Web Directories as Topical Context

Rianne Kaptein[1] Jaap Kamps[1,2]

[1] Archives and Information Studies, Faculty of Humanities, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

ABSTRACT

In this paper we explore whether the Open Directory (or DMOZ) can be used to classify queries into topical categories on different levels and whether we can use this topical context to improve retrieval performance. We have set up a user study to let test persons explicitly classify queries into topical categories. Categories are either chosen freely from DMOZ, or from a list of suggestions created by several automatic topic categorization techniques. The results of this user study show that DMOZ categories are suitable for topic categorization. Either free search or evaluation of a list of suggestions can be used to elicit the topical context. Free search leads to more specific topic categories than the list of suggestions. Different test persons show moderate agreement between their individual judgments, but broadly agree on the initial levels of the chosen categories. When we use the topic categories selected by the free search as topical context, this leads to significant improvements over the baseline retrieval results. The more general topic categories selected from the suggestions list, and top level categories do not lead to significant improvements.

1. INTRODUCTION

One of the main bottle-necks in providing more effective information access is the poverty of the query end. With an average query length of about two terms, 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. Associating the query with a topical category can help to disambiguate the query. 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 non-relevant topic categories or to cluster search results. In this paper we will investigate how to get and use topical context on different levels of granularity. Queries can be associated with a topical category by using implicit or explicit techniques. Examples of identifying topical context implicitly are using a user profile built on previous information seeking behavior, previous issued queries, or automatic classification of query words or retrieved documents. We will elicit the context explicitly, i.e. ask the user to classify a query into a topical category.

Several large directories on the web have organised their information into topical categories, usually in a hierarchical way e.g. DMOZ (also known as ODP Open Directory Project) [5], Yahoo! Directory [18] and Wikipedia [17]. There has been a stream of papers that use some form of topical model or context use the DMOZ directory, or a part of it, to represent topical categories (see Section 2 below). DMOZ has a lot of attractive features, it is hierarchical, large, and created by human users especially for the web. In a previous study [11] we have used a small number of self-defined categories, that did not cover a wide range of query topics. By using a considerable part of the DMOZ category we can cover a wide range of topics. For these reasons this paper uses the DMOZ directory to represent topical categories.

Being large also has some disadvantages, for users it might not be so easy find the category they are searching for. There is a trade-off between the quality of the user categorization, i.e. whether the category covers exactly the query topic, and the effort that is needed. Searching or browsing the complete directory requires the most effort from the user, but can result in finding categories an automatic classifier can not find. Choosing from a list of suggestions takes less effort from the user, but there is always a risk that the best possible topic category is not included in the list of suggestions. We will 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.

In this paper we address the following main research question:

- Can we effectively use the DMOZ directory as a source of topical context?

We break up our main research question, into the following two research questions:

1. Can the DMOZ directory be used to effectively categorize query topics into topic categories?

We carry out a user study that identifies topical context explicitly in order to answer our first research question. We explore whether the topic categories in DMOZ are representative for query topics. Furthermore, we compare two different forms of deriving context explicitly, i.e. free search or browsing on the DMOZ site, and evaluation of topic categories from a list of suggestions. Agreement on the relevance of DMOZ categories between different test persons is also considered. To answer our second research question, we use the results from our user study to look at the effects of using topical context on retrieval performance:

2. Can we use topical context to improve retrieval effectiveness?

We compare performance of runs using topical context in addition to the query on different levels in the DMOZ directory.

The rest of this paper is organized as follows. In the next section we discuss related work. In Section 3 we describe the data, i.e. the

DIR 2009, February 2-3, 2009, Enschede, the Netherlands.
.

queries that we use and the DMOZ directory. Section 4 describes the language models that we are using for topic categorization and retrieval. In Section 5 we discuss the user study we have executed. Section 6 describes experiments where we use the topical context that we got from our user study to try to improve retrieval effectiveness. Finally in Section 7 we discuss the results and draw our conclusions.

2. RELATED WORK

There is a range of studies that use topical models to improve retrieval performance or retrieval effectiveness [1, 2, 4, 6]. Two approaches are commonly used, one approach creates some kind of user profile that does not depend on the query. These user profiles can be built in different ways. Chirita et al. [4] lets users pick multiple topic categories from DMOZ to create user profiles which best fit their interests. At run-time the output of the search engine is reranked by using a calculated distance from the user profile to each output URL.

