Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 12 (19): 2035-2043, 2012
ISSN 1812-5654 / DOI: 10.3923/jas.2012.2035.2043

Combination of Dependence, Relevance and Structure for Effective Web Retrieval

[1]I. Agavriloaei, [2]A. Rauber and [1]M. Craus
[1]Department of Computer Science and Engineering, Gheorghe Asachi, Technical University of Iasi, 27 Mangeron Blvd, Iasi, 700050, Romania
[2]Institute of Software Technology and Interactive Systems Favoritenstrasse 9-11/188, A-1040 Vienna, Austria

Abstract: This study presents a new four-model based approach regarding the *Ad hoc* retrieval task. The approach combines dependence and relevance models with filed-based techniques and the resulted retrieval lists are processed by applying a spam filtering phase. The parameters' estimation for each model is accomplished by optimizing the Mean Average Precision (MAP). All experiments are performed on the Category B part of ClueWeb09 collection. Experimental results indicate that significant improvements are possible with regards to the retrieval effectiveness by adding a field-based modeling and a spam filtering to dependence and relevance models. Also, the results of this study show that the additive feature between methods is conserved.

Key words: *Ad hoc* retrieval, term dependence model, query expansion, field-based model, spam filter, performance evaluation

INTRODUCTION

Ad hoc retrieval is an important research topic and its applications cover a wide range of Information Retrieval (IR) problems. *Ad hoc* retrieval investigates the performance of finding a static set of documents that are topically relevant using some given standard and unseen topics ranked by decreasing probability of relevance. Document relevance is considered to be independent of other documents that appear before it in the ranking list (Clarke *et al*., 2004).

Language Modeling (LM) is a principled statistical model that has proven to work well in different areas such as Speech Recognition, Machine Translation and Information Retrieval. In IR, the model was proposed by Ponte and Croft (1998) and it was followed by a number of studies and developments dedicated to this topic (Croft and Lafferty, 2003). Some of these approaches have led to the Lemur Project that will be used in the experiments of this study.

Many field-based retrieval methods (Metzler *et al*., 2004) that were applied in the past to Web collections are rather simplistic, isolated and have a poor contribution on the system performance which is not the case for the Indri's default field retrieval model as shown by Nguyen and Callan (2010) and Zhao and Callan (2008). Moreover, all of the initial approaches were only tested with much smaller Web collections, like the WT10 g and the GOV2. This is problematic because these collections have some particular features such as a specific domain (the. gov domain is very different in nature from the. com domain) or fewer incoming links per page. Also, they contain little redundancy and do not reflect the real Web, where there is lots of spam and the pages are much more interconnected. Koolen and Kamps (2010) concluded that a good Web collection needs to be sufficiently large, very various in terms of domains and have sufficiently high inter- and intra-server link densities. The new collection ClueWeb09 which reflects a fraction of a commercial search engine, comes with these needed features. Also, in the last years of TREC it was observed that the Category B generally contains higher quality documents than the rest of the collection and has a higher prior probability of relevance (Smucker *et al*., 2009) due to the fact that it incorporates the Wikipedia corpus and it includes only English pages.

The Web can be viewed as a system of structured documents where they are connected through hyperlinks. The difference between the Web search and the traditional text search is given by these new connections (and the text around them) and the way that the information is perceived, understood and used by the final user.

In order to build a highly effective retrieval system that emphasizes the important items of information, several studies at TREC (Allan *et al*., 2008; Bendersky *et al*., 2010; Clarke *et al*., 2011; He *et al*., 2010; Smucker *et al*., 2009) have used the Markov Random Field

Corresponding Author: Ioan Agavriloaei, Department of Computer Science and Engineering, Gheorghe Asachi, Technical University of Iasi, 27 Mangeron Blvd, Iasi, 700050, Romania

retrieval model (Metzler and Croft, 2005) to reflect term dependencies. In previous TREC years it was shown that this technique is very effective and that significant improvements are possible by using full dependencies with shorter queries, especially on Web collections which are less homogenous.

