# Combining Positive and Negative Query Feedback in Passage Retrieval

**Taoufiq Dkaki(*,**) & Josiane Mothe(*,***)**
(*) Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 04, France
(**) Isycom, Université Toulouse le Mirail,Toulouse
(***)Institut Universitaire de Formation des Maîtres, 56 Avenue de l'URSS, 31400 Toulouse
{dkaki,mothe}@irit.fr tel: 33 5 61 55 63 22

## Abstract

Information Retrieval Systems aim at retrieving relevant documents according to the information needs which users express. Most Information Retrieval Systems focus on passage retrieval where the granularity of information retrieved is not the document but a smaller unit such as a sentence or passage. These systems try to better answer the users' needs by giving more importance to the most relevant document parts. This paper addresses the problem of passage retrieval as defined by the TREC novelty track, subtask 1 where the aim is retrieving relevant sentences from relevant documents. We define a new term weighting function that takes non relevancy information into account and which is based on query evidence only meaning that it does not need global parameters such as tf.idf term weights. Our method is evaluated on both the 2002 and 2003 TREC novelty collection where we show that taking into account the narrative part that describes non-relevant documents is useful as well as is emphasising terms from topic titles.

## 1 Introduction

Information Retrieval Systems (IRS) traditionally consider documents as the atomic units of information and documents correspond to indexing units and, more importantly, to the units to be retrieved. The fact that most systems choose this level of granularity is probably linked to the fact that the first IRS handled short documents which were generally secondary documents, abstracts of the primary documents. IRS helped users to locate relevant documents but once identified, users accessed paper versions manually. Nowadays, most systems handle electronic documents directly and following this evolution, the size of documents handled by IRS have increased from several paragraphs to several pages or more. These differences in the nature of information handled by an IRS lead to different retrieval mechanisms and interfaces. One of the issues that this throws up is that users may not be satisfied by a system that indicates that a document may be relevant but does not help to retrieve the specific parts within the documents that really answer information needs.

Different approaches have been proposed to deal with this problem. For long documents, giving prominence to the query terms in the retrieved texts (by highlighting them for example) offers valuable help to the user in that it "explains" why a given document has been retrieved while on the other hand it guides the user to document parts to focus on. Retrieving passages rather than documents is another alternative way to help the user. This approach has been largely developed in the 1990s, following the expansion of interest in the SGML language. The structural mark-up of documents has driven retrieval of document parts and more precisely document component retrieval (Wilkinson, 1994), (Corral, 1995). Other approaches aimed at retrieving chunks or fixed-size windows of texts when non-structured documents are involved (Salton, 1994).

More recently, the expansion of interest in the XML language has given rise to interest in document component retrieval (INEX, 2003) whereas the evaluation program TREC (TExt Retrieval Conference), within the novelty track, has been leading the development of new research in passage retrieval within non structured documents at an even lower granularity level, the sentence level (Harman, 2002). The TREC novelty track is composed of two different subtasks namely (1) retrieving relevant sentences from relevant documents and (2) selecting the sentences that bring new information (information not seen before).

In this paper, we detail a method we developed to address the first sub-task (retrieving relevant passages at the sentence level). We discuss the results we obtained using the TREC 2002 and 2003 collections but we note that these experiments, and these results, do derive from an artificially elaborate definition of a user's information need. TREC topic definitions, as used in our experiments, are extremely rich in their information content in that they describe aspects of the user's information need in great detail and indeed they also describe aspects of an information need which a user does not want. This is indeed far richer in terms of detail than is usually available in, say, a web search engine and TREC topic definitions can be criticised for this in some respect. However, our work is based on the premise that a user has indeed expressed their information need in such detail and we leave ignore, for now, the difficulties of how to capture this detail in a real, operating IRS.

## 2 Related works

Information retrieval systems are based on two main processes: information representation (indexing) and information searching (matching a user's query and documents). The indexing process aims at extracting document representations, generally through a set of weighted terms. During the search process, the similarity between the document representations and the query representation is computed in order to decide documents to retrieve.

Passage retrieval as opposed to document retrieval was introduced following the expansion of the SGML language. SGML defines document structure and by doing this, defines meaningful passages. In (Wilkinson, 1994), (Corral, 1995), information units or document parts are extracted using the DTD (Document Type Definition), which is the generic document structure. Information retrieval systems handle these units at each step: indexing, searching and displaying (information units are passages instead of documents). Other studies focus on combining different types of similarities in the searching process: local similarity, which refers to similarity computed at the passage level, and global similarity, which is computed at the document level. (Salton, 1994) developed a search process in two steps: documents are first indexed at the global level and from a user's query, matching documents are selected and re-indexed but at the passage level. The similarity between the query and the passages leads to the selection of the supposed best passages. Other definitions of passage have been proposed in the case of non structured documents such as fixed-size windows (Stanfill, 1992) or sentences (Harman, 2002).