Liu et al. [12] builds user profiles automatically by using the search history, that consists of the issued queries, relevant documents and related categories. A new query is mapped to a set of categories using the user profile, a general profile, or a combination of user and general profile. The categories are ranked, and the top 3 categories are chosen to reflect the user's search intention.

Also Trajkova and Gauch [14] builds user profiles 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 DMOZ directory. Only the top 3 levels of the directory are used. To classify a Web page, the highest weighted 20 words are used to represent the content of the Web page. Classification consists of comparing the vector created for the Web page with each category vector (created and stored during training) using the cosine similarity measure.

The other approach, which is also employed in this paper, is to use topical models that depend on the query. Wei and Croft [16] manually assign topic categories according to some basic rules. Haveliwala [6] considers two scenarios to assign topical categories to queries. Both scenarios use personalization vectors calculated for the 16 top-level DMOZ categories. 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. 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 topic. 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 the users' browsing patterns, bookmarks, and e-mail archives.

Bai et al. [2] compares the automatic and the manual assignment of topical domains. Here, the topic domains do not come from an existing topic hierarchy, but the users can define their own domains. Domain models are created by either using the relevant documents for the in-domain queries, or by using the top 100 documents retrieved with the in-domain queries. TREC queries 51-150 are used, since these query topics also include a manually assigned topic domain. 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 query domain is assigned manually.

Besides topical context, a well-studied form of context is genres of webpages. For example, Rosso [13] explores user-based identification of web genres. He defines genre as: essentially a document type based on purpose, form, and context. Examples of genres are resumes, scientific articles or tax income forms. The study contains of three parts, first information is obtained on what genres users perceive. Secondly, all used terminology is refined into a tentative genre palette. Finally, the genre palette is validated by letting users classify pages into the defined genres.. The study is restricted to pages from the edu domain to increase the chance of developing a recognizable palette.

3. DATA

In this paper we investigate whether we can use the DMOZ directory as a source of topical context. We use ad hoc topics from the TREC Terabyte tracks as test data. The TREC Terabyte track ran for three 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 keyword 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 information about which documents are considered relevant and/or irrelevant. All topics are created by NIST assessors [3].

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, around 1% of the sites listed in 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 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.

The DMOZ directory is organized as a tree, where the topic categories are inner nodes and pages are leaf nodes. 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 dataset, which is available in RDF. Google for example uses DMOZ as basis for its Google Directory service [4].

The complete DMOZ directory contains over 590,000 categories. Categories selected by test persons during the free search can be any of the 590,000 categories, except categories under the "World" category, that contains categories in languages other than English. It does not matter if the category contains links to webpages or not. We allow multiple DMOZ categories to be assigned to one topic.

To produce the list of suggestions, we focus on a part of the DMOZ directory in order to reduce complexity. That is, we use mainly categories from the first four levels of DMOZ, which still comprises around 30,000 categories. In addition we consider a classification on the top level of the DMOZ directory, which comprises of 15 topic categories.

4. MODELS

Topical context can be derived either implicitly or explicitly. In this paper we focus on explicitly derived topical context that is obtained from a user study. We first describe the language modeling

approach, followed by the models for topic categorization that are used to generate a list of suggested categories. These same models could be used to derive topical context implicitly. In the last part of this section, we describe the model we use to incorporate topical context in our retrieval model.

4.1 Language Modeling

We use unigram language models [7] for topic categorization as well as for retrieval. Our standard model for document retrieval uses Jelinek-Mercer smoothing [19] in a mixture of the document model with a general collection model as follows, i.e., for a collection C, document D, query Q and smoothing parameter λ:

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

where

$$\begin{aligned} P_{mle}(t|D) &= \frac{tf_{t,D}}{\sum_t tf_{t,D}} \\ P_{mle}(t|C) &= \frac{\text{doc\_freq}(t,C)}{\sum_{t' \in C} \text{doc\_freq}(t',C)} \end{aligned}$$

Instead of using maximum likelihood estimation to estimate the probability $P(t|D)$, it can also be estimated using parsimonious estimation. The parsimonious model concentrates the probability mass on fewer terms than a standard language model. Terms that are better explained by the general language model $P(t|C)$ (i.e. terms that occur about as frequent in the document as in the whole collection) can be assigned zero probability, thereby making the parsimonious language model smaller than a standard language model. The model automatically removes stopwords, and words that are mentioned occasionally in the document [8].

