Jonathan Furner, School of Information and Media Studies, and David Harper, School of Computer and Mathematical Studies, The Robert Gordon University, Aberdeen, Scotland. (Eds)

Information Retrieval Research

Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research,
Aberdeen, Scotland, 8-9 April 1997

Paper:

Where the Linked Dependence Assumption Fails and How to Move Beyond It

B. Mateev, E. Mittendorf and P. Schäuble

Published in collaboration with the
British Computer Society

Where the Linked Dependence Assumption Fails and How to Move Beyond It.

Bojidar Mateev

Institute for Operations Research

ETH Zentrum

CH-8092 Zurich

{mateev}@ifor.math.ethz.ch

Elke Mittendorf and Peter Schäuble

Institute for Information Systems

ETH Zentrum

CH-8092 Zurich

{mittendorf, schauble}@inf.ethz.ch

Abstract

The linked dependence assumption, a widely used simplification in probabilistic retrieval, is briefly reviewed and its validity is investigated. We show that the linked dependence assumption is violated particularly if the query features are good discriminators between relevant and non-relevant documents. We then propose the *manual concept generation* method, which takes into account the dependence between different query concepts. For a limited set of routing queries, experiments show a significant improvement when compared with retrieval methods which do not take into account feature dependencies. We consider this method to be the first successful step which progresses beyond the linked dependence assumption.

1 Introduction

For the past twenty years, the application of probability theory in information retrieval has resulted in powerful, and quite effective, retrieval models and methods. However, in order to derive practicable models, researchers have had to tread a careful path between theoretical purity and pragmatism. Simplifying the assumptions underlying a given model are typical of such an approach. In this way, a number of linear ranking functions have been derived, including the well-known feature frequency times inverse document frequency weighting schemes, (Fuhr 1992).

Common assumptions relate to the independence of the occurrence of indexing features, the independence of features given relevance, and the independence of features given non-relevance. These simplifying assumptions are not only crude and poor reflections of data (Gey 1993), but if assumed simultaneously they even lead to logical inconsistencies (Robertson 1974). Fortunately, Cooper proved that the linked dependence assumption (LDA)—a weakening of the simultaneous assumption of independence given relevance and independence given non-relevance—suffices for most probabilistic retrieval functions (Cooper 1995).

During the last few years, especially while observing results from interactive retrieval, a growing number of research groups have established a “feeling” that there is something wrong with the ranking algorithm in particular for short queries, e.g. (Hearst 1996) and (Rose & Stevens 1997). The co-occurrence of query features in particular seems to receive a disproportionate emphasis. Refiltering a list of probabilistically ranked documents with a boolean filter has proved to be successful (Hearst 1996). This idea has also been carried further for routing retrieval (Gey et al. 1997).

The probability ranking principle has been proved optimal from a theoretical point of view, but it is obviously not optimal in practice. Thus there must be something wrong with the theoretical assumptions. We know that the probability ranking principle is not optimal if the relevance of documents is not assessed independently (Gordon & Lenk 1991, Robertson 1977). This accounts for some of the ranking weaknesses in interactive retrieval when a user wants to address different aspects of a query (e.g. “name different alternative

sources of energy”, “which countries have weak anti-pollution laws”, etc.) However, the dependence of relevance assessments can not be the reason for the probabilistic ranking weakness in the routing task in the TREC-environment (Harman 1994), where relevance assessments are purely independent. If we rely on the probability estimates, then the question arises: how well does the linked dependence assumption fit the data? To our knowledge, this has never been investigated. We estimated values on TREC data to determine how well the data fits the linked dependence assumption, and we have concluded that the linked dependence assumption does not fit the data very well at all. Moreover, the LDA is more significantly violated for “good” query features, i.e. features possessing great discriminating power between relevance and non-relevance.

There have been several attempts to include feature dependencies into retrieval models and weighting formulas, e.g. (van Rijsbergen 1977, Harper & van Rijsbergen 1978, Eftimiadis 1993). So far these attempts have had only limited success. We believe that different reasons can be given for this. One reason is the fact that these approaches only tried to move beyond independence and not beyond linked dependence. A second reason is that the approaches did not have enough training data to estimate different dependencies. The approach of using non-linear neural networks for the routing task (Hull et al. 1996a, Hull et al. 1996b) is one case which does not rely on the LDA, but it has not yielded great success yet. The difficult estimation of too many parameters making it prone to overfitting may be responsible for this.

Inspired by our observations about the violation of the LDA, and learning from the related work, we aim to find weighting formulas that move beyond the linked dependence assumption, i.e. beyond a simple sum of feature weights. Such a function must account for the interaction between groups of features (concepts). To be able to estimate the interaction we must restrict our experiments to the routing task where training data is available. Our *feature dependence retrieval model* (FDRM) for the routing task requires a grouping into different sets, which we call the concepts of the query. The concepts are required to be independent from each other in a colloquial sense. An example: for a query about “joint ventures involving Japanese companies” a user identified the two independent concepts “joint ventures” and “Japanese companies”, and grouped the features “joint” and “venture” into the first concept and features such as “Japan”, “Tokyo”, “Kobe” into the second concept. The grouping into concepts is not trivial and is best performed by a searcher with domain knowledge. The FDRM shows a substantial improvement in terms of recall and precision compared with other successful retrieval functions. As far as we know, this is the first approach that goes beyond the linked dependence assumption.

