Information Retrieval, 3, 243–251, 2000

Using the Co-occurrence of Words for Retrieval Weighting

ELKE MITTENDORF elke.mittendorf@systor.com
Systor AG, CH-8048, Zürich, Switzerland

BOJIDAR MATEEV mateev@eurospider.ch
PETER SCHÄUBLE schauble@eurospider.ch
Eurospider Information Technology AG, CH-8006 Zürich, Switzerland

Received August 30, 1999; Revised August 30, 1999; Accepted June 12, 2000

Abstract. We have applied the well-known Robertson-Sparck Jones weighting to sets of indexing features that are different from word-based features. Our features describe the co-occurrences of words in a window range of predefined size. The experiments have been designed to analyse the value of features that are beyond word-based features but all used retrieval methods can be motivated strictly in the probabilistic framework. Among the several implications of our experiments for weighted retrieval is the surprising result that features that describe the co-occurrences of words in sentence-size or paragraph-size windows are significantly better descriptors than purely word-based indexing features.

Keywords: probabilistic term weighting, word concurrences, term phase weighting, retrieval routing

1. Introduction

The more query words co-occur in a document the more likely is this document relevant to the query. This fact is widely used in information retrieval for the development of ranking functions (retrieval status values, RSV) that determine the order in which the documents are presented to the user. That the closeness of query words, i.e., the co-occurrence of words in a phrase or the co-occurrence of words in a sentence, paragraph, or more generally in a text window is a very good indicator for relevance, is undisputed. Moreover, if indexing is n-gram based (instead of word based) the incorporation of phrases or other co-occurrences into the retrieval process seems to be particularly protruding (Huang and Robertson 1997). Experienced users of advanced boolean search engines are very fond of operators, such as adj, with, or near, that exploit the closeness of word co-occurrences. Another reason that encourages us to look at co-occurrences as descriptions is that in areas such as texture detection in image analysis co-occurrences of features are the best-known descriptions (Hug 1996).

In contrast to the many promising aspects of co-occurrences no satisfying way to exploit the closeness of query words has been presented for weighted retrieval. Robertson (1997) points out that, so far, there is no method consistent with the probabilistic retrieval framework that can employ simultaneously words and phrases (or other types of co-occurrences). If we discard single words and use only phrases for retrieval then the precision of the search can be improved but at the same time its recall (Roth 1994) is impaired.

In weighted retrieval phrases are often used simultaneously with words for retrieval. Phrases are indexed and selected in a syntactical or just a statistically-justified way and than added to the basic indexing vocabulary based on words. The addition of phrases yields, if at all, only a moderate improvement over purely word-based retrieval. This result is disappointing because of the high significance that phrases seem to have intuitively. The lack of success is often attributed to the primitive statistical selection of phrases or the error-prone linguistic selection. However, weighted retrieval shows a high robustness against weak feature selection. We therefore believe that rather the above-mentioned probabilistic inconsistencies cause the lack of success of applying—simultaneously—phrases and words in weighted retrieval.

Other studies that involve co-occurrences of words are: (1) The work of Rijsbergen (1977) who tried to improve probabilistic retrieval by incorporating co-occurrence data into the weighting formula. The motivation of his work was to correct the “wrong” independence assumption that has to be made to derive a probabilistic weighting formula. The independence assumptions can however be replaced by the weaker “linked dependence assumption” (Fuhr 1992, Cooper 1995), which explains that the suggested use of co-occurrences does not improve retrieval. (2) There is a long history of incorporating phrases into weighted retrieval, e.g., (Fagan 1987, Croft et al. 1991) that accounts for some improvements above pure word-based retrieval. Although the success of phrases was very limited. (3) Several approaches tried to improve retrieval by adding a passage retrieval component. The aim is to exploit the closeness of query terms in long documents for a better (Knaus et al. 1994, Moffat et al. 1993, Salton et al. 1994). (4) The study of Haas and Losee (1994) shows that words co-occurring in a window of size 7 to 11 represent a natural grouping of words, which can be successfully used to improve search, though their experiments were not performed with weighted retrieval.

