Université de Neuchâtel
Faculté des Sciences
Institut d'Informatique

Domain Specific Information Retrieval Social Science, Blogsphere and Biomedicine

par

Claire Fautsch

Thèse

présentée à la Faculté des Sciences
pour l'obtention du grade de Docteur ès Sciences

Acceptée sur proposition du jury:

Prof. Jacques Savoy, directeur de thèse
Université de Neuchâtel, Suisse

Prof. Pascal Felber, rapporteur
Université de Neuchâtel, Suisse

Prof. Patrick Ruch, rapporteur
Hôpitaux Universitaires de Genève, Suisse

Prof. Rolf Ingold, rapporteur
Université de Fribourg, Suisse

Soutenue le 22 Octobre 2009

Université de Neuchâtel

FACULTE DES SCIENCES
Secrétariat-Décanat de la faculté
- Rue Emile-Argand 11
- CP 158
- CH-2009 Neuchâtel

IMPRIMATUR POUR LA THESE

Domain Specific Information Retrieval : Social Science, Blogsphere and Biomedicine

Claire FAUTSCH

UNIVERSITE DE NEUCHATEL

FACULTE DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel,
sur le rapport des membres du jury

MM. J. Savoy (directeur de thèse), P. Felber,
R. Ingold (Université de Fribourg) et
P. Ruch (HEG et Université de Genève)

autorise l'impression de la présente thèse.

Neuchâtel, le 30 octobre 2009

Le doyen :
F. Kessler

UNIVERSITE DE NEUCHATEL
FACULTE DES SCIENCES
Secrétariat - décanat de la faculté
Rue Emile-Argand 11 - CP 158
CH-2009 Neuchâtel
Felix Kessler

"Do not go where the path may lead, go instead where there is no path and leave a trail."

Ralph Waldo Emerson (1803-1882)

Abstract

Keywords: Information Retrieval, Domain-Specific, Blogsphere, Evaluation, Information Retrieval Models

Nowadays information retrieval is widely known and used in the context of online web search engines. Information retrieval however also presents many other fields of applications, one of which is domain-specific information retrieval. This thesis summarizes our work in this field by presenting a selection of our research papers.

In the presented work the challenges of information retrieval in three different domains, namely Blogsphere, social science and biomedicine and our solutions to improve retrieval effectiveness in these domains are presented. For each domain we evaluate the standard retrieval procedures first and then adapt them in order to meet domain-specific issues. We finally compare and discuss our results by participating in various evaluation campaigns.

Furthermore we present an approach for opinion mining in blogs as well as a proposal for a domain independent retrieval model taking account of domain-specific information. Finally we also present a more general study on algorithmic stemmers and morphological analysis for the English language.

Résumé

Mots-Clés: Recherche d'Information, Domaine Spécifique, Blogosphère, Evaluation, Modèles de Recherche d'Information

Aujourd'hui la recherche d'information est bien connue et utilisée dans le contexte des moteurs de recherche en ligne. Or la recherche d'information présente aussi beaucoup d'autres applications, tel que la recherche d'information dans les domaines spécifiques. Cette thèse résume nos travaux effectués dans ce champ en présentant une sélection de nos articles scientifiques.

Dans ce travail les défis de la recherche d'information dans trois domaines différents - la Blogosphère, la science sociale et la biomédecine - ainsi que nos solutions pour améliorer la recherche d'information dans ces domaines sont présentés. Pour chaque domaine on évalue d'abord les approches standards avant de les adapter afin de satisfaire aux besoins spécifiques du domaine. Enfin on présente, compare et discute nos résultats en participant à diverses campagnes d'évaluation.

En plus on a présenté une approche pour la détection d'opinions dans des blogs ainsi qu'une proposition pour un modèle pour la recherche d'information dans les domaines spécifiques, indépendant du domaine tout en tenant compte des spécificités du domaine. Finalement on présente une étude plus générale sur les enracineurs et l'analyse morphologique pour la langue anglaise.

Kurzfassung

Schlüsselwörter: Informationssuche, Domän Spezifisch, Blogs, Evaluation, Modelle der Informationssuche

Heutzutage ist Informationssuche vor allem bekannt durch die Benutzung von Suchmaschinen bei der Websuche. Allerdings hat die Informationssuche ein weitaus grösseres Anwendungsspektrum, unter anderem die Informationssuche in spezifischen Domänen. Diese Dissertation fasst unsere Arbeit in diesem Bereich zusammen.

In der hier vorgestellten Arbeit werden die Herausforderungen der Informationssuche in drei verschiedenen Gebieten - Blogsphere, Sozial Wissenschaft und Biomedizin - ausgearbeitet und anschliessend Lösungsansätze vorgeschlagen um die Informationssuche in diesen Domänen zu verbessern. Zuerst werden gewöhnliche Prozeduren der Informationssuche ausgewertet und dann angepasst um den spezifischen Charakteristiken gerecht zu werden. Anhand der Teilnahme an diversen Evaluationskampagnen werden schlussendlich die erzielten Resultate diskutiert und verglichen.

Des Weiteren wird eine Methode zum Erfassen von Meinungen in Blogs sowie ein Modell zu Informationssuche in spezifischen Domänen vorgestellt. Schlussendlich wird noch auf eine allgemeine Studie von Stemming und morphologischer Analyse für die Englische Sprache eingegangen.

Acknowledgements

Completing this work has been an enriching experience from scientific and human points of view. There are many people I would like to thank for their contributions, help and support throughout the last three years.

First of all I would like to thank my supervisor Prof. Jacques Savoy for offering me the opportunity to complete this Ph.D. thesis and discover the world of information retrieval and for his guidance, support and help.

I would also like to thank the members of the jury, namely Prof. Rolf Ingold (University of Fribourg), Prof. Patrick Ruch (University and Hospitals of Geneva) and Prof. Pascal Felber (University of Neuchâtel), for accepting to examine this thesis. I extend my thanks to the Department of Computer Science at University of Neuchâtel for providing me with a an excellent working environment and I would like to acknowledge and thank the Swiss National Science Foundation which financially supported part of this research under Grant #200021-113273.

Furthermore, many thanks go to my former colleague Samir Abdou. His help and support, especially at the beginning of this thesis, contributed a great deal to the success of this thesis. Many thanks also go to Ljiljana Dolamic, my friend, colleague and office mate for the interesting discussions, her patience and for the nice time we spent together.

Special thanks go to my friend and colleague Heiko Sturzrehm for his moral support, his proofreading and in memory of the great times we spent together discussing, playing, laughing.

Last but not least, I would like to thank my family, my parents Monique and Marcel, my sisters Tessy and Lis, my grandmother Yvonne as well as my godmother Rougie for their unconditional love, patience and endless support.

Mama a Papa, Merci dass dir fir mech do sidd. Tessy a Lis, Merci fir dei schein (an maner schein) Zeiten dei mir schon mateneen verbruecht hun an hoffentlech och an Zukunft nach mateneen verbrengen wärten.

Contents

List of Figures

List of Tables

Dedicated to my family

Chapter 1

Introduction

1.1 Motivation

Although information retrieval (IR) was already an issue in the early 1960's mainly in libraries, several factors stimulated research in IR in the 1990's and the beginning of the 21st century.

One of these factors (probably a major one) is certainly the growth of the internet. In March 2009, 23.8%[1] of the world population used the internet compared to only 11.5% in 2004. For Europe the percentage of online users changed from 28.1% in 2004 to 48.9% in 2009 and for North America from 66.1% to 74.4%. Figure 1.1 shows an evolution of the internet usage from 2002 to 2009. According to Nielsen-Online[2], searches submitted to the ten most popular U.S. search engines increased from 5.1 million in October 2005 to 7.8 million in May 2008 and to 9.4 million in May 2009. In their study "How much information? 2003" [1] the authors estimated the size of the internet in 2002 to 532,897 Terabytes.

The constant growth of the World Wide Web and consequently larger amount of available data and the rising number of connected users indirectly entail the development of search engines and thus also bring up new challenges to information retrieval, still the main purpose of a search engine. At the beginning of the internet boom in the early 1990's web directories separating the websites into various categories, such as Yahoo![3], were quite popular for internet searches. With the increasing amount of data available this technique became however unfeasible and full text search engines proved to be more adequate. On the other side, the growth of the internet was amongst others supported

[1] `http://www.internetworldstats.com/`, data from March 31st 2009
[2] `http://www.nielsen-online.com/`
[3] `http://www.yahoo.com`

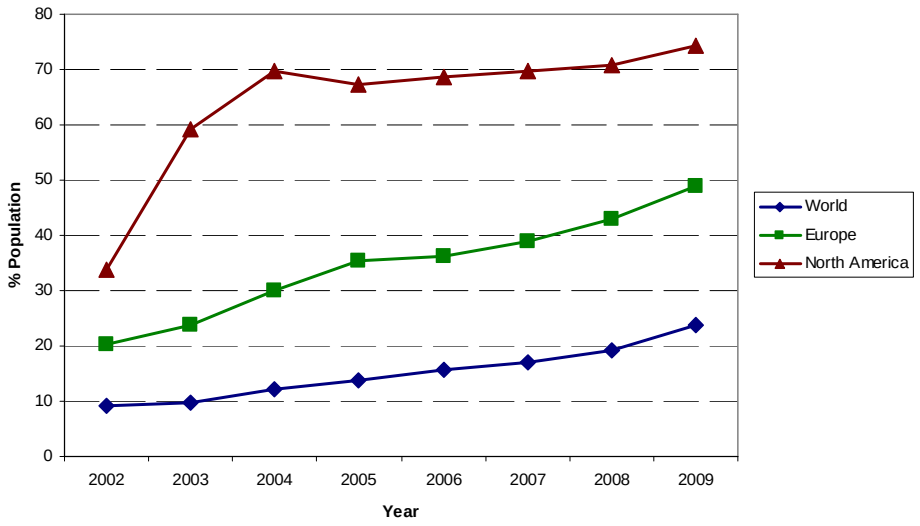

FIGURE 1.1: Internet usage statistics for the World, Europe and North America from 2002 to 2009 (source: internetworldstats.com)

by the genesis of powerful and efficient search engines. They were the key for the internet as we know it today, given that it is useless to make a large amount of data available if it is not possible to search it for information. Consequently, the growth of the internet and development of search engines are at least in part mutually dependent.

Other factors favoring the need of high-performance information retrieval systems are the increasing performance of personal computers, growing capacity of storage media and soaring existence of digital content such as video or audio data, but also the increasing use of smart phones making the data everywhere and constantly available. With this increasing amount of available data, manual search would become inefficient and time consuming. Thus information retrieval systems become essential even on home computers and are not just limited to text retrieval anymore but also face the complexity of retrieving relevant information from multimedia libraries for example, such as YouTube[4] or Flickr[5].

Beyond these popular and most visible facets of information retrieval however, lie different other useful applications of IR, such as domain-specific IR in digital libraries and bibliographic records, one of the original appliances of IR and still an important issue nowadays.

For scientific journals for example increasing printing costs as well as the much easier distribution of electronic copies than of printed media favor electronic dissemination of scientific results. However once the information is available the user also wants to search it in a simple and efficient way. In the case of scientific articles for example, the

[4] `http://www.youtube.com/`
[5] `http://www.flickr.com/`

user is generally only looking for information in a given domain and thus retrieval can be restricted to this domain. Especially for researchers it is important to find relevant information in a fast and efficient way in order to minimize time and effort spend on non relevant resources. Examples of available domain-specific search engines are for example PubMed[6] for the biomedical domain or LexisNexis[7] and SwissLex[8] for the legal domain.

An other interesting and challenging part of IR is its multidisciplinarity. A typical retrieval system is not only based on computer science but also for example on mathematics, statistics, linguistics or library and information science.

1.2 Objectives

Based on the previously described motivations, one of the main objectives of this thesis is the analysis of domain-specific textual IR. A first aim is to study the behavior of standard IR procedures, such as retrieval models, query expansion or indexing techniques on different domains. In the next step, the idea is to elaborate the properties of different domains and adapt the standard procedures to satisfy the needs of the underlying domain and improve retrieval on this domain. In a third step we then eventually want to propose a generic retrieval model adapted to domain-specific IR, independent of the underlying domain.

Furthermore we plan to use different external resources in order to enhance retrieval of relevant documents. Finally we want to go beyond simple document retrieval, such as opinion mining, question answering or prior art search. The ultimate goal is to evaluate the different elaborated strategies by participating in various evaluation campaigns.

1.3 Organization of this Thesis

The remainder of this thesis is organized as follows. The aim of this first chapter is to give a brief overview on IR and present the methodology used in our work. In the following section, we briefly define several concepts of IR for the better understanding of this thesis. In Section 1.5 we present in detail the experimental setup used for our work and finally in Section 1.6 we give a short outline of related work.

[6] `http://www.ncbi.nlm.nih.gov/pubmed/`
[7] `http://www.lexisnexis.com`
[8] `https://www.swisslex.ch`

Chapter 2 aggregates the different papers representing the core part of this thesis. For each paper a short outline and overview of the contributions is given. Finally in Chapter 3, the work is concluded and an outlook on possible future work is given.

1.4 Methodology

A simple, yet appropriate description of information retrieval could be :

*"Information retrieval is the art of finding **relevant** information to a given **query** in a collection of **documents**".*

This simple definition of IR raises three questions: "What means relevant?", "What is a query?" and "What is a document?". In this section we will try to give an answer to these questions and more generally describe the basic concepts of IR. Furthermore we will present the essential background for reading and understanding this thesis.

1.4.1 Information Retrieval

IR is born from a need for information about a given topic. Usually this need is expressed by a human user. To possibly find the wanted information, the user has to formulate a query expressing his/her need. In a second step the query is fed to an information retrieval system (or search engine) where it is matched against the available documents. The documents considered relevant by the system are then returned to the user, who finally selects the documents he/she considers relevant to his/her information need. A general representation of an information retrieval process is shown in Figure 1.2.

To perform a search, we first need documents in which the relevant information should be searched. In modern retrieval systems the notion of *document* is very vast. Everything from simple sentences, to paragraph, section, article and book over chemical formulae, images, to audio and video data, basically everything containing *information* can be considered a document. In this thesis however, we will limit ourselves to text retrieval and therefore the term *document* will be synonym for *text document*. Under this view point the presented procedures are also restrained to text retrieval.

One possibility to search for the information in a document, is to simply sequentially scan the text file and compare it with the query. This would however be time consuming and very inefficient especially for more complicated queries. Therefore the collected documents[9] are generally indexed. This step is presented in more detail in the next

[9] We will not describe the different techniques to collect the documents such as crawling for example as this would go beyond the limits of this thesis

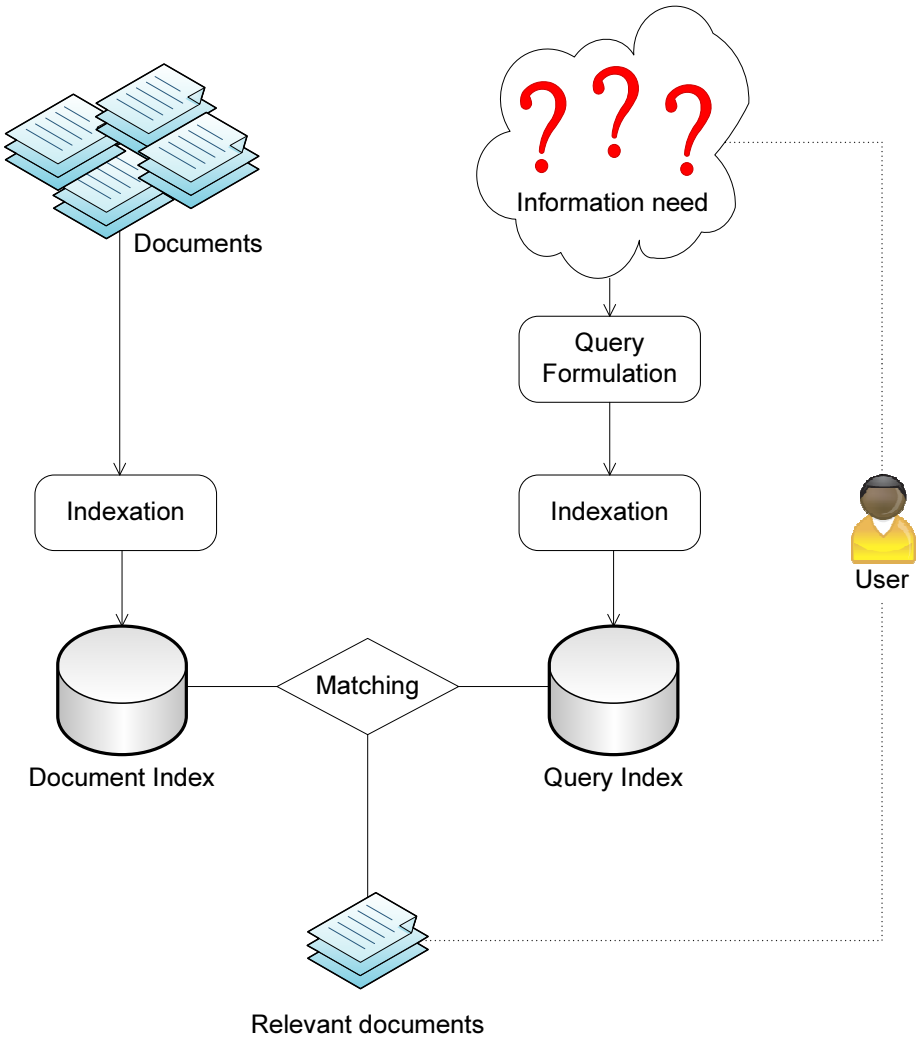

FIGURE 1.2: IR Process

subsection. The query formulated by the user is then also indexed and matched against the document index. During the matching process the retrieval system filters out documents considered relevant and returns them to the user. The matching process is generally based on retrieval models as described in Section 1.4.3.

1.4.2 Indexing

To simplify retrieval of relevant information in a given document collection and to speed up the information retrieval process the collection is generally first indexed, i.e., a set of features (indexing units) such as terms, n-grams or noun phrases, representing the document's content is selected and stored. In this section we will describe the different steps from a full text document to a reduced size index as implemented in our retrieval system. A more complete overview of different indexing techniques and more detailed information can be found for example in [2].

First of all we have to distinguish between three indexing strategies, manual, automatic and semi-automatic. For manual indexing a human person selects for each document the indexing units. Usually a person knowledgeable in the domain to which the document is related reads the document and then chooses the appropriate terms for indexing. An other possibility is that the author added keywords. Even though this approach generally generates high quality retrieval results, it has several inconveniences. Two human annotators for example might choose different indexing terms for the same document. Furthermore manual indexing is time consuming, expensive and a person knowledgeable in the domain has to be available or even several if the collection covers more than one domain. The opposite of manual indexing is automatic indexing. An in-between solution is semi-automatic indexing. Nowadays in most retrieval systems automatic indexing is used and several approaches are possible. In the remains of this section we will describe the different steps used in our automatic indexing procedure.

Even if some indexing steps might change depending on the language or collection, a generic framework for our automatic indexing can be proposed. It follows the bag of words assumption, which means that the order in which the words occur in the document is ignored but their frequencies are stored. The first step is to tokenize the document. This can be done for example by splitting the input text at whitespaces and punctuation marks. Once this is done, we have a list of tokens. Each token is then transformed to lowercase and special words such e-mail addresses, acronyms, etc. are handled according to the chosen analyzer. For example "U.S.A" might be transformed to "u.s.a" and then "usa". Using an other analyzer, we might for example keep the uppercase form ("USA"). In the following steps accents and stopwords are usually removed. Stopwords are words that are considered *useless* for information retrieval, i.e., not containing any information. These words vary depending on the collection and the language. For the English language for example words such as "the", "a" or "is" could be removed, because having "the" or "a" in common with the query is not a good criteria to separate relevant documents from non-relevant documents. Finally each token undergoes language and collection specific analysis, such as stemming, decomposition or separation into n-grams. The resulting tokens after these six steps are the final indexing terms representing the document (or the query).

Figure 1.3 shows possible steps to create an index from an English document applying a stemming strategy. We observe that two exactly the same token (e.g. "president" or "read") are only indexed once. The main reason for this is to minimize the index and speed up searching. However we generally keep track in the index that the *term frequency* (tf) of the token "president" in the given document is 2.

Step 0 (Original Sentence)
The new president of the U.S.A's e-mail address is president@whitehouse.org. It is however a cliché that he is reading and even answering e-mails from the nation. He prefers reading the newspaper.

Step 1 (Tokenization)
The_ new_ president_ of_ the_ U.S.A's_ e-mail_ address_ is_ president@whitehouse.org_ It_ is_ however_ a_ cliché_ that_ he_ is_ reading_ and_ even_ answering_ e-mails_ from_ the_ nation_ He_ prefers_ reading_ the_ newspaper

Step 2 (Transform to lowercase)
the new president of the u.s.a's e-mail address is president@whitehouse.org it is however a cliché that he is reading and even answering e-mails from the nation he prefers reading the newspaper

Step 3 (Handle special words)
the new president of the usa s email address is president whitehouse org it is however a cliché that he is reading and even answering emails from the nation he prefers reading the newspaper

Step 4 (Remove accents)
the new president of the usa s email address is president whitehouse org it is however a cliche that he is reading and even answering emails from the nation he prefers reading the newspaper

Step 5 (Remove stopwords)
president usa email address president whitehouse cliche reading answering emails nation prefers reading newspaper

Step 6 (Stemming)
president usa email address president whitehouse cliche read answer email nation prefer read newspaper

Final indexing terms
president usa email address whitehouse cliche read answer nation prefer newspaper

FIGURE 1.3: Indexing Steps

1.4.3 Retrieval Models

The main challenge in classical IR is to find all documents relevant to a query. As seen in section 1.4.2 documents and queries will be represented by a set of indexing terms. Based on this representation the retrieval system has to decide whether a document is relevant or not to a given query. This decision is usually based on a *retrieval model*. The aim of a retrieval model is to define how documents and queries are represented and how the similarity between a document and a query is computed. A ranked list containing the documents with the highest similarities (scores) is then returned to the user. In classical IR, we might usually distinguish between four classes of retrieval models, the

Boolean model, the logical model, the vector-space model and the probabilistic model. A complete overview of these classes can be found for example in [3], [4] or [5].

1.4.4 Evaluation

The last question which is still open is "What means relevant?". A document is considered *relevant* when the information it contains satisfies the user's information need. However if two users formulate the same query they might not consider the same documents as relevant. If for example the travel agent Anne and the biologist Ben formulate the query "malaria", Anne might be looking for information about regions targeted by this disease while Ben might be looking for the transmitter and thus they do not necessarily judge the same documents as relevant.

Normally, the main assignment of an information retrieval system is to retrieve documents relevant to a query. One of the important research purposes in information retrieval is therefore to optimize the ability of the system to fulfill this assignment. And because *"If you can not measure it, you can not improve it"* (Lord Kelvin) an evaluation measure is needed. But also a test reference collection consisting of a collection of documents (*corpus*), a set of queries (*topics*) and a list of relevant documents (*relevance judgments*) for each query are needed.

The first major test-collection in information retrieval was built in the late 1960's during the *Cranfield* project [6]. The collection contains 1,398 abstracts of aerodynamics journal articles as well as a set of 225 queries and relevance judgments. From the Cranfield experiments ([6] and [7]) emerged the *Cranfield paradigm* [8] which constitutes the base of evaluation in the research field of information retrieval. To evaluate a retrieval system according to the *Cranfield paradigm* a test-collection and an evaluation methodology are needed. Furthermore under this paradigm, two systems tested on the same test-collection can then be evaluated and their performance statistically compared. In her article, Vorhees [8] gives an overview of the *Cranfield paradigm* and reviews its fundamental assumptions and appropriate uses in the context of modern information retrieval.

Over the years different test-collections have been build and made available, mostly through evaluation campaigns such as TREC or CLEF (described in Section 1.4.6). The different collections used for the empirical studies of this thesis are described in Section 1.5.3.

Two key statistics used to evaluate retrieval systems and used as base for various other evaluation measures are *precision* and *recall*. While *precision* quantifies the fraction

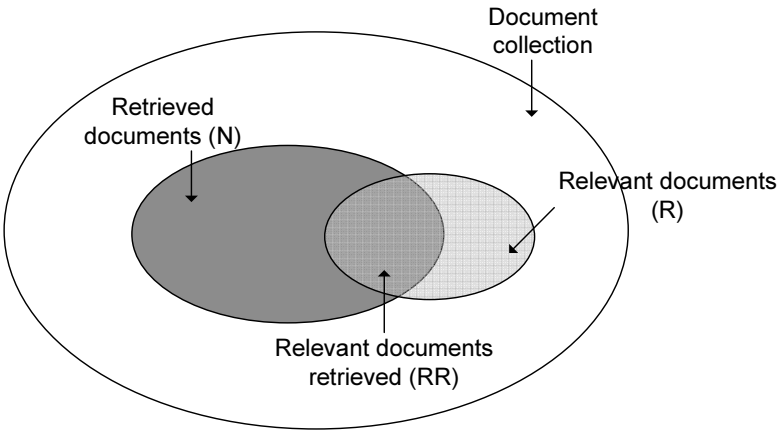

FIGURE 1.4: Recall and Precision

of the returned results relevant to the query, *recall* expresses the fraction of relevant documents retrieved. Let N be the total number of documents retrieved for the query Q, RR the number of relevant documents retrieved by the system and R the number of relevant documents for Q. *Precision* and *recall* can then be defined as follows.

$$Precision = \frac{RR}{N} \tag{1.1}$$

$$Recall = \frac{RR}{R} \tag{1.2}$$

Figure 1.4 shows an illustration of *precision* and *recall*. Different evaluation measures for information retrieval have been presented, mostly based on *precision* and *recall* but eventually also on the rank of the retrieved documents ([4], [2], [9]). The evaluation measure used in our work is presented in Section 1.5.2.1.

All the experiments described in this thesis are conducted and evaluated under the assumptions of the *Cranfield paradigm*.

1.4.5 Relevance Feedback

An additional feature implemented in some retrieval systems to improve performance is relevance feedback. As seen before not necessarily all documents returned to the user contain the wanted information. The idea of relevance feedback is to reformulate the query and recompute rankings of documents based on the user's feedback which documents are relevant and which are not. The different steps of an IR procedure including relevance feedback would then be

1. Information need expressed by a query and send to an IR system
2. Ranked list of retrieved documents returned to the user according to the query

3. Some documents are marked as relevant or non relevant by the user (not necessarily all)
4. Reformulation of the query by the IR system
5. New set of results returned to user

We can in general distinguish between three categories of relevance feedback, explicit, implicit and blind (pseudo) relevance. While explicit relevance feedback is based on the explicit judgment of the user if a document is relevant or not, implicit relevance feedback uses indirect evidences to estimate relevancy of documents, such as the time spend to read a given document. Blind relevance feedback does not depend on the user's judgments but automates this step of relevance feedback. The basic idea of blind relevance feedback is to assume that the top k documents are relevant and eventually that the documents occurring at the bottom of the retrieved list are not relevant.

In our work, we used two pseudo-relevance feedback approaches, the first based on Rocchio's method [10] and the second proposed by Abdou *et al.* [11]. For both approaches the system would add the m most *important* terms (defined by the used method) extracted from the top k documents retrieved for the original query.

1.4.6 Evaluation Campaigns

Information retrieval has a long empirical tradition starting with the *Cranfield* experiments ([6] and [7]) in the 1960's. Many experiments were conducted and several other test-collections have been built. However at the beginning of the 1990's the lack of comparable and reproducible results on a large scale, the stagnation of test-collection to a relatively small size, as well as the questionable evaluation methodology led to a growing dissatisfaction [12]. Furthermore by that time the National Institute of Standards and Technologies (NIST) build a large test-collection for the TIPSTER project supported by the Defense Advanced Research Projects Agency (DARPA). Under these premises the first evaluation campaign was launched in 1992, under the name of TREC[10] (Test REtrieval Conference) by the NIST. The aim of the founders was to continue the Cranfield tradition by offering the necessary requirements, such as test-collections, queries and relevance judgments as well as an appropriate evaluation methodology.

The first TREC campaign was held in November 1992. Over the years different "de facto standards" arouse from the TREC campaigns, such as the MAP evaluation measure as described in Section 1.5.2.1 and the statistical testing (Section 1.5.2.2) or the common

[10] `http://trec.nist.gov/`

convention to formulate topics in three parts (title, description and narrative). These "de facto standards" are nowadays frequently used in other evaluation campaigns.

As of today, various evaluation campaigns offering different tracks exist, not only for text retrieval (CLEF[11], NTCIR[12] or FIRE[13]) but also for example on digital libraries (INEX[14]) or video and multimedia (IMAG-EVAL[15], TRECVid[16]).

Basically the main goal of the evaluation campaigns is to propose an environment to exchange and discuss research ideas but also to offer the proper tools such as test-collections to evaluate and compare the different information retrieval systems of the participants. The campaigns are experimental workshops and the benchmarking experiments are not supposed to be competitive but rather to identify strengths and weaknesses of the different systems. Furthermore the objective is to propose useful real-world scenarios.

A large part of the experiments presented in this thesis were conducted as part of the TREC and CLEF (Cross Language Evaluation Forum) evaluation campaigns.

1.5 Experimental Setup

Since information retrieval is mainly based on empirical studies, the primary part of this work is also based on empirical studies. In the following sections, we describe the models (1.5.1) and the measures (1.5.2) used to evaluate and compare our different retrieval experiments as well as the test-collections and domains on which we mainly worked (1.5.3) and the search engine library used (1.5.4).

1.5.1 Retrieval Models

In the following paragraphs we will give a short overview on the different models used during our research. All models were issued from the classes of vector-space model and probabilistic model.

1.5.1.1 Vector-Space Model

The vector-space model has been presented in the late 1960's and been first used in the SMART (System for the Mechanical Analysis and Retrieval of Text) retrieval system [9].

[11] `http://www.clef-campaign.org/`
[12] `http://research.nii.ac.jp/ntcir/`
[13] `http://www.isical.ac.in/~fire/`
[14] `http://inex.is.informatik.uni-duisburg.de/`
[15] `http://www.imageval.org/`
[16] `http://www-nlpir.nist.gov/projects/trecvid/`

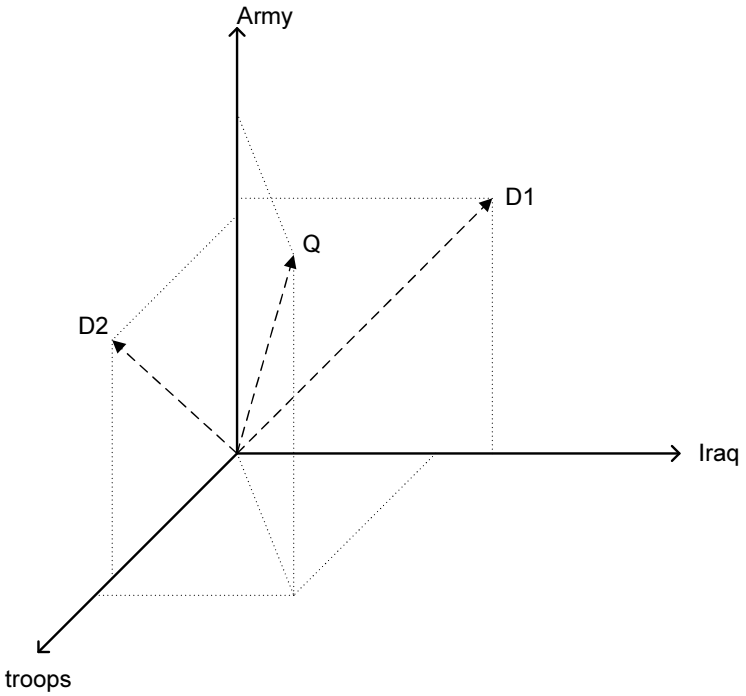

FIGURE 1.5: Representation of the vector-space model

Being aware of several disadvantages of the Boolean model, especially exact matching, Salton *et al.* [13] presented the vector-space model proposing a framework allowing partial matching. Based on geometric intuition, the idea of the model is to attach weights to query terms and represent documents and queries as vectors and calculate the degree of similarity between those vectors. The coordinates of the vectors are the weights of the query terms in the document (for the document vector) and the query. If the query contains v terms, documents and the query are represented in a v-dimensional vector-space. It is generally supposed that the terms are independent and thus term-vectors are orthogonal.

Figure 1.5 shows a representation of the query "Army troops in Iraq" and the two documents $D1$ and $D2$. The first document contains only the terms "Iraq" and "Army", while the second document contains the terms "troops" and "Army".

Among many others, one possibility to calculate the degree of similarity between a document D_k and a query Q is to calculate the cosine of the angle between the two

vectors according to following formula.

$$\begin{aligned} score(D_k, Q) &= sim(D_k, Q) && (1.3)\\ &= \frac{\vec{D_k} \cdot \vec{Q}}{\left\|\vec{D_k}\right\| \left\|\vec{Q}\right\|} && (1.4)\\ &= \frac{\sum_{i=1}^{v} w_{ik} \cdot w_{iQ}}{\sqrt{\sum_{i=1}^{v} w_{ik}^2} \cdot \sqrt{\sum_{i=1}^{v} w_{iQ}^2}} && (1.5) \end{aligned}$$

where w_{ik} represents the weight of term t_i in the document D_k, w_{iQ} the weight of t_i in the query and v the number of terms in the query. Equations 1.4 and 1.5 are one possibility to compute similarity between documents and query. Other possibilities are for example the Jaccard and Dice similarity measures.

In [9] Salton and McGill give a description of various schemes for term weighting. A commonly used method is the *tf idf* weighting scheme, in which the weights are calculated by

$$w_{ij} = tf_{ij} \cdot idf_i \tag{1.6}$$

where tf_{ij} represents the frequency of the term t_i in the document D_j and idf_i the *inverse document frequency* of the term t_i. Generally the *idf* is calculated as

$$idf_i = \log \frac{n}{df_i} \tag{1.7}$$

where n is the total number documents in the collection and df_i the document frequency of the term t_i in the collection, i.e., the number of documents containing it. We will reference to this model in the remain of this thesis as "standard *td idf* (vector-space) model with cosine normalization".

One drawback of the vector-space model is its lack of a solid theory justifying several aspects. For example no theoretical justification on which similarity measure to chose or which term weighting scheme to use is given. However the vector-space model proved to be quite efficient and is nowadays still one of the most used models.

1.5.1.2 Probabilistic Model

The first probabilistic model in information retrieval has already been presented in 1960 [14]. The principle of probabilistic model has been summarized in 1977 by the *Probabilistic Ranking Principle* (*PRP*) formulated by Robertson [15]. The PRP, common idea behind the different implementations of the probabilistic retrieval model, states that documents should be ranked according to their estimated probability of relevance to a given query. The PRP does however not explain how this ranking should be computed

but just suggest using probability theory to establish the ranking. The "*How?*" is given by various implementations of the probability model, and the difference between these implementations is mainly how the probability of relevance is estimated.

The PRP can be formalized by following formula.

$$score(D, Q) = \frac{P(R|D)}{P(R^c|D)} \tag{1.8}$$

where $P(R|D)$ represents the probability that the document D is relevant with respect to the query Q and $P(R^c|D)$ the probability that D is not relevant. By applying the Bayes theorem we obtain

$$\begin{aligned} score(D, Q) &= \frac{P(R|D)}{P(R^c|D)} && (1.9) \\ &= \frac{P(D|R) \cdot P(R)}{P(D|R^c) \cdot P(R^c)} && (1.10) \end{aligned}$$

and finally since $P(R)$ and $P(R^c)$, representing the probabilities of a document being respectively relevant or not, are usually the same for all documents in the collection, the documents can be ranked using

$$score(D, Q) = \frac{P(D|R)}{P(D|R^c)} \tag{1.11}$$

In the following sections we present three different implementations of the probabilistic model used during our research. The presented models make the assumption that all terms are independent.

Okapi The first implementation we used is the BM25 (Best Match 25) model, better known as Okapi. The model was developed at *City University of London* under the assumption that an efficient implementation of the probabilistic retrieval model should account for the term frequency of the query terms in a given document and the length of the document to calculate its relevancy probability. Under these assumptions different weighting strategies have been tested finally resulting in the Okapi model ([16] and [17]).

In this thesis following scoring formula is used when we reference to the Okapi model.

$$score(D_k, Q) = \sum_{t_i \in Q} qtf_i \cdot \log \left[\frac{n - df_i}{df_i} \right] \cdot \frac{(k_1 + 1) \cdot tf_{ik}}{K + tf_{ik}} \tag{1.12}$$

with $K = k_1 \cdot [(1 - b) + b \cdot \frac{l_k}{avdl}]$ and where qtf_i denotes the frequency of term t_i in the query Q, n the number of documents in the collection, df_i the number of documents in which the term t_i appears, l_k the length of the document D_k and $avdl$ the average

document length. The constants b and k_1 are generally set empirically according to the underlying collection.

Divergence from Randomness Repeatedly during our research we used implementations of the probabilistic model issued from the *Divergence from Randomness* paradigm (DFR) as presented in [18]. One of the main ideas behind the DFR paradigm is that terms which entail only little information are randomly distributed over the collection. Consequently the more information a term adds to a given document, the more its term frequency in the given document diverges from its frequency within the collection.

The relevancy score is given by

$$score(D_k, Q) = \sum_{i=1}^{q} w_{ik} \cdot qtf_i \tag{1.13}$$

where qtf_i represents the frequency of the term t_i in the query, w_{ik} the weight of the term in the document and q the number of terms in the query. According to the DFR framework, w_{ik} is based on two information measures.

$$w_{ik} = Inf^1_{ik} \cdot Inf^2_{ik} \tag{1.14}$$

Inf^1 is defined as $Inf^1 = -\log_2(Prob_1)$ where $Prob_1$ is the probability of having tf_{ik} occurrences of the term t_i in the document D_k according to a chosen model of randomness (e.g., Poisson distribution or Bose-Einstein statistics). According to the core idea of the paradigm, the smaller this probability is the more information the given term is carrying since it diverges from the randomness model. Inf^1 thus formalizes this idea.

Inf^2 is defined as $Inf^2 = 1 - Prob_2$. $Prob_2$ represents the *risk of accepting a term as document descriptor* [18]. It indicates the probability to come across another occurrence of term t_i in the document D_k given already tf occurrences and it is obtained by observing only the set of documents in which the term occurs (*elite set*). This probability is for example modelized using a Laplace model or a ratio of two Bernoulli processes. Often Inf^2 is referenced to as the *first normalization of the informative content.*

One crucial factor to calculate $Prob_1$ and $Prob_2$ is the term frequency (tf). However the tf value depends on the length of the documents, and thus the term frequency is generally normalized. In [18] the authors propose two hypotheses to normalize the term frequency. This normalization is generally referenced as *second normalization.*

To summarize, a model issued from the DFR paradigm has three components, a basic randomness model, *first normalization* and *second normalization*. From these three components arrived the nomenclature for the DFR models. For example the model *PL2* would use a Poisson distribution as basic randomness model, a Laplace law as *first normalization* and hypothesis 2 as *second normalization*.

In Appendix E a complete overview of the components used in this work is given. A detailed description and explanation of all used DFR models is given in [18].

Language Models Even though language models have a long history ([19]) in natural language processing (NLP) and particularly in speech recognition, only in 1998 they have been applied to IR [20]. The classical probabilistic model (as described previously) uses the probability $P(R|D)$ of relevancy given a document and a query as ranking score. The idea behind the language models in information retrieval is that a user would use words he/she supposes to find in a relevant document to formulate the query. Under this assumption, a document would be a good match to a given query, if the document language model is likely to generate the query. Therefore for each document D_k a probabilistic language model M_{D_k} is build and the relevancy score of the document is calculated as

$$score(D_k, Q) = P(Q|M_{D_k}) \tag{1.15}$$

Various strategies have been proposed to calculate $P(Q|M_D)$. In our work we used a unigram model (all terms are independent), as proposed by Hiemstra [21] and defined by following formula.

$$\begin{aligned} score(D_k i, Q) &= P(Q|M_{D_k}) && (1.16)\\ &= \prod_{t_j \in Q} P(t_j|M_{D_k}) && (1.17)\\ &= P(D_k) \prod_{t_j \in Q} (\lambda_j \cdot P(t_j|D_k) + (1-\lambda_j) \cdot P(t_j|C)) && (1.18) \end{aligned}$$

with $P(t_j|D_k) = tf_{jk}/l_k$ and $P(t_j|C) = df_j/lc$ with $lc = \sum_k df_k$, where λ_j is a smoothing factor. The probability $P(D_k)$ is usually constant and equal for all documents and can thus be ignored to compute ranking scores.

	Query A		Query B		Query C	
Rank	Relevant	Precision	Relevant	Precision	Relevant	Precision
1	1	1	1	1	1	1
2	0		1	2/2	1	2/2
3	1	2/3	1	3/3	1	3/3
...						
10	1	3/10	0		0	
...						
50	1	4/50	1	4/50	0	

TABLE 1.1: Average Precision

1.5.2 Evaluation and Comparison

1.5.2.1 Evaluation

As described in Section 1.4.4 different measures to evaluate information retrieval systems are available. During our work we used mainly one measure, namely the Mean Average Precision (MAP) defined as

$$MAP = \frac{1}{k} \sum_{j=1}^{k} AP(Q_j) \tag{1.19}$$

where k is the total number of queries and $AP(Q_j)$ the average precision for query Q_j defined as

$$AP(Q_j) = \sum_{i=1}^{N} \frac{\mathbb{1}_{rel(i)} \cdot P@i}{R_j} \tag{1.20}$$

where N is the number of documents retrieved for the query Q_j, R_j the number of relevant documents for Q_j and $\mathbb{1}_{rel(i)}$ and $P@i$ defined as follows:

$$\mathbb{1}_{rel(i)} = \begin{cases} 1 & \text{if the document at rank } i \text{ is relevant} \\ 0 & \text{else} \end{cases}$$

$$P@i = \frac{\text{number of relevant documents found until rank } i}{i}$$

$P@i$ represents the *precision at i*.

Table 1.1 shows an example of three queries and their retrieval results. All queries have four relevant documents and for each query 50 documents have been retrieved. For the

third query not all relevant documents have been retrieved. Therefore we have

$$AP(Q_A) = \frac{1}{4} \cdot \left(\frac{1}{1} + \frac{2}{3} + \frac{3}{10} + \frac{4}{50} \right) = 0.52 \quad (1.21)$$

$$AP(Q_B) = \frac{1}{4} \cdot \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{3} + \frac{4}{50} \right) = 0.77 \quad (1.22)$$

$$AP(Q_C) = \frac{1}{4} \cdot \left(\frac{1}{1} + \frac{2}{2} + \frac{3}{3} \right) = 0.75 \quad (1.23)$$

$$MAP = \frac{1}{3} \cdot (0.52 + 0.77 + 0.75) = 0.68 \quad (1.24)$$

The MAP for the tested retrieval system would then be 0.68.

Even though this measures presents several inconveniences (e.g., hard to interpret for the final user), it is the principal measure used in various evaluation campaigns and thus our first choice to evaluate our test runs. Furthermore this measure takes account of the *precision*, the *recall* as well as of the rank of the retrieved and relevant document. We used the TREC_EVAL[17] tool made available through TREC to calculate MAP values. MAP calculations are based on a maximum of 1,000 retrieved documents.

1.5.2.2 Comparison Statistics

Our goal is not only to evaluate one single retrieval strategy (or retrieval model) but also to compare the different approaches. One solution is to simply assume if $MAP_A > MAP_B$ then system A performs better than system B. This might be true in a crude view, but if we examine the systems closer this simple comparison does not necessarily hold. We could imagine for example that we have two systems tested on 1,000 queries. For each query both systems obtain the same average precision (AP) except for one query, where system A produces a slightly better result. In this case $MAP_A > MAP_B$ but we can not say in good conscience that system A performs better than system B.

Therefore to test if two systems are statistically different, we opted for the *bootstrap* methodology as presented by Savoy [22]. In our case the null hypothesis H_0 states that both retrieval systems produce a similar performance. If the two systems show statistically the same performance, H_0 would be accepted, otherwise it would be rejected. We applied a two-sided test with significance level $\alpha = 5\%$.

[17] `http://trec.nist.gov/trec_eval/`

1.5.3 Test-Collections and Domains

The main focus of this thesis lies on domain-specific information retrieval and in particular in the domains of social science, biomedicine and Blogsphere. We therefore briefly present the collections used for each of these domains. Furthermore we also seize the opportunity and give a brief overview of the particularities of each domain and the motivations to work on those domains.

1.5.3.1 Social Science

For our first domain of interest, social science, we used the GIRT (German Indexing and Retrieval Test database) collection, more precisely the GIRT4-DE corpus.

The original GIRT corpus was created in 1997 to serve as base for comparison between different retrieval systems for the German language. Most collections available until then were mainly English press corpora. The GIRT collection should become a German collection with focus on domain-specific bibliographic references.

