Finding content-bearing terms using term similarities

Justin Picard
Institut Interfacultaire d'Informatique
University of Neuchâtel
SWITZERLAND
justin.picard@seco.unine.ch

Abstract

This paper explores the issue of using different co-occurrence similarities between terms for separating query terms that are useful for retrieval from those that are harmful. The hypothesis under examination is that useful terms tend to be more similar to each other than to other query terms. Preliminary experiments with similarities computed using first-order and second-order co-occurrence seem to confirm the hypothesis. Term similarities could then be used for determining which query terms are useful and best reflect the user's information need. A possible application would be to use this source of evidence for tuning the weights of the query terms.

1 Introduction

Co-occurrence information, whether it is used for expanding automatically the original query (Qiu and Frei, 1993), for providing a list of candidate terms to the user in interactive query expansion, or for relaxing the independence assumption between query terms (van Rijsbergen, 1977), has been widely used in information retrieval. Nevertheless, the use of this information has often resulted in reduction of retrieval effectiveness (Smeaton and van Rijsbergen, 1983), a fact sometimes explained by the poor discriminating power of the relationships (Peat and Willet, 1991). It was not until recently that a more elaborated use of this information resulted in consistent improvement of retrieval effectiveness. Improvements came from a different computation of the relationships named "second-order co-occurrence" (Schutze and Pedersen, 1997), from an adequate combination with other sources of evidence such as relevance feedback (Xu and Croft, 1996), or from a more careful use of the similarities for expanding the query (Qiu and Frei, 1993).

Indeed, interesting patterns relying in co-occurrence information may be discovered and, if used carefully, may enhance retrieval effectiveness. This paper explores the use of co-occurrence similarities between query terms for determining the subset of query terms which are good descriptors of the user's information need . Query terms can be divided into those that are useful for retrieval and those that are harmful, which will be named respectively "content" terms and "noisy" terms. The hypothesis under examination is that two content terms tend to be more similar to each other than would be two noisy terms, or a noisy and a content term. Intuitively, the query terms which reflect the user's information need are more likely to be found in relevant documents and should concern similar topic areas. Consequently, they should be found in similar contexts in the corpus. A similarity measures the degree to which two terms can be found in the same context, and should be higher for two content terms.

We name this hypothesis the "Cluster Hypothesis for query terms", due to its correspondence with the Cluster Hypothesis of information retrieval which assumes that relevant documents "are more like one another than they are like non-relevant documents" (van Rijsbergen and Sparck-Jones, 1973, p.252). Our middle-term objective is to verify experimentally the hypothesis for different types of co-occurrences, different measures of similarity and different collections. If a higher similarity between content terms is indeed observed, this pattern could be used for tuning the weights of query terms in the absence of relevance feedback information, by increasing the weights of the terms which appear to be content terms, and inversely for noisy terms. Next section is about the verification of the hypothesis on the CACM collection (3204 documents, 50 queries).

2 Verifying the Cluster Hypothesis for query terms

2.1 The Cluster Hypothesis for query terms

The hypothesis that similarities between query terms is an indicator of the relevance of each term to the user's information need is based on an intuition. This intuition can be illustrated by the following request:

`Document will provide totals or specific data on changes to the proven reserve figures for any oil or natural gas producer.`

It appears that the only terms which appear in one or more relevant documents are *oil*, *reserve* and *gas*, which obviously concern similar topic areas, and are good descriptors of the information need [1]. All the other terms retrieve only non-relevant documents, and consequently reduce retrieval effectiveness. Taken individually, they do not seem to specifically concern the user's information need. Our hypothesis can be formulated this way:

- Content terms which are representative of the information need (like *oil*, *reserve*, and *gas*) concern similar topics and are more likely to be found in relevant documents;
- Terms which concern similar topics should be found in similar contexts of the corpus (documents, sentences, neighboring words...);
- Terms found in similar contexts have a high similarity value. Consequently, content terms tend to be similar to each other.

2.2 Determining content terms and noisy terms

Until now, we have talked of "content" or "noisy" terms, as terms which are useful or harmful for retrieval. How can we determine this? First, terms which do not occur in any relevant document can only be harmful (at best, they have no impact on retrieval) and can directly be classified as "noisy". For terms which occur in one or more relevant documents, the usefulness depends on the total number of relevant documents and on the number of occurrences of the term in the collection. We use the χ_2 test of independence between the occurrence of the term and the relevance of a document to determine if the term is a content or a noisy term. For terms which fail the test at the 95% confidence level, the hypothesis of independence is rejected, and they are considered content terms. Otherwise, they are considered noisy terms.

