6. Global collocation analysis experiments

6.1 Introduction

The aim of this set of experiments was to examine the hypothesis that global long-span collocates of query terms, used for query expansion, can lead to performance improvements over the initial query.

Initial stage was to prepare the necessary platform for the experiments, i.e. to build a resource of global collocates. Then two stages of experiments were conducted:

1. Retrieval experiments on query expansion with global collocates;
2. Statistical analysis of collocation association measures MI and Z, which were used for selecting collocates for query expansion.

The first stage of experiments is directly targeted at testing the above hypothesis. The objective was to evaluate the retrieval performance of the expanded queries and to compare it to the performance of the initial queries. The objective of the second experiment stage was to get a statistical evidence about the capability of collocation association measures (MI and Z) applied to global collocates, of predicting their usefulness as relevance-discriminating terms. The second stage is seen as a deeper quantitative analysis of the global collocates, which throws light on the interrelationship of their statistical characteristics, such as Offer Weight, MI, Z and collection frequency.

6.2 Construction of the database of global collocates

The first step in the construction of the database of global collocates was to create a corpus from FT 96 collection. FT 96 has been indexed as an Okapi database. The corpus was built by extracting every record from the database and parsing it using the Okapi parsing algorithm, consisting of GSL look-up and stemming (see section *5.5.2* above). The corpus was recorded as a single text file, consisting of concatenated records. Each record, delimited with record boundaries, is represented as a sequence of tokens – stems and GSL synonym codes of the original wordforms appearing in text. The order of tokens is the same as the order of their corresponding wordforms in the text of each record.

If we were to build a fully operational system supporting query expansion with global collocates, we would have to identify global collocates for each index term in the database. Since our task was laboratory evaluation of this technique, in which we used 50 request formulations, we had to identify global collocates only for the query terms from these request formulations. This significantly reduced our time for building the global collocates database, yet did not impose any limitations on the experiments conducted using this collocation database.

Request formulations were constructed from 50 TREC topics as described in section *5.5.3*. Each request formulation is represented as a list of stems and/or GSL synonym codes of the corresponding query terms. For convenience we will refer to stems/GSL codes of words in the request formulation simply as query terms.

The next step was to extract for every occurrence of every query term in the corpus all its collocates using the windowing technique presented in section *5.3*. The window size set for global collocates extraction was 201 (for rationale of this decision also see section *5.3*). For each identified collocate of a query term instance we recorded a collocation pair: [*query term*, *collocate*].

After all collocation pairs have been recorded, the next step was to rank collocates of each query term by two association measures – MI and Z (see sections *5.4.1* and *5.4.2* above). MI and Z scores were calculated only for those collocates which have individual collection frequency greater than 30 (see section *5.4.2*). Terms with smaller frequencies were discarded from the collocation lists. For each query term two files were created – one with its collocates ranked in descending order of their MI scores, the other – containing collocates in descending order of their Z scores. For samples of MI- and Z-ranked collocation lists see Appendix A.1.

In the MI-ranked lists, collocates with MI > 0 are considered to have association greater than chance. The higher the collocate's MI score, the stronger is its association with the node. It was observed that on average the number of collocates with MI > 0 is quite large for most query terms, well exceeding 100 terms. For most query terms, top 100 collocates ranked by MI also tended to have significant Z scores. As noted in section *5.4.2* we consider Z > 1.6 as significant.

In the Z-sorted list, again the number of significant collocates was quite large – most query terms had more than 100 collocates with Z > 1.6. Z scores in the top of the list tended to be quite high, top collocates of some query terms had Z > 100. However, some terms in the top 100 of the Z-ranked lists had rather low MI scores (MI ≈ 0).

The next decision to be made was whether to use a fixed number of top-ranked collocates in MI- and Z-ranked lists for query expansion, or collocates with scores above a certain threshold. The problem with the latter approach is that the scores are highly variable for collocates of different query terms. Top collocates of some query terms have, for example, Z scores well above 100 (table 6.1), others have much lower scores (table 6.2). MI scores have smaller variance than Z, but the values of top scores and the numbers of collocates with MI above a certain value are also too variable. This strategy would result in a huge number of collocates selected for some terms, and none for others. For this reason we decided to select a fixed number of top ranked collocates for query expansion from both MI- and Z-ranked lists. A very large margin of significant collocates in both MI- and Z-ranked will still exceed the fixed number of top collocates we need for query expansion. This means that all collocates we choose either from Z-, or MI-ranked lists will have significant values of the corresponding statistic.

collocate	Z score
tobacco	383.515
marlboro	218.875
smoke	178.732
bat	172.081
smoker	158.420
morris	154.348
nicotin	145.943
rjr	145.569
seita	136.464
@0018	118.345
filip	111.361
rothman	107.858
nabisco	103.962
brand	102.078
bppc	88.9071
morriss	80.7002
packet	70.8662
rj	66.7835
habgood	65.9124
bunzl	64.0986

Table 6.1 Top 20 collocates of the query term *cigarette* (topic 257) ranked by Z score

collocate	Z score
narva	43.6465
curzon	39.3772
seti	33.4198
estonia	30.8962
estonian	30.5858
immigr	28.7626
parti	27.7749
hindus	24.8815
camacho	24.8815
kashmiris	24.5734
voter	23.8587
zoe	23.3030
karajan	23.1119
baird	22.8774
haider	22.7896
@0111	21.9208
poe	21.4189
thian	21.1312
boarder	21.0979
cultur	20.5486

Table 6.2 Top 20 collocates of the query term *alien* (topic 252) ranked by Z score

6.3 Lexical-semantic Analysis of Collocations

To understand the tendencies MI and Z statistics show in the collocates ranking a comparative lexical-semantic analysis of collocations has been carried out. Specifically we looked into the following aspects:

- differences between MI and Z in selecting collocates for general and more specific terms;
- collocates of polysemantic words;
- differences between collocates and terms found in manually engineered term structures.