The first GIRT collection (GIRT1) was composed of about 13,000 documents from the years 1990 to 1996 extracted from various social science sources. It has been made available through the TREC evaluation campaign in 1999. Over the years, the collection evolved and we finally use the fourth version. GIRT4 was the first of the GIRT corpora available in two languages, German (GIRT4-DE) and English (GIRT4-EN). The English version is a manual translation of the German collection. The GIRT-4 corpora have been used in the domain-specific tasks of the CLEF campaigns from 2003 to 2008.

The GIRT4-DE corpus is composed of bibliographic records extracted from two social science sources, SOLIS (social science literature) and FORIS (current research in the field of social science), covering the German language. The GIRT4-DE corpus now contains more than 150,000 documents. For more information about the evolution and description of the different GIRT corpora see [23].

A typical record of the GIRT4-DE corpus consists of author name, title, document language, publication year and abstract. Furthermore, for each document descriptors extracted from a controlled vocabulary are manually added.

To evaluate our different systems on this collection, we used the queries deployed in the domain-specific track in CLEF from 2004 to 2008. This gives us a total of 125 queries, developed for the GIRT4-DE corpus, as well as relevance assessments for these queries.

According to the standard TREC format, each topic is structured into three logical sections, namely a title, a description and a narration. Additionally to the documents

and the queries, a machine readable version[18] of the German-English thesaurus for social science [24] was made available. This version of the thesaurus contains a total of 10,624 entries. Each entry represents a German descriptor, given with related, broader or narrower terms. In the Appendix B examples of documents, topics and thesaurus are shown.

Our work on the GIRT collection was motivated by different aspects. First this collection contains manually added descriptors for each document in the collection. This is a good premise to compare manually and automatically indexing techniques as well as the benefit of manually added keywords. Second this collection consists of bibliographic records and search in domain-specific literature represents a main challenge in "real world" information retrieval, besides web retrieval. Finally for this collection a machine readable thesaurus is available, simplifying empirical studies on automatic document and query extension. Last but not least an interesting factor is the language of the collection. Compared to English, German is a morphologically much more complex language and thus presents several additional challenges.

1.5.3.2 Blogsphere

Over the last years blogs (web logs in journal style) became more and more frequent among internet users. Nowadays web space is often provided freely to the user (financed by advertisement) and various tools make it very simple to put its own web log online. These facilities foster personal blogs and motivate users to share their opinion, view point or simply their life with other users. The constant growth of the Blogsphere, but also the subjective nature of the blogs as well as their particular language and writing style are factors that in our opinion justify investigating further in the direction of information retrieval in the Blogsphere, going even beyond simple text retrieval (e.g., opinion mining).

A large part of the presented thesis focuses on the Blogsphere. To analyze IR in this domain, we used of the Blogs06 collection described below.

For the TREC 2006 campaign, the organizers wanted to promote retrieval in the Blogsphere and thus an adequate collection was needed. Between December 2005 and February 2006, members from the *Department of Computing Science* from *University of Glasgow* crawled the web and created the wanted collection. The final collection, named Blogs06, contains a total of 4,293,732 documents (148 GB) separated into 753,681 feeds (38.6 GB), 3,215,171 permalinks (88.8 GB) and 324,880 homepages (20.8 GB).

[18] `http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/`

Compared to standard websites, blogs usually have XML feeds describing recent postings. Nowadays two standard formats are used for feeds, RSS and Atom. In the Blogs06 collection, both standards were considered equally and thus the collection contains feeds in both formats. Generally the feeds contain information about new blog posts (permalinks) as well as the content of these posts. However the feed does not necessarily include the full text of the posts. For this reason besides the feeds also the permalinks have been added to the collection. Literally, a permalink is a link to a given blog post. The permalink `http://weblog.greenpeace.org/nuclear-reaction/2009/07/the_epr_at_olkiluoto_from_disa.html` for example links to a post with the title *The EPR at Olkiluoto: from disaster to farce* from the blog of the organization *Greenpeace*. In the context of the Blogs06 collection, permalink is synonym for the documents referenced by the permalink. Additionally to the feeds and the permalinks, the collection also contains homepages representing the entry point to a blog. For the previously given example, this would be `http://weblog.greenpeace.org/`. Since the goal was to mirror in the best possible way the Blogsphere, the collection also contains spam blogs (splogs) and spam comments have not been removed. A more detailed description of the Blogs06 collection and its creation can be found in [25].

For our experiments, we only considered the permalink part of the Blogs06 collection. Currently a total of 150 topics with relevance assessments are available for this collection. These topics have been created during the three years this collection has been used at the TREC evaluation campaign (2006-2008). To take account of the subjective nature of the blogs and to promote opinion mining, the relevance assessments also judge the subjective character of the blogs. In the Appendix C examples of a permalink document and topics are shown.

1.5.3.3 Biomedicine

Last but not least, we worked on the biomedical domain. Among others, a particularity of this domain is the very specific language. Generally each specific domain has its specific vocabulary, but while for the two previously described domains for example, terms from their vocabulary might still be found in a general dictionary (as long as they are correctly spelled), this is not necessarily the case for the biomedical domain. An other particularity is for example the presence of orthographic variants (e.g., “Krohn” or “Crohn”) and naming variations (e.g., “BSE” or “mad cow disease”). Furthermore in the biomedical domain, scientific literature is an important source of evidence for researchers having different background and interests. This corpus is crucial to share their knowledge and is not only resource for research but also for diagnosis. Together

with the growing size of available biomedical literature, it has become an important challenge to improve information retrieval on this particular domain.

For our studies on the biomedical domain, we used two different collections, both made available through TREC and used in the Genomics tracks from 2004-2005 respectively from 2006-2007.

The first collection, used in 2004 and 2005, contains a 10-year subset from MEDLINE[19] (Medical Literature Analysis and Retrieval System Online) bibliographic database. The corpus contains 4,591,008 records which in 2004 represented about one third of the entire database. In addition to standard information such as identifier (PMID), title and abstract, the documents also contain keywords, so-called Medical Subject Headings (MeSH) headings, extracted from the MeSH[20] thesaurus. From the two years the collection was used for TREC, we gathered 100 topics with their relevance judgments. More information on the documents and the topics can be found in [26] and [27].

The second biomedical collection comes from HighWire Press[21] an electronic distribution of journals. The corpus contains 162,259 documents (12.3 GB) from 49 journals related to the biomedical domain. Each document is a full text paper formatted in HTML. Used during the 2006th and 2007th editions of the Genomics track in TREC, a total of 64 (28 from 2006 and 36 from 2007) queries is available for this collection. Further information about this collection can be found in [28] and [29]. In the Appendix D examples of documents and queries are shown for both collections.

1.5.4 Lucene

The basic system on which all our experiments are conducted is implemented using the Lucene[22] Java API. The Lucene project is an open-source information retrieval software supported by the Apache Software Foundation including various components, such as a crawler (Droids), a suite of scalable machine learning libraries (Mahout) or the used Java API (Lucene Java) which is the main component of the Lucene project.

The Java API provides not only full text indexing and searching technologies but also libraries for spellchecking or advanced analysis and tokenization. The libraries can easily be extended to satisfy one's own needs. The freely available Lucene source-code

[19] `http://www.nlm.nih.gov/pubs/factsheets/medline.html`
[20] `http://www.nlm.nih.gov/mesh/`
[21] `http://highwire.stanford.edu/`
[22] `http://lucene.apache.org/`

has been slightly modified to even better satisfy our needs, especially to facilitate the implementation of new retrieval models[23].

A complete description of Lucene can be found on the project homepage or in the corresponding book [30].

1.6 Related Work

The main work of this thesis consists of analyzing and improving domain-specific information retrieval on primarily three domains. In this section, we give a short overview on work previously done in the different domains. A more detailed and specific related work overview can be found in each of the papers on which this thesis is constructed and presented in Chapter 2

As described in 1.5.3.1 the collection used for our experiments on social science contains manually added descriptors for each document. It has been shown ([31]) that the use of descriptors extracted from a controlled vocabulary could improve retrieval results if they were used additionally to the document content. In [32] Savoy showed on a corpus containing French bibliographic notices, that for short queries including manually assigned descriptors might significantly improve retrieval results. For the GIRT corpus, Petras [33] obtained the same conclusion. She suggested including manually added descriptors in order to disambiguate fuzzy topics. Both studies however just analyze the impact of manually added keywords. One of our objectives is to automatically enhance the documents using the available thesaurus and compare both enhancements.

But also in the biomedical domain the use of manually assigned descriptors has been proven. Abdou *et al.* ([11]) showed that extending documents with descriptors extracted from the MeSH thesaurus could enhance retrieval performance by up to 13.5%. Furthermore the authors showed that expanding queries by adding automatically generated orthographic variants from the query words would slightly enhance retrieval performance.

Research on information retrieval with respect to the biomedical domain was highly supported by the Genomics tracks in the TREC campaigns from 2003 to 2007. During these five years different subtasks have been offered, such as ad hoc retrieval, information extraction, text categorization, passage retrieval and question answering. One objective of the tasks was to represent real-world problems and thus for example real information

[23]The author would like to thank again at this place her former colleague Samir Abdou for his enormous work done on extending the original Lucene API. Without this valuable tool much of the research done for this thesis would have been by far more circuitous

needs from biologists would be used for queries. Various approaches have been proposed to satisfy the needs of the different tasks. A complete overview of the different strategies proposed can be found in the respective proceedings ([34], [35], [36], [37] and [38]).

Research on information retrieval in the Blogsphere became an interesting topic mainly due to the growth of the Blogsphere but also due to the subjective nature of blog posts providing consequently an ideal playground for research in opinion mining.

Query expansion based on external resources has become a common strategy to retrieve blogs relevant to a given topic. In [39] Zhang *et al.* for example use Wikipedia and web feedback for query expansion, while Weerkamp *et al.* [40] propose to use a news corpus covering the same timeframe as the blog corpus to extend queries.

Also for opinion mining different strategies have been proposed, mainly based either on weighted dictionaries or precompiled lists of subjective terms.

As for the biomedical domain, information retrieval and opinion mining in the Blogsphere has been considerably developed thanks to the Blog tracks at TREC (2006 until today, [37], [38] and [41]).

Information retrieval on various other domains has also been promoted, such as in the "legal" domain (legislation, regulations and judicial decisions for example), chemistry or intellectual property.

We observe that the trend in domain-specific information retrieval is to analyze the underlying domain and adapt the retrieval system in order to maximize performance on the given domain. We will essentially follow this movement in our work, and mainly focus on the tree presented domains. On the other side, one goal is however to try to propose a generic framework with the goal of improving retrieval in specific domains taking account of particularities of the given domain but usable on all domains without adjustments due to the domain.

Chapter 2

Overview of Selected Papers

2.1 Introduction

In this section we will briefly present the different papers on which this thesis is drawn. These papers have been published in different conferences, workshops and journals and cover a large part of our work on domain-specific information retrieval and beyond.

Following papers have been selected, mainly to underline our work in three different domains (social science, biomedical and blogs):

- C. Fautsch, J. Savoy
  IR-Specific Searches at TREC 2007: Genomics and Blog Experiments
  In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings* Gaithersburg, Maryland, November 5-9, 2007.

- C. Fautsch, J. Savoy
  Stratégies de recherche dans la blogosphère
  In *Document Numérique* Volume 11, pages 109-132, Hermès-Lavoisier, Paris, France, 2008.

- C. Fautsch, J. Savoy
  UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere
  In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings* Gaithersburg, Maryland, November 18-21, 2008.

- C. Fautsch, L. Dolamic, J. Savoy
  UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches
  In *Working Notes for the CLEF 2008 Workshop* Aarhus, Denmark, September 17-19, 2008.
  Revised version in *Evaluating Systems for Multilingual and Multimodal Information Access*, Lecture Notes in Computer Science, *to appear*

- C. Fautsch, J. Savoy
  Comparison Between Manually and Automatically Assigned Descriptors Based on a German Bibliographic Collection
  In *Proceedings of 6th International Workshop on Text-based Information Retrieval* Linz, Austria, August 31 - September 4, 2009.

- C. Fautsch, J. Savoy
  Adapting the *tf idf* Vector-Space Model to Domain-Specific Information Retrieval
  To appear in *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC)*, Sierre, Switzerland, March 22-26, 2010.

- C. Fautsch, J.Savoy
  Algorithmic stemmers or morphological analysis? An evaluation
  In *Journal of the American Society for Information Science and Technology*, Volume 60, pages 1616-1624, Wiley InterScience, 2009.

In the following sections, we will give a short overview of the content and the contributions for each of these papers. The complete papers containing all references, results and discussions can be found in Appendix A. The complete reference list to all published papers is presented in Appendix E. Copyrights for the presented papers are held by the respective publishers.

2.2 IR-Specific Searches at TREC 2007: Genomics and Blog Experiments

This paper describes our participation in the sixteenth edition of TREC (2007). We participated in the Genomics and Blog tasks of this campaign. Both tasks go beyond simple document retrieval and raise problems specific to the respective domains.

In the Genomics track the participants had to retrieve relevant document passages from publications extracted from various biomedical journals. The segmentation of the original documents into passages was left to the participant. This segmentation added an additional challenge to this track besides the retrieval of relevant information. The queries were based on real information needs gathered from biologists. Each topic refers to one of fourteen possible entities (aspects) such as antibodies, pathways or symptoms for example. The systems of the participating groups should return passages covering these entities, surrounded by supporting text. The retrieved passages have been evaluated based on a character-based MAP (called Passage2 MAP).

Second, we also participated in the Blog track. Nowadays more and more Internet users maintain a personal web log, a so-called blog. They publish content varying from their diary to complete product reviews, either in an objective way or with a personal touch. The aim of the Blog track was therefore to emulate an Internet user searching factual information as well as opinions on a specific target in blogs. This task could be divided into two subtasks. First the system should retrieve relevant information (facts). In a second step the retrieved documents should be classified according to the included opinion regarding the topic target (positive, negative or mixed).

In the paper, we first gave a short overview on both domains and the different collections used. We then described the indexing and retrieval techniques used. Finally we analyzed the obtained results and drew conclusions from the observations made.

2.2.1 Genomics

Information retrieval in the biomedical domain presents some particular challenges compared to information retrieval on newspaper collections for example. A major characteristic, proper to this domain, is for example the presence of several orthographic variants representing the same name [42]. In the 2006 Genomics track Abdou *et al.* [43] presented their method to extend queries with orthographic variants of the query terms. In 2007 we reused this work for query expansion and additionally completed it by using the WordNet[1] thesaurus to broaden the queries and extract specific words.

Since the requirement was to return relevant text passages rather then whole documents, we also had to deal with segmentation of the documents into passages. We used two different passage sets generated from the original collection. The first set was described by Abdou *et al.* [43], where passages are based on different HTML tags. The mean passage length for this set is 63 indexing terms per passage. The second set was made available

[1] http://wordnet.princeton.edu/

by the *Erasmus MC-University Medical Center Rotterdam*[2] and is based on a sentence level division (mean passage length 14 terms). The use of these two different passage sets, allowed us to compare retrieval performance on longer and shorter passages. To underline our results, we applied two probabilistic retrieval models, namely the Okapi model and the *InB2* model derived from the DFR Paradigm.

To be able to compare the different query expansion techniques as well as the two document segmentations, we performed different runs. First, we presented the results using only the single IR models and then combined these with the different query expansion options, such as spelling and WordNet thesaurus. From these results we drew different conclusions. For both models, the longer passage formulation performed better than segmentation at the sentence level. For both models we had a relative difference of about +113% for the Passage2 MAP (Okapi: 0.0190 vs. 0.0089, *InB2*: 0.2036 vs. 0.0952). Furthermore these results showed us that on this collection the *InB2* model performed better than the Okapi model.

A further observation was that while using the WordNet thesaurus for query expansion improved retrieval performance considerably, extending queries using orthographic variants slightly hurt performance. As implemented in our system the enlargement of queries with different spelling variants dropped the passage based MAP from 0.2533 to 0.2510 (-0.1%) for the *InB2* model. Expansion with synonyms however boosted the score from 0.2533 to 0.2777 (+9.63%) for the same model.

These experiments concluded our work on the biomedical domain.

2.2.2 Blog

The main objective of the Blog track was not only to promote search engines specialized in blog retrieval, but also to improve opinion detection in blogs. For our first participation in this task however, we focused on simple fact retrieval, i.e., retrieving documents containing information about a given target (opinionated or not). Our goal was to do a first performance comparison between different probabilistic models and different query lengths on the given collection. In a second stage we analyzed these results to identify characteristic issues related to blog retrieval. This would allow us in our future work to adapt our system to these domain-specific problems.

To fulfill these objectives, we evaluated the Okapi model, four models from the DFR Paradigm (*PL2*, *IneC2*, *InB2* and *PB2*) and one language model. All models have been evaluated on three different query lengths. While short queries only contained

[2] http://www.biosemantics.org

the title part of the topic, medium queries have been extended by the description part. The third and longest query formulation took into account all three topic parts (title, description and narrative). The obtained runs show a slightly better performance for the Okapi and *PL2* models, but all models perform on the same scale for all three query formulations. While the difference between short and medium query formulations is considerable (+12.5% in mean), increasing query formulations from medium to long does not necessarily improve retrieval performance (-2.2%).

In a second step we tried to figure out why our system failed for some queries. By analyzing the poorly performing queries, we realized that especially for topic-only queries, the main problem is the very high term frequency of one of the query terms in the retrieved but non-relevant documents. The used retrieval models would then boost these documents in front of relevant documents containing both query terms but less frequently. A particularity we noticed on the topic titles is that they often reference to a target by a noun phrase (e.g. "larry summer", "brokeback mountain" or "New York Philharmonic Orchestra"). We therefore concluded that in our future work, the topic titles should rather be considered as an entity, to ensure the presence of all equally important query terms.

2.3 "Stratégies de recherche dans la Blogosphère"

Our participation in the TREC Blog track 2007, described in the previous section and the corresponding paper, unveiled several problems related to IR in the Blogsphere. In the paper "Stratégies de recherche dans la Blogosphère" ("Searching strategies in the Blogsphere"), we presented these problems more thoroughly and proposed different approaches to face these domain-specific issues. We evaluated different retrieval models, various stemming procedures, a blind query expansion technique and an alternative indexing procedure to analyze their behavior on this specific collection. Furthermore we used two different performance measures to analyze the results, mean average precision (MAP) and mean reciprocal rank (MRR). All experimental runs were conducted on the Blogs06 collection using 100 queries, made available through the TREC 2006 and 2007 evaluation campaigns.

In the introductory section of this paper, we gave an overview of the Blogsphere with its particularities and difficulties and presented the document collection. We gave some statistics on the collection and how it has been generated, as well as examples of several documents and queries. In the remain of the paper we presented the indexing techniques and retrieval models used as well as the results of the different experiments conducted.

At first, we turned our attention to the analysis of different stemming strategies. We assumed that the collection of blogs we are facing deviates from the newspaper collections generally used for research purposes in information retrieval. While newspaper articles are generally written in well-formed and grammatically correct English, blogs are spontaneous and written in a more simple language. Furthermore the title part of the available queries is often very short, containing only one or two terms. Under these considerations, we concluded that no or a light stemming procedure might show better results than a more aggressive one.

To underline and prove these theoretic assumptions, we analyzed two different stemming strategies and compared them to an approach using no stemming. As a light stemmer, we used the algorithm proposed by Harman [44]. As a second more aggressive approach, we used the algorithm proposed by Porter [45] based on about 60 rules. We tested all three strategies using five different retrieval models, such as a vector-space model, a language model and three parametric probabilistic models.

The obtained results confirmed the expected results. For the Blogs06 collection all tested models show better performance than without stemming (neither light nor aggressive) or morphological analysis. Compared to both applied stemming strategies these differences are statistically significant. These results are particular for the given collection. Generally for the English language, stemming strategies show at least small overall improvements as shown for example by Hull *et al.* in [46] or [47], even though stemming might be more efficient on highly inflectional languages such as German.

The available queries were composed of three parts, a title, a description and a narrative part. This structure made it possible to compare the retrieval effectiveness of different query lengths. For the first query formulation, we used only the title part while for medium formulations we included also the description. Finally for the third and longest query formulation, we included all three topic sections. We evaluated the three query formulations using five retrieval models. We observed that longer queries improve retrieval results compared to shorter ones, but there is no real difference between medium and long query formulations. For medium length queries we have a mean improvement of +13.3% over short queries, while for the longest query formulation this improvement is just +12.7%.

In a next step, we evaluated Rocchio's blind query expansion approach. We expanded the queries using ten or twenty terms extracted from three, five or ten documents. The query expansion however did only change MAP and MRR values slightly, without any statistically significant difference to retrieval using no query expansion feature. We also

observed that MAP and MRR values are note necessarily related. While one query expansion approach might improve MAP values for example, the same approach eventually decreases MRR values.

In a second part of this paper, we presented our solutions to the problems uncovered during our participation in the last TREC Blog track. During the search for the failure causes of our systems for some queries, we realized that often the title part of the query should be considered as an entity rather than as a set of single terms. We therefore re-indexed the collection, completing the single indexing terms with couples of terms forming indexing units. The phrase "Big love in Paris" for example would generate following indexing units after stopword removal: "big love", "love paris", "big", "love", "paris". For short queries, i.e., considering only the title part, this new indexing strategy considerably improved retrieval of relevant queries (MAP from 0.3395 to 0.3657 (+7.7%) for the Okapi model). To apply a short stoplist containing only 9 words instead of 571 however does not improve retrieval results.

To conclude the paper we summarized the observations and gave a short outlook on future work that could be done in order to improve retrieval of relevant documents from blog posts.

2.4 UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere

In this paper we summarized our second participation in the TREC Blog track. As in the previous editions of this track, it could be separated into two parts, factual retrieval and opinion mining with a polarity detection subtask. We wanted to fulfill two objectives for our participation. The first goal was to merge all our observations made since our first participation in the Blog track to improve factual retrieval. Our second ambition was to set up a first, basic system for opinion mining and test it on the opinion retrieval and polarity detection tasks.

2.4.1 Factual Retrieval

The intent of the factual retrieval task was once again to retrieve relevant information about a given target entity. Based on our previous experiences, we decided to use two different indexing strategies. One index used only single words as indexing units, while the other also took account of compound constructions (word pairs respecting the order of the words). No stemming was used for both indexes. We also evaluated three

different retrieval models (Okapi, *PB2* and *PL2*) and two query expansion approaches. The first is a blind query expansion based on Rocchio's method, while for the second approach we used Wikipedia[3] to enhance queries. Finally we studied three different query formulations. The simplest query formulation uses only the title part, whereas the second employs also the description part. The last formulation uses also topic and description parts but the title part is formulated as phrase query. We did not use the narrative part to formulate queries, since our previous evaluations showed that long query formulations do not improve retrieval performance compared to medium length query formulations.

The results confirmed that using the compound indexing approach increases the performance. The same is valid for phrase queries. These results underline our assumption that it is important to keep the title part of the query together. We also observed that the performance for the three tested models is almost the same.

Accordingly to our results obtained in previous experiments concerning blind query expansion where we obtained only a small or no improvement using this expanding technique, we proposed to use an external source for query expansion (Wikipedia) rather then pseudo-relevance feedback. Each query title was send to Wikipedia and expanded with the ten most frequent terms from the first article returned. Using this method, we obtain an average improvement of +2.75% on MAP.

2.4.2 Opinion Retrieval and Polarity Detection

The opinion mining task was separated into two parts. First the participants' systems needed to identify blog posts containing an opinion. In a second step the polarity of each posts had to be detected, i.e., the posts had to be classified as containing positive, negative or mixed opinions. Since both tasks were closely related, we used one approach to classify the documents as positive, negative, mixed or neutral. A neutral document would then be considered as not opinionated and therefore be eliminated from the runs.

We proposed two different methods to classify the documents based on standard score (Z-score) and characteristic vocabulary. The idea was to determine which terms are characteristic for a given document polarity (positive, negative, mixed and neutral). Based on this vocabulary, we build an additive model and another one based on logistic regression to estimate the polarity. For each document to be classified and each possible polarity, our system calculated a value (depending on the used model) estimating the odds of the document having the given polarity. The polarity having the highest value is then attributed to the document.

[3] http://www.wikipedia.org/

For opinion retrieval we considered all documents classified as positive, negative or mixed as opinionated. The results showed us that adding an opinion detection feature to the baseline runs, does not necessarily improve results in our case. This is partly due to the fact that applying opinion detection eliminated documents considered neutral and thus the run contained less than the usual 1000 retrieved documents per query. However if these documents were wrongly eliminated, retrieval performance was hurt. We also observed that the additive model performs better than the logistic regression model. For the polarity task we made basically the same observations as for the opinion retrieval, however the MAP is relatively low compared to opinion retrieval.

If the results for factual retrieval were satisfactory, the results for opinion mining were not. Our next goal in the domain of blogs is therefore to adapt and improve our models for polarity and opinion detection.

The three papers presented so far document our work on the Blogsphere, from the initial exploring of the domain to the proposition of solutions to satisfy domain-specific issues and considerable improvements of retrieval performance. Our future work on this domain will lie in ameliorating our opinion mining techniques.

2.5 UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches

This paper summarizes our participation in the domain-specific track of the CLEF 2008 campaign. In this task, the participants' systems should retrieve relevant documents from a specific collection covering the social science domain. While this task included three languages, namely German, English and Russian, our main interest was in the German language and in presenting two new query expansion approaches.

From our previous participations in the CLEF domain-specific task [48], we learned that for the German language a decomposition algorithm should be applied before indexing due to its morphology. Furthermore, we applied our light stemmer and a stopword list containing 603 words. We used different parametric probabilistic retrieval models (Okapi and different variants from the DFR paradigm) as well as a language model and a vector-space model. We also used three different lengths of query formulations. We observe that all tested probabilistic models show statistically the same performance, while the vector-space model is less efficient. Once again, our results showed that for longer queries the retrieval results improve.

Additionally to the standard retrieval, we presented and applied two new query expansion approaches. The first one is based on Abdou's IDFQE approach ([11]). Using the

IDFQE algorithm, k terms from the top m documents are selected to expand the queries. The expansion terms are selected from the document, ignoring their position compared to the original query term. We assumed however that this relative position might be important and thus restrained the selection of expansion terms to a window of ten on both sides around the query terms in the top m retrieved documents. The idea is that related terms would be in a certain proximity to each other. Using the Okapi model and different values for k and m, we have a mean improvement of +2.44% compared to an approach using no query expansion and the Okapi model. To compare, using the IDFQE decreases results by -7.52% on the given collection.

The second query expansion approach extends the query using an external resource. The query title is send to the Google search engine and the first two snippets returned by the search engine are added to the query. This approach improves performance by +3.12%. In addition, we also presented our results on the English and Russian domain-specific tasks.

2.6 Comparison Between Manually and Automatically Assigned Descriptors Based on a German Bibliographic Collection

Our main goal of the paper presented in this section, was to empirically answer the question "Is it worth to spend human resources to manually add keywords to documents in order to improve retrieval results or does a manual expansion eventually show the same performance?". For this purpose, we studied the German GIRT collection, containing manually added descriptors for each document and coming with a domain-specific thesaurus.

After a short introduction and a brief overview on related work, we presented the corpus and the different IR models used. One of our objectives was to automatically extend documents and queries with keywords and thus needed to appropriately select keywords from the thesauri. We therefore decided to compare different similarity measures (Jaccard, mutual information and probabilities) presented at the end of the first, introductory part of this paper. In the second part we presented and analyzed the obtained results.

Based on the German GIRT Corpus, our first objective was to analyze the impact on retrieval performance if we take account of the available descriptors for retrieval. For each document a person knowledgeable in the domain added at least one keyword issued from a controlled vocabulary. In a first step we evaluated four different retrieval models (*tf idf*, language model, *InB2* and Okapi) on the documents excluding descriptors

from search. This run served as baseline to demonstrate the impact of keywords. We then perform retrieval over the documents including descriptors and observed that performance would improve considerably (between +10.6% and +17.94%) compared to an approach ignoring them. In view of this improvement, we wanted to analyze if eventually automatically added keywords might yield the same performance improvements.

For the manual expansion, we selected two different thesauri. First we used the domain-specific thesaurus for social science and second the general, freely available thesaurus, OpenThesaurus. This allowed us simultaneously to compare the performance differences of a general and a specific thesaurus. In our view, this was interesting because a domain-specific thesaurus might not always be available. To expand documents with the appropriate keywords, we selected the Jaccard similarity to calculate similarity between the documents and the thesaurus terms. Each document would contain at most n additional keywords. For our experiments, we fixed n at 50. This is the average number of manually added descriptors per document. Once again we applied all four retrieval models. The obtained results showed that this automatic expansion considerably hurt retrieval performance compared to an approach where no keywords are used. Degradation was between -30.47% and -20.94% for the GIRT-thesaurus and between -2.85% and -23.79% for the general thesaurus. The impact of keywords from the general thesaurus is less important and thus less hurting retrieval.

Furthermore we wanted to test the effect of extending queries rather than documents. We presented three measures used in natural language processing to calculate textual entailment between two terms. We used these measures to select expansion candidates. Using four retrieval models, we observe that all three measures perform the same and thus only presented results for the Jaccard similarity. The proposed query expansion technique does not bring any significant improvements. For the GIRT-thesaurus and short query formulations, the change for MAP values was between -0.23% and +0.78% while for the OpenThesaurus it varied between -0.29% and +0.62%. For longer queries, improvement was between +0.13% and +0.42% for specific thesaurus respectively between +0.08% and +0.28% for OpenThesaurus. By having a closer look at the results however, we observed that for some queries the impact of query expansion is quite high, just as well in a positive as in a negative way. For the specific thesaurus for example, we have an improvement for 52 queries and a decrease for 72 queries and no change for one query.

Finally, we concluded that it is worth to spend human resources to expand documents with keywords. Compared to an automatic expansion algorithm, a human person is able to take into account context and meaning of a document to select appropriate keywords. Thus these keywords would considerably improve retrieval. Furthermore,

we deduced that the benefit of query expansion is dependent on the given query and unpredictable. While for some queries, retrieval performance is much higher using query expansion, for others it considerably hurts the retrieval performance. Once again it is very complicated to choose the right terms to add to the original query in order to not hurt retrieval performance.

2.7 Adapting the *tf idf* Vector-Space Model to Domain-Specific Information Retrieval

The solutions generally proposed to improve domain-specific IR are either based on domain-specific features or involve external resources, such as thesauri or Wikipedia for example. Our aim in this paper was to present an extension to the well known *tf idf* retrieval model, which would on one side take account of domain-specific information, but on the other side be independent of the underlying domain. The objective was to present a model usable on all domains (and languages) without any adjustments.

In the two introductory sections we presented the general problems of domain-specific IR and related work conducted to face these challenges. We then presented our extension to the *tf idf* model.

The main idea was to add one component to the model in order to take account of the specificity of each query term. The assumption is that specific terms in the query are more important than general ones and thus their presence in the retrieved documents should be weightier.

We presented a total of four adapted vector-space models, based on different measures issued from information theory and corpus based linguistics. Since the proposed models are supposed to be domain and language independent, we evaluated them on three different domains (social science, Blogsphere, biomedicine) and two languages (English and German). Finally we also evaluated the standard *tf idf* and Okapi models, to compare our new approach to these familiar models.

The first observation we made, was that all models clearly outperform the classical *tf idf* approach. For the adapted models we had a mean improvement between 59.13% and 66.13% depending on the extension measure used. For the biomedical collection, three out of the four newly presented models showed a statistically similar performance as the Okapi model. For the German social science corpus all four models even showed a statistically significant better performance then the Okapi model. For the two other corpora the adapted models performed slightly worse than the Okapi model.

These first, very promising results, allowed us to conclude that it is worth to take account of specificity for term weighting. The advantage of the presented approaches is their autonomy. The models can easily be used on any collection and language. A second advantage is the absence of parameters compared to several other retrieval models, such as Okapi or models based on the DFR framework.

2.8 Algorithmic Stemmers or Morphological Analysis: An Evaluation

This last paper contributing to the presented thesis breaks ranks. The topic is not exactly about domain-specific IR, however the presented results and evaluations might as well be relevant for domain-specific IR.

The main objective of this paper was to present an evaluation of different stemming approaches and a morphological analysis on the English language. We studied the impact of algorithmic stemmers as well as a morphological approach. Additionally, we also evaluated the benefit of word sense disambiguation techniques and stopword removal. All evaluations were done on a relatively large set of queries to underline the obtained results.

In IR stemming is a common approach and considered to be an effective mean to enhance retrieval performance. For the English language, different algorithmic stemming strategies have been proposed. Various previous studies already investigated and proved the utility of applying stemming strategies. In our evaluation, we wanted to see if for a large set of queries these results still hold and if suffixing is really better than a non stemming approach. We also compared different available stemmers for the English language.

A second approach, is to apply a morphological analysis rather than use an algorithmic stemmer. In this case, instead of taking the stem of a given term as indexing unit, we used the corresponding lemma (dictionary entry). This raised a second question. Is a morphological analysis more efficient than an algorithmic stemmer? We tried to analyze this question in our evaluation.

Accessorily, we wanted to see if the use of word sense disambiguation (WSD) such as thesaurus class numbers or part of speech (POS) would prove useful to increase retrieval effectiveness.

All our evaluations were based on the English test-collections used during the CLEF ad hoc tasks from 2001 to 2006. Additionally in 2008 WSD data and morphological

analysis was made available for the given collection. Out of the 310 collected topics, we used 284 for our evaluation, having at least one relevant document.

We tested five different retrieval models and four different stemmers having different levels of aggressiveness (S-Stemmer, Porter, Lovins and SMART). We compared the stemmers to an approach using no stemming and to one using lemma as indexing units. The difference between stemming and morphological analysis is that when applying a stemming algorithm to a word we transform it into a stem not necessarily found in a dictionary while when applying a morphological analysis we obtain a lemma representing a dictionary entry. We can thus not directly compare stemming and morphological analysis but we can compare their effect on information retrieval.

The first observation we made was that all approaches would perform better than if no stemming and no morphological analysis is used. Mean improvements are between +4.9% (Lovins) and +9.2% (SMART). Except for Lovins' stemmer, all approaches performed significantly better than the non stemming approach. For the morphological approach (lemma) we had a mean improvement of +7.1%.

When comparing the different algorithmic stemmers and the morphological analysis, we selected the SMART stemmer as baseline to compare the other approaches. We observed that only the Lovins' stemmer showed a statistically significant worse performance. The second observations was that the morphological analysis did not clearly outperform any algorithmic stemming approach.

In the second part, we evaluated two different WSD techniques, POS and thesaurus. To each lemma was associated a POS tag and a Synset number, representing the corresponding number of the thesaurus class in the WordNet thesaurus. Both information depended on the context. We included this information in our retrieval and evaluated four different runs (lemma, lemma & POS, lemma & Synset and lemma & Synset & POS). While if including only POS we have a mean improvement of +1.5% (5 models), if including Synset we have a decrease of -3.4% compared to using only lemma. The POS information thus showed more benefit than a thesaurus based WSD.

Finally we evaluated the impact of stopword removal. We therefore tested three approaches, one using no stoplist, the second using a short stoplist composed of nine words and finally a long stoplist coming from the SMART retrieval system containing 571 words. Taking the SMART list as a baseline, we have a decrease of -0.6% if using a short stoplist and of -14.5% using no stoplist.

Based on the presented empirical study, we finally concluded that for the English language stemming is important even though if for other languages such as German or Finnish the impact is much higher. This is due to the more complex morphology of

these languages. However it is also important to consider the end-user and not use a too aggressive stemming approach since this obscures the results (e.g., Lovins). Using WSD techniques did not really improve results compared to using only a morphological analysis. Furthermore we concluded that it can however be useful to remove stopwords during indexing, even if the stopword list is very short.

Chapter 3

Conclusion

In this final chapter of the thesis, we will recapitulate our conclusions and give an outlook on possible future work which might follow this thesis.

3.1 Contributions and Conclusions

In the previous chapter we presented an overview of our research done in domain-specific information retrieval. Our main focus laid on three different domains, namely social science, biomedicine and Blogsphere. We did however not restrained our work to these domains. In this section, we will first summarize our contributions for the three main domains. In the second part we will present more general conclusions derived from our work and analyze if and how far the objectives of this thesis have been fulfilled.

Blogsphere

Information retrieval in the Blogsphere presents several challenges, such as factual retrieval, opinion mining, recurring topic search or faceted IR. In our work on this domain, we mainly concentrated on factual retrieval but also presented a first approach for opinion mining and polarity detection. Our research in the Blogsphere contributed the largest part to this thesis.

In a first step we tested several well known and widely used retrieval models to analyze their behavior on the Blogsphere and evaluated various query lengths and formulations. We could observe that models performing well on general collections, such as Okapi or models issued from the DFR framework, would also outperform other models (language model, vector-space model) on this specific collection. The same also applies for query

lengths. Medium length queries show the best retrieval results. Furthermore we evaluated different query expansion techniques using either pseudo-relevance feedback or external resources, such as Wikipedia or Google. On the Blogs06 collection all query expansion techniques show only slight or no improvement, but the query expansion techniques using external resources perform better than pseudo-relevance feedback.

In a second step we analyzed the obtained results and brought out various problems and characteristics of the domain. We realized for example that it might be efficient to ignore stemming and consider topic titles as an entity and take account of the order of the words in the query. By considering these properties and including them to the retrieval process, we could actually considerably improve retrieval of relevant blog posts.

Being consequently able to provide a relatively stable and efficient baseline, we finally dared a first opinion mining approach based on characteristic vocabulary. Even tough the results of this approach are not yet completely satisfactory, they are promising and allowed a first valuable insight in the field of opinion mining.

Social Science

Our work in the domain of social science was mainly motivated by the particularities of the collection and the availability of the domain-specific thesaurus for social science. Each document contains manually added descriptors selected from a controlled vocabulary. We analyzed the impact of these keywords on retrieving relevant documents, concluding that they considerably improve retrieval. Furthermore the available thesaurus allowed us to evaluate automatic document and query expansion. We consequently could compare automatic and manual document expansion and concluded that manual expansion clearly outperforms automatic expansion, at least as implemented and evaluated in this thesis. We assert that when tasks are complex, such as taking into account context information, humans still outperform automatic algorithms.

In the same series of evaluations, we compared a specific and a general thesaurus and their impact on document and query expansion. The results showed that in our case no real difference could be seen. This might however be different in a domain using a more scientific and complex language such as in biomedicine or mathematics.

We also developed and tested a new blind query expansion technique on this collection based on the idea that the terms used for expansion should be close to the original query terms in the pseudo-relevant documents. We could slightly improve results using this approach. Furthermore the German language added a second difficulty to our studies on this domain, which we solved by using available stemming and decomposition algorithms. Especially decomposition would considerably improve retrieval performance.

Finally we evaluated various models on the English and German social science collections. The probabilistic retrieval models (Okapi, DFR, language model) perform on a similar level, while the vector-space model proves to be less efficient.

Biomedicine

In the biomedical domain, we mainly evaluated different query expansion techniques using external resources such as for example WordNet thesaurus and various retrieval models. Once again we observed that the same models as for general IR show the best retrieval performances. We also investigated on different query and text segmentation lengths and their impact on retrieval. We observed that longer text passages and queries improve retrieval effectiveness.

General Conclusion

This first part of our research fulfills the first and major objective of this thesis, namely to analyze various domains, evaluate the effectiveness of standard and widely used retrieval procedures on these domains, use various external resources (Google, Wikipedia, WordNet, OpenThesaurus, ...), figure out particularities of the respective domain and finally adapt the system to meet domain-specific issues.

So far we can conclude that approaches working well in general still do so in domain-specific IR. For example probabilistic retrieval models, such as Okapi, the language model or models issued from the DFR paradigm generally outperform vector-space approaches. Depending on the collection and query lengths either Okapi either one of the DFR models performs best but generally these differences are not statistically significant. The used language model approach would always lag a bit behind but still outperforming the used vector-space model. We observed that also in domain-specific information retrieval longer query formulations tend to improve retrieval effectiveness. In some special cases however, an approach proven to improve retrieval in general should not be applied on a specific domain. For example for retrieval on the Blogsphere applying even a light stemming strategy decreases retrieval performance. The main conclusion is therefore that in a first step it is important to have an efficient baseline retrieval. For the baseline retrieval general information retrieval approaches can be applied. In a second step a domain-specific layer can then be added.

We also evaluated several query expansion approaches on the various domains and observed that in general query expansion is not as efficient as it has been showed to be

on general domains. Blind query expansion showed either only minimal or no improvement at all. In the cases where it failed, query expansion using external resources would eventually improve retrieval effectiveness.

Our second main objective was to propose a domain-specific retrieval model, independent of the language and the domain, but taking account of domain-specific information. We proposed a first approach satisfying these conditions. Based on different information measures, we extended the standard vector-space model in order to take account of the specificity of the query terms. This model can be easily applied to various domains and languages and has been tested on three domains and two languages (English and German). The first results were promising and motivate further research in this direction.

With our first opinion mining and polarity detection approach, we also partially fulfill the objective to go beyond simple document retrieval.

Even though we mainly concentrated on the three described domains, we also worked on various other domains and bordering areas of domain-specific IR. We investigated for example in retrieving relevant information from bibliographic records extracted from The European Library (TEL) in English, German and French ([49]) or in establishing prior arts search on intellectual property ([50]). Finally as already described in the previous chapter, we evaluated different algorithmic stemmers and a morphological analysis for the English language ([51]).

We presented our work on different conferences and participated in various evaluation campaigns such as TREC and CLEF and thus also reached our last objective. These participations allowed us to compare our techniques with other approaches and discuss the obtained results with other participants.

3.2 Future Work

Finally, to conclude this thesis we will give an outlook on possible future work.

As we have seen, in domain-specific IR it is a relatively common approach to study a given domain and elaborate its characteristics to adapt existing retrieval systems in order to improve retrieval effectiveness. It is also quite common to use domain specific thesauri to enlarge queries and documents in order to improve retrieval. Effectiveness of these approaches varies, but generally retrieval can be improved if the available system has been sufficiently tuned to satisfy the needs of the underlying domain and if the baseline system itself is efficient. Since results are generally satisfying, this tradition

will probably be continued as new domains will be studied. Other examples of such domain-specific IR procedures currently worked on are for example IR models based on ontologies ([52], [53]) mainly used in the Semantic Web or context sensitive stemming ([54]).

We think however that it might also be interesting and worthwhile to investigate further in the direction to develop a domain-specific retrieval model independent of the underlying domain. Such a model would avoid considerable knob tuning and be easily deployable on new domains. Our first research in this direction, by adapting an existing model, showed promising results. This model however still carries some disadvantages and discrepancies which might be eliminated in prospective work.

An other field in research which in our opinion merits further attention is opinion mining and polarity detection, not only in the Blogsphere. Being promoted by several evaluation campaigns such as TREC and NTCIR, opinion mining is not only an interesting field of research but it might also be interesting for a producer or a market analyst. After a new product has been launched, the producer might want to know the feedback of the first customers in order to better satisfy the needs of future consumers and eventually adapt the product in a second (third, fourth, ...) version. He/she would therefore for example search blog posts containing opinions on his product and perform polarity detection to detect negative and/or positive features of the product. We proposed a first polarity detection algorithm, still being in its first stage but showing prospective results. This algorithm might also be further developed in our forthcoming work. One possibility might for example be to use approaches from natural language processing or from machine learning rather than from statistics or combine several of these approaches.

An other interesting angle to investigate in the Blogsphere would be to implement and evaluate models which do not make the term independence assumption such as for example dependence language models ([55]) or Markov Random Field models (MRF, [56]). We showed that for the blog queries, the title part should be considered as an entity and that the query terms are related in this particular case and thus assume that models not ignoring term dependencies might improve retrieval effectiveness. For a first empirical evaluation to see if this assumption holds, we implemented a MRF model combined with the Okapi model and could observe that as a matter of fact on this domain we could improve the results.

Last but not least our interest in domain-specific information retrieval will eventually go toward prior art search in intellectual property (IP). IP summarizes copyright laws and patents in various domains, commercial or not. A person or team wanting to submit their work for protection under intellectual property will do a prior art search to find similar and related work to avoid violating existing copyright laws or patents. Or before

granting a patent, the European Patent Office will for example do a *novelty search* to establish the novelty of a patent's claim. A prior art search can also simply be conducted to get an overview on existing inventions on a specific filed. A main challenge in prior art search in the IP domain is that the queries are generally whole documents, already formulated as a patent which have to be matched against a collection of patents. Our first step in this domain, focused on formulating an appropriate query out of the query document. In future work, we might want to continue on this angle but also explore several other ideas such as summarizing documents or multilingual searches since patents might be written in various languages.

As this short outlook shows, research in domain-specific information retrieval and information retrieval in a broader sense is not yet exhausted. In contrary, different compelling challenges are still to be solved and new ones will certainly arise out of the growing availability of electronic data, the growth of the World Wide Web and various other causes, eventually yet unforeseeable.