The strategy of our work is to avoid the inconsistencies of the simultaneous use of words and phrases or other co-occurrences. We shall work either with a solely word-based indexing vocabulary or with an indexing vocabulary that consists of solely of features that describe co-occurrences of words. We apply the well-studied probabilistic framework to co-occurrence-based features and use routing retrieval as a testbed. We do, however, combine rankings achieved with a word-based vocabulary and a co-occurrence-based vocabulary by logistic regression, since we know that logistic regression is (in contrast to the probabilistic weighting) robust in the presence of stochastic dependencies. The idea behind our approach is to only change the set of indexing features and remain completely in the well-known probabilistic framework. The aim of this work is to assess the value of co-occurrences of words for retrieval purposes, to asses, which kind of co-occurrences are useful and to find indicators how a probabilistic model incorporating co-occurrence should look like.

We have structured this paper as follows: in Section 2 we recall the probabilistic retrieval model, define new feature sets that describe different kinds of co-occurrences, and then apply the probabilistic retrieval to the new feature sets. Section 3 presents experiments with probabilistic retrieval based on co-occurrence features in the routing environment. The experiments provide some surprising and interesting results which are concluded in Section 4.

2. The co-occurrence of words in the probabilistic framework

Our approach is a straightforward approach of generalizing the Binary Independence Retrieval model and the Robertson-Sparck Jones (RSJ) weighting formula (Robertson 1977). We chose the RSJ approach because it is very well understood from a theoretical point of view. It ranks documents according to the widely-accepted probability ranking principle. Its underlying assumptions have been well studied and during the history of probabilistic retrieval the assumption have been reduced to weak and realistic assumptions (Cooper 1995, Fuhr 1992). The RSJ approach has been generalized and yields retrieval functions that are proven to yield very effective rankings (Robertson et al. 1995). We consider the probabilistic model a justification for weighted retrieval in general (Robertson and Walker 1994, Fuhr 1992). All these properties proffer the probabilistic framework for an analysis of the use of co-occurrences in weighted retrieval.

Retrieval ranking functions operate on a set of *indexing features* Φ. A feature $\varphi_i \in \Phi$ represents an equivalence class of words, such as e.g., all non-stopwords which are reduced by the Porter algorithm to the same stem. A particular occurrence of a feature is named a *token*. To distinguish these basic, word-based features from the features that we shall define on co-occurrences we call a feature $\varphi_i \in \Phi$ a *first-order* feature. Features that describe the co-occurrence of two first-order features are called *second-order* features.

Probabilistic retrieval with first-order features Assume that a given query q contains s first-order features. Let them be denoted (without loss of generality) by $\varphi_0, \dots, \varphi_{s-1}$. The RSJ approach operates on a binary description vector of the documents representing the absence or presence of query features, $d_j := (a_{j,0}, \dots, a_{j,(s-1)})' \in \{0, 1\}^s$ with $a_{j,i} = 1$ if $\varphi_i \in d_j$ and 0 else. The $(\cdot)'$ denotes the vector or matrix transpose.

Let R denote the set of relevant documents and $\bar{R}$ the complement of R. Define $p_i := P(\varphi_i \mid R)$ as the probability that the feature $\varphi_i \in \Phi$ occurs in a relevant document and by $q_i := P(\varphi_i \mid \bar{R})$ the probability that the feature $\varphi_i \in \Phi$ occurs in an irrelevant document. The query then is represented by a feature-weight vector $b := (b_0, \dots, b_{s-1})'$, with

$$b_i = \frac{p_i(1 - q_i)}{q_i(1 - p_i)}.$$

The retrieval status value according to the RSJ weighting then is the vector product of the vector descriptions of document and query.

$$\mathrm{RSV}_{\mathrm{basic}}(q, d_j) := b' d_j. \tag{1}$$

Documents that are ranked by this function are ranked according to the probability ranking principle.