TREC 2002 (Harman, 2002) defined passage retrieval at the sentence level in a more general framework of redundancy detection. In the TREC framework, the sentence retrieval task was based on documents that human evaluators had selected as relevant. This approach simplified the problem as a system could wrongly select sentences from non-relevant documents. Thus this simplification leads to better precision than if all documents had been considered. On the other hand, this simplification aims at keeping the focus on the evaluation of passage retrieval. Most of the systems used in TREC consider sentences as documents and applied their system modules at the sentence level instead of at the document level as usually used. Individual sentences were indexed and then systems computed the similarity between sentences and queries. Blind relevance feedback was often used.

(Larkey, 2002) used the traditional tf.idf and Language Modelling (LM) based models (Ponte, 1998) combined with blind relevance feedback. (Collins, 2002) also used tf.idf-based retrieval to measure the similarity between sentences and query and a query was expanded according to pseudo-relevance feedback. They studied different types of classifiers based on semantic and lexical features extracted from text analysis in order to remove possible non-relevant sentences. (Schiffman, 2002) chose to expand the initial query adding the terms that were semantically equivalent to the query terms and terms that co-occurred with the query terms. (Zhang, 2003) combined blind relevance feedback and automatic sentence categorisation based on a Support Vector Machine. (Dkaki, 2002) introduced a new approach based on term characteristics. Terms were categorised according to three classes: highly relevant, scarcely relevant, non-relevant (stop words).

This paper modifies this latter approach. It introduces a fourth category of terms: highly non-relevant terms. In the case of the experiments within the TREC framework, these terms are extracted from topic sentences that describe what will be considered as a non-relevant sentence. We detail this approach and discuss the results obtained. We also study the influence of the query sections in which terms occur and term weighting functions. The model is evaluated on the TREC collections (2002 and 2003).

One of the advantages of our approach is that it uses a very simple algorithm and results are similar to those obtained using more demanding methods such as SVM. Indeed, our approach is based on a decision function without any learning phase, which opens up a high degree of possible improvements. Moreover, no global parameters from the document collection are needed (e.g. idf-like measure or a LM-like representation), meaning that incorporating new documents or new sentences is trivial.

## 3 Task definition and evaluation

In this paper, passage retrieval is defined according to subtask 1 of the TREC novelty track, (Harman, 2002) where information granularity is the sentence level. Sentences are extracted from relevant documents and the task consists of retrieving those sentences that are relevant.

In our experiments we used TREC collections to evaluate the relevance at the sentence level. In TREC 2002, 49 topics were used from the TREC collection. For each, NIST selected relevant documents with a maximum of 25 documents per topic, which were given to participants with sentences marked-up. After runs were submitted, NIST evaluators decided which sentences were relevant. In 2003, the same type of collection was built using 50 different topics.

Figure 1 corresponds to an example of a TREC topic. It is composed of three textual parts: a title that is supposed to correspond to a typical user's query. It is composed of just a few words. It is written under the form of keywords and not necessarily in real natural language. The two other parts are written in natural language. The descriptive part explains the title whereas the narrative part describes what will be a relevant sentence and a non-relevant sentence.

> Topic: 35
> **Title**: NATO, Poland, Czech Republic, Hungary
> **Descriptive**: Accession of new NATO members: Poland, Czech Republic, Hungary, in 1999.
> **Narrative**: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant. Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

**Figure 1**: topic 35 (TREC 2003)

Table 1 reports some features of the two TREC collections.

| | NIST-2002 | NIST-2003 |
|---|---|---|
| Number of topics | 49 | 50 |
| Number of documents per topic (avg) | 22.3 | 25 |
| Number of sentences per topic (avg) | 1321 | 796.4 |
| Relevant sentences per topic (avg) | 27.9 | 311.14 |
| % of relevant sentences (avg) | 2.1 | 39.1 |

**Table 1**: TREC collections

According to these figures, a system that would retrieve all the sentences (recall equal to 1) would get a precision of about 0.02 (TREC 2002) and of 0.39 (TREC 2003).

We used the evaluation measures proposed in TREC 2003, which are based on commonly used measures of recall and precision. In the general framework of document retrieval, recall and precision are defined in terms of numbers of documents and when considering the sentence level, these measures become (Harman, 2002):

$$Rs = \frac{Number\ of\ relevant\ retrieved\ sentences}{Number\ of\ relevant\ sentences}$$

$$Ps = \frac{Number\ of\ relevant\ retrieved\ sentences}{Number\ of\ retrieved\ sentences}$$

$$Fs = \frac{2 \cdot Ps \cdot Rs}{Ps + Rs}$$

We also use the Fs measure, defined above in terms of Precision and Recall. These measures are averaged over all topics.

## 4 Detection of relevant sentences

### 4.1. Method

Our approach to the detection of relevant sentences in an IRS application consists of considering the sentence as an information unit and most of the TREC participants use this same principle (Harman, 2002), (Soboroff, 2003). Retrieving relevant sentences is based on three phases:

1. texts are analysed to extract indexing terms which are weighted.
2. a similarity function is used to match topics and sentences,
3. sentences are then filtered according to a decision function

The originality of the method we propose is the way texts are represented which emphasises the distinction between terms more than traditional term weighting functions. More precisely, terms are categorized into different levels of relevancy and non relevancy which are used to compute the topic/sentence similarity. These three steps are described in detail in the next sections.