Appendix A

Selected Papers

IR-Specific Searches at TREC 2007: Genomics & Blog Experiments

Claire Fautsch, Jacques Savoy
Computer Science Department, University of Neuchatel
Rue Emile-Argand, 11, CH-2009 Neuchatel (Switzerland)
{Claire.Fautsch, Jacques.Savoy}@unine.ch

ABSTRACT

This paper describes our participation in the TREC 2007 Genomics and Blog evaluation campaigns. Within these two tracks, our main intent is to go beyond simple document retrieval, using different search and filtering strategies to obtain more specific answers to user information needs. In the Genomics track, the dedicated IR system has to extract relevant text passages in support of precise user questions. This task may also be viewed as the first stage of a Question/Answering system. In the Blog track we explore various strategies for retrieving opinions from the blogsphere, which in this case involves subjective opinions about various targets entities (e.g., person, location, organization, event, product or technology). This task can be subdivided in two parts: 1) retrieve relevant information (facts) and 2) extract positive, negative or mixed opinions about the specific entity being targeted.

To achieve these objectives we evaluate retrieval effectiveness using the Okapi (BM25) and various other models derived from the *Divergence from Randomness* (DFR) paradigm, as well as a language model (LM). Through our experiments with the Genomics corpus we find that the DFR models perform clearly better than the Okapi model (relative difference of 70%) in terms of mean average precision (MAP). Using the blog corpus, we found the opposite; the Okapi model performs slightly better than both DFR models (relative difference around 5%) and LM (relative difference 7%) model.

1. INTRODUCTION

The biomedical domain presents the information retrieval (IR) community with a number of challenging problems. For the first Genomics campaign [1] for example the main objective was to retrieve bibliographic references (composed mainly of title, author names and abstract) from a large subset of the MEDLINE repository, in order to meet real user needs. Last year [2], the main goal was to retrieve text fragments or passages rather than the entire scientific article. From an IR point of view, this task lies somewhere between classical text retrieval in which search responses consists of documents (or references to these documents) and question/answering where responses consist of very short passages extracted from documents. The term "passage" is in fact not very precise, given it could refer to a paragraph, sentence, or a short window of *n* characters.

For the Blog track [3], the IR system has to retrieve relevant information from different permalink documents (URLs pointing to a specific blogging entry), representing various points of view on various domains. Unlike traditional document collections used in the IR domain, a blog is more subjective, while also being characterized by more diverse document structures and writing styles. Even though the blogsphere may contain objective information (facts), the objective of the Blog track is to find answers based on opinions rather than relevant factual information. As such, relevant answers to the request "IPhone" may include factual and technological information (relevant but unopinioned answers) but also more personalized (and subjective) aspects of the product (why it is useful, complaints about this new tool, drawbacks of using a specific function, personal experiences concerning new product, etc.). Thus, in a first step the answer would contain a ranked list of relevant documents, but in a second stage a classification procedure would subdivide them into documents not based on opinion (factual information or descriptions), or documents expressing positive, mixed or negative opinion about the target entity.

The rest of this paper is organized as follows. Section 2 depicts the main characteristics of the Genomics test-collection and how passages are derived from an article according to our definition while Section 3 describes the main features of the Blog test-collection. Section 4 describes the indexing approach and Section 5 briefly presents the three probabilistic models used to search the genomics or blogsphere. Section 6 evaluates the three IR models by applying different conditions. Finally, the main findings of this paper are presented in Section 7.

2. GENOMICS TEST-COLLECTION

The document collection used this year contains approximately 12 GB of uncompressed data, made up of 162,259 full-text publications extracted from 49 biomedical journals (for more details, see the Web site at http://ir.ohsu.edu/genomics/2006data.html). To facilitate the effective retrieval of relevant passages and not documents, the IR literature [4] defines passages according to their various types, based mainly on delimiters such as text, window or semantic markers.

In a first approach to defining passages, we processed each article in order to generate its corresponding passages. As passage delimiters, we assigned the following HTML tags: H1, H2, H3, H4, H5, H6, P, BR, HR, TABLE, TD, TH, TR, OL, and UL.

```
<PASSAGE>
<FN> /raid/Genomics/peds/12118078.html
<ID> 12118078.23
<SO> 28541
<L> 978
<TGN> p
<R> false
<TITLE> Alterations in the Mouse and Human
Proteome Caused by Huntington's disease
<TX> In addition to the cytoplasmic brain
fraction that was used in the above experiments,
proteins solubilized by urea and detergent
treatment, yielding an extract enriched in
membrane proteins, as well as DNA-binding
proteins released by DNase, were screened to
expand the range of protein classes studied. In
both fractions no additional proteins were
consistently different between R6/2 and control
mice (data not shown). AAT was present at low
amounts in the membrane fraction and
undetectable in the fraction of proteins
released by DNase in control mice, arguing for a
mainly cytoplasmic localization of the protein
(data not shown). ABC was found in all three
fractions. A consistently lower expression of
ABC and AAT expression below the detection limit
were detected in R6/2 samples in all three
fractions (data not shown).
</PASSAGE>
```

Figure 1. Example of generated passage

Figure 1 shows an example of a passage that might be generated. All our passages are structured according to the following set of fields.

- FN (article filename path),
- ID (passage identifier),
- SO (start offset),
- L (passage length in bytes),
- TGN (tag name from which the passage was extracted),
- R (indicates whether or not the passage is identified as a reference),
- TITLE (title of article),
- TX (passage contents).

Following the filtering of all passages containing fewer than 10 words, the resulting collection contained exactly 10,700,925 passages from which 1,275,132 (11.9%) were marked as references.

For a second passage definition we used the sentence level and reused the subdivision structure applied at Erasmus MC - University Medical Center Rotterdam (the Netherlands) (see the Web site www.biosemantics.org).

This collection consisted also of 36 topics (numbered #200 to #235) corresponding to the real information needs commonly expressed by biologists (see Figure 2 for examples). Each topic relates to one of the 14 possible biological entity types (e.g., antibodies, diseases, mutations, pathways, tumor types, signs or symptoms). This information could thus be used to automatically enlarge the submitted query.

```
<ID> 200
<QUESTION> What serum [PROTEINS] change
expression in association with high disease
activity in lupus?

<ID> 214
<QUESTION> What [GENES] are involved axon
guidance in C.elegans

<ID> 232
<QUESTION>  What [DRUGS] inhibit HIV type 1
infection?
```

Figure 2. Examples of three topics (genomics corpus)

3. BLOG TEST-COLLECTION

The Blog test collection contains approximately 148 GB of uncompressed data, made up of 4,293,732 documents extracted from three sources: 753,681 feeds (or 17.6%), 3,215,171 permalinks (74.9%) and 324,880 homepages (7.6%). Their size is as follows; 38.6 GB for feeds (or 26.1%), 88.8 GB for permalinks (60%) and 20.8 GB for the homepages (14.1%). In this evaluation campaign only the permalink part is used. This corpus was crawled between Dec. 2005 and Feb. 2006 (for more information see: http://ir.dcs.gla.ac.uk/test_collections/).

Figure 3 depicts two examples of blog documents, showing their date, URL source and permalink structure at the beginning of each document. Some information extracted during the crawl is placed after the **<DOCHDR>** tag. Additional pertinent information follows after the **<DATA>** tag, along with ad links, name sequences (e.g., authors, countries, cities) plus various menu or site map items. Finally there is some factual information, such some of the locations where various different opinions can be found.

```
<DOC>
<DOCNO> BLOG06-20051212-051-0007599288
<DATE_XML> 2005-10-06T14:33:40+0000
<FEEDNO> BLOG06-feed-063542
<FEEDURL> http://
contentcentricblog.typepad.com/ecourts/index.rdf
<PERMALINK>
http://contentcentricblog.typepad.com/ecourts/20
05/10/efiling_launche.html#
<DOCHDR> …
Date: Fri, 30 Dec 2005 06:23:55 GMT
Accept-Ranges: bytes
Server: Apache
Vary: Accept-Encoding,User-Agent
Content-Type: text/html; charset=utf-8
…
<DATA>
electronic Filing &amp;amp; Service for Courts
…
October 06, 2005
eFiling Launches in Canada
Toronto, Ontario, Oct.03 /CCNMatthews/ -
LexisNexis Canada Inc., a leading provider of
comprehensive and authoritative legal, news, and
business information and tailored applications
to legal and corporate researchers, today
announced the launch of an electronic filing
pilot project with the Courts
…
```

Figure 3. Example of LexisNexis blog page

```
<DOC>
<DOCNO>  BLOG06-20060212-023-0012022784
<DATE_XML> 2006-02-10T19:08:00+0000
<FEEDNO> BLOG06-feed-055676
<FEEDURL>  http://
lawprofessors.typepad.com/law_librarian_blog/ind
ex.rdf#
<PERMALINK>
http://lawprofessors.typepad.com/law_librarian_b
log/2006/02/free_district_c.html#
<DOCHDR> …
Connection: close
Date: Wed, 08 Mar 2006 14:33:59 GMT …
<DATA>
Law Librarian Blog

Blog Editor
Joe Hodnicki
 Associate Director for Library Operations
 Univ. of Cincinnati Law Library
…
News from PACER   :

&amp;amp;quot;In the spirit of the E-Government Act
of 2002, modifications have been made to the
District Court CM/ECF system to provide PACER
customers with access to written opinions free
of charge

The modifications also allow PACER customers to
search for written opinions using a new report
that is free of charge. Written opinions have
been defined by the Judicial Conference as
&amp;amp;quot;any document issued by a judge or
judges of the court sitting in that capacity,
that sets forth a reasoned explanation for a
court's decision.&amp;amp;quot; …
```

Figure 4. Example of blog document

During this evaluation campaign a set of 50 topics (Topics #901 to #950) was created from this corpus. Like last year (Topics #851 to #900) they express user information needs extracted from a commercial search engine blog log, such as the examples shown in Figure 5.

```
<ID>  916
<TITLE>  dice.com
<DESC>  Find opinions concerning dice.com,
an on-line job search site.
<NARR>  Opinions on dice.com's effectiveness
are relevant.  Mention of its problems is
relevant. Recounting an experience using
dice.com is relevant. Simply mentioning it
as a possible tool is not relevant.

<ID>  928
<TITLE>  "big love"
<DESC> Find opinions regarding the HBO
television show "Big Love".
<NARR>  All statements of opinion regarding
the HBO production "Big Love" are relevant.
Statements of opinion about HBO or actors
in the show are relevant provided that "Big
Love" is mentioned.

<ID>  937
<TITLE>  LexisNexis
<DESC> Find opinions about the information
service LexisNexis.
<NARR> Relevant documents will provide
opinions about the information service
LexisNexis. Documents that are obviously
sponsored by LexisNexis are considered to
be spam and not relevant.
```

Figure 5. Three examples of Blog track topics

Based on relevance assessments (relevant facts & opinions, or relevance value ≥ 1) made on this test collection, we listed 12,187 correct answers. The mean number of relevant web pages per topic is 243.74 (median: 208; standard deviation: 186.0). Topic #939 ("Beggin' Strips") returned the minimal number of pertinent passages (16) while Topic #903 ("Steve jobs") produced the greatest number of relevant passages (710).

Based on opinion-based relevance assessments ($2 \leq$ relevance value ≤ 4), we found 7,000 correct opinions. The mean number of relevant web pages per topic is 140.0 (median: 109.5; standard deviation: 123.456). Topic #910 ("Aperto Networks") and Topic #950 ("Hitachi Data Systems") returned a minimal number of pertinent passages (4) while Topic #903 ("Steve jobs") produced the most relevant passages (496).

The polarity of opinions pertaining to target entities could be divided into three groups: negative (relevance value = 2), mixed (relevance value = 3) or positive (relevance value = 4) opinion. From an analysis of negative opinions only (relevance value = 2), we found 1,844 correct answers (mean: 40.087, median: 22.5, min: 1 (Topic #909 "Barilla", #934 "cointreau", #948 "sorbonne" or #950 "Hitachi Data Systems"), max: 189

(Topic #912, "nasa"), standard deviation: 45.12). Topic #901 ("jstor"), #910 ("Aperto Networks"), #914 ("northernvoice") and #925 ("mashup camp") obtained no positive opinions.

For positive opinions only (relevance value = 4), we found 2,960 correct answers (mean: 59.2, median: 49.5, min: 1 (Topic #950, "Hitachi Data Systems"), max: 234 (Topic #903, "Steve jobs"), standard deviation: 53.98). Finally for mixed opinions only (relevance value = 3), we found 2,196 correct answers (mean: 47.74, median: 22, min: 1 (Topic #901, "jstor", and Topic #925, "mashup camp"), max: 196 (Topic #946, "tivo"), standard deviation: 50.74).

4. INDEXING APPROACHES

To index documents or queries, we applied the indexing method described in Section 4.1. To derive orthographic variations of protein or gene names that could be included in topics, we used the algorithm described in Section 4.2.

4.1 Document Indexing

As a natural approach to indexing and searching both corpora, we chose words as the indexing units. As such our lexical analyzer applies the followings steps to process the input. First, the text is tokenized (using spaces or punctuation marks), simple acronyms are normalized (e.g., D.N.A. is converted into DNA) and hyphenated terms are also broken up into their components. For example, a word such as "COUP-TF1" generates three different forms, namely "COUP", "TF1" and the original form "COUP-TF1". Second, uppercase letters are transformed into their lowercase forms. Third, stopwords are filtered out using the SMART list (571 entries). Fourth, with the *S-stemmer* algorithm [5] based on three rules, we remove the final '-s' (the most common plural suffix for the English language). This choice is based on the experiments we did over previous years [6], [7] which demonstrate that out of the four evaluated stemmers (Lovins, *S-stemmer*, Porter and SMART) the *S-stemmer* provided the best retrieval effectiveness.

For the Blog task we also considered a second tokenization procedure. For example we noticed that in certain blogs there are rather long sequences of identical letters such as "aaaaah" and thus we retained only the first three letters, transforming it into "aaah".

4.2 Generation of Orthographic Variants

As is known, in biomedical literature there can be several orthographic variants [8] representing a given name, generally introduced for a variety of reasons:

1) Typographic errors and misspellings (e.g. "retreival" and "retrieval") or cognitive (e.g., "ecstasy", "extasy", or "ecstacy"; "occurence" or "occurrence");
2) Alternative punctuation and tokenization, mainly due to the lack of a naming convention (e.g. "Nur77", "Nurr-77" or "Nurr 77");
3) Regional language variations, such as British and American English (e.g. "colour" or "color", "grey" or "gray", etc.)
4) Transliteration of foreign names (e.g., "Crohn" and "Krohn" or "Creutzfeld-Jakob" and "Creutzfeldt-Jacob");
5) Morphological variations (inflections or derivations) which could be resolved by using a stemmer.

During previous TREC campaigns, many methods were proposed for resolving problems with orthographic variations, as for example [9]. The algorithms proposed were usually rule-based and were essentially concerned with secondary causes such as those described above (e.g., see [10]).

In order to automatically find a ranked list of alternative spellings for each search word, we modified the Lucene [11] Spell Checker[1]. In its initial stage this tool required a lexicon containing the correct spelling, so in our case we used the words extracted from the TREC 2005 corpus, a large subset of the MEDLINE collection. We then introduced a single term or a short sequence of words, limited in the current case to two terms. The spellchecker thus responded by returning a ranked list of the top 100 hits extracted from the given lexicon. In our case we used the following formula to re-ranked this list according to the minimal *edit-distance* measure and its length, calculated for each candidate considered a variant of the original (misspelled) term submitted:

Score = 1 – [edit-distance / length(term)]

When the two similar candidates were deemed to be equal (which occurred relatively frequently), they were ordered according to popularity (or *df*, document frequency), ranging from most to less frequent.

For each topic available in this TREC campaign, we submitted each search word or group of two successive words to the spellchecker engine. As shown in Figure 6, the spelling candidates were then re-sequenced by the *edit* and *df* measure and automatically added to the topic following the <BISPLELL-*n*> tag (followed by the alternative number).

In Figure 6, the *input* attribute describes the term submitted to the spellchecker. The *score* attribute refers to the final score achieved by the alternative term.

We then used the WordNet thesaurus to automatically enlarge the query. As shown in Figure 6 for the entity in question and the tag <ENTITY-EXPANSION> we could add

[1] http://wiki.apache.org/jakarta-lucene/SpellChecker

synonyms (e.g. "dna" for Topic #214) or morphologically related terms (e.g., "signal signaling signalize singnalise" to the term "signal"), and modifications such as these were made for 30 out of 50 queries. Finally for the tag <MEDICAL-TERM> we added synonyms from the question words extracted from the WordNet thesaurus. The number of added synonyms is relatively low (e.g., 20 words for the 50 queries under the tag <MEDICAL-TERM>).

```
<ID>  200
<ENTITY>  PROTEINS
<ENTITY-EXPANSION>
<QUESTION>  What serum PROTEINS change
expression in association with high disease
activity in lupus
<MEDICAL-TERM>
<BISPELL-1 input="serum proteins" score="0.86"
freq="1">  serum-proteina
<BISPELL-2 input="serum proteins" score="0.85"
freq="15">  serum-protein
<BISPELL-1 input="disease activity" score="0.94"
freq="3">  disease-activity

<ID>  214
<ENTITY>  GENES
<ENTITY-EXPANSION>  dna
<QUESTION>  What GENES are involved axon
guidance in C.elegans
<MEDICAL-TERM>
<BISPELL-1 input="axon guidance" score="0.92"
freq="5">  axon-guidance
```

Figure 6. Example of two topics, their orthographic variants and their WordNet expansions

5. RETRIEVAL MODELS

In our evaluations we conducted experiments by applying the single IR models described in Section 5.1 or by merging the result lists computed by various single IR models as explained in Section 5.2 (data fusion).

5.1 Single IR Models

To begin our evaluation we considered three probabilistic retrieval models. As a first approach, we used the Okapi (BM25) model [12], evaluating the document D_i score for the current query Q using the following formula:

$$Score(D_i, Q) = \sum_{t_j \in q} qtf \cdot \log\left(\frac{n - df_j}{df_j}\right) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}, \quad (1)$$

$$\text{where } K = k_1 \cdot \left[(1 - b) + b \cdot (l_i / avdl)\right]$$

in which the constant *advl* was fixed at 839 for the Blog corpus and 14 with sentences (Genomics) or 63 with our passage delimitation (Genomics), *b* was set to either 0.4 (Blog), 0.55 (Genomics, passages), or 0.35 (Genomics, sentences) and $k_1 = 1.4$ (Blog) or 1.2 (Genomics).

As a second approach, we implemented various models derived from the *Divergence from Randomness* (DFR) paradigm [13]. In this case, the document score was evaluated as:

$$Score(D_i, Q) = \sum_{t_j \in q} qtf \cdot w_{ij} \quad (2)$$

where *qtf* denotes the frequency of term t_j in query Q, and the weight w_{ij} of term t_j in document D_i is based on combining two information measures as follows:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf))$$

As a first model, we implemented the PB2 scheme, defined by the following equations:

$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{with } \lambda_j = tc_j / n \quad (3)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)] \quad (4)$$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* is the average document length (fixed at 839 for the Blog, or 63 for the Genomics), *n* the number of documents in the corpus, and *c* a constant (= 5 for the Blog or the Genomics sentences or to 9.5 for the Genomics passages).

For the second model PL2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and $Prob^2_{ij}$ by Equation 4, as shown below:

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad (4)$$

where λ_j and tfn_{ij} were defined previously.

For the third model called IneC2, the implementation is given by the following two equations:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e + 0{,}5)] \quad \text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \quad (5)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))] \quad (6)$$

where *n*, tc_j and tfn_{ij} were defined previously, and df_j indicates the number of documents in which the term t_j occurs.

A third approach we considered was based on a statistical language model (LM) [14], [15], where probability estimates would be estimated directly, based on occurrence frequencies in document D_i or corpus C. According to this language model paradigm, various implementation and smoothing methods could be considered, although in this study we adopted the model proposed by Hiemstra [15] as described in Equation 7, combining an estimate based on document ($P[t_j | D_i]$) and on corpus ($P[t_j | C]$).

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1 - \lambda_j) \cdot P[t_j | C]]$$

$$\text{with } P[t_j | D_i] = tf_{ij}/l_i \text{ and } P[t_j | C] = df_j/lc \text{ and with } lc = \sum_k df_k \quad (7)$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and *lc* an estimate of the size of the corpus C.

5.2 Combining Different IR Models

It is assumed that combining different search models would improve retrieval effectiveness, due to the fact that each document representation might retrieve pertinent items not retrieved by others and thus increase overall recall [16]. In this current study we combined three probabilistic models representing both the parametric (Okapi and DFR) and non-parametric (language model or LM) approaches. Various fusion operators have been suggested to perform these combinations, such as the "Sum RSV" operator, where the combined document score (or the final retrieval status value) is simply the sum of the retrieval status value (RSV_k) for the corresponding document D_k computed by each single indexing scheme [17].

$$\text{Z-score } RSV_k = [((RSV_k - Mean^i) / Stdev^i) + \delta^i],$$
$$\delta^i = ((Mean^i - Min^i) / Stdev^i) \quad (8)$$

This year, we only used the Z-Score operator (shown in Eq. 8) to combine two or more single runs. To do this we needed to compute the average RSV_k value (denoted $Mean^i$) and the standard deviation (denoted $Stdev^i$) for each *i*th result list. These values could then be used to normalize the retrieval status for each document D_k found in the *i*th result list through computing the deviation for RSV_k with respect to the mean ($Mean^i$). Of course another method would be to weight the relative contribution of each retrieval scheme by assigning a different α_i value to each retrieval model.

6. EVALUATION

To evaluate our various search strategies, we used the tool provided by the organizers, based on the TREC_EVAL method to measure retrieval effectiveness. Based on the retrieval of 1,000 passages per query, this program computed different performance measures (e.g., the MAP). For the Blog collection, we limited our investigation to the opinion-finding task, namely the retrieval of information on the target entities without classifying them as positive, negative or mixed. For the Genomics task, the MAP was used in three different types of granularity at the document, passage and passage2 levels, and also at the feature level.

6.1 Genomics Official Runs

Table 1 provides a description of our three official runs within the Genomics task. These runs were based on the two probabilistic models (Okapi & I(n)B2) and include some of the search features described previously. First we listed the I(n)B2 model with the WordNet expansions (see Figure 6 for an example). In our second official run we applied WordNet thesaurus expansions and for our third we considered orthographic variants resulting from WordNet expansions.

Run name	IR model	Passage defined by
UniNE1	I(n)B2 + WordNet Exp.	<P </P
UniNE2	Okapi + WordNet Okapi + reranking I(n)B2 + WordNet	sentence
UniNE3	I(n)B2 + WordNet + Spell. Okapi + WordNet I(n)B2 + WordNet	<P </P

Table 1. Description of official runs (Genomics track)

Run name	MAP document	MAP passage2	MAP aspect	Passage defined by
Okapi	0.1486	0.0190	0.0633	<P </P
Okapi	0.1289	0.0089	0.0740	sentence
I(n)B2	0.2533	0.0907	0.2036	<P </P
I(n)B2	0.1508	0.0193	0.0952	sentence
Okapi+WN	0.1690	0.0287	0.0388	<P </P
Okapi+WN	0.1566	0.0166	0.0896	sentence
I(n)B2+WN	0.2777	0.0998	0.2177	<P </P
I(n)B2+WN	0.1978	0.0347	0.1227	sentence
Okapi+Spell	0.1462	0.01883	0.0602	<P </P
Okapi+Spell	0.1219	0.0084	0.0683	sentence
I(n)B2+Spell	0.2510	0.0902	0.2019	<P </P
I(n)B2+Spell	0.1538	0.0179	0.0850	sentence
Okapi+WN+Sp	0.1671	0.02819	0.0707	<P </P
Okapi+WN+Sp	0.1509	0.0159	0.0875	sentence
I(n)B2+WN+S	0.2765	0.0983	0.2177	<P </P
I(n)B2+WN+S	0.1961	0.0328	0.1188	sentence
UniNE1	**0.2777**	**0.0988**	**0.2189**	<P </P
UniNE2	0.1903	0.0278	0.1102	sentence
UniNE3	0.2710	0.0978	0.2043	<P </P

Table 2. Official Genomic track results and their components

Table 2 lists the evaluation results for our three official runs, together with their various components. Listed first in this table are the single IR models (Okapi & I(n)B2), and then these same models with the WordNet (WN) query expansion option (lines 5 to 8). In lines 9 and 12 we used the Okapi and I(n)B2 models along with spelling variations of the search terms, and finally we evaluated the Okapi and I(n)B2 approaches with both WordNet and orthographic variant expansions (lines 13 and 16). Our three official runs thus combined IR models based on the Z-score approach (see Section 5.2).

The results listed in Table 2 show that through using the WordNet thesaurus, we could enlarge the query (both with synonyms and morphological related terms) and improve the MAP results (from 9.6% to 31.2% in relative values). For example, with the I(n)B2 model, the MAP increases from 0.2533 to 0.2777 (+9.6%). Including orthographic variants tend to hurt slightly the MAP values (from -5.4% to 2%). When compared to the use of passage segmentation (denoted **<P** **</P** in Table 2), the use of sentences as passages was clearly not a good idea. Applying the document-based MAP, our best run (UniNE1) produced performances that were 30 times better than the median of all submitted runs.

6.2 Opinion-Finding Official Runs

To search information in the blogsphere, we based our official runs on three IR systems, namely the probabilistic Okapi model, the language model (LM) and models derived from the *Divergence from Randomness* (DFR) paradigm. See Table 3 for an evaluation of these different IR approaches and three query formulations (T, TD and TDN). In this case we considered all factorial web pages to be relevant (relevance value, rv=1) and all documents comprising various opinions (negative rv=2, mixed rv=3 or positive rv=4) concerning the specified target entity.

IR Model	T	TD	TDN
Okapi	**0.3585**	0.4003	0.3965
DFR-PL2	0.3568	**0.4033**	**0.3942**
DFR-IneC2	0.3398	0.3849	0.3771
DFR-I(n)B2	0.3397	0.3770	0.3606
DFR-PB2	0.3365	0.3767	0.3617
LM	0.3331	0.3808	0.3812

Table 3. Fact and opinion evaluations of the single IR models (Blog, three query formulations)

This table illustrates how the Okapi or the DFR-PL2 approaches produced the best results, albeit with rather small differences. Through adding the descriptive part in the query formulation we might improve the MAP by 12.5% in mean. Also worth noting is that increasing the query from TD to TDN does not necessarily improve the MAP values (mean decrease of -2.2%). Table 4 lists our six official runs for the Blog track Table 5 lists our official results.

Our official results for the Blog track tend to indicate that simple IR models perform better than more complex search strategies. With the TD query formulation for example, combining two IR models for the UniNEblog3 run produced an MAP of 0.4034, while under the same conditions the DFR-PB2 by itself model achieved an MAP of 0.4033 (see Table 3).

Run name	IR model
UniNEblog1	Okapi
UniNEblog2	DFR-PL2
UniNEblog3	DFR-PB2 + Okapi & Rocchio 5/50
UniNEblog4	LM (λ=0.35) + DFR-PL2
UniNEblog5	DFR In2C2 + Okapi (5-gram) + LM (λ=0.35, three letters)
UniNEblog6	LM (λ=0.35)

Table 4. Description of official Blog track results

Run name	QUERY	RELEVANT	POLARITY
UniNEblog1	T	0.3585	0.2770
UniNEblog2	TDN	0.3942	0.2898
UniNEblog3	TD	**0.4034**	**0.3049**
UniNEblog4	T	0.3467	0.2659
UniNEblog5	TD	0.3892	0.2972
UniNEblog6	TD	0.3808	0.3016

Table 5. Official results of the Blog track results

6.3 Difficult Topics in the Blog Track

Table 6 lists the top five most difficult topics of our best performing runs and also provides a better picture of the problems encountered when our systems searched the Blog track (UniNEblog3).

Topic ID	AP	Main explanation
#916	0.0005	Too many spam
#937	0.0049	Discrimination fails
#928	0.0177	Stopword list too large
#921	0.0373	Discrimination fails
#929	0.0571	Discrimination fails

Table 6. The most difficult topics in our best runs (UniNEblog3)

Because this search model does not account for noun phrases, there was a decrease in retrieval effectiveness due to our inability to impose the presence of two (or more) search terms. With title-only queries such as Topic #929 ("Brand manager"), Topic #921 ("Christianity Today") or Topic #928 ("Big Love") for example, the presence of both terms in the web page should be imposed and thus ensure their retrieval. Our IR models tend to extract many documents because one of the search terms has a high term frequency.

A second problem is our extended stopword list. In order to ignore HTML-tags (which may have passed the parsing step) and also to remove very frequent blog words, we added a few terms to our stopword list (e.g., big, com). In

Topic #928 (“Big Love”) or Topic #916 (“dice.com”) however this reduced the underling query to the single term “love” or “dice”, meaning that such a query would not effectively retrieve and rank highly relevant web pages.

For Topic #916 (“dice.com”), our IR systems encountered a problem related to spam. Given that “dice.com” was reduced to “dice”, most retrieved documents at the top of the result list assigned very high term frequency to the term “dice”. Most of the spam blogs retrieved thus had the same content, being a list of popular internet searches containing terms such as “dice game”, “dodecahedron dice” or “Dice Games and Rules”, all of which originate from the same server (newgreatblogs.com).

For Topic #937 (“LexisNexis”) most of the highly ranked yet non-relevant web pages were retrieved from the same blog (lawprofessors.typepad.com/law_librarian_blog), which contains numerous links to the LexisNexis web site. The outcome was an increase in the *tf* component for those pages, providing them with higher ranks. Unfortunately we cannot simply ignore these pages because they originate from a blog that also contains some relevant documents.

7. CONCLUSION

During this TREC 2007 Genomic evaluation campaign we evaluated various indexing and search strategies. The empirical evidence collected shows that the DFR-I(n)B2 model tends to perform better than the Okapi probabilistic model (0.2533 vs. 0.1486, document-based MAP). The inclusion of orthographic variants for search words (or two-word query sequences) does not really improve retrieval effectiveness, at least as implemented in our system (e.g., with the I(n)B2 model, from 0.2533 to 0.2510). Enlarging query formulations by adding synonyms or morphological related words extracted from the WordNet thesaurus results in better MAP (e.g., from 0.2533 to 2777 using the I(n)B2 model). Our passage segmentation approach was clearly more efficient than an approach based on sentences.

In the Blog track (limited in our case to retrieving opinions on a target entity), we find that the Okapi or the DFR-PL2 search models tend to produce the best MAP for certain query formulations. For example with the T query formulation we obtained a MAP of 0.3585 for the Okapi model compared to 0.3331 for the language model (-7.1%). By including the topic's descriptive part, this formulation increases the MAP by around 12% in mean (e.g., Okapi 0.3585 vs. 0.4003). Including the narrative part however tends to hurt the MAP (mean decrease around -2%). Moreover, simple IR models tend to produce retrieval performance similar to that of more complex IR strategies, such as those combining two ranked lists. When using TD queries for example the DFR-PL2 produces a MAP of 0.4033 while with a combined run (DFR-PB2 and Okapi plus pseudo-relevance feedback) a similar MAP (0.4034) resulted. In an effort to improve the MAP, we analyzed various difficult topics and their result lists. From an analysis of these resultant ranked lists we concluded that accounting for noun phrases (e.g., “Brand manager”, “Big Love”) or at least accounting for the presence of the two (or more) search terms in the retrieved web page may improve the MAP.

ACKNOWLEDGMENTS

This research was supported in part by the Swiss NSF under Grant #200021-113273.

8. REFERENCES

[1] Hersh, W.R., Cohen, A.M., Yang, J., Bhuptiraju, R.T., Roberts, P., & Hearst, M. TREC 2005 genomics track overview. In *Proceedings of TREC-2005.* Gaithersburg (MA), 2006.

[2] Hersh, W.R., Cohen, A.M., Roberts, P., & Rekapalli, H.K. TREC 2006 genomics track overview. In *Proceedings of TREC-2006.* Gaithersburg (MA), NIST Publication #500-272, 2007.

[3] Ounis, I., de Rijke, M., Macdonald, C., Gilad Mishne, G., & Soboroff, I. Overview of the TREC-2006 blog track. In *Proceedings of TREC-2006.* Gaithersburg (MA), NIST Publication #500-272, 2007.

[4] Kaszkiel, M., & Zobel, J. Effective ranking with arbitrary passages. *Journal of the American Society for Information Science and Technology*, 52(4), 2001, 344-364.

[5] Harman, D. How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 1991, 7-15.

[6] Abdou, S., Ruch, P., & Savoy, J. Evaluation of stemming, query expansion and manual indexing approaches for the genomic task. In *Proceedings TREC-2005*, Gaithersburg (MA), 2006, 863-871.

[7] Abdou, S., & Savoy, J. Report on the TREC 2006 genomics experiment. In *Proceedings TREC-2006*, Gaithersburg (MA), NIST Publication #500-272, 2007.

[8] Yu, H., & Agichtein, E. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(1), 2003, i340-i349.

[9] Huang, X., Zhong, M., & Si, L. York University at TREC 2005: Genomics track. In *Proceedings of TREC-2005*. Gaithersburg (MA), 2006.

[10] Cohen, A.M. Unsupervised gene/protein named entity normalization using automatically extracted

dictionaries. In *Proceeding ACL-ISMB*, Detroit (MI), 2005, 17-24.

[11] Gospodnetic, O., & Hatcher, E. *Lucene in Action*. Manning Publications, 2004

[12] Robertson, S.E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 2000, 95-108.

[13] Amati, G., & van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 2002, 357-389.

[14] Hiemstra, D. Using language models for information retrieval. CTIT Ph.D. Thesis, 2000.

[15] Hiemstra, D. Term-specific smoothing for the language modeling approach to information retrieval. In *Proceedings of the ACM-SIGIR*, 2002, 35-41.

[16] Vogt, C.C. & Cottrell, G.W. Fusion via a linear combination of scores. *IR Journal*, 1(3), 1999, 151-173.

[17] Fox, E.A. & Shaw, J.A. Combination of multiple searches. In *Proceedings TREC-2*, Gaithersburg (MA), NIST Publication #500-215, 1994, 243-249.

Stratégies de recherche dans la *blogosphère*

Claire Fautsch — Jacques Savoy

Institut d'informatique
Université de Neuchâtel
rue Emile Argand 11
2009 Neuchâtel, Suisse
{Claire.Fautsch, Jacques.Savoy}@unine.ch

RÉSUMÉ. *Cette communication présente les principaux problèmes liés à la recherche d'information dans la* blogosphère. *Recourant au modèle vectoriel* tf idf, *ainsi qu'à trois approches probabilistes et un modèle de langue, cet article évalue leur performance sur un corpus TREC extrait de la* blogosphère *et comprenant 100 requêtes. Les raisons expliquant les faibles performances sont exposées. Basés sur deux mesures de performance, nous démontrons que l'absence d'enracineur s'avère plus efficace que d'autres approches (enracineur léger ou celui de Porter). Imposer la présence côte à côte de deux mots recherchés dans la réponse fournie permet d'accroître significativement la performance obtenue.*

ABSTRACT. *This paper describes the main retrieval problems when facing blogs. Using the classical tf idf vector-space model together with three probabilistic and one statistical language model, we evaluate them using a TREC test-collections composed of 100 topics. We analyze the hard topics. Using two performance measures, we show that ignoring a stemming approach results in a better performance than other indexing strategies (light or Porter's stemmer). Taking account of the presence of two search words in the retrieved documents may significantly improve the retrieval performance.*

MOTS-CLÉS : *blogosphère, domaine spécifique, évaluation, modèle probabiliste, TREC.*

KEYWORDS: *Blogs, Domain-specific IR, Evaluation, Probabilistic model, TREC.*

DOI:10.3166/DN.11.1-2.109-132

110 DN – 11/2008. Documents et web 2.0

1. Introduction

Internet, outil de communication par excellence, continue à progresser en offrant ses services à un nombre croissant de personnes. Celles-ci ne se contentent plus d'être de simples consommateurs d'information gravitant autour des moteurs de recherches (Boughanem et Savoy, 2008), consultant des horaires ou la météo, réservant des chambres d'hôtel, ou achetant livres et CDs de musique. Les internautes occupent également un rôle de producteur d'information. Rédiger son journal intime et le diffuser, donner son opinion personnelle, écrire son carnet de bord d'artiste ou partager ses émotions face aux événements, tous ces exemples se retrouvent dans les *blogs*[1].

Enumérer la liste possible des thèmes abordés par les *blogs* s'avère impossible car ils tentent de couvrir toutes les activités et préoccupations humaines. Cependant, si chaque *blog* possède, originellement pour le moins, un caractère autobiographique prononcé, on peut attribuer également à ces journaux électroniques les qualificatifs de « subjectif » et « d'opinion »[2]. Parfois centré exclusivement sur une personne, le *blog* possède très souvent un aspect d'interaction. En effet, chaque billet publié peut faire l'objet de commentaires de lecteurs, parfois de manière continue. Si le *blog* s'ouvre d'emblée avec une vocation communautaire (e.g., celui des grévistes ou sur un projet de construction controversé), les personnes seront plus portées à discuter ou commenter les événements ou les billets postés précédemment dans cette tribune.

La *blogosphère* n'a pas laissé indifférent les acteurs du marketing et les entreprises ont compris qu'elles pouvaient en tirer parti de multiples manières (influencer, apprendre, communiquer, se montrer, conseiller, vendre) (Malaison, 2007). Les acteurs politiques ont également suivi cette tendance avec des différences notables entre les Etats-Unis et la France, voire entre politiciens français (Jereczek-Lipinska, 2007 ; Véronis *et al.*, 2007). Ainsi on peut se limiter à reporter les discours officiels à l'image des quotidiens. Dans ce cas, on perd la dimension personnelle, la vision privée du politicien (avec simplicité et franc-parler) ainsi que toute interactivité avec les citoyens qui font du *blog* un nouveau média tant dans la forme que dans le contenu. Est-ce que ce nouveau moyen de communication favorisera réellement une démocratie participative, souhait exprimé par S. Royal lors de la campagne présidentielle de 2007 ?

---

1. Ce terme provient de la contraction de « web » et « log ». L'équivalent français apparu dans le *Journal officiel* est « bloc-notes » ou « bloc » mais nous avons une préférence pour la forme anglo-saxonne indiquant son aspect électronique. D'autres termes anglais comme « feed » (flux d'information) ou « permalink » (permalien) utilisés dans cet article ne disposent pas encore d'un équivalent officiel. Nous avons choisi souvent de conserver la forme originale tout en proposant un terme français qui nous semble adapté.

2. Voir le site www.LoicLemeur.com/france/ maintenu par Loïc Le Meur, un des pionniers du domaine en France.

Ce nouveau média s'accompagne de ses propres faiblesses et abus. Par exemple, lors d'une campagne électorale, un site de *blogs* peut être pris d'assaut par les partisans d'un des candidats afin d'imposer leur point de vue. De manière similaire, MySpace.com peut être submergé de vidéos soulignant les mérites d'un parti ou d'un candidat. Afin de mieux cerner les principales tendances on peut également essayer de dresser une cartographie numérique de la *blogosphère* comme le propose le site www.USandUS.eu pendant les élections américaines de 2008.

Cette progression de l'ensemble de ces écrits formant la *blogosphère* a été grandement facilitée par la simplicité d'accès et d'édition proposée par des logiciels spécifiques[3]. Ces derniers s'appuient sur des compléments au standard XML (e.g., RSS, ATOM) afin de faciliter la gestion de l'aspect dynamique des flux d'information comme, par exemple, pour insérer un billet dans la bonne rubrique et selon l'ordre chronologique ou pour avertir les abonnés dès qu'une nouvelle est insérée. Structurée selon des auteurs et des thématiques, la consultation s'effectue habituellement selon l'ordre chronologique inverse (le dernier billet rédigé se retrouvant souvent dans la page d'accueil).

Si le contenu textuel domine, celui-ci peut s'accompagner de dessins, d'images, de photos (*photoblog*), voire d'éléments audio (*podcasting* ou baladodiffusion) ou vidéo (*videoblog*). Evidement, parfois l'élément audio ou visuel prend clairement le dessus et la distinction entre *blog* et service de diffusion s'estompe (voir, par exemple, les sites Flickr.com, FaceBook.com ou YouTube.com). Dans d'autres cas, l'attention se porte sur les liens entre pages personnelles à l'image des réseaux sociaux (Facebook.com ou Linkedin.com). Parfois, la distinction entre un *blog* collectif se nourrissant de son propre bavardage et un forum de discussion peut s'atténuer.

Faire partager ses émotions, donner son avis ou convaincre l'autre ne signifie pas liberté de dire n'importe quoi. Le contenu d'un *blog* reste sous la responsabilité de son éditeur. De plus, si de nombreux *blogs* voient le jour, de nombreux autres tombent dans l'oubli ou sont délaissés rapidement par leur créateur. Si leur volume croît suivant une courbe exponentielle, comment retrouver l'information pertinente dans la *blogosphère* (Witten, 2007) ? Le moteur *Google* s'y intéresse et il a ajouté à son inventaire de services un moteur de recherche[4] dédié à ce contenu particulier. Pour une présentation des défis et solutions particulières de la recherche sur le web, on peut se référer à (Boughanem et Savoy, 2008).

Le contenu et le style de la *blogosphère* possèdent des caractéristiques distinctes des corpus d'articles scientifiques ou de presse utilisés habituellement en recherche d'information. Par nature subjectif, le *blog* possède comme premier objectif de

3. Ou systèmes de gestion de contenu (SGC ou CMS). Les sites Blogger.com, MySpace.com ou Skyrock.com proposent également aux internautes de concevoir et de rédiger leur propre *blog*. Ils se chargeront ensuite de les héberger.

4. Disponible à l'adresse http://blogsearch.google.fr

diffuser des opinions ou de faire partager des émotions. Les fautes d'orthographe et d'accord, avec une syntaxe hésitante, vont connaître une plus grande fréquence. Le lexique lui-même va laisser transparaître une classe sociale donnée et le recours à l'argot ou au langage SMS[5] (Fairon *et al.*, 2006) n'est pas une exception. La connaissance précise de la langue dans laquelle est rédigée un document ne sera plus acquise de manière certaine (Singh, 2006). A ceci s'ajoute la prise en compte de plusieurs codages possibles pour une écriture voire pour des lettres accentuées. Le style distinct entre les deux types de corpus peut également soulever de nouveaux problèmes.

Le document traditionnel (livre, périodique, thèse, carte, partition) se caractérise par son support auquel s'associe une trace d'inscription. Sa conception tend habituellement à favoriser une lecture linéaire. La subdivision logique apporte un élément structurant sur le contenu véhiculé et favorise des accès intradocument. Le document numérique désire proposer une nouvelle gestion des documents, souvent par le biais d'un accès moins linéaire et en favorisant l'intégration d'autres médias (image, son, vidéo) à l'écriture. Dans la *blogosphère*, le billet d'information peut certes posséder sa propre structure mais l'attention se porte également sur la réaction des lecteurs qui ont la possibilité d'y inclure leurs commentaires voire des remarques sur ces derniers. Encourager une écriture collective peut également favoriser l'effacement des noms des auteurs, à l'image de l'encyclopédie Wikipédia.

Les requêtes reflètent clairement les intérêts de la communauté des internautes avec une prépondérante d'interrogations comportant uniquement le nom d'une personne, d'un lieu ou d'un produit. La réponse attendue doit comporter souvent un point de vue personnel sur une question (« la guerre en Irak ») ou correspondre aux expériences personnelles concernant un produit (« iPhone »). Retourner de simples faits ne constitue pas toujours une réponse idéale. De plus, le temps joue un rôle crucial et toute information obsolète doit être ignorée. Le dépistage de la bonne réponse peut également s'appuyer sur les étiquettes descriptives (les meta-informations) spécifiant le thème d'un flux d'information ou en admettant qu'un même auteur rédige des blocs ayant des sujets reliés. Finalement, le monde des *blogs* contient également son lot de contenu commercial non désiré, le *spam*.

Afin d'analyser empiriquement une partie de ces questions, la piste « *blog* » a été créée lors de la campagne d'évaluation TREC en 2006 (Ounis *et al.*, 2006) et poursuivie en 2007 (Macdonald *et al.*, 2007). Dans cette communication, nous désirons présenter le corpus utilisé (section 2). Afin de travailler avec les meilleures stratégies de dépistage, nous avons décidé d'implémenter le modèle Okapi, deux approches tirées de la famille *Divergence from Randomness* (DFR) et un modèle de langue (voir section 3). La section 4 présente notre méthodologie d'évaluation et l'appliquera à nos divers modèles de recherche en fonction de différentes stratégies

5. De nombreuses possibilités sont offertes comme la suppression des voyelles (« bjr » pour « bonjour »), les abréviations, le rébus (« K7 » pour « cassette »), des sigles (« mdr » pour « mort de rire ») ou l'écriture phonétique (« jtm » pour « je t'aime »).

d'indexation ou de longueur de requêtes. Notre proposition décrite dans la cinquième section s'appuie sur la prise en compte de plusieurs mots de la requête dans la réponse retournée à l'internaute.

