Integration of Collocation Statistics into the Probabilistic Retrieval Model

Olga Vechtomova

Centre for Interactive Systems Research,
Department of Information Science,
City University, London, UK

Stephen Robertson

Centre for Interactive Systems Research,
Department of Information Science,
City University, London, UK
and
Microsoft Research Ltd,
Cambridge, UK

Abstract

The paper presents a method of combining corpus information on word collocations with the probabilistic model of information retrieval. Corpus term dependencies are used to modify the probabilistic retrieval based on the term independence assumption. Collocates are derived from windows around term occurrences in the corpus. Statistical measures of mutual information and Z score are applied to select significantly associated collocates which are later used in query expansion. The results of the lexico-semantic analysis of significant collocates and their comparison with engineered term networks and thesauri are also discussed.

1 Introduction

The setting in which information retrieval operates is dynamic. Emerging factors, such as bigger and heterogeneous document collections, growing dominance of full-text retrieval and an increase in demand for higher precision in retrieval stimulate re-assessment of well established methods and investigation of new approaches. The general probabilistic approach to information retrieval, developed by Maron and Kuhns, and specifically the probabilistic model developed by Robertson and Sparck Jones [17], are flexible enough to provide for a range of retrieval variables and conditions. The probabilistic model has been extended through time into a number of variations, which use additional sources of information, such as document length, best-matching passages and statistically defined phrases. The ongoing project reported in this paper is an attempt to integrate information on word collocations in corpus into the probabilistic model. The paper presents the motivation and need for such integration and details the process of the research work underway.

The probabilistic model was built with a strong assumption of independence of attributes describing documents, i.e. terms [17]. This is, of course, a conventional statement, which is necessary for implementation of the model. The assumption is obviously not justified in reality, particularly from a linguistic point of view. It is clear that words are interdependent units in text. Information on term interdependence in a corpus, i.e. what words show significant association levels with other words is potentially important for IR, as it creates another dimension by which documents can be compared to queries. In other words, instead of estimating document-query likeness by independent presence of query terms in documents, we will take account of corpus associations of these terms to achieve more precise matching.

The experimental platform of the reported research work is Okapi -- an experimental information retrieval system, which implements the probabilistic model in term weighting, document ranking and relevance feedback mechanisms.

It may be noted that the methods of selecting collocates used here might also be applicable within other general models of information retrieval (such as the vector space model). However, in this paper we discuss only the probabilistic model.

2 Probabilistic Information Retrieval. The Formal Probabilistic Model and its Extensions

The underlying principle of the probabilistic model is to rank the documents retrieved in response to the user's request according to the probability of their relevance to this request. Each document is assigned a score, which is a sum of weights of individual terms in this document. Term weighting takes into consideration information on term incidence in relevant and non-relevant documents, provided some relevance feedback is obtained from the user. Relevance feedback information is obtained by the system each time the user inspects a document and makes a yes/no relevance judgement. After several documents are known to be relevant, the system extracts all terms from these documents, weights them and composes a new query from the terms with the highest weights in both the relevant documents and in the previous query. Before any relevance judgements are made, terms are weighted on the assumption that the number of relevant documents in the collection is very small. In this case probability of a term occurrence in non-relevant documents is estimated from the number of documents in the collection containing the term and probability of a term occurrence in relevant documents is a constant. The term weights are constantly updated throughout the search session as more relevance data is known [17].

The raw data used in term relevance weighting function of the probabilistic model are numbers of documents: relevant documents containing the term, non-relevant documents containing the term, relevant documents not containing the term and non-relevant documents not containing the term. This illustrates the inherently non-linguistic nature of probabilistic retrieval. The probabilistic model is built with a clear assumption of term independence in text. Probabilities are calculated for single terms, assuming that they occur independently in texts. This assumption is not justified by linguistics, which postulates word co-dependence and influence in text. However, the independence assumption is indispensable for the operability of the probabilistic model: if the probabilistic model took account of term dependency without making an independence assumption, probabilities would have to be calculated for each possible combination of terms, which is practically impossible [17]. One possible way of eliminating this problem is to use corpus-derived term co-occurrence information to modify independence based term weighting schemes.

Although the probabilistic model weights terms independently, not accounting for their linguistically-motivated relatedness, it actually implies some term dependency. Probabilities of occurrence of two terms are independent in the sets of relevant and non-relevant documents, however if probabilities of both terms are higher in the relevant documents, then independence in each of the sets implies relevance-based dependence in the collection as a whole [17]. It is also noteworthy to mention that Cooper's generalisation of the independence model, called linked dependence, leads to the same model [4].

The only linguistic evidence about document contents considered by the original probabilistic model is binary data of the presence/absence of a term in the text. The extended variants of the model also take into account frequency of word occurrence in the text and, optionally, positional information to support adjacency/proximity search [15]. The explanation why such poor linguistic knowledge is sufficient for effective implementations of probabilistic IR systems lies in the fact that the natural language text is inherently redundant, i.e. the plane of expression of text has a greater dimensionality than the plane of contents. Therefore loss of one type of linguistic data is roughly compensated by redundancy of the other.

Notwithstanding the consistently effective performance of the probabilistic model in evaluation experiments such as TREC there is still much scope for improvement. During recent years the Centre for Interactive Systems Research (CISR) has experimented with different means of using linguistic knowledge within the formal probabilistic model, notably -- various methods of indexing with compound terms and phrases [7].