In most cases the user queries are usually short, very general and ambiguous. Therefore, an automatic query expansion mechanism can improve this weak point and can provide some increase in performance. A formal probabilistic approach to estimate a relevance model with no training data was proposed by Lavrenko and Croft (2001). In the last three years of the TREC Web Track, the query expansion was explored by different approaches. While several studies used a specific and potentially much cleaner source such as Wikipedia (He *et al*., 2010; McCreadie *et al*., 2010; Nguyen and Callan, 2010) or results from commercial search engines (Smucker *et al*., 2009), others tried to optimize the query by using pseudo-relevant documents of the same collection (Algarni *et al*., 2010) or even larger ones (Diaz and Metzler, 2006). The reason for such different treatments was that the initial results in finding representative terms have the risk of including noisy terms such as spam terms. The usage of the Wikipedia corpus has the advantage of being relatively clean, very objective, with topics that have a good coverage and are presented in a scientific manner. But, the Wikipedia articles do not often cover human subjectivity and some particular viewpoints over the topics on which the regular Web user might be interested in. Along with this strictness, the corpus adds some specific terms such as "article", "edit", "Wikipedia" which is not desirable. The findings of these methods was that spam is an important issue in the *Ad hoc* task and that optimization techniques help to further the retrieval performance which is assumed to potentially reduce spam in the top ranked results.

There is substantial previous research (Agavriloaei *et al*., 2011; He *et al*., 2010; Nguyen and Callan, 2010) about combining evidence from different fields of a document (e.g. title, inlink and body text of HTML). A generative retrieval model for structured documents is proposed by Zhai and Lafferty (2001), where structured queries are used to retrieve structured documents. The performance of the best structured retrieval models is not worse and sometimes outperforms the keyword retrieval models when the structured queries are accurate enough. Also, combining the field-based model with the PageRank model proved to be unsatisfactory (Chen *et al*., 2010).

Spam detection and filtering was once again recognized as an important component of the Web retrieval in the TREC Web Track 2009. The Waterloo Spam Ranking proposed by Cormack *et al*. (2011) showed that the use of spam filtering or re-ranking significantly improves the retrieval effectiveness of practically all the TREC Web Track 2009 participants. Thus, this shows that the presence of spam and other low-quality pages substantially influence the overall results.

In this study are explored several retrieval techniques proved to work well in the past-the dependence model and the relevance model. These methods are combined providing a strong baseline to which now it is added a retrieval method based on fields. The study analyses how the performance of the *Ad hoc* retrieval is changed once the field-based model is added to the dependence and relevance models in LM. This analysis helps to gain insight into the role that Web structured documents (that contain HTML fields or defined annotations) can play in increasing the effectiveness of current retrieval systems. Moreover, the study investigates the effect of spam over the results. The aim of this study is to improve the Mean Average Precision (MAP) and the precision at 10 (P@10). The experiments of the study were carried out on the Category B subset of the ClueWeb09 collection.

FOUR-MODEL BASED APPROACH

Here, it is described a new proposed approach to the *Ad hoc* retrieval task. The study considers four major constituents of IR (Information Retrieval) on the Web scale: Query term dependence, query expansion, spam filtering and field-based retrieval. The first three features have been approved as a significant part of the Web retrieval systems by past TREC participants while the fourth one is not as well explored when it is combined with the former models.

Modeling term dependence: The Dependence Model (DM) (Metzler and Croft, 2005) is a formal framework for rewriting a query using Markov Random Fields to get a better representation of the original query. The model allows using several types of text evidence (features) such as occurrences of single terms, ordered phrases (ordered windows) and unordered phrases (unordered windows) as real-valued feature functions. These techniques are outlined in Table 1.

In order to address the dependence model, the study makes use of the full dependence model variant proposed by Metzler and Croft (2005). The full dependence model goes beyond single and adjacent query terms and attempts to incorporate dependencies between every subset of query terms and thus also consists of both exact phrase matches and proximity matches. A three term query like "who invented music", would have three single terms and the following phrases: "who invented",

Table 1: Techniques addressed in the study's experiments

Technique	Description
Stopping	Standard stopping list of 418 stop words included in Indri
Stemming	Krovetz stemmer (Krovetz, 1993)
Term smoothing	Dirichlet smoothing (Zhai and Lafferty, 2004) with μ parameter tuned for each method
Ordered windows	Ordered proximity windows - terms must appear ordered with no other terms between each occurrence, scored for every sequence of 2 or 3 query terms (Metzler and Croft, 2005)
Unordered windows	Unordered proximity windows-terms must appear in any order with a maximum windows size of four times the number of terms being scored, for every combination of 2 or 3 query terms (Metzler and Croft, 2005)
Pseudo-relevance	Indri's relevance modeling with 20 terms selected feedback automatically from the top 10 documents, weighting the original query as 0.5 and the expanded as 0.5
Field-based windows	A weighted model implemented to build field dependent queries to retrieve structured text
Spam Filtering	Spam filtering as described by Cormack *et al.* (2011) as a post retrieval phase

"invented music", "who music" and "who invented music". The terms, the ordered and unordered windows are each assigned different weights (λ_T for terms, λ_O for ordered windows and λ_U for unordered windows) which are tuned to optimize the impact on the system's effectiveness. In the Indri language model the query can be rewritten as:

```
#weight (
    λT #combine (who invented music)
    λO #combine (
        #1 (who invented)
        #1 (who music)
        #1 (invented music)
        #1 (who invented music)
    )
    λU #combine (
        #uw8 (who invented)
        #uw8 (who music)
        #uw8 (invented music)
        #uw12 (who invented music)
    )
)
```

where, the weights for the different query parts are determined by Metzler and Croft (2005) and have as a goal the optimization of the mean average precision.