2. Regard sur le corpus d'évaluation et la *blogosphère*

Créée par l'Université de Glasgow, la collection de *blogs* dénommée Blogs06 a été extraite du web entre décembre 2005 et février 2006. Elle comprend un volume d'environ 148 Go pour 4 293 732 documents. Trois sources composent ce corpus soit 753 681 *feeds* ou flux d'information représentant environ 17,6 % du total, 3 215 171 *permalinks* (*permalien* ou lien permanent) (74,9 %) et 324 880 pages d'accueil (pour environ 7,6 %). Dans cet ensemble d'articles, ce corpus contient également des *spams* (ou *pourriel*), des documents à contenu essentiellement publicitaire cherchant à tromper les moteurs de recherche afin d'être dépisté en réponse à des mots-clés fréquents.

Les flux d'information correspondent bien à un outil de la *blogosphère*. Si l'on analyse le volume de l'information mémorisée au lieu du nombre d'entrées, la partie *feed* représente 38,6 Go (pour environ 26,1 %). Par exemple, la page d'accueil du site de Sarah Carey (en Irlande) est disponible à l'adresse http://www.sarahcarey.ie/. Depuis cette adresse, de nombreux flux de discussion peuvent s'ouvrir comme, par exemple sur la presse en général (http://www.sarahcarey.ie/wordpress/feed/#). Depuis un flux d'information, plusieurs *permaliens* peuvent être obtenus et suivis.

La partie des *permaliens* comprend 88,8 Go pour environ 60 % du volume total. Un *permalien* correspond à une URL utilisée pour référer l'entrée d'un élément d'information (billet) à caractère dynamique relié à une discussion précise. Comme cet élément voit son volume croître avec le temps, donner comme référence la page elle-même conduirait les internautes au début de la discussion et non directement à l'élément visé par la citation. Le format associé aux *permalien*s n'est pas standardisé mais ce dernier se compose de l'adresse URL, suivi souvent d'une date (e.g., quatre chiffres pour l'année, deux chiffres pour le mois et deux chiffres pour le jour). Il se termine par le nom ou numéro de l'article (voire de l'ajout). Parfois le nom de l'utilisateur est inclus dans le *permalien* (e.g., en tête de l'URL si la personne concerné à ouvert son journal électronique sur un système dédié comme http://*tintin*.blogspot.com). Si nous reprenons notre exemple précédent, nous avons les *permaliens* "http://www.sarahcarey.ie/wordpress/archives/2005/11/29/women-and-work-2#" ou "http://www.sarahcarey.ie/wordpress/archives/2005/12/03/radio-appearance#".

Finalement la partie des pages d'accueil se compose de 20,8 Go soit environ 14,1 % du total. On y retrouve ceux de personnes désirant offrir une simple carte de visite électronique ou une première page offrant l'accès à divers flux d'information. Des détails plus complets sur cette collection sont disponibles à l'adresse http://ir.dcs.gla.ac.uk/test_collections/. Dans la suite de nos expériences, de même

que lors des deux campagnes TREC (Ounis *et al.*, 2006 ; Macdonald *et al.*, 2007), seule la partie des *permaliens* a été retenue pour l'évaluation.

```
<DOC>
<DOCNO> BLOG06-20051206-000-0024615414
<DATE_XML> 2005-12-06T07:06:00+0000
<FEEDNO> BLOG06-feed-000088
<FEEDURL> http://www.houseoffusion.com/cf_lists/RSS.cfm/forumid=4#
<PERMALINK>http://www.houseoffusion.com/cf_lists/message.cfm/forumid:4/
messageid:226252#
…
<DOCHDR> …
http://www.houseoffusion.com/cf_lists/message.cfm/forumid:4/messageid:
226252# 0.0.0.0 2005122020386 15533
Date: Tue, 20 Dec 2005 21:39:33 GMT
Server: Microsoft-IIS/5.0
Content-Language: en-US
Content-Type: text/html; charset=UTF-8
… </DOCHDR>
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd"> …
<HTML>
<HEAD>
  <LINK rel="shortcut icon" href="/_/favicon.ico" >
  <LINK rel="stylesheet" type="text/css" href="/_/hof.css">
  <TITLE> CFEclipse</TITLE>
  <META name="Keywords" content="CFEclipse, CF-Talk">
  <META name="Description" content="CFEclipse was asked on the Talk">…
  <SCRIPT src="http://www.google-analytics.com/urchin.js"
type="text/javascript"></SCRIPT>  …
</HEAD>
<BODY bgcolor="#ccccff">
<TABLE width="1008" border="0" cellspacing="0" cellpadding="0" >
<TR> <TD rowspan="1" width="130" height="20" valign="top"> <IMG
src="/_/hof125.gif" alt="House of Fusion Logo" border="0"
style="position: absolute; top: 1px; left: 2px; z-index: 1;"></TD>
   <TD height="20" width="400"> </TD>
…
<UL class="sideList">
<LI><A class="navbar-left" href="/cf_lists/threads.cfm/4"
title="ColdFusion technical mailing list">CF-Talk</A></LI>
<LI><A class="navbar-left" href="/cf_lists/threads.cfm/13"
title="ColdFusion Jobs mailing list">CF-Jobs</A></LI>
…
</UL> <UL class="sideList">
<LI><A href="/cf_lists/threads.cfm/51" class="navbar-
left">CFUnit</A></LI>
<LI><A href="/cf_lists/threads.cfm/47" class="navbar-
left">AJaX</A></LI>
…
<TD colspan="2" valign="top" class="textmain" width="748"
bgcolor="#ffffff">
<div> As the year draws to an end and your company finds itself with a
budget excess that it has to get rid of, please remember to support
the resource that supports you. <BR>
Thank you </div>
…
```

Figure 1. *Exemple d'un document du corpus Blogs06*

La figure 1 présente un exemple de document extrait du corpus Blogs06. Dans cette figure, on constate que chaque document débute par la balise <DOC>. Nous retrouvons ensuite la balise l'identificateur unique de l'article (<DOCNO>) puis la date de sa récupération depuis internet (balise <DATE_XML>). Suivent diverses balises spécifiant le *feed*, le *permalien* ainsi que l'en-tête de réponse du serveur lors de la récupération de ce document (après la balise <DOCHDR>). Dans cette partie, on retrouve la date, l'URL source, le type de logiciel utilisé par le serveur, le type de codage, etc.

Les données pertinentes pour la recherche d'information suivent la balise fermante </DOCHDR>. On y retrouve l'en-tête de la page (encadré par les balises <HEAD>) avec la présence assez récurrente des balises métas "Keywords" et "Description" ainsi que <TITLE>. Contrairement à diverses autres collections-tests, on y retrouve beaucoup d'éléments pas ou peu pertinents pour la recherche d'information comme des programmes Javascript, des redirections, la référence à des feuilles de style, etc. Ces divers éléments peuvent être exploités par d'autres applications ou par des systèmes d'information ayant un objectif différent du nôtre consistant à proposer un accès efficient par le contenu.

Le contenu de la page visible sur l'écran de l'internaute est encadré par les balises <BODY>. Selon les documents, la densité de commandes HTML est variable mais elle s'avère supérieure à nos attentes. De larges passages mémorisent de simples menus, listes ou font appel à des scripts. Les documents n'ont pas été nettoyés et il n'est pas rare d'obtenir des documents rédigés dans d'autres langues comme l'espagnol ou le japonais.

La figure 2 illustre un second document que notre système de dépistage a récupéré en réponse à la demande « LexisNexis ». Cet article correspond à une annonce faite par l'entreprise pour un nouveau produit. Le contenu se composera d'une information dite « objective » dans le sens que l'on a une description d'un produit sans véritable jugement personnel. Dans la figure 2, le document a été analysé et les balises HTML inutiles pour le dépistage de l'information ont été éliminées. La balise <DATA> a été ajoutée pour indiquer le début du texte retenu pour l'indexation.

Avec ces documents, nous disposons de 100 requêtes numérotées de 851 à 900 pour l'année 2006 et de 901 à 950 pour l'année 2007. Dans la présente étude, nous avons fusionné ces deux sous-ensembles pour former un lot relativement important de requêtes. En effet, aucune modification majeure n'a été apportée en 2007 par rapport à 2006. De plus, le doublement du nombre de requêtes (ou d'observations) permet ainsi une analyse plus fine des résultats. Limiter nos analyses à 50 cas n'a pas de sens alors que nous pouvons disposer d'un volume deux fois plus conséquent. De plus, sur la base de 50 observations il s'avère plus difficile de détecter des différences statistiquement significatives.

116 DN – 11/2008. Documents et web 2.0

```
<DOC>
<DOCNO> BLOG06-20051212-051-0007599288
<DATE_XML> 2005-10-06T14:33:40+0000
<FEEDNO> BLOG06-feed-063542
<FEEDURL> http://contentcentricblog.typepad.com/ecourts/index.rdf
<PERMALINK>
http://contentcentricblog.typepad.com/ecourts/2005/10/efiling_launche.
html#
<DOCHDR> …
Date: Fri, 30 Dec 2005 06:23:55 GMT
Accept-Ranges: bytes
Server: Apache
Vary: Accept-Encoding,User-Agent
Content-Type: text/html; charset=utf-8
…  </DOCHDR>
<DATA>
electronic Filing &amp;amp; Service for Courts
…
October 06, 2005
eFiling Launches in Canada
Toronto, Ontario, Oct.03 /CCNMatthews/ - LexisNexis Canada Inc., a
leading provider of comprehensive and authoritative legal, news, and
business information and tailored applications to legal and corporate
researchers, today announced the launch of an electronic filing pilot
project with the Courts
```

Figure 2. *Exemple d'un document concernant le service LexisNexis après notre nettoyage*

Suivant le modèle habituel des diverses campagnes d'évaluation, chaque requête possède principalement trois champs logiques, à savoir un titre bref (<TITLE> ou T), une phrase décrivant le besoin d'information (<DESC> ou D) et une partie narrative (<NARR> ou N) spécifiant plus précisément le contexte de la demande ainsi que des critères de pertinence permettant de mieux évaluer les opinions dépistées. La figure 3 présente trois exemples. Dans nos évaluations, nous avons retenu souvent uniquement la partie "titre" (T) pour construire les requêtes. Avec cette contrainte, le nombre moyen de termes d'indexation par requête s'élève à 1,72 (min :1, max : 5, médiane : 2) tandis que le recours aux deux champs "titre" et "descriptif" (TD) produisent une longueur moyenne de 6,53 mots (min : 2, max : 12, médiane : 6). Finalement pour les requêtes longues (TDN), le nombre moyen de mots pleins par requête s'élève à 17,4 (min : 8, max : 34, médiane : 16,5).

Les thèmes des demandes couvrent des domaines variés comme la recherche d'opinions, commentaires ou recommandations touchant la culture (n° 851 "March of the Penguins", n° 875 "american idol", n° 913 "sag awards", ou n° 928 "big love"), les produits et services (n° 862 "blackberry", n° 883 "heineken", n° 900 "mcdonalds", n° 909 "Barilla", ou n° 937 "LexisNexis"), les personnalités (n° 880 "natalie portman", n° 935 "mozart" ou n° 941 "teri hatcher"), la politique (n° 855 "abramoff bush", n° 878 "jihad", n° 887 "World Trade Organization" ou n° 943 "censure"), la science et la technologie (n° 896 "global warming", n° 902 "lactose

gas", ou n° 923 "challenger"), les faits divers (n° 869 "muhamad cartoon"), voire des thématiques plus variées (n° 861 "mardi gras" ou n° 889 "scientology").

```
<NUM>  853
<TITLE>  state of the union
<DESC>  Find opinions on President Bush's 2006 State of the Union
address
<NARR>  All statements of opinion on the address are relevant.
Descriptions of the address, quotes from the address without comment,
and comedians' jokes about the address are not relevant unless there
is a clear statement of opinion.  Announcements that the address will
take place or has taken place are not relevant.  Schedules of events
or discussion groups to support or oppose the address are not
relevant.  Predictions of what will be in the address are not relevant

<NUM>  901
<TITLE>  jstor
<DESC>  Find opinions on JSTOR, the system developed to make scholarly
journals available from a digital archive
<NARR>  Reports of difficulty or ease in using JSTOR are relevant
opinions.  A statement that one is lucky to have access or wishes to
have access to JSTOR is a relevant opinion.  A statement that
information is available in JSTOR is not an opinion.  Simply citing
JSTOR as a reference for a document is not an opinion.

<NUM>  903
<TITLE>  "Steve jobs"
<DESC>  Find documents stating opinions about Apple CEO Steve Jobs.
<NARR>  Relevant documents will state opinions about Steve Jobs, the
head of Apple Computer.  Documents will include comments on his great
success with the iPod, his management style, and his unusual keynote
presentations he gives at the introduction of new products..
```

Figure 3. *Exemples de trois requêtes de notre corpus*

Elles incluent des questions présentant un caractère ambigu indéniable comme la demande n° 905 "king funeral" (concernant Coretta Scott King et non Elvis Presley ou un autre roi). Notons également que cette demande est reliée à la requête n° 874 "coretta scott king". Les thèmes relèvent essentiellement de la culture nord-américaine et correspondent, pour la partie "titre", à des demandes formulées en l'état par des internautes. Ce biais en faveur des Etats-Unis laisse toutefois apparaître des requêtes portant sur des thèmes internationaux comme la ville indienne de "varanasi" (n° 918) ou la "sorbonne" (n° 948). Les sujets abordés correspondent assez bien aux interrogations les plus populaires adressées aux moteurs de recherche commerciaux comme Google[6] ou Yahoo!

Si l'on analyse les jugements de pertinence correspondant à 32 078 documents pertinents, on remarque que le nombre moyen d'articles pertinents par requête

6. La liste des requêtes les plus fréquemment soumises à Google lors de l'année 2006 est disponible à l'adresse http://www.google.com/intl/en/press/zeitgeist2006.html tandis que celles adressées à Yahoo.com se trouve à http://f1.buzz.re2.yahoo.com/topsearches2006/

s'élève à 320,78 (médiane : 263,5, min : 16 (n° 939 "Beggin Strips"), max : 872 (n° 872 "brokeback mountain") avec un écart type de 225,84).

Les jugements de pertinence sont notés sur une échelle de 1 à 4. Une valeur unitaire indique que le document répond à la requête de manière objective ou d'une manière adéquate pour être repris dans une réponse que devrait rédiger l'internaute. Les valeurs supérieures indiquent que l'article répond à la requête mais qu'il possède également une opinion personnelle sur le sujet. Ainsi, la valeur quatre indique que l'article présente clairement un jugement positif concernant la requête tandis qu'une valeur de deux signifie une appréciation négative concernant le thème de la demande. La valeur de trois indique des jugements mélangés, tantôt positifs, tantôt négatifs voire ambigus ou peu clairement tranchés. Dans nos évaluations, nous avons admis comme bonne réponse tous les articles ayant une valeur de pertinence supérieure ou égale à un. Nous n'avons donc pas fait de distinctions entre un document objectif ou subjectif d'une part et, d'autre part, entre une opinion négative, mixte ou positive.

3. Les stratégies d'indexation et modèles de recherche d'information

Nous désirons obtenir une vision assez large de la performance de divers modèles de dépistage de l'information (Boughanem et Savoy, 2008). Dans ce but, nous avons indexé les billets d'information de la *blogosphère* (et les requêtes) en tenant compte de la fréquence d'occurrence (ou fréquence lexicale notée tf_{ij} pour le j^{e} terme dans le i^{e} document). Ainsi, si un terme possède une fréquence d'occurrence plutôt forte pour un document (*tf* élevé), il décrira bien le contenu sémantique de celui-ci et doit donc posséder une forte pondération.

En complément à cette première composante, une pondération efficiente tiendra compte de la fréquence documentaire d'un terme (notée df_j, ou plus précisément de l'$idf_j = \log(n/df_j)$ avec n indiquant le nombre de documents inclus dans le corpus). Ainsi, si un terme dispose d'une fréquence documentaire très élevée, il apparaît dans presque tous les documents (comme, par exemple, les mots "dans" ou "http"). Dans ce cas, sa présence dans la requête ne s'avère pas très utile pour discriminer les documents pertinents des articles sans intérêt. A l'inverse, si ce terme dispose d'une fréquence documentaire faible, il apparaît dans un nombre restreint de pages web (et sa valeur *idf* sera élevée). Dans ce cas, ce mot permet d'identifier un ensemble restreint de documents dans le corpus. Une pondération élevée permettra à ces quelques articles d'être classés au début de la liste des résultats retournés.

Afin de tenir compte de ces deux premières composantes, on multiplie les deux facteurs pour obtenir la formulation classique $tf \cdot idf$ donnant naissance à un premier modèle vectoriel. Comme troisième composante nous pouvons tenir compte de la longueur du document, en favorisant, *ceteris paribus*, les documents les plus courts comme le proposent plusieurs modèles probabilistes.

En effet, ces derniers ont été proposés afin d'améliorer la pondération $tf \cdot idf$. Dans le cadre de notre étude, nous avons considéré le modèle Okapi (Robertson e*t al*., 2000) utilisant la formulation suivante :

$$w_{ij} = [(k_1+1) \cdot tf_{ij}] / (K + tf_{ij}) \quad \text{avec } K = k_1 \cdot [(1-b) + ((b \cdot l_i) / mean\ dl)] \qquad [1]$$

dans laquelle l_i est la longueur du i^e article (mesurée en nombre de termes d'indexation), et b, k_1, *mean dl* des constantes fixées à $b = 0{,}4$, $k_1 = 1{,}4$ et *mean dl* = 787. Au niveau de la requête, les termes de celle-ci sont pondérés selon la formulation classique $tf \cdot idf$, soit :

$$w_{qj} = tf_{qj} \cdot idf_{qj} \qquad [2]$$

Remarquons que les requêtes étant des expressions brèves, la composante *tf* se limite très souvent à l'unité. Dès lors, la formule [2] correspond essentiellement à une pondération des termes de la requête selon leur valeur *idf*. Le score de chaque document D_i par rapport à la requête Q est calculé selon l'équation [3], soit :

$$Score\ [D_i, Q] = \sum_j w_{ij} \cdot w_{qj} \qquad [3]$$

Comme deuxième modèle probabiliste, nous avons implémenté le modèle PL2, un des membres de la famille *Divergence from Randomness* (DFR) (Amati et van Rijsbergen, 2002). Dans ce dernier cas, la pondération w_{ij} combine deux mesures d'information, à savoir :

$$\begin{aligned} &w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = Inf^1_{ij} \cdot (1 - Prob^2_{ij}) \quad \text{et} \\ &Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad \text{avec } tfn_{ij} = tf_{ij} \cdot \ln[1+((c \cdot mean\ dl) / l_i)] \\ &Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{avec } \lambda_j = tc_j / n \end{aligned} \qquad [4]$$

dans laquelle tc_j représente le nombre d'occurrences du j^e terme dans la collection, n le nombre d'articles dans le corpus et c une constante fixée à 5. La pondération des termes de la requête dans les divers modèles DFR se limite à la fréquence d'occurrence (soit $w_{qj} = tf_{qj}$) et le score de chaque document en fonction de la requête Q est calculé selon la formule [3].

Comme troisième modèle probabiliste, nous avons retenu le modèle I(n_e)C2 également issu de la famille DFR se basant sur la formulation suivante.

$$\begin{aligned} &Prob^2_{ij} = 1 - [(tc_j +1) / (df_j \cdot (tfn_{ij}+1))] \\ &Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0{,}5)] \quad \text{avec } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \end{aligned} \qquad [5]$$

Enfin, nous avons repris un modèle de langue (LM) (HIEMSTRA, 2000), dans lequel les probabilités sont estimées directement en se basant sur les fréquences d'occurrences dans le document D ou dans le corpus C. Dans cet article, nous avons repris le modèle de Hiemstra (2000) décrit dans l'équation [6] qui combine une estimation basée sur le document (soit $Prob[t_j \mid D_i]$) et sur le corpus ($Prob[t_j \mid C]$).

$$\text{Prob}[D_i \mid Q] = \text{Prob}[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot \text{Prob}[t_j \mid D_i] + (1-\lambda_j) \cdot \text{Prob}[t_j \mid C]] \quad [6]$$

$$\text{avec } \text{Prob}[t_j \mid D_i] = tf_{ij} / nt_i \quad \text{et } \text{Prob}[t_j \mid C] = df_j / lc \quad \text{avec } lc = \sum_k df_k \quad [7]$$

dans laquelle λ_j est un facteur de lissage (une constante pour tous les termes t_j, fixée à 0,35) et *lc* correspond à une estimation de la taille du corpus C.

4. Evaluation

La recherche d'information possède une longue tradition empirique visant à confirmer ou infirmer les modèles et techniques proposés. Dans cet esprit, nous décrirons notre méthodologie d'évaluation dans la section 4.1. La section 4.2 évalue l'emploi d'un enracineur (*stemmer*) plus ou moins agressif permettant d'augmenter la moyenne des précisions (MAP). Le recours à des requêtes plus longues permet habituellement d'augmenter la qualité du dépistage de l'information. Cette affirmation sera examinée dans la section 4.3. Dans la suivante, nous évaluons l'impact d'une procédure d'enrichissement automatique de la requête. Finalement, la section 4.5. propose de comparer les résultats de nos approches avec les meilleures performances obtenues lors des deux dernières campagnes d'évaluation TREC.

4.1. *Méthodologie d'évaluation*

Afin de connaître la performance d'un système de dépistage de l'information, nous pouvons tenir compte de divers facteurs comme la vitesse de traitement de la réponse, la qualité de l'interface, l'effort exigé par l'usager afin d'écrire sa requête ou la qualité de la liste des réponses fournies. En général, seul le dernier critère est pris en compte.

Pour calculer la qualité de la réponse associée à une requête, la communauté scientifique a adopté comme mesure principale la précision moyenne (PM) (Buckley et Voorhees, 2005). Son calcul s'opère selon le principe suivant. Pour chaque requête, on détermine la précision après chaque document pertinent. Cette dernière correspond au pourcentage de bonnes réponses (documents pertinents) dans l'ensemble des articles retournés à l'usager. Par exemple, dans le tableau 1, après trois documents retournés, la précision serait de 2/3 tandis que la précision après 35 documents serait de 3/35. Ensuite on calcule une moyenne arithmétique sur l'ensemble de ces valeurs. Si une interrogation ne dépiste aucun document pertinent, sa précision moyenne sera nulle. Dans le tableau 1, la précision moyenne de la requête A possédant trois documents pertinents s'élève à $(1/3)\cdot(1/2 + 2/3 + 3/35) = 0{,}4175$.

Rang	Requête A	Requête B
1	NP	P 1/1
2	P 1/2	P 2/2
3	P 2/3	NP
…	NP	NP
35	P 3/35	NP
…	NP	NP
108	NP	P 3/108
PM	0,4175	0,6759

Tableau 1. *Précision moyenne de deux requêtes ayant trois documents pertinents (notés P) et non pertinents (NP) présentés dans des rangs différents*

Pourtant la précision moyenne (PM) possède quelques inconvénients. En premier lieu cette valeur reste difficile à interpréter pour un usager. Que signifie une précision moyenne de 0,3 ? Ce n'est pas la précision après 5 ou 10 documents dépistés, valeur qui serait simple à interpréter pour l'utilisateur. Deuxièmement, comme l'illustre le tableau 1, des différences de précision moyenne importantes comme par exemple 0,6759 *vs*. 0,4175 (variation relative de 60 %) ne semblent pas correspondre à une différence aussi significative pour un usager. En effet, le classement proposé par la requête A ne s'éloigne pas beaucoup de la liste obtenue avec la requête B. En tout cas, l'usager n'attribuerait pas à cette variation une amplitude aussi élevée que 60 %.

Pour un ensemble de requêtes, nous pouvons opter pour la moyenne arithmétique (MAP) des précisions moyennes individuelles (PM). Cette mesure a été adoptée par diverses campagnes d'évaluation (Voorhees *et al.*, 2007) pour évaluer la qualité de la réponse à un ensemble d'interrogations. Afin de savoir si une différence entre deux modèles s'avère statistiquement significative, nous avons opté pour un test bilatéral non paramétrique (basé sur le rééchantillonnage aléatoire ou *bootstrap* (Savoy, 1997), avec un seuil de signification $\alpha = 5\ \%$). Comme nous l'avons démontré empiriquement, d'autres tests statistiques comme le *t*-test ou le test du signe aboutissent très souvent aux mêmes conclusions (Savoy, 2006).

Pour compléter la précision moyenne, nous pourrions également recourir à l'inverse du rang moyen de la première bonne réponse (MRR ou *mean reciprocal rank*), mesure reflétant mieux le comportement des internautes souhaitant uniquement une seule bonne réponse. A l'aide de cette mesure, la requête A du tableau 1 posséderait la valeur 1/2 = 0,5 tandis que la requête B obtiendrait une valeur de 1/1 = 1,0. Notons toutefois que ces diverses mesures de performance sont fortement corrélées (Buckley et Voorhees, 2005). Le choix de la MAP ou du MRR ne présente pas un éclairage biaisé dans l'analyse des résultats.

4.2. *Enracineurs*

Nous savions que le style et le lexique utilisés dans la *blogosphère* s'avéreraient différents des corpus d'agence de presse que nous avons l'habitude de traiter. Comme les interrogations sont souvent très courtes et se limitent à un ou deux termes précis (souvent un nom propre), nous pensons que le recours à un enracineur léger devrait fournir de meilleures performances qu'une approche plus agressive comme l'algorithme de Porter (1980) basé sur environ 60 règles. Dans ce but nous avons évalué la suppression de la consonne finale '-s' indiquant souvent la forme pluriel de la langue anglaise (Harman, 1991).

```
Si la finale est '-ies' mais pas '-eies' ou '-aies'
        alors remplacez  '-ies' par '-y', fin;
Si la finale est '-es' mais pas '-aes', '-ees' ou '-oes'
        alors remplacez  '-es' par '-e', fin;
Si la finale est '-s' mais pas '-us' ou '-ss'  alors éliminez  '-s';
fin.
```

Tableau 2. *Les trois règles de l'enracineur léger suggéré par Harman (1991)*

Prenons note toutefois que ces enracineurs fonctionnent sans connaissance de la langue et génèrent des erreurs. Ainsi, l'algorithme de Porter ne réduit pas sous la même racine l'adjectif "European" et le nom "Europe", tandis que l'approche proposée par Harman (voir tableau 2) retourne "speeche" pour le pluriel "speeches".

Comme autre possibilité, nous pouvons ignorer tout traitement morphologique (évaluation donnée sous la colonne "aucun" dans le tableau 3). Comme troisième choix, nous avons repris l'algorithme de Porter (1980) afin d'éliminer les suffixes flexionnels et certains suffixes dérivationnels. Comme autre approche, nous pourrions recourir à une analyse morphologique plus poussée capable de nous retourner le lemme ou l'entrée correspondante dans le dictionnaire. Dans ce dernier cas, en réponse au terme "eating", le système retournerait "eat". Toutefois, la couverture du dictionnaire sous-jacent n'est jamais complète et la présence de noms propres soulève la délicate question du traitement des mots pas reconnus par une telle analyse morphologique. Enfin la préférence pour l'emploi d'enracineurs s'explique par la nécessité d'un traitement peu coûteux en espace mémoire et en temps de calcul.

L'application d'un enracineur offre une première forme de normalisation des mots permettant un meilleur appariement entre les termes de la requête et ceux des documents. Ainsi, si la requête inclut la forme « jeux », il semble naturel de dépister des sites web décrit par les mots « jeux » ou « jeu ». Par contre, l'application d'un

enracineur même une approche simple peut provoquer des appariements erronés. Ainsi à la requête « Jeu de Nim », le moteur de recherche Google[7] nous a retourné dans les rangs deux et trois des sites proposant des informations sur les « jeux à Nîmes ». Les variations morphologiques entre les deux formes « Nim » et « Nîmes » peuvent être assimilées, de manière incorrecte dans cet exemple, à des variations de genre et de nombre. La présence d'accent et leur élimination automatique reste un problème que l'on rencontre dans plusieurs langues mais pas de manière significative en langue anglaise (qui connaît quelques expressions comme "résumé" ou "cliché").

	Moyenne des précisions (MAP)		
Enracineur	aucun	léger (-'s')	Porter
Okapi	**0,3395**	**0,3325 ***	**0,3242** *
DFR-PL2	0,3375	0,3310 *	0,3215 *
DFR-I(n_e)C2	0,3258	0,3202 *	0,3122 *
LM (λ=0,35)	0,2518	0,2464 *	0,2390 *
tf · idf	0,2129	0,2088 *	0,2033 *

Tableau 3. *Evaluation de nos divers modèles de dépistage selon trois algorithmes de suppression des séquences terminales (100 requêtes « titre »)*

Les évaluations de le tableau 3 indiquent que le modèle Okapi propose la meilleure qualité de réponse. Dans ce tableau, les différences de performance par rapport à la meilleure approche notée en gras et statistiquement significatives seront soulignées. Comme on le constate, la performance du modèle Okapi ne s'écarte pas significativement du modèle DFR-PL2. Les différences de performance avec les trois autres modèles s'avèrent par contre statistiquement significatives.

Si l'on pose comme référence la performance obtenue en l'absence de tout traitement morphologique, la suppression des suffixes tend à réduire la performance, et les différences avec un enracineur léger ou plus sophistiqué sont statistiquement significatives (notées par un astérisque '*' dans le tableau 3). En moyenne, ces différences de performance sont relativement faibles, soit de -1,9 % avec un enracineur léger ou -4,6 % avec l'algorithme de Porter. La recherche dans la *blogosphère* ne doit pas, contrairement à la recherche dans les dépêches d'agence, recourir à un enracineur même dans une version limitée à la suppression de la lettre finale '-s'.

Afin de connaître les problèmes particuliers de nos diverses stratégies de dépistage, nous avons analysé quelques interrogations présentant une performance

7. Notons que les fonctionnalités d'un moteur commercial comme Google peuvent changer sans que les internautes en soient avertis. Selon nos dernières analyses, la présente requête retourne, dans les dix premiers résultats, uniquement des sites ayant trait au jeu de Nim.

très faible avec notre meilleure approche (Okapi et sans enracineur). La demande n° 916 “dice.com” possède une précision moyenne de 0,0 et aucune bonne réponse n’a été dépistée. La forme interne de la requête se limitait aux deux termes “dice” et “com” provoquant l’extraction d’un nombre considérable de pages ayant un caractère *spam* indéniable ou pointant vers des sites de jeux enligne (“dice game”). Or l’internaute désirait spécifiquement des commentaires ou jugements concernant spécifiquement le site « dice.com ».

Avec l’interrogation n° 928 (précision moyenne de 0,0005, première bonne réponse au rang 115), l’internaute souhaitait recevoir des opinions concernant l’émission de télévision de HBO “Big Love” et ses participants. Avec une représentation interne {“big”, “love”}, le système de dépistage n’est pas arrivé à retourner en premier des *blogs* liés spécifiquement à l’émission de télévision concernée.

Avec la requête n° 937 “LexisNexis”, notre système a retourné en première place une bonne réponse mais la précision moyenne de cette requête demeure faible (0,0355) car il existe de nombreuses bonnes réponses (précisément 210). De nombreuses pages web contiennent la forme exacte apparaissant dans la demande mais souvent sous la forme d’un lien vers le site de l’entreprise LexisNexis. L’usager désirait lui des informations concernant la qualité de service du système LexisNexis et non des offres promotionnelles ou des annonces de nouveaux produits ou services liés à cette firme.

Si l’on compare l’indexation sans suppression des suffixes et l’algorithme de Porter, nous pouvons illustrer les différences avec la requête n° 936 “grammys”. Avec l’absence de tout traitement morphologique, cette requête obtient une précision moyenne de 0,0513 et le premier document pertinent se trouve au dixième rang (il existe 420 bonnes réponses à cette interrogation). Avec l’algorithme de Porter, la précision moyenne baisse à 0,0445 mais la première bonne réponse se situe au 616e rang. En fait la fréquence documentaire passe de 3 456 (terme “grammys” sans enracineur) à 9 899 (terme d’indexation “grammi” avec l’approche de Porter).

4.3. *Evaluation avec des requêtes plus longues*

En utilisant les cinq modèles de recherche et sans enracineur, le tableau 4 indique que le modèle le plus performant dépend de la longueur de la requête, soit l’approche Okapi pour des requêtes très courtes (T), ou le modèle DFR-PL2 si l’on considère des requêtes de longueur moyenne (TD) ou longue (TDN).

Les différences de performance entre le modèle Okapi d’une part et, d’autre part, l’approche DFR-PL2 ne s’avèrent pas statistiquement significatives. Par contre ces variations sont significatives entre le modèle le plus performant et le modèle de langue (LM) ou l’approche classique $tf \cdot idf$ (valeurs soulignées dans le tableau 4).

Comparé à l'emploi des requêtes très courtes (T), les requêtes « titre & descriptif » (ou TD) permettent d'accroître la performance moyenne de l'ordre de 13,3 % tandis que pour les requêtes longues (TDN), cette augmentation s'élève à 12,7 %. Comparées aux requêtes très courtes (T), ces différences de performance sont toujours statistiquement significatives sauf pour le modèle $tf \cdot idf$ (différence significative indiquée avec le symbole '*'). Dans ce dernier cas, la variation de la longueur de la requête n'a pas vraiment d'impact sur la performance obtenue. Pour l'ensemble des modèles, la différence de performances entre les requêtes TD et TDN demeure marginale et souvent contradictoire (pour les modèles les plus performants, les requêtes TD apportent une meilleure performance).

	Moyenne des précisions (MAP)		
Type de requite	T	TD	TDN
Nombre moyen termes	1,73	6,62	18,43
Okapi	**0,3395**	0,3786 *	0,3686 *
DFR-PL2	0,3375	**0,3821 ***	**0,3693 ***
DFR-I(n_e)C2	0,3258	0,3722 *	0,3630 *
LM (λ=0,35)	0,2518	0,3166 *	0,3357 *
$tf \cdot idf$	0,2129	0,2132	0,2178
Moyenne		+ 13,3 %	+ 12,7 %

Tableau 4. *Evaluation de nos divers modèles de dépistage selon trois types de requêtes (100 requêtes, sans enracineur)*

Si l'on regarde la figure 3, on comprend bien que les termes ajoutés par la partie descriptive (nombre moyen de termes par requête passe de 1,73 à 6,62) s'avèrent, en général, plus adéquats afin de discriminer entres les pages abordant le thème sous-jacent à la demande et celles plus périphériques à la requête. La partie narrative (longueur moyenne des requêtes de 18,43 mots) ajoute beaucoup de mots dans la requête sans que ces derniers apportent des éléments autorisant un meilleur classement des documents pertinents.

4.4. *Pseudo-rétroaction positive*

Lorsque l'on mesure la performance par la précision moyenne, le recours à une pseudo-rétroaction (Efthimiadis, 1996 ; Buckley *et al.*, 1996) afin d'élargir automatiquement les requêtes courtes permet d'augmenter la qualité du dépistage. Une telle approche semble, *a priori*, aussi attractive dans le contexte de la *blogosphère* puisque l'augmentation de la longueur des requêtes décrite dans la section précédente apportait une augmentation de la précision moyenne. Cependant,

cet enrichissement était fait manuellement par l'usager. De plus, après une augmentation de la performance, l'accroissement de la taille des requêtes conduisait à une légère dégradation (voir tableau 4).

Afin de procéder à une expansion automatique des interrogations soumises, nous avons implémenté l'approche de (1971) avec les constantes $\alpha = 0{,}75$ et $\beta = 0{,}75$ et en incluant entre 10 et 20 nouveaux termes extraits des 3 à 10 premiers *blogs* dépistés. Les résultats obtenus sont indiqués dans le tableau 5 et les différences de performance demeurent relativement faibles et ne sont habituellement pas significatives (les variations significatives sont indiquées par un soulignement). On remarque également que les deux mesures, la moyenne des précisions moyennes ou MAP et score du premier document pertinent dépisté (MRR) ne corroborent pas parfaitement. Les variations entre ces deux mesures étant toutefois mineures.

Requête « titre » seulement	MAP	MRR
Modèle avant (Okapi)	0,3395	0,7421
3 documents / 10 termes	0,3298	0,7590
3 documents / 20 termes	0,3142	0,7359
5 documents / 10 termes	**0,3472**	0,7753
5 documents / 20 termes	0,3313	0,7635
10 documents / 10 termes	0,3456	**0,8122**
10 documents / 20 termes	0,3394	0,8006

Tableau 5. *Evaluation avant et après l'expansion automatique des requêtes*

Pour expliquer cette faible variation de performance, nous pouvons suivre les indications données par Peat et Willett (1991). Ces auteurs indiquent, qu'en moyenne, les termes des requêtes tendent à avoir une fréquence plus importante que la moyenne. Selon la formule de Rocchio, les nouveaux termes à inclure dans la requête ont tendance à être présents dans plusieurs documents classés au début des réponses retournées et donc ils possèdent également une fréquence d'apparition importante. L'injection de ces termes n'améliore pas la discrimination entre les articles pertinents et ceux qui ne le sont pas. Le résultat final aboutit alors à une dégradation de la performance de la recherche.

4.5. *Comparaison avec les résultats de TREC*

Notre méthodologie d'évaluation s'appuie sur la présence de 100 requêtes afin de déterminer l'efficacité du système de recherche proposé. Afin d'avoir une idée de leur efficacité comparée aux meilleurs systèmes de dépistage proposés lors des

campagnes d'évaluation en 2006 et en 2007, le tableau 6 indique la précision moyenne calculée séparément pour les deux années.

Pour l'année 2006 (requêtes n° 851 à 900), le meilleur système de dépistage a été proposé par Indiana University (Yang, 2006) avec une précision moyenne de 0,2983. Pour l'année 2007 (requêtes n° 901 à 950), la meilleure approche a été proposée par l'Illinois University à Chicago (Zhang et yu, 2007). Comme l'indiquent les valeurs du tableau 6, le modèle Okapi présente une meilleure performance pour l'année 2006 mais qui s'avère inférieure pour l'année 2007. Signalons que pour 2007, l'accroissement de la précision moyenne provient d'expansions automatiques des requêtes. En effet, les participants ont remarqué que les requêtes correspondaient bien à des nouvelles et thématiques récurrentes sur le *web*. Ainsi, on a proposé d'utiliser directement le moteur *Google* (ou sa version adaptée aux *blogs*) afin d'extraire des termes appropriés afin d'élargir les interrogations. Parfois, on propose d'utiliser le corpus de nouvelles AQUAINT (Ernsting *et al*., 2007) ou le site Wikipédia (Zhang et yu, 2007) pour extraire les termes adéquats (Ernsting *et al*., 2007). Signalons également que pour certains participants (Ernsting *et al*., 2007), toutes les pages n'ont pas la même probabilité *a priori* d'être pertinente et que le nombre de billet inclus dans un article pourrait être un indicateur de la popularité et donc de la pertinence de la page sous-jacente.

	Moyenne des précisions (MAP)	
Modèle	TREC 2006	TREC 2007
Okapi (T)	0,3091	0,3699
& 5 documents / 10 termes (T)	0,3111	0,3834
& deux termes (T) (cf. section 5)	**0,3202**	0,4112
Indiana Univ. (TDN) (Yang, 2006)	0,2983	
Illinois Univ. (T) (Zhang et yu, 2007)		**0,4819**

Tableau 6. *Evaluation des deux meilleurs systèmes lors des campagnes TREC comparés au modèle Okapi*

Il faut cependant prendre garde de ne pas comparer les niveaux de performance d'une collection à une autre. Dans le cas présent, on ne peut pas inférer que les systèmes de recherche pour la *blogosphère* se sont sensiblement améliorés en comparant directement la performance obtenue en 2006 à celle obtenue en 2007 comme l'indique Macdonald *et al*. (2007).

En effet, toute comparaison doit être faite avec les mêmes données (collection *et* requêtes) et comme l'indique Buckley et Voorhees (2005), toute mesure de performance (MAP ou MRR) reste relative.

> "The primary consequence of the noise is the fact that evaluation scores computed from a test collection are *relative* scores only. The only valid use for such scores is to compare them to scores computed for other runs using the exact same collection." (BUCKLEY, 2005, p. 73).

Nous ne pouvons donc pas comparer directement les deux colonnes chiffrées du tableau 6. Par contre, nous pouvons clairement indiquer que l'élargissement des requêtes *via* des ressources externes (corpus similaires) tend à apporter des améliorations sensibles de la performance moyenne.

5. Favoriser des réponses possédant plusieurs mots recherchés

Suite à l'analyse des requêtes pour lesquelles notre système de dépistage de l'information présentait des lacunes, nous avons décidé d'améliorer notre algorithme de recherche. Dans ce dessein, nous avons désiré conserver une grande rapidité dans le traitement des requêtes. De plus, nous avons également décidé de renoncer à recourir à diverses sources externes que nous ne maitrisons pas (e.g., Google) ou que nous ne possédons pas.

Dans un premier temps, nous avons décidé d'étudier l'impact de la présence d'une liste de mots-outils très brève à la place de notre liste de 571 formes. En effet, pour quelques requêtes, la présence de mots inclus dans une liste trop longue peut diminuer sensiblement la performance. Ainsi, notre liste de 571 formes contenait les mots « big » et « com » dont l'importance est indéniable dans les interrogations « big love » ou « dice.com». Comme alternative, nous avons sélectionné les neuf mots retenus par le système DIALOG (soit les mots « an », « and », « by », « for », « from », « of », « the », « to », « with ») (Harter, 1986).

Comme deuxième voie d'amélioration, nous tenons à favoriser les réponses dépistées ayant deux mots (ou plus) appartenant à la requête. Notre intention consiste à améliorer le classement des pages *web* ayant, par exemple, les deux termes "dice" et "com" ou "big" et "love" apparaissant de manière adjacente. Pour atteindre cet objectif, notre indexation par termes isolés se complétera d'une indexation par paires de termes d'indexation adjacents. Ainsi la phrase « Big love in Paris » sera indexée par les termes « big, love, paris, big+love, love+paris ». On y retrouve les termes simples et les paires de termes adjacents après suppression des mots-outils.

Requête « titre » seulement	MAP	MRR
Modèle de reference (Okapi)	0,3395	0,7421
avec une stop liste brève	0,3221	0,7372
avec deux mots	0,3657	0,7835

Tableau 7. *Evaluation avant et après l'emploi d'une liste de neuf mots-outils ou favorisant la présence d'au moins deux mots de la requête (modèle Okapi, requête « titre » uniquement)*

Dans le tableau 7, nous avons repris en deuxième ligne le modèle Okapi en utilisant uniquement les requêtes très courtes (T) et sans suppression des suffixes. Ensuite nous avons évalué la performance obtenue avec notre liste brève de mots-outils. La différence de performance s'avère faible (0,3395 *vs*. 0,3221, -5,1 %) mais elle est tout de même significative.

En dernière ligne, nous avons reporté la performance de notre système qui favorise la présence de deux mots adjacents de la requête dans les pages retournées. Comme plusieurs requêtes sont composées que d'un seul terme (moyenne : 1,73 ; médiane : 2), cet accroissement ne peut pas être extrêmement fort. Selon la précision moyenne, l'accroissement s'élève à 0,3657, soit une augmentation statistiquement significative de 7,7 %.

En analysant quelques demandes, on constate que le plus souvent l'effet s'avère favorable comme pour l'interrogation n° 928 "Big Love". Dans ce cas, la précision moyenne passe de 0,0005 avec la première bonne réponse au rang 115 à une précision moyenne de 0,182 (la première bonne réponse se place au premier rang). Le scénario est le même pour la requête n° 916 "dice.com" qui ne dépistait aucune bonne réponse (précision moyenne = 0,0). Après le traitement des couples de mots, cette demande obtient une précision moyenne de 0,1997 et le premier article dépisté s'avère pertinent. Par contre pour la demande n° 927 "oscar fashion" la première réponse pertinente passe du deuxième rang au rang 50 après notre traitement des termes adjacents. La précision moyenne se dégrade également puisqu'elle passe de 0,0261 à 0,018. Dans ce cas, les autres articles présentés entre le premier et le 50e rang contiennent bien les mots "oscar" et "fashion" côte à côte (en fait il s'agit du syntagme "oscar fashion 2003") mais cette conjonction appartiennent à un menu et ne s'avère pas être un descripteur pertinent de la page web considérée.

6. Conclusion

Sur la base d'un corpus extrait de la *blogosphère* et accompagné de 100 requêtes, nous avons démontré que le modèle Okapi ou une approche dérivée du paradigme *Divergence from Randomness* apporte la meilleure performance. Afin d'obtenir de bonnes performances, il est recommandé de ne pas supprimer les séquences terminales, que ce soit uniquement la marque du pluriel avec un enracineur léger ou en éliminant également certains suffixes dérivationnels (voir tableau 3).

Si les internautes rédigent des demandes plus longues, une augmentation moyenne de la précision d'environ 13 % est attendue (voir tableau 4, colonne « TD »). Mais après l'inclusion d'un certain nombre de termes, l'accroissement de la requête par l'usager tend à diminuer la précision moyenne (tableau 4, colonne « TDN »). Le recours à un enrichissement automatique par pseudo-rétroaction n'apporte pas toujours d'amélioration de la précision moyenne ou du rang de la première bonne réponse (voir tableau 5). De plus, il demeure délicat de fixer de manière optimale les paramètres sous-jacents à cette approche.