### 4.2. Text analysis and Term Weighting

Each sentence from a document is considered as an information unit and is indexed as follows:

- Stop word removal,
- Stemming where a term is processed to extract its root. We use a resource that provides a list of English terms and its associated root, composed of 21291 entries.
- Phrases are extracted: we use statistically extracted phrases (Crimmins, 1999). Phrases correspond to highly frequent sequences of stems that occur in topics. Once phrases are extracted, topics and sentences are represented by terms. A term can be either a phrase or a stem.
- Weighting: a weight is associated with each term for each sentence as detailed later.

Topics are analysed according to the same process as documents. Each query is represented as a set of weighted terms.

One of the novel aspects of our approach is the way we take into account the topic structure. A TREC topic consists of three query parts (see section 3) where the narrative part indicates what will be considered as a relevant element and what will not. One of the main ideas of the method is to distinguish these two aspects of the narrative part in order to extract terms that are supposed to characterize relevance and terms supposed to characterize non-relevance. Thus we split the narrative section into two sections: NarrativeRel and NarrativeNonRel (see Figure 2). We will return to how this can be applied to more generic retrieval, outside the framework of TREC, at the end of this paper.

**Topic**: 35
**Title**: NATO, Poland, Czech Republic, Hungary
**Descriptive**: Accession of new NATO members: Poland, Czech Republic, Hungary, in
1999.
**NarrativeRel**: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant.
**NarrativeNonRel**: Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

**Figure 2:** topic 35 (TREC 2003) after pre-processing

We can now define a term weight based on its relevancy nature in a topic definition rather than on its frequency of use in documents.

According to our approach, each term extracted from a text (either a document sentence or a topic) is weighted. When extracted from a document sentence, the term weight is set to the term frequency. On the other hand, when extracted from a query, its weight depends on the term category based on its frequency in the topic and on the topic part it occurs in. We defined four classes of terms:

- Highly relevant (HTk),
- Scarcely relevant (STk),
- Non relevant (iTk)
- Highly non-relevant (ITk)

The weight of a term in a topic and its weight in a sentence are then combined when choosing sentences to be retrieved. We now present our definition of the term weighting functions we use.

1. Weight of terms from document sentences

Given $S_j$ a sentence, $t_i$ a term : $tf_{i,j}$ is the frequency of $t_i$ in $S_j$ .

$$weight(t_i, S_j) = tf_{i,j} \qquad (1)$$

2. Weight of topic terms and classes of terms

According to our approach, a term is associated with a term class according to its topic frequency and to its use in the topic definition.

Given $T_k$ a topic and $t_i$ a term,

$T_k = TT_k \cup TD_k \cup TNR_k \cup TNN1_k \cup TNN2_k$ where $TT_k$ corresponds to the set of terms extracted from the Title of the topic, $TD_k$ from the Descriptive, $TNR_k$ from the NarrativeRel and $TNN1_k \cup TNN2_k$ corresponds to terms extracted from the NarrativeNonRel topic part ; they are defined as follows: $TNN2_k \cap (TT_k \cup TD_k \cup TNR_k \cup TNN1_k) = \phi$. TNN1k corresponds to terms that can also occur in the title, descriptive or relevant narrative topic parts whereas terms from TNN2k only occur in the irrelevant narrative part. We distinguish TNN1k from TNN2k, because the latter terms can be considered as defining non relevancy whereas for the former terms a decision cannot be taken.

The term weight regarding a topic is computed as follows:

$$\omega_{i,k} = \sum_{P \in \{T,D,NR,NN1,NN2\}} \mu_P \cdot tf_{i,k,P}$$

$$\begin{aligned} weight(t_i, T_k) &= \omega_{i,k} \quad if \quad \omega_{i,k} \geq \tau_H \quad or \quad \omega_{i,k} < 0 \qquad (2) \\ \tau_0 < w_{i,k} &< \tau_H \; if \quad 0 < \omega_{i,k} < \tau_H \\ &= 0 \qquad otherwise \end{aligned}$$

where $tf_{i,k,P}$ is the frequency of $t_i$ in $TP_k$ part, $P \in \{T, D, NR, NN1, NN2\}$, $\tau_L \leq w_{i,k}$

$\tau_L \leq \tau_H$ , $\mu_{NN2} \leq 0$

This formula puts into practice, the following hypothesis:

- A query term should be weighted according to its frequency of use in the different query parts (tfi,k,P),
- Query parts can contribute differently to the term weighting: typically the terms from the title are more important than the terms from the narrative part. μP is used to weight the importance of the different topic parts (typically μT ≥ μD). wi,k is the contribution of ti to the topic Tk.
- A term which contributes highly to the topic definition (wi,k ≥ τH) should be included in the topic representation at the level of its contribution (Weight (ti,Tk)=wi,k).
- A term which contributes somehow to the topic (wi,k > τ0) but which has a low contribution to the topic definition (wi,k<τH) should contribute less in the topic representation, typically less than its frequency. We set its weight to a constant $\tau_L$ such as $\tau_L \leq w_{i,k}$. $\tau_H$, $\tau_L$ and $\tau_0$ are set according to the term distribution in the topics (see section 5.1).
- A term that occurs only in the 'NarrativeNonRel', i.e. terms from TNN2k (see Figure 3) should be weighted in a negative way or deleted (μNN2≤0); in such case wi,k may become negative.
- A term that occurs in the 'NarrativeNonRel' but also in other topic parts could be theoretically weighted in a positive or in a negative way. Intuitively, we prefer to give a positive contribution to these terms (μNN1≥0).