One of the main outcomes of this model is that these three features capture just general aspects of the text since their assigned weights ($\lambda_T = 0.8$, $\lambda_O = 0.1$, $\lambda_U = 0.1$) generalize well on multiple collections. This study assumes that more domain-specific features may bring further improvements.

Modeling relevance and query expansion: The Relevance Model (RM) (Lavrenko and Croft, 2001) is an effective formal method for estimating probabilities of new terms in the unknown set of documents relevant to the query terms with no training data. The method typically begins with an initial query, does some processing and then returns a list of expansion terms. These terms are then added to the original query and the system is rerun using the new expanded query. Given the fact that the Web collections can be very noisy and that the goal is to improve the estimation of the new query model by avoiding this noise, some researchers selected the new terms from a much less noisy external collection such as Wikipedia or results from search engines, results that are already cleaned. For the experiments, in order to study the effect of the pseudo-relevance feedback on the ClueWeb09, we chose to use the whole collection as the source for the expansion terms.

For this purpose, the study creates the RM query terms by selecting the top 20 terms from the top 10 pages and mixes them with the original query terms. This is done by using Indri's pseudo-relevance feedback mechanism as an adaptation of the Lavrenko proposed model (Lavrenko and Croft, 2001). Even more, each part of the mix model was weighted properly in order to maximize the MAP. In the Indri language modeling we can rewrite the obtained query as:

```
#weight (
    λDM dependence_model
    λRM pseudo-relevance_model
)
```

where, the weights for each model are tuned to optimize the effectiveness.

Modeling specific fields: Since, the dependence model covers general aspects of a query and the relevance model provides an expansion of terms of the same type, it is a challenge to investigate if more specific features can bring some improvements in the system's effectiveness. The Web documents are structured pages and contain elements defined by author(s) or annotations assigned by other people or processes (Zhai and Lafferty, 2001). The field-based retrieval model (FM) could be a solution for a more specific focus on the Web collection characteristics, e.g,. BM25F formula (Robertson *et al.*, 2004). Furthermore, by aggregating a field-based model with the previous two models we also expect the system's effectiveness to increase. With these premises a combined model can be build that incorporates dependence, relevance and field-based aspects. Therefore, the study can take into account five specific fields (i.e., body, title, heading, inlink, strong) which were indexed at the index time. In Indri, for example, now the query "horse hooves" can be rewritten to restrict the retrieval both on the field title as well as on the field body:

```
#weight(
    λDM #weight(
        λT #weight(
            λTITLE #combine (horse.(title) hooves.(title) )
            λBODY #combine( horse hooves )
        )
        λO #weight(
            λTITLE #combine(#1(horse hooves).(title))
            λBODY #combine(#1(horse hooves))
        )
        λU #weight(
            λTITLE #combine(#uw8(horse hooves).(title))
            λBODY #combine(#uw8(horse hooves))
        )
    λRM pseudo-relevance_model
)
```

where, the optimal weight for the fields title and body is tuned using the same approach as with the dependence model.

Modeling spam filter: The ClueWeb09 collection has a lot of spam documents thus, the results may also contain many spam documents that would have a negative influence on the system's performance. The study uses the spam filtering as described by Cormack *et al.* (2011) to delete spam documents from the retrieved document list as a post retrieval step. This process consists in assigning a percentile score p(D) to each of the documents in the ClueWeb09 collection. Then, the score is used in combination with a threshold to classify a page as "spam" or "not spam", or may be used to rank the page with respect to others by "spamminess".

From the Category B it was excluded 10% of the documents computed by the fusion method which are most likely "spam"; this rate was found to optimize the MAP in the optimization experiments with the Web Track 2009 queries.

EXPERIMENTS

First, it is briefly described the indexing of the ClueWeb09 collection, then how the Dirichlet smoothing parameters, the weight values of specific fields and the optimal percent value for the spam filter is tuned. Afterwards, optimal parameter values are used to run each of the above methods and their combinations.