Our paper is structured as follows: In Section 2 we give an overview of independence assumptions in probabilistic retrieval and discuss critically the linked dependence assumption. Section 3 explains how we check the validity of the LDA and discusses the results on TREC data, Section 4 describes the feature dependence retrieval model for the routing task, Section 5 and Section 6 discuss the application of this model and the experimental results.

2 Independence and Dependence Assumptions

In this section we briefly review the different independence assumptions that are used in probabilistic retrieval models and connect them to the linked dependence assumption (Cooper 1995). We then introduce some scenarios where the linked dependence assumption is violated.

The indexing features are denoted by φ_i . In our experiments for example, features are porter-reduced non-stopwords (Porter 1980). We abbreviate the event of pairs (q, d_j) where d_j contains the feature φ_i simply by φ_i . The set of all query–document pairs (q, d_j) with d_j relevant to query q is denoted by R and analogously $\bar{R}$ denotes the set of pairs (q, d_j) such that d_j is not relevant to q . See the symbol table in Appendix A for an overview of the notation. The notation is chosen according to (Schäuble 1997).

There are three independence assumptions about the occurrence of indexing features which have been used in probabilistic retrieval to get tractable formulas. The independence assumptions are:

- The independence of features $\varphi_{i_1}, \dots, \varphi_{i_n}$,

$$P(\varphi_{i_1}, \dots, \varphi_{i_n}) = P(\varphi_1) \cdots P(\varphi_n). \quad (1)$$

- The conditional independence given relevance

$$P(\varphi_{i_1}, \dots, \varphi_{i_n} | R) = P(\varphi_{i_1} | R) \cdots P(\varphi_{i_n} | R). \quad (2)$$

- The conditional independence given non-relevance

$$P(\varphi_{i_1}, \dots, \varphi_{i_n} | \overline{R}) = P(\varphi_{i_1} | \overline{R}) \cdots P(\varphi_{i_n} | \overline{R}). \quad (3)$$

These assumptions are problematic. If the features are distributed in the collection and among the relevant documents in a manner that $P(\varphi_{i_1}, R) \cdots P(\varphi_{i_n}, R) > P(\varphi_{i_1}) \cdots P(\varphi_{i_n})(P(R))^{(n-1)}$ then—no matter how the co-occurrence of features is distributed—Equations (1) and (2) can NOT hold simultaneously. This can be proved for example by contradiction in a straightforward manner. Data that is distributed in this manner exists and so is a realistic example. That data with these logical inconsistencies might exist was first mentioned in (Robertson 1974). The simultaneous assumption of (1) and (3) leads to inconsistencies in an analogous way. However, the performance of newer probabilistic models does not suffer from these kinds of inconsistencies since the assumption of (1) is superfluous.

Not all problems are solved by relying only on the assumptions of conditional independence (2) and (3). In several papers it was claimed more or less intuitively that the two conditional independence assumptions must be incorrect. The first verification of this intuition was given by F. Gey in his PhD-thesis (Gey 1993). He tested the hypothesis of independence for all three assumptions (1), (2), and (3) on CACM data. The result is that all three hypotheses can be rejected for a majority of randomly chosen query feature pairs $(\varphi_{i_1}, \varphi_{i_2})$.

Fortunately, the simultaneous assumption of (2) and (3) can be replaced by the weaker assumption of *linked dependence*:

$$\begin{aligned} P(\varphi_{i_1}, \dots, \varphi_{i_n} | R) &= C_r P(\varphi_{i_1} | R) \cdots P(\varphi_{i_n} | R) \text{ and} \\ P(\varphi_{i_1}, \dots, \varphi_{i_n} | \overline{R}) &= C_{nr} P(\varphi_{i_1} | \overline{R}) \cdots P(\varphi_{i_n} | \overline{R}) \text{ and} \\ C_r &= C_{nr}. \end{aligned} \quad (4)$$

In words: If the linked dependence assumption holds for the features $\varphi_{i_1}, \dots, \varphi_{i_n}$ then the degree of dependence of the features is the same in both the set of relevant documents and the set of non-relevant documents. It is obvious that the linked dependence assumption is weaker than the conditional independence assumptions because (2) and (3) implicate (4). Thus, it is clear that the LDA fits any data collection at least as well as the assumption of (2) and (3).

Although the LDA is a weaker, and thus a better, assumption than the conditional independence assumption, it does still not mean that it is a realistic one. To our knowledge the LDA has never been tested. What is the reason for this lack of research results? One reason might be that there are no standard test procedures for testing linked dependence as there are for testing independence. Another reason is that one needs to have this simplification to get tractable formulas. Without LDA, retrieval formulas can not be as simple as a sum over weights of query features. This is what retrieval functions nowadays almost always are. The LDA definitely helped to develop powerful retrieval models. However this is no reason to stick with this assumption forever, especially if we know it leads to non-optimal rankings.

As we stated in the introduction, there are experiments which indicate that the LDA is not always a good approximation of reality. In addition, it is very interesting for us to observe that a closer look at Gey's experiments (Gey 1993) reveals that the rejection of (2) is not as significant as the rejection of (3). If this observation is not due to an artifact from widely differing sizes of R and $\overline{R}$, it places further doubt on the validity of the LDA, because the LDA presumes an equivalent degree of dependence for both sets R and $\overline{R}$.

Figures 1, 2, and 3 demonstrate different dependence scenarios. The enclosing rectangle symbolizes the set of all documents. To the left of the vertical line the set of all relevant documents is situated, to the right the non-relevant documents. For simplicity sake, the query is limited to two features φ_1 and φ_2 . The vertical grey rectangles represent the sets of documents containing φ_1 and the horizontal grey rectangles represent the sets of documents containing φ_2 (within the subset of relevant or the subset of non-relevant documents).