At the same time, this function is generic enough to make it possible to invalidate the contribution of some query parts, for example considering title and description only by setting the other μP to zero.

**Topic**: 35
**Title**: NATO, Poland, Czech Republic, Hungary
**Descriptive**: Accession of new NATO members: Poland, Czech Republic, Hungary, in
1999.
**NarrativeRel**: Identity of current and newly-invited members, statements of support for and opposition to NATO enlargement and steps in the accession process and related special events are relevant. Impact on the new members, i.e., requirements they must satisfy, and their expectations regarding the implications for them are relevant. Progress in the ratification process is relevant.
**NarrativeNonRel**: Future plans for NATO expansion, identification of nations admitted on previous occasions, and comments on future NATO structure or strategy are not relevant.

TNN2$_{35}$: STRUCTURE, STRATEGY, PLAN, NATION, IDENTIFICATION, EXPANSION, COMMENT

NATO, for example, is not part of TNN2$_{35}$ but is part of TNN1$_{35}$ (has its occurs not only in NarrativeNonRel but in other sections too).

**Figure 3**: Example of TNN2

Based on its final weight in the query, each term is then categorized into one of the groups defined as follows:

$$HT_k = \{t_i / t_i \in T_k \text{ and } weight(t_i, T_k) > \tau_L\}$$

$$ST_k = \{t_i / t_i \in T_k \quad and \quad weight(t_i, T_k) = \tau_L \quad and \quad w_{i},_{k} > 0\}$$

$$IT_k = \{t_i \ / \ weight(t_i, T_k) < 0\}$$

$$iT_k = \{t_i \ / \ weight(t_i, TP_k) = 0 \quad \forall P \in \{T, D, NR, NN1, NN2\}\}$$

HTk corresponds to the set of terms that are highly relevant for the topic; their wi,k is higher than τH, thus their weight(ti,Tk) is higher than τL.
STk corresponds to the set of terms that are scarcely relevant for the topic, their contribution to the topic was too low and their weight(ti,Tk) has been set to τL. The condition wi,k>0 is used for mathematical reasons as the constraint τL>0 is not part of our model. However, in practice, intuitively, τL should be positive.
ITk is the set of terms representative of non-relevant terms; because μNN2≤0, weight(ti,Tk) is negative for these terms.
iTk can either correspond to stop words or to terms that do not occur in the treated topic.

These groups are used when deciding the relevance of a given sentence.

## 4.3. Detection of relevant sentences

In order to decide if a sentence is relevant, we associate three components to each sentence:

- a score that reflects the sentence/topic matching : the similarity between a topic and a sentence are computed according to the traditionally used vector space similarity measure (Salton, 1971). (3) is not normalized because texts are short (sentences and topics) and we believe normalisation will have no effect. Once topic/sentence similarity is computed, a decision function decides which sentences will be retrieved (see (4)).
  Given a topic $T_k$ and a sentence $S_j$

$$Score(S_j, T_k) = \sum_{t_i} \left( weight(t_i, S_j) \cdot weight(t_i, T_k) \right) \qquad (3)$$

using the term weights as defined in paragraph 4.2.

- and two groups of terms:

$$HS_j = \{t_i / t_i \in (Sj \cap HT_k)\}$$

$$SS_j = \{t_i / t_i \in (Sj \cap ST_k)\}$$

$HS_j$ corresponds to the highly relevant terms from the topic that occur in the sentence,

$SS_j$ corresponds to the scarcely relevant terms from the topic that occur in the sentence.

A given sentence $S_j$ is then considered as relevant iff:

$$Score(S_j,T_k) > \psi(|HS_j|,|SS_j|,|HT_k|,|ST_k|) \qquad (4)$$

where $|X|$ is the number of elements of $X$

A decision function is used to select the sentences to be retrieved.
It is based on the fact that the score the sentence gets should be higher than a given threshold. This threshold depends on the number of relevant terms (highly and scarcely relevant) from topics and sentences (4).
The ψ function has been defined according to the definition of Score (Sj,Tk)

$$Score(S_j,T_k) = \sum_{t_i}\left(weight(t_i,S_j)\cdot weight(t_i,T_k)\right)$$

Because Sj is one sentence (i.e. term frequency is 1 for most of terms from a sentence, apart from the stop words which are removed anyway), the following approximation can be done: weight (ti,Sj)∈ {0,1}. The score can be approximated to:

$$Score(S_j,T_k) \approx \sum_{t_i\in S_j}\left(weight(t_i,T_k)\right)$$

$$\approx \sum_{t_i/T_i\in HS_j} w_{ik} + \tau_L\cdot|SS_j|$$

Ideally, $HT_k$=$HS_j$ and $ST_k$=$SS_j$, that is to say $|HT_k|$=$|HS_j|$ and $|ST_k|$=$|SS_j|$. Thus, for $S_j$ to be a very good candidate for $T_k$, $Score(S_j,T_k) = \sum_{t_i\in HTk} w_{i,k} + \tau_L\cdot|ST_k|$ should be verified. Additionally, we consider the way $w_{i,k}$ is computed: $\sum_{t_i\in HSj} w_{i,k} \leq \tau_H\cdot|HS_j|$ with ideally $HS_j$=$HT_k$.

As a result, the decision function is relaxed so that : $\psi(|LS_j|,|HS_j|,|HT_k|,|LT_k|) < \tau_H\cdot|HT_k| + \tau_L\cdot|ST_k|$

The function we defined is based on this assumption and is defined by:

$$\psi_0 = (\tau_H - 1)\cdot|HT_k| + (\tau_L - 0.15)\cdot|ST_k|$$

Because we want to encourage sentences that have a high proportion of highly relevant terms when the topic has a lot of relevant terms or a high proportion of scarcely relevant terms when the topic has a few highly relevant terms, we relax more the constraint and defined ψ as:

$$\psi_1 = (\tau_H - 1)\cdot|HT_k| + (\tau_L - 0.15)\cdot|ST_k| - \frac{3}{2}\cdot\frac{|SS_j|}{|SS_j|+|HS_j|}\cdot|HT_k| - \frac{1}{2}\cdot\frac{|HS_j|}{|SS_j|+|HS_j|}\cdot|ST_k| \qquad (5)$$

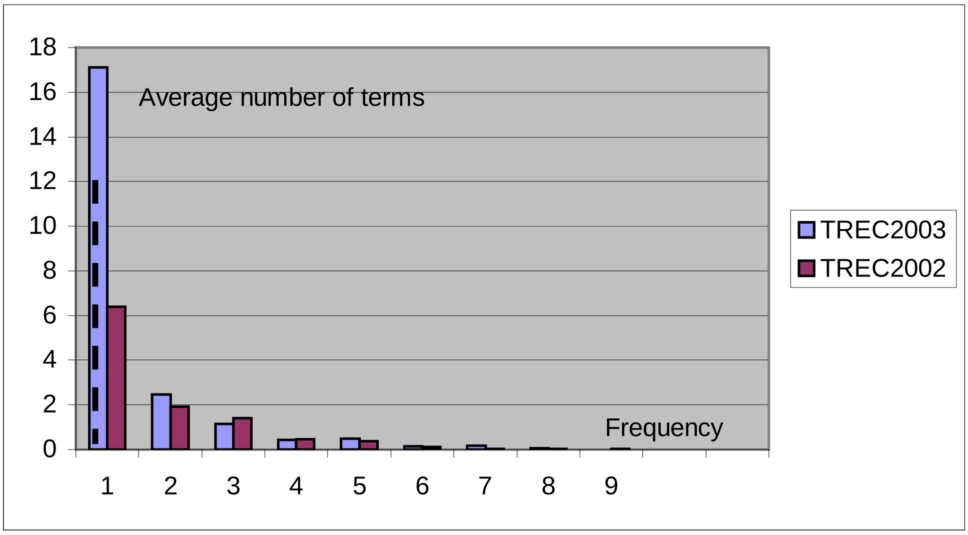
## 5 Experiments

Some of the parameters we defined earlier in this paper were defined in a intuitive way and so our evaluation will take into account ranges of values for some of these parameters. Because different parameters values have been used, we obtained about 200 different experimental runs. However, we present only some of these results and comment on them.

### 5.1. Defining $\tau_0$, $\tau_L$ and $\tau_H$

The definition of τ0, τL and τH (defined formula (2)) is based on the analysis of the topic term frequencies. Graph 1 indicates the number of terms (after stop word removal and stemming) for each term frequency within the topics.

Graph 1 show that in TREC 2003, on average, a query is composed of about 22 terms, 17 of which have a frequency of 1, 2.5 terms have a frequency of 2 and the 2.5 other terms have a frequency between 3 and 8. In TREC 2002, on average, a query is composed of about 11 terms, 6.5 having a frequency of 1, 2 a frequency of 2 and the remaining 2.5 terms have a frequency which is between 3 and 9. Thus, most of the topic terms have a frequency of 1 (59% in TREC 2002 and 78% in TREC 2003).



**Graph 1**: Number of terms for each term frequency – Average over the topics

τ0 is a constant below which the contribution of the terms to the topics are not considered. According to the topic term distribution, if τ0 is set to 1, more than half of the topic terms will be considered as stop words. We set τ0 as follows: τ0 =0.