Face à des requêtes très courtes (en moyenne 1,73 mots), l'indexation par paire de termes adjacents permet d'accroître significativement la précision moyenne (voir tableau 7). On passe ainsi d'une précision moyenne de 0,339 à 0,366 tandis que le score de l'inverse de la première page pertinente retournée passe de 0,74 à 0,78. Notre proposition améliore clairement le rang du premier document pertinent dépisté comme le confirme l'analyse de quelques requêtes difficiles.

Ces premiers résultats ouvrent la porte vers de nouvelles analyses afin de répondre à l'ensemble de nos questions. Nous n'avons pas vraiment l'impression que la qualité orthographique des documents de la *blogosphère* était nettement inférieure à celle que l'on retrouve dans des corpus de presse (Jereczek-Lipinska, 2007). Existe-t-il donc une certaine continuité des caractéristiques linguistiques entre les quotidiens et la *blogosphère*, ou, au contraire, une rupture existe mais n'a pas de réel impact sur les systèmes de dépistage ? Une réponse plus complète mériterait une analyse plus approfondie.

De même, l'inclusion de documents rédigés dans d'autres langues que l'anglais n'a pas perturbé de manière significative la qualité du dépistage de l'information. La présence de *spam* mériterait une analyse plus détaillée car ce sujet n'a pas vraiment été abordé avec l'attention qu'il mériterait lors des deux dernières campagnes d'évaluation TREC (Ounis *et al.*, 2006 ; Macdonald *et al.*, 2007). La présence des métabalises (e.g., « Keywords » et « Description ») mériterait également une analyse afin de connaître leur impact lors de la recherche d'information. Finalement, nous n'avons pas tenu compte de la date à laquelle les *blogs* sont apparus sur internet, une composante qui doit certainement jouer un rôle dans l'appréciation faite par l'internaute.

Comme autre perspective ouverte par la *blogosphère*, nous pouvons signaler que la nature subjective des billets d'information mériterait un intérêt plus important de la communauté du traitement automatique de la langue naturelle. Ainsi, la réponse à une requête (e.g., « IKEA », « G. Bush », « tour Eiffel ») ne serait pas une simple liste de billets sur la thématique souhaitée mais une réponse distinguant clairement les faits des opinions. De plus, ces dernières, par nature subjective, pourraient être distinguées entre les avis positifs, ceux franchement négatifs ou des opinions plus nuancées sur le thème de la requête. L'emploi d'outil plus fin en traitement automatique de la langue pourrait même définir précisément l'auteur de l'opinion et sa délimitation à l'intérieur d'une phrase ou d'un paragraphe.

Remerciements

Cette recherche a été financée en partie par le Fonds national suisse pour la recherche scientifique (subside n^0 200021-113273).

7. Bibliographie

Abdou S. et Savoy J., « Considérations sur l'évaluation de la robustesse en recherche d'information », *Actes CORIA'07*, St-Etienne, 2007, p. 5-30.

Amati G., van Rijsbergen C.J., "Probabilistic models of information retrieval based on measuring the divergence from randomness", *ACM-Transactions on Information Systems*, vol. 20, n° 4, 2002, p. 357-389.

Boughanem M., Savoy J., *Recherche d'information. Etat des lieux et perspectives*, Hermès, Paris, 2008.

Buckley C., Singhal A., Mitra M., Salton G., "New retrieval approaches using SMART", *Proceedings of TREC-4*, NIST Publication #500-236, Gaithersburg (MD), 1996, p. 25-48.

Buckley C., Voorhees E., "Retrieval system evaluation", *TREC, Experiment and Evaluation in Information Retrieval*, The MIT Press, Cambridge (MA), 2005, p. 53-75.

Efthimiadis E.N., *Query expansion*, Annual Review of Information Science and Technology, 31, 1996, p. 121-187.

Ernsting B., Weerkamp W., Yude Rijke M., "The university of Amsterdam at TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

Fairon C., Klein J.R., Paumier S., *Le langage SMS*, Presses universitaires de Louvain, Louvain-la-Neuve, 2006.

Harter S.P., *Online information retrieval. Concepts, principles, and techniques*, Academic Press, San Diego, 1986.

Harman D., "How effective is suffixing?", *Journal of the American Society for Information Science*, vol. 42, n° 1, 1991, p. 7-15.

Hiemstra D., Using language models for information retrieval, CTIT Ph.D. Thesis, 2000.

Jereczek-Lipinska J., « Le blog en politique. Outil de démocratie électronique participative ? », *Grottopol*, vol. 10, 2007, p. 159-172.

Macdonald C., Ounis, I., Soboroff I., "Overview of the TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

Malaison C., *Pourquoi bloguer : dans un contexte d'affaires ?*, Editions IQ, Montréal, 2007.

Ounis I., de Rijke M., Macdonald C., Mishne G., Soboroff I., "Overview of the TREC-2006 blog track", *Proceedings of TREC-2006*, NIST Publication #500-272, Gaithersburg (MD), 2006, p. 17-32.

Peat H.J., et Willett P., "The limitations of term co-occurrence data for query expansion in document retrieval systems", *Journal of the American Society for Information Science*, vol. 42, n° 5, 1991, p. 378-383.

Porter M.F., "An algorithm for suffix stripping", *Program*, vol. 14, 1980, p. 130-137.

Robertson S.E., Walker, S., Beaulieu M., "Experimentation as a way of life: Okapi at TREC", *Information Processing & Management*, vol. 36, n° 1, 2000, p. 95-108.

132 DN – 11/2008. Documents et web 2.0

Rocchio J.J.Jr., "Relevance feedback in information retrieval", G. Salton (Ed.), *The SMART Retrieval System*. Prentice-Hall Inc., Englewood Cliffs (NJ), 1971, p. 313-323

Sakai T., "On the reliability of information retrieval metrics based on graded relevance", *Information Processing & Management*, vol. 43, n° 2, 2007, p. 531-548.

Savoy J., "Statistical inference in retrieval effectiveness evaluation", *Information Processing & Management*, vol. 33, n° 4, 1997, p. 495-512.

Savoy J., « Un regard statistique sur l'évaluation de performance. L'exemple de CLEF 2005 », *Actes CORIA'06*, Lyon, 2006, p. 73-84.

Singh A. K., "Study of some distance measures for language and encoding idientification", *Proceedings Workshop Coling-ACL* "Linguistic distances", Sydney, 2007, p. 63-724.

Véronis E., Véronis J., Voisin N., *Les politiques mis au net*, Max Milo Editions, Paris, 2007.

Voorhees E.M., "TREC: Continuing information retrieval's tradition of experimentation", *Communications of the ACM*, vol. 50, n° 11, 2007, p. 51-54.

Yang K., Yu N., Valerio A., Zhang H., "WIDIT in TREC-2006 blog track", *Proceedings of TREC-2006*, NIST Publication #500-272, Gaithersburg (MD), 2006.

Witten I.H., Gori M., Numerico T., "*Web Dragons*", Elsevier, Amsterdam, 2007.

Zhang W., Yu C., "UIC at TREC-2007 blog track", *Notebook of TREC-2007*, Gaithersburg (MD), 2007.

UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere

Claire Fautsch, Jacques Savoy
Computer Science Department, University of Neuchatel
Rue Emile-Argand, 11, CH-2009 Neuchatel (Switzerland)
{Claire.Fautsch, Jacques.Savoy}@unine.ch

ABSTRACT

This paper describes our participation in the Blog track at the TREC 2008 evaluation campaign. The Blog track goes beyond simple document retrieval, its main goal is to identify opinionated blog posts and assign a polarity measure (positive, negative or mixed) to these information items. Available topics cover various target entities, such as people, location or product for example. This year's Blog task may be subdivided into three parts: First, retrieve relevant information (facts & opinionated documents), second extract only opinionated documents (either positive, negative or mixed) and third classify opinionated documents as having a positive or negative polarity.

For the first part of our participation we evaluate different indexing strategies as well as various retrieval models such as Okapi (BM25) and two models derived from the *Divergence from Randomness* (DFR) paradigm. For the opinion and polarity detection part, we use two different approaches, an additive and a logistic-based model using characteristic terms to discriminate between various opinion classes.

1. INTRODUCTION

In the Blog track [1] the retrieval unit consists of permalink documents, which are URLs pointing to a specific blog entry. In contrast to a corpus extracted from scientific papers or a news collection, blogposts are more subjective in nature and contain several points of view on various domains. Written by different kinds of users, a post retrieved following the request "TomTom" for might contain factual information about the navigation system, such as software specifications for example, but it might also contain more subjective information about the product such as ease of use. The ultimate goal of the Blog track is to find opinionated documents rather than present a ranked list of relevant documents containing either objective (facts) or subjective (opinions) content. Thus, in a first step the system would retrieve a set of relevant documents but then a second step this set would be separated into two subsets, one containing the documents without any opinions (facts) and the second containing documents expressing positive, negative or mixed opinions on the target entity. Finally the mixed-opinion documents would be eliminated and the positive and negative opinionated documents separated. Later in this paper, the documents retrieved during the first step will be referenced as a baseline or factual retrieval.

The rest of this paper is organized as follows. Section 2 describes the main features of the test-collection used. Section 3 explains the indexing approaches used and Section 4 introduces the models used for factual retrieval. In Section 5 we explain our opinion and polarity detection algorithms. Section 6 evaluates the different approaches as well as our official runs. The principal findings of our experiments are presented in Section 7.

2. BLOG TEST-COLLECTION

The Blog test collection contains approximately 148 GB of uncompressed data, consisting of 4,293,732 documents extracted from three sources: 753,681 feeds (or 17.6%), 3,215,171 permalinks (74.9%) and 324,880 homepages (7.6%). Their sizes are as follows: 38.6 GB for feeds (or 26.1%), 88.8 GB for permalinks (60%) and 20.8 GB for the homepages (14.1%). Only the permalink part is used in this evaluation campaign. This corpus was crawled between Dec. 2005 and Feb. 2006 (for more information see: http://ir.dcs.gla.ac.uk/test_collections/).

Figures 1 and 2 show two blog document examples, including the date, URL source and permalink structures at the beginning of each document. Some information extracted during the crawl is placed after the **<DOCHDR>** tag. Additional pertinent information is placed after the **<DATA>** tag, along with ad links, name sequences (e.g., authors, countries, cities) plus various menu or site map items. Finally some factual information is included, such as some locations where various opinions can be found. The data of interest to us follows the <**DATA**> tag.

```
<DOC>
<DOCNO> BLOG06-20051212-051-0007599288
<DATE_XML> 2005-10-06T14:33:40+0000
<FEEDNO> BLOG06-feed-063542
<FEEDURL> http://
contentcentricblog.typepad.com/ecourts/index.rdf
<PERMALINK>
http://contentcentricblog.typepad.com/ecourts/20
05/10/efiling_launche.html#
<DOCHDR> …
Date: Fri, 30 Dec 2005 06:23:55 GMT
Accept-Ranges: bytes
Server: Apache
Vary: Accept-Encoding,User-Agent
Content-Type: text/html; charset=utf-8
…
<DATA>
electronic Filing &amp;amp; Service for Courts
…
October 06, 2005
eFiling Launches in Canada
Toronto, Ontario, Oct.03 /CCNMatthews/ -
LexisNexis Canada Inc., a leading provider of
comprehensive and authoritative legal, news, and
business information and tailored applications
to legal and corporate researchers, today
announced the launch of an electronic filing
pilot project with the Courts
…
```

Figure 1. Example of LexisNexis blog page

```
<DOC>
<DOCNO>  BLOG06-20060212-023-0012022784
<DATE_XML> 2006-02-10T19:08:00+0000
<FEEDNO> BLOG06-feed-055676
<FEEDURL>  http://
lawprofessors.typepad.com/law_librarian_blog/ind
ex.rdf#
<PERMALINK>
http://lawprofessors.typepad.com/law_librarian_b
log/2006/02/free_district_c.html#
<DOCHDR> …
Connection: close
Date: Wed, 08 Mar 2006 14:33:59 GMT …
<DATA>
Law Librarian Blog

Blog Editor
Joe Hodnicki
 Associate Director for Library Operations
 Univ. of Cincinnati Law Library
…
News from PACER   :

In the spirit of the E-Government Act of 2002,
modifications have been made to the District
Court CM/ECF system to provide PACER customers
with access to written opinions free of charge

The modifications also allow PACER customers to
search for written opinions using a new report
that is free of charge. Written opinions have
been defined by the Judicial Conference as any
document issued by a judge or judges of the
court sitting in that capacity, that sets forth
a reasoned explanation for a court's decision. …
```

Figure 2. Example of blog document

During this evaluation campaign a set of 50 new topics (Topics #1001 to #1050) as well as 100 old topics from 2006 and 2007 (respectively Topics #851 to #900 and #901 to #950) were used. They were created from this corpus and express user information needs extracted from the log of a commercial search engine blog. Some examples are shown in Figure 3.

```
<num> Number: 851
<title> "March of the Penguins"
<desc> Description:
Provide opinion of the film documentary
"March of the Penguins".
<narr> Narrative:
Relevant documents should include opinions
concerning the film documentary "March of
the Penguins".  Articles or comments about
penguins outside the context of this film
documentary are not relevant.

<num> Number: 941
<title> "teri hatcher"
<desc> Description:
Find opinions about the actress Teri
Hatcher.
<narr> Narrative:
All statements of opinion regarding the
persona or work of film and television
actress Teri Hatcher are relevant.

<num> Number: 1040
<title> TomTom
<desc> Description:
What do people think about the TomTom GPS
navigation system?
<narr> Narrative:
How well does the TomTom GPS navigation
system meets the needs of its users?
Discussion of innovative features of the
system, whether designed by the
manufacturer or adapted by the users, are
relevant.
```

Figure 3. Three examples of Blog track topics

Based on relevance assessments (relevant facts & opinions, or relevance value ≥ 1) made on this test collection, we listed 43,813 correct answers. The mean number of relevant web pages per topic is 285.11 (median: 240.5; standard deviation: 222.08). Topic #1013 ("Iceland European Union") returned the minimal number of pertinent passages (12) while Topic #872 ("brokeback mountain") produced the greatest number of relevant passages (950).

Based on opinion-based relevance assessments ($2 \leq$ relevance value ≤ 4), we found 27,327 correct opinionated posts. The mean number of relevant web pages per topic is 175.99 (median: 138; standard deviation: 169.66). Topic #877 ("sonic food industry"), Topic #910 ("Aperto Networks") and Topic #950 ("Hitachi Data Systems") returned a minimal number of pertinent passages (4) while Topic #869 ("Muhammad cartoon") produced the greatest number of relevant posts (826).

The opinion referring to the target entity and contained in a retrieved blogpost may be negative (relevance

value = 2), mixed (relevance value = 3) or positive (relevance value = 4). From an analysis of negative opinions only (relevance value = 2), we found 8,340 correct answers (mean: 54.08; median: 33; min: 0; max: 533; standard deviation: 80.20). For positive opinions only (relevance value = 4), we found 10,457 correct answers (mean: 66.42, median: 46; min: 0; max: 392; standard deviation: 68.99). Finally for mixed opinions only (relevance value = 3), we found 8,530 correct answers (mean: 55.48; median: 23; min: 0; max: 455; standard deviation: 82.33). Thus it seems that the test collection tends to contain, in mean, more positive opinions (mean 66.42) than it does either mixed (mean: 55.48) or negative opinions (mean: 54.08) related to the target entity.

3. INDEXING APPROACHES

We used two different indexing approaches to index documents and queries. As a first and natural approach we chose words as indexing units and their generation was done in three steps. First, the text is tokenized (using spaces or punctuation marks), hyphenated terms are broken up into their components and acronyms are normalized (e.g., U.S. is converted into US). Second, uppercase letters are transformed into their lowercase forms and third, stop words are filtered out using the SMART list (571 entries). Based on the result of our previous experiments within the Blog track [2] or Genomics search [3], we decided not to use a stemming technique.

In its indexing units our second indexing strategy uses single words and also compound constructions, with the latter being those composed of two consecutive words. For example in the Query #1037 "*New York Philharmonic Orchestra"* we generated the following indexing units after stopword elimination: "york," "philharmonic," "orchestra," "york philharmonic," "philharmonic orchestra" ("new" is included in the stoplist). We decided to use this given the large number of queries containing proper names or company names such as "David Irving" (#1042), "George Clooney" (#1050) or "Christianity Today" (#921) for example should be considered as one single entity for both indexing and retrieval. Once again we did not apply any stemming procedure.

4. FACTUAL RETRIEVAL

The first step in the Blog task was factual retrieval. To create our baseline runs (factual retrieval) we used different single IR models as described in Section 4.1. To produce more effective ranked results lists we applied different blind query expansion approaches as discussed in Section 4.2. Finally, we merged different isolated runs using a data fusion approach as presented in Section 4.3. This final ranked list of retrieved items was used as our baseline (classical *ad hoc* search).

4.1 Single IR Models

We considered three probabilistic retrieval models for our evaluation. As a first approach we used the Okapi (BM25) model [4], evaluating the document D_i score for the current query Q by applying the following formula:

$$Score(D_i, Q) = \sum_{t_j \in q} qtf_j \cdot \log\left(\frac{n - df_j}{df_j}\right) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}, \quad (1)$$

$$\text{where } K = k_1 \cdot \left[(1 - b) + b \cdot (l_i / avdl)\right]$$

in which the constant *avdl* was fixed at 837 for the word-based indexing and at 1622 for our compound-based indexing. For both indexes the constant *b* was set to 0.4 and k_1 to 1.4.

As a second approach, we implemented two models derived from the *Divergence from Randomness* (DFR) paradigm [5]. In this case, the document score was evaluated as:

$$Score(D_i, Q) = \sum_{t_j \in q} qtf_j \cdot w_{ij} \quad (2)$$

where qtf_j denotes the frequency of term t_j in query Q, and the weight w_{ij} of term t_j in document D_i was based on a combination of two information measures as follows:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf))$$

As a first model, we implemented the PB2 scheme, defined by the following equations:

$$Inf^1_{ij} = -\log_2[(e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}!] \quad \text{with } \lambda_j = tc_j / n \quad (3)$$

$$Prob^2_{ij} = 1 - [(tc_j + 1) / (df_j \cdot (tfn_{ij} + 1))]$$
$$\text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i)] \quad (4)$$

where tc_j indicates the number of occurrences of term t_j in the collection, l_i the length (number of indexing terms) of document D_i, *mean dl* the average document length (fixed at 837 for word-based respectively at 1622 for compound-based indexing approach), *n* the number of documents in the corpus, and *c* a constant (fixed at 5).

For the second model PL2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and $Prob^2_{ij}$ by Equation 5, as shown below:

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \quad (5)$$

where λ_j and tfn_{ij} were defined previously.

4.2 Query Expansion Approaches

In an effort to improve retrieval effectiveness, various query expansion techniques were suggested [6], [3], and in our case we chose two of them. The first uses a blind query expansion based on Rocchio's method [7], wherein

the system would add the top *m* most important terms extracted from the top *k* documents retrieved in the original query. As a second query expansion approach we used Wikipedia[1] to enrich those queries based on terms extracted from a source different from the blogs. The title of the original topic description was sent to Wikipedia and the ten most frequent words from the first retrieved article were added to the original query.

4.3 Combining Different IR Models

It was assumed that combining different search models would improve retrieval effectiveness, due to the fact that each document representation might retrieve pertinent items not retrieved by others. On the other hand, we might assume that an item retrieved by many different indexing and/or search strategies would have a greater chance of being relevant for the query submitted [8], [9].

To combine two or more single runs, we applied the Z-Score operator [10] defined as:

$$Z-Score\ RSV_k = \sum_j \left[\frac{RSV_k^j - Mean^j}{Stdev^j} + \delta_j \right] \qquad (6)$$

with $\delta^i = ((Mean^j - Min^j) / Stdev^j)$

In this formula, the final document score (or its retrieval status value RSV_k) for a given document D_k is the sum of the standardized document score computed for all isolated retrieval systems. This later value was defined as the document score for the corresponding document D_k achieved by the *j*th run (RSV_k^j) minus the corresponding mean (denoted $Mean^j$) and divided by the standard deviation (denoted $Stdev^j$).

5. OPINION AND POLARITY DETECTION

Following the baseline retrieval, the goal was to separate the retrieved documents into two classes, namely opinionated and non-opinionated documents, and then in a subsequent step assign a polarity to the opinionated documents.

In our view, opinion and polarity detection are closely related. Thus, after performing the baseline retrieval, our system would automatically judge the first 1,000 documents retrieved. For each retrieved document the system may classify it as positive, negative, mixed or neutral (the underlying document contains only factual information). To achieve this we calculated a score for each possible outcome class (positive, negative, mixed, and neutral), and then the highest of these four scores determined the choice of a final classification.

[1] http://www.wikipedia.org/

Then for each document in the baseline we looked up the document in the judged set to obtain its classification. If the document was not there it was classified as *unjudged.*

Documents classified as positive, mixed or negative were considered to be opinionated, while neutral and unjudged documents were considered as non-opinionated. This classification also gave the document's polarity (positive or negative).

To calculate the classification scores, we used two different approaches, both being based on Muller's method for identifying a text's characteristic vocabulary [11], as described in Section 5.1. We then presented our two suggested approaches, the additive model in Section 5.2 and the logistic approach in Section 5.3.

5.1 Characteristic Vocabulary

In Muller's approach the basic idea is to use Z-score (or standard score) to determine which terms can properly characterize a document, when compared to other documents. To do so we needed a general corpus denoted *C*, containing a documents subset *S* for which we wanted to identify the characteristic vocabulary. For each term *t* in the subset *S* we calculated a Z-Score by applying Equation (7).

$$Z-Score(t) = \frac{f' - n' \Pr ob(t)}{\sqrt{n' \cdot \Pr ob(t) \cdot (1 - \Pr ob(t))}} \qquad (7)$$

where *f'* was the observed number of occurrences of the term *t* in the document set *S*, and *n'* the size of *S*. Prob(t) is the probability of the occurrence of the term *t* in the entire collection *C*. This probability can be estimated according to the Maximum Likelihood Estimation (MLE) principle as Prob(t) = *f*/*n*, with *f* being the number of occurrences of *t* in *C* and *n* the size of *C*. Thus in Equation 7, we compared the expected number of occurrences of term *t* according to a binomial process (mean = $n' \cdot$ Prob(t)) with the observed number of occurrences in the subset *S* (denoted *f'*). In this binomial process the variance is defined as $n' \cdot$ Prob(t) $\cdot$ (1-Prob(t)) and the corresponding standard deviation becomes the denominator of Equation 7.

Terms having a Z-score between –ε and +ε would be considered as general terms occurring with the same frequencies in both the entire corpus *C* and the subset *S*. The constant ε represents a threshold limit that was fixed at 3 in our experiments. On the other hand, terms having an absolute value for the Z-score higher than ε are considered overused (positive Z-score) or underused (negative Z-score) compared to the entire corpus *C*. Such terms therefore may be used to characterize the subset *S*.

In our case, we created the whole corpus *C* using all 150 queries available. For each query the 1,000 first retrieved documents would be included in *C*. Using the relevance assessments available for these queries (queries #850 to

#950), we created four subsets, based on positive, negative, mixed or neutral documents, and thus identified the characteristic vocabulary for each of these polarities. For each possible classification, we now had a set of characteristic terms with their Z-score.

Defining the vocabulary characterizing the four different classes in one step, and in the second step it is to compute an overall score, as presented in the following section.

5.2 Additive Model

In our first approach we used characteristic term statistics to calculate the corresponding polarity score for each document. The scores were calculated by applying following formulae:

$$\begin{aligned} Pos\_score &= \frac{\#PosOver}{\#PosOver + \#PosUnder} \\ Neg\_score &= \frac{\#NegOver}{\#NegOver + \#NegUnder} \\ Mix\_score &= \frac{\#MixOver}{\#MixOver + \#MixUnder} \\ Neutral\_score &= \frac{\#NeuOver}{\#NeuOver + \#NeuUnder} \end{aligned} \quad (8)$$

in which *#PosOver* indicated the number of terms in the evaluated document that tended to be overused in positive documents (i.e. Z-score > ε) while *#PosUnder* indicated the number of terms that tended to be underused in the class of positive documents (i.e. Z-score < -ε). Similarly, we defined the variables *#NegOver, #NegUnder, #MixOver, #MixUnder, #NeuOver, #NeuUnder,* but for their respective categories, namely negative, mixed and neutral.

The idea behind this first model is simply assigning the category to each document for which the underlying document has relatively the largest sum of overused terms. Usually, the presence of many overused terms belonging to a particular class is sufficient to assign this class to the corresponding document.

5.3 Logistic Regression

As a second classification approach we used logistic regression [12] to combine different sources of evidence. For each possible classification, we built a logistic regression model based on twelve covariates and fitted them using training queries #850 to #950 (for which the relevant assessments were available). Four of the twelve covariates are *SumPos, SumNeg, SumMix, SumNeu* (the sum of the Z-scores for all overused and underused terms for each respective category). As additional explanatory variables, we also use the 8 variables defined in Section 5.2, namely *#PosOver, #PosUnder, #NegOver, #NegUnder, #MixOver, #MixUnder, #NeuOver,* and *#NeuUnder.* The score is defined as the logit transformation $\pi(x)$ given by each logistic regression model is defined as:

$$\pi(x) = \frac{e^{\beta_0 + \sum_{i=1}^{12} \beta_i x_i}}{1 + e^{\beta_0 + \sum_{i=1}^{12} \beta_i x_i}} \quad (9)$$

where β_i are the coefficients obtained from the fitting and x_i the variables. These coefficients reflect the relative importance of each explanatory variable in the final score.

For each document, we compute the $\pi(x)$ corresponding to the four possible categories and for the final decision we need simply to classify the post according to the maximum $\pi(x)$ value. This approach accounts for the fact that some explanatory variables may have more importance than others in assigning the correct category. We must recognize however that the length of the underlying document (or post) is not directly taken into account in our model. Our underling assumption is that all documents have a similar number of indexing tokens. As a final step we could simplify our logistic model by ignoring explanatory variables having a coefficient estimate (β_i) close to zero and for which a statistical test cannot reject the hypotheses that the real coefficient β_i = 0.

6. EVALUATION

To evaluate our various IR schemes, we adopted mean average precision (MAP) computed by `trec_eval` software to measure the retrieval performance (based on a maximum of 1,000 retrieved records). As the Blog task is composed of three distinct subtasks, namely the *ad hoc* retrieval task, the opinion retrieval task and the polarity task, we will present these subtasks in the three following sections.

6.1 Baseline Ad hoc Retrieval Task

A first step in the Blog track was the *ad hoc* retrieval task, where participants were asked to retrieve relevant information about a specified target. These runs also served as baselines for opinion and polarity detection. In addition the organizers provided 5 more baseline runs to facilitate comparisons between the various participants' opinion and polarity detection strategies. We based our official runs on two different indexes (single words under the label "W" and compound construction under the label "comp." see Section 3) and on two different probabilistic models (see Section 4). We evaluated these different approaches under three query formulations, T (title only), TD (title and description) and TD$^+$. In the latter case, the system received the same TD topic formulation as previously but during the query representation process the system built a phrase query from the topic description's title section. Table 1 shows the results and Table 2 the results of our two different query expansion techniques.

Model	T		TD		TD^+	
	comp.	W	comp.	W	comp.	W
Okapi	**0.374**	0.337	**0.403**	0.372	0.400	0.390
PL2	0.368	0.336	0.398	0.378	0.396	0.392
PB2	0.362	0.321	0.394	0.358	0.374	0.380

Table 1. MAP of different IR models (*ad hoc* search) (Blog, T & TD query formulations)

As shown in Table 1 the performance for the Okapi and the DFR schemes is almost the same, with the Okapi perhaps having a slight advantage. This table also shows that using compound indexing approach (word pairs) or phrase (from the title section of the query) increases the performance. This can be explained by the fact that in the underling test collection numerous queries contain statements that should appear together or close together in the retrieved documents, such as names (e.g. #892 "Jim Moran", #902 "Steve Jobs" or #931 "fort mcmurray") or concepts (e.g. #1041 "federal shield law"). Finally it can also be observed that adding the descriptive part (D) in the query formulation might improve the MAP.

	T	
	comp.	W
Okapi (baseline)	0.374	0.336
Rocchio 5 doc/ 10 terms	0.387	**0.344**
Rocchio 5 doc/ 20 terms	0.386	0.331
Rocchio 5 doc/ 100 terms		0.253
Rocchio 10 doc/ 10 terms	0.384	0.343
Rocchio 10 doc/ 20 terms	**0.390**	0.339
Rocchio 10 doc/ 100 terms		0.277
Wikipedia	0.387	0.342

Table 2. Okapi model with various blind query expansions

Table 2 shows that Rocchio's blind query expansion might slightly improve the results, but only if a small number of terms is considered. When adding a higher number of terms to the original query, the system tends to include more frequent terms such as navigational terms (e.g. "home", "back", "next") that are not related to the original topic formulation. The resulting MAP tends therefore to decrease. Using Wikipedia as an external source of potentially useful search terms only slightly improves the results (an average improvement of +2.75% on MAP).

Table 3 lists our two official baseline runs for the Blog track and Table 4 the MAP for both the topic (or *ad hoc*) search and opinion search for our two official baseline runs, as well as for the additional five baseline runs provided by the organizers.

Run Name	Query	Index	Model	Expansion
UniNEBlog1	T	comp.	Okapi	Rocc. 5/20
	TD	comp.	PL2	none
	TD^+	W	PB2	none
UniNEBlog2	T	comp.	Okapi	Wikipedia
	T	comp.	Okapi	Rocc. 5/10

Table 3. Description of our two official baseline runs for *ad hoc* search

Run Name	Topic MAP	Opinion MAP
UniNEBlog1	**0.424**	**0.320**
UniNEBlog2	0.402	0.306
Baseline 1	0.370	0.263
Baseline 2	0.338	0.265
Baseline 3	0.424	0.320
Baseline 4	**0.477**	**0.354**
Baseline 5	0.442	0.314

Table 4. *Ad hoc* topic and opinion relevancy results for baseline runs

6.2 Opinion retrieval

In this subtask participants were asked to retrieve blog posts expressing an opinion about a given entity and then to discard factual posts. The evaluation measure adopted for the MAP meant the system was to produce a ranked list of retrieved items. The opinion expressed could either be positive, negative or mixed. Our opinion retrieval runs were based on our two baselines described in Section 6.1 as well as on the five baselines provided by the organizers. To detect opinion we used two approaches: Z-Score (denoted Z in the following tables) and logistic regression (denoted LR). This resulted in a total of 14 official runs. Table 5 lists the top three results for each of our opinion detection approaches.

Compared to the baseline results shown in Table 4 (under the column "Opinion MAP"), adding our opinion detection approaches after the factual retrieval process tended to hurt the MAP performance. For example, the run UniNEBlog1 achieved a MAP of 0.320 without any opinion detection and only 0.309 when using our simple additive model (-3.4%) or 0.224 with our logistic approach (-30%).

This was probably due to the fact that during the opinion detection phase we removed all the documents judged by our system to be non-opinionated. Ignoring such

documents thus produced a list clearly comprising less than 1,000 documents. Finally, Table 5 shows that having a better baseline also provides a better opinion run and that for opinion detection our simple additive model performed slightly better than the logistic regression approach (+36.47% on opinion MAP).

RunName	Baseline	Topic	Opinion
UniNEopLR1	UniNEBlog1	0.230	0.224
UniNEopLRb4	baseline 4	0.228	0.228
UniNEopLR2	UniNEBlog2	0.220	0.212
UniNEopZ1	UniNEBlog1	0.393	0.309
UniNEopZb4	baseline 4	**0.419**	**0.327**
UniNEopZ2	UniNEBlog2	0.373	0.296

Table 5. MAP of both ad hoc search and opinion detection

6.3 Polarity Task

In this third part of the Blog task, the system retrieved opinionated posts separated into a ranked list of positive and negative opinionated documents. Documents containing mixed opinions were not to be considered. The evaluation was done based on the MAP value, and separately for documents classified as positive and negative. As for the opinion retrieval task, we applied our two approaches in order to detect polarity in the baseline runs. Those documents that our system judged as belonging to either of the mixed or neutral categories were eliminated.

Table 6 lists the three best results (over 12 official runs) for each classification task. It is worth mentioning that for the positive classification task, we had 149 queries and for the negative opinionated detection only 142 queries provided at least one good response. The resulting MAP values were relatively low compared to the previous opinionated blog detection run (see Table 5).

For our official runs using logistic regression, we did not classify the documents into four categories (positive, negative, mixed and neutral) but instead into only three (positive, negative, mixed). This meant that instead of calculating four polarity scores, we calculated only three and assigned polarity to the highest one. Table 7 shows the results for the logistic regression approach, with three (without neutral) and four (with neutral) classifications.

RunName	Baseline	Positive MAP	Negative MAP
UniNEpolLRb4	baseline 4	0.102	0.055
UniNEpolLR1	UniNEBlog1	**0.103**	0.057
UniNEpolLRb5	baseline 5	0.102	0.055
UniNEpolZb5	baseline 4	0.070	**0.061**
UniNEpolZ5	baseline 5	0.067	0.058
UniNEpolZ3	baseline 3	0.067	0.063

Table 6. MAP evaluation for polarity detection

Baseline	With neutral		Without neutral	
	Positive	Negative	Positive	negative
UniNEBlog1	0.065	0.046	0.103	0.057
UniNEBlog2	0.064	0.042	0.102	0.051

Table 7. Logistic regression approach with three or four classifications

Using only three classification categories instead of four had a positive impact on performance, as can be seen from an examination of Table 7 (logistic regression method only). Most documents classified as "neutral" in the four-classification approach were then eliminated. When we considered only three categories, these documents were mainly classified as positive. This phenomenon also explains the differences in positive and negative MAP in our official runs when logistic regression was used (see Table 6).

7. CONCLUSION

During this TREC 2008 Blog evaluation campaign we evaluated various indexing and search strategies, as well as two different opinion and polarity detection approaches.

For the factual or *ad hoc* baseline retrieval we examined the underlying characteristics of this corpus with the compound indexing scheme that would hopefully improve precision measures. Compared to the standard approach in which isolated words were used as indexing units, in the MAP we obtained there was a +11.1% average increase for title only queries, as well as a +7.7% increase for title and description topic formulations. These results strengthen the assumption that for Blog queries such a precision-oriented feature could be useful. In further research, we might consider using longer tokens sequences as indexing unit, rather than just word pairs. Longer queries such as #1037 "New York Philharmonic Orchestra" or #1008 "UN Commission on Human Rights" might for example obtain better precision.

For the opinion and polarity tasks, we applied our two approaches to the given baselines as well as to two of our

own baselines. We noticed that applying no opinion detection provides better results than applying any one of our detection approaches. This was partially due to the fact that during opinion detection we eliminated some documents, either because they were judged "neutral" or because they were not contained in the judged pool of documents ("unjudged").

In a further step we will try to rerank the baselines instead of simply removing documents judged as non-opinionated. A second improvement to our approach could be judging each document at the retrieval phase instead of first creating a pool of judged documents. In this case we would no longer have any documents classified as "unjudged" although more hardware resources would be required. Polarity detection basically suffers from the same problem as opinion detection. Finally, we can conclude that having a good factual baseline is the most important part of opinion and polarity detection.

ACKNOWLEDGMENTS

The authors would also like to thank the TREC Blog task organizers for their efforts in developing this specific test-collection. This research was supported in part by the Swiss NSF under Grant #200021-113273.

8. REFERENCES

[1] Ounis, I., de Rijke, M., Macdonald, C., Gilad Mishne, G., & Soboroff, I. Overview of the TREC-2006 blog track. In *Proceedings of TREC-2006*. Gaithersburg (MA), NIST Publication #500-272, 2007.

[2] Fautsch C, & Savoy J. IR-Specific Searches at TREC 2007: Genomics & Blog Experiments. In *Proceedings TREC-2007*, NIST publication #500-274, Gaithersburg (MD), 2008.

[3] Abdou S., & Savoy J. Searching in Medline: Stemming, Query Expansion, and Manual Indexing Evaluation. *Information Processing & Management*, 44(2), 2008, 781-789.

[4] Robertson, S.E., Walker, S., & Beaulieu, M. Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), 2000, 95-108.

[5] Amati, G., & van Rijsbergen, C.J. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 2002, 357-389.

[6] Efthimiadis, E.N. Query expansion. *Annual Review of Information Science and Technology*, 31, 1996, p: 121-187

[7] Rocchio, J.J.Jr. Relevance feedback in information Retrieval. In G. Salton (Ed.): *The SMART Retrieval System*. Prentice-Hall Inc., Englewood Cliffs (NJ), 1971, 313-323.

[8] Vogt, C.C. & Cottrell, G.W. Fusion via a linear combination of scores. *IR Journal*, 1(3), 1999, 151-173.

[9] Fox, E.A. & Shaw, J.A. Combination of multiple searches. In *Proceedings TREC-2*, Gaithersburg (MA), NIST Publication #500-215, 1994, 243-249.

[10] Savoy, J. Combining Multiple Strategies for Effective Cross-Language Retrieval. *IR Journal*, 7(1-2), 2004, 121-148.

[11] Muller, C. *Principe et methodes de statistique lexicale.* Honoré Champion, Paris, 1992.

[12] Hosmer D., Leneshow S., *Applied Logistic Regression*. Wiley Interscience, New York, 2000.

UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches

Claire Fautsch, Ljiljana Dolamic, Jacques Savoy

Computer Science Department
University of Neuchatel, Switzerland
{Claire.Fautsch, Ljiljana.Dolamic, Jacques.Savoy}@unine.ch

Abstract

Our first objective in participating in this domain-specific evaluation campaign is to propose and evaluate various indexing and search strategies for the German, English and Russian languages, in an effort to obtain better retrieval effectiveness than that of the language-independent approach (*n*-gram). To do so we evaluate the GIRT-4 test-collection using the Okapi, various IR models derived from the *Divergence from Randomness* (DFR) paradigm, the statistical language model (LM) together with the classical *tf·idf* vector-processing scheme.

Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing**]: Indexing methods, Linguistic processing. I.2.7 [**Natural Language Processing**]: Language models. H.3.3 [**Information Storage and Retrieval**]: Retrieval models. H.3.4 [**Systems and Software**]: Performance evaluation.

General Terms

Experimentation, Performance, Measurement, Algorithms.

Additional Keywords and Phrases

Natural Language Processing with European Languages, Digital Libraries, German Language, Russian Language; Manual Indexing, Thesaurus.

1 Introduction

Domain-specific retrieval is an interesting task, one in which we access bibliographic notices (usually composed of a title and an abstract) extracted from two German social science sources and one Russian corpus. The records in these notices also contain manually assigned keywords extracted from a controlled vocabulary by librarians who are knowledgeable of the discipline to which the indexed articles belong. These descriptors should be helpful in improving document surrogates and consequently the extraction of more pertinent information, while also discarding irrelevant abstracts. Access to the underlying thesaurus would also improve retrieval performance.

The rest of this paper is organized as follows: Section 2 describes the main characteristics of the GIRT-4 (written in the German and English languages) and ISISS (Russian) test-collections. Section 3 outlines the main aspects of our stopword lists and light stemming procedures, along with the IR models used in our experiments. Section 4 explains different blind query expansion approaches and evaluates their use with the available corpora. Section 5 provides our official runs and results.

2 Overview of Test-Collections

In the domain-specific retrieval task, the two available corpora are composed of bibliographic records extracted from various sources in the social sciences domain. Typical records (see Figure 1 for a German example) in this corpus consist of a title (tag <TITLE-DE>), author name (tag <AUTHOR>), document language (tag <LANGUAGE-CODE>), publication date (tag <PUBLICATION-YEAR>) and abstract (tag <ABSTRACT-DE>). Manually assigned descriptors and classifiers are provided for all documents. An inspection of this German corpus reveals that all bibliographic notices consist of a title and 96.4% of them include an abstract. In addition to this information provided by the author, a typical record contains on average 10.15 descriptors

(“<CONTROLLED-TERM-DE>”), 2.02 classification terms (“<CLASSIFICATION-TEXT-DE>”), and 2.42 methodological terms (“<METHOD-TEXT-DE>“ or “<METHOD-TERM-DE>“). The manually assigned descriptors are extracted from the controlled list known as the “Thesaurus for the Social Sciences”. Finally, associated with each record is a unique identifier (“<DOCNO>”). Kluck (2004) provides a more complete description of this corpus.

<DOC>
<DOCNO> GIRT-DE19909343
<TITLE-DE> Die sozioökonomische Transformation einer Region : Das Bergische Land von 1930 bis 1960
<AUTHOR> Henne, Franz J.
<AUTHOR> Geyer, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> DE
<CONTROLLED-TERM-DE> Rheinland
<CONTROLLED-TERM-DE> historische Entwicklung
<CONTROLLED-TERM-DE> regionale Entwicklung
<CONTROLLED-TERM-DE> sozioökonomische Faktoren
<METHOD-TERM-DE> historisch
<METHOD-TERM-DE> Aktenanalyse
<CLASSIFICATION-TEXT-DE> Sozialgeschichte
<ABSTRACT-DE> Die Arbeit hat das Ziel, anhand einer regionalen Studie die Entstehung des "modernen" fordistischen Wirtschaftssystems und des sozialen Systems im Zeitraum zwischen 1930 und 1960 zu beleuchten; dabei geht es auch um das Studium des "Sozial-imaginären", der Veränderung von Bewußtsein und Selbst-Verständnis von Arbeitern durch das Erlebnis und die Erfahrung der Depression, des Nationalsozialismus und der Nachkriegszeit, welches sich in den 1950er Jahren gemeinsam mit der wirtschaftlichen Veränderung zu einem neuen "System" zusammenfügt.
<DOC> …

Figure 1: Example of record written in German

<DOC>
<DOCNO> GIRT-EN19901932
<TITLE-EN> The Socio-Economic Transformation of a Region : the Bergische Land from 1930 to 1960
<AUTHOR> Henne, Franz J.
<AUTHOR> Geyer, Michael
<PUBLICATION-YEAR> 1990
<LANGUAGE-CODE> EN
<CONTROLLED-TERM-EN> Rhenish Prussia
<CONTROLLED-TERM-EN> historical development
<CONTROLLED-TERM-EN> regional development
<CONTROLLED-TERM-EN> socioeconomic factors
<METHOD-TERM-EN> historical
<METHOD-TERM-EN> document analysis
<CLASSIFICATION-TEXT-EN> Social History
<DOC> …

Figure 2: English translation of the record shown in Figure 1

<DOC>
<DOCNO> ISISS-RAS-ECOSOC-20060324-41210
<AUTHOR-RU> Мартынова, М.Ю.
<TITLE-RU> Нормы и правила межличностного общения в культуре народов России
<KEYWORDS-RU> Россия; межличностные отношения; межкультурные отношения; коммуникация
<DOC> …

Figure 3: Example of a record extracted from the ISISS corpus

The above-mentioned German collection was translated into British English, mainly by professional translators whose native language was English. Included in all English records is a translated title (listed under "<TITLE-EN>" in Figure 2), manually assigned descriptors ("<CONTROLLED-TERM-EN>"), classification terms ("<CLASSIFICATION-TEXT-EN>") and methodological terms ("<METHOD-TERM-EN>"). Abstracts however were not always translated (in fact they are available for only around 15% of the English records).

In addition to this bilingual corpus, we may also access the GIRT thesaurus, containing 10,623 entries (all including both the <GERMAN> and <GERMAN-CAPS>) tags together with 9,705 English translations. It also contains 2,947 <BROADER-TERM> relationships and 2,853 <NARROWER-TERM> links. The synonym relationship between terms is expressed through <USE-INSTEAD> (2,153) links, <RELATED-TERM> (1,528) or <USE-COMBINATION> (3,263).

As a third language, we access bibliographic records written in the Russian language composed of the ISISS (Russian Economic and Social Science) bibliographic data collection (see Figure 3 for an example of a record extracted from the Russian collection). Using a pattern similar to that of the other two corpora, records include a title ("<TITLE-RU>" in Figure 3), sometimes an abstract ("<ABSTRACT-RU>"), and certain manually assigned descriptors ("<KEYWORDS-RU>").

Table 1 below lists a few statistics from these collections, showing that the German corpus has the largest size (326 MB), the English ranks second and the Russian third, both in size (81 MB) and in number of documents (145,802). The German corpus has the larger mean size (89.71 indexing terms/article), compared to the English collection (54.86), while for the Russian corpus the mean value is clearly smaller (18.77). The English corpus includes also the *CSA Sociological Abstracts* (20,000 documents, 38.5 MB).

During the indexing process, we retained all pertinent sections in order to build document representatives. Additional information such as author name, publication date and the language in which the bibliographic notice was written are of less importance, particularly from an IR perspective, and thus they will be ignored in our experiments.