Techniques for identification of more complex indexable units than single terms, developed for both pre-coordinate and post-coordinate indexing processes, included thesauri, phrase lists, adjacency and proximity operators. These methods are aimed at compensating for the underlying assumption of term independence by taking into account various term relations: paradigmatic - in thesauri usage, within-term relations -- in compound terms, and inter-term relations -- in phrases, brought to the system in the form of phrase-lists or by means of adjacency/proximity operators. The effectiveness of these additions to the probabilistic model did not, however, prove consistent. The problem with phrase-lists lies in the difficulty of automating their compilation. Linguistically motivated phrases, i.e. which represent complex concepts, can be identified either manually or using NLP. Their

manual construction is a labour-demanding task and is impractical for very large cross-domain collections, whereas automatic construction using NLP techniques is, first, computationally demanding, and, secondly, complicated due to wide subject range within collections, dealt with by modern IR systems. More crude statistical methods have been tried by various research groups, leading to some positive results. For example SMART system builds phrase-lists by extracting all adjacent pairs of non-stopwords that occur more than 25 times in the collection [2].

CISR experiments in thesaurus-assisted query expansion were conducted only in interactive conditions. Early TREC experiments by various research groups in automatic query expansion using manual thesauri were not successful. Experiments on the use of automatically constructed thesauri in query expansion produced positive results mainly in combination with relevance information [16].

We hypothesize that the use of automatically constructed corpus-derived term dependency structures is more efficient for automatic query expansion than usage of human-engineered term networks for the reason set out below. Engineered constructions represent the conceptual structure of the domain and reflect abstract and conceptual associations between terms. Contrarily, corpus-derived relations represent contextual and topical dependencies between terms in a particular collection. Theoretically, query expansion using collection-specific term relations is more likely to result in higher precision, than less discriminating approaches to query expansion with conceptually-related terms. Our initial comparison of the statistically built collocation lists and engineered term networks is suggestive of notable difference between the two kinds of term sets. The results obtained will be discussed in greater detail in part 5 of this article.

To summarise, the motivation to complement the probabilistic retrieval with collocation statistics methods is twofold: first, collocation dependency information compensates for the term independence assumption of the probabilistic model; secondly, corpus-derived term dependencies theoretically more effectively describe the context than sets of related terms in engineered term networks.

3 Corpus Linguistics Methods and their Application in IR

Corpus linguistics is a branch of linguistics which studies language phenomena by analysing their patterns of behaviour in large collections of text -- corpora. Corpus linguistics as a source of techniques appeals to probabilistic IR research for several reasons. First of all, it has an established and tested toolkit of statistical methods, which can be used to complement the statistics of the probabilistic model. Secondly, corpus linguistics techniques are well suited to deal with such intrinsic features of IR as large document collections and broad subject domains, which have always been an obstacle for in-depth NLP techniques, aiming at semantical text representation.

Corpus-based approaches to IR tasks, specifically to query expansion, also proved to be a viable alternative to knowledge-based techniques [5, 9, 18]. One of their main advantages is that they do not involve construction of complex knowledge bases. Moreover, the selection of terms extracted from text on the basis of their co-occurrence in discourse is hypothesised to be more efficient than engineered networks of terms, put together by their conceptual similarity, since text collocations have a better chance of representing topically motivated relations.

It is common knowledge that language is non-stationary. This is manifested by the fact that local statistics of word occurrences throughout the text is variable with the topic development. A word appearance in text is not a random event but influenced by its context, i.e. its neighbouring words. Words in text are "co-selective, and dependent on each other to realise a relevant sense in the text" [12]. Each word in text exerts a certain amount of influence on its environment. The question of how far the environment of a word stretches in text remains arguable. Most researchers so far have made empirical decisions about the span of a word's environment, mainly out of practical considerations of the particular research task. The span is measured either in syntactic units, such as phrases, sentences, paragraphs or even entire texts, or by the number of words to the left and right of the node. Such lack of uniformity can be partly attributed to different interpretations of the concept *the environment of a word*. Therefore, before proceeding further we clarify our perception of the notion. We emphasise the difference between the short-span environment of a word where lexico-syntactic factors determine word relations, and the long-span environment where terms are topically associated. Our initial investigations of statistical parameters of the environment of a word indicated a marked difference between word distributions in the immediate span of the node (4-5 word span either sides) and the bigger span. The statistical approach to analysing the environment of a word was chosen because a word occurrence changes the global statistics of word distribution within a certain span. The experiment consisted in counting the number of observed tokens for each location within 50 words both sides of the node. The results showed that the diversity of collocates within the immediate span was much lower than further away from the node. Immediate collocates maintained stable joint co-occurrence with the node, indicative of either

syntactic (e.g. phrasal verbs), or lexico-semantic factors (e.g. compound terms). Term diversity further away from the node was distinctly higher. However, the results obtained in the later experiments -- high MI and Z scores for the long-span collocates -- are evident of the term co-dependence in this distance range.