The model is estimated using *Expectation-Maximization*:

$$\begin{aligned} \text{E-step:} \quad & e_t = tf_{t,D} \cdot \frac{\alpha P(t|D)}{\alpha P(t|D) + (1-\alpha) P(t|C)} \\ \text{M-step:} \quad & P(t|D) = \frac{e_t}{\sum_t e_t}, \text{i.e. normalize the model} \end{aligned}$$

In the initial E-step, the maximum likelihood estimates are used to estimate $P(t|D)$. The E-step benefits terms that occur relatively more frequent in the document as in the whole collection. The M-step normalizes the probabilities. After the M-step terms that receive a probability below a certain threshold are removed from the model. In the next iteration the probabilities of the remaining terms are again normalized. The iteration process stops after a fixed number of iterations or when the probability distribution does not change significantly anymore. For $\alpha = 1$, and a threshold of 0, the algorithm produces the maximum likelihood estimate $P_{mle}(t|D)$ as defined before. Lower values of α result in a more parsimonious model. We will denote the resulting estimate by $P_{pars}(t|D)$.

4.2 Topic Categorization

We will discuss three methods to automatically categorize topics. The first two methods are similar, and consist of two steps. In the first step we create topical models of the DMOZ topic categories. Secondly, we assign a topical category to each query by using either the query title, or the top 10 retrieved documents.

To create a topical model for a topic category, we crawl the sites from the category, and of all its available direct sub categories. If that results in at least 10 sites a parsimonious language model is created using the Expectation-Maximization equation described in the previous section. Instead of document D we now have TM, the topical model, that consists of the raw text of the web sites belonging to the category. The background collection C here is the DMOZ background corpus. It consists of the raw text of all web pages up to level 4 we were able to crawl. All terms with term frequency 1 are excluded from the background corpus. The corpus consists of 459,907 documents and a total number of 350,041,078 terms.

The websites used to create the topical model are spread over the category and all its subcategories. For efficiency reasons we have crawled only the upper four levels of the DMOZ directory, so we can create topical models up until the third level of the hierarchy using also the subcategories. The topical models on the fourth level use only the sites on that level.

The second step is to assign a topical category to each query. Our first method is based on classifying documents.

Top ranking documents We use the top 10 results of a baseline model run, and select categories fitting these documents best.

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

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

The topical models ranked by their probabilities are 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 sub categories from the 20 highest ranked topic categories are used. In the end, the topical category belonging to the topical model with the highest probability, no matter on what level, is assigned to the query.

Our second method is directly classifying the query.

Query We simply classify the short topic statement in the title field, and select best matching categories

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

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

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

The third method we use to categorize the query is simple.

Title match We match the query words with the label of the topic category.

If all query words are present in the topic category label, the topic category is assigned to the query. When a topic category matches all query words, all its descendants automatically also match all query words. However, we then only assign the highest level topic category. Both the query words and the topic category labels are stemmed using a Porter stemmer. This method only assigns a topic category to a query topic if there is an exact match.

To produce a list of suggestions for a topic, we merge the top 10 ranked categories from the three categorization methods

4.3 Retrieval

For retrieval we use not only the query, but also a topical model assigned to the query topic. To produce a ranking a mixture of the query model and the topical model is used as follows:

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

Table 1: Coverage of topics: taking all evaluations, and taking the best evaluation per topic.