Probabilistic retrieval of second-order features To be able to capture co-occurrences we need an intermediate description of a document that preserves the relative order of features, e.g., as a sequence of tokens: $d_j := \langle y_0, \dots, y_{l-1} \rangle$ where $y_i \in \Phi$. A second-order feature $\psi^{\Delta}_{kl} \in \Psi^{\Delta}$ is a feature that describes the co-occurrence of two first-order features φ_k and φ_l from the set Φ in a predefined distance range $\Delta = (\delta_0, \delta_1]$. Or more formally,

$\psi^{\Delta}_{kl} \in d_j$ if and only if there exists an occurrence y_{t_0} of φ_l and an occurrence y_{t_1} of φ_k with their distance to each other $\delta_0 < t_1 - t_0 \leq \delta_1$. If we allow the lower bound of the Δ to be negative, i.e. $\delta_0 < 0$, then Ψ^{Δ} includes the first-order features into Ψ^{Δ}. The additional condition $\delta_0 \geq 0$ or $\delta_1 = 0$ guarantees that we do not have to handle first-order features and second-order features simultaneously (in order to avoid inconsistencies in the probabilistic model, see Section 1).

Similarly to the first-order case we define document description matrices and query description matrices as a generalization of description vectors. Given is a query $q = \{\varphi_0, \ldots, \varphi_{s-1}\}$. A document description matrix D^{Δ}_j of d_j is a binary $s \times s$ matrix with elements $a_{j,kl} = 1$ if $\psi^{\Delta}_{kl} \in d_j$ and $a_{j,kl} = 0$ else. The query description matrix B^{Δ} is a $s \times s$ matrix with elements $b_{kl} = \frac{p_{kl}(1-q_{kl})}{q_{kl}(1-p_{kl})}$, p_{kl} and q_{kl} are the probabilities that $\psi^{\Delta}_{kl} \in \Psi^{\Delta}$ occurs in a relevant document or in an irrelevant document, respectively.

The retrieval status value is then defined as

$$RSV^{\Delta}(q, d_j) = \mathrm{Tr}\big(B^{\Delta\prime} \cdot D^{\Delta}_j\big), \tag{2}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace, i.e., the sum of diagonal elements, of a matrix. For each range Δ this defines a different set of features and a different retrieval method.

The following remarks are supposed to motivate and clarify this set of retrieval methods.

- The RSJ weighting method assumes that the features are linked dependent. Thus, for any possible query we have to assume linked dependence. Experiments have to reveal whether linked dependence is an appropriate assumption.
- The only difference between standard RSJ retrieval and the co-occurrence retrieval is the definition of the feature set. The word-based character of the features is, however, not a presumption of the RSJ weighting. Thus, each retrieval function RSV^{Δ} ranks the documents according to the probability ranking principle.
- For $\Delta = (-1, 0]$, the feature set Ψ^{Δ} consists of first-order features and thus RSV^{Δ} and $\mathrm{RSV}_{\text{basic}}$ are equivalent.
- For $\Delta = (0, 1]$, Ψ^{Δ} consists of phrases. This definition of a phrase is however a primitive one; it is not a definition of a phrase in a syntactic sense.
- Our definition of second-order features is order sensitives, i.e., $\psi^{\Delta}_{lk} \neq \psi^{\Delta}_{kl}$, “venetian blind” and “blind venetian” represent different features, as well as “information retrieval” and “retrieval of information”.
- RSV^{Δ} does not use feature-frequency information, an information that is very valuable for a good estimation of probability of relevance. Implicitly feature frequencies influence retrieval if the upper bound of Δ is greater than one: A feature with high frequency has a better chance to co-occur with another feature.
- It is widely accepted that locality information, such as derived from passage retrieval is valuable information. Since only co-occurrences in a window of predefined length are used, the co-occurrence indexing preserves valuable local information if the upper bound of Δ is not too large.

3. Experiments and results

An essential question that is linked with every probabilistic retrieval method is the question of *robust parameter estimation*, for RSJ weighting of second-order features of $\log(\frac{p_{lk}(1-q_{lk})}{q_{lk}(1-p_{lk}}))$, $l, k = 0, \ldots, s-1$ have to be estimated. The event of the occurrence of a particular feature is a rare event even for most first-order features φ_i (Zipf's Law). It is a hindrance for robust parameter estimation that the occurrence of a second-order feature is an even rarer event.