τL is a constant that corresponds to the weight of the scarcely relevant terms whereas τH is the minimum value of a term contribution for the term to become a highly relevant term. As a result, τL<0 and τL≥τH would be nonsense. On the other hand, queries should not be composed of only scarcely relevant terms without any highly relevant terms. As 10% (TREC 2002) and 5% (TREC 2003) of terms have a frequency higher than 3, τH >3 would set almost all the contributing query terms to τL invalidating the effect of term weighting, which is not our aim. Taking these points into consideration we decided on ranges of values for these parameters and conducted experiments.

### 5.2. Initial query sections

A topic is initially composed of different parts: title, descriptive and narrative. The usefulness of the narrative part has already been shown in previous TRECs when document retrieval was involved. We evaluate the hypothesis that this remains true when sentence retrieval is considered. In these experiments, the narrative part is considered in the usual way (not as a negative contribution to the topic representation).

Table 2 reports the results we obtained. TD in the *Topic part* column means the title and descriptive only are considered (TD: $\mu_T=\mu_D=1, \mu_{NR}=\mu_{NN1}=\mu_{NN2}=0$), T2DN corresponds to $\mu_T=2, \mu_D=\mu_{NR}=\mu_{NN1}=\mu_{NN2}=1$). In these experiments, $\tau_H$=3, $\tau_L$=0 and $\tau_0$=0.

Intuitively, terms extracted from the title of a topic should be more important than terms from the other topic parts.

| | TREC 2002 | | | TREC 2003 | | |
|---|---|---|---|---|---|---|
| $\tau_0=0$ $\tau_L=0$ $\tau_H=3$ Topic part | Rs | Ps | $\frac{2*Ps*Rs}{Ps+Rs}$ | Rs | Ps | $\frac{2*Ps*Rs}{Ps+Rs}$ |
| TD_P | 0.42 | 0.15 | 0.180 | 0.54 | 0.63 | 0.491 |
| TDN_P | 0.47 | 0.15 | 0.186 | 0.58 | 0.63 | 0.522 |
| T2DN_P | 0.5 | 0.15 | 0.193 | 0.61 | 0.62 | 0.543 |
| T3DN_P | 0.5 | 0.14 | 0.190 | 0.64 | 0.62 | 0.552 |
| T4DN_P | 0.50 | 0.14 | 0.190 | 0.64 | 0.62 | 0.556 |
| T5DN_P | 0.50 | 0.14 | 0.190 | 0.64 | 0.61 | 0.555 |

**Table 2**: Initial topic sections – average over topics

Regarding TREC 2002, there is no significant difference on the experiments reported above. The best results are obtained in making title terms contribute twice to the topic representation.

With regard to the TREC 2003 collection, precision does not vary, but there is a significant improvement of recall (about 12%) and the best results are obtained when terms from the title are considered as four times the other terms.

The fact that the title should be considered differently according to the collection could be linked to the fact the topics have different length according to the collection (22 terms for TREC 2003 and 11 terms for TREC 2002) whereas title length does not vary from TREC 2002 to TREC 2003. These results show the not unsurprising result that giving more importance to the title section enables relevant sentences to be filtered successfully by the decision function. Additionally, these results are consistent with the literature reports of the use of the narrative section of the topics.

### 5.4. Non-Relevance

The narrative part of each topic corresponds to a description of what an evaluator should consider as a relevant document and the query part is supposed to represent the user's point of view. We applied a simple analysis process on the narrative part in order to separate:

- sentences that correspond to the description of relevance and,
- sentences that describe non-relevance.

These two types of narrative are used in order to build new topic representations where highly irrelevant terms are taken into account (ITk) using $\mu_{NR}, \mu_{NN1}, \mu_{NN2}$. Table 3 reports the results we obtained when:

- T4DN_P: $\mu_T=4$ $\mu_D=\mu_{NR}=\mu_{NN1}=\mu_{NN2}=1$, (irrelevance is not taken into account), phrases are used.
- T4DN_P_I: $\mu_T=4$, $\mu_D=\mu_{NR}=\mu_{NN1}=1$ and $\mu_{NN2}=0$, (irrelevant terms are considered like stop words "I" is the run name indicates that the NN2 section is considered differently).
- T4DN_P_IN: $\mu_T=4$, $\mu_D=\mu_{NR}=\mu_{NN1}=1$ and $\mu_{NN2}=-7$, ("IN" in the run name indicates the NN2 section is considered in a negative way).

| | TREC 2002 | | | TREC 2003 | | |
|---|---|---|---|---|---|---|
| $\tau_0=0$ $\tau_L=0$ $\tau_H=3$ | Rs | Ps | $\frac{2*Ps*Rs}{Ps+Rs}$ | Rs | Ps | $\frac{2*Ps*Rs}{Ps+Rs}$ |
| TDN_P | 0.47 | 0.15 | 0.186 | 0.58 | 0.63 | 0.522 |
| T4DN_P | 0.50 | 0.14 | 0.190 | 0.64 | 0.62 | 0.556 |
| T4DN_P_I | 0.50 | 0.14 | 0.190 | 0.65 | 0.62 | 0.561 |
| T4DN_P_IN | 0.50 | 0.14 | 0.190 | 0.63 | 0.62 | 0.554 |

