methods based on similarity of the documents in the category and either the query or the top-ranked documents generate categories up to level 4 in the hierarchy, thereby still missing out of a large number of relevant categories.

Participants preferences. We now turn to compare the preferences of the participants of the two ways of eliciting explicit category feedback: either by evaluating a list of suggestions, or by freely searching the DMOZ hierarchy.

Table 4 compares free search with the evaluation of the suggestions on different variables. Variables "Quick" (I directly found the selected category(ies), and did not browse in several categories), "Confident" (I am confident I selected the best possible category(ies)), and "Easy" (It was easy to select categories) are measured on a Likert-scale from 1–5, where 1 means "Strongly Disagree" and 5 means "Strongly Agree". Averages are calculated over all participants and all queries. The post experiment numbers in the second and fourth column are averages over all participants on answers in the post-experiment questionnaire.

When comparing free search with the evaluation of suggested categories, we have to consider a bias that occurs because the participants evaluate the list of suggested categories first and then do the free search. In close to 50% of the cases, the participants say that the list of suggestions helped them to select a category from the DMOZ site using free search. In 55% of the cases, the participants think that the category they selected freely from the DMOZ site is better than all the suggestions in the list.

How easy and how efficient are both methods of eliciting explicit topical context? The average time spent per query for the free search is significantly higher than the average time spent for the evaluation of the suggested categories (2.0 and 1.3 min, respectively). The participants however perceive both methods to be quick. The confidence in their classifications is the same on average, and in the final evaluation for both methods. The participants find the evaluation of the list of suggested categories slightly easier than the free search.

When asked which method the participants prefer, the replies are mixed. Three participants prefer free search, four participants prefer evaluation of a list of suggested categories, and seven participants prefer to look at a list of suggested categories, and then search freely on the DMOZ site.

Agreement between participants. We now look at the agreement between different participants categorizing the same query. Although it is shown that people do not agree much on tasks like this (Furnas, Landauer, Gomez, & Dumais, 1987; Saracevic & Kantor, 1988), we can still assume that the easier the task, the higher the agreement between participants will be. In addition, agreement should not be considered as an indication of the quality of the category assignment. We calculate pairwise agreement between participants. Strict agreement means that there is agreement on the relevant categories, and on the degree of relevance ("Relevant, but too broad," "Relevant, but too specific," and "Excellent"). Weak agreement means that there is agreement on the relevant categories, but the degree of relevance is not taken into account. Categories that are evaluated as not relevant by all participants are not included.

For the list of suggested categories, two types of agreements are calculated. "All evaluations" calculates agreement for each category on the suggestions list when at least one participant considers the category relevant. "Best match" only calculates agreement for the category of the list of suggested categories with the best agreement, i.e., there is an overlap between the categories evaluated as relevant for a query by two participants. Similarly, when free search is used, and two categories are selected, only the best matching categories are used to calculate agreement. For the majority of cases, participants select only one category in the free search; therefore, we omit the calculation of all evaluations of the free search. The results are presented in Table 5.

Strict agreement for all evaluations of the list of suggested categories is low (0.14), and is comparable to strict agreement for the best matching categories selected using free search, which has an agreement of 0.15. Agreement on the best matching categories from the list of suggested categories is high, i.e., a strict agreement of 0.61. This means that for most queries, the participants agree on at least one relevant category. This relevant category will be used in our retrieval experiments that follow. Categories selected by free search receive somewhat higher weak agreement than all evaluations of the list of suggested categories, 0.20 and 0.34, respectively.

DOI: 10.1002/asi

TABLE 5. Strict and weak agreement between participants over all relevant judgments, and over best matching relevant judgements.

	# Queries	Strict agr.	Weak agr.
All evaluations			
Title match	6	0.69	0.89
Top docs sim.	49	0.14	0.18
Query sim.	44	0.12	0.22
List of suggested categories	50	0.14	0.20
Best match			
List of suggested categories	50	0.61	0.75
Free search	50	0.15	0.34

TABLE 6. Weak agreement on different levels between participants over best matching relevant judgements.