The described experiment allowed us only to identify the boundary between the short-span and the long-span influences of the word. The spread of long-span topically-motivated influence of the word was not established because the span we analysed was too short. In our identification of the boundary for a long-span environment we rely on the research undertaken by Beeferman et al. [1] who studied the effect of distance on the triggering and prediction capabilities of collocates (or trigger-words in their terminology). They analysed distance distributions in two groups of collocates: self-collocates and non-self collocates, using technique which consisted in calculating for each distance k the probability that two trigger words are separated by exactly $k \pm 2$ words. The results showed that a word's influence on statistical distribution of words around it stretches as far as several hundred words, levelling off by the position of 400 words. In our IR experiments we reduced the long span to 100 words for two reasons: first, the probability that two trigger words are separated by exactly $k \pm 2$ words becomes too low further away from the node, secondly, 100 word span is likely to provide the system with sufficient number of collocates. Further on in our research other span lengths will be tested with the primary aim of finding the smallest window size that will provide the sufficient amount of useful collocates for query expansion.

The long-span collocates of a word (or node -- in corpus linguistics terms) can be viewed as having a type of transitive co-occurrence with it: they are related to the node, because altogether they belong to the same topic. Since our aim is to develop a method of obtaining more information on the topic described initially by query terms, we focus our attention on wide-span topically motivated associations.

Not all words in the environment of the node term are influenced or triggered by its occurrence. The environment of a word consists of several types of units: high-frequency words (e.g. prepositions, articles, auxiliary verbs), which also co-occur frequently with the node; one-off collocates, with negative joint frequency and significant collocates with very high positive joint frequency. Renouf observed that the majority of words in text are related to more than one context. She distinguished eight types of such words: very common words, discourse organising words, homonyms, semi-technical words, words with several technical senses, metaphors, typographically ambiguous words [12]. Genuine topically-dependent words, as opposed to high-frequency or context-independent words, are expected to have significant amount of association with the node. *Significant collocates* [10] which can trigger each other's occurrence are distinguished from the chance word pairs by using various statistical measures of association. In our work we use modified *mutual information* and *Z score* statistics, which will be described in more detail in the next part of this article.

4 Combination of the Probabilistic Model with Corpus-derived Knowledge

The primary aim of our current research is to investigate the hypothesis that usage of collocated words, extracted from the collection can positively contribute to the performance of probabilistic retrieval. Specifically we hope to improve probabilistic IR performance in two areas: expansion of initial user requests and iterative querying techniques: user's relevance feedback and blind expansion, whereby the top ranked documents are considered relevant and are used to update the query contents and modify query term weights.

The method consists in building a database, where each indexing term in the collection will be bound with a list of its significant collocates. An experimental collocational database which will be used to test the new query expansion technique has been constructed using a set of TREC topics and FT-96 database. 50 TREC topics were parsed using Okapi, with all indexing terms extracted and used as node terms for the retrieval of collocates. The database is implemented as an Okapi database, where each record is indexed by the node term - indexing term, while a list of its significant collocates is the returned contents of the record.

The Okapi collocation database will be an additional intermediate layer in the existing querying technique, whereby the query -- whether user- or system-generated -- will be modified by expanding all query terms with significant collocates from the database (figure 1).

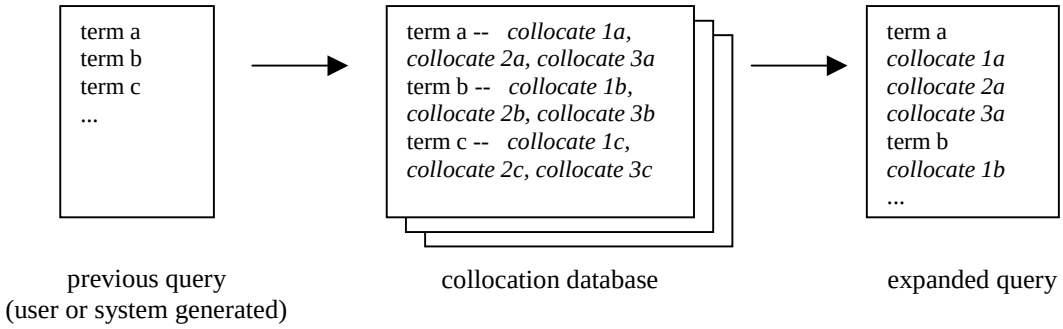

Figure 1: Query expansion mechanism using Okapi collocation database

The collocation database will be used at both the initial stage of the query to expand the list of terms submitted by the user, and in iterative querying following relevance feedback, where the query is expanded by adding significant collocates of those query terms, which occur in the relevant document. As a possible extension, transitive collocations -- collocates of the first-level collocates -- can also be identified. This presupposes that instead of submitting the combined list to the text database, it will be resubmitted to the collocation database for the retrieval of second-level collocates.

In the last thirty years in both linguistics and IR a number of studies involving word co-occurrence information were undertaken. These studies can be categorised by their understanding of what word co-occurrence is. We distinguish three main approaches to word co-occurrence:

- document as bag of words;
- ordered co-occurrence;
- unordered co-occurrence.

Rijsbergen in his research of term clustering in probabilistic IR considered two terms co-occurring if they both occurred in the same document [13]. Church et al. [3] estimated association ratio of ordered pairs of words occurring next to each other. In our research we consider two words co-occurring if they occur in either order within a maximum distance of 100 words.

