The system RELIEFS : a new approach for information filtering

Christophe Brouard and Jian-Yun Nie

Laboratoire RALI
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
CP. 6128 succ. Centre Ville
Montréal QC H3C 3J7, Canada

{brouardc,nie}@iro.umontreal.ca

Abstract

In this year's filtering track, we implemented a system called RELIEFS that tries to learn about the prediction capability of words or conjunctions of words for the relevance of documents. The novelty of the system resides in two main points. First, the features used in the prediction involve both : the implication D->Q (from document to query), and the reverse implication Q->D. This is different from usual approaches where only the first of the implication is used. Therefore, the relevance estimation of a document combines the probability that a document containing a term is relevant, and the reverse probability - the probability that a term appears in relevant documents. The second novelty is that, in addition to the use of words as prediction elements, we also consider word combinations (conjunctions). However, not all combinations are significant. Therefore, an incremental algorithm is developped to select only the meaningful conjunctions. To limit the number of conjunctions, we do not use a cut on conjunction length. Rather, we eliminate the conjunctions A&B that bear the same information as A or as B. Our first results prove the feasibility of the approach. Other experiments are ongoing in order to fully evaluate this approach.

1. Introduction

The goal of our participation in TREC9 is to experiment the following two ideas for information filtering :

The first idea is about the use of the two implications D->Q (from document to query) and Q->D. Usually, in Information Retrieval, relevance evaluation is based on the evaluation of D->Q (van Rijsbergen, 1986). If one considers a document as a set of terms, and a query as a specification of what we are looking for, the implication D->Q may be decomposed to the judgment of "if the term is present then the document is relevant" for each term of the document.

Even if some authors signal the importance of the reverse implication (Q->D) (Nie, 1988), the relation has not been integrated in relevance evaluation. This relation has been taken into account in probabilistic models as a way to calculate D->Q. In our approach we will consider both implications simultaneously. the consideration of the reverse implication Q->D means in practice that we have to consider the relation "if the document is relevant then the term is present in the document". From a pragmatic point of view, the use of Q->D may be justified by the fact that it allows us to favour terms which have been met many times in relevant documents, comparing to rare terms for which the presence in relevant document is a coincidence. The two implications may be illustrated as two relationships between terms and relevance as in Fig 1. The two relationships are different in nature and both are important for judging the relevance of a document. Therefore, we will integrate both of them in our approach.

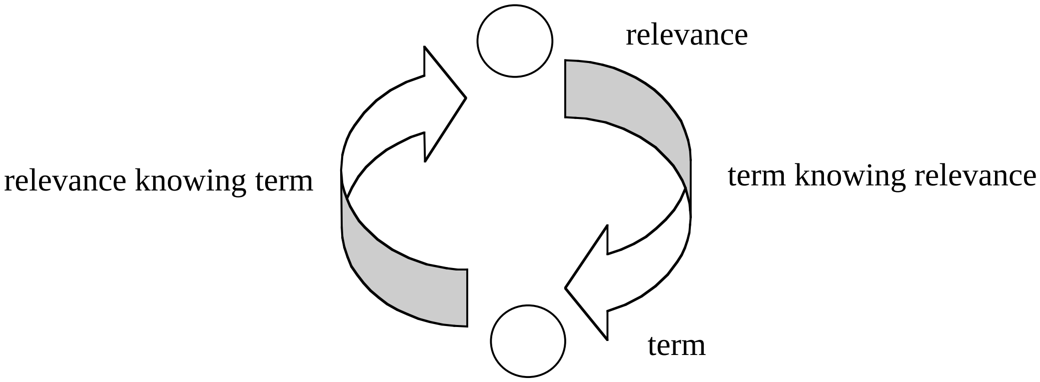

Figure 1 - Relationships between terms and relevance

The second important aspect in our approach is the use of term combinations. Usually, the learning for adaptive filtering system consists of updating the weights of terms and not term conjunctions. This is because the assumption that terms are independent. It is also due to the fact that considering term combinations would lead to a combinatory explosion. Some methods tried to consider term combinations, but usually limited themselves to a certain length. This solution is not totally satisfactory since the length constraint can not be completely justified. Morever, the number of combinations is still very high. We propose here to update all the implications whithout loosing any information. The economy principle we propose is based on the observation that if two terms t_1 and t_2 are always present simultaneously (in the same documents), it is useless to create the information t_1&t_2 since this information is the same as t_1 (or t_2). In this way, many combinations can be eliminated. We use an incremental algorithm (Brouard, 2000) to determine whether a combination should be added and its weight be updated.

2. System description

The goal of RELIEFS system is to find words or conjunctions of words that are good predictors of document relevance. The RELIEFS processing can be decomposed in three steps: 1. Selection of N document words from the document, 2. Estimation of the document's relevance, 3. Revision of word's predictability.

2.1 Step 1 : Selection of N document words

All the document words are compared with the words which have been extracted from the query, the document examples given for learning and the documents which have been previously selected. The considered words or word conjunctions are elements of prediction p_i. They are sorted by the value of their predictability of relevance. This predictability is estimated as the product between the relative frequency of relevance knowing p_i and the reverse frequency, i.e. $F(R/w_i).F(w_i/R)$. If less than N words (in our experiments we choose N=20) can be selected in this way, this selection is completed first by the document words which are related to the query words and finally by the document words in their lecture order. The relatedness between words is estimated using both implications on the training set (Ohsumed 87). In our solution of additional words, those that are related to sereval query words are given priority.