	List of suggested categories		Free search	
	# Queries	Weak agr.	# Queries	Weak agr.
Level 1	50	0.75	50	0.74
Level 2	50	0.73	50	0.64
Level 3	48	0.67	50	0.58
Level 4	37	0.48	50	0.50
Complete	50	0.75	50	0.34

What is the difference in agreement over the different category suggestion methods? From the three methods used to produce categories for the list of suggestions, the query title match produces the categories that best cover the query, and that receives the most agreement. The drawback of this method is that only for a small percentage of queries (10.4%), there is an exact match with a DMOZ category label. Expanding this method to include nearly exact matches could be beneficial. Differences between the top docs similarity method and the query similarity method are small.

We also calculate agreement over best-matching categories on different levels, e.g., agreement on level 1 means that the categories have the same top-level category. The results are presented in Table 6. The "Complete" row gives agreement on the complete categories.

A problem in DMOZ is that category names are ambiguous when the full path in the category hierarchy is not taken into account. For example, in DMOZ there are four fruit categories in different places in the directory: ("Shopping: Home and Garden: Plants: Fruit," "Home: Gardening: Plants: Fruit," "Science: Agriculture: Horticulture: Fruits," and "Shopping: Food: Produce: Fruit").

On the positive side, every chosen category in the DMOZ hierarchy is subcategory of a whole path up to the root node. Hence, different categories may still share the same top-level categories. What is the agreement over the levels of the DMOZ hierarchy? We look here at the best-matching relevant category only. For the free search, agreement on levels 1–4 of the DMOZ directory is much higher, from an agreement of 0.74 on the first level, to an agreement of 0.50 on the fourth level. For the list selection, the agreement for the best-matching relevant category is very similar with 0.75 at the top-level, and 0.48 at level 4.

Discussion

We conducted this user study to answer our first research question: *How well can users categorize queries into DMOZ categories?* We conclude that the DMOZ directory can be considered suitable to categorize queries into categories. Using either free search or the suggestions list for 98.5% of the queries, a relevant DMOZ category is found. This category can however be too broad or too specific. When participants evaluate categories from a list of suggestions, only 19.9% of the categories is evaluated to be relevant. The relevant categories are usually too broad. For many queries, the categories till level 4 of the DMOZ category are not specific enough to categorize queries appropriately, because when we look at the categories selected by the free search, in 61% of the cases, the selected category is at level 5 or deeper.

Considering the method used to elicit the topical context, there is no clear preference from the participants' point of view. In our setup, there is however a difference in the quality of the query categorization. The list of suggestions only retrieves categories until level 4, thereby excluding a large part of the DMOZ directory. When free search is used, most often a category on level 5 is selected. Extending the automatic categorization to produce suggestions to the fifth or a even deeper level thus has clear potential to improve the quality of the list of suggested categories. The participants in our user study now consider the evaluation of the suggested categories as easier, and faster. It would be interesting to see whether these advantages still hold when deeper level categories are also shown in the suggested categories list.

Looking at the different methods of automatic query categorization, the title match of the query words with DMOZ category labels produces high-quality suggestions, but not for many queries. Using a less stringent title match, where not all query words have to occur in the category title, could provide us with more possible relevant categories. The categories produced by the classification of the query differ substantially from the categories produced by the classification of the top 10 documents. Differences in the agreement and the coverage of queries are however still small. To make the list of suggestions, classification of the query, the top 10 retrieved documents, and the query title match can all produce different useful suggestions. We do not have to choose between these methods, since users can easily review the list of suggestions and make decisions on relevance.

What is the agreement on the relevance of DMOZ categories between different participants? Considering that the participants can choose from 590,000 categories, the weak agreement of 0.34 for the free search is quite good. For the list-based suggestions, the weak agreement over all categories deemed relevant by any of the participants is 0.20. A problem with the evaluation of the list of suggested categories is that some participants tend to select only one or two categories, whereas other participants evaluate substantially

DOI: 10.1002/asi

more categories as relevant, but too broad, leading to a lot of disagreement. That is, if we consider only the best-matching category assigned by both judges, the weak agreement is as high as 0.75.