Collocates of indexing terms are extracted from the entire collection within the span of 100 words both sides of the node term. The choice of a wide word span was motivated by the aim to identify topical term relations, rather than lexico-syntactical types, taking place closer to the node. As already mentioned in part 2 the decision to set the upper boundary of the span to 100 was prompted by research results achieved by Beeferman et al. [1] on the one hand, and by pragmatics of the query expansion task, on the other.

A window-based approach is theoretically more appropriate for selection of significant collocates than the document-wide term extraction used in current implementations of the probabilistic model. The supporting argument is that the topic flow changes throughout the text; therefore while one part of text may be relevant to the user's request, the other may not. As the topic changes, so does the distribution of terms. In document-wide extraction of collocates, terms from non-relevant parts of the document will skew the obtained statistical data on term co-occurrence, which may negatively affect selection of significant collocates.

The association strength of each collocate from the initially gathered pool of collocates of a term's instances throughout the collection is estimated by modified mutual information (MI) and Z score statistics. The criteria for selection of significant collocates is based on the combined results of these two statistics. These scores allow us to differentiate between significant collocations - pairs of words, which have some consistent linguistic (semantic, syntactic or lexical) relations, and chance pairs of words.

The mutual information score between a pair of words or any other linguistic units "compares the probability that the two words are used as a joint event with the probability that they occur individually and that their co-

occurrences are simply a result of chance" [10, p.71]. The mutual information score grows with the increase in frequency of word co-occurrence. If two words co-occur mainly due to chance their mutual information score will be close to zero. If they occur predominantly individually, then mutual information will be a negative number.

The standard formula for calculating mutual information score is:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

where *P(x,y)* is the probability that words *x* and *y* occur together, while *P(x)* and *P(y)* are the probabilities that they occur individually.

The standard MI formula is valid where term *x* immediately follows term *y* in text, i.e. for ordered pairs [3]. If we need to take into account backward co-occurrence or co-occurrence within a distance more than one word, adjustments need to be made to the baseline formula. In our work we use a modified MI statistic, which takes into account the window size for collocate extraction and backward co-occurrence. The modified MI formula is:

$$I_v(x, y) = \log_2 \frac{P_v(x,y)}{P(x)P(y)} = \log_2 \frac{\frac{f(x,y)}{Nv_x}}{\frac{f(x)f(y)}{N^2}}$$

where *f(x,y)* - frequency of joint occurrence of *x* and *y*, *f(x)* and *f(y)* - frequencies of independent occurrence of *x* and *y*, v_x - average window size for *x* (see below for a discussion of v_x), $P_v(x,y)$ - probability of occurrence of *y* in the windows around *x*, *N* - corpus size.

While mutual information is useful in filtering out pairs of words whose joint probability of occurrence is greater than chance, it gives very limited information as to how far joint probability differs from chance. Very high mutual information scores generally indicate strong bond between two words, whereas lower scores can be misleading, especially with low frequencies. Therefore it is not safe to make assumptions about the strength of words' association without knowing how much of that association is due to chance. For this reason we use a combined measure of significance, applying Z score as a secondary filter to the list of collocates selected by MI statistic.

Z score is a more reliable statistic: it gives us indication with varying degrees of confidence that an association is genuine by measuring the distance in standard deviations between the observed frequency of occurrence of *y* around *x*, and its expected frequency of occurrence under the null hypothesis. For a chance pair of words in the conditions of low word frequencies we may misleadingly get a high mutual information score, whereas the Z score will not be high since the difference between the observed and the expected joint frequencies of these two words in standard deviations will be small.

Our calculation of Z score as a secondary filter in the selection of significant collocates is similar to t-score used by Church et al. [3]. The data we operate on has the following characteristics, which determined some major differences in the formula:

- We operate on a very large corpus size - 43 million words;
- The span for collocate extraction is 100 words either sides of the node, hence the window size is 201;

Another important divergence is the method we used for identification of windows in the collection. Although the ideal window size is 201, in practice it is usually much smaller for the following reasons. We count the window size by the number of words both sides from the node. If we meet the same word as the node in the right-hand half of the window (i.e. to the right of the node), we stop counting further, if the same word is found in the left-hand half of the window, we ignore this half altogether. This condition is necessary to avoid duplicate extraction of the same

collocates for several occurrences of the term which are close in the text. Another case when the window is truncated takes place if the node occurs near the beginning or end of the document: i.e. if the actual number of words to the left or right of the node is less than 100. Therefore, as the real window sizes for the term *x* are different from the ideal window, in our Z formula we use the average window size for the term *x* - denoted v_x.

The above described method of counting window sizes leads to another specificity of our method - asymmetry. For the node *x* and its collocate *y* we calculate the expected number of occurrences of *y* within the windows of every occurrence of *x* under the null hypothesis. The null hypothesis means that only chance is affecting the co-occurrence of *y* with *x*. The expected number of occurrences of *y* will be $v_x f(x) f(y)/N$, where $v_x f(x)$ is the number of locations which might contain *y* collocated with *x* and $f(y)/N$ is the probability that *y* occurs in any of these locations under the null hypothesis. The asymmetry of the method becomes evident if we perceive *y* as the node and *x* as its collocate: the expected number of occurrences of *x* near *y* will be a different figure, since v_y will also be different.