	Not relevant		Too broad		Too specific		Excellent	
	All Evals	Best Eval	All Evals	Best Eval	All Evals	Best Eval	All Evals	Best Eval
Suggested:								
Query	78.7% (1,193)	14.1% (19)	15.8% (239)	45.2% (61)	3.6% (54)	15.6% (21)	2.0% (30)	25.2% (34)
Top Docs	77.2% (1,188)	11.1% (15)	19.8% (304)	60.7% (82)	1.9% (29)	15.6% (21)	1.1% (17)	12.6% (17)
Rel Pages	79.4% (1,212)	4.0% (2)	18.1% (276)	54.0% (27)	1.6% (25)	18.0% (9)	0.9% (13)	24.0% (12)
Title Match	17.9% (5)	0.0% (0)	17.9% (5)	0.0% (0)	21.4% (6)	14.3% (2)	42.9% (12)	85.7% (12)
Total	80.1% (2,861)	1.5% (2)	15.8% (563)	45.2 % (61)	2.6% (93)	17.8% (24)	1.6% (56)	35.6% (48)
Free Search:								
First Cat.	3.4% (8)	1.5% (2)	14.8% (35)	9.0% (12)	43.5% (103)	35.3% (47)	38.4% (91)	54.1% (72)
Second Cat.	5.2% (3)	4.3% (2)	22.4% (13)	13.0% (6)	56.9% (33)	63.0% (29)	15.5% (9)	19.6% (9)
Total	3.7% (11)	1.5% (2)	16.3% (48)	9.0% (12)	46.1% (136)	35.3% (47)	33.9% (100)	54.1% (72)

$P(TM|D)$ is estimated similarly to $P(Q|D)$ as described before.

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

5. USER STUDY

In this section we describe the user study that has been executed in order to let test persons assign topic categories to query topics.

5.1 Design

The user study is designed as follows. Test persons first read an instruction, and do a training task. Before starting the actual tasks, test persons fill out a pre-experiment questionnaire that consists of some demographic questions. The main part consists of 15 tasks. Each task corresponds to one topic. At the beginning of each task the topic, 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.
3. Search or browse on the DMOZ site to find the best category.
4. Post-task questions

In the second and third task also some questions are asked on how easy the task was, and how confident the test persons are about their categorization. After the 15 tasks each test person fills out a post-experiment questionnaire that consists of questions on how they experienced and liked the different tasks. At each stage of the user study, there are open questions for comments of any kind.

In subtask 2 the test person evaluates a list of suggested categories. For each suggestion the test person evaluates how relevant the category is to the topic by answering the question: "For each suggested category evaluate how relevant it is to the topic". The four options are: "Not at all", "Relevant, but too broad", "Relevant, but too specific", and "Excellent". The list of suggestions is composed of the categories resulting from the three topic categorization methods described in the previous subsection.

In subtask 3 the test person is free to select a category from the DMOZ site that he or she thinks applies best to the topic. Categories can be found by browsing the DMOZ site, or by using the search function on the DMOZ site. If the test person finds more than one category that applies best to the query topic, there is a possibility to add a second DMOZ category. The test person evaluates again the relevance of the found category to the topic. We do not rotate subtask 2 and 3 because our goal is to obtain good human feedback. Seeing the list of suggestions first means there is a learning effect which can improve the quality of the categories selected in the free search.

5.2 Set-up

The user study is done using the queries from the three TREC Terabyte tracks 2004, 2005 and 2006 (.GOV2 collection of 25M documents) [15]. Topics 801-850 are done by two to four test persons, all other topics are done by one test person. In total 135 out of the 150 Terabyte topics are covered. The order and the selection of topics is randomized. Each test persons gets assigned 15 topics.

For the automatic query topic categorization we have to set some parameters. We use the topic categorization methods as described in Section 4.2, where $P(t|TM)$ is calculated according to the parsimonious model, $P_{pars}(t|TM)$. Stopwords are removed according to a standard stopword list. Stemming is not applied.

The standard value of the smoothing parameter λ in the language model is 0.15. In the TREC Terabyte tracks, it is known that the .GOV2 collection requires little smoothing [9], i.e. a value of 0.9 for λ gives the best results.

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 topic set.

The online user study records all answers, and also the time it takes test persons 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 they are all in the same format.

5.3 Results

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

Demographics

The user study has been filled out by 14 test persons, of which 9 male and 5 female. Two test persons participated twice in the user study, so they did 30 instead of 15 topics. The main part of the test persons is studying or working within the field of information retrieval. Average age is 31 years. Half of them are familiar with the DMOZ directory, and 3/4 of them are familiar with the subject of topic categorization.