Since best-matching categories can be deeply nested in DMOZ, getting the initial levels of these categories right can be very important. That is, each category also represents all their ancestors' categories in the DMOZ's hierarchy. Agreement on levels 1–4 of the directory is much better hence, at least participants start out on the same path to a category. Finally, they may select different categories at different levels of granularity. Overall, free search results in the best and most specific categories, considering agreement and coverage of the query. However, the categories in the list of suggested categories can still be improved by including more of the DMOZ hierarchy. From the participants point of view, there is no agreement on a preference for one of the methods. Hence, a good option will be to use a combination of both methods so that users can decide for themselves per query how they want to select a category.

Summarizing, from our user study we can conclude that for nearly all queries a relevant DMOZ category can be found. Categories selected in the free search are more specific than the categories from the list of suggestions. For the participants, there are no large differences between selecting categories from a list of suggestions and the free search considering speed, confidence, difficulty, and personal preference. Agreement between participants is moderate, but increases considerably when we look only at the top-level categories.

Retrieval using Topical Feedback

In this section, we report on our experiments that exploit the topical context as retrieved from our user study.

Experimental Setup

To test our topical feedback approach, we use Terabyte topics 800–850 that have been classified by at least two participants in our user study. All parameters for the topic models are the same as used in the user study. Only for retrieval, we do use a Porter stemmer, because our initial results indicate that stemming leads to better results. In some of our experiments, we also use a document length prior to favor longer documents. For parameter β, we try values from 0 to 1 with steps of 0.1. For computational efficiency, we rerank results. The run we are reranking is created by using a standard language model, with Jelinek-Mercer smoothing ($\lambda = 0.9$). We rerank the top 1,000 results.

From our user study, we extract query classifications on three levels. The deepest level topic models are based on the categories selected most frequently in the free search, hence, on any level in the directory (Free Search). The middle level consists of the categories selected most frequently from the suggested categories of levels 1–4 of the directory (Suggestions). We add a third classification on the top level, where one of the 13 top-level categories is picked. For the top-level category, we use the top category that occurs most frequently in the list of suggested categories (Top Level). When there is a tie between categories, we decide randomly.

TABLE 7. Retrieval results using topical context.

Topical Context	β	MAP	P10
Baseline	0.0	0.2932	0.5540
Top level	1.0	0.0928$^{\bullet}$	0.1000$^{\bullet}$
Suggestions	1.0	0.1388$^{\bullet}$	0.2160$^{\bullet}$
Free search	1.0	0.2179*	0.3640*
Top level	0.7	0.2937	0.5700
Suggestions	0.6	0.2984	0.5720
Free search	0.6	**0.3238**$^{\bullet}$	**0.6140***

Note. Significance of increase or decrease over baseline according to t-test, one-tailed, at significance levels 0.05($^{\circ}$), 0.01(*), and 0.001($^{\bullet}$).

We want to know whether applying topical feedback can improve the results obtained with relevance feedback. We therefore compare the results of topical feedback with relevance feedback results, and combine topical feedback with relevance feedback to see if that leads to additional improvements. To compare topical feedback with relevance feedback, we use odd-numbered topics 800–850 from the terabyte track, which have been used as training data in the TREC relevance feedback track. Besides the standard topic query expansion (Topic QE), we also give the results of the weighted topic query expansion (W. Topic QE). To create a parsimonious topic model, we use a λ of 0.01, and a threshold of 0.001. When blind feedback is used, the top 50 terms from the top 10 documents are used. We also experiment with applying a document length prior.

Experimental Results

In this section, we describe our experimental results, which are split into two parts; first, we discuss the influence of the query categorization and second, the relation between topical feedback and relevance feedback.