The expected frequency of occurrence of *y* near *x* - $v_x f(x) f(y)/N$ - is the mean of a binomial distribution. Because the probability of occurrence of *y* near x - $f(y)/N$ is very small, the mean square error of the variance of the binomial distribution will be approximately the same value, i.e. $v_x f(x) f(y)/N$.

We use Z statistic instead of the t-score because, due to large sample size, we deal with normal distribution. We can interpret this statistic as a t-score with a very large number ($v_x f(x)-1$) of degrees of freedom, which means that the distribution becomes normal.

The formula for the Z score calculates the difference between the actually observed joint frequency $f(x,y)$ and the expected frequency of *y* as a collocate of *x* under the null hypothesis in standard deviations:

$$Z = \frac{f(x,y) - \frac{v_X f(x) f(y)}{N}}{\sqrt{\frac{v_X f(x) f(y)}{N}}}$$

The statistic, however, will be inaccurate if the frequency of *x* is too small, e.g. one, since the locations we consider relate to a single occurrence of *x*. We alleviate this problem by applying the Z statistic to terms with sufficient frequency of occurrence ($f(x)>30$).

Z score and mutual information bring to the top different kinds of collocations: Z score tends to pick frequent word combinations, and may have a drawback of showing syntactical collocations with functional words like 'by sea'. Mutual information highlights word combinations that are specific to both words, e.g. fixed phrases, some compound terms and proper names. It tends to find low-frequency domain-specific combinations. For lexicographic tasks, where a high degree of inter-corpus generalisation is important, this property of MI may be considered a drawback, whereas in IR it may prove useful as the more specific the phrase, the better potential it has for differentiating the document.

The lexico-semantic analysis described in the next part highlights some interesting observations related to the linguistical and conceptual characteristics of the collocates selected by the MI and Z statistics.

5 Lexico-semantic Analysis of Collocations

5.1 Comparison of Collocations Selected by MI and Z Statistics

In the process of building the collocation database two sets of significant collocates for each indexing term were formed. The first set is a list of collocates, sorted by their MI score; the second list is comprised of collocates ranked by the magnitude of their Z score. The question of how query expansion terms will be selected from the two collocation lists remains open, requiring systematic testing. Several combinatorial patterns for the selection of collocates for query expansion will be tried at the next stage - retrieval experiments. The main points to be investigated at that stage will be the significance cut-off points in either of the collocation lists and the method of merging the two lists.

To understand the tendencies the two statistics show in the collocates selection and ranking we have undertaken a comparative lexico-semantic analysis of collocation lists formed by MI and Z. Some of the observations made and patterns identified are described further in this part.

The majority of the top terms ranked by Z score are related to the node semantically and their associations tend to be collection-independent. This tendency was observed in collocation lists for both general and specific terms. The picture is distinctly different in the MI-ranked term lists. For general terms MI tends to pick rare collection-dependent collocates, which are predominantly proper names, whereas for more specific terms the results yielded by MI statistic were resemblant to those of Z score, i.e it selected collection-independent conceptually associated terms. For example the types of significant collocates for the term *acquire* in MI-ranked and Z-ranked lists have marked difference:

MI	Z
Noorda 4.19 (surname)	**acquisition** 141.64
Huntsman 4.18 (company name)	**pound** 114.25
Nextel 4.18 (company name)	**stake** 104.267
Gartland 4.14 (surname)	**company** 102.65
Revco 4.08 (company name)	**group** 99.77
Viglen 4.07 (company name)	**purchase** 84.33
Tampella 4.02 (company name)	**share** 84.17
Cinzano 4.02 (company name)	**profit** 66.67
Conspress 3.98 (company name)	**operation** 61.70
CCL 3.91 (company name)	**business** 59.141

Table 1. Lists of top collocates for the term *acquire* sorted by MI and Z statistics

The table illustrates that all top MI-ranked terms are proper names. The inspection of documents containing the instances of the node term together with the listed company names or surnames showed that their topics are all related to the idea of acquisition/purchase of companies by other enterprises. All listed company names and surnames denote parties in company acquisition transactions.

The Z-sorted list in contrast contains more general terms, some similar to the type of terms found in a manually constructed thesaurus.

Another example illustrates the same tendency observed in the collocate lists for the synonym group *(environment, environmental)* :

MI	Z
tribal 5.98	**waste** 152.12
GEF 4.34 (Global Environment Facility)	**pollution** 150.42
ecolabel 4.22	**emission** 146.35
Lalonde 4.17 (B. Lalonde, Environment Minister, France)	**recycle** 131.15
VOC 4.15 (Volatile Organic Compounds)	**energy** 101.45
Meana 4.14 (Ripa di Meana, EC Environment Commissioner)	**carbon** 100.39
Topfer 4.13 (C.Topfer, Environment Minister, Germany)	**water** 95.49
CPRE 4.13 (The Council for the Protection of Rural England)	**pollute** 93.37
Ripa 4.11 (Ripa di Meana)	**dioxide** 92.26
DSD 4.03 (Duales System Deutschland - scheme adopted by companies in Germany to recover waste from households and reuse the raw materials)	**Gummer** 86.27 (R.Gummer, the Minister of Agriculture)
UNEP 4.01 (UN Environment Programme)	**forest** 86.10
deforest 4.00	**Greenpeace** 84.20
LRB 3.98 (London Residuary Body)	**clean up** 79.43

Table 2. Lists of top collocates for the synonym group (*environment, environmental)* sorted by MI and Z statistics

In this example the MI-sorted list is also dominated by very specific collocations, characteristic of the particular collection, most of which are again proper names and also abbreviations of compound terms. Z score highlighted predominantly typical collection-independent associations. Proper names in the MI group, as evident from the comments, are all topically related to the node term. Having low collection frequency proper names can be good contents discriminators, thus their presence in the expanded query is desired.