Figure 1: A scenario where the linked dependence assumption holds.

Figure 2: The linked dependence assumption is violated for a feature φ_2 that discriminates well between relevant and non-relevant documents (“good feature”).

The intersection is represented in black. Figure 1 illustrates an example where the LDA holds (in this case even the two independence assumptions (2) and (3) hold). We want to illustrate how the LDA might be violated. Figure 2 and Figure 3 indicate two different examples where the LDA does not hold. This can be compared with the example in Figure 1. We choose examples with $C_r = 1$ since in our experiments C_r is often estimated to be close to 1. We show that there is a connection between the ability of the features to discriminate between relevant and non-relevant documents and the violation of the LDA. Therefore Figures 2 and 3 show examples of different kinds of features.

- If we compare Figure 2 with Figure 1, the probability $P(\varphi_1, \varphi_2 | \bar{R})$ remains constant and $P(\varphi_2 | \bar{R})$ decreases. This means that the feature φ_2 in the case of Figure 2 is a better discriminator than in the case of Figure 1. In Figure 2 the LDA is violated because $C_{nr} > 1$.
- If we compare Figure 3 with Figure 1, the probability $P(\varphi_2 | \bar{R})$ remains constant but $P(\varphi_1, \varphi_2 | \bar{R})$ increases. This means that the feature φ_2 in the case of Figure 2 is a worse discriminator than in the case of Figure 1. Again, in Figure 3 the LDA is violated because $C_{nr} > 1$.

3 Exploring the Linked Dependence Assumption

In this section we explain how the linked dependence assumption is examined. We describe two extreme cases for the set of features (Section 3.1), and two extreme cases for the set of documents (Section 3.2)

Figure 3: The linked dependence assumption is violated for a feature φ_2 that does not discriminate well between relevant and non-relevant documents (“bad feature”).

for which the estimations of C_r and C_{nr} are performed. In Section 3.3 we present our idea for examining the LDA according to the different cases of features and documents. Experiments are then described in Section 3.4.

3.1 “Good” and “Standard” Features

In this section we describe how we select the features in $\Phi(q)$ for which the LDA is tested. We shortly describe a feature selection method (U-features) which has been statistically derived in (Mateev 1996) and (Ballerini et al. 1996).

The set $\Phi(q)$ can be obtained in many different ways: e.g. from the original topic (Lewis et al. 1996), from Rocchio expansion (Buckley et al. 1995), using the χ^2 -test (Hull et al. 1996a), etc. The quality of the features is always w.r.t. a given weighting scheme, but normally features which perform well for one weighting scheme are also suitable for others. We examine the LDA in two extreme cases. The first is where $\Phi(q)$ consists of features belonging to the original query topic (**standard features**). The second case is where $\Phi(q)$ consists of features selected through a very good feature selection algorithm (**good features**).

Intuitively a good feature is one which discriminates well between relevant and non-relevant documents, i.e. the feature occurs more often in relevant documents than in non-relevant ones. We are interested in how this influences the coefficients C_r and C_{nr} . We decided to work with U-features. The method consists of building a contingency table (Table 1) for each feature $\varphi_i \in \Phi^{D_T}$, where Φ^{D_T} is the set of all features from the training documents D_T .

	doc. relevance	doc. non-relevance	Total
φ_i presence	x_{11}	x_{12}	$x_{1.}$
φ_i absence	x_{21}	x_{22}	$x_{2.}$
Total	$x_{.1}$	$x_{.2}$	$x_{..}$

Table 1: 2x2 contingency table

Using these contingency tables one can rank the features according to (Andersen 1992):

$$U(x_{11}, x_{12}, x_{21}, x_{22}) = \sqrt{x_{..}} \frac{x_{11}x_{22} - x_{12}x_{21}}{\sqrt{x_{1.}x_{.1}x_{2.}x_{.2}}}. \quad (5)$$

The N features with the highest value are selected. These features build the set $\Phi(q)$.

We already compared the U-features with other feature selection methods in (Mateev 1996) and (Ballerini et al. 1996). The U-features out-performed all other strategies tested. Our experiments show that the LDA is much more significantly violated for good features than for the standard ones.

3.2 “Difficult” and “Random” Non-relevant Documents

For a given query, usually the methods for routing retrieval do not use the whole set of non-relevant documents but only a small part of it because the set is too large. This raises the question of the impact of the chosen subset on the LDA. We distinguish between subsets containing random non-relevant documents and subsets containing difficult non-relevant documents. A **random non-relevant document** is a document which is randomly selected from the set of all non-relevant documents w.r.t. query q . A **difficult non-relevant document** is a non-relevant document which is hard to distinguish from relevant ones. Typically, a difficult non-relevant document contains many of the features used in the query. Since the features from our feature set $\Phi(q)$ occur more often in such difficult non-relevant documents than in random ones, the coefficient C_{nr} might be affected. The set of relevant documents remains constant in the experiments.

Our experiments use documents with relevance assessments from the TREC collection. Harman (1995) describes how such documents with relevance assessments are obtained. Our experiments show that the LDA is much more significantly violated in collections where the non-relevant documents are difficult ones, than in collections with random non-relevant documents.