Influence of query categorization. Table 7 shows the retrieval results. The baseline run does not use topical context. First, we look at how well the topical context captures the information need of the queries. As expected, when only the topical context is used ($\beta = 1.0$), the results are significantly worse than the baseline. The free search categories still perform quite reasonably, showing that the DMOZ categories can capture the information request at hand to some degree. Second, we look at combining the baseline run with the topical context. In the table, only the best runs are shown. We show MAP and P10 over different values of β in Figures 4 and 5. The results start degrading only at a high value of β at around 0.8 or 0.9, suggesting that the topical context is quite robust. There is however no clear optimal value for β, which leads to best MAP and P10 results.

DOI: 10.1002/asi

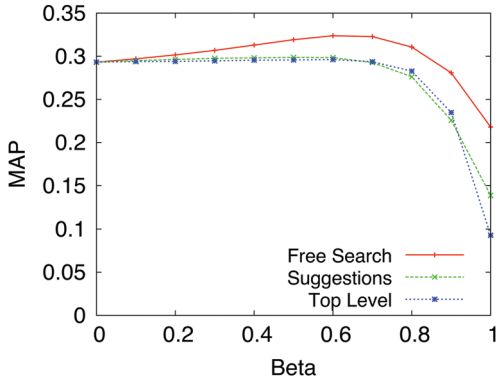

FIG. 4. Topical context: MAP. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

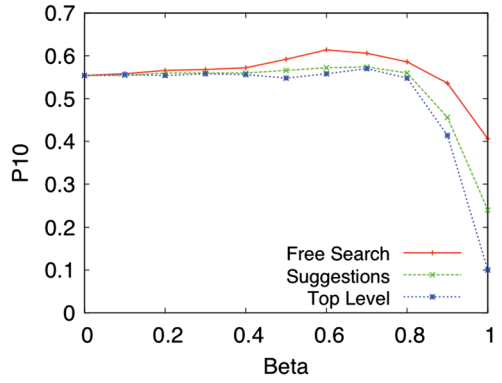

FIG. 5. Topical context: P10. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Topical context using the top-level categories or the suggested categories only leads to small, not significant improvements in early precision. We see that the topical context on the deepest level retrieved using free search in the DMOZ directory leads to the best results with significant improvements over the baseline where no topical context is used. There is no difference in the performance between categories evaluated as excellent by the participants and categories evaluated as relevant, but too broad or too specific.

Topical context in the form of a DMOZ category significantly improves the retrieval results when the DMOZ categories are selected using free search allowing categories at any level of the directory to be selected. It is difficult to compare our results to previous work, since the test collection is different. Similar to previous work (Bai et al., 2007; Ravindran & Gauch, 2004; Wei & Croft, 2007), we achieve significant improvements in average precision.

Topical feedback versus relevance feedback. We conduct experiments to get a better idea about the value of topical feedback compared with (blind) relevance feedback. First of all, we look at the relation between topical feedback and blind relevance feedback. The results of runs with and without topical as well as blind relevance feedback can be found

TABLE 8. Results of topical and blind relevance feedback.

Topical FB	Relevance FB	Prior	MAP	P10
None	No	No	0.2902	0.5680
None	Yes	No	0.3267	0.6120
Topic	No	No	0.2694	0.5560
Topic	No	Yes	0.2789	0.5160
Topic	Yes	No	0.3069	0.5760
W. Topic	No	Yes	0.3023	0.5560
W. Topic	Yes	Yes	**0.3339**	**0.6360**