Data set, index and runs: For experiments it was used the ClueWeb09 Category B dataset which is a sample of 50 million English pages from the larger ClueWeb09 collection. This collection contains many of the most important Web pages since the Web pages were crawled based on a large seed of URLs with a high PageRank. This idea is also supported by the findings of the TREC 2010 which suggest that the Category B subset generally contains higher quality documents than the rest of the collection (Clarke *et al.*, 2011). In all runs the parameters are tuned with the TREC Web Track 2009 queries and the experiments are performed with queries provided by the TREC Web Track 2010.

The system chosen for experiments is Indri 5.1 from the Lemur Project that was modified to index the ClueWeb09. An index on one machine with 1.4 GHz CPU, 30 GB of RAM and approximately 5 TB of free disk storage was built. Building the index for Category B took about 68 h. During the indexing process, a stop word removal and stemming with the Krovetz (1993) stemmer was performed because Krovetz stemmer does not over generalize as much as the Porter stemmer does. Also, some HTML tags was indexed as fields: the tag title as a title field, the tags h1, h2, h3 and h4 as a heading field, the tag anchor as an inlink field and the tags strong and bold as a strong field. Except the field title, the entire document content is referred as the body field. Thus, the fields can be searchable through the Indri query language. Finally, the obtained index contains the whole document text and specified fields.

Armstrong *et al.* (2009) identified that there are six techniques that provide some improvements on the retrieval performance. These techniques are presented in the first six rows of Table 1 with some changes and adaptations for experiments. It was observed that there is an additive relationship between the number of used techniques and the achieved retrieval effectiveness. We expected that, by adding a specific field-based method and a spam filtering to the previous methods, the additive feature will be conserved and the retrieval effectiveness will increase properly.

All retrieval experiments were performed using Indri as well. The Indri query language is used to create full dependence model queries, as well as to perform query expansions and field-based queries. The first two techniques from Table 1, stopping and stemming are applied to all the methods since, they were done at index time. The smoothing is performed on each model according to the tuning phase. Each of the described methods is run separately, i.e., the DM, RM and FM runs. Then, we combine them one by one leading us to the DM+RM run also referred to as RM3 (Smucker *et al.*, 2009), the DM+FM and RM+FM runs and in a final experiment, we aggregate all the three previous techniques resulting the DRF3 run. Spam filtering process is applied as an optional post retrieval step.

Tuning parameters: The method selected to find the optimal values of the parameters was the direct maximization of the Mean Average Precision (MAP).

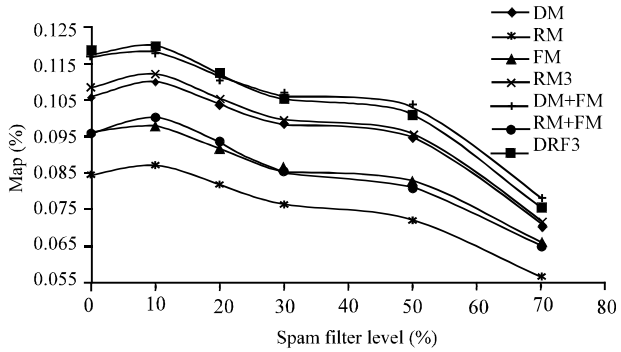

Fig. 1: Mean average precision (MAP) evolution for each model and combination against the Web Track 2009 queries on different spam filtering levels

Table 2: Optimal smoothing Dirichlet parameter for the Web Track 2009 and 2010 queries in the ClueWeb09 Cat. B

	Dirichlet (μ)						
Queries	DM	RM	FM	RM3	DM+FM	RM+FM	DRF3
wt2009	3000	1500	3000	2000	3000	2500	2000
wt2010	4000	2000	4000	2500	4000	3000	2500

Since the models addressed in this study have few parameters, it is possible to do a linear search to reach the optimal smoothing parameter, the weight of expanded query and the spam filter level. It is possible also to perform a simple coordinate-level hill climbing search to find the optimal fields weight values by starting at the default parameter settings ($\lambda_{BODY} = 1$, $\lambda_{TITLE} = 0$, $\lambda_{HEADING} = 0$, $\lambda_{INLINK} = 0$, $\lambda_{STRONG} = 0$).

The smoothing is a major problem in the LM estimation. Smoothing techniques try to balance the probability of terms which appear in a document with the ones which are missing. Some works (Zhai and Lafferty, 2001) concluded that a simple smoothing scheme based on the Dirichlet priors gives a very good performance, due to the way that it effectively normalizes the query terms for the document length. All experiments use the Indri language model approach and the Dirichlet smoothing rule. This study has also chosen to smooth the test collection by this rule as (Zhai and Lafferty, 2001) showed that the Dirichlet smoothing significantly outperforms the Jelinek-Mercer smoothing on the structured retrieval tasks. To learn the optimal smoothing parameter the study performed a linear search by tuning it between 1000 and 8000 with a step of 500 and aiming to maximize the mean average precision. The optimal values are listed in Table 2. As discussed further below, the reason for different smoothing parameter values being optimal in the Web Track 2009 and 2010 queries may be attributed to the fact that the Web Track 2010 topics contain 12% more single-term queries than the Web Track 2009 topics.