In contrast to the previous examples of general terms, specific terms have a different distribution of MI and Z significance scores. In contrast to the very insignificant overlap of top collocates selected by the two statistics for general terms, MI- and Z-sorted collocation lists for specific terms demonstrate more similarity. The term *gene* is a typical example of this pattern:

MI	Z
genome* 9.11	**genetic*** 426.08
PCR 8.86 (polymerase chain reaction)	**DNA*** 261.99
transgenic* 8.79	**genome*** 187.77
NIH* 8.77 (National Institute of Health)	**therapy** 165.13
Venter 8.77 (Craig Venter, one of America's leading gene researchers)	**protein** 156.88
fibrosis* 8.77	**cell** 145.23
cystic* 8.77	**fibrosis*** 144.57
chromosome* 8.68	**cystic*** 144.57
DNA* 8.65	**transgenic*** 132.89
Lockhart 8.60 (Gene Lockhart)	**chromosome*** 131.04
genetic* 4.39	**NIH*** 123.71 (National Institute of Health)

***** - terms are found in both groups

Table 3. Lists of top collocates for the term *gene* sorted by MI and Z statistics

The table shows that 72% of the top 11 significant collocates for the term *gene* occur in both MI and Z groups. The majority of collocates in both groups represent domain-independent semantic associations. The tendency of MI to assign high scores to specific terms like abbreviations and proper names is present, though to a less degree, in this example as well. Selection of low frequency collocates, specific to both words has, however, its downside. One of the proper names picked by MI - *Lockhart* - is a surname of a person whose first name is *Gene*. Although words *gene* and *lockhart* co-occur in the collection only nine times, the term *gene* is very specific to *lockhart*, i.e. the latter term has a very distinct pattern of co-occurrence with the former among its collocates. Z also ranked *lockhart* high, but its position is relatively low as compared to the position in the MI-sorted list.

5.2 Collocates of Polysemantic Words

The example mentioned in the previous passage spotlights a significant problem of automatic term extraction, namely multiple word senses. Statistical methods of collocate extraction do not provide for word sense disambiguation, therefore both MI- and Z-sorted collocation lists for polysemantic node words feature collocates related to different word senses of the node. Collocation lists for the term *pyramid* provide a good illustration of the co-occurrence patterns of polysemantic words.

MI	Z
Amway[1] 9.24 (the name of the pyramid selling scheme) **Caritas**[1] 9.06 (the name of the Romanian pyramid scheme) **Mavrodi**[1] 8.97 (surname of the head of Moscow pyramid company MMM) **Cluj**[1] 8.63 (hometown of Caritas) **MMM**[1] 8.32 (the name of Moscow pyramid company) **Projet**[2.2] 8.17 (Grands Projets) **Alchemy**[1] 8.12 (the name of the pyramid company) **Louvre**[2.2] 7.84 **Angkor**[2.1] 7.79 **Tyzack**[1] 7.75 (surname of the head of a pyramid selling scheme)	**MMM**[1] 184.46 **Caritas**[1] 156.58 **Mavrodi**[1] 143.59 **Amway**[1] 125.19 **Louvre**[2.2] 122.75 **Alchemy**[1] 88.09 **Cluj**[1] 84.28 **Projet**[2.2] 71.98 **Luxor**[2.1] 50.94 **Egypt**[2.1] 50.31

[1, 2.1, 2.2] - collocates related to the corresponding senses of the word *pyramid*

Table 4. Lists of top collocates for the term *pyramid* sorted by MI and Z statistics

The collocates listed in the table refer to the following senses of the word *pyramid*:

1. financial scheme or company
2. architectural construction *(2.1. Egyptian pyramids; 2.2. Glass pyramids in Louvre courtyard)*

Because the main topical domain of the FT collection is financial and business news, the predominant number of significant collocates refer to the first sense of the word *pyramid -- financial scheme or company.*

One possible method to minimise the presence in the extended query of collocates related to irrelevant senses of the query word is to make use of other query terms. A high association ratio with two or more terms from the query indicates that the collocate is most likely to be related to that sense of the query terms, which was meant by the user. The method consists in merging the collocation lists for all query terms and applying PROJECT operator to select only those terms which appear in more than one list.

Another possibility for sense disambiguation of each occurrence of the node in retrieved documents is to use the information provided by its short-span collocates. This is a *local* analysis, i.e. analysis of collocates of a word in a relevant document, which can be integrated into our method of *global*, i.e. collection-wide, collocate extraction. After initial search by significant global collocates of the query terms is done and the user judges some documents relevant, the query term instances are located in these documents and their short-span collocates are extracted. The next step is to compare the global collocate lists for the short-span collocates and their corresponding query term. If a term is a significant collocate of more than one term, this is an indicator of its relatedness to the relevant sense of the query term. But this method also faces a potential difficulty -- the short-span collocates need to be related to the node semantically to have comparable collocation lists, therefore an accurate method for their selection need to be devised.