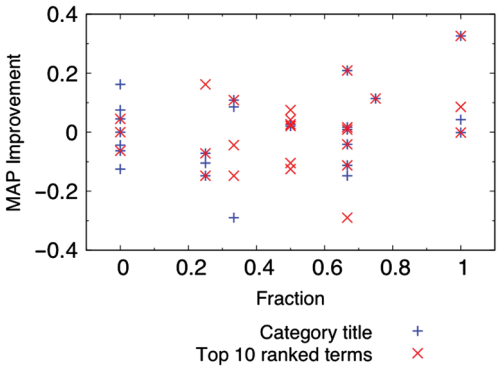

FIG. 6. MAP improvement correlations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

in Table 8. In the first column, the type of topical feedback is given, in the second column is shown whether additional blind relevance feedback is also applied. On average the topical feedback only leads to a small improvement of MAP over the baseline without blind relevance feedback. Applying only blind relevance feedback (second row in the table) leads to better results than applying only topical feedback (third row in the table). In the run Weighted Topic QE, we reweigh the original query terms according to the inverse fraction of query terms that occur in the category title, i.e., if half of the query terms occur in the category title, we double the original query weights. These runs lead to better results and to small improvements over blind relevance feedback, but they are not significant on our set of 25 queries.

The weighted topic query expansion works because there is a weak (non-significant) correlation between improvement in MAP when topic query expansion is used, and the fraction of query terms in either the category title, or the top-ranked terms of the topic language model, as can be seen in Figure 6. Applying a document length prior does not lead to consistent improvements or declines in retrieval performances. In the results table, we show the runs that gave the best results.

Furthermore, it is interesting to see that the topical feedback and blind relevance feedback are complementary. When blind relevance feedback is applied in combination with topical feedback, this leads to additional performance improvements.

Besides blind relevance feedback, we also consider explicit relevance feedback, where one or more document

DOI: 10.1002/asi

TABLE 9. Number of queries for which a feedback method gives the best results.

Model	Baseline		Relevance FB		Topical FB	
Blind Relevance FB	No	Yes	No	Yes	No	Yes
# Queries with best MAP	1	5	3	8	6	2
# Queries with best P10	4	7	9	12	4	10

is marked as relevant by the users. It is difficult however to make a fair comparison between topical feedback and explicit relevance feedback because of the evaluation. If the given relevant documents for relevance feedback are included in the ranking that is evaluated, it gives an unfair advantage to the relevance feedback approach. But if the given relevant documents are excluded from the ranking to be evaluated, it gives an unfair disadvantage compared with topical feedback. To compare explicit relevance feedback with topical feedback, we will therefore not look at the average retrieval scores, but look at it per query. As explicit relevance feedback we use one relevant document, which is provided in the relevance feedback track.

To compare the results of implicit and explicit relevance feedback and topical feedback, we look at which type of feedback gives the best results on our test set of 25 queries. Again we also consider the option to apply blind relevance feedback in combination with the other feedback methods. As can be seen in Table 9, each of the retrieval techniques works best for some of the queries. In case multiple retrieval techniques have the same best P10, they are all counted as best. Although additional blind feedback leads to significant improvements on average, there is a considerable number of queries where applying blind feedback leads to lower values of MAP and P10. It is hard to predict which kind of feedback will work best on a particular query. If we would be able to perfectly predict which feedback should be used, MAP would be 0.3917—an improvement of 42.3% over the baseline.This almost doubles the improvement that is achieved with the best single feedback technique.

We do find indicators to predict whether the topical feedback technique will improve over the baseline results or not. It turns out the user provided factors "confidence" and the "fit of the category" (based on the user study) do not have a strong correlation to performance improvement. The factors "fraction of query terms in category title" and "fraction of query terms in top-ranked terms" do have a weak correlation with performance improvements, as we have seen before. When the weight of the feedback is adjusted according to the query terms in the category title or the top-ranked terms, we see an improvement in the results. For pseudo-relevance feedback and explicit feedback, there is no such correlation between the fraction of query terms in top-ranked terms of the feedback model and the performance improvement. Since the feedback is based on top-ranked documents, the query terms always occur frequently in these documents.

There is also a positive side to the fact that the fit of the category does not correlate much with performance improvement. Sometimes categories that are clearly broader than the query lead to improvements. The queries "handwriting recognition" and "Hidden Markov Model HMM" both improve considerably when the topical model of category "Computers–Artificial Intelligence–Machine Learning" is applied. Hence, it seems that categories on more general levels than specific queries are useful and one topical model can be beneficial to multiple queries.