2.2 Step 2 : Evaluation of the relevance

Considering the elements of prediction which appear in the document, the score of the document is computed as follows :

$$\frac{\sum_{i^*=1}^{k} F(R/p_{i^*}).F(p_{i^*}/R)}{\sum_{i=1}^{n} F(R/p_i).F(p_i/R)}$$

where $F(R/p_i)$is the relative frequency of relevance given the presence of the element of prediction $p_i$in this document, $F(R/p_i)$is the reverse relative frequency and i* are the indices of the elements of prediction which are present in the document. In RELIEFS, the relevance of a document is estimated as the sum of the implication products for all the elements of prediction present in the document divided by the sum of the implication products for all the elements of prediction. In the example of Fig 2, word5 and word8 are elements of prediction and appear in the document, the implication products of these elements are taken into account and increase the score of the document.

2.3 Step 3 : Updating relative frequencies

If the evaluation of step2 is larger than a defined threshold, then the N words selected in the first step are submitted to an updating process on their relative frequencies, and new conjunctions are also built. The building condition of a conjunction A&B is that F(A/B) and

F(B/A) are different from 1. This condition allows us to avoid building useless conjunctions (i.e A&B is equivalent to A or/and to B).

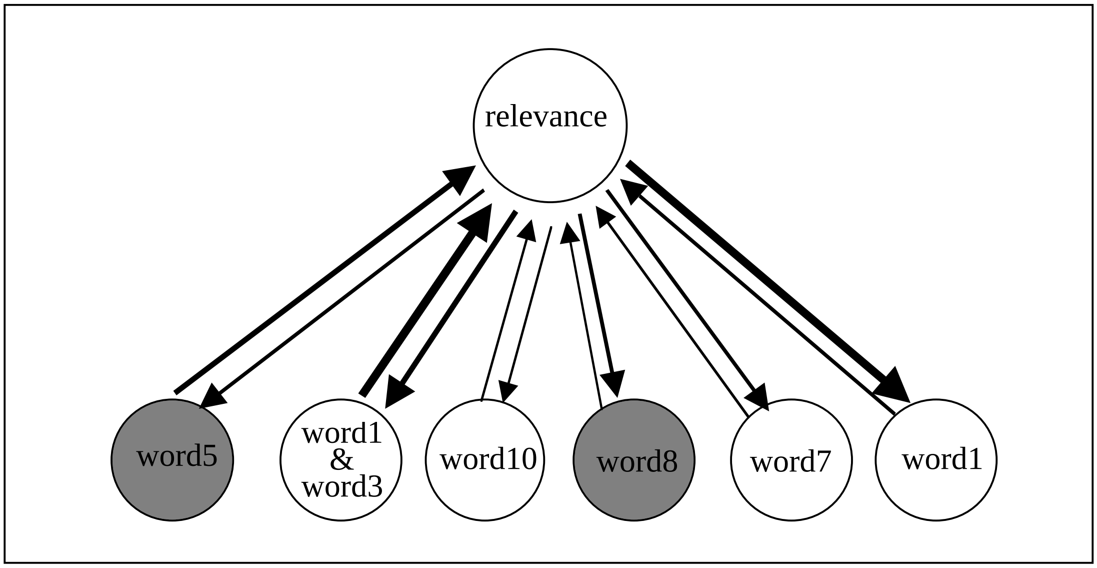

Figure 2 - Relevance evaluation in RELIEFS

2.4 Threshold adaption

We tried to adjust the thresholds with a very simple mechanism. When a selected document is irrelevant, the threshold is increased with a small value. Each time a document is not selected, the threshold is decreased. The initial threshold is computed on the basis of the score of the two first relevant documents and the amplitude of the threshold modification is based on the difference between the average score of the two last relevant documents and the two last irrelevant documents. Initially, we considered an average of 0 for irrelevant documents. So the change tends to be larger at the begining than at the end. Morever, the product of the change scale by a constant allows us to vary more globally all the thresholds.

3. Results & Discussion

We have submitted two runs on Ohsumed collection. The first one considers higher thresholds than the second one (the constant used in the product with the change scale is larger). Its utility score is positive (+1.1). We submitted it for comparison on utility criteria. The comparison is favourable since about 60% of the scores are above the median (table1).

	below-median	at-median	above-median
relief1	12	14	37

Table 1 : Comparison on utility criteria of adaptive filtering run.

We considered also smaller thresholds (decreasing the constant) for a second run in order to increase the number of selected documents which was too small for optimizing precision since when less than 50 document are selected a penalty is applied. This time, the utility score is approximately -1 and the corrected precision is approximately 0.17 (0.28 if not corrected). The comparison of our result with other systems optimized for precision is not favourable (table 2). However, it is to be noted that our system is not tuned to optimized the precision but utility.We think that the results could be improved if we set lower thresholds in order to keep more documents and then to avoid the under-50 penalty.

	below-median	at-median	above-median
reliefs2	42	11	10

Table 2 : Comparison on precision criteria of adaptive filtering run.

Globally, these very first results are encouraging, in particular for utility. They show that using a small number of words (20) to represent documents can perform as well as traditional information filtering systems in which much more words are considered. However, it is also necessary to consider word conjunctions.

4. Conclusion

In our information filtering approach, we take into account two implications, D->Q and Q->D. Morever, we developped a solution in order to take into account word conjunctions. Further experiments will be done to evaluate more precisely the avantages of each of these aspects.

Acknowledgment

This research has been funded by INRIA (Institut National de Recherche en Informatique et Automatique).

References

Brouard, C.(2000). *Construction et exploitation de réseaux Sémantiques Flous pour l'Extraction d'Information Pertinente : Le système RELIEFS*. Thèse de l'université Paris 6.

Nie, J. Y. (1988). *An outline of a general model for information retrieval*. Proceedings of the 11th Annual ACM Conference on Research and Development in Information Retrieval, Grenoble.

van Rijsbergen, C. J. (1986). A non-classical logic for information retrieval. *The Computer Journal, 29(6),* 481-485.