5.3 Collocation Lists vs. Engineered Term Networks

Comparison of statistically formed collocation lists with engineered term dependency structures -- thesauri and term networks -- so far confirmed our earlier hypothesis that discourse (syntagmatic) relations between terms are essentially different from conceptual (paradigmatic) term relations. We compared MI- and Z-ranked collocation lists with WordNet term relations and INSPEC thesaurus entries. WordNet is a database of the general lexicon of the English language in which words are grouped in synonym sets (synsets) and are connected to other synsets via hyponymical and holonymical hierarchies [11]. Different senses of polysemantic words form separate synonym sets in WordNet. INSPEC thesaurus is limited to the physics domain, therefore only senses belonging to this domain could be correlated in the comparative analysis.

Although FT database contains documents related to various subject domains, its overall subject profile is skewed towards business, economics and financial topics. It is especially evident through the ratio of

economics/finance related collocates in the collocation lists for general terms. For example all top collocates of the polysemantic term *pressure* in both MI and Z lists are related to either of two general senses of the node:

1. **pressure** - a force that compels;
2. **pressure** - imperativeness, insistence, press (the state of urgently demanding notice or attention)[1].

But although *pressure* is used in its two general senses, there is no overlap between the sets of related terms in WordNet and the collocation lists. While WordNet showed terms related to the node *conceptually* (as synonyms, hyponyms, hypernyms and holonyms), collocation lists featured words related to the node *situationally*. We see this as the essential difference between paradigmatic[2] and syntagmatic types of relationships between words. From the paradigmatic viewpoint the word is an element of an artificial system -- a conceptual paradigm -- which is an abstraction, a systematisation of the human knowledge built on concrete situations. Two or more words belong to the same conceptual paradigm if they have some semantic similarity.

In discourse the words are arranged in sequence and the type of relations they acquire are called syntagmatic, i.e. relations based on the linear nature of the discourse. The word in discourse can be said to be an instantiation of the word in the paradigmatic sense as an element of several paradigms (grammatical, lexical, conceptual). The properties of the word instance in text are formed by inheriting the properties of the corresponding elements in all paradigms. From the semantic viewpoint the word in text is an instantiation in a concrete situational setting of the properties of the corresponding element in the conceptual paradigm. Theoretically it is possible that some of the word relations in the conceptual paradigm (e.g. broader, narrower words) are also used in the same situation in discourse. However, the results obtained so far showed little evidence on the presence of paradigmatic neighbours of terms in the syntagmatic environment.

Comparison of the collocation lists and WordNet/INSPEC entries for technical terms, e.g. *fuel, plutonium, uranium* also showed little overlap. There were no matching terms for the top 12 collocates of *plutonium* and *uranium* in INSPEC and hyponymical, holonymical and synonymical relations in WordNet. Collocates of the term *fuel*, top ranked by MI, did not have any matching terms in either INSPEC, or WordNet , however top 12 collocations selected by Z score for the term *fuel* contained 1 word which matched an entire term in INSPEC and 3 words which matched parts of the compound terms in INSPEC or WordNet. This highlights one more important difference between engineered term networks which represent conceptual relations and collocation relations -- each term in either INSPEC, or WordNet represents a concept, therefore complex concepts are represented by compound terms. Collocation lists, in contrast, always contain single words, which do not necessarily denote a concept.

Collocates of the term *fuel* selected by Z score	Completely/partially matching terms in INSPEC and WordNet
diesel 156.83	**diesel oil** (hyponym) WordNet
energy 138.17	**energy resources** (broader term) INSPEC
reactor 137.04	**fission reactor fuel** (narrower term) INSPEC
coal 118.01	**coal** (narrower term) INSPEC
	coal gas (hyponym) WordNet

Table 5. Terms in Z-ranked collocation list matching INSPEC and WordNet terms related to the term *fuel*

The described comparison of automatically generated collocation lists and human engineered term dependency structures is a part of the very initial experiments, the results of which are only suggestive of the trends, but do not provide any fundamental evidence necessary to draw final conclusions. In the future systematic comparative experiments with a wider sample range will be undertaken to explore this correlation in greater depth.

[1] definitions are taken from WordNet

[2] here by paradigmatic relations we mean relations in the conceptual paradigm of the word, different from lexical or grammatical paradigms.

6 Conclusion

This paper presented an approach to compensate for the essentially non-linguistic nature of the probabilistic model of information retrieval with the corpus-derived data on word co-occurrence. Extending probabilistic query-document matching and term weighting to take account of significant word collocations can possibly improve performance of the basic probabilistic model. There are several considerations to support this statement:

(a) The document/query representations in the probabilistic model are strings of unrelated terms. Such representations give a weak evidence of the underlying contents. We hope that by identifying significantly associated collocates of a term throughout the collection, we can obtain richer representation of the contents of its occurrence and hence provide the system with more variables by which a query can be mapped to document representations.
(b) Significant collocates are selected from the windows of 200 words around the node terms. A window-based technique has lower probability of deriving topically irrelevant terms than document-wide collocate extraction. This is of particular significance for term extraction from long multi-topic documents.
(c) Collocations are extracted globally, i.e. for each occurrence of a term in the collection. Analysis of the global environment of a word has a better likelihood of bringing forward potentially useful collocates than analysis of a single context of a query term instance in the relevant document
(d) A two-level selection criterion is applied to filter significantly associated collocates: MI and Z statistics.