Summarizing, topical feedback can lead to significant improvements in retrieval performance when categories selected through free search in the DMOZ directory are used. High-level categories do not help to improve retrieval performance on average. The results of applying feedback vary per query, but in most cases topical feedback is complementary to blind relevance feedback.

Conclusion

In this article, we investigated methods to extract and use topical context using the DMOZ directory as our category hierarchy. We defined three research questions, the first one being: *How well can users categorize queries into DMOZ categories?* We conclude that the DMOZ directory is a good option to use as a source of categories, since for the vast majority of queries at least one relevant category is found. Two methods to elicit topical context are compared, free search on the DMOZ site to select the best category, and evaluation of a list of suggested categories. To create the list of suggestions, a combination of classification of query, top 10 retrieved documents, and a query title match is used. Free search leads to more specific categories than the list of suggestions. A problem in DMOZ is that category names are ambiguous when the full path in the category hierarchy is not taken into account. Different participants show moderate agreement between their individual judgments, but broadly agree on the initial levels of the chosen categories. Free search is most effective when agreement and coverage of queries is considered. According to the participants none of the methods is clearly better.

Second, we examined the question: *How can we use topical feedback to improve retrieval results?* Our experimental results show that topical feedback can indeed be used to improve retrieval results, but the DMOZ categories need to be quite specific for any significant improvements. Top-level categories, and the suggested categories from our list that go up to the fourth level, do not provide enough information to improve average precision. These categories could however be useful to cluster search results.

Our third research question: *Does topical feedback improve retrieval results obtained using standard relevance feedback?* A common and effective way to improve retrieval effectiveness is to use (blind) relevance feedback. On our dataset we find that combining topical context and blind relevance feedback on average leads to better results than applying either of them separately. Looking at a

query-by-query basis, we see that there is a large variance in which type of feedback works best. Topical feedback regularly outperforms explicit relevance feedback based on one relevant document and vice versa. For other queries using any type of feedback, only degrades the results. Hence, while topical context alone might not outperform (blind) relevance feedback on average, applying topical feedback does lead to considerable improvements for some queries. Finally, our main research question: *How can we explicitly extract and exploit topical context from the DMOZ directory?* From our experiments with the DMOZ directory, we can conclude that DMOZ is a good resource to use to interact with users on the topical categories applicable to their query. The large size of the directory means that specific categories applicable to queries can be found. The average improvements in the performance of topical feedback are small and not always significant in our experiments. While for some queries using topical context from the DMOZ directory greatly improves the retrieval results, it is probably not worth the effort to apply it blindly to each and every query. Besides using topical context to improve retrieval results, topical context can be used for suggestion of topically related query terms, or to cluster results into subtopics.

We can conclude that DMOZ is a good resource to use for topical feedback, but we do not know whether it is better than using the Yahoo! directory, or the category hierarchy from Wikipedia. The methods described in this article can be applied to any category hierarchy containing documents. Further experiments can be conducted to determine which category hierarchy is most appropriate for topical feedback. Especially as Wikipedia is growing at a fast pace and has a large user base, it is an interesting alternative.

In this study, we have made some adjustments to our methods to improve efficiency, i.e., we rerank 1,000 results in the feedback algorithms, our query categorization methods expand only the top 20 subcategories of each category, and only classify categories up to level 4 in the category hierarchy. Reranking results have limited influence on early precision, it is not likely that documents below rank 1,000 in the initial ranking end up in the top 10 by applying feedback. Some improvements in average precision might occur when more documents are considered for feedback. Expanding 20 subcategories of each category during query categorization covers a large part of all categories in the hierarchy, and therefore we do not expect including the small number of most likely irrelevant categories will not lead to any improvements. Classifying only up to level 4 categories is a big limitation for the automatic query categorization, as we have seen that in the free search the participants select categories below level 4 in more than half of the cases.