Table 3: Optimal weight values for specific fields for the Web Track 2009 and 2010 queries in the ClueWeb09 Category. B

Queries	Body	Title	Heading	Inlink	Strong
Raw values					
wt2009	0.95	0.12	0.18	0.14	0.17
wt2010	1.12	0.21	0.13	0.15	0.17
Normalized values					
wt2009	0.6	0.08	0.12	0.09	0.11
wt2010	0.63	0.12	0.07	0.08	0.1

In Table 2 it is provided parameter tuning results for both the Web Track 2009, 2010 queries to show the differences in the optimal parameter values. Further, for system evaluation, we will obviously use the optimal parameter setting obtained on the Web Track 2009 queries (training set), performing the evaluation on the Web Track 2010 queries (test set).

To find the best weight for the expanded query the study applied the same principle as was done for the Dirichlet smoothing parameter. The weight of the expanded query was tuned between 0 and 1 with a step of 0.1. The value that provided the best results for each run that involved a query expansion was found to be 0.5. The optimal values according to the Web Track 2009 queries are then used throughout the remainder of the experiments.

In order to determine the weight values for specific fields it was performed a hill-climbing search having as objective the maximization of the MAP. During this search the study use the Dirichlet smoothing parameter with a default value, $\mu = 2500$. The optimal found weight values for specific fields (body, title, heading, inlink and strong) that maximize the MAP are listed in Table 3. By the field body we refer to the document content that is bordered by the body HTML tag.

The optimal spam filter level is determined in the same manner as the smoothing parameter. The value that optimizes the MAP was found to be 10%. In Fig. 1 we illustrate the evolution of the MAP against the Web Track 2009 queries and the different spam levels for each of the mentioned runs.

Table 4: Mean Average Precision (MAP) and precision at 10 (P@10) using optimal parameter settings for each model or combination and optimal spam filter percentile score

	No spam filter				Spam Filter at 10%			
Runs	MAP	ΔRM3	P@10	ΔRM3	MAP	ΔRM3	P@10	ΔRM3
Indri.base	0.0917	-12.95%**	0.2062	-20.8%**	0.1004	-12.1%*	0.2542	-15.28%*
DM	0.0972	-7.6%	0.2354	-9.6%	0.1064	-6.86%	0.2833	-5.56%
RM	0.0941	-10.65%**	0.2250	-13.6%	0.1000	-12.42%**	0.2625	-12.5 %*
FM	0.1025	-2.6%	0.2583	-0.8%	0.1095	-4.14%	0.3083	+2.78%
RM3 (baseline)	0.1053	-	0.2604	-	0.1142	-	0.3000	-
DM+FM	0.1091	+3.6%	0.2979	+14.4%	0.1151	+0.75	0.3312	+10.42%
RM+FM	0.1048	-0.46%	0.2792	+7.2%*	0.1089	-4.64%	0.3062	+2.08%
DRF3	**0.1181**	+12.14%**	**0.3188**	+22.4%**	**0.1246**	+9.03%*	**0.3438**	+14.58%*

The highest score for each metric are in bold, ΔRM3: Relative improvement of each metric with respect to RM3 baseline, *,**Statistically significant at confidence levels of 90 and 95%, respectively

RESULTS

Since, the study are dealing with three different models, by combining them, is obtained seven unique combinations that are run against the collection and the resulting runsets are scored using the MAP and the P@10 before and after the post retrieval spam filtering stage. The baseline was set to be the combination of the dependence model with the relevance model, i.e., the RM3 run. Also, a standard Indri run is provided as a bottom baseline, i.e. Indri.base run. The standard Indri model is based on a combination of the language modeling (Ponte and Croft, 1998) and inference network (Turtle and Croft, 1991). The outcomes are shown in Table 4. Statistical significance was computed using Paired Randomization Test (Smucker *et al*., 2007).

The DRF3 run provides the best results and leads to significant improvements in the MAP and the P@10 with respect to the RM3 run showing that the Web document structure can be used to improve the *Ad hoc* search. All DRF3 results are statistical significant. Before spam filtering phase the confidence level is at 95% while after removing spam documents the confidence level decrease to 90%. The spam filtering phase brings a substantial improvement even to the best run of the study: the MAP increases by 5.5% from 0.1181-0.1246 and the P@10 grows with 7.84% from 0.3188- 0.3438.