Another way of verifying if a term is useful for retrieval would be to compare the retrieval efficiency of the query with and without the term. This method is appealing since our final objective is better retrieval efficiency. But it has some drawbacks: (1) there are several measures of retrieval effectiveness, and (2) the classification of a term will depend in part on the retrieval system itself.

A point deserves discussion: terms which do not appear in any relevant documents and which are classified noisy may sometimes be significant of the content of the query. This may happen for example if the number of relevant documents is small and if the vocabularies used in the request and in the relevant documents are different. Anyway, this does not change the fact that the term is harmful to retrieval. It could still be used for finding expansion terms, but this is another problem. In any case, a rough classification of terms between "content" and "noisy" can always be discussed, the same way that a binary classification of documents between relevant and non-relevant is a major controversy in the field of information retrieval .

2.3 Preliminary experiments

Once terms are classified as either content or noisy, three types of term pairs are considered: content-content, content-noisy, and noisy-noisy. For each pair of query terms, different measures of similarity can be computed, depending on the type of co-occurrence, the association measure, and so on. Each of the three classes of term pairs has an a-priori probability to appear. We are interested in verifying if the similarity has an influence on this probability.

One problem with first-order co-occurrence is that the majority of terms never co-occur, because they occur too infrequently. We decided to select terms which occur more than ten times in the corpus. The same term pairs were used for first and second-order co-occurrence. Term pairs come from selected terms of the same query. For example, take a query with 10 terms of which 5 are classified content. Then for this query, there are $\frac{10*(10-1)}{2} = 45$ term pairs, of which $\frac{5*(5-1)}{2} = 10$ are content-content, 10 are noisy-noisy, and the other 25 are noisy-content.

On the 50 queries used for experiments, there are 7544 term pairs, of which 1340 (17.76%) are

[1] Remark that we do not consider here phrases such as 'natural gas', but the argument can be extended to phrases.

of class content-content, 3426 (45.41%) of class content-noisy, and 2778 (36.82%) of class noisy-noisy. 40.47% of the terms are content terms. Obviously, a term can be classified content in a query and noisy in another. In the following subsections, we present our preliminary experiments on the CACM collection.

2.3.1 First-order co-occurrence

First-order co-occurrence measures the degree to which two terms appear together in the same context. If the vectors of weights of t_i and t_j in documents d_1 to d_n are respectively $(w_{i1}, w_{i2}, ..., w_{in})^T$ and $(w_{j1}, w_{j2}, ..., w_{jn})^T$, the cosine similarity is:

$$\frac{\sum_{k=1}^{n} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \sqrt{\sum_{k=1}^{n} w_{jk}^2}} \quad (1)$$

The weight w_{ij} was set to 1 if t_i occured in d_j, and to 0 otherwise, and within document frequency and document size were not exploited. Figure 1 shows the probability to find each of the classes vs similarity. The probabilities are computed from the raw data binned in intervals of similarity of 0.05, and for the 0 similarity value. The values associated on the graph are 0 for the 0 similarity value, 0.025 for interval]0,0.05], 0.075 for]0.05,0.1], etc. The similarities after 0.325 are not plotted because there are very few of them.

There is a neat increase of probability of the class 'content-content' with increasing similarity. It is interesting to remark that if high values of similarities are evidence that the terms are content terms, small values can be taken as negative evidence for the same conclusion. By using smaller and more reliable contexts such as sentences, paragraphs or windows, it is expected that the measures of similarity should be more reliable, and the observed pattern should be stronger.

2.3.2 Second-Order co-occurrence

Second-order co-occurrence measures the degree to which two terms occur with similar terms. Terms are represented by vectors of co-occurrences where the dimensions correspond to each of the m terms in the collection. The value attributed to dimension k of term t_i is the number of times that t_i occurs with t_k. More elaborated measures take into account a weight for each dimension, which represent the discriminating value of the corresponding term. Term t_i is represented here by $(w_{i1}, w_{i2}, ..., w_{im})^T$, where w_{ij} is the number of time that t_i and t_j occur in the same context.

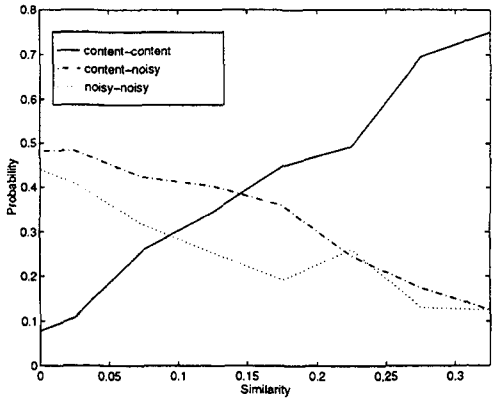

Figure 1: Probability of term pairs classes vs First-order similarity

We used again Equation 1 for computing similarities between query terms. The similarity values were in general higher than for first-order co-occurrence. Remark that the same data (term pairs) were taken for first and second-order co-occurrence. For the computation of probabilities, data were binned in intervals of 0.1, on the range $[0, 0.925]$ (not enough similarities higher than 0.925). Figure 2 represents the probabilities of the class vs similarity.