For future work, we would like to extend our query categorization methods to suggest categories from the complete directory, and experiment with completely automatically generated topical feedback. Furthermore, we would like to experiment with using the documents in the directory directly as search results, besides using the textual content of these documents as a source for query expansion terms. While there is little overlap with the documents in DMOZ and the .GOV2 collection, a new web collection has recently become available. The Clueweb (Carnegie Mellon University, Language Technologies Institute, 2010) document collection contains one billion web pages and will contain considerably more DMOZ pages opening up new opportunities.

References

Azzopardi, L., Girolami, M., & Rijsbergen, C.V. (2004). Topic based language models for ad hoc information retrieval. In Proceedings of the IEEE International Joint Conference on Neural Networks (pp. 3281–3286). Washington, DC: IEEE Press.

Bai, J., Nie, J.-Y., Bouchard, H., & Cao, G. (2007). Using query contexts in information retrieval. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '07) (pp. 15–22). New York: ACM Press.

Balog, K., Vries, A.D., Serdyukov, P., Thomas, P., & Westerveld, T. (2009). Overview of the TREC 2009 entity track. Paper presented at The 18th Text REtrieval Conference Notebook (TREC '09), Gaithersburg, MD. Retrieved from http://krisztianbalog.com/files/talks/trec2009-entity-overview.pdf

Blei, D.M., Ng, A.Y., Jordan, M.I., & Lafferty, J. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Carnegie Mellon University, Language Technologies Institute. (2010). The ClueWeb09 Dataset. http://boston.lti.cs.cmu.edu/Data/clueweb09/ (Accessed 11-3-2011).

Chirita, P., Nejdl, W., Paiu, R., & Kohlshuetter, C. (2005). Using ODP metadata to personalize search. In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05) (pp. 178–185). New York: ACM Press.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Demartini, G., Iofciu, T., & Vries, A.P.D. (2009). Overview of the INEX 2009 entity ranking track. In Proceedings of Focused Retrieval and Evaluation: Eighth International Workshop of the Initiative for the Evaluation of XML Retrieval (INEX '09)(pp. 254–264). Berlin, Germany: Springer Verlag.

Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S.T. (1987). The vocabulary problem in human–system communication. Communications of the ACM, 30, 964–971.

Haveliwala, T.H. (2002). Topic-sensitive pagerank. In Proceedings of the 11th International Conference on World Wide Web (WWW '02) (pp. 517–526). New York: ACM Press.

Hearst, M.A., & Pedersen, J.O. (1996). Reexamining the cluster hypothesis: Scatter/gather on retrieval results. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96) (pp. 76–84). New York: ACM Press.

Hiemstra, D., Robertson, S., & Zaragoza, H. (2004). Parsimonious language models for information retrieval. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04) (pp. 178–185), New York: ACM Press.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99) (pp. 50–57). New York: ACM Press.

Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, 44(3), 161–174.

Jansen, B.J., Spink, A., & Koshman, S. (2007). Web searcher interaction with the Dogpile.com metasearch engine. Journal of the American Society for Information Science and Technology, 58(5), 744–755.

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. Information Processing & Management, 36, 207–227.

Kamps, J. (2006). Effective smoothing for a terabyte of text. Paper presented at the 14th Text REtrieval Conference (TREC '05). Gaithersburg, MD.

DOI: 10.1002/asi

Retrieved from http://trec.nist.gov/pubs/trec14/papers/uamsterdam.tera.kamps.pdf

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference: a bibliography. SIGIR Forum, 37, 18–28.

Kim, J., & Croft, W.B. (2010). Ranking using multiple document types in desktop search. In Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10) (pp. 50–57). New York: ACM Press.

Lau, T., & Horvitz, E. (1999). Patterns of search: Analyzing and modeling web query refinement. In Proceedings of the Seventh International Conference on User Modeling (pp. 119–128), Secaucus, New Jersey: Springer-Verlag, Inc.