Appropriateness of DMOZ categories

We first look at the question: does an appropriate DMOZ category exists for the topics? In Table 1 we present the coverage of the query topics, that we get from the answers to the question of how relevant the suggested and the free search categories are to the top-

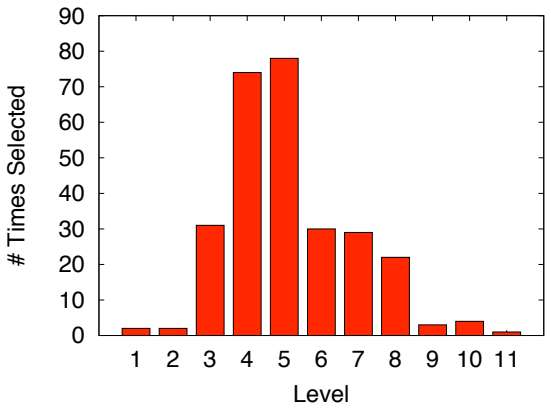

Figure 1: Levels of DMOZ categories selected by free search

Table 2: Free search vs. Suggestions list results

	Free Search		Suggestions	
	Avg.	Final	Avg.	Final
Time in min.	2.0		1.3	
Quick		3.5		3.5
Confident	3.5	3.4	3.5	3.4
Easy	3.0	3.2	3.2	3.5

ics. The columns with 'All Evals' are all evaluations per topic of all test persons taken together. The 'Best Eval' columns take only the best evaluation of all test persons and categories per topic, where a 'Too specific' category is rated above 'Too broad'. To produce the suggested categories a fourth categorization method was included. Similar to the 10 top ranked documents, 10 randomly chosen relevant documents are used for categorization (Rel Pages). We can see that when the list of suggestions is used, for only 1.5% of the topics no relevant DMOZ category is found. When the category is relevant, it is usually too broad (45.2% of the topics). Still, for 35.6% of the topics and excellent matching category is found. When free search is used, also for 1.5% of the topics no relevant category is found. For more than half of the topics (54.1%) an excellent matching category is found.

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 test persons can select a category on any level of the DMOZ directory. Figure 1 shows the distribution of categories over the level of the DMOZ hierarchy. We see that the deepest level that is chosen is 11. The median level is 5.

List Selection versus Free Search

We now turn to the two ways of eliciting explicit category feedback: either by selecting from a list of suggestions, or by freely searching the DMOZ hierarchy.

Table 2 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 to 5, where 1 means 'Strongly Disagree' and 5 means 'Strongly Agree'. Average numbers are averaged over all test persons and all topics. The final numbers are averages over all test persons on answers in the post-experiment questionnaire. When comparing the free search with the evaluation of suggested categories, we have to consider a bias that occurs because the test persons always first evaluate the list of suggested categories and then do the free search. In close to 50% of the cases, the test persons say the list of suggestions helped them to select a category from the DMOZ site using free search. In 55% of the cases the test persons 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 topic for the free search is higher than the average time spent for the evaluation of the suggested categories (2.0 minutes and 1.3 minutes respectively). The test persons however perceive both methods to be as quick. The confidence in their classifications is the same on average, and in the final evaluation for both methods. The test persons find the evaluation of the suggestions list slightly easier than the free search.

When asked what method the test persons prefer, the replies are mixed. 3 test persons prefer free search, 4 test persons prefer evaluation of a suggestions list, and 7 test persons prefer to look at a list of suggestions, and then search freely on the DMOZ site.

Agreement between Test Persons

We now look at the agreement between different test persons categorizing the same topic. We calculate pairwise agreement between test persons. Strict agreement means there is agreement on the relevant categories, and on the degree of relevance ('Relevant, but too broad', 'Relevant, but too specific', and 'Excellent'). Lenient agreement means 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 test persons are not included.

For the suggestions list two types of agreements are calculated. 'All evaluations' calculates agreement for each category on the suggestions list when at least one test person considers the category relevant. One combination of different methods is used on the suggestions list, i.e. a category is only selected if both the classification of top 10 retrieved documents and the query produce the category (see Combination in Table 3). 'Best match' only calculates agreement for the category of the suggestions list with the best agreement. Similarly, when free search is used, and two topic categories are selected, only the best matching categories are used to calculate agreement. Agreement is calculated on different levels, where categories are simply cut off at the desired level. The 'Complete' row gives agreement on the complete topic categories without cut off. The results are presented in Table 3.