In the DM+FM run, the improvement is smaller in MAP (3.6 and 0.75%) and bigger in P@10 (14.4 and 10.42%) but without statistical significance. In the RM+FM run the results are worse than the baseline, with statistical significance only for P@10 before the spam filtering phase.

To put the study's results into perspective, the best run was incorporated within the top ten results of the best official submissions of the TREC 2010. The results are shown in Table 5. A direct comparison with the results presented in the TREC 2010 (Clarke *et al*., 2011) might be a bit forced due to a number of reasons. The primary effectiveness measure for the *Ad hoc* task at the TREC

Table 5: Top 10 results of the TREC 2010 web track *Ad hoc* task against the best result ordered by the mean average precision

Runs	MAP	P@20	ERR@20
cmuWiki10	0.157	0.400	0.112
umassSDMW	0.148	0.484	0.138
irra10b	0.133	0.443	0.126
IvoryL2Rb	0.133	0.379	0.134
uogTrA42	0.127	0.411	0.127
DRF3	**0.1246**	**0.305**	**0.112**
THUIR10QaHt	0.112	0.331	0.128
msrsv3	0.082	0.344	0.166
UMa10IASF	0.080	0.293	0.119
TREC.baseline	0.069	0.374	0.164
UAMSA10mSF30	0.043	0.237	0.110

The results of the study are in bold

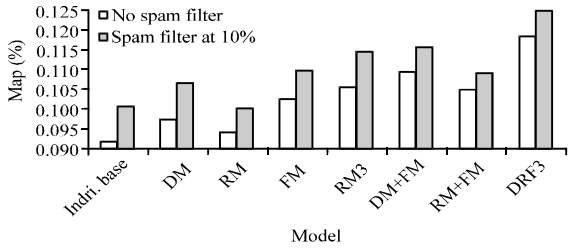

Fig. 2: Mean average precision (MAP) evolution as a function of combined models for the Indri environment running

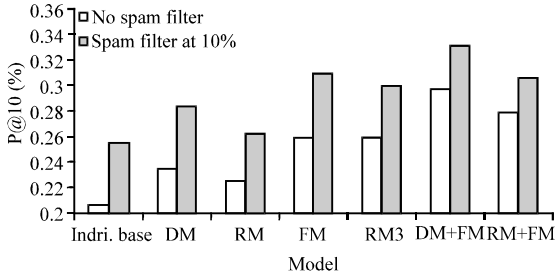

Fig. 3: Precision at 10 (P@10) evolution as a function of combined models for the Indri environment running against the ClueWeb09 Cat. B collection before and after the spam filtering phase

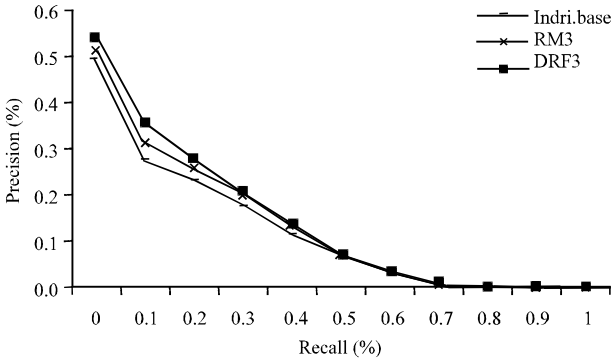

Fig. 4: Averaged 11-point precision-recall graph after spam filtering phase against the Web Track 2010 queries. The MAP for DRF3 run is 0.1246

Table 6: Mean average precision (MAP) and Precision at 10 (P@10) for each specific field in the field-based model. The field body is identical with the Indri.base run

	No Spam Filter		Spam Filter at 10%	
Runs	MAP	P@10	MAP	P@10
title	0.0191	0.0896	0.0195	0.0938
heading	0.0208	0.1146	0.0214	0.1208
inlink	0.0205	0.1646	0.0218	0.1832
strong	0.0152	0.1146	0.0158	0.1125
body	0.0917	0.2062	0.1004	0.2542

2010 was Expected Reciprocal Rank (ERR) while this study aimed to optimize the MAP. Also, the baseline was beaten just by one submitted run and most of the participants made use of the entire ClueWeb09 collection. Even so, a conclusion can be drawn about the effectiveness of the approach which is above average in terms of the MAP.

Figure 2 and 3 plot the MAP and the P@10 scores achieved by the different Indri configurations in the Indri environment, as a function of the combined techniques.