Lavrenko, V., & Croft, W.B. (2001). Relevance-based language models. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '01) (pp. 120–127). New York: ACM Press.

Lee, K.-F. (2008). Delighting Chinese users: The Google China experience. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (p. 1). New York: ACM Press.

Liu, X., & Croft, W.B. (2004). Cluster-based retrieval using language models. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '04) (pp. 186–193). New York: ACM Press.

Liu, F., Yu, C., & Meng, W. (2002). Personalized web search by mapping user queries to categories. In Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM '02) (pp. 558–565). New York: ACM Press.

Meij, E., Trieschnigg, D., De Rijke, M., & Kraaij, W. (2010). Conceptual language models for domain-specific retrieval. Information Processing & Management, 46, 448–469.

Miller, G.A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38, 39–41.

Oxford English Dictionary (2011). The Oxford English dictionary—Relaunched. Retrieved from http://oxforddictionaries.com/page/oedrelaunch/the-oxford-english-dictionary-relaunched/

Porter, M.F. (1997). An algorithm for suffix stripping. Readings in Information Retrieval (pp. 313–316). San Francisco: Morgan Kaufmann Publishers Inc.

Ravindran, D., & Gauch, S. (2004). Exploiting hierarchical relationships in conceptual search. In Proceedings of the 13th ACM International Conference on Information and Knowledge Management (CIKM '04) (pp. 238–239). New York: ACM Press.

Rocchio, Jr., J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), The SMART Retrieval System: Experiments in Automatic Document Processing, Prentice-Hall Series in Automatic Computation, Chapter 14 (pp. 313–323). Englewood Cliffs, New Jersey: Prentice-Hall.

Rosso, M.A. (2008). User-based identification of web genres. Journal of the American Society for Information Science and Technology, 59(7), 1073–1092.

Ruthven, I., & Lalmas, M. (2003). A survey on the use of relevance feedback for information access systems. Knowledge Engineering Review, 48(2), 95–145.

Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science, 41(4), 288–297.

Saracevic, T., & Kantor, P. (1988). A study in information seeking and retrieving, II. Users, questions, and effectiveness. Journal of the American Society for Information Science, 39(3), 176–195.

Sondhi, P., Chandrasekar, R., & Rounthwaite, R. (2010). Using query context models to construct topical search engines. In Proceedings of the Third Symposium on Information Interaction in Context (IIiX '10) (pp. 75–84). New York: ACM Press.

Strohman, T., Metzler, D., Turtle, H., & Croft, W.B. (2005). Indri: A language-model based search engine for complex queries. Paper presented at the International Conference on Intelligent Analysis. Retrieved from https://analysis.mitre.org//proceedings/index.html

Trajkova, J., & Gauch, S. (2004). Improving ontology-based user profiles. In Proceedings of the Recherche d' Information Assiste par Ordinateur (RIAO '04) (pp. 380–389). Paris: C.I.D.

TREC (2011). Text REtrieval Conference. URL: http://trec.nist.gov/.

Trieschnigg, D., Pezik, P., Lee, V., De Jong, F., Kraaij, W., & Rebholz-Schuhmann, D. (2009). MeSH Up: Effective MeSH text classification for improved document retrieval. Bioinformatics, 25(11), 1412–1418.

Van Rijsbergen, C.J. (1979). Information Retrieval. Newton, MA: Butterworth-Heinemann

Voorhees, E.M. (1994). Query expansion using lexical-semantic relations. In Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94) (pp. 61–69). New York: ACM Press.

Wei, X., & Croft, W.B. (2006). LDA-based document models for ad-hoc retrieval. In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06) (pp. 178–185). New York: ACM Press.

Wei, X., & Croft, W.B. (2007). Investigating retrieval performance with manually-built topic models. Large Scale Semantic Access to Content (Text, Image, Video, and Sound) (RIAO '07) (pp. 333–349). Paris: C.I.D.

Zhai, C. (2008). Statistical language models for information retrieval, a critical review. Foundations and Trends in Information Retrieval, 2, 137–213.