6.3.1 Differences between collocates selected with MI and Z statistics

Many top collocates ranked by Z score tend to be related to the nodes semantically and there is a large number of collection-independent collocates. 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 resembling those of Z score, i.e. it selected more general conceptually associated terms. For example the types of significant collocates for the term *acquire* in MI-ranked and Z-ranked lists have marked difference (table 6.3).

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 6.3. 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, of the type expected to be found in manually engineered term structures. They are all related to the sense of *acquire* as *obtaining by means of a financial transaction*. This is explained by a large number of economy/finance related documents in the FT 96 collection.

Another example is collocations for the synonym group @0104 (*environment, environmental*). It illustrates similar tendency of MI to select specific and rare terms and of Z to select higher frequency general terms (table 6.4).

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

Table 6.4. Lists of top collocates for the synonym group @0104 (*environment, environmental*) sorted by MI and Z statistics

Here MI-sorted list is also dominated by low-frequency collocates, most of which are proper names, specific to this particular collection. All of them, as evident from the comments, are topically related to the sense of the term *environment* as:

> *the complex of physical, chemical, and biotic factors (as climate, soil, and living things) that act upon an organism or an ecological community and ultimately determine its form and survival*[*]

Z score highlights more general terms with higher collection frequency for the above sense of the node. Many of these terms, as seen from the table, are conceptually related to this sense and could be considered key terms in the textual topics about environmental protection issues.

Similar pattern is evident in the collocation lists for the term *education* (table 6.5).

MI	Z
Ofsted (The Office for Standards in Education) 5.00365 **GNVQ** (General National Vocational Qualifications) 4.98525 **Blatch** (Lady Blatch, schools minister) 4.91188 **Natfhe** (The University and College Lecturers' Union) 4.85370 **Gruchy** (Nigel de Gruchy, general secretary of the National Association of Schoolmasters/Union of Women Teachers) 4.79819 **GCSE** 4.79289 **Educationalist/educationally** 4.78483 **truancy** 4.76496 **A-level** 4.72806 **Nasuwt** (National Association of Schoolmasters/Union of Women Teachers) 4.71952	**school** 351.413 **teacher** 220.438 **student** 169.629 **pupil** 155.444 **curriculum** 155.220 **university** 147.055 **college** 144.692 **A-level** 141.422 **vocation** 130.109 **Patten** (Christopher Patten, chairman, Conservative party) 120.885

Table 6.5. Lists of top collocates for the term *education* sorted by MI and Z statistics

In contrast to the previous examples of general terms, specific terms have a different distribution of MI and Z significance scores, with more overlap between the top collocates. Typical examples are collocation lists of the terms *nitrogen* (table 6.6) and *gene* (table 6.7).

[*] the definition is taken from Merriam-Webster Online Dictionary (http://www.m-w.com)

MI	Z
NOx* (nitrogen oxides) 9.26984	**oxide*** 200.956
monoxide* 9.24756	**emission** 181.221
particulate* 9.20253	**monoxide*** 168.745
superconductor* 9.18177	**particulate*** 162.550
legume 8.68129	**dioxide** 154.997
gasification 8.65821	**NOx*** 154.909
urea 8.56510	**sulphur** 152.579
ammonia* 8.24317	**superconductor*** 115.374
autocatalyst 8.17606	**carbon** 111.171
superconduct(ing/ivity) 8.11193	**diesel** 102.248
soot 8.05053	**pollution** 102.000
oxide* 7.99091	**fertiliser** 94.1899
co-generation 7.90757	**hydrocarbon** 89.7964
phosphor 7.90368	**pollute** 89.7602
nitrate* 7.85498	**nitrate*** 84.3515
lolly 7.81349	**ammonia*** 81.3761

Table 6.6. Lists of top collocates for the term *nitrogen* sorted by MI and Z statistics (* terms are top ranked in both lists)

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
	cell 145.23
fibrosis* 8.77	**fibrosis*** 144.57
cystic* 8.77	**cystic*** 144.57
chromosome* 8.68	**transgenic*** 132.89
DNA* 8.65	**chromosome*** 131.04
Lockhart 8.60 (Gene Lockhart)	**NIH*** 123.71 (National Institute of Health)
genetic* 4.39	

Table 6.7. Lists of top collocates for the term *gene* sorted by MI and Z statistics (* terms are top ranked in both lists)

As it can be seen from the two above tables, 7 out of 11 top collocates of the term *nitrogen* and 8 out of 11 collocates of the term *gene* occur in both MI and Z lists. The majority of collocates in MI and Z lists of both nodes represent domain-independent semantic associations. The tendency of MI to give high scores to low-frequency collocates like proper names applies, though to a less extent, to specific terms too (as evident from table 6.7). Selection of low-frequency associations has, however, its downside. One of the collocates picked by MI for the term *gene* – *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 compared to the position in the MI-ranked list.

6.3.2 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-ranked collocation lists for polysemantic query terms can feature collocates related to different word senses of these terms. Collocates of the term *pyramid* provide a good illustration of the co-occurrence patterns of polysemantic words (table 6.8).

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 the Russian pyramid company MMM) **Cluj**[1] 8.63 (hometown of Caritas) **MMM**[1] 8.32 (the name of the Russian 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

Table 6.8. Lists of top collocates for the term *pyramid* sorted by MI and Z statistics

The collocates listed in table 6.8 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 