**Table 3**: Irrelevancy – average over topics

The results reported table 3 show that taking into account non relevancy in terms improves the Fs measure for the TREC 2003 collection (T4DN vs T4DN_P_I 9%) where non relevancy in terms should make them be considered as stop words. An important feature of this is how many terms are in NN2 and table 4 reports these numbers.

| | TREC 2002 | TREC 2003 |
|---|---|---|
| Number of terms in NN2 | 40 | 130 |
| Number of phrases in NN2 | 0 | 1 |
| Number of terms in NN2 per topic | 0.8 | 2.6 |
| % of NN2 compared to topic size | 6% | 11% |

**Table 4**: number of NN2 terms

In TREC 2002, only a third of the topics contain a narrative part that indicates non-relevancy (35% of the topics) whereas 66% of the topics in TREC 2003 contain such a description. However, in Table 3, the results are averaged over the 49/50 topics.

When averaging the results over the different runs, we found that both precision and recall are almost doubled when acknowledging the existence of irrelevant parts in topics narrative sections on TREC

2003. Average precision over the runs is 0.656 (when irrelevancy is taking into special account) vs 0.325 and recall is 0.437 vs 0.228.

### 5.5. Comparison with TREC Runs

The TREC novelty track was defined in 2002 and in 2003, 14 groups participated in this track (55 runs submitted). A comparison of results obtained by participants is presented Table 5.

| | *TREC 2003* | | |
|---|---|---|---|
| | *Ps* | *Rs* | $\frac{2 * Ps * Rs}{Ps + Rs}$ |
| Best TREC | 0.60 | 0.79 | 0.619 |
| Our Official TREC run | 0.64 | 0.58 | 0.526 |
| T4DN_P_I | 0.65 | 0.62 | 0.561 |
| T4DN_P_I_BFB | 0.59 | 0.75 | 0.593 |

**Table 5**: TREC results

Most of the systems used in TREC used pseudo-relevance feedback (Mitra, 1997). When applied to document retrieval, pseudo-relevance feedback considers the top retrieved documents as relevant and modifies the initial query modification according to Rocchio's algorithm (Rocchio, 1971).
This process can be included in our model, modifying the term weight (weight(ti,Tk)) as follows:

$$\text{weight}(t_i,T_k)' = \text{weight}(t_i,T_k) + \beta \sum_{S_j\, rel} weight(t_i,S_j)$$

where Sj are the sentences considered as relevant.
In table 5, this process has been included using the 10 first retrieved sentences ("first" is based on chronological order) based on T4DN_P_I ranking and given the mnemonic T4DN_I_BFB. When compared to the best results obtained in TREC 2003 (first line in table 5, using SVM), the Fs-measure only differs by 4%.

## 6 Conclusions and Perspectives

Because IR systems now handle long documents it is of great importance to help users to focus on the most useful parts of retrieved documents. Similarly, it is of great importance to help users in describing their information need in as much detail as possible. Traditionally, IRS have assumed a user's query to be a short statement of an information need and this assumption has been taken to an extreme in the development of WWW search engines where average query lengths are only a few search terms. TREC topic definitions have a level of detail in their description which is not normally available in an IRS and this leads to valid criticism of experiments such as ours which takes advantage of this detail. However, we are not concerned here with developing mechanisms which can capture this level of detail and we are only interested in proving that if this level of detail is available, then it can be used profitably to improve retrieval effectiveness, and we have achieved that goal.

In this paper we described a method to retrieve sentences that are relevant according to a given topic. This method has two original points: first it aims to differentiate terms according to the role they play in the topic definition. We defined several classes of query terms to capture this namely highly and scarcely relevant terms and non and highly non-relevant terms. Secondly, the approach to sentence retrieval that we have taken is based on a decision function where the threshold for retrieval we consider depends on the distribution of the types of terms in the topics only and is independent of global parameters. To evaluate this method we used the two collections constructed in the framework of the novelty detection task in TREC. We found that when using this method on the TREC 2003 collection, 3 of 4 of the retrieved sentences are relevant and about 60% of the relevant sentences are retrieved. Our best results have been obtained by weighting terms from the title as four times terms from the other sections and deleting terms that occur only in the narrative part that describes what will be considered as a non relevant element.

Our study, as well as previous studies related to passage retrieval, should open a discussion on the question of what is the granularity of a relevant object: paragraphs or shorter elements such as sentences or larger elements such as sections? The choice of granularity level should be linked to the type of application or use of the retrieved information. Retrieving sentences out of their context is probably of little use for final users however, knowing how helpful systems that just highlight the query terms in the retrieved documents are, one can imagine how useful more sophisticated interfaces that highlight relevant sentences could be. We have developed such an interface described elsewhere (Mothe, 2003) and using this interface a user can browse retrieved documents and retrieved sentences where the level of highlighting is related to the supposed level of sentence relevance.