We decided to make the parameter estimation as easy as possible and chose the routing task as defined for TREC (Harman 1996) as our testbed for evaluating the defined methods. We must emphasize that the aim of these experiments is not to find an ultimate routing method. Routing is a very difficult task, where one has to use sophisticated methods for the selection of training material, the selection of features, and for the weighting of features, incorporating feature frequencies, optimal length normalization, etc. The experiments in our case have been designed to understand probabilistic retrieval based on second-order features and not to optimize routing retrieval, in the first place.

Training and test documents To train our methods for the 50 routing queries for TREC-4 and for the 45 routing queries for TREC-5 (Harman 1996)—the documents and the relevance information provided by TREC disks, disk1, disk2, and disk3 have been used. To test the methods we run the queries with the trained methods either against the test documents provided for TREC-4 (R4, news group data) and for TREC-5 (R5, data from the Foreign Broadcast Information Service, FBIS).

Feature selection We did not work on the original queries. In order to control the query size and to analyze its influence we selected those s first-order features that are selected by the u-measure (Mateev 1996).

Pre-selection of documents Unfortunately, the version of the SPIDER system we used for these experiments (Ballerini et al. 1997) did not administer position information. To be able to perform efficient experiments with position information we pre-selected for each query 1000 documents by a method that is known to perform well, i.e., selected 50 first-order features and 20 phrases with the u-measure and ranked the documents according to the so-called Lnu.ltn weighting (Singhal et al. 1996). All further experiments simply re-rank these lists. Average precision for the list on the R4 test set for this method is 0.3103 and for R5 it is 0.2046.

The size of the window ranges We chose a sequence of disjoint window ranges: $\Delta_0 := (-1, 0]$, $\Delta_1 := (0, 1]$, $\Delta_2 := (1, 10]$, $\Delta_3 := (10, 30]$, $\Delta_4 := (30, 200]$. The ideas behind these ranges are, for RSV^{Δ_0} we have the normal RSJ-weighting, for RSV^{Δ_1} we have probabilistic retrieval on phrases, for RSV^{Δ_2} we have retrieval on feature occurrences that describe approximately the size of a sentence, for RSV^{Δ_3} the size of a paragraph and for RSV^{Δ_4} the size of a document.

3.1. *Experiments with different query sizes*

The first set of experiments compares the performances of RSV^{Δ_0} to RSV^{Δ_4} for different query sizes. The results on R4 are shown in Table 1 and on R5 in Table 2, respectively. The best performances for each window range Δ are printed bold.

Table 1. The influence of the query size: Experiments on R4.

Query size	RSV^{Δ_0}	RSV^{Δ_1}	RSV^{Δ_2}	RSV^{Δ_3}	RSV^{Δ_4}
3	0.2595	0.1760	0.2352	0.2470	0.2520
10	0.2827	0.2614	0.2910	0.2945	0.2857
20	**0.2867**	0.2803	0.3099	0.3073	**0.2835**
40	0.2728	0.2902	0.3130	0.3043	0.2655
50	0.2681	0.2944	0.3162	0.3038	0.2617
60	0.2646	**0.2976**	**0.3168**	**0.3090**	0.2610

Table 2. The influence of the query size: Experiments on R5.

Query size	RSV^{Δ_0}	RSV^{Δ_1}	RSV^{Δ_2}	RSV^{Δ_3}	RSV^{Δ_4}
20	0.1839	0.1578	0.1915	0.1863	0.1812