As shown in Appendix 2, the available topics cover various subjects (e.g., Topic #206: "Environmental justice," Topic #209: "Doping and sports," Topic #221: "Violence in schools," or Topic #211: "Shrinking cities"), and some of them may cover a relative large domain (e.g. Topic #212: "Labor market and migration").

	German	English	Russian
Size (in MB)	326 MB	235 MB	81 MB
# of documents	151,319	171,319	145,802
# of distinct terms	10,797,490	6,394,708	40,603
Number of distinct indexing terms per document			
Mean	71.36	37.32	14.89
Standard deviation	32.72	25.35	7.54
Median	68	28	13
Maximum	391	311	74
Minimum	2	2	1
Number of indexing terms per document			
Mean	89.71	54.86	18.77
Standard deviation	44.5	42.41	9.32
Median	85	39	17
Maximum	629	534	98
Minimum	4	4	2
Number of queries	25	25	24
Number rel. items	2290	2133	292
Mean rel./ request	91.6	85.32	12.17
Standard deviation	90.85	59.95	17.45
Median	72	89	5
Maximum	431 (T #218)	206 (T #201)	73 (T #204)
Minimum	7 (T #204)	4 (T #218)	1 (T #215)

Table 1: CLEF GIRT-4 and ISISS test collection statistics

3 IR Models and Evaluation

3.1 Indexing and IR Models

For the English, German and Russian language, we used the same stopword lists and stemmers that we selected for our previous CLEF participation (Fautsch *et al.*, 2008). Thus for English it was the SMART stemmer and stopword list (containing 571 items), while for the German we apply our light stemmer (available at http://www.unine.ch/info/clef/) and stopword list (603 words). For all our German experiments we also apply our decompounding algorithm (Savoy, 2004). For the Russian language, the stopword list contains 430 words and we apply our light stemming procedure (based on 53 rules to remove the final suffix representing gender (masculine, feminine, and neutral), number (singular, plural) and the six Russian grammatical cases (nominative, accusative, genitive, dative, instrumental, and locative)).

In order to obtain a broader view of the relative merit of various retrieval models, we may first adopt the classical *tf idf* indexing scheme. In this case, the weight attached to each indexing term in a document surrogate (or in a query) combines the term's occurrence frequency (denoted tf_{ij} for indexing term t_j in document D_i) and also the inverse document frequency (denoted idf_j).

In addition to this vector-processing model, we may also consider probabilistic models such as the Okapi model (or BM25) (Robertson *et al.*, 2000). As a second probabilistic approach, we may implement four variants of the DFR (*Divergence from Randomness*) family suggested by Amati & van Rijsbergen (2002). In this framework, the indexing weight w_{ij} attached to term t_j in document D_i combines two information measures as follows.

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}(tf)] \cdot (1 - Prob^2_{ij}(tf))$$

The first model PB2 is based on the following equations:

$$Prob^1_{ij} = (e^{-\lambda_j} \cdot \lambda^{tfij}) / tf_{ij}! \qquad \text{with } \lambda_j = tc_j / n \tag{1}$$

$$Prob^2_{ij} = 1 - [(tc_j+1) / (df_j \cdot (tfn_{ij}+1))] \qquad \text{with } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl) / l_i) \tag{2}$$

where tc_j represents the number of occurrences of term t_j in the collection, df_j the number of documents in which the term t_j appears, and n the number of documents in the corpus. Moreover, c and *mean dl* (average document length) are constants whose values are given in the Appendix 1.

The second model GL2 is defined as:

$$Prob^1_{ij} = [1 / (1+\lambda_j)] \cdot [\lambda_j / (1+\lambda_j)]^{tfn_{ij}} \tag{3}$$

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \tag{4}$$

For the third model I(n)B2, we still use Equation 2 to compute $Prob^2_{ij}$ but the implementation of Inf^1_{ij} is modified as:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (df_j+0.5)] \tag{5}$$

For the fourth model I(n_e)C2 the initial value of $Prob^2_{ij}$ is obtained from Equation 2 and for the value Inf^1_{ij} we use:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2[(n+1) / (n_e+0.5)] \quad \text{with } n_e = n \cdot [1 - [(n-1) / n]^{tcj}] \tag{6}$$

Finally, we also consider an approach based on a statistical language model (LM) (Hiemstra 2000; 2002), known as a non-parametric probabilistic model (both Okapi and DFR are viewed as parametric models). Thus, the probability estimates would not be based on any known distribution (as in Equations 1, or 3), but rather be estimated directly based on the occurrence frequencies in document D or corpus C. Within this language model (LM) paradigm, various implementations and smoothing methods might be considered, and in this study we adopt a model proposed by Hiemstra (2002) as described in Equation 7, which combines an estimate based on document ($P[t_j | D_i]$) and on corpus ($P[t_j | C]$) (Jelinek-Mercer smoothing method).

$$P[D_i | Q] = P[D_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot P[t_j | D_i] + (1-\lambda_j) \cdot P[t_j | C]]$$
$$\text{with } P[t_j | D_i] = tf_{ij}/l_i \text{ and } P[t_j | C] = df_j/lc \quad \text{with } lc = \sum_k df_k \tag{7}$$

where λ_j is a smoothing factor (constant for all indexing terms t_j, and usually fixed at 0.35) and *lc* an estimate of the size of the corpus C.

3.2 Overall Evaluation

To measure the retrieval performance, we adopted the mean average precision (MAP) (computed on the basis of 1,000 retrieved items per request by the new TREC-EVAL program). In the following tables, the best performances under the given conditions (with the same indexing scheme and the same collection) are listed in bold type.

Table 2 shows the MAP obtained by the seven probabilistic models and the classical *tf idf* vector-space model using the German or English collection and three different query formulations (title-only or T, TD, and TDN). In the bottom lines we reported the MAP average over the best 6 IR models (the average is computed without the *tf idf* scheme), and the percent change over the medium (TD) query formulation. The DFR I(n)B2 model for the German and also for the English corpus tend to produce the best retrieval performances.

	Mean average precision				
Query Model \ # of queries	German T 25 queries	German TD 25 queries	German TDN 25 queries	English T 25 queries	English TD 25 queries
DFR PB2	0.3877	0.4177	0.4192	0.2620	0.3101
DFR GL2	0.3793	0.4000	0.4031	0.2578	0.2910
DFR I(n)B2	**0.3940**	**0.4179**	0.4202	**0.2684**	**0.3215**
DFR $I(n_e)C2$	0.3935	0.4170	0.4121	0.2662	0.3191
LM (λ=0.35)	0.3791	0.4130	**0.4321**	0.2365	0.2883
Okapi	0.3815	0.4069	0.4164	0.2592	0.3039
tf idf	0.2212	0.2391	0.2467	0.1715	0.1959
Mean (top-6 best models)	0.3859	0.4121	0.4172	0.2584	0.3057
% change over TD queries	-6.37%		+1.24%	-15.48%	

Table 2: Mean average precision of various single searching strategies (monolingual, GIRT-4 corpus)

Table 3 lists the evaluations done for Russian (word-based indexing & *n*-gram indexing (McNamee & Mayfield, 2004)). The last three lines in this table indicate the MAP average computed for the 4 IR models, the percent change compared to the medium (TD) query formulation, and the percent change when comparing word-based and 4-gram indexing approaches.

From this table, we can see that when using word-based indexing, the DFR $I(n_e)B2$ or the LM models tend to perform the best. With the 4-gram indexing approach, the LM model always presents the best performing schemes. The short query formulation (T) tends to produce a better retrieval performance than medium (TD) topic formulation. As shown in the last line, when comparing the word-based and 4-gram indexing systems, the relative difference is seen to be rather short (around 4.6%) and favors the 4-gram approach.

Using our evaluation approach, evaluation differences occur when comparing with values computed according to the official measure (the latter always takes 25 queries into account).

	Mean average precision			
Query type Indexing / stemmer IR Model	Russian T word / light 24 queries	Russian TD word / light 24 queries	Russian T 4-gram 24 queries	Russian TD 4-gram 24 queries
DFR GL2	0.1515	0.1332	0.1617	0.1570
DFR $I(n_e)B2$	0.1470	**0.1468**	0.1402	0.1358
LM (λ=0.35)	**0.1528**	0.1337	**0.1688**	**0.1669**
Okapi	0.1418	0.1349	0.1499	0.1440
tf idf	0.1047	0.1089	0.1098	0.1132
Mean	0.1484	0.1372	0.1552	0.1509
% change over T	baseline	-7.5%	baseline	-2.72%
over stemming	baseline	baseline	+4.64%	+10.04%

Table 3: Mean average precision of various single search strategies (monolingual, ISISS corpus)

4 Blind-Query Expansion

To provide a better match between user information needs and documents, various query expansion techniques have been suggested. The general principle is to expand the query using words or phrases having similar meanings to, or related to those appearing in the original request. To achieve this, query expansion approaches consider various relationships between these words, along with term selection mechanisms and term weighting schemes. Specific answers regarding the best technique may vary, thus leading to a variety of query expansion approaches (Efthimiadis, 1996).

In our first attempt to find related search terms, we might ask the user to select additional terms to be included in an expanded query. This could be handled interactively through displaying a ranked list of retrieved items returned by the first query. As a second strategy, Rocchio (1971) proposed taking the relevance or non-relevance of top-ranked documents into account, as indicated manually by the user. In this case, a new query would then be built automatically in the form of a linear combination of the term included in the previous query and terms automatically extracted from both relevant (with a positive weight) and non-relevant documents (with a negative weight). Empirical studies have demonstrated that such an approach is usually quite effective.

As a third technique, Buckley *et al*. (1996) suggested that even without looking at them or asking the user, it could be assumed that the top-*k* ranked documents would be relevant. This method, denoted as the pseudo-relevance feedback or blind-query expansion approach does not require user intervention. Moreover, using the MAP as performance measure is a strategy that usually tends to enhance performance measures.

In the current context, we used Rocchio's formulation (denoted "Rocchio") with $\alpha = 0.75$, $\beta = 0.75$, whereby the system was allowed to add *m* terms extracted from the *k* best ranked documents from the original query. For the German corpus (Table 4, third column), such a search technique does not seem to enhance the MAP. For the English collection (Table 5, second and third column), Rocchio's blind query expansion may improve the MAP from +9.3% (DFR PB2, 0.3101 vs. 0.3392) or hurt the retrieval performance -8.72% (Okapi model, 0.3039 vs. 0.2774). For the Russian language (Table 6, second and forth column), blind query expansion improves the MAP (e.g., +28.98% with the Okapi model, 0.1740 vs. 0.1349 or +2.3% with the DFR I(n_e)B2 model, 0.1503 vs. 0.1468).

	Mean average precision			
Query TD PRF model IR Model / MAP	German idf PB2 **0.4177**	German Rocchio DFR I(n)B2 **0.4179**	German idf DFR I(n)B2 **0.4179**	German idf LM **0.4130**
k doc. / *m* terms	5/70 0.4149 10/100 0.4068 10/200 0.4078	5/70 0.3965 10/100 0.3965 10/200 0.3992	5/70 0.4120 10/100 0.4025 10/200 0.4104	5/70 0.3818 10/100 0.3879 10/200 0.3941

Table 4: Mean average precision using blind-query expansion (German GIRT-4 collection)

	Mean average precision			
Query TD PRF model IR Model / MAP	English Rocchio Okapi **0.3039**	English Rocchio DFR PB2 0.3101	English idf DFR PB2 **0.3101**	English idf LM **0.2883**
k doc. / *m* terms	10/50 0.2774 10/100 0.2776 10/200 0.276**7**	10/50 **0.3392** 10/100 0.3366 10/200 0.3324	10/50 0.3023 10/100 0.3032 10/200 0.3006	10/50 0.2672 10/100 0.2725 10/200 0.2746

Table 5: Mean average precision using blind-query expansion (English GIRT-4 collection)

	Mean average precision			
Query TD PRF model IR Model / MAP	Russian Rocchio Okapi 0.1349	Russian idf Okapi 0.1349	Russian Rocchio DFR I(n_e)B2 0.1468	Russian idf DFR I(n_e)B2 **0.1468**
k doc. / *m* terms	3/50 0.1737 5/70 **0.1740** 10/100 0.1733	3/50 **0.1612** 5/70 0.1245 10/100 0.1251	3/50 0.1457 5/70 0.1284 10/100 **0.1503**	3/50 0.1433 5/70 0.1366 10/100 0.1391

Table 6: Mean average precision using blind-query expansion (Russian, ISISS corpus)

Rocchio's query expansion approach however does not always significantly improve the MAP. Such a query expansion approach is based on term co-occurrence data and tends to include additional terms that occur very frequently in the documents. In such cases, these additional search terms will not always be effective in discriminating between relevant and non-relevant documents, and the final effect on retrieval performance could be negative.

As another pseudo-relevance feedback technique we may apply an *idf*-based approach (denoted "idf" in following tables) (Abdou & Savoy, 2008). In this query expansion scheme, the inclusion of new search terms is based on their *idf* values, tending to enlarge the query with more infrequent terms. Overall this *idf*-based term selection performs rather well and usually its retrieval performance is more robust.

For example, with the Russian language (Table 6, third and fifth column), this idf-based blind query expansion may also improve the MAP (e.g., +19.5% with the Okapi model, 0.1612) but, on the other hand, with the DFR I(n_e)B2 model, the MAP is slightly reduced (-2.3% from 0.1468 to 0.1433).

However, the *idf*-based query expansion tends to include rare terms, without considering the context. Thus among the top-*k* retrieved documents such a scheme may add terms appearing far away from where the search terms occurred. The single selection criterion is based only on *idf* values, not the position of those additional terms in the top-ranked documents. This year we investigated retrieval effectiveness when including a second criterion in the selection of terms to be included in the new expanded query. We considered it to be important to expand the query using terms appearing close to a search term (fixed at 10 indexing terms in the current experiments). This short window includes 10 terms to the right and 10 terms to the left of each query term. This type of query expansion method is denoted as "idf-window" in Table 7.

Finally, to find words or expressions related to the current request, we considered using commercial search engines (e.g., Google) or online encyclopedia (e.g., Wikipedia). In this case, we submitted a query containing the short topic formulation (T or title-only) to each information service. When using Google, we fetched the first two text snippets and added them as additional terms to the original topic formulation, forming a new expanded query. When using Wikipedia, we fetched the first returned article and added the ten most frequent terms (*tf*) contained in the extracted article.

	Mean average precision			
Query TD PRF model IR Model / MAP	German Rocchio Okapi **0.4069**	German idf Okapi **0.4069**	German idf + window Okapi 0.4069	German with Google Okapi 0.4096
k doc. / *m* terms	5/50 0.3801 10/50 0.3783 10/200 0.3822	5/50 0.3726 10/50 0.3696 10/200 0.3868	5/50 0.4110 10/50 0.4146 10/200 **0.4247**	**0.4196**

Table 7: Mean average precision using four blind-query expansions (German GIRT-4 collection)

The retrieval effectiveness of our two new query expansion approaches is depicted in Table 7 (German collection) and is compared to two other query expansion techniques. Compared to the performance before query expansion (0.4096), Rocchio's and the idf-based blind query expansion cannot improve the MAP. On the other hand, the variant "idf-window" presents a better retrieval performance (+4.9%, from 0.4069 to 0.4247). Using the first two text snippets returned by Google, we may also enhance slightly the MAP (from 0.4096 to 0.4196, or +2.4%). The MAP variation varied according to approaches and parameter settings, while the largest enhancement could be found using the idf+window technique (forth column in Table 7). Finally, using Google to find related terms or phrases implied that we required more processing time.

5 Official Results

Table 8 describes our 9 official runs in the monolingual GIRT task. In this case each run was built using a data fusion operator "Z-Score" (see (Savoy & Berger, 2005)). For all runs, we automatically expanded the queries using the blind relevance feedback method of Rocchio (denoted "Roc"), our IDFQE approach (denoted "idf"), or our new window-based approach (denoted "idf-win"). Finally Table 8 depicts the MAP obtained for the Russian collection when considering 24 queries and in parenthesis, the official MAP computed for 25 queries.

As a complementary search technique, we used two stemmers when defining the official run UniNEDSde3. In this case we first applied our light stemming approach and then a more aggressive one. If the same term was produced by the two stemmers, we only kept one occurrence. On the other hand, if the returned stem differed, we added the two forms to the query formulation.

Run name	Language	Query	Index	Model	Query expansion	MAP	Comb.MAP
UniNEDSde1	German	TD	dec	I(n)B2	Roc 10 docs / 200 terms	0.3992	Z-score
		TD	dec	LM	Google	0.4265	**0.4537**
		TD	dec	PB2	idf-win 10 docs / 150 terms	0.4226	
UniNEDSde2	German	TD	dec	PB2	idf 5 docs / 200 terms	0.4151	Z-score
		TD	dec	I(n)B2		0.4179	0.4399
		TD	dec	I(n)B2	idf-win 10 docs / 200 terms	0.4248	
UniNEDSde3	German	T	dec	I(n)B2		0.3940	Z-score
	special	TD	dec	I(n)B2	idf-win 10 docs / 200 terms	0.4319	0.4251
		TD	dec	I(n_e)C2		0.4170	
UniNEDSde4	German	TD	dec	Okapi	idf-win 5 docs / 50 terms	0.4110	Z-score
		TD	dec	IneC2		0.4170	0.4343
		TD	dec	PB2	idf 10 docs / 200 terms	0.4078	
UniNEDSen1	English	TD	N-stem	InB2	Roc 10 docs / 100 terms	0.3140	Z-score
		TD	N-stem	InB2		0.3562	**0.3770**
		TD	N-stem	LM	Roc 5 docs / 150 terms	0.3677	
UniNEru1	Russian	TD	word/light	I(n_e)B2	Roc 3 docs / 50 terms	0.1457	Z-score
		TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	0.1594
							(0.1531)
UniNEru2	Russian	TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	Z-score
		TD	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1284	0.1628
		TD	word/light	Okapi	Roc 3 docs / 50 terms	0.1737	(0.1563)
UniNEru3	Russian	TD	4-gram	I(n_e)B2	Roc 5 docs / 150 terms	0.1164	Z-score
		TD	word/light	I(n_e)B2	idf 5 docs / 70 terms	0.1366	0.1655
		TD	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1284	(0.1589)
UniNEru4	Russian	TD	4-gram	I(n_e)B2	Roc 3 docs / 150 terms	0.1129	Z-score
		TD	word/light	I(n_e)B2	Roc 5 docs / 70 terms	0.1652	**0.1890**
		TD	word/light	I(n_e)B2	idf 3 docs / 70 terms	0.1739	(**0.1815**)

Table 8: Description and mean average precision (MAP) of our official GIRT runs

5 Conclusion

For our participation in this domain-specific evaluation campaign, we evaluated different probabilistic models using the German, English and Russian languages. For the German and Russian languages we applied our light stemming approach and stopword list. The resulting MAP (see Tables 2 and 3) show that the DFR I(n)B2 or the LM model usually provided in the best retrieval effectiveness. The performance differences between Okapi and the various DFR models were usually rather small.

In our analysis of the blind query expansion approaches (see Tables 4 to 6), we find that this type of automatic query expansion we used can sometimes enhance the MAP. Depending on the collection or languages however, this approach will not provide the same degree of improvement or can sometimes hurt the retrieval effectiveness. For example this search strategy results in less improvement for the English corpus than it does for the Russian collection. For the German collection however, this search strategy clearly hurt the MAP.

This year we suggest two new query expansion techniques. The first, denoted "idf-window", is based on co-occurrence of relatively rare terms in a close context (within 10 terms from the occurrence of a search term in a retrieved document). As a second approach, we add the first two text snippets found by Google to expand the query. Compared to the performance before query expansion (e.g., with Okapi the MAP is 0.4096), Rocchio's and the idf-based blind query expansion cannot improve this retrieval performance. On the other hand, the variant “idf-window” presents a better retrieval performance (+4.9%, from 0.4069 to 0.4247). Using the first two text snippets returned by Google, we may also enhance slightly the MAP (from 0.4096 to 0.4196, or +2.4%).

Acknowledgments

The authors would like to also thank the GIRT - CLEF-2008 task organizers for their efforts in developing domain-specific test-collections. This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

References

Abdou, S., & Savoy, J. (2008). Searching in Medline: Stemming, query expansion, and manual indexing evaluation. *Information Processing & Management*, 44(2), p. 781-789.

Amati, G. & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems*, 20(4), p. 357-389.

Buckley, C., Singhal, A., Mitra, M. & Salton, G. (1996). New retrieval approaches using SMART. In *Proceedings of TREC-4*, Gaithersburg: NIST Publication #500-236, p. 25-48.

Efthimiadis, E.N. (1996). Query expansion. *Annual Review of Information Science and Technology*, 31, p. 121-187.

Fautsch, C., Dolamic, L., Savoy, J., (2008). Domain-Specific IR for German, English and Russian Languages. In C. Peters, P. Clough, F.C. Gey, J. Karlgen, B. Magini, D.W. Oard, M. de Rijke & M. Stempfhuber (Eds.), *8th Workshop of the Cross-Language Evaluation Forum*. LNCS #5152, Springer-Verlag, Berlin, p. 196-199.

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. In *Proceedings of the ACM-SIGIR*, The ACM Press, Tempere, p. 35-41.

Kluck, M. (2004). The GIRT data in the evaluation of CLIR systems - from 1997 until 2003. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237. Springer-Verlag, Berlin, 2004, p. 376-390.

McNamee, P. & Mayfield, J. (2004). Character *n*-gram tokenization for European language text retrieval. *IR Journal*, 7(1-2), p. 73-97.

Robertson, S.E., Walker, S. & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing & Management*, 36(1), p. 95-108.

Rocchio, J.J.Jr. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.): *The SMART Retrieval System*. Prentice-Hall Inc., Englewood Cliffs (NJ), p. 313-323.

Savoy, J. (2004). Report on CLEF-2003 monolingual tracks: Fusion of probabilistic models for effective monolingual retrieval. In C. Peters, J. Gonzalo, M. Braschler, M. Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems*. LNCS #3237. Springer-Verlag, Berlin, 2004, p. 322-336.

Savoy, J., & Berger, P.-Y. (2005): Selection and merging strategies for multilingual information retrieval. In: Peters, C., Clough, P., Gonzalo, J., Jones, G.J.F., Kluck, M., Magnini, B. (Eds.): Multilingual Information Access for text, Speech and Images. Lecture Notes in Computer Science: Vol. 3491. Springer, Heidelberg, p. 27-37.

Appendix 1: Parameter Settings

	Okapi			DFR	
Language	*b*	k_1	*avdl*	*c*	*mean dl*
German GIRT	0.55	1.2	200	1.5	200
English GIRT	0.55	1.2	53	4.5	53
Russian word	0.55	1.2	19	1.5	19
Russian 4-gram	0.55	1.2	113	1.5	113

Table A.1: Parameter settings for the various test-collections

Appendix 2: Topic Titles

C201	Health risks at work	C213	Migrant organizations
C202	Political culture and European integration	C214	Violence in old age
C203	Democratic transformation in Eastern Europe	C215	Tobacco advertising
C204	Child and youth welfare in the Russian Federation	C216	Islamist parallel societies in Western Europe
		C217	Poverty and social exclusion
C205	Minority policy in the Baltic states	C218	Generational differences on the Internet
C206	Environmental justice	C219	(Intellectually) Gifted
C207	Economic growth and environmental destruction	C220	Healthcare for prostitutes
C208	Leisure time mobility	C221	Violence in schools
C209	Doping and sports	C222	Commuting and labor mobility
C210	Establishment of new businesses after the reunification	C223	Media in the preschool age
C211	Shrinking cities	C224	Employment service
C212	Labor market and migration	C225	Chronic illnesses

Table A.2: Query titles for CLEF-2008 GIRT test-collections

Comparison Between Manually and Automatically Assigned Descriptors Based on a German Bibliographic Collection

Claire Fautsch, Jacques Savoy
Computer Science Department
University of Neuchatel
2009 Neuchâtel, Switzerland
{*Claire.Fautsch,Jacques.Savoy*}*@unine.ch*

***Abstract*—This paper compares and illustrates the use of manually and automatically assigned descriptors on German documents extracted from the GIRT Corpus. A second objective is to analyze the usefulness of both specialized or general thesauri to automatically enhance queries.**
To illustrate our results we use different search models such as a vector space model, a language model and two probabilistic models. We also proposed different measures to compute textual entailment between two terms allowing us to hopefully select appropriate keywords from thesauri to expand documents or queries automatically.

I. Introduction

During the last years electronic bibliographic tools gained more and more importance, partly due to the fact that electronic copies of printed media are made available on a large scale. For scientific journals, the growing printing cost especially when colors are required tends to favor electronic versions. Furthermore the distribution of electronic copies is nowadays much easier and faster than of printed media.

The information has not only to be made available, but the user must also be able to search the records easily and find pertinent information in an user-friendly way. For scientific papers, often only title and abstract are freely available in the bibliographic records database. This is mainly due to copyright issues. Hence these scientific documents often contain manually assigned keywords added to increase the matching possibilities between authors and information searchers. These keywords usually extracted from a controlled vocabulary can either be added during indexing by a person having a good knowledge in the given domain and/or by the author. An example for such an online bibliographic records database is ERIC[1], providing access to scientific literature for the educational world.

In this paper, we want to see whether manually added keywords can enhance retrieval. Moreover, we want to verify whether automatically added keywords might yield an improvement. Since domain-specific thesauri are not always available, we also use a general thesaurus for query and document expansion. We may thus see the differences between specific and general thesauri for the German language. In a second part we are interested in the impact of enhancing queries rather than documents. This is especially interesting if the searcher does not have a strong knowledge in the domain of interest and does not use domain specific terms in his/her query formulation. Expanding queries using a domain specific thesaurus might fill this gap between general and specific language more appropriate than a general thesaurus (e.g., WordNet [1]).

The rest of this paper is organized as follows. Section II presents related works, while in Section III we describe the test-collection and the thesauri used. Section IV gives a short overview of the different information retrieval (IR) models used for our evaluations and Section V explains the different lexical entailment measures used for term selection. Section VI shows the results of our different test runs. Finally in Section VII we summarize the main findings of this paper.

II. Related Work

For the various manual-indexing strategies used in the IR domain, their retrieval impact was studied and evaluated during the well-known Cranfield experiments. For example in the context of the the Cranfield II test (1,400 documents, 221 queries), Cleverdon [2] reported that single-word indexing was more effective than using terms extracted from a controlled vocabulary, where both indexing schemes were done by human beings.

Rajashekar & Croft [3] used the INSPEC test collection (12,684 documents, 84 queries) to evaluate retrieval effectiveness of various document representations. This study showed that automatic indexing based on article titles and abstracts performed better than any other single indexing schemes. While the controlled vocabulary terms by themselves were not effective representations, their presence as an additional source of evidence on document content could improve retrieval performance. Based on a corpus containing French bibliographic notices, in [4] we demonstrated that including manually assigned descriptors for title-only queries might significantly enhance MAP, compared to an approach that ignores them.

[1]Education Resources Information Center, http://www.eric.ed.gov/

In order to obtain better retrieval performance with the GIRT corpus, Petras [5] suggested adding manually assigned subject keywords in order to help make fuzzy topic descriptions less ambiguous. She later also showed that combining pseudo-relevance feedback and thesaurus-based query expansions could also improve retrieval performance.

Descriptors assigned manually represent significant cost increases for information providers and their utility must be analyzed and evaluated. In this perspective, we are concerned with the following question: Do such descriptors statistically improve the information retrieval process? The rest of this paper will try to provide answers to this question.

III. Test-Collection

The test collection we used for our different experiments is composed of the German GIRT corpora, 125 queries and two thesauri, a domain-specific thesaurus and a general thesaurus, described in the following sections.

A. GIRT Corpus

The GIRT (German Indexing and Retrieval Test database) corpus was made available through the CLEF[2] evaluation campaign. Over the years, the corpus has been enlarged to contain more than 150,000 documents, and an English translation is also available. More information about the GIRT corpora can be found in Kluck [6].

A typical record of the GIRT corpus consists of author name, title, document language, publication year and abstract and may as well contain manually added keyword terms. The document parts relevant for our experiments can be separated into two categories, on the one hand we have the title and abstract and on the other manually added keywords. The remaining fields (such as publication year) are not considered important for our experiments and will thus be ignored.

B. Topics

For our test runs, we used the queries deployed in the domain-specific track in the CLEF campaigns from 2004 to 2008. This gives us a total of 125 queries (i.e. 25 per year). Each topic is structured into three logical sections. The first part of a topic is a short title (T) followed by a brief description (D) of what the user is looking for, generally consisting of one short sentence. While these two sections represent the real user's needs, the last part (N) is a longer characterization of the user's needs indicating relevance assessment criteria. All topics have been judged on the same GIRT corpus.

[2]Cross Language Evaluation Forum, http://www.clef-campaign.org/

C. Thesauri

One of our objectives in this paper is to analyze the improvements in retrieval if additional keywords are added either manually or automatically to the documents or the queries. As a second objective we want to see if automatically added keywords, extracted from a thesaurus, add the same benefit to documents as those manually added.

1) Domain Specific Thesaurus: For the domain specific track in CLEF, a machine readable version of the German-English thesaurus for social science [7] was made available. The manually added controlled vocabulary terms were extracted from this thesaurus. We use this thesaurus as a domain specific thesaurus for automatically expanding documents or queries with keywords. The machine readable version is formatted in XML and contains 10,624 entries. Each entry represents a German descriptor, given with narrower and/or broader terms as well as with related terms. Other attributes that might also be given for a descriptor are *use-instead*, *use-combination* and *scope note*.

2) General Thesaurus: As a second, general thesaurus we use OpenThesaurus, freely available from http://www.openthesaurus.de/[3]. This thesaurus contains 17,619 entries, but on the contrary to the social science thesaurus each entry is just a set of words with a similar meaning. As the name implies, this thesaurus is "open" and regularly enlarged from different users through a collaborative effort. More information can be found in [8].

IV. IR Models

For indexing the documents and queries, we first normalize each indexing unit by transforming it to lowercase letters and removing diacritics (e.g., "Überraschung" would be normalized to "uberraschung"). We then apply our light stemmer[4], a decompounding algorithm for the German language [9] and remove words occurring in a stopword list (603 words, e.g., "der", "in", "ist").

To give a solid base to our empirical studies, we used different models to retrieve relevant information. As a baseline approach, we use a standard *tf idf* weighting scheme with a cosine normalization. As a second approach we used the Okapi (BM25) model proposed by Robertson *et al.* [10], evaluating the document D_i score for the query Q by applying the following formula:

$$Score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot \log(\frac{n - df_j}{df_j}) \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}} \quad (1)$$

with $K = k_1 \cdot [(1 - b) + b \cdot \frac{l_i}{avdl}]$ where qtf_j denotes the frequency of term t_j in the query Q, n the number of documents in the collection, df_j the number of documents in which the term t_j appears and l_i the length of the document

[3]We use the image from November 19th 2008, 00:47

[4]Freely available at http://www.unine.ch/info/clef/

D_i. The constant b was set to 0.55 and k_1 to 1.2. $avdl$ represents the average document length.

As a third model we used $InB2$ derived from the *Divergence of Randomness* paradigm [11]. In the *Divergence of Randomness* framework two information measures are combined to obtain the weight w_{ij} of the term t_j in the document D_i. We then obtain following formula for the document score:

$$Score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot w_{ij} \tag{2}$$

where

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2(Prob^1_{ij}(tf_{ij})) \cdot (1 - Prob^2_{ij}(tf_{ij}))$$

For $InB2$, the two information measures are defined as follows:

$$Inf^1_{ij} = tfn_{ij} \cdot \log_2\left((n+1)/(df_j + 0.5)\right) \tag{3}$$

$$Prob^2_{ij} = 1 - [(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))] \tag{4}$$

with $tfn_{ij} = tf_{ij} \cdot \log_2\left(1 + ((c \cdot mean\_dl)/l_i)\right)$ where tc_j represents the number of occurrences of the term t_j in the collection. Moreover, c is a constant, fixed at 1.5 for our test cases and $mean\_dl$ is the mean document length.

To complete our models, we use a language model. Contrary to the Okapi and $InB2$ model, the language model approach is a non-parametric probabilistic model. We adopt a model proposed by Hiemstra [12] and described in Equation 5

$$P(D_i|Q) = P(D_i) \prod_{t_j \in Q} (\lambda_j \cdot P(t_j|D_i) + (1 - \lambda_j) \cdot P(t_j|C)) \tag{5}$$

with $P(t_j|D_i) = tf_{ij}/l_i$ and $P(t_j|C) = df_j/lc$ with $lc = \sum_k df_k$, where λ_j is a smoothing factor fixed at 0.35 for our experiments, and lc an estimate of the size of the corpus C.

V. Textual Entailment and Similarity Measures

In natural language processing, different measures are used to calculate textual entailment between two terms. We retain three measures.

As a first and simple measure we use the Jaccard similarity. Let u and v be the two terms for which we want to calculate similarity, and U and V the set of documents where they occur. We denote by $|U|$ (resp $|V|$) the cardinal of these sets. The Jaccard similarity between u and v is defined by following equation:

$$Jaccard(u, v) = \frac{|U \cap V|}{|U \cup V|} \tag{6}$$

The advantage of this similarity measure is that it is easy to calculate. As drawback it is known that this measure does not take into account the frequencies of the terms u and v in a document or in the collection. Under this consideration we use two other measures to compute lexical entailment. The first is a simple probability, defined as

$$P(v|u) = \sum_{d \in D} P(v|d)P(d|u) \tag{7}$$

where D is the set of documents in the collection and $P(v|u)$ is the probability of finding v in a document knowing this document contains u. $P(d|u)$ cannot be calculated easily, but we can assume that $P(d)$ is uniform (constant) and that if $u \notin d$, $P(d|u) = 0$. We can also assume that the length of d does not play any role. With these assumptions Equation 7 can be rewritten as

$$P(v|u) \propto \sum_{d \in D : u \in d} P(v|d)P(u|d) \tag{8}$$

As third and last measure we will use an average mutual information (MI) between two terms, defined as follows

$$I(u, v) = \sum_{X \in \{u, \widetilde{u}\}} \sum_{Y \in \{v, \widetilde{v}\}} P(X, Y) \log_2 \frac{P(X, Y)}{P(X)P(Y)} \tag{9}$$

where $\widetilde{u}$ (respectively $\widetilde{v}$) stands for the absence of u (respectively v). We note that if u and v are independent, $I(u, v) = 0$.

VI. Results

First we want to analyze the impact of manually or automatically assigned descriptors and see the difference in efficiency of human selected keywords versus automatically selected keywords. In a second step we automatically expand queries using a thesaurus with the intention of improving retrieval effectiveness. We used the four IR models described in Section IV and to measure the retrieval performance, we used MAP (Mean Average Precision) values computed on the basis of 1000 retrieved documents per query using TREC_EVAL[5].

A. Manually Indexing Evaluation

As a baseline we first evaluate a simple run searching only in the title and abstract part of the documents and using short (title only, T) query formulations (MAP depicted in second column of Table I).

To analyze the effect of manually added keywords, we then perform retrieval over the complete document, i.e searching for relevant information not only in the title and abstract part, but also in the keywords. In the third column of Table I (label "+Manual") we depicted the MAP when searching in title, abstract and keywords. The last column shows the performance difference before and after considering manually assigned descriptors.

This table shows that the inclusion of manually added keywords from a controlled vocabulary considerably improves retrieval results. This is a first indication showing

[5]http://trec.nist.gov/trec_eval/

	MAP		
Model	Title & Abstract	+Manual	%Change
tf idf	0.1929	0.2275	17.94
LM2	0.2865	0.3215	12.22
InB2	**0.3157**	0.3493	10.64
Okapi	0.3042	**0.3494**	14.86

Table I
MAP WITH AND WITHOUT MANUALLY ASSIGNED DESCRIPTORS FOR SHORT QUERIES (T-ONLY)

	MAP		
Model	Title & Abstract	+Automatic	%Change
tf idf	0.1929	0.1404	-27.22
LM2	0.2865	0.1992	-30.47
InB2	**0.3157**	**0.2496**	-20.94
Okapi	0.3042	0.2151	-29.29

Table II
DOCUMENT EXPANSION USING GIRT-THESAURUS

	MAP		
Model	Title & Abstract	+Automatic	%Change
tf idf	0.1929	0.1874	-2.85
LM2	0.2865	0.238	-16.93
InB2	**0.3157**	0.2406	-23.79
Okapi	0.3042	**0.2654**	-12.75

Table III
DOCUMENT EXPANSION USING OPENTHESAURUS

	MAP			MAP	
Model	No Exp.	GIRT	%Change	OpenThes.	%Change
tf-idf	0.2275	0.2285	0.44	0.2289	0.62
LM2	0.3215	0.324	0.78	0.3233	0.56
InB2	**0.3493**	0.3485	-0.23	0.3483	-0.29
Okapi	0.3494	**0.3503**	0.26	0.3510	0.46

Table IV
MAP AFTER QUERY EXPANSION WITH SHORT QUERIES (T)

us that adding keywords to bibliographic resources might be helpful. If we have a closer look at our results, we observe that for the *InB2* model for example, we have an improvement for 78 queries, but also a decrease for 45 queries. The question that then comes up, is if it is worth to spend human resources to add this keywords. Manually added keywords require time and people qualified in the given domain. Therefore we want to analyze if automatically added keywords based on a thesaurus might yield the same performance improvement.

B. Automatic Document Expansion

In this section we presented the results obtained when extending documents automatically with keywords. For manual expansion, an expert selects appropriate keywords from the thesaurus based on its knowledge of the domain and the context of the documents. With a computer we need an algorithm to select the controlled terms to be added. Our expansion procedure is mainly based on the textual entailment measures proposed in Section V and can be divided into four steps. First we select the part of the document (or query) to be extended. Then for each term t_i, we do a search in the thesaurus. For each retrieved thesaurus entry for the term t_i we retain all the terms w_j^i contained in the entry and compute their similarity score $score_{ij}$ with their related term t_i using one of the similarity measures described in Section V. Once we have finished this step for all terms t_i, we have a set of couples $(w_j^i, score_{ij})$. The terms w_j^i are candidates for expansion. Finally, since the number of potential candidates might be elevated, we select the N_{Best} terms with the highest score to extend the documents. For some documents there might be less than N_{Best} terms available. In this case all the candidate terms are added. Since the number of documents to expand is quite high, after some empirical analysis we selected the $Jaccard$ similarity for the expansion procedure, and fixed N_{Best} at 50 (which also equals the mean number of controlled vocabulary terms per documents). Table II shows the results of the retrieval using documents expanded with GIRT-thesaurus, and Table III using OpenThesaurus for expansion. We observe that automatically enhancing documents does not improve retrieval. Compared to manually added keywords, we only have improvement for 22 queries (vs. 78). However for 22 queries automatic document expansion performs better than manual expansion. For example with Query #153 ("Kinderlosigkeit in Deutschland"), the MAP is 0.6030 with manual expansion and 0.0839 with our suggested automatic expansion, while for Query #204 ("Kinder- und Jugendhilfe in der russischen Föderation") we have a MAP of 0.0494 after manual and 0.1930 after automatic expansion.

Compared to the GIRT-thesaurus the results are slightly better for OpenThesaurus, but we still have an important decrease compared to the retrieval results without keywords.

C. Query Expansion

In this part we present our results obtained when extending queries. After several tests, we decided to fix N_{Best} at 5, i.e. to each query are added at most 5 terms extracted form the thesaurus. We used the three measures presented in Section V to measure textual entailment between terms and chose expansion terms, as well as four retrieval models and the two thesauri and two query formulations, a short one using only the title part (T) and a longer using title and description (TD). We search in the complete document (title, abstract and keyword). Since our test runs show that all textual entailment measures perform the same, we only present results for the Jaccard measure.

Table IV shows a recapitulation of query expansion using Jaccard similarity for short query formulations (T) for both thesauri and the comparison to the baseline. We observe that query expansion does not bring any significant improvement.

	MAP			MAP	
Model	No Exp.	GIRT	%Change	OpenThes.	%Change
tf-idf	0.2428	0.243	0.08	0.2431	0.12
LM2	0.3606	0.3621	0.42	0.3616	0.28
InB2	0.379	0.3795	0.13	0.3793	0.08
Okapi	**0.3856**	**0.3861**	0.13	**0.3865**	0.23

Table V
MAP AFTER QUERY EXPANSION WITH LONG QUERIES (TD)

The small variations in the MAP are due to minor changes in the order of the retrieved documents rather than in the expected better retrieval of relevant documents for expanded queries. We make the same observations for longer query formulations as seen in Table V.

If we have a closer look at the results query-by-query for the *InB2* model and short queries (T), we see that for GIRT-thesaurus we have an improvement for 52 queries and decrease for 72. For OpenThesaurus, the use of thesaurus improves retrieval for 36 queries and decreases for 38. Query #44 ("Radio und Internet") for example has MAP 0.3986 if we do not use any query expansion. Using GIRT-thesaurus boosts MAP to 0.4509 (added words are "Rundfunk", "Datennetz", "Datenaustausch" and "Welle"), while OpenThesaurus even performs a MAP of 0.4846 ("Hörfunk", "Rundfunk", "Netze", "Netz" and "Funk"). For Query #118 ("Generationsunterschiede im Internet") however, the use of the GIRT-thesaurus drops MAP from 0.4789 to 0.4408 (added "Datennetz", "Datenaustausch" and "Intranet") while OpenThesaurus improves MAP to 0.4827 ("Netze", "Netz", "Web", "WWW" and "World"). This last example also shows that the choice of terms to expand the query is important, some expansion terms might hurt performance while some others might improve.

VII. CONCLUSION

In this paper we present the use of manually added keywords for searching relevant information in bibliographic records database written in the German language. While the manually assigned descriptors extracted from a controlled vocabulary considerably improve retrieval performance ($+13.9\%$ in mean), automatically added terms either from the same controlled vocabulary or from a general thesaurus hurt the retrieval performance. In a second part we tried to enhance queries rather than documents. The inclusion of keywords to the query however does not improve retrieval results.

We can conclude that adding terms extracted from a controlled vocabulary may improve retrieval performance. The problem however is to choose the right keyword terms to add to the documents. We tried different techniques to select expansion terms, but all show the same performance. Human specialists seem to be more accurate in selecting the appropriate keywords to enhance retrieval performance. In contrary to machines, a human person having a good knowledge in the given domain can take into account the semantics and pragmatics as well as the importance of a keyword term in the underlying corpus. Although if for some queries even manual indexing does not help to improve retrieval, it seems to be worth to invest time and human resources to gain in the overall performance for finding relevant information.

Acknowledgments: This research was supported in part by the Swiss NSF under Grant #200021-113273.

REFERENCES

[1] E. Voorhees, "Using WordNet™ to disambiguate word senses for text retrieval," in *Proceedings ACM-SIGIR'93*, 1993, pp. 171–180.

[2] C. W. Cleverdon, "The Cranfield Tests on Index Language Devices," *Aslib Proceedings*, vol. 19, pp. 173–194, 1967.

[3] T. B. Rajashekar and W. B. Croft, "Combining automatic and manual index representations in probabilistic retrieval," *Journal of the American Society for Information Science*, vol. 46, pp. 272–283, 1995.

[4] J. Savoy, "Bibliographic database access using free-text and controlled vocabulary: an evaluation," *Information Processing & Management*, vol. 41, pp. 873–890, 2005.

[5] V. Petras, "How one word can make all the difference - using subject metadata for automatic query expansion and reformulation," in *Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria.*, 2005.

[6] M. Kluck, "Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation," in *ISI*, ser. Schriften zur Informationswissenschaft, B. Bekavac, J. Herget, and M. Rittberger, Eds., vol. 42. Hochschulverband für Informationswissenschaft, 2004, pp. 247–268.

[7] H. Schott, Ed., *Thesaurus Sozialwissenschaften*. Informationszentrum Sozialwissenschaften, Bonn, 2002.

[8] D. Naber, "OpenThesaurus: ein offenes deutsches Wortnetz," in *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn*. Peter-Lang-Verlag, Frankfurt, 2005, pp. 422–433.

[9] J. Savoy, "Combining multiple strategies for effective monolingual and cross-lingual retrieval," *IR Journal*, vol. 7, pp. 121–148, 2004.

[10] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, pp. 95–108, 2000.

[11] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Trans. Inf. Syst.*, vol. 20, pp. 357–389, 2002.

[12] D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term," in *25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, 2002, pp. 35–41.

Adapting the tf idf Vector-Space Model to Domain Specific Information Retrieval

Claire Fautsch
Computer Science Department
University of Neuchatel, Switzerland
claire.fautsch@unine.ch

Jacques Savoy
Computer Science Department
University of Neuchatel, Switzerland
jacques.savoy@unine.ch

ABSTRACT

The default implementation in Lucene, an open-source search engine, is the well-known vector-space model with *tf idf* weighting. The objective of this paper is to propose and evaluate additional techniques that can be adapted to this search model, in order to meet the particular needs of domain-specific information retrieval (IR). In this paper, we suggest certain specificity measures derived from either information theory or corpus-based linguistics. As an additional feature we suggest accounting for the number of search terms that a query and retrieved documents have in common. To integrate these methods we design and implement four extensions to the classical *tf idf* model and then evaluate the new IR models by applying them to four different domain-specific collections and comparing them to results found by a probabilistic retrieval model. The results tend to demonstrate that the adapted vector-space models clearly outperform the baseline approach (*tf idf*) and that performance levels obtained even surpass those found in the Okapi model.