We can observe that the additive feature is conserved and the retrieval effectiveness increases accordingly when more models are combined.

Figure 4 plots the precision-recall graph by using the 11 cutoff values. We can notice that the DRF3 run is superior and indicate the best performance in terms of precision.

The Web Track 2009 and the Web Track 2010 queries contain many single-term queries: the former has 17 single-term queries (34%) while the latter has 23 single- term queries (46%). Taking into account how the dependence and the relevance models are designed, we can state that they perform well with queries that have at least two terms. In the DM a single-term query reduces the model to a standard Indri language modeling which is the same as the query likelihood. Moreover, by modeling a single-term query with the RM we can obtain some very general and ambiguous expanded query terms which could have a higher potential harmful impact on the system performance. The field-based model does not consider the dependencies between the terms of the query therefore, it is not affected by the query length. In Fig. 2 and 3, this advantage leads to a better MAP and P@10 for the FM at the expense of the first two models when they are run individually. When the models are put together the poor performance of the first two models with single-term queries is compensated by a good coverage of this feature given by the FM.

The field-based model contains the title, heading, inlink, strong and body as fields. The results per field are shown in Table 6. The weight values for each of the above fields that optimize the MAP are listed in Table 2. Taken separately, the inlink field has a height precision which was also observed by others (Koolen and Kamps, 2010; Nguyen and Callan, 2010). However, the MAP obtained by the first four fields is much lower compared to the one obtained by the body field. The reason for this score could be the fact that not all pages contain these fields and that the body field, besides the fact that includes the fields heading, inlink and strong, incorporates the rest of the useful information. Furthermore, when we apply a spam filter, the P@10 of the body field improves substantially while the title, heading and inlink precision improve much less. Also, the precision of strong field decreases after the spam filtering stage. This indicates that the first four fields have less spam while the body field is very predisposed to spam.

Overall, the DRF3 run that combines all four models has excellent performance in terms of the effectiveness of the *Ad hoc* retrieval task. To conclude, it can be observed that the use of the spam filtering significantly improves the accuracy of all study results.

CONCLUSION

In this study a four-model based approach for *Ad hoc* retrieval was proposed. The proposed method (named as DRF3) directly optimize the Mean Average

Precision (MAP) over a set of queries. The study analyzed the proposed approach using the ClueWeb09 Category B collection. In the study's experiments, the four-model based approach significantly outperformed the standard Indri method and the RM3 method.

Regarding the effectiveness of the performance, the results were quite encouraging given the fact that the experiments did not rely on any external resources, such as specific collections or results from commercial search engines for query expansions. This shows that the field-based retrieval model is competitive when it is added to the dependence and relevance models, despite the fact that it is simple, it can be modeled easily using the Indri modeling language and it involves tuning parameters. But to achieve this improvement, parameters have to be carefully tuned.

Dealing with the field-based technique and aggregating it with the dependence and the relevance model helps to establish the generality of its benefit. In addition, the methods that include the field-based technique performed well for many single-terms queries, where both the dependence model and the retrieval with expanded queries yielded non relevant documents at the top ranks. Moreover, the spam filtering has proved its usefulness by improving the accuracy of all results, even if the Category B collection contains higher quality documents.

Thus, the experimental results of this study have confirmed the assertion made by Armstrong *et al*. (2009) by adding the field-based and spam filtering model to the dependence and relevance model, the additivity of improvements is conserved and the system's effectiveness increase properly.

There are a number of problems that need to be addressed in future research, including finding a way to refine the system by including spam filtering as a pre-retrieval step, by reducing the noise from the query expansion and by finding a way to effectively employ the PageRank scores.

ACKNOWLEDGMENTS

This study was realized with the support of EURODOC "Doctoral Scholarships for research performance at European level" project, financed by the European Social Found and Romanian Government and the Institute of Software Technology and Interactive Systems (ISIS) at Vienna University of Technology, Austria.

REFERENCES

Agavriloaei, I., A. Alexandrescu and M. Craus, 2011. Improving web clustering through a new modeling for web documents. Proceedings of the 15th International Conference on System Theory, Control and Computing (ICSTCC), October 14-16, 2011, Sinaia, Romania, pp: 1-6.

Algarni, A., Y. Li and X. Tao, 2010. Mining specific and general features in both positive and negative relevance feedback. Proceedings of the 19th Text REtrieval Conference: Relevance Feedback Track (TREC), November 16-19, 2010, Gaithersburg, USA, pp: 1-9.

Allan, J., J.A. Aslam, V. Pavlu and E. Kanoulas, 2008. Million Query Track 2008 overview. Proceedings of the 17th Text REtrieval Conference (TREC), November 18-21, 2008, Gaithersburg, USA.