3.3 How to Examine the Linked Dependence Assumption

We are not aware of any methods to statistically test the LDA. We have therefore developed statistics to indicate how well the LDA is met. The first step is to estimate the coefficients C_r and C_{nr} :

$$C_r = \frac{|\{d_j \in R | \varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_n} \in d_j\}|}{|R| \frac{|\{d_j \in R | \varphi_{i_1} \in d_j\}|}{|R|} \dots \frac{|\{d_j \in R | \varphi_{i_n} \in d_j\}|}{|R|}}$$

$$C_{nr} = \frac{|\{d_j \in \overline{R} | \varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_n} \in d_j\}|}{|\overline{R}| \frac{|\{d_j \in \overline{R} | \varphi_{i_1} \in d_j\}|}{|\overline{R}|} \dots \frac{|\{d_j \in \overline{R} | \varphi_{i_n} \in d_j\}|}{|\overline{R}|}}$$

where $|R|$ is the number of relevant documents, $|\{d_j \in R | \varphi_{i_1} \in d_j\}|$ the number of relevant documents containing φ_{i_1} , and $|\{d_j \in R | \varphi_{i_1}, \varphi_{i_2}, \dots, \varphi_{i_n} \in d_j\}|$ is the number of relevant documents where the features occur together. The notations for $\overline{R}$ are analogous.

Our idea is to examine the LDA for different numbers of features n , and for features obtained from different feature selection algorithms. We experiment with a set of features $\Phi(q)$, for a query q , and a fixed value for n , where $n < N = |\Phi(q)|$. It makes sense to experiment only with small values for n , otherwise we can not estimate $P(\varphi_{i_1}, \dots, \varphi_{i_n} | R)$ and $P(\varphi_{i_1}, \dots, \varphi_{i_n} | \overline{R})$.

Suppose that the case where $C_r < C_{nr}$ occurs on average as often as the case where $C_r > C_{nr}$. This would mean that the LDA is a priori not a bad assumption, and that there is no reason to reject it, since we can hope that the generated errors can compensate for each other. If one of the two inequalities $C_r < C_{nr}$ or $C_r > C_{nr}$ receives a significantly higher count of events, then we have good reason to assume that the LDA is not a reliable assumption for retrieval. Note that this is not a powerful statistic like for example a statistical test, rather it is a conservative decision rule that is very careful in rejecting the LDA.

According to our considerations in Section 3.1 and Section 3.2 we distinguish between a set $\Phi(q)$ containing good features (w.r.t. q) and a set containing standard features. Moreover the set of non-relevant documents needed to estimate C_{nr} can consist of randomly selected documents or of difficult ones. Four cases are checked:

1. good features and difficult non-relevant documents
2. standard features and difficult non-relevant documents
3. good features and random non-relevant documents
4. standard features and random non-relevant documents.

Our results show that the LDA behaves differently in each of these four cases.

3.4 Results

Experiments are performed with a subset of the TREC-4 routing queries as a first setup. This subset consists of all TREC-4 routing queries with TREC topic numbers smaller than 50. There are 23 such queries. These experiments use all available relevance information on disk1, disk2 and disk3 (Harman 1996). Another experiment setup involved TREC queries 202–250, and the relevance information from disk2 and disk3. Since the results from the two setups are similar, the results from the second experiment setup are not listed here. The LDA is examined for two and three features. The coefficients C_r and C_{nr} are computed for all pairs (resp. triplets) from $\Phi(q)$, where the feature set $\Phi(q)$ consists of:

- all features from the original queries
- the set of the best 20 U-features.

The set of documents where C_r and C_{nr} are estimated is:

- the TREC training set – this set contains many difficult non-relevant documents
- the randomized TREC training set: all non-relevant documents from the original set which are not relevant for any other query are assigned at random to the different queries – we assume that for each query the non-relevant documents obtained in this way are random.

Combining the different possibilities for features and sets of documents gives us four cases to investigate: (good features, difficult documents), (standard features, difficult documents), (good features, random documents), (standard features, random documents).

In order to examine the LDA we observe:

1. the coefficients C_r and C_{nr}
2. the median of $\frac{C_r}{C_{nr}}$
3. the average values of the estimated probabilities: $P(\varphi_1, \varphi_2 | R)$, $P(\varphi_1 | R)$, $P(\varphi_2 | R)$, $P(\varphi_1, \varphi_2 | \overline{R})$, $P(\varphi_1 | \overline{R})$, $P(\varphi_2 | \overline{R})$ over all pairs of features.

Table 2 shows some statistics about the coefficients C_r and C_{nr} over all pairs (resp. triplets) which were built. The size for good features and random non-relevant documents in Table 2 is smaller than that for good features and difficult non-relevant documents. Fewer pairs occur together in these documents, i.e. there are more cases where the coefficient C_{nr} can not be estimated.

The results for the good features show that there are many more cases where $C_r < C_{nr}$ than where $C_r > C_{nr}$. For standard features there is less asymmetry between both cases. This means that the LDA is more significantly violated for the good features than for the standard ones. For the same reason, the LDA is more significantly violated for document sets containing difficult non-relevant documents than for document sets containing random non-relevant documents. The LDA is also more significantly violated for triplets than for pairs.