Keywords

Domain Specific, Vector-Space Model, Term Specificity

1. INTRODUCTION

Due to the current growth in electronic resources, an increasing number of scientific journals and web sites and the emergence of blogs, efficient domain-specific information retrieval is more and more important. Biologists looking for interactions between DNA sequences and a given disease will limit their searches to the bio-medical domain, and eventually even to the most recent publications in this domain. Other users looking for specifications and opinions regarding a new cell phone would probably rather search in technical blogs. Among these users, the common need is relevant, domain-specific information. Given the extensive range of scientific publications and their use of technical language and formulae, collections on cooking recipes for example and other specialized subjects tend to use simple but specific terminology.

To make domain-specific information retrieval more efficient, the given domain is usually studied to unveil its underlying properties. For example Yu & Agichtein [1] showed orthographic variants found in the bio-medical domain are an important issue. Examples that might be encountered include spelling variants (e.g., "ecstasy", "extasy", or "ecstacy"), alternative punctuation and tokenization (e.g., "Nurr 77", "Nurr77" or "Nurr-77") or alternative names (e.g., the same protein could be named as "LARD", "Apo3", "DR3", "TRAMP", "wsl" or "TnfRSF12"). In order to improve retrieval effectiveness and account for the underlying characteristics of the corresponding domain these variations may be incorporated into the search strategy (e.g., extending the query with spelling variations extracted from a dedicated database). Such approaches cannot be easily applied to other fields.

Moreover, previous studies [2] in domain-specific IR tended to demonstrate that it is important to assign higher rankings to retrieved documents having many terms in common with the submitted query. In fact when a term occurs rarely its presence in a document surrogate may promote this document to the top of the ranked list, a phenomenon that may occur even if the document does not include additional search terms, able to more precisely specify the meaning of user's information need. Additionally, this problem tends to appear more frequently in domain-specific IR due to the fact that the precise meaning is given by a sequence of terms while a single term or a bigram may be too ambiguous (e.g., "algorithm" vs. "parallel sorting algorithm").

In this paper our intention is to propose and evaluate a variety of methods that can be applied to all domains in the same way, through accounting for both the specificity of these search terms and on the number of terms that the query and the retrieved documents have in common. Finally we also take account for the constraint to work with the classes defined in the Lucene open-source search engine[1] [3]. It is easier to extend the implemented weighting scheme base on *tf idf* weighting then to implement more complex search models such as probabilistic models.

The rest of the paper is organized as follows. Section 2 presents related work while Section 3 describes the test collections used to evaluate our proposed models. In Section 4 we expose our extended information retrieval models and describe the evaluation methodology used. Finally in Section 5 we analyze the results and draw some conclusions.

[1]http://www.lucene.apache.org/

SAC'10 March 22-26, 2010, Sierre, Switzerland.

2. RELATED WORK

In this section, we describe the previous research done on term specificity estimation and its use in domain-specific information retrieval.

Three approaches for measuring term specificity are presented in [4]. These measures are based on measures found in information theory and adapted for use in automatic hierarchy construction. As a method of identifying the domain-specific vocabulary, in a related paper Drouin [5] suggested comparing a domain-specific corpus to a more general reference corpus.

Over the last years domain specific information retrieval has become an important issue and thus has been the subject of various evaluation campaigns and tracks such as TREC[2] (Text REtrieval Conference) or CLEF[3](Cross Language Evaluation Forum). In domain-specific information retrieval, an approach currently being evaluated is document or query enhancement, using a controlled vocabulary based on a domain-specific thesaurus. For the GIRT corpus for example Petras [6] suggested using manually assigned keywords to improve retrieval results. In a second step she showed that combining pseudo-relevance feedback and query expansion by using a thesaurus could improve retrieval performance. Abdou *et al.* [7] described the impact of manually assigned descriptors taken from the MeSH thesaurus, illustrating how these descriptors could enhance retrieval performance by up to 13.5%. In the same paper they showed that extending the queries by applying automatically generated orthographic variants would slightly enhance overall retrieval effectiveness, although the outcome was less successful than expected. Other possibilities might include the use of a controlled vocabulary based on a domain-specific thesaurus to extend document representation [8] or even the generation of more specific indexing methods [2]. These suggested methods would however require various adaptations from domain to domain.

One of the goals in this paper is to go beyond extending documents and queries by means of specialized thesauri or other related lexical structures. In our opinion it would be preferable to enhance the overall retrieval performance by identifying specific search terms and thus enhancing their importance when matching queries and document surrogates.

3. TEST COLLECTIONS

To evaluate the retrieval models proposed in this paper, we use four different test corpora covering three different domains: biomedical, social sciences and blogosphere. We also consider two natural languages, namely English and German.

3.1 Genomics

The first collection was made available through the TREC evaluation campaign and had been used for the *ad hoc* Genomics retrieval track in 2004 and 2005. This corpus contains a 10-year subset from MEDLINE, a collection of abstracts and citations from publications in the bio-medical domain (containing 4,591,008 records or about 10.6 GB of compressed data), and includes a set of 100 topics as well as their relevance judgments. More information on these documents and topics can be found in [9].

[2]http://trec.nist.gov/

[3]http://www.clef-campaign.org/

3.2 Blog

The second collection was also used in the TREC evaluation campaign from 2006 to 2008 during the Blog tracks. This corpus was crawled between December 2005 and February 2006 and contains a total of 148 GB of data (or 4,293,732 documents), consisting of 753,681 feeds (38.6 GB), 3,215,171 permalinks (88.8 GB) and 324,880 homepages. In our evaluation we used only the permalink part and a total of 150 queries available for this collection. More information about this collection can be found in [10].

3.3 GIRT

The last two specific collections cover the social sciences domain. The GIRT (German Indexing and Retrieval Test database) was made available through the CLEF evaluation campaign. The original German collection contains 151,319 records taken from the social sciences while the English version is a translation of the German collection. For each language we applied a total of around 125 queries used in the CLEF domain-specific tasks from 2004 to 2008. For more information on this collection see [11].

3.4 General Corpora

Finally we needed a German and an English general reference corpus. The German reference collection was linked to a newspaper corpus containing 294,809 articles published in the *Frankfurter Rundschau* (1994), *Der Spiegel* (1994 and 1995) as well as articles from 1994 provided by the Swiss news agency (SDA).

The English corpus contains 169,477 articles extracted from the *Glasgow Herald* for 1995 as well as news articles extracted from the *Los Angeles Time* (1994). More information about both collections can be found in [12].

4. IR MODELS AND EVALUATION

The well-known vector-space model with *tf idf* weighting scheme has been adopted as the default IR model in Lucene, an open source search engine written in Java. Based on the implementation proposed, expanding this IR model is a rather straightforward procedure, but implementing the Okapi model requires substantial work. From several evaluation campaigns, e.g., TREC or CLEF, it is known that the retrieval effectiveness of the vector-space model with *tf idf* weighting is lower than certain implementations of the probabilistic model, such as Okapi [13]. As described in this section, we will extend the vector-space model to meet both the challenges of domain-specific information retrieval and those involved in improving its overall retrieval performance levels.

4.1 Vector-Space Model

As a baseline approach we used a standard *tf idf* weighting scheme with a cosine normalization. The score for the document D_i given the query Q_k was calculated by applying the following formula:

$$Score(D_i, Q_k) = \sum_{t_j \in Q_k} w_{ij} \cdot w_{kj}$$

where w_{ij} and w_{kj} respectively represent the weights of term t_j in the document D_i and in the query Q_k and are defined

as follows:

$$w_{ij} = \frac{tf_{ij} \cdot idf_j}{\sqrt{\sum_k (tf_{ik} \cdot idf_k)^2}}$$

where tf_{ij} is the frequency of the term t_j in the document D_i (or the query), idf_j the inverse document frequency computed as $\log(n/df_j)$, with n indicating the number of documents in the collection and df_j is the number of documents containing the term t_j.

4.2 Adapted Vector-Space Model

Our aim is to extend the previously described vector processing model to account for the particularities of domain-specific information retrieval, as well as an extension that would remain domain-independent. The underlying idea here is to discriminate between general and specific terms in topic formulation and as such attribute greater importance to more specific terms in the matching score. The *idf* measure can be viewed as a term specificity measure but only based on document frequency information (the number of documents in which a given term occurs) and not for example the occurrence frequency in the corpus, in a given document or compared to a general corpus.

Moreover, we wanted to assign more weight to documents having more than one search term in common with the submitted query. We therefore extended the *tf idf* model by using following formula:

$$Score(D_i, Q_k) = \sum_{t_j \in Q_k} (w_{ij} \cdot w_{kj} + spec_C(t_j))$$

where $spec_C(t_j)$ measures the specificity of the term t_j in the collection C. To measure this specificity we used the various methods described in the following paragraphs. In the proposed scoring function we adopted an addition operator to combine the *tf idf* model with the supplementary weight attached to the specificity of each search term. The advantage of this additive process is that it allows us to increase the matching score directly, according to the number of search terms that the query and the retrieved items have in common.

4.2.1 Mutual Information

Mutual information (MI) is widely used in Natural Language Processing (NLP) [14] to measure the association between two terms. We will use the MI measure presented in [15] estimating a relevance score of the term t across the collection, and calculated as follows:

$$MI(t) = \sum_i P(D_i) \cdot \log\left(\frac{P(t|D_i)}{P(t)}\right)$$

where $P(D_i) = 1/n$ is the probability of selecting the document D_i in the corpus, $P(t|D_i) = tf/l_i$ the probability of term t occurring in the document D_i, and $P(t) = cf/cl$ the probability of the term t in the collection, with cf the number of occurrences of t in the collection, cl the total number of indexing terms in the collection and l_i the length of D_i. In the current context, $spec_C(t)$ is then defined as $spec_C(t) = MI(t)$ and we denote this model as *tf idf* + MI.

4.2.2 Information Gain

Information Gain (IG) is a measure borrowed from information theory and NLP to estimate the relevance of a term t to a given document. We use the formula presented in [15] and defined as follows:

$$\begin{aligned} IG(t) \quad &= P(t) \textstyle\sum_i (P(D_i|t) \cdot \log\left(\frac{P(D_i|t)}{P(D_i)}\right) \\ &+ P(t^c) \textstyle\sum_i (P(D_i|t^c) \cdot \log\left(\frac{P(D_i|t^c)}{P(D_i)}\right) \end{aligned}$$

where $P(t^c)$ is the probability of the term t not occurring (i.e., $1 - P(t)$). The probabilities are calculated as defined in the previous paragraph. Finally we define $spec_C(t) = 1 - IG(t)$. We will reference to this model as *tf idf* + IG.

4.2.3 Relative Frequency Ratio

This third measure is based on the comparison of two corpora: a general one and a specific one. We assume here that domain-specific words are more frequent in a specific collection than in a general corpus, and based on this assumption, we calculate the specificity of the term t as follows:

$$spec_C(t) = \begin{cases} 1 & \text{if } freqR(t) \leq 1 \\ 2 & \text{if } 1 < freqR(t) < \infty \\ 3 & \text{if } freqR(t) = \infty \end{cases}$$

where $freqR(t) = \frac{f_{spec}}{f_{gen}}$ with f_{spec} and f_{gen} are the relative frequencies of the term t in the specific and the general corpus respectively (with similar size). If for a given term the relative frequency is greater than or equal to the general corpus compared to the specific one, the $spec_C(t)$ value will be one, while for the inverse it will have value of two. At the limit, when the frequency in the general corpus is null (the term does not occur), we get $freq(t) = \infty$ and thus $spec_C(t) = 3$. We reference to this measure as *tf idf* + RFR.

4.2.4 Index of Peculiarity

The final measure is based on 3-gram segmentation, normally used to detect spelling errors. In this case, each term is subdivided into tokens of length 3 (e.g., the word "house" generates the tokens "hou", "ous", and "use"). For each 3-gram (e.g., "xyz"), a *index of peculiarity* (IP) is calculated as follows:

$$IP(xyz) = \frac{\log(f(xy) - 1) - \log(f(yz) - 1)}{2} - \log(f(xyz) - 1)$$

where $f(xy)$ indicates the frequency of the bigram "xy" in the corpus, and $f(xyz)$ the frequency of the 3-gram "xyz".

In a spelling detection context, a large IP would indicate a misspelled word, while in our case a large IP means that a given term has a very specific meaning. For a given word t, the $spec_C(t)$ is calculated using the following formula.

$$spec_C(t) = \max_{xyz \in t} IP(xyz)$$

where "xyz" is a 3-gram extracted from word t. Finally, if the length of the given term is less than 3, $spec_C(t)$ is fixed at 0. We reference this model as *tf idf* + IP.

4.3 Okapi Model

To compare the results of the standard vector-space model to our new adapted models, we also used a probabilistic information retrieval model. To do so we implemented the Okapi (BM25) model proposed by Robertson *et al.* [13]. The document score was evaluated using the following formula:

$$Score(D_i, Q) = \sum_{t_j \in Q} qtf_j \cdot \log\left[\frac{n - df_j}{df_j}\right] \cdot \frac{(k_1 + 1) \cdot tf_{ij}}{K + tf_{ij}}$$

with $K = k_1 \cdot [(1 - b) + b \cdot \frac{l_i}{avdl}]$ where qtf_j denotes the frequency of term t_j in the query Q, df_j the number of documents in which the term t_j appears, l_i the length of the document D_i, and $avdl$ represents the average document length. To obtain the best retrieval performance the constants b and k_1 were set empirically according to the underlying collection.

4.4 Evaluation Methodology

To evaluate retrieval performance we used MAP [16] (Mean Average Precision), computed using the TREC_EVAL[4], using at most 1,000 retrieved documents per query to calculate MAP values.

To determine whether or not a given search strategy was statistically better than another, we applied the bootstrap methodology [17], with the null hypothesis H_0 stating that both retrieval schemes produced similar performance. In the experiments presented in this paper statistically significant differences were detected by applying a two-sided test (significance level $\alpha = 5\%$). Such a null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected.

5. EVALUATION

For the four collections we presented and tested four adapted vector-space models, as well as the classical *tf idf* and the Okapi model. Table 1 lists the results of our tests and Table 2 presents the mean improvement of each adapted model compared to the baseline approach.

The second column in Table 1 lists the results obtained on the Genomics collection, the third column those obtained for the Blog collection and the two last columns the results for the English and German GIRT corpora. Each column lists the results of the best performing model in bold print. Based on the statistical tests, we always found that there were statistically significant performance differences when comparing the classical *tf idf* and all other approaches. In all cases, the adapted vector space models perform better then the classical *tf idf* model (a difference that was always greater than +50%). When the Okapi model was used as a baseline, we used a "*" to denote those models showing statistically significant differences in the retrieval performances obtained.

For the Genomics collection, all adapted models except the IP model performed statistically at the same level as the Okapi model. To understand the effect of term specificity, we may analyze the performance of some queries. A closer look at the *tf idf* and the *tf idf*+IG models for example showed that for the query *"Comparison of Promoters of GAL1 and SUC1"* the MAP varied from 0.0323 to 1.0 when we accounted for term specificity. Indeed, this query retrieved only a single relevant document, and ranked it in position 31, when no specificity information was used (i.e., using the *tf idf* model). In total we obtained improvements for 89 of the 99 queries having at least one relevant document (he remaining query does not have any relevant document in the collection). The biggest decrease resulting from the application of specificity information was for the query *"Proteins involved in the nerve growth factor pathway"*. We noticed that except for proteins, this query did not contain any words which could be considered as belonging to biomedical domains.

[4]http://trec.nist.gov/trec_eval

	Mean Average Precision			
	Genomics	Blog	GIRT-EN	GIRT-DE
Queries	99	150	124	125
tf idf	15.58 *	19.33 *	20.79 *	23.92 *
tf idf + MI	29.41	30.91 *	31.34 *	**38.36**
tf idf + IG	**30.58**	30.82 *	32.82	38.05
tf idf+ IP	27.00 *	30.99 *	31.31 *	37.40
tf idf + RFR	30.25	30.94 *	31.66 *	37.52
Okapi	30.26	**33.57**	**33.70**	37.40

Table 1: MAP of the adapted *tf idf* vector-space models and Okapi probabilistic model

For the Blog collection, Okapi model resulted in the best performance. All other models resulted in statistically different retrieval performances when compared to Okapi. The adapted vector-space models showed considerable improvement over the standard vector-space model. When comparing the *tf idf* model and the *tf idf*+IP model for example, we observed that for 97 queries there were improvements while for 6 queries the classical *tf idf* produces a better performance. There was no change for the other 47 queries. The greatest improvement occurred with the query "Ruth Rendell" (from 0.0008 to 0.7737) in which the presence of both search terms improves the retrieved performance.

Upon an analysis of the results from the English GIRT collection, we observed that all models except the *tf idf*+IG model showed statistically significant performance differences when compared to Okapi. Even for this collection however the adapted models improved retrieval performance considerably. For the *tf idf*+IG model, the highest improvement was obtained for the query "Advertising and Ethics" where the MAP improved from 0.0374 to 0.7657 while for the query "Mortality rate" the MAP dropped from 0.5867 to 0.3702. For this model we obtained improvements for 107 queries, out of a total of 124 queries producing at least one relevant document.

Finally for the German GIRT collection we observed that all adapted vector-space models produced better performance than the Okapi model, yet these performance differences were not statistically significant. When comparing the *tf idf*+MI model to the classical *tf idf*, we observed improvements for 103 queries, with the highest improvement occurring for the query "Minderheitenpolitik im Baltikum" (MAP from 0.0957 to 0.7391) while for the query "Vaterrolle" we observed the greatest MAP decrease (from 0.7573 to 0.5545).

As depicted in Table 2, all four suggested approaches resulted in better performances than the classical *tf idf* vector-space model.

6. CONCLUSION

In this paper we presented four different extensions to the *tf idf* information retrieval model, that would allow it to be better adapted to specific domains. We have worked with the constraint to use the *tf idf* vector-space model because it is frequently used as well as implemented by the open-

Model	Mean MAP	%Change
tf idf	19.91	
tf idf + MI	32.51	+63.30%
tf idf + IG	33.07	+66.13%
tf idf + IP	31.68	+59.13%
tf idf + RFR	32.59	+63.74%

Table 2: Average improvement of the various adapted *tf idf* vector-space models compared to the baseline *tf idf*

source engine Lucene. Moreover this scheme does not have any parameters that need to be tuned.

Second, this work focused on detecting specific search terms and increasing their matching value. Various measures were proposed to identify specific terms, thus making it possible to derive various implementations. In the suggested additive scheme derived from the classical *tf idf* model, we also accounted for the number of terms that the query and the retrieved documents had in common.

We compared the suggested models to the Okapi model, a probabilistic approach, and then tested all models by applying them to four different collections written in English and German. The experiment showed that the adapted vector-space models significantly improved retrieval performances when compared to the classical *tf idf* approach. For the German collection the four adapted vector-space models even outperformed the Okapi model, while for the Genomics Collection at least three out of four adapted models produced retrieval performances that were statistically similar to those obtained by the Okapi approach.

We can thus conclude that for information retrieval in specific domains accounting for term the specificity is indeed worth the effort. One advantage of the adapted vector-space models is that no parameters are required, while this is not the case for the Okapi model. A second advantage is that these models can easily be adapted to each collection, regardless of the underlying domain or language.

Acknowledgments.

This research was supported in part by the Swiss National Science Foundation under Grant #200021-113273.

7. REFERENCES

[1] H. Yu and E. Agichtein, "Extracting synonymous gene and protein terms from biological literature," *Journal of Bioinformatics*, vol. 19, pp. 340–349, 2003.

[2] C. Fautsch and J. Savoy, "UniNE at TREC 2008: Fact and opinion retrieval in the blogsphere," in *The Seventeenth Text REtrieval Conference Proceedings (TREC 2008)*, 2008.

[3] E. Hatcher and O. Gospodnetic, *Lucene in Action (In Action series)*. Manning Publications, December 2004.

[4] P. M. Ryu and K. S. Choi, "Measuring the specificity of terms for automatic hierarchy construction," in *ECAI-2004 Workshop on Ontology Learning and Population*.

[5] P. Drouin, "Detection of domain specific terminology using corpora comparison," in *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

[6] V. Petras, "How one word can make all the difference - using subject metadata for automatic query expansion and reformulation," in *Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria*, 2005.

[7] S. Abdou and J. Savoy, "Searching in Medline: Query expansion and manual indexing evaluation," *Information Processing & Management*, vol. 44, pp. 781–789, 2008.

[8] C. Fautsch and J. Savoy, "Comparison between manually and automatically assigned descriptors based on a German bibliographic collection," in *Proceedings of the 6th International Workshop on Text-based Information Retrieval (TIR 2009)*, 2009.

[9] W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst, "TREC 2005 Genomics Track Overview," in *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2005.

[10] C. Macdonald and I. Ounis, "The TREC Blogs06 collection : Creating and analysing a blog test collection," *DCS Technical Report Series*, 2006.

[11] M. Kluck, "Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation," in *ISI* (B. Bekavac, J. Herget, and M. Rittberger, eds.), vol. 42 of *Schriften zur Informationswissenschaft*, pp. 247–268, Hochschulverband für Informationswissenschaft, 2004.

[12] C. Peters, P. Clough, J. Gonzalo, G. J. F. Jones, M. Kluck, and B. Magnini, eds., *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*, vol. 3491 of *LNCS*, Springer, 2005.

[13] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, pp. 95–108, 2000.

[14] P. M. Nugues, *An Introduction to Language Processing with Perl and Prolog*. Berlin: Springer, 2006.

[15] C. Orăsan, V. Pekar, and L. Hasler, "A comparison of summarisation methods based on term specificity estimation," in *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004)*, 2004.

[16] C. Buckley and E. M. Voorhees, "Retrieval system evaluation," in *TREC: Experiment and Evaluation in Information Retrieval*. (E. M. Voorhees and D. K. Harman, eds.), (Cambridge, MA), pp. 53–75, MIT Press, 2005.

[17] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Information Processing & Management*, vol. 33, pp. 495–512, 1997.

Algorithmic Stemmers or Morphological Analysis: An Evaluation

Claire Fautsch, Jacques Savoy

Computer Science Department
University of Neuchatel
2009 Neuchâtel, Switzerland
{Claire.Fautsch, Jacques.Savoy}@unine.ch

Abstract. It is important in information retrieval, information extraction or classification tasks that morphologically related forms be conflated under the same root or lemma. To achieve this for the English language, both algorithmic stemming and various morphological analysis approaches have been suggested. Based on CLEF test-collections containing 284 queries and various IR models, this paper evaluates these word-normalization proposals. We found that the *Divergence from Randomness* paradigm tends to result in slightly better retrieval effectiveness than the Okapi, and significantly better than the language model or *tf idf* IR schemes. Stemming improves the MAP significantly by around 7%, while performance differences are not significant when comparing various algorithmic stemmers nor algorithmic stemmers and morphological analysis. Accounting for thesaurus class numbers during indexing does not modify overall retrieval performances. Finally, we demonstrate that including a stopword list, even one containing only around ten terms, might significantly improve retrieval performance, depending on the IR model.

Keywords. Stemming Strategies, Morphological Analysis, Algorithmic Stemmers, Thesaurus, English Language, WordNet, Evaluation.

1 Introduction

Stemming refers to the conflation of word variants into a common stem (or form when the string cannot be found in the language). In information retrieval (IR) the application of a stemming procedure when indexing documents or requests is assumed to be a good practice (Manning *et al.* 2008); although the *N*-gram indexing strategy is typically an exception to this rule (McNamee & Mayfield 2004). For example when a query contains the word "horse," it seems reasonable to also retrieve documents containing the related word "horses," a practice which usually tends to improve retrieval effectiveness. Designing effective stemming procedures may also be helpful for other purposes, such as text mining, natural language processing or gathering statistics on a document corpus.

For the English language, various authors have proposed algorithmic stemmers based on the morphological rules of this language (e.g., see Lovins (1968) and Porter (1980)). An alternative is to apply a more complex morphological analysis requiring additional computational resources and a dictionary able to return the correct lemma (or dictionary entries). Moreover, as a means of defining better matches between terms occurring in the query and the document we might also make use of part-of-speech (POS) information (Krovetz 1993), (Savoy 1993) . Finally, once a word's corresponding lemma has been found, we could also consider the word's various synonyms, making use of synset numbers (thesaurus class number) available in the WordNet™ thesaurus (Fellbaum 1998).

The main objective of this paper is to analyze and evaluate various stemming strategies using a relatively large number of queries. The rest of the paper is organized as follows: Section 2 describes related stemming approaches while Section 3 depicts the main characteristics of our test-collection. Section 4 briefly describes the IR methods applied during our experiments. Section 5 evaluates the performance of various IR models along with different algorithmic stemmers or morphological analysis. The use of POS information and thesaurus class numbers is also evaluated and analyzed. The main findings of this paper are presented in the conclusion.

2 Related Work

In the IR domain stemming is usually considered as an effective means of enhancing retrieval performance through conflating several different word variants into a common form. As a first approach to designing a stemmer, we begin by removing only inflectional suffixes so that singular and plural word forms (e.g., "dogs" and "dog") or feminine and masculine variants (e.g., "actress" and "actor") will conflate to the same root. Suffix removal is also controlled through the adjunct of quantitative restrictions (e.g., '-ing' would be removed if the resulting stem had more than three letters, as in "running," but not in "king") or qualitative restrictions (e.g., '-ize' is removed if the resulting stem does not end with "e" as in "seize"). Moreover, certain *ad hoc* rules are used to correct spelling and improve conflation accuracy (e.g., "running" becomes "run" and not "runn"), due to certain irregular grammar rules usually applied to a language to facilitate pronunciation. Of course, one should not stem proper nouns such as "Collins" or "Hawking", at least when the system can recognize them.

These suffix-removal methods are based on a set of rules known as algorithmic stemmers, and thus ignore word meaning and part-of-speech categories. Other stemming techniques that remove only morphological inflections are termed "light" suffix-stripping algorithms, such as the S-stemmer (Harman 1991), apply three rules to remove the plural morpheme '-s' . There are also more sophisticated approaches that remove derivational suffixes (e.g., '-ment', '-ably', '-ship' in the English language). Those suggested by Lovins (1968) are based on a list of over 260 suffixes while Porter's algorithm (1980) looks for about 60 suffixes.

Stemming methods are usually designed to work with general texts and work with any given language. Certain stemming procedures may however be especially de-

signed for a specific domain (e.g., medicine) or a given document collection, such as that of Xu & Croft (1998). They suggest developing stemming procedures using a corpus-based approach which more closely reflects the language used (including word frequencies and other co-occurrence statistics), instead of using a set of morphological rules in which the frequency of each rule (and therefore its underlying importance) is not precisely known.

Algorithmic stemming procedures tend to make errors, usually due to over-stemming (e.g., "general" becomes "gener", and "organization" is reduced to "organ") or to under-stemming (e.g., with Porter's stemmer, the words "create" and "creation" or "European" and "Europe" do not conflate to the same root). In general however stemming tends to improve recall, yet these examples show it may also decrease precision, rendering web search strategies problematic. In this case, Peng *et al.* (2007) suggest applying context sensitive stemming methods to search terms based on a statistical language model. Another method of reducing stemming errors involving an on-line dictionary was suggested, in order to produce better conflations (Krovetz 1993).

Continuing in this vein but requiring more computational resources is a more complex morphological analysis capable of precisely defining the corresponding lemma (or entry in the dictionary) for a given word. Flexions can be removed to obtain the lemma (e.g., "houses" becomes "house"), and the resulting part-of-speech information could then be used to further enhance the quality of the suffix-removal process. For example, the derivational suffix '-able' is used to form an adjective from a verb stem as in "readable" or "thinkable", and for the French language a similar process is available (Savoy 1993). This stemming strategy, based on more complex morphological analysis is used very infrequently in information retrieval, and mostly for large collections.

Based on an analysis of IR stemming performances, Harman (1991) demonstrated that no statistically significant improvement would result from applying three different algorithmic stemmers, namely that of Lovins (1968), Porter (1980) and the light S-stemmer (that conflates only singular and plural English word forms). A query-by-query analysis revealed that stemming did affect the performance, with the number of queries showing improved performance almost equaling the number of queries showing poorer performance. Other studies (Hull 1996) based on a single IR model (a variant of the classical *tf idf* method) showed that stemming resulted in modest improvements, ranging from 1% to 3%. This analysis revealed however that stemming tends to make a difference for many individual queries. According to Hull's study (1996), all stemmers resulted in statistically superior average precision than a non-stemming approach. Moreover, the S-stemmer proved to be less effective than the Lovins or Porter methods.

Based on these facts, the rest of this paper will address the following questions: 1) With a large set of queries (around 300), is suffixing really better than a non-stemming approach? 2) Is it possible to obtain improved retrieval effectiveness when applying a morphological analysis instead of an algorithmic stemmer? 3) Is it possible to obtain statistically significant differences between various algorithmic stemmers? 4) Does the use of thesaurus class numbers or simple POS information prove useful in increasing retrieval effectiveness?

3 Test-Collections

The evaluations reported in this paper were based on the English test collections built during the CLEF 2001 through CLEF 2006 evaluation campaigns (Peters *et al*. 2008) and regrouped into the Robust track in CLEF-2008. This corpus consists of articles published in 1994 in the *Los Angeles Times*, as well as others extracted from the *Glasgow Herald*, newspapers published in 1995. This collection contains a total of 169,477 documents (or about 579 MB of data), and each article contains about 250 on average (median: 191) content-bearing terms (not counting commonly occurring words such as "the," "of" or "in"). Typically, documents in this collection are represented by a short title plus one to four paragraphs of text, and both American and British English spellings can be found in the corpus.

	2001	2002	2003	2004	2005	2006
Source	*LA Times*	*LA Times*	*LA Times Glasgow H.*	*Glasgow H.*	*LA Times Glasgow H.*	*LA Times Glasgow H.*
Size	425 MB	425 MB	579 MB	154 MB	579 MB	579 MB
# docs	113,005	113,005	169,477	56,472	169,477	169,477
# topics	47	42	54	42	50	49
Topics	#41 - #90	#91 - #140	#141 - #200	#201 - #250	#251 - #300	#301 - #350

Table 1. A few CLEF test-collections statistics

This collection contains 310 topics, each subdivided into a brief title (denoted as T), a full statement of the information need (called description or D), plus any background information that might help assess the topic (narrative or N). An example is given in Table 2. These topics cover various subjects (e.g., "El Niño and the Weather," "Chinese Currency Devaluation," "Eurofighter," "Victories of Alberto Tomba," "Marriage Jackson-Presley" or "Computer Animation"), including both regional ("Films Set in Scotland," "Area of Kaliningrad") and international coverage ("Oil Prices," "Sex in Advertisements"). In our evaluations we built the queries based on the title (T) and descriptive (D) parts of the topic formulation, corresponding to the official query format in the CLEF evaluation campaigns.

Relevance judgments (correct answers) were supplied by human assessors throughout the various CLEF evaluation campaigns. As shown in Table 1, the entire corpus was not used during all the evaluation campaigns and thus pertinent articles had to be searched in different parts of the corpus. For example, Topics #201 to #250 were created in 2004 and responses resulting from searches in the *Glasgow Herald* (1995) collection, a subset representing 56,472 documents. Of the 50 queries originally available in 2004, we found that only 42 returned at least one correct answer.

In all, 26 queries were removed because there were no relevant documents in the corpus, meaning only 284 (310 minus 26) topics were used in our evaluation. Upon an inspection of these relevance assessments, the average number of correct responses for each topic was 22.46 (standard deviation: 28.9, median: 11.5), with Topic #254 ("Earthquake Damage") obtaining the greatest number of relevant documents (229).

<NUM> C062 </NUM>
<EN-TITLE> Northern Japan Earthquake </EN-TITLE>
<EN-DESC> Find documents that report on an earthquake on the east coast of Hokkaido, northern Japan, in 1994. </EN-DESC>
<EN-NARR> Documents describing an earthquake with a magnitude of 7.9 that shook Hokkaido and other northern Japanese regions in October 1994 are relevant. Also of interest are tidal wave warnings issued for Pacific coastal areas of Hokkaido at the time of the earthquake. Documents reporting any other earthquakes in Japan are not relevant. </EN-NARR>
...
<NUM> C062 </NUM>
<EN-TITLE>
<TERM ID="C062-1" LEMA ="northern" POS = "NNP">
<WF> Northern </WF>
<SYNSET SCORE="1" CODE ="05210354-n"/> </TERM >
<TERM ID ="C062-2" LEMA ="japan" POS = "NNP">
<WF> Japan </WF>
<SYNSET SCORE ="0.4451194309593595" CODE ="06520317-n"/>
<SYNSET SCORE ="0.5548805690406404" CODE ="06519251-n"/> </TERM>
<TERM ID ="C062-3" LEMA ="earthquake" POS = "NN">
<WF> Earthquake </WF>
<SYNSET SCORE ="1" CODE ="05526375-n"/> </TERM>
</EN-TITLE> ...

Table 2. Example of a query with and without lemma, WordNet thesaurus number (synset), and part-of-speech (POS) tag

During the Robust track at CLEF-2008, the organizers also provided an extended version of both documents and topic descriptions (for an example see the bottom part of Table 2) with additional information that could be used to verify whether word-sense disambiguation (WSD) might improve retrieval effectiveness. To achieve this objective, each surface word (after the label <WF>) was preceded by its corresponding lemma (under the tag <TERM>, a value placed after the keyword LEMMA) with its corresponding part-of-speech (POS) tag. The latter information was given according to a variant of the Penn Treebank tag set (Marcus *et al*. 1993). As seen in our example, the tag "NN" was used to indicate noun, "NNP" for proper noun. With the lemma information the morphological analysis results becomes available and therefore a stemming procedure is no longer needed.

As shown in Table 2 synset number(s) are placed after the string corresponding to the surface word, and linking it to the entry in the WordNet thesaurus (version 1.6) (information given after the tag <SYNSET>). This entry could be unique (as in our example with the word “whale”). For a proper noun (e.g., personal, geographic or product name), no pertinent entry in the WordNet thesaurus can be found and the corresponding synset information is thus not given. Finally, a term may belong to different synsets (the noun “reserve” belongs to three synsets). In such cases, each possible entry is preceded by probability estimation that the corresponding synset is the correct one.

Not all of this information is introduced manually. The MXPOST (Maximum Entropy POS Tagger[1]) (Ratnaparkhi 1996) identifies the part-of-speech of each word, and then the corresponding lemma is extracted using the JWNL (Java WordNet Library), an API used to provide easy access to the WordNet relational thesaurus. Based on this information along with local collocations and surrounding words, the NUS-PT WSD system (Chan *et al*. 2007) disambiguates the word-type based on Vector Support Machine (VSM) approach trained with the SemCor corpus, as well as other training examples extracted from parallel texts serving as training data. In a related study with WordNet, Voorhees (1993) attaches only a single synset number to each noun, using the most frequently occurring synset number found in the surrounding text, should there be multiple possibilities.

4 IR Models

To evaluate various stemming strategies with respect to different IR models, we first used the classical *tf idf* model wherein the weight attached to each indexing term was the product of its term occurrence frequency (tf_{ij} for indexing term t_j in document d_i) and the logarithm of its inverse document frequency ($idf_j = \log(n/df_j)$). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights (Manning *et al*. 2008). This IR model was used for example in (Voorhees 1993) or (Hull 1996).

To complement this vector-space model, we implemented certain probabilistic models, such as the Okapi (or BM25) approach (Robertson *et al*. 2000), and two models derived from *Divergence from Randomness* (DFR) paradigm (Amati & van Rijsbergen 2002) wherein the two information measures formulated below are combined:

$$w_{ij} = Inf^1_{ij} \cdot Inf^2_{ij} = -\log_2[Prob^1_{ij}] \cdot (1 - Prob^2_{ij}) \quad (1)$$

in which $Prob^1_{ij}$ is the probability of finding by pure chance the tf_{ij} occurrences of the term t_j in a document. On the other hand, $Prob^2_{ij}$ is the probability of encountering a new occurrence of term t_j in the document, given that tf_{ij} occurrences of this term had already been found. To calculate these two probabilities, we used the $I(n_e)C2$ model, based on the following estimates:

$$Prob^1_{ij} = \left(\frac{n+1}{n_e+1}\right)^{tfn_{ij}}$$

$$\text{and } Prob^2_{ij} = 1 - \left(\frac{tc_j + 1}{df_j \cdot (tfn_{ij} + 1)}\right) \quad (2)$$

$$\text{with } tfn_{ij} = tf_{ij} \cdot \ln\left(1 + \frac{c \cdot \text{mean dl}}{l_i}\right) \quad \text{and } n_e = n \cdot \left(1 - \left(\frac{n-1}{n_i}\right)^{tc_j}\right)$$

[1] Freely available at http://www.inf.ed.ac.uk/resources/nlp/local_doc/MXPOST.html

where tc_j is the number of occurrences of term t_j in the collection, df_j indicates the number of documents in which the term t_j occurs, n the number of documents in the corpus, l_i the length of document d_i, *mean dl* (= 212), the average document length, and c a constant (fixed empirically at 1.5).

For our second DFR model called DFR-PL2, the implementation of $Prob^1_{ij}$ is given by Equation 3, and $Prob^2_{ij}$ is given by Equation 4, as follows:

$$Prob^1_{ij} = (e^{-\lambda_j} \cdot \lambda_j^{tf_{ij}})/tf_{ij}! \quad \text{with } \lambda_j = tc_j / n \tag{3}$$

$$Prob^2_{ij} = tfn_{ij} / (tfn_{ij} + 1) \tag{4}$$

Finally, we also applied a language model (LM) approach (Hiemstra 2000), known as a non-parametric probabilistic model. Within this language model paradigm, various implementation and smoothing methods might also be considered. In this paper we adopted a model proposed by Hiemstra (2000; 2002) as described in Equation 5 and using the Jelinek-Mercer smoothing method (Zhai & Lafferty 2004), a combined estimate based on both the document ($P[t_j \mid d_i]$) and the entire corpus ($P[t_j \mid C]$).

$$Prob[d_i \mid q] = Prob[d_i] \cdot \prod_{tj \in Q} [\lambda_j \cdot Prob[t_j \mid d_i] + (1-\lambda_j) \cdot Prob[t_j \mid C]]$$

$$\text{with } Prob[t_j \mid d_i] = \left(\frac{tf_{ij}}{l_i} \right)$$

$$\text{and } Prob[t_j \mid C] = \left(\frac{df_j}{lc} \right) \quad \text{with } lc = \sum_{k=1}^{t} df_k \tag{5}$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j indicates the number of documents indexed with the term t_j, and lc is a constant related to the size of the underlying corpus C.

5 Evaluation

In order to measure retrieval performance (Buckley & Voorhees 2005), we adopted the mean average precision (MAP) computed by TREC_EVAL based on a maximum of 1,000 retrieved items. To statistically determine whether or not a given search strategy is statistically better than another, we applied the bootstrap methodology (Savoy 1997), with the null hypothesis H_0 stating that both retrieval schemes produce similar performance. In the experiments presented in this paper, statistically significant differences were detected by applying a two-sided test (significance level α=5%). This null hypothesis would be accepted if two retrieval schemes returned statistically similar means, otherwise it would be rejected.

5.1 IR Models Evaluation

Based on the methodology previously described, the MAP obtained from applying six stemming approaches to five IR models are shown in Table 3. The second column (labeled "None") lists the retrieval performances obtained when ignoring the stem-

ming stage during the indexing or query processing. The "S-stemmer" column lists the performance obtained by the light stemmer based on three rules (Harman 1991) while the MAP obtained by either Porter's (1980) or Lovins' (1968) stemmer are shown in the next two columns. The SMART system (Salton 1981) also proposes another English language stemmer and its evaluation is shown in the sixth column. Finally, the last column reports the retrieval performance obtained by applying a morphological analysis returning the lemma of each surface word.

	Mean Average Precision (MAP)					
	None	S-stemmer	Porter	Lovins	SMART	Lemma
Okapi	**0.4345**	0.4648†	0.4706†	0.4560 ‡	0.4755†	0.4663†
DFR-PL2	0.4251	0.4553†	0.4604†	0.4499†‡	0.4634†	0.4608†
DFR-I(n_e)C2	0.4329	**0.4658**†	**0.4721**†	**0.4565** ‡	**0.4783**†	**0.4671**†
LM	0.4240	0.4493†	0.4555†	0.4389 ‡	0.4568†	0.4444†
tf idf	0.2669	0.2811†	0.2839†	0.2650 ‡	0.2860†	0.2778†
Average	0.4291	0.4588	0.4647	0.4503	0.4685	0.4597
% change		+6.9%	+8.3%	+4.9%	+9.2%	+7.1%

Table 3. Mean average precision (MAP) of various IR models and different stemmers (284 TD queries)

In Table 3 and in the following tables, the best performance under a given condition is depicted in bold. Using this performance as a baseline, we then underlined those MAP values (in the same column) depicting statistically significant differences. Except for the second column, the DFR-I(n_e)C2 model always obtains the best results, statistically outperforming the classical *tf idf* vector-space model or the language model (LM) scheme, from a statistical point of view. The same is usually true for a DFR-PL2 variant when compared to the best model. On the other hand, the MAP differences between Okapi and DFR-I(n_e)C2 are never statistically significant, implying that these two probabilistic models tend to perform at the same level.

When comparing two retrieval schemes, each overall statistical measure, such as the MAP may hide performance irregularities among certain queries. For example when comparing the DFR-I(n_e)C2 model with the classical *tf idf* model, we found that the DFR-I(n_e)C2 model performed better for 245 queries, while for 27 queries the classical *tf idf* provided better AP and for the remaining 12 queries, both models maintained the same retrieval performances. To understand performance differences between these two models, we examined the largest difference obtained with Topic #62 (“Northern Japan Earthquake”). In this case the DFR-I(n_e)C2 model obtained an AP of 1.0 while the classical *tf idf* model obtains an AP of 0.0062, . For this query the *tf idf* model’s poor performance resulted from the fact that for some query terms, the term frequency was relatively high in the query. In this model the documents at the higher ranks often contained one single search term with also a very high *tf* value (e.g., “Japan” with *tf* = 125 or *tf* = 85). Within the *tf idf* model, this property ranked these documents higher compared to articles containing more query terms but having lower frequencies. For example a relevant document having the search terms “earthquake”

(*tf* = 3), “Japan” (*tf*=1) and “report”(*tf* = 1) was ranked in the first position with the DFR-I(n_e)C2 but only in the 213th with the *tf idf* method.

5.2 Differences Between Stemming and Non-Stemming Approaches

Table 3 also lists the results of following a verification of whether a stemmer’s application might improve retrieval performance when compared to a search strategy ignoring this type of word normalization procedure. As shown in the second to last row of Table 3, we computed the average performance achieved by each of the five retrieval models in order to obtain an overview of the performance of each stemming approach. The last row shows the percent change when compared to an approach ignoring the stemming procedure. This value shows that the SMART stemmer obtains the highest average value of 0.4685 (or a relative improvement of +9.2% over the non-stemming method). The difference is rather small when comparing the SMART stemmer with other approaches such as that of Porter (0.4647) or when applying the morphological analysis (under the label "Lemma", 0.4597). In fact for 159 queries, the S-stemmer improved retrieval performances while for the other 93, the non-stemming approach resulted in a better AP.

To verify whether these differences were statistically significant, we chose the performance labeled "None" as baseline. When using a stemmer, if retrieval effectiveness was statistically significant, we placed the symbol "†" after the corresponding MAP value. For example when using the DFR-I(n_e)C2 IR model without stemming, we obtained a MAP of 0.4329 compared to 0.4658 when applying the S-stemmer. This difference was statistically significant, and was denoted by a "†" after the MAP value of 0.4658. Except for the Lovins' stemmer, all stemming approaches performed significantly better than the non-stemming approach. The Lovins' stemmer tended to produce retrieval performances that were statistically similar to a non-stemming approach.

In order to obtain an overview of the precise effect of stemming, we analyzed concrete examples. With the DFR-I(n_e)C2 model, we saw that Topic #306 (“ETA Activities in France”) retrieved a single relevant document and obtained an AP of 0.333 without stemming, and after applying the S-stemmer, the AP was 1.0. The difference was due to the term “activities” which after stemming is reduced to “activity”. The relevant document contains the term “activity” three times and “activities” two times. When conflated under the same stem, this search term was helpful in ranking the relevant article in first position after stemming.

Topic #98 (“Films by the Kaurismäkis”) on the other hand retrieved only one single relevant document, with an AP of 1.0 before stemming and 0.5 after applying the S-stemmer. In this case, the single relevant document contains the term “films” 9 times and “film” 12 times. After applying the S-stemmer, a non-relevant document was ranked higher than the relevant article.