Sentence selection can also be considered as a first step in other tasks. It can be used in novelty detection as defined in TREC (Harman, 2002). The novelty task aims at detecting sentences that bring novel information or in other words sentences that are not redundant. We propose a method based on our model of relevance sentence retrieval (Dkaki, 2003).

Our future work will focus on pseudo-relevance feedback where the top ranked documents are considered as relevant for the purposes of applying relevance feedback. In the case of sentence retrieval, chronological order is of importance, specifically if sentence retrieval is part of novelty detection. However, the first retrieved sentences are not necessarily the best ones to be used in pseudo-relevance feedback, especially if they are extracted from a single document. We will investigate the usefulness of a sample of sentences extracted from different documents.

## 7 References

(Allan, 2003) J. Allan, C. Wade, A. Bolivar, Retrieval and Novelty Detection at the Sentence Level, Research and Development in Information Retrieval, SIGIR'03, pp 314-321, 2003.

(Collins, 2002) K. Collins-Thompson, P. Ogilvie, Y. Zhang, J. Callan, Information filtering, Novelty detection, and named-page finding, Text Retrieval Conference TREC 2002, pp 107-118, 2002.

(Corral, 1995) M.-L. Corral, J. Mothe, How to retrieve and display long structured documents ?, Basque International Workshop on Information Technology, BIWIT'95, pp 10-19, 1995.

(Crimmins, 1999) F. Crimmins, T. Dkaki, J. Mothe, A. F. Smeaton, TétraFusion: Information Discovery on the Internet, IEEE Intelligent Systems & their applications, 14 (4), pp 55-62, IEEE Computer Society, 1999.

(Dkaki, 2002) T. Dkaki, J. Mothe, J. Augé, Novelty track at IRIT-SIG, Text Retrieval Conference TREC 2002, pp 332-336, 2002.

(Dkaki, 2003) T. Dkaki, J. Mothe, Novelty track at IRIT-SIG, Text Retrieval Conference TREC 2003, pp 413-418, 2003.

(Harman, 2002) D. Harman, Overview of the TREC 2002 novelty track, Text Retrieval Conference TREC 2002, pp 46-55, 2002.

(INEX, 2003) Initiative for the Evaluation of XML retrieval (http://qmir.dcs.qmw.ac.uk/INEX/).

(Kazawa, 2002) H. Kazawa, T. Hirao, H. Isozaki, E. Maeda, A machine learning approach for QA and Novelty tracks: NTT system description, Text Retrieval Conference TREC 2002, pp 472-475, 2002.

(Larkley, 2002) L.S. Larkey, J. Allan, M.E. Connell, A. Bolivar, C. Wade, UMASS at TREC 2002: Cross Language and Novelty Tracks, Text Retrieval Conference TREC 2002, pp 721-732, 2002.

(Mitra, 1997) M. Mitra, C. Buckley, A. Singhal, C. Cardie, An analysis of Statistical and Syntactic Phrases, RIAO, pp 200-214, 1997.

(Mothe, 2003) J. Mothe, T. Dkaki, C. Mhamedi, Restituer l'information utile à l'utilisateur : visualisation de la pertinence et de la nouveauté dans les textes, Journées de l'innovation, to appear in 2004.

(Ponte, 1998) J.M. Ponte, W.B. Croft, A language modelling approach to information retrieval, Research and Development in Information Retrieval, SIGIR'98, pp 275-281, 1998.

(Salton, 1994) G. Salton, J. Allan, C. Buckley, Automatic structuring and retrieval of large text files, communication ACM, 37(2), pp 97-108, 1994.

(Schiffman, 2002) B. Schiffman, Experiments in Novelty Detection at Columbia University, Text Retrieval Conference TREC 2002, pp 188-196, 2002.

(Soboroff, 2003) I. Soboroff, D. Harman, Overview of the TREC 2003 Novelty track, pp 96-115, 2003.

(Stanfill, 1992) C. Stanfill, D.L. Waltz, Statistical methods, artificial intelligence, and information retrieval, Text-based intelligent systems: current research and practice in information extraction and retrieval, Ed. P.S. Jacobs, pp 215-226, 1992.

(Wilkinson, 1994) R. Wilkinson, Effective retrieval of structured documents, Research and Development in Information Retrieval, SIGIR'94, pp 311-317, 1994.

(Zhang, 2002) M. Zhang, R. Song, C. Lin, S. Ma, Z. Jiang, Y. Jin, Y. Liu, L. Zhao, et S. Ma, Expansion-based technologies in finding relevant and new information: THU TREC2002: Novelty Track Experiments, Text Retrieval Conference TREC 2002, pp 586-590, 2002.

(Zhang, 2003) M. Zhang, C. Lin, Y. Liu, L. Zhao, L. Ma and S. Ma, THUIR at TREC 2003: Novelty, Robust, Web and HARD, Text Retrieval Conference TREC 2003, pp 137-147, 2003.