What is the agreement between test persons? Strict agreement for the suggestions list total and the free search is almost the same, 0.14 and 0.15 respectively. Categories selected by free search receive somewhat higher lenient agreement than the categories from the list of suggestion, 0.20 and 0.34 respectively.

What is the difference in agreement over the different list 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 topic, and that receive the most agreement. The drawback of this method, is that only for a small percentage of topics, there is an exact match with a DMOZ category label (6). Expanding this method to include nearly exact matches could be beneficial. The combination of methods also achieves better agreement than the separate methods, on a larger number of topics (23).

Every chosen category in the DMOZ hierarchy is subcategory of a whole path up to the root node. So different categories may still share the same top-level categories. What is the agreement over levels of the DMOZ hierarchy? We look here at the best matching

Table 3: Strict and lenient agreement between test persons over all relevant judgments, and over best matching relevant judgements.

	# topics	Strict	Lenient
List (All evaluations)			
Query	44	0.12	0.22
Top Docs	49	0.14	0.18
Rel Pages	48	0.15	0.18
Combination	23	0.28	0.38
Title Match	6	0.69	0.89
Total	50	0.14	0.20
List (Best match)			
Level 1	50	–	0.75
Level 2	50	–	0.73
Level 3	48	–	0.67
Level 4	37	–	0.48
Complete	50	0.61	0.75
Free Search (Best match)			
Level 1	50	–	0.74
Level 2	50	–	0.64
Level 3	50	–	0.58
Level 4	50	–	0.50
Complete	50	0.15	0.34

relevant category only. For the free search, agreement on levels 1 to 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.

Summarizing, from our user study we can conclude that for nearly all topics 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 test persons 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 test persons is moderate, but increases considerably when we look only at the top-level categories.

6. EXPERIMENTS

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

6.1 Experimental Set-Up

To test our topical feedback approach, we use Terabyte topics 800 to 850 that have been classified by at least two test persons in our user study.

All parameters for the topical models are the same as used in the user study. However, for retrieval we do use a Porter stemmer, because our initial results indicate that stemming leads to better results. We also experimented with document length normalization, but that does not lead to any improvements. 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 (λ = 0.9). We rerank the top 1,000 results.

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

Table 4: Retrieval results using topical context

Topical Context	Beta	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	$0.2179^{\circledcirc}$	$0.3640^{\circledcirc}$
Top Level	0.7	0.2937^{-}	0.5700^{-}
Suggestions	0.6	0.2984^{-}	0.5720^{-}
Free Search	0.6	$0.3238^{\bullet}$	$0.6140^{\circledcirc}$

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

6.2 Experimental Results

Table 4 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 topics. As expected, when only the topical context is used ($\beta = 1.0$), results are significantly worse than the baseline. The free search categories do still perform quite reasonably, showing that the DMOZ categories can capture the information request at hand. Second, we look at combining the baseline run with topical context. In the table only the best runs are shown. Topical context using the top level categories or the suggested categories only leads to small, not significant improvements in early precision. We see that 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. We show MAP and P10 over different values of β in Figure 2. 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.

In terms of effectiveness, there seems to be a relation with the depth in the DMOZ hierarchy. Figure 3 shows the correlation between the level of the category used as topical context, and the improvement in MAP as the result from using the free search categories as topical context. Besides the average MAP improvements per level, we added the MAP improvements per query topic.

Topical context in the form of a DMOZ category significantly improves retrieval results when the DMOZ categories are selected using free search allowing categories at any level of the directory to be selected.