As another example, we could compare retrieval effectiveness using the DFR-I(n_e)C2 model and the SMART stemmer with the non-stemming approach (0.4329 vs. 0.4783). For Topic #180 (“Bankruptcy of Barings”) using the SMART stemmer, the AP was 0.0082 while without stemming the AP was 0.7652. In this case, the word

"Barings" was stemmed to "bare" which hurt retrieval performance. For Topic #63 ("Whale Reserve") using the stemmer, the AP was 1.0 meaning that the single relevant document was placed in the first position. Without stemming the AP was only 0.0286 and the single relevant document was ranked 35th. Using the SMART stemmer, the word "Antarctic" occurring in the topic description was stemmed to "antarct" which would then match the word "Antarctica" appearing in the relevant document.

Similar findings can be obtained with other IR models, such as the Okapi. For Topic #198 ("Honorary Oscar for Italian Directors") returning a single relevant document obtains an AP of 0.5 without stemmer and 1.0 with the SMART stemmer. Important changes in the query included the search terms "Honorary" (reduced to "honor") and "awarded" (stemmed to "award").

5.3 Algorithmic Stemmers or Morphological Analysis

As shown in the last line of Table 3, the percent of change obtained when comparing an approach ignoring the stemming procedure was rather similar across the different algorithmic stemmers or when applying the morphological analysis (under the label "Lemma"). To verify whether these differences are statistically significant, we selected the retrieval performance achieved with the SMART stemmer as a baseline. When using a stemmer, if the retrieval effectiveness was statistically significant, we indicated this by adding the symbol "‡" after the corresponding MAP value. For example, when using the DFR-I(n_e)C2 IR model with the SMART stemmer, we obtained a MAP of 0.4783 compared to 0.4565 when applying the Lovins' stemmer. Its statistically significant difference was denoted by an "‡" after the MAP 0.4565. Performance differences were also significant for the other IR models, leading to the conclusion that the Lovins' stemmer results in lower performance levels than the SMART stemmer.

During a query-by-query performance analysis comparing the Lovins and SMART stemmer, for Topic #98 ("Films by the Kaurismäkis") the AP was 0.1429 with Lovins' stemmer, while for the SMART stemmer the AP was 1.0. The single relevant document was ranked in the seventh position with the Lovins' stemmer and in the first by the SMART method. An analysis of the various stems produced by the two stemmers, shows that with the Lovins method the stems were "ak" and "mik" while with SMART stemmer they were "aki" and "mika". These two names came from the descriptive part of the topic formulation ("Search for information about films directed by either of the two brothers Aki and Mika Kaurismäki"). The stems produced by the Lovins' method were shorter and thus matched other terms in the rest of the collection.

On the other hand, Topic #231 ("New Portuguese Prime Minister") obtained an AP of 1.0 with the Lovins stemmer and only 0.5 with the SMART stemmer. In this case, the single relevant item contained the noun "elections," which the Lovins method reduced to the same stem as the adjective "electoral" appearing in the descriptive part of the topic. With the SMART stemmer, the noun and the adjective did not conflate under the same form and thus the relevant item was not ranked first on the list.

The main conclusion therefore is that there are no statistically significant differences between efficient algorithmic stemmers such as Porter, SMART or S-stemmer

and the morphological analysis which returns the dictionary entry (or lemma) for each surface word. Thus a light suffix-stripping algorithm such as the S-stemmer can achieve, *in mean*, a retrieval performance comparable to both the more aggressive algorithmic stemmers (Porter, SMART) or systems based on advanced natural language processing that correctly removes all inflexional suffixes.

5.4 Morphological Analysis, Part-Of-Speech and Thesaurus

In Table 4, we reported the MAP obtained using morphological analysis to produce the corresponding lemma for each surface word (same values as last column of Table 3). In the third column we increased the document score when lemma common to the query and the retrieved item had the same part-of-speech (POS) tag. This feature could be useful in determining the precise meaning attached to a form. In the English language, the same term may have different meanings, depending on its part-of-speech, such as "lean" as adjective (thin, mean lacking charm) or verb (to recline or bend). The word "face" (or "form," "bank," "stem") may have a different meaning as a noun (a happy face) or as a verb (to deal with). To do so, for each indexing term a string composed of the term and its POS tag (e.g., with the adjective "alien", we added "alieanJJ" in which "JJ" is the POS tag for the adjectives (Marcus *et al.* 1993)).

In the fourth column we listed retrieval performances achieved by increasing the document score when query and documents had the same synset numbers. To do so, we added all synset numbers attached to an article or a query to its corresponding surrogate. Finally in the last column of Table 4, we combined the two previous enhancements, which in turn in turn assigned more weight when the terms common to both the retrieved records and the query were also the same POS and shared the synset numbers.

	Mean Average Precision			
	Lemma	Lemma & POS	Lemma & Synset	Lemma & POS & Synset
Okapi	0.4663	0.4720†	0.4395†	0.4482†
DFR-PL2	0.4608	0.4634	0.4365†	0.4433†
DFR-I(n_e)C2	**0.4671**	**0.4740**†	**0.4665**	**0.4705**
LM	0.4444	0.4562†	0.4342†	0.4458
tf idf	0.2778	0.2879†	0.2834	0.2888†
Average	0.4597	0.4664	0.4442	0.4520
% change		+1.5%	-3.4%	-1.7%

Table 4. Mean average precision (MAP) for various IR models and different morphological analysis variants (284 TD queries)

The results depicted in Table 4 confirm the conclusions we had drawn regarding the data shown in Table 3. The best IR model was still the DFR-I(n_e)C2 and the performance differences were always statistically significant (MAP underlined in Table 4) with both the LM or *tf idf* models.

Compared to the morphological analysis only (performance under the label "Lemma" in Table 4 used as baseline), we might use the POS information to partly remove the ambiguity attached to search keywords. This additional information slightly improves the MAP and the performance differences are always significant (MAP followed by the symbol "†" in Table 4), except for the DFR-PL2 model. For example with the DFR-I(n_e)C2, the POS data increased the AP for 138 queries, decreased it for 98 (and for the remaining 48, we obtained the same performance). Using this IR model and Topic #217 (“AIDS in Africa”), the AP was of 0.1944 under "Lemma" yet when we added the POS information, the AP increased to 0.5526. When inspecting the corresponding query, we first found that the stemming converted “AIDS” into “aid,” and this increased the possibility of matches. When accounting for the POS tag, the stem “aid” was tagged as a proper noun, and thus improved the ranking of articles containing this abbreviation compared to document containing either the singular noun “aid” or the plural form “aids”.

Adding the thesaurus numbers to document and query representations (retrieval performance listed under the label "Lemma & Synset") tended to slightly decrease the MAP. For the three IR models, the differences were even statistically significant. With the Okapi model for example, Topic #76 (“Solar Energy”) obtained an AP of 0.663 under the "Lemma" condition but only obtained an AP of 0.0722 under "Lemma & Synset". In this case, the description part of the topic contained the form “is” and “being” twice. The corresponding lemma “be” belongs to the ten synsets added in the query surrogate (with a frequency of three). For each document containing a verbal form related to the verb “to be”, we will thus have ten query matches through the sysnset numbers, thus rendering discrimination between relevant and non-relevant items more difficult.

5.5 Stopword Lists

Finally, we have compared the retrieval effectiveness of the various IR models using different stopword lists. These lists contain words serving no purpose for retrieval purposes, but very frequently found in the documents. Upon removing these terms, each match between a query and a document would thus be based on good indexing terms. In other words, retrieving a document because it contains words such as “the,” “has,” “in,” or “your” in the corresponding request does not constitute an intelligent search strategy. These non-significant words represent noise, and may actually damage retrieval performance because they do not discriminate between relevant and non-relevant documents. Hopefully we would also reduce the inverted file’s size, by from 30% to 50%.

In the second column of Table 5, we reported the retrieval performance achieved using the S-stemmer with the SMART stopword list containing 571 entries. This list may be viewed as relatively large but Fox (1990) also suggested a relatively long list with 421 words. Next, we used the same stemming approach but without any stopword list. The performance differences were small (around 1%) for the last three retrieval models when compared to SMART stopword list but relatively large for the Okapi (a relative decrease of -26.8%) and DFR-PL2 (-30.1%) approaches. In the last

column of Table 5 we used the short stopword list composed of nine words ("an," "and," "by," "for," "from," "of," "the," "to," "with") found in the DIALOG search engine (Harter 1986). The average difference with the SMART stopword list is rather small (-0.6%) tending to indicate that the important point is to ignore only a short number of very frequent terms without any important meanings.

	Mean Average Precision		
	SMART	None	Short
Okapi	0.4648	0.3403†	0.4581
DFR-PL2	0.4553	0.3185†	0.4526
DFR-I(n_e)C2	**0.4658**	**0.4661**	**0.4665**
LM	0.4493	0.4433	0.4462
tf idf	0.2811	0.2831	0.2830
Average	0.4588	0.3921	0.4559
% change		-14.5%	-0.6%

Table 5. Mean average precision (MAP) for various stopword lists using the S-stemmer (284 TD queries)

When applying the statistical tests (significant differences are underlined while best performances are shown in bold), we can still conclude that the DFR-I(n_e)C2 model is the best. When using the retrieval performance with the SMART stopword list as baseline (second column), we found two cases in which the performance differences are statistically significant (MAP value followed by the symbol "†"). For example, when using the Okapi model, the MAP using the SMART stopword list is 0.4648, yet it decreases significantly to 0.3403 when accounting for all frequently occurring words (performances listed under the label "None"). Clearly, the performance achieved by either the Okapi or DFR-PL2 is sensitive to the presence of very frequent words.

Through analyzing an example, we discover the main reasons for this phenomenon. Based on the Okapi model and applying the SMART stopword list we obtained better retrieval performances for 223 queries while for 37, indexing all terms produced better AP (for the remaining 24 queries the same AP was produced). From an analysis of the extreme cases, we saw that Topic #136 ("Leaning Tower of Pisa") obtained an AP of 1.0 with SMART stopword list yet the AP was 0.0 when we accounted for all word forms. In the underlying query, the presence of many stopwords (e.g., "of," "the," "is," "what") ranked many non-relevant documents higher than the single relevant document.

On the other hand, with Topic #104 ("Super G Gold medal") we obtained an AP of 0.6550 when ignoring the stopword list yet an AP of 0.4525 with a stopword list. In this case, the search term "G" included in the stopword list was removed during the query processing. After this stopword removal, the final query was more ambiguous ("super gold medal") and could not rank the articles higher up on the result list.

6 Conclusion

It has been recognized that the stemming procedure is an important component in modern IR systems and an inappropriate stemmer may generate unexpected results to be presented to the user (Buckley 2007; Savoy 2007). Contrary to previous evaluations based only on the classical *tf idf* vector-space model, we have shown that the same problem occurs with modern probabilistic models (e.g. Okapi, language model or *Divergence from Randomness* (DFR) paradigm), which perform significantly better than the *tf idf* approach.

Using a large set of queries (284) extracted from the CLEF test-collections, we also demonstrated that some algorithmic stemmers or morphological analyses tend *in mean*, to result in similar retrieval performances, at least for the English language. For medium-sized queries, the enhancement is around 7% greater than a search technique without stemming. For a language having a rather simple inflectional structure this mean improvement is relatively high, as compared to other languages. Using similar test-collections (newspapers articles and comparable queries) Tomlinson (2004) obtained the following average improvements after stemming: +4% for Dutch, +7% Spanish, +9% French, +15% Italian, +19% German, +29% Swedish and +40% Finnish.

Among the various stemming approaches suggested for the English language, we found that the SMART (Salton 1971), Porter (1980) and S-stemmer (Harman 1991) methods as well as morphological analyses returning the corresponding lemma resulted in similar performance levels. Retrieval performance for the latter is significantly better than a non-stemming approach or the Lovins' stemmer (1968). In our opinion this latter method removes too many final letters and thus is too aggressive, resulting in relatively short stems having high document frequencies. The examples presented in Section 5 demonstrate some of these aspects.

When comparing stemming procedures, in our opinion it is important to consider the final user. A non-stemming or a light stemming approach is better understood than a more aggressive approach returning unexpected results. For this same reason, in the English language we suggest using the S-stemmer (Harman 1991) which only removes the plural form associated with English nouns.

We also tried to improve retrieval effectiveness by considering part-of-speech (POS) information and thesaurus class numbers. Compared to the morphological stemmer, accounting for the POS information will significantly improve the MAP. The presence of the synset (or thesaurus class) numbers does not however significantly modify mean retrieval performance, at least as implemented in this paper.

Finally it must be recognized that stopword lists were developed on the basis of certain arbitrary decisions (Fox 1990). This is the case for example in commercial information systems, which tend to adopt a very conservative approach involving only a few stopwords. According to our evaluations, the presence of a short stopword list containing around 10 terms produces retrieval effectiveness similar to that of longer stopword lists with 571 terms. Thus, not removing these very frequent terms with no real meaning may significantly hurt retrieval performance for some IR models (e.g., Okapi and DFR-PL2 in our experiments), when compared even to short stopword lists.

Acknowledgments
This research was supported in part by the Swiss NSF under Grant #200021-113273.

References

Amati, G., & van Rijsbergen, C.J. (2002). Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM-Transactions on Information Systems*, 20(4), 357-389.

Buckley, C. (2007). Why current IR engines fail. In *Proceedings ACM-SIGIR'2004*, Sheffield (UK), 584-585.

Buckley, C., & Voorhees, E.M. (2005). Retrieval system evaluation. In E.M. Voorhees, D.K. Harman (Eds): *TREC. Experiment and Evaluation in Information Retrieval* (pp. 53-75). The MIT Press, Cambridge (MA).

Chan, Y.S., Ng, H.T., & Zhong, Z. (2007). NUS-PT: Exploiting Parallel Texts for Word Sense Disambiguation in the English All-Words Tasks. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (SemEval 2007), Prague, 253–256.

Fellbaum, C. (1998). *WordNet. An Electronic Lexical Database*. The MIT Press, Cambridge (MA).

Fox, C. (1990). A stop list for general text. *ACM-SIGIR Forum*, 24, 19-35.

Harman, D. (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1), 7-15.

Harter, S.P. (1986). *Online Information Retrieval. Concepts, Principles, and Techniques.* Academic Press, San Diego (CA).

Hiemstra, D. (2000). Using language models for information retrieval. CTIT Ph.D. Thesis.

Hiemstra, D. (2002). Term-specific smoothing for the language modeling approach to information retrieval. The importance of a query term. In *Proceedings of the ACM-SIGIR*, Tempere, 35-41.

Hull, D. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1), 70-84.

Krovetz, R (1993). Viewing morphology as an inference process. In *Proceedings of the ACM-SIGIR'93*. Pittsburgh (PA), 191-202

Lovins, J.B. (1968). Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1), 22-31

McNamee, P., & Mayfield, J. (2004). Character *N*-gram tokenization for European language text retrieval. *Information Retreival Journal*, 7(1-2), 73-97

Marcus, M.P., Santorini, B., & Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2), 313-330.

Peng, F., Ahmed, N., Li, X., & Lu, Y. (2007). Context sensitive stemming for web search. In *Proceedings of the ACM-SIGIR*, Amsterdam, 639-646.

Peters, C., Jijkoun, V., Mandl, T., Müller, H., Oard, D.W., Peñas, A. & Santos, D. (Eds.). (2008). *Advances in Multilingual and Multimodal Information Retrieval*. LNCS #5152, Springer-Verlag, Berlin.

Porter, M.F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130-137.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, 133-142.

Robertson, S.E., Walker, S., & Beaulieu, M. (2000). Experimentation as a way of life: Okapi at TREC. *Information Processing &Management*, 36(1), 95-108.

Salton, G. (1971). *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall Inc., Englewood Cliffs (NJ).

Savoy, J. (1993). Stemming of French words based on grammatical category. *Journal of the American Society for Information Science*, 44(1), 1-9.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management*, 33(4), 495-512.

Savoy J. (2007). Why do successful search systems fail for some topics? In *Proceedings ACM-SAC*, Seoul, 872-877.

Sproat, R. (1992). *Morphology and Computation*. The MIT Press, Cambridge (MA).

Tomlinson, S. (2004). Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003. In *Comparative Evaluation of Multilingual Information Access Systems* (pp. 286-300). LNCS #3237, Springer-Verlag, Berlin.

Voorhees, E.M. (1993): Using WordNet™ to disambiguate word senses for text retrieval. In *Proceedings ACM-SIGIR'93*, Pittsburgh (PA), 171–180.

Xu, J., & Croft, B. (1998). Corpus-based stemming using cooccurrence of word variants. *ACM-Transactions on Information Systems*, 16(1), 61-81.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems*, 22(2), 179-214.

Appendix B

Examples from the GIRT4-DE Collection

```
<DOC>
<DOCNO>GIRT-DE19907042</DOCNO>
<DOCID>GIRT-DE19907042</DOCID>
<TITLE-DE>Vergleichende Studie zu Methoden der Auswertung von
kulturpolitischen Maßnahmen in Europa.
</TITLE-DE>
<AUTHOR>Bontinck, Irmgard</AUTHOR>
<AUTHOR>Angerer, Marie-Luise</AUTHOR>
<PUBLICATION-YEAR>1990</PUBLICATION-YEAR>
<LANGUAGE-CODE>DE</LANGUAGE-CODE>
<CONTROLLED-TERM-DE>Kulturpolitik</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>internationaler Vergleich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Bewertung</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Forschungsbericht</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Erhebungsmethode</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Forschungsansatz</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Österreich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Schweiz</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Jugoslawien</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Frankreich</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Schweden</CONTROLLED-TERM-DE>
<CONTROLLED-TERM-DE>Europa</CONTROLLED-TERM-DE>
<METHOD-TERM-DE>empirisch</METHOD-TERM-DE>
<METHOD-TERM-DE>internationaler Vergleich</METHOD-TERM-DE>
<METHOD-TERM-DE>Aktenanalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Inhaltsanalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Sekundäranalyse</METHOD-TERM-DE>
<METHOD-TERM-DE>Querschnitt</METHOD-TERM-DE>
<CLASSIFICATION-TEXT-DE>spezielle Ressortpolitik
</CLASSIFICATION-TEXT-DE>
<ABSTRACT-DE>Vergleich von fünf Länderstudien zur jeweiligen
 Kulturpolitik. Inwieweit decken sich Anspruch und Realisierung,
welche Sparten werden als Kulturpolitik begriffen und behandelt?
Wie ist die methodische Vorgehensweise der Forschungsberichte?
Vorläufige Schlußfolgerung: der hohe theoretische Anspruch läßt
ich nur partiell einlösen. Die kulturellgesellschaftspolitische
Unterschiedlichkeit der Länder spiegelt sich in der
Art und Weise des untersuchten Feldes wider.</ABSTRACT-DE>
</DOC>
```

FIGURE B.1: Example document from the German GIRT Collection

```
<top lang="de">
<num>10.2452/176-DS</num>
<title>Geschwisterbeziehungen</title>
<desc>Suchen Sie Dokumente, die die Entwicklung von Beziehungen
Zwischen Schwestern und Brüdern näher beschreiben.</desc>
<narr>Alle Dokumente, die die Beziehungen unter Geschwistern
in den verschiedenen Lebenslagen untersuchen sind relevant:
die Rolle von Geschwistern in der Familie, in der Schule, in der
Freizeit beziehungsweise die Veränderung der Beziehungen von der
Kindheit zum Erwachsenen sowie Unterschiede zwischen großen und
kleinen Familien.</narr>

</top>
<top lang="de">
<num>10.2452/177-DS</num>
<title>Arbeitslose Jugendliche ohne Berufsausbildung</title>
<desc>Suchen Sie Veröffentlichungen, die sich auf Jugendliche
beziehen, die arbeitslos sind und keine abgeschlossene
Berufsausbildung haben.</desc>
<narr>Relevante Dokumente geben einen Überblick über den
Umfang und die Probleme von Jugendlichen, die arbeitslos sind und
keine abgeschlossene Berufsausbildung haben. Nicht relevant sind
Dokumente, die sich ausschließlich mit Maßnahmen der Jugendhilfe
und Jugendpolitik beschäftigen.</narr>
</top>
```

FIGURE B.2: Two example topics from the German GIRT collection

```
<entry>
<german>Abbrecher</german>
<german-caps>ABBRECHER</german-caps>
<related-term>Abgänger</related-term>
<related-term>Aussteiger</related-term>
<related-term>drop out</related-term>
<english-translation>drop-out</english-translation>
</entry>
<entry>
<entry>
<german>Vergleich</german>
<german-caps>VERGLEICH</german-caps>
<scope-note-de>nicht im Sinne einer Regelung von
Rechtsstreitigkeiten, dann
Rechtsvergleich;</scope-note-de>
<narrower-term>internationaler Vergleich</narrower-term>
<narrower-term>interkultureller Vergleich</narrower-term>
<narrower-term>Modellvergleich</narrower-term>
<narrower-term>Ost-West-Vergleich</narrower-term>
<narrower-term>Soll-Ist-Vergleich</narrower-term>
<narrower-term>Leistungsvergleich</narrower-term>
<narrower-term>Kostenvergleich</narrower-term>
<narrower-term>Systemvergleich</narrower-term>
<narrower-term>Theorievergleich</narrower-term>
<narrower-term>regionaler Vergleich</narrower-term>
<narrower-term>Methodenvergleich</narrower-term>
<english-translation>comparison</english-translation>
</entry>
<entry>
<german>Verkehrsbelastung</german>
<german-caps>VERKEHRSBELASTUNG</german-caps>
<broader-term>Belastung</broader-term>
<broader-term>Umweltbelastung</broader-term>
<english-translation>traffic load</english-translation>
</entry>
```

FIGURE B.3: Example entries from the GIRT thesaurus

Appendix C

Examples from the Blogs06 Collection

```
<DOC>
<DOCNO>BLOG06-20060221-002-0000076733</DOCNO>
<DATE_XML>2006-02-14T10:39:00-05:00</DATE_XML>
<FEEDNO>BLOG06-feed-001593</FEEDNO>
<FEEDURL>http://covertoperations.blogspot.com/atom.xml#</FEEDURL>
<BLOGHPNO></BLOGHPNO>
<BLOGHPURL></BLOGHPURL>
<PERMALINK>
http://covertoperations.blogspot.com/2006/02/treason.html#
</PERMALINK>
<DOCHDR>
http://covertoperations.blogspot.com/2006/02/treason.html#
0.0.0.0 200639999 38527
Connection: close
Date: Thu, 09 Mar 2006 16:09:53 GMT
Accept-Ranges: none
ETag: "6bdbfe-9525-43f208d8"
Server: Apache
Vary: Accept-Encoding
Content-Length: 38181
Content-Type: text/html
Last-Modified: Tue, 14 Feb 2006 16:44:08 GMT
Client-Date: Thu, 09 Mar 2006 16:09:09 GMT
Client-Peer: 66.102.15.101:80
Client-Response-Num: 1
Test: %{HOSTNAME}e
</DOCHDR>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en"
lang="en">
……
<title>Humint Events Online: Treason</title>
……
<div id="content">
<h1 id="blog-title">
<a href="http://covertoperations.blogspot.com">
Humint Events Online
</a>
</h1>
<p id="description">The 9/11 hijacking attacks were very likely
Facilitated by a rogue group within the US government that created
an Islamic terrorist "Pearl Harbor" event as a catalyst for the
military invasion of Middle Eastern countries. This weblog will
explore the incredibly strange events of 9/11/01, and other issues
of US government responsibility.</p>
</div>
……
</html>
</DOC>
```

FIGURE C.1: Example document from the Blogs06 Collection

```
<top>
<num> Number: 851
<title> "March of the Penguins"
<desc> Description:
Provide opinion of the film documentary "March of the Penguins".
<narr> Narrative:
Relevant documents should include opinions concerning the film
documentary "March of the Penguins". Articles or comments about
penguins outside the context of this film documentary are not
relevant.
</top>

<top>
<num> Number: 852
<title> larry summers
<desc> Description:
Find opinions on Harvard President Larry Summers' comments on
Gender differences in aptitude for mathematics and science.
<narr> Narrative:
Statements of opinion on Summers' comments are relevant.
Quotations of Summers without comment or references to Summers'
Statements without discussion of their content are not relevant.
Opinions on innate gender differences without reference to Summers'
statements are not relevant.
</top>
```

FIGURE C.2: Two example topics from the Blogs06 Collection

Appendix D

Examples from the Genomics Collection

```
<DOC>
<PMID>10605436</PMID>
<DA>20000107</DA>
<DCOM>20000107</DCOM>
<LR>20031114</LR>
<IS>0021-9525</IS>
<VI>76</VI>
<IP>2</IP>
<DP>1978 Feb</DP>
<TI>Concerning the localization of steroids in centrioles and basal
bodies by im munofluorescence.</TI>
<PG>255-60</PG>
<AB>Specific steroid antibodies, by the immunofluorescence
technique, regularly reveal fluorescent centrioles and cilia-
bearing basal bodies in target and nonta rget cells. Although the
precise identity of the immunoreactive steroid substance e has not
yet been established, it seems noteworthy that exogenous steroids
can be vitally concentrated by centrioles, perhaps by exchange with
steroids already present at this level. This unexpected
localization suggests that steroids may affect cell growth and
differentiation in some way different from the two-step receptor
mechanism.</AB>
<AD>Istituto di Anatomia e Istologia Patologica, Universita di
Ferrara, Italy.</AD>
<FAU>Nenci, I</FAU>
<AU>Nenci I</AU>
<FAU>Marchetti, E</FAU>
<AU>Marchetti E</AU>
<LA>eng</LA>
<PT>Journal Article</PT>
<PL>UNITED STATES</PL>
<TA>J Cell Biol</TA>
<JID>0375356</JID>
<RN>0 (Steroids)</RN>
<SB>IM</SB>
<MH>Animals</MH>
<MH>Centrioles/*ultrastructure</MH>
<MH>Cilia/ultrastructure</MH>
<MH>Female</MH>
<MH>Fluorescent Antibody Technique</MH>
<MH>Human</MH>
<MH>Lymphocytes/*cytology</MH>
<MH>Male</MH>
<MH>Organelles/*ultrastructure</MH>
<MH>Rats</MH>
<MH>Rats, Sprague-Dawley</MH>
<MH>Respiratory Mucosa/cytology</MH>
<MH>Steroids/*analysis</MH>
<MH>Trachea</MH>
<SO>J Cell Biol 1978 Feb;76(2):255-60. </SO>
</DOC>
```

FIGURE D.1: Example document from the 2004/2005 Genomics collection

```
<DOC>
<html>
<body>
<H2>
Metabolic Studies with Radioactive Carbon, <SUP>11</SUP>
C: A. Baird Hastings
</H2>
<STRONG>
</NOBR><NOBR>Robert D. Simoni</NOBR>,
<NOBR>Robert L. Hill</NOBR>,  and
<NOBR>Martha Vaughan</NOBR>
</STRONG><P>
<p>
<B>Metabolism of Lactic Acid Containing Radioactive Carboxyl
Carbon<SUP> </SUP><BR>(Conant, J. B., Cramer, R. D., Hastings,
A. B., Klemperer, F.<SUP> </SUP>W., Solomon, A. K., and
Vennesland, B. (1941) <I>J. Biol. Chem.</I><SUP> </SUP>137,
557&#150;566)</B><SUP> </SUP><P>
<B>The Participation of Carbon Dioxide in the Carbohydrate
Cycle<SUP> </SUP><BR>(Soloman, A. K., Vennesland, B.,
Klemperer, F. W., Buchanan,<SUP> </SUP>J. M., and Hastings,
A. B. (1941) <I>J. Biol. Chem.</I> 140, 171&#150;182)</B><SUP>
</SUP><P>
A. Baird Hastings (1895&#150;1987) was born in Dayton, Kentucky
<SUP> </SUP>but lived in Indianapolis until he went to college
<A HREF="#REF1">1</A>).<SUP><A NAME="RFN1"></A><SUP>
<A HREF="#FN1">1</A></SUP></SUP> A high<SUP> </SUP>school teacher,
Ella Marthens, was strongly influential and<SUP> </SUP>encouraged
 his interests in biology and in going to college.<SUP> </SUP>He
chose the University of Michigan and decided to major in chemical
 engineering primarily because after graduation he would be able
to get a job quickly and help support his family. After a time at
 Michigan, Hastings gravitated toward physical chemistry and was
asked by Professor Floyd Bartell to serve as his course assistant.
 As graduation approached, Hastings was prepared to get a job but
 was encouraged by Bartell to consider graduate school. He
entered the University of Michigan graduate school in 1916, but
with the beginning of World War I his graduate training was
interrupted and his advisor, Bartell, joined the Chemical Warfare
 Service. Hastings' persistent efforts to enlist in the military
were rejected primarily because he was under weight. He took a
job as a "sanitary chemist" with the Public Health Service to
study fatigue, convinced that it would be a contribution to the
 war effort. It was a notable opportunity because it introduced
 Hastings to the study of physiology, which eventually became
his life's work.<P>
……
</BODY>
</HTML>
```

FIGURE D.2: Example document from the 2006/2007 Genomics collection

```
<TOPIC>
<ID>1</ID>
<TITLE>Ferroportin-1 in humans</TITLE>
<NEED>Find articles about Ferroportin-1, an iron transporter, in
humans.</NEED>
<CONTEXT>Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1;
HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1;
MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled
divalent metal ion transporters), member 3) may play a role in
iron transport.
</CONTEXT>
</TOPIC>

<TOPIC>
<ID>100</ID>
<METHOD>How to "open up" a cell through a process called
"electroporation"</METHOD>
</TOPIC>

<160>What is the role of PrnP in mad cow disease?

<200>What serum [PROTEINS] change expression in association with
high disease
activity in lupus?
```

FIGURE D.3: Example topics example from the Genomics collection Genomics

Appendix E

Divergence from Randomness – Formulae

Let be

- cf the number of occurences of term t in the collection
- tf the term frequency of the term t in the document D
- N the number of documents in the collection
- df the number of documents containing term t

Basic Randomness Models

Poisson approximation of the binomial model (P)

$$Inf_1(tf) = tf \cdot \log_2(\frac{tf}{\lambda}) + (\lambda + \frac{1}{12 \cdot tf} - tf) \cdot \log_2 e + 0.5 \cdot \log_2(2\pi \cdot tf)$$

where

$$\lambda = \frac{cf}{N}$$

Approximation of the binomial model with the divergence (D)

$$Inf_1(tf) = cf \cdot D(\phi, p) + 0.5 \log_2(2\pi \cdot tf(1 - \phi))$$

where

$$\phi = \frac{tf}{cf}$$

$$p = \frac{1}{N}$$

and

$$D(\phi, p) = \phi \cdot \log_2(\frac{\phi}{p}) + (1 - \phi) \cdot \log_2(\frac{1 - \phi}{1 - p})$$

Geometric as limiting form of Bose-Einstein (G)

$$Inf_1 = -\log_2(\frac{1}{1 + \lambda}) - tf \cdot \log_2(\frac{\lambda}{1 + \lambda})$$

where

$$\lambda = \frac{cf}{N}$$

Limiting form of Bose-Einstein (B_E)

$$Inf_1(tf) = -\log_2(N - 1) - log_2(e) + f(N + cf - 1, N + F - tf - 2) - f(F, F - tf)$$

where

$$f(n, m) = (m + 0.5) \cdot log_2(\frac{n}{m}) + (n - m)\log_2(n)$$

Inverse document frequency ($I(n)$)

$$Inf_1(tf) = tf \cdot \log_2(\frac{N + 1}{df + 0.5})$$

Mixture of Poisson and inverse document frequency ($I(n_e)$)

$$Inf_1 = tf \cdot \log_2(\frac{N + 1}{n_e + 0.5})$$

where

$$n_e = N \cdot (1 - (\frac{N - 1}{N})^{cf})$$

Approximation of I(n_e) ($I(F)$)

$$Inf_1(tf) = tf \cdot \log_2(\frac{N + 1}{cf + 0.5})$$

First Normalization

Laplace's law of succession (L)

$$Prob_2(tf) = \frac{tf}{tf + 1}$$

$$Inf_2 = \frac{1}{tf + 1}$$

Ratio of two Bernoulli processes (B)

$$Inf_2 = \frac{cf + 1}{df \cdot (tf + 1)}$$

Second Normalization

If a second normalization is applied, tf in the previous formulae is replaced by tfn.

Uniform distribution of the term frequency ($H1$)

$$tfn = tf \cdot \frac{avdl}{l}$$

where $avdl$ is the average document length and l the length of the document.

The term frequency density is inversely related to the length ($H2$)

$$tfn = tf \cdot \log_2(1 + \frac{c \cdot avdl}{l})$$

where c is a constant fixed depending on the collection.

Special Case

As a special case we used for the first normalization a Ratio of two Bernoulli processes with second normalization

$$tfn = tf \cdot \ln(1 + \frac{c \cdot avdl}{l})$$

We reference to this first normalization as C (instead of B, for example $I(n)C2$)

Appendix F

Published Papers

2009

- Ljiljana Dolamic, Claire Fautsch, Jacques Savoy
  UniNE at CLEF 2009: Persian ad hoc and CLEF-IP
  In *Working Notes for the CLEF 2009 Workshop* Corfum Greece, September 30 - October 2, 2009.

- Claire Fautsch, Jacques Savoy
  Adapting the *tf idf* Vector-Space Model to Domain-Specific Information Retrieval
  To appear in *Proceedings of the 2010 ACM Symposium on Applied Computing (SAC)*, Sierre, Switzerland, March 22-26, 2010.

- Claire Fautsch
  Challenges in Domain Specific Information Retrieval
  Technical Report BENEFRI Summer School 2009

- Ljiljana Dolamic, Claire Fautsch, Jacques Savoy
  UniNE at CLEF 2008: TEL, Persian and Robust IR
  In *Carol Peters et al. (Eds.): Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, Revised Selected Papers*, Lecture Notes in Computer Science, Springer Berlin/Heidelberg *to appear*

- Claire Fautsch, Ljiljana Dolamic, Jacques Savoy
  UniNE at Domain-Specific IR - CLEF 2008: Scientific Data Retrieval: Various Query Expansion Approaches
  In *Carol Peters et al. (Eds.): Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, Revised Selected Papers*, Lecture Notes in Computer Science, Springer Berlin/Heidelberg *to appear*

- Claire Fautsch, Jacques Savoy
  Comparison between manually and automatically assigned descriptors based on a German bibliographic collection
  In *Proceedings of 6th International Workshop on Text-based Information Retrieval* August 31 - September 4 2009, Linz, Austria (peer reviewed)

- Claire Fautsch, Jacques Savoy
  Algorithmic stemmers or morphological analysis? An evaluation
  In *Journal of the American Society for Information Science and Technology*, Volume 60, pages 1616-1624, Wiley InterScience, 2009

- Claire Fautsch, Jacques Savoy
  Evaluation de diverses stratégies de désambiguïsation lexicale
  In *Proceedings CORIA 2009* Mai 5-7 2009, Presqu'île de Giens, France, pages 19-31

2008

- Claire Fautsch, Jacques Savoy
  UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere
  In *The Seventeenth Text REtrieval Conference (TREC 2008) Proceedings* November 18-21, 2008, Gaithersburg, Maryland, U.S.A.

- Claire Fautsch, Jacques Savoy
  Stratégies de recherche dans la blogosphère
  In *Document Numérique* Volume 11, pages 109-132, Hermès-Lavoisier, Paris, France

- Claire Fautsch, Jacques Savoy
  Recherche d'informations dans la blogosphère : Défis et premières évaluations
  In *Proceedings CORIA 2008*, Mars 12-13 2008, Trégastel, France, pages 441-448

- Claire Fautsch, Ljiljana Dolamic, Samir Abdou, Jacques Savoy
  Domain-Specific IR for German, English and Russian Languages
  In *Carol Peters et al. (Eds.): Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum, CLEF 2007, Revised Selected Papers, Lecture Notes in Computer Science*, Volume 5152/2008, pages 196-199, Springer Berlin/Heidelberg

2007

- Claire Fautsch, Jacques Savoy
  IR-Specific Searches at TREC 2007: Genomics and Blog Experiments
  In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings* November 5-9, 2007, Gaithersburg, Maryland, U.S.A.

Bibliography

[1] P. Lyman, H. R. Varian, P. Charles, N. Good, L. L. Jordan, and J. Pal, "How much information? 2003," 2003.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[3] M. Lalmas and C. J. van Rijsbergen, *Information Retrieval: Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[4] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*. Addison Wesley, 1999.

[5] M. Boughanem and J. Savoy, *Recherche d'information : Etat des lieux et perspectives*. Hermes Science Publications, 2008.

[6] C. Cleverdon, "The Cranfield tests on index language devices," *Readings in information retrieval*, pp. 47–59, 1997.

[7] C. W. Cleverdon, "The significance of the Cranfield tests on index languages," in *SIGIR '91: Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 3–12, ACM, 1991.

[8] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, (London, UK), pp. 355–370, Springer-Verlag, 2002.

[9] G. Salton and M. J. Mcgill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[10] J. J. Rocchio, "Relevance feedback in information retrieval," in *The SMART Retrieval System - Experiments in Automatic Document Processing* (G. Salton, ed.), Englewood, Cliffs, New Jersey: Prentice Hall, 1971.

[11] S. Abdou and J. Savoy, "Searching in Medline: Query expansion and manual indexing evaluation," *Information Processing & Management*, vol. 44, pp. 781–789, 2008.

[12] E. M. Vorhees, "Trec: Improving information access through evaluation." Bulletin of the American Society for Information Science and Technology, 2005.

[13] G. Salton and M. E. Lesk, "Computer evaluation of indexing and text processing," *Journal of the ACM*, vol. 15, no. 1, pp. 8–36, 1968.

[14] M. E. Maron and J. L. Kuhns, "On relevance, probabilistic indexing and information retrieval," *Journal of the ACM*, vol. 7, no. 3, pp. 216–244, 1960.

[15] S. E. Robertson, "The probability ranking principle in *IR*," *Readings in information retrieval*, pp. 281–286, 1997.

[16] S. E. Robertson, S. Walker, and M. Beaulieu, "Experimentation as a way of life: Okapi at TREC," *Information Processing & Management*, vol. 36, no. 1, pp. 95–108, 2000.

[17] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: development and comparative experiments," *Information Processing & Management*, vol. 36, no. 6, pp. 779–808, 2000.

[18] G. Amati and C. J. V. Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 357–389, 2002.

[19] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[20] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval*, (New York, NY, USA), pp. 275–281, ACM, 1998.

[21] D. Hiemstra, "Term-specific smoothing for the language modeling approach to information retrieval: The importance of a query term," in *Proceedings of the 25th ACM Conference on Research and Development in Information Retrieval (SIGIR'02)*, pp. 35–41, 2002.

[22] J. Savoy, "Statistical inference in retrieval effectiveness evaluation," *Information Processing & Management*, vol. 33, pp. 495–512, 1997.

[23] M. Kluck, "Die GIRT-Testdatenbank als Gegenstand informationswissenschaftlicher Evaluation," in *ISI* (B. Bekavac, J. Herget, and M. Rittberger, eds.), vol. 42 of *Schriften zur Informationswissenschaft*, pp. 247–268, Hochschulverband für Informationswissenschaft, 2004.

[24] H. Schott, ed., *Thesaurus Sozialwissenschaften.* Informationszentrum Sozialwissenschaften, Bonn, 2002.

[25] C. Macdonald and I. Ounis, "The TREC Blogs06 collection : Creating and analysing a blog test collection," *DCS Technical Report Series*, 2006.

[26] W. R. Hersh, R. T. Bhupatiraju, L. Ross, A. M. Cohen, D. Kraemer, and P. Johnson, "TREC 2004 Genomics Track Overview," in *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*, 2004.

[27] W. Hersh, A. Cohen, J. Yang, R. T. Bhupatiraju, P. Roberts, and M. Hearst, "TREC 2005 Genomics Track Overview," in *The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2005.

[28] W. Hersh, A. M. Cohen, P. Roberts, and H. K. Rekapalli, "TREC 2006 Genomics Track Overview," in *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2006.

[29] W. Hersh, A. Cohen, L. Ruslen, and P. Roberts, "TREC 2007 Genomics Track Overview," in *The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)*, 2007.

[30] E. Hatcher and O. Gospodnetic, *Lucene in Action (In Action series).* Manning Publications, 2004.

[31] T. B. Rajashekar and W. B. Croft, "Combining automatic and manual index representations in probabilistic retrieval," *Journal of the American Society for Information Science*, vol. 46, pp. 272–283, 1995.

[32] J. Savoy, "Bibliographic database access using free-text and controlled vocabulary: an evaluation," *Information Processing & Management*, vol. 41, no. 4, pp. 873–890, 2005.

[33] V. Petras, "How one word can make all the difference - using subject metadata for automatic query expansion and reformulation," in *Working Notes for the CLEF 2005 Workshop, 21-23 September, Vienna, Austria.*, 2005.

[34] E. M. Vorhees and L. P. Buckland, eds., NIST *Special Publication 500-255: The Twelfth Text REtrieval Conference Proceedings (TREC 2003)*, 2003.

[35] E. M. Vorhees and L. P. Buckland, eds., *NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*, 2004.

[36] E. M. Vorhees and L. P. Buckland, eds., *NIST Special Publication 500-266: The Fourteenth Text REtrieval Conference Proceedings (TREC 2005)*, 2005.

[37] E. M. Vorhees and L. P. Buckland, eds., *NIST Special Publication 500-272: The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2006.

[38] E. M. Vorhees and L. P. Buckland, eds., *NIST Special Publication 500-274: The Sixteenth Text REtrieval Conference Proceedings (TREC 2007)*, 2007.

[39] C. Y. Wei Zhang, "UIC at TREC 2006 blog track," in *The Fifteenth Text REtrieval Conference(TREC 2006) Proceedings*, 2006.

[40] W. Weerkamp, K. Balog, and M. de Rijke, "A generative blog post retrieval model that uses query expansion based on external collections," in *Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-ICNLP 2009)*, (Singapore), 2009.

[41] E. M. Vorhees and L. P. Buckland, eds., *NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008)*, 2008.

[42] H. Yu and E. Agichtein, "Extracting synonymous gene and protein terms from biological literature.," *Bioinformatics*, vol. 19 Suppl 1, 2003.

[43] S. Abdou and J. Savoy, "Report on the TREC 2006 Genomics Experiment," in *The Fifteenth Text REtrieval Conference Proceedings (TREC 2006)*, 2006.

[44] D. Harman, "How effective is suffixing," *Journal of the American Society for Information Science*, vol. 42, pp. 7–15, 1991.

[45] M. F. Porter, "An algorithm for suffix stripping," *Program*, vol. 3, no. 14, pp. 130–137, 1980.

[46] D. A. Hull and G. Grefenstette, "A detailed analysis of English stemming algorithms," tech. rep., Xerox Research and Technology, 1996.

[47] D. A. Hull, "Stemming algorithms - a case study for detailed evaluation," *Journal of the American Society for Information Science*, vol. 47, pp. 70–84, 1996.

[48] C. Fautsch, L. Dolamic, S. Abdou, and J. Savoy, "Domain-specific ir for German, English and Russian languages," in *Advances in Multilingual and Multimodal Information Retrieval, 8th Workshop of the Cross-Language Evaluation Forum,*

CLEF 2007, Budapest, Hungary, September 19-21, 2007, Revised Selected Papers, vol. 5152 of *Lecture Notes in Computer Science*, pp. 196–199, Springer, 2008.

[49] L. Dolamic, C. Fautsch, and J. Savoy, "UniNE at CLEF 2008: TEL, Persian and Robust IR," in *Carol Peters et al. (Eds.): Evaluating Systems for Multilingual and Multimodal Information Access: 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus, Denmark, September 17-19, Revised Selected Papers*, (Springer Berling/Heidleberg), LNCS, 2009.

[50] L. Dolamic, C. Fautsch, and J. Savoy, "UniNE at CLEF: Persian ad hoc and CLEF-IP," in *to appear*, 2009.

[51] C. Fautsch and J. Savoy, "Algorithmic stemmers or morphological analysis? an evaluation," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 1532–2882, 2009.

[52] P. Castells, M. Fernandez, and D. Vallet, "An adaptation of the vector-space model for ontology-based information retrieval," *IEEE Trans. on Knowl. and Data Eng.*, vol. 19, no. 2, pp. 261–272, 2007.

[53] S. Jun-feng, Z. Wei-ming, X. Wei-dong, L. Guo-hui, and X. Zhen-ning, "Ontology-based information retrieval model for the semantic web," in *EEE '05: Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05) on e-Technology, e-Commerce and e-Service*, (Washington, DC, USA), pp. 152–155, IEEE Computer Society, 2005.

[54] F. Peng, N. Ahmed, X. Li, and Y. Lu, "Context sensitive stemming for web search," in *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 639–646, ACM, 2007.

[55] J. Gao, J.-Y. Nie, G. Wu, and G. Cao, "Dependence language model for information retrieval," in *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 170–177, ACM, 2004.

[56] D. Metzler and W. B. Croft, "A markov random field model for term dependencies," in *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, (New York, NY, USA), pp. 472–479, ACM, 2005.