Again, the probability of having the class content-content increases with similarity, but to a lesser degree than with first-order similarity. More experiments are needed to see if first-order co-occurrence is in general stronger evidence of the quality of a term than second-order co-occurrence. However, a second-order similarity can be computed for nearly all query terms, while first-order similarities can only be computed for frequent enough terms.

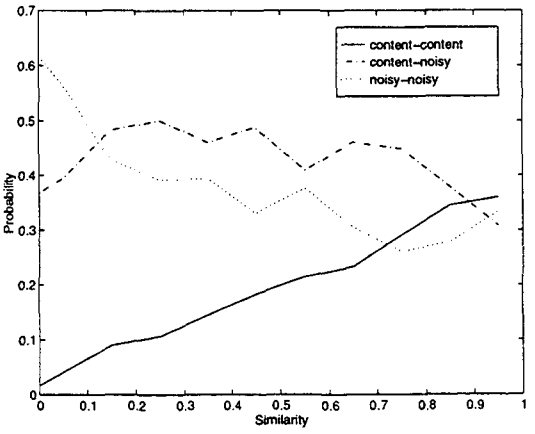

Figure 2: Probability of term pairs classes vs Second-order similarity

3 Discussion

In this paper, we have formulated the hypothesis that query terms which are good descriptors of the information need tend to be more similar to each other. We have proposed a method to verify if the hypothesis holds in practice, and presented some preliminary investigations on the CACM collection which seem to confirm the hypothesis. But many other investigations have to be done on bigger collections, involving more elaborate measures of similarity using weights, different contexts (paragraphs, sentences), and not only single words but also phrases. Experiments are ongoing on a subset of the TREC collection (200 Mb), and preliminary results seem to confirm the hypothesis. Our hope is that investigations on this large test collection should yield better results, since the computed similarities are statistically more reliable when they are computed on larger data sets.

In a way, this work can be related to word sense disambiguation. This problem has already been addressed in the field of the information retrieval, but it has been shown that the impact of word sense disambiguation is of limited utility (Krovetz and Croft, 1992). Here the problem is not the determination of the correct sense of a word, but rather the determination of the usefulness of a query term for retrieval. However, it would be interesting to see if techniques developed for word sense disambiguation such as (Yarowsky, 1992) could be adapted to determine the usefulness of a query term for retrieval.

From our preliminary investigations, it seems that similarities can be used as positive and as negative evidence that a term should be useful for retrieval. The other part of our work is to determine a technique for using this pattern in order to improve term weighting, and at the end improve retrieval effectiveness. While simple techniques might work and will be tried (e.g. clustering), we seriously doubt about it because every relationship between query terms should be taken into account, and this leads to very complex interactions. We are presently developing a model where the probability of the state (content/noisy) of a term is determined by uncertain inference, using a technique for representing and handling uncertainty named Probabilistic Argumentation Systems (Kohlas and Haenni, 1996). In the next future, this model will be implemented and tested against simpler models. If the model allows to predict reasonably well the state of each query term, this information can be used to refine the weighting of query terms and lead to better information retrieval.

Acknowledgements

The author wishes to thank Warren Greiff for comments on an earlier draft of this paper. This research was supported by the SNSF (Swiss National Scientific Foundation) under grants 21-49427.95.

References

J. Kohlas and R. Haenni. 1996. Assumption-based reasoning and probabilistic argumentation systems. In J. Kohlas and S. Moral, editors, *Defeasible Reasoning and Uncertainty Management Systems: Algorithms*. Oxford University Press.

R. Krovetz and W.B. Croft. 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141.

H.J. Peat and P. Willet. 1991. The limitations of term co-occurence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, pages 378–383, June.

Y. Qiu and H.P. Frei. 1993. Concept based query expansion. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 160–169.

H. Schutze and J.O. Pedersen. 1997. A cooccurrence-based thesaurus and two applications to information retrieval. *Information Processing & Management*, 33(3):307–318.

A.F. Smeaton and C.J. van Rijsbergen. 1983. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246.

C.J. van Rijsbergen and K. Sparck-Jones. 1973. A test for the separation of relevant and non-relevant documents in experimental retrieval collections. *Journal of Documentation*, 29(3):251–257, September.

C.J. van Rijsbergen. 1977. A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33(2):106–119.

J. Xu and W.B. Croft. 1996. Query expansion using local and global document analysis. In *Proc. of the Int. ACM-SIGIR Conf.*, pages 4–11.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *COLING-92*, pages 454–460.