7. DISCUSSION AND CONCLUSIONS

In this paper we investigated methods to get and use topical context from users where the DMOZ directory provides topic categories. We investigated two research questions, our first one being: *Can the DMOZ directory be used to effectively categorize query topics into topic categories?* We conclude that the DMOZ directory can be considered suitable to categorize query topics into cat-

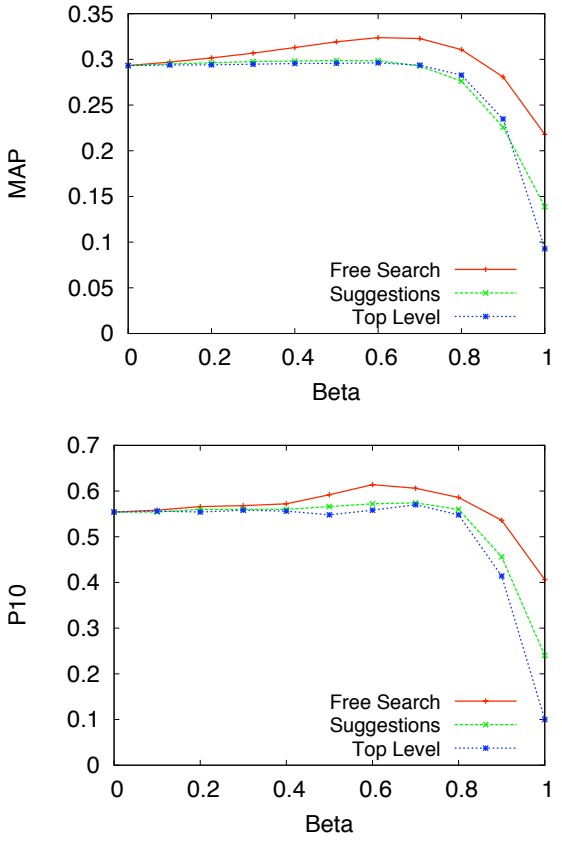

Figure 2: Topical context: MAP and P10

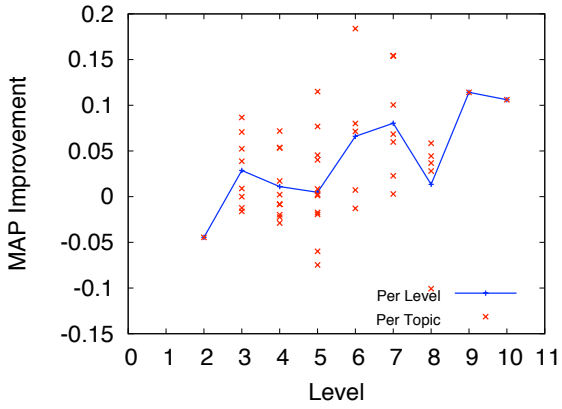

Figure 3: Correlation between level of free search topic category and MAP improvement

egories. Using either free search or the suggestions list for 98.5% of the query topics a relevant DMOZ category is found. This category can however be too broad or too specific. When test persons 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 topics, the categories till level 4 of the DMOZ category are not specific enough to categorize topics 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 to use to elicit the topical context, there is no clear preference from the test persons point of view. In our set-up there is however a difference in the quality of the topic categorization. The list of suggestions only retrieves topic 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 used to produce suggestions to the fifth or a even deeper level, thus has clear potential to improve the quality of the suggestions list. Our test persons now consider evaluation of suggested categories easier, and they are also faster. It would be interesting to see if these advantages still hold when deeper level categories are also shown in the suggested categories list.

Looking at the different methods of automatic topic categorization, the title match of the query words with DMOZ category labels produces high quality suggestions, but not for many topics. Using a more lenient title match, where not all query words have to occur in the category title could provide us with more possible relevant topic 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 agreement and the coverage of query topics, 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 three 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 test persons? Considering the test persons can choose from 590,000 categories, the lenient agreement of 0.34 for the free search is quite good. For the list based suggestions, the lenient agreement over all categories deemed relevant by any of the test persons is 0.20. A problem with the evaluation of the suggestions list is that some test persons tend to select only one or two categories, while other test persons evaluate substantially 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 lenient 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 to 4 of the directory is much better, so at least test persons start out on the same path to a topic category. They may only in the end 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 topic. However, the categories in the suggestions list can still be improved by including more of the DMOZ hierarchy. From the test persons point of view, there is no agreement on a preference for one of the methods. So, 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.

Our second research question was: *Can we use topical context to improve retrieval effectiveness?* Our experimental results show that topical context can indeed be used to improve retrieval effectiveness, but the topical 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.

Looking at the level of the topic category in correlation to MAP improvement, we find a weak positive correlation. Deeper levels of categorization are likely to lead to better MAP improvements, but we need more data for statistical proof.