Let us conclude some interesting observations

- Conventional retrieval is word based, i.e., based on first-order features such as RSV^{Δ_0}. It is very *surprising* that for query sizes larger than 10 features, retrieval on occurrences of sentence-size and paragraph-size windows is significantly better than retrieval on first-order features.
- The best RSJ-ranking is achieved by RSV^{Δ_2} that is the sentence-size window range.
- The retrieval on the second-order features composed of phrases (RSV^{Δ_1}), of co-occurrences in a sentence-size window (RSV^{Δ_2}), and of co-occurrences in a paragraph-size window (RSV^{Δ_3}) is the better the larger the queries are.
- Phrases (RSV^{Δ_1}) are in general worse than retrieval on second-order features belonging to a larger window size. For large queries however, the phrases as well as the other second-order features are better than first-order features.
- Although the second-order features perform so significantly better than first-order features, the sophisticated feature-frequency based weighting that was used as a baseline (average precisions of 0.3103 and 0.2046) is still better.
- It is remarkable that RSV^{Δ_4} for all experiments leads to a similar average precision than RSV^{Δ_0}. In addition, the influence of the query length on retrieval performance of RSV^{Δ_4} resembles its influence on the performance for RSV^{Δ_0}. We conclude from this that Δ_4 is not local enough to provide anything more than just information about presence or absence of features. Neither the preservation of the information about the order in which features occur nor about the co-occurrence itself can improve retrieval above $\mathrm{RSV}_{\mathrm{basic}}$.

Combining different window ranges with logistic regression. On the one hand, we do not want to use first-order and second-order features simultaneously to avoid inconsistencies in the probabilistic framework. On the other hand, each RSV^{Δ_i} is based on a different kind of indexing, and the descriptions preserve rather different information, although this information might not be stochastically independent. Logistic regression is robust even if the explaining variables (here $\mathrm{RSV}^{\Delta_0}, \ldots, \mathrm{RSV}^{\Delta_4}$) are not independent. We thus decided to combine the different rankings by logistic regression, i.e.,

$$\mathrm{RSV}_{comb} = \beta + \alpha_0 \mathrm{RSV}^{\Delta_0} + \cdots + \alpha_4 \mathrm{RSV}^{\Delta_4}.$$

The combination parameters $\beta, \alpha_0, \ldots, \alpha_4$ are determined with multiple logistic regression on the relevance information provided by the first three TREC disks. The statistics package S-Plus (Venables and Ripley 1994) was used. We chose the query size 20.

β	α_0	α_1	α_2	α_3	α_4
–3.2832	0.8549	0.2034	0.0304	0.1320	–0.0833

The average precision on the R4 is **0.3001**, which is a disappointing result since it is even less than the performance of RSV^{Δ_2} alone for the same query size. Even omitting RSV^{Δ_4}, which does not yield rankings highly different from RSV^{Δ_0}, did not improve the combination results.

We then decided to make a *query-specific estimation* of the parameters $\beta, \alpha_0, \ldots, \alpha_3$ (omitting the obsolete RSV^{Δ_4}). The test with query-specific parameters against R4 yields an average precision of **0.3372**, which presents a significant improvement and an even acceptable quality for a routing function.

The most interesting result of this combination experiment is that for different queries different methods $\mathrm{RSV}^{\Delta_0}, \ldots, \mathrm{RSV}^{\Delta_3}$ must be weighted differently. These different weights $\beta, \alpha_0, \ldots, \alpha_3$ are query specific but not collection specific, otherwise they would not improve retrieval on the test collection. A possible explanation for this behavior is that for different queries the query words may present sometimes rather constituents of phrases, words rather co-occur in a sentence, or are rather loosely coupled in paragraphs.

4. Conclusions

We have pointed out that so far there is no consistent way to handle first-order features and second-order features simultaneously in the probabilistic framework. This lack of appropriate models is in our opinion the reason that the success of phrases and other second-order features in weighted retrieval only small, despite the high probability with which second-order features indicate relevance. In our study we treated different feature sets separate from each other. We applied the probabilistically-derived and well-studied RSJ weighting to a new definition of indexing features. Different kinds of second-order features have been defined. We used phrase-like co-occurrences of words, order-sensitive co-occurrences of words in sentence-size windows, paragraph-size windows, or windows of a size that corresponds to an average document length.

As a testbed for the newly-defined feature sets we chose the TREC routing environment. First-order features are the basis of conventional retrieval methods. It is therefore a surprising result of our study that second-order features describing co-occurrences in a sentence-size window or in a paragraph-size window perform *significantly better* (in terms of average precision) than first-order features. It is not a new result that co-occurrences of words are important—but it is a new result that they are better than word-based features. On the other hand phrases—though they are not syntactically extracted phrases—are only better than first-order features when the query size is sufficiently large.