Our lexico-semantic analysis of collocates selected by MI and Z demonstrated the viability of application of these statistics to large word spans. In both MI- and Z-ranked collocation lists all top collocates were related to the node topically and/or semantically. MI proved to have different trends in collocation retrieval from Z: while MI tends to select specific terms - proper names and compound term abbreviations, Z ranked highly a larger number of typical and more semantically predicted types of collocates.

The initial comparison of statistically formed collocation lists with human engineered term networks showed little overlap, which is suggestive of the inherent differences between the syntagmatic relations occurring in text and the paradigmatic relations in the conceptual trees of terms.

The next stage of the presented research work will be evaluation and testing of the new query expansion method in three different settings: query expansion of the initial user requests, query expansion following user's relevance feedback and blind query expansion. To take better advantage of the collocation information in the probabilistic model, we will need to modify term weighting function to take account of a term's value in representing the context of the query term(s). Various combinations for selecting and merging sets of significant collocates will also be tested at the next stage.

The research into integration of linguistically motivated corpus-based techniques into the purely statistical IR model like the probabilistic model is our answer to the current changes in operational conditions of IR, namely larger and more eclectic document collections and increased demand for IR to yield higher precision. Although the probabilistic model achieves the highest performance level so far, this is still a rather modest performance, given the scope for potential improvement. The presented approach increases the amount of information available to the model by complementing the probabilistic algorithms with corpus linguistics methods with the aim of achieving more accurate query-document matching and document ranking.

7 References

1. Beeferman, D., Berger, A., Lafferty, J. A Model of Lexical Attraction and Repulsion. Proc. ACL-EACL Joint Conference, 1997 Madrid, Spain.

2. Buckley, C.; Salton, G.; Allan, J.; Singhal, A. Automatic Query Expansion Using SMART: TREC 3. In: The third Text REtrieval Conference, Gaithersburg, Maryland, 1995, p.69.

3. Church, K.; Gale, W.; Hanks, P.; Hindle, D. Using Statistics in Lexical Analysis. In: U.Zernik (ed) Lexical Acquisition: Using On-line Resources to Build a Lexicon. Englewood Cliffs, NJ: Lawrence Elbraum Associates, 1991, pp.115-164.

4. Cooper, W.S. Inconsistencies and Misnomers in Probabilistic IR. In: Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Chicago, 1991, pp. 57-62.

5. Hearst, M. and Grefenstette, G. A Method for Refining Automatically-Discovered Lexical Relations: Combining Weak Techniques for Stronger Results. In: Carl Weir (ed) Statistically-Based Natural Language Programming Techniques: Papers from the 1992 Workshop Technical Report W-92-01, ANNAI Press, Menlo Park, CA., 1992

6. Jing, Y and Croft, B. An Association Thesaurus for Information Retrieval, Technical Report UMASS-CS-94-17 (IR-47), University of Massachusetts, 1994

7. Jones, S. A thesaurus data model for an intelligent retrieval system. Journal of Information Science 1993; 19, pp. 167-178.

8. Lewis, D. and Sparck Jones, K. Natural Language Processing for Information Retrieval. Communications of the ACM, 1996; 39(1), pp. 92-101.

9. McDonald, J., Ogden, W., Foltz, P. Interactive Information Retrieval Using Term Relationship Networks. New Mexico State University, Las Cruces, NM, USA. The Sixth Text REtrieval Conference (TREC-6), Gaithersburg, MD, USA, 1997

10. McEnery, T. and Wilson, A. Corpus Linguistics. Edinburgh, 1996

11. Miller, G., Beckwith, R., Fellbaum, C. Gross, D., Miller, K. Five Papers on WordNet. Cognitive Science Laboratory, Princeton University. CSL Technical Report no. 43, 1990

12. Renouf, A. What the linguist has to say to the information scientist. Journal of Document and Text Management, 1993; vol.1, no. 2., pp.173-190.

13. Van Rijsbergen C.J. A Theoretical Basis for the Use of Co-occurrence Data in Information Retrieval. Journal of Documentation, 1977; vol.33, no. 2, pp. 106-119.

14. Robertson, S., Walker, S., Jones, S., Hancock-Beaulieu, M. and Gatford, M. Okapi at TREC-3. In: D. Harman (ed) The Third Text REtrieval Conference (TREC-3), Gaithersburg, MD, NIST, 1995

15. Robertson, S. Overview of the Okapi Projects. Journal of Documentation, 1997; vol. 53, no. 1, pp. 3-7.

16. Salton, G. Automatic Term Class Construction Using Relevance - A Summary of Word in Automatic Pseudoclassification, IP&M, vol. 16 (1), 1980, pp. 1-15.

17. Sparck Jones, K., Walker, S. and Robertson, S. A Probabilistic Model of Information Retrieval: Development and Status. University of Cambridge Computer Laboratory Technical Report no. 446, 1998

18. Xu, J. and Croft, B. Query Expansion Using Local and Global Document Analysis. Proc. 19th International Conference on Research and Development in Information Retrieval (SIGIR '96), Zurich, Switzerland, 1996; pp. 4-11.