Table 3 shows information about the median, the 0.25- and the 0.75-quantile for $\frac{C_r}{C_{nr}}$ over all pairs (resp. triplets) which were built. Results are shown only for difficult non-relevant documents.

	feature set	document set	size	$C_r < C_{nr}$	$C_r > 1$	$C_{nr} > 1$
pairs	standard	difficult	957	57.4 %	80.9 %	87.4 %
pairs	standard	random	957	76.3 %	81.0 %	99.2 %
pairs	good	difficult	3713	89.5 %	77.4 %	90.8 %
pairs	good	random	3506	96.4 %	77.5 %	97.4 %
triplets	standard	difficult	2404	57.8 %	89.6 %	92.4 %
triplets	good	difficult	18261	94.5 %	84.9 %	95.7 %

Table 2: Examining the LDA through statistics on C_r and C_{nr} .

	feature set	0.25-quantile	median	0.75-quantile
pairs	standard	0.83	0.96	1.14
pairs	good	0.48	0.70	0.84
triplets	standard	0.69	0.91	1.08
triplets	good	0.2	0.41	0.60

Table 3: Examining the LDA by comparing the median of C_r/C_{nr} for difficult documents.

One can see that for the standard features the median is close to one, which means that the LDA holds on average. Moreover, the 0.25- and the 0.75-quantiles are placed more or less symmetrically w.r.t. the median. For the good features, the median is much lower than one, and again both quantiles are placed more or less symmetrically w.r.t. the median. These results confirm our observations in Table 2.

We also estimate the probabilities $P(\varphi_1, \varphi_2|R)$, $P(\varphi_1|R)$, $P(\varphi_2|R)$ and the analogous for $\bar{R}$ over all pairs of features for the four cases described above. We then compute the averages of these estimations. Figure 4 shows the results in percentages.

One can see that the C_r remains more or less constant. The constant C_{nr} is larger both for good features than for standard ones, and for random non-relevant documents than for difficult ones. Having C_r constant means that larger C_{nr} values lead to considerable infringements of the LDA. This once again confirms the results from the previous two experiments.

All three experiments show that the LDA is more significantly violated for good features than for standard ones, and for random non-relevant documents than for difficult ones.

4 A Feature Dependence Retrieval Model

Inspired by the investigations described in the previous section, we tried to find weighting formulas that move beyond the LDA. Considering all possible dependencies, however, yields intractable formulas. In addition, we would like to deal with only a few free parameters in order to avoid overfitting, which is a severe problem for routing (Hull et al. 1996b). Thus we try to move beyond LDA only in cases where we can expect the LDA to be violated heavily. Encouraged by the work of Hearst (1996), we decided to group features together into so-called *concepts*. The dependencies between the concepts (which is captured in Hearst’s approach by the Boolean AND-operator) must be modelled in the probabilistic retrieval model.

A query is represented by a set of features, e.g. the U-features described in Section 3.1. The partitioning of the features is called the *concept generation*. It can be performed manually—*manual concept generation*—or automatically. We focus here on the manual concept generation because we do not yet know a good automatic partitioning algorithm which serves our goals.

The aim of probabilistic retrieval is to estimate the probability of relevance as exactly as possible, or

	"Difficult" non-relevant documents	"Random" non-relevant documents																
"standard" features	<table><tr><td>R</td><td>36.6</td><td>31.3</td><td>R</td></tr><tr><td>36.1</td><td>15</td><td>12.6</td><td>31.9</td></tr></table> Cr=1.14Cnr=1.26	R	36.6	31.3	R	36.1	15	12.6	31.9	<table><tr><td>R</td><td>36.7</td><td>21.2</td><td>R</td></tr><tr><td>35.8</td><td>15</td><td>7.3</td><td>20.8</td></tr></table> Cr=1.14Cnr=1.66	R	36.7	21.2	R	35.8	15	7.3	20.8
R	36.6	31.3	R															
36.1	15	12.6	31.9															
R	36.7	21.2	R															
35.8	15	7.3	20.8															
"good" features	<table><tr><td>R</td><td>45.8</td><td>23.8</td><td>R</td></tr><tr><td>45</td><td>23.1</td><td>8.3</td><td>22.7</td></tr></table> Cr=1.12Cnr=1.53	R	45.8	23.8	R	45	23.1	8.3	22.7	<table><tr><td>R</td><td>46.9</td><td>11.7</td><td>R</td></tr><tr><td>46.2</td><td>24.2</td><td>2.7</td><td>10.9</td></tr></table> Cr=1.12Cnr=2.11	R	46.9	11.7	R	46.2	24.2	2.7	10.9
R	45.8	23.8	R															
45	23.1	8.3	22.7															
R	46.9	11.7	R															
46.2	24.2	2.7	10.9															

Figure 4: Examining the LDA through estimations of the average probabilities. Results shown in percentages.

rather to estimate a monotonous function of the probability of relevance. A commonly used function is:

$$\text{logit}(P(R|q, d_j)) := \log \left(\frac{P(R|q, d_j)}{P(\bar{R}|q, d_j)} \right).$$

It holds that

$$\text{logit}(P(R|q, d_j)) = \log \left(\frac{P(R|q, d_j)}{P(\bar{R}|q, d_j)} \right) = \log \left(\frac{P(R, q, d_j)}{P(\bar{R}, q, d_j)} \right) = \log \left(\frac{P(d_j|R, q)}{P(d_j|\bar{R}, q)} \right) + K(q, R), \quad (6)$$

where $K(q, R)$ is a constant not depending on d_j . At this point usually the LDA is applied and the probability is split into a product over features (Robertson 1977), (Robertson & Walker 1994). We intervene here, however, to consider the query $q = C_1 \cap C_2$ as the intersection of two dependent, but not linked dependent, concepts:

$$\begin{aligned} \text{logit}(P(R|q, d_j)) &= \log \left(\frac{P(R, C_1 \cap C_2, d_j)}{P(\bar{R}, C_1 \cap C_2, d_j)} \right) = \log \left(\frac{C_r(d_j)P(R, C_1, d_j)P(R, C_2, d_j)}{C_{nr}(d_j)P(\bar{R}, C_1, d_j)P(\bar{R}, C_2, d_j)} \right) \\ &= \log \left(\frac{P(R, C_1, d_j)}{P(\bar{R}, C_1, d_j)} \right) + \log \left(\frac{P(R, C_2, d_j)}{P(\bar{R}, C_2, d_j)} \right) + \log \left(\frac{C_r(d_j)}{C_{nr}(d_j)} \right) \\ &= \log \left(\frac{P(d_j|R, C_1)}{P(d_j|\bar{R}, C_1)} \right) + \log \left(\frac{P(d_j|R, C_2)}{P(d_j|\bar{R}, C_2)} \right) + \log \left(\frac{C_r(d_j)}{C_{nr}(d_j)} \right) + K(C_1, R) + K(C_2, R). \end{aligned} \quad (7)$$

The last step in this derivation uses Equation (6). Note that constants $C_r(d_j)$ and $C_{nr}(d_j)$ which fulfill the preceding equation always exist (as long as we have not to deal with zero probabilities). These constants capture the dependencies of the concepts C_1 and C_2 . Note also that these constants are not the same as C_r and C_{nr} . In contrast to C_r and C_{nr} which have been estimated in Section 3.3, the coefficients $C_r(d_j)$ and $C_{nr}(d_j)$ are NOT independent from the document d_j , and thus the document d_j must influence the estimation of $\log \left(\frac{C_r(d_j)}{C_{nr}(d_j)} \right)$.

5 Towards the Application of the Feature Dependence Retrieval Model

A reliable estimation of $\log\left(\frac{C_r(d_j)}{C_{nr}(d_j)}\right)$ relative to a document is a difficult and crucial step in going beyond the linked dependence assumption. Assuming LDA for the features in C_1 and C_2 , the first and second terms in Equation (7) can be treated as usual in probabilistic retrieval, and can be approximated by a well known retrieval function. A very successful RSV-function is the so-called Lnu.ltn weighting scheme (Singhal et al. 1996). Using this weighting scheme we approximate Equation (7) by

$$\text{logit}(P(R|q, d_j)) = \alpha \text{RSV}(C_1, d_j) + \beta \text{RSV}(C_2, d_j) + \gamma \text{RSV}(C_1, d_j) \text{RSV}(C_2, d_j) + \delta,$$

where $\alpha, \beta, \gamma, \delta \in \mathbb{R}$. In this equation, $\log(C_r(d_j)/C_{nr}(d_j))$ is estimated as $\gamma \text{RSV}(C_1, d_j) \text{RSV}(C_2, d_j)$. We need α and β in the formula since we do not assume that RSV is normalized over the queries or concepts. The unknown parameters α, β, γ , and δ are estimated with logistic regression on the training data for the routing task.

Unfortunately, experiments with this formula have not been very promising. We observe that some features from C_1 and C_2 seem to belong to both concepts, and that other features can be classified neither to C_1 nor to C_2 . A “tuning” of the manual concept generation method is necessary. We add two “pseudo-concepts” C_{12} and C_{garb} to our algorithm. The pseudo-concept C_{12} contains all features which belong to C_1 and C_2 in the opinion of the person who generated the concepts. The features from C_{12} are only used to improve the estimation of the dependence term $\log(C_r(d_j)/C_{nr}(d_j))$. The pseudo-concept C_{garb} (“garbage” concept) contains features which can be classified neither to C_1 nor to C_2 . The concept C_{garb} can also contain features which the person who generates the concepts does not understand. The features from C_{garb} are not used in the retrieval function. Altogether we have

$$\begin{aligned} \text{logit}(P(R|q, d_j)) = & \alpha \text{RSV}(C_1, d_j) + \beta \text{RSV}(C_2, d_j) \\ & + \gamma \text{RSV}(C_1, d_j) \text{RSV}(C_2, d_j) \text{RSV}(C_{12}, d_j) + \delta. \end{aligned} \quad (8)$$

The design strategy for this formula is to keep the model as simple as possible (fewer parameters for estimation), and at the same time model the dependency between C_1 and C_2 in the way described above. We do not include a linear term for $\text{RSV}(C_{12}, d_j)$ in the function above because C_{12} has meaning only w.r.t. C_1 and C_2 . One could include the terms $\text{RSV}(C_1, d_j) \text{RSV}(C_{12}, d_j)$ and $\text{RSV}(C_2, d_j) \text{RSV}(C_{12}, d_j)$. This would yield more degrees of freedom to model, for example, a stronger dependence between C_{12} and C_1 than between C_{12} and C_2 . We did not experiment with such terms because of lack of time and also because more parameters in the logit function require more training documents.

Here is a small example which illustrates the ideas described above: Consider the TREC-4 query Topic 3 which concerns the announcement of new joint ventures involving Japanese companies and also the activities of these joint ventures. One of the authors plays the role of the questioner. The subjective opinion of the questioner is that one can define two more or less independent concepts “joint ventures” and “Japan” by putting automatically selected features into the sets C_1 and C_2 respectively. An arbitrary feature selection algorithm can be used to obtain features. The U-features described in Equation (5) are used for this experiment. For Topic 3 the top 20 features are: {ventur, japan, nippon, prm, joint, tokyo, corp, rim, japanes, sumitomo, cyc, kobe, laser, osaka, kawasaki, yen, mitsubishi, pacif, nihon, steel}. The features “joint” and “venture” can—in the questioner’s subjective opinion—be put in C_1 , and features like “japan”, “tokyo” or “japanes” in C_2 . Features like “nippon” (another name for Japan, but also a first word in some japanese company and joint-venture names) or “kawasaki” (japanese company but also partner in joint-venture companies) are common to both concepts mentioned above (probably depending on the linguistic context). For this reason, one can put such features in a pseudo-concept C_{12} which is justified only given C_1 and C_2 . There are features, like “prm” or “rim”, which could not be identified by the questioner and which were put in another concept C_{garb} . Again we emphasize that this concept generation is subjective and other questioners might perform it in a quite different manner.