Armstrong, T., A. Moffat, W. Webber and J. Zobel, 2009. Improvements that don't add up: *Ad hoc* retrieval results since 1998. Proceedings of the 18th ACM Conference on Information and Knowledge Management, November 2-6, 2009, Hong Kong, pp: 601-610.

Bendersky, M., D. Fisher and W.B. Croft, 2010. TREC 2010 web track notebook: Term dependence, spam filtering and quality bias. Proceedings of the 19th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Chen, X., Z. Peng, J. Wang, X. Yu, Y. Liu, H. Xu and X. Cheng, 2010. ICTNET at web track 2010 *Ad hoc* task. Proceedings of the 19th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Clarke, C., N. Craswell and I. Soboroff, 2004. Overview of the TREC 2004 terabyte track. Proceedings of the 13th Text REtrieval Conference (TREC), November 16-19, 2004, Gaithersburg, USA.

Clarke, C., N. Craswell, I. Soboroff and G. Cormack, 2011. Overview of the TREC 2010 web track. Proceedings of the 19th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Cormack, G.V., M.D. Smucker and C.L.A. Clarke, 2011. Efficient and effective spam filtering and re-ranking for large web datasets. Inform. Retrieval, 14: 441-465.

Croft, W. and J. Lafferty, 2003. Language Modeling for Information Retrieval. Kluwer, Netherlands.

Diaz, F. and D. Metzler, 2006. Improving the estimation of relevance models using large external corpora. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 6-11, 2006, Seattle, USA, pp: 154-161.

He, J., K. Balog, K. Hofmann, E. Meij, M. de Rijke, E. Tsagkias and W. Weerkamp, 2010. Heuristic ranking and diversification of web documents. Proceedings of the 18th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Koolen, M. and J. Kamps, 2010. The importance of anchor text for *Ad hoc* search revisited. Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, July 19-23, 2010, Geneva, Switzerland, pp: 122-129.

Krovetz, R., 1993. Viewing morphology as an inference process. Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, June 27-July 1, 1993, Pittsburgh, USA., pp: 191-202.

Lavrenko, V. and W. Croft, 2001. Relevance based language models. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and development in Information Retrieval, September 9-12, 2001, New Orleans, USA, pp: 120-127.

McCreadie, R., M. Craig, O. Iadh, P. Jie and L. Rodrygo, 2010. University of Glasgow at TREC 2009: Experiments with terrier - blog, entity, million query, relevance feedback and web tracks. Proceedings of the 18th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Metzler, D. and W. Croft, 2005. A Markov random field model for term dependencies. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 15-19, 2005, Salvador, Brazil, pp: 472-479.

Metzler, D., T. Strohman, H. Turtle and W. Croft, 2004. Indri at TREC 2004: Terabyte track. Proceedings of the 13th Text REtrieval Conference (TREC), November 16-19, 2004, Gaithersburg, USA.

Nguyen, D. and J. Callan, 2010. Combination of evidence for effective web search. Proceedings of the 19th Text REtrieval Conference (TREC), November 16-19, 2010, Gaithersburg, USA.

Ponte, J.M. and W.B. Croft, 1998. A language modeling approach to information retrieval. Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, August 24-28, Melbourne, Australia, pp: 275-281.

Robertson, S., H. Zaragoza and M. Taylor, 2004. Simple BM25 extension to multiple weighted fields. Proceedings of the 13th ACM International Conference on Information and Knowledge Management, May 17-20, 2004, NY., USA., pp: 42-49.

Smucker, M.D., C.L.A. Clarke and G.V. Cormack, 2009. Experiments with ClueWeb09: Relevance feedback and web tracks. Proceedings of the 18th Text Retrieval Conference, November 17-20, 2009, Gaithersburg, USA.

Smucker, M.D., J. Allan and B. Carterette, 2007. A comparison of statistical significance tests for information retrieval evaluation. Proceedings of the 16th ACM Conference on In-formation and Knowledge Management, November 6-9, 2007, Lisbon, Portugal, pp: 623-632.

Turtle, H. and W.B. Croft, 1991. Evaluation of an inference network-based retrieval model. ACM Trans. Inf. Syst., 9: 187-222.

Zhai, C. and J. Lafferty, 2001. A study of smoothing methods for language models applied to *Ad hoc* information retrieval. Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, September 9-13, 2001, New Orleans, USA., pp: 334-342.

Zhao, L. and J. Callan, 2008. A generative retrieval model for structured documents. Proceedings of the 17th ACM Conference on Information and Knowledge Management, October 26-30, 2008, Napa Valley, USA., pp: 1163-1172.