A common and effective way to improve retrieval effectiveness is to use (pseudo) relevance feedback. On this TREC data set it is found that combining topical context and pseudo relevance feedback leads to better results than applying either of them separately [10]. So while topical context alone might not outperform (pseudo) relevance feedback, their contributions to performance are complementary.

Finally, our main research question: *Can we effectively use the DMOZ directory as a source of topical context?* We can conclude that the DMOZ directory is a good option to use as a source of topic categories, since for the vast majority of query topics at least one relevant topic 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 categories. Free search is most effective when agreement and coverage of query topics is considered. According to the test persons none of the methods is clearly better. To create the list of suggestions a combination of classification of query, top 10 retrieved documents, and a query title match can be used. Looking at retrieval effectiveness the more specific free search categories are to be preferred, since these categories are the only categories that lead to significant improvements over the baseline.

In future work we want to address the question whether automatic categorization into topic categories can also benefit retrieval. In that case no input from the user is required. So far, free search categories chosen by test persons seem to be of a better quality than suggested categories obtained by automatic categorization, but extending the automatic categorization into deeper levels of the hierarchy might lead to better results.

Acknowledgments.

We would like to thank Rongmei LI and Djoerd Hiemstra for their cooperation, and all test persons for their efforts. Rianne Kaptein was supported by the Netherlands Organization for Scientific Research (NWO, grant # 612.066.513). Jaap Kamps was supported by NWO (grants # 612.066.513, 639.072.601, and 640.-001.501).

REFERENCES

[1] L. Azzopardi, M. Girolami, and C. van Rijsbergen. Topic based language models for ad hoc information retrieval. In *IEEE International Joint Conference on Neural Networks*, pages pp. 3281–3286, Budapest, 2004.

[2] J. Bai, J.-Y. Nie, H. Bouchard, and G. Cao. Using query contexts in information retrieval. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 15–22. ACM Press, New York NY, 2007.

[3] S. Buttcher, C. Clarke, and I. Soboroff. The trec 2006 terabyte track. In *The Fifteenth Text REtrieval Conference The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*, 2006.

[4] P. Chirita, W. Nejdl, R. Paiu, and C. Kohlshuetter. Using odp metadata to personalize search. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2005.

[5] DMOZ. The Open Directory Project, 2008. URL `http://www.dmoz.org`.

[6] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the Eleventh International World Wide Web Conference*, 2002.

[7] D. Hiemstra. *Using Language Models for Information Retrieval*. PhD thesis, Center for Telematics and Information Technology, University of Twente, 2001.

[8] D. Hiemstra, S. Robertson, and H. Zaragoza. Parsimonious language models for information retrieval. In *Proceedings SIGIR 2004*, pages 178–185. ACM Press, New York NY, 2004.

[9] J. Kamps. Effective smoothing for a terabyte of text. In *The Fourteenth Text REtrieval Conference (TREC 2005)*. National Institute of Standards and Technology. NIST Special Publication, 2006.

[10] R. Kaptein, J. Kamps, R. LI, and D. Hiemstra. Experiments with positive, negative and topical relevance feedback. In *The Seventeenth Text REtrieval Conference (TREC 2008) Notebook.*, 2008.

[11] R. LI, R. Kaptein, D. Hiemstra, and J. Kamps. Exploring topic-based language models for effective web information retrieval. In *Proceedings DIR 2008*, 2008.

[12] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages pp. 558 – 565. ACM Press, New York NY, 2002.

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

[14] J. Trajkova and S. Gauch. Improving ontology-based user profiles. In *Proceedings of RIAO 2004*, 2004.

[15] TREC. Text REtrieval Conference, 2008. `http://trec.nist.gov/`.

[16] X. Wei and W. B. Croft. Investigating retrieval performance with manually-built topic models. In *Proceedings of RIAO 2007 - 8th Conference - Large-Scale Semantic Access to Content (Text, Image, Video and Sound)*, 2007.

[17] Wikipedia. The Free Encyclopedia, 2008. URL: `http://www.wikipedia.org`.

[18] Yahoo! Directory, 2008. URL `http://search.yahoo.com/dir`.

[19] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 49–56. ACM Press, 2001.