Now we formalize the way manual concepts are generated and used. It is assumed that the query topic can be described by two more or less independent concepts:

1. Let S be the set of the top 20 features selected for a query topic using a feature selection algorithm (for example U-features described in Section 3.1).
2. Find two concepts C_1 and C_2 in the query topic which are as independent as possible in a colloquial sense. Describe these concepts through features from S and mark these features as used.
3. Group all unused features from S which have something in common with C_1 **and** C_2 in a pseudo-concept C_{12} .
4. Put the remaining features from S into a pseudo-concept C_{garb} . These features are either not familiar to us or they have nothing to do with C_1 or C_2 (in the questioner’s subjective estimation).
5. Compute the coefficients for the model (8) using logistic regression over the training data.
6. Perform ranking on the test data using the computed model.

The advantage of manual concept generation is that human knowledge about the topic is included in the query. Disadvantages are that the manual generation of concepts is time consuming and that concept identification is very subjective. There are problems determining the two most important and independent concepts for a query topic and defining these concepts through the given features. These problems are illustrated in the following experiment:

Different people were asked to build concepts for query Topic 3 described above. They had to describe with few words the concepts C_1 and C_2 , and to put the given 20 features into the four different groups. Most people described C_1 and C_2 as “joint venture” and “Japan”. Some interpreted these concepts in a very broad sense and others very strictly. Although it appears that there is similarity between the described concepts there are great differences in the way the features are put into these concepts. Such experiments also show us that very often we need the knowledge of an expert on the given topic. Another disadvantage is that the process of stemming words to features is not bijective, and sometimes it is difficult to determine the meaning of the original word from the selected feature.

6 Preliminary Experiments

Our experiments are performed using the SPIDER information retrieval system described in (Knaus et al. 1996). We experiment with the query topics 3, 5, 14 and 20 from the TREC-4 routing task. The query set is very limited for two reasons: First, the method cannot be applied for every query because some queries might consist of only one or more than two different concepts; and Second, the manual concept generation is a very time consuming process and we did not have more time to generate queries. Note however that the four queries have not been selected because they yield the best results, but were selected arbitrarily.

The training data to estimate the regression parameters and to select optimal features consists of all available TREC-4 routing training documents for these queries on disk1, disk2 and disk3 (about 1800 relevant and non-relevant documents per query). Feature selection selected the 20 best U-features (Equation 5). For each topic we generate concepts from the selected features. For each query the coefficients α , β and γ in Equation 8 are computed using logistic regression over the training data, with $RSV_{Lnu.ttn}$ as the basic weighting function. The test data is the TREC-4 routing test data.

For efficiency reasons, we perform re-ranking on an existing list of ranked documents from the TREC-4 routing task instead of retrieving documents from the whole test set. The retrieval method used to build this list is described in (Buckley et al. 1996), and is obtained by dynamic feedback optimization (DFO). The DFO was one of the best TREC-4 routing methods. We refer to this list as **DFO-list**. The average precision per query is shown in Figure 5. It is compared with the average precision per query of the original DFO-list,

Figure 5: Manual concept generation for TREC-4 query topics

and with the best average precision per query obtained by re-ranking on the DFO-list using *Lnu.ltn* on the top 3, 20, or 50 features from the U-features list.

The manual concept generation, together with *Lnu.ltn* as the weighting function for the different RSV's, shows very good results. It performs always better than the poor *Lnu.ltn*, and in three of four cases better than the DFO method, which is much more sophisticated.

It is still an open question how this algorithm can be extended for N such concepts. When there are more than two concepts then the most important two can be used, and the features for the others can be put in C_{12} . If there is only one concept then the algorithm is reduced to the normal weighting function applied over all features.

7 Summary and Future Work

We question the linked dependence assumption, and have been able to present evidence indicating that the dependence of features in the set of relevant documents, and the dependence in the set of non-relevant documents, are not as tightly linked as the linked dependence assumption supposes. The dependencies have been analysed on feature pairs and on feature triplets. The violation of the LDA is usually produced by a stronger dependence in the set of non-relevant documents. So although one may hope that different violations of the linked dependence assumption compensate for each other, in fact they do not. It is alarming that the violation is more significant on “good” query features, i.e. features that have a strong capability to discriminate between relevant and non-relevant documents.

We have also introduced a first approach to move beyond the linked dependence assumption for the routing task: the feature dependence model with manual concept generation. Manual concept generation is a very careful manual grouping of query features into two concepts, a set of features that interact with both concepts, and a set of garbage features. Together with the feature dependence model, a retrieval method that pays attention to the dependency of the two concepts— on a small query set the manual concept generation yields a substantial increase in average precision over other very good weighting functions which do not pay attention to these dependencies.