It is also interesting to note that features that describe the order-sensitive co-occurrence in a document-length size window yield almost the same results as first-order features. In contrast to conventional first-order features, which convey information only about the presence and absence of words, second-order features in a document-length size window convey information about the co-occurrence of words and about their relative order. We can conclude that neither the information about the co-occurrence of words (if it is not a local co-occurrence) nor the order in which they occur can improve retrieval performance.

An experiment that combined the rankings of documents on different feature sets with query-specific logistic regression yields a routing retrieval method that has an good performance, though for a better routing-retrieval method still a lot of tuning has to be done. A logistic regression across all queries is not suggestible. This difference between query-specific and unspecific combination is the most interesting result of the combination experiments: For different queries different feature sets, i.e., the co-occurrences in differently-sized windows, describe the information need in the best way and must be weighted differently.

Ideally, we want to develop a consistent probabilistic model that can deal simultaneously with first-order features and second-order features. Our study shows that, if possible, such a model is promising. It also helps to decide which kinds of word co-occurrences convey valuable information and which do not.

References

Ballerini J-P, Büchel M, Domenig R, Knaus D, Mateev B, Mittendorf E, Schäuble P, Sheridan P and Wechsler M (1997) SPIDER retrieval system at TREC-5. In: TREC-5 Proceedings.

Cooper W (1995) Some inconsistencies and misidentified modeling assumptions in probabilistic information retrieval. ACM-Transactions on Information Systems, pp. 100–111.

Croft WB, Turtle HR and Lewis DD (1991) The use of phrases and structured queries in information retrieval. In: ACM SIGIR Conference on R&D in Information Retrieval, pp. 32–45.

Fagan JL (1987) Automatic phrase indexing for document retrieval: An examination of syntactic and non-syntactic methods. In: ACM SIGIR Conference on R&D in Information Retrieval, pp. 91–101.

Fuhr N (1992) Probabilistic models in information retrieval. The Computer Journal, 35(3):243–255.

Haas SW and Losee RM jr. (1994) Looking in text windows: Their size and composition. Information Processing & Management, 30(5):619–629.

Harman D (1996) Overview of the fifth text retrieval conference (TREC-5). In: TREC-5 Proceedings.

Huang X and Robertson SR (1997) Application of probabilistic methods to Chinese text retrieval. Journal of Documentation, 53(1):74–49.

Hug J (1996) Analyse und synthese der textur von organoberflächen. Master's thesis, Institute for Communication Systems.

Knaus D, Mittendorf E and Schäuble P (1994) Improving a basic retrieval method by links and passage level evidence. In: TREC-3 Proceedings, pp. 241–246.

Mateev B (1996) Stochastic dependence of indexing features and the routing problem. Diploma Thesis, Department of Computer Science, ETH Zürich.

Moffat A, Sacks-Davis R, Wilkinson R and Zobel J (1993) Retrieval of partial documents. In: TREC-2 Proceedings.

Robertson SE (1977) The probability ranking principle in IR. Journal of Documentation, 33(4):294–304.

Robertson SE and Walker S (1994) Some simple effective approximations of the 2-Poisson model for probabilistic weighted retrieval. In: ACM SIGIR Conference on R&D in Information Retrieval. pp. 232–241.

Robertson SE, Walker S, Jones S, Hancock-Beaulieu MM and Gatford M (1995) OKAPI at TREC-3. In: TREC-3 Proceedings, pp. 109–126.

Roth M (1994) Analyse von Indexierungsmerkmalen in grossen Dokumentenkollektionen. Master's Thesis, ETH Zurich.

Salton G, Allan J, Buckley C and Singhal A (1994) Automatic analysis, theme generation, and summarization of machine-readable texts. Science, 264(3):1421–1426.

Singhal A, Buckley C and Mitra M (1996) Pivoted document length normalization. In: ACM SIGIR Conference on R&D in Information Retrieval, pp. 21–29.

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

Venables WN and Ripley BD (1994) Modern applied statistics with S-plus. In: Statistics and Computing, Springer-Verlag, New York.