In our future work we must get beyond the limited usability of the feature dependence method based on manual concept generation. Methods for more than two concepts must be generated. Automatic methods to split features into concepts are needed in order to enable users without a very deep domain knowledge to use this manual concept generation. And finally one should get beyond the linked dependence assumption for ad-hoc queries, especially for short queries. For the ad-hoc task we would like to exploit the result that C_r is usually close to one and that C_{nr} is considerably greater than one for “good” features and for “random” documents.

A Symbols used in the paper

The symbols used in this report are shown in Table 4.

Symbol	Definition
Φ^{D_T}	set of features in the training collections
$\Phi(q)$	set of features for query q on which the LDA is checked
C_r, C_{nr}	coefficients in the LDA (Equation 4)
C_1, C_2	main concepts in the manual query generation
C_{12}	additional concept in the manual query generation
C_{garb}	“garbage” concept in the manual query generation
$\mathbb{R}$	the set of the real numbers

Table 4: Used symbols

References

- Andersen, A. (1992), *The Statistical Analysis Of Categorical Data*, Springer - Verlag , Berlin.
- Ballerini, J., Büchel, M., Domenig, R., Knaus, D., Mateev, B., Mittendorf, E., Schäuble, P., Sheridan, P. & Wechsler, M. (1996), SPIDER Retrieval System at TREC5, *in* ‘TREC-5 Proceedings’.
- Buckley, C., Salton, G., Allan, J. & Singhal, A. (1995), Automatic Query Expansion Using SMART: TREC-3, *in* ‘TREC-3 Proceedings’, pp. 69–80.
- Buckley, C., Singhal, A. & Mitra, M. (1996), New Retrieval Approaches Using SMART: TREC-4, *in* ‘TREC-4 Proceedings’.
- Cooper, W. (1995), ‘Some Inconsistencies and Misidentified Modeling Assumptions in Probabilistic Information Retrieval’, *ACM Transactions on Information Systems* **13**(1), 100–111.
- Eftimiadis, E. (1993), A User-Centered Evaluation of Ranking Algorithms for Interactive Query Expansion, *in* ‘ACM SIGIR Conference on R&D in Information Retrieval’, pp. 146–159.
- Fuhr, N. (1992), ‘Probabilistic Models in Information Retrieval’, *The Computer Journal* **35**(3), 243–255.
- Gey, F. (1993), Probabilistic Dependence and Logistic Inference in Information Retrieval, PhD thesis, University of California at Berkeley.
- Gey, F., Chen, A., He, J., Xu, L. & Meggs, J. (1997), Term importance, Boolean conjunct training, negative terms, and foreign language retrieval: probabilistic algorithms at TREC-5, *in* ‘TREC-5 Proceedings’. to appear.
- Gordon, M. & Lenk, P. (1991), ‘A Utility Theoretic Examination of the Probability Ranking Principle in Information Retrieval’, *Journal of the ASIS* **42**(10), 703–714.
- Harman, D. (1994), Overview of the Second Text REtrieval Conference (TREC-2), *in* ‘TREC-2 Proceedings’, pp. 1–20.
- Harman, D. (1995), Overview of the Third Text REtrieval Conference, *in* ‘TREC-3 Proceedings’, pp. 1–19.
- Harman, D. (1996), Overview of the Fourth Text REtrieval Conference (TREC-4), *in* ‘TREC-4 Proceedings’, pp. 1–23.

- Harper, D. J. & van Rijsbergen, C. (1978), 'An evaluation of feedback in document retrieval using co-occurrence data', *Journal of Documentation* **34**(3), 189–216.
- Hearst, M. (1996), Improving Full-Text Precision on Short Queries Using Simple Constraints, in 'Symposium on Document Analysis and Information Retrieval', pp. 217–232.
- Hull, D., Pedersen, J. & Schütze, H. (1996a), Document Routing as Statistical Classification, in 'AAAI Spring Symposium on Machine Learning in Information Access', pp. <http://www.parc.xerox.com/istl/projects/mlia/mlia-papers.html>.
- Hull, D., Pedersen, J. & Schütze, H. (1996b), Method Combination for Document Filtering, in 'ACM SIGIR Conference on R&D in Information Retrieval', pp. 279–287.
- Knaus, D., Mittendorf, E., Schäuble, P. & Sheridan, P. (1996), Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System, in 'TREC-4 Proceedings'.
- Lewis, D., Shapire, R., Callan, J. & Papka, R. (1996), Training algorithms for linear text classifiers, in 'ACM SIGIR Conference on R&D in Information Retrieval', pp. 298–306.
- Mateev, B. (1996), Stochastic Dependence of Indexing Features and the Routing Problem. Diploma Thesis, Department of Computer Science, Institute for Information Systems ETH Zürich.
- Porter, M. F. (1980), 'An Algorithm for Suffix Stripping', *Program* **14**(3), 130–137.
- Robertson, S. E. (1974), 'Specificity and weighted retrieval', *Journal of Documentation* **30**, 41–46.
- Robertson, S. E. (1977), 'The Probability Ranking Principle in IR', *Journal of Documentation* **33**(4), 294–304.
- Robertson, S. E. & 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.
- Rose, D. & Stevens, C. (1997), V-twin: A lightweight engine for interactive use, in 'TREC-5 Proceedings'. to appear.
- Schäuble, P. (1997), *Multimedia Information Retrieval—Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer Academic Publishers, Boston/London/Dordrecht.
- Singhal, A., Buckley, C. & Mitra, M. (1996), Pivoted Document Length Normalization, in 'ACM SIGIR Conference on R&D in Information Retrieval', pp. 21–29.
- van Rijsbergen, C. (1977), 'A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval', *Journal of Documentation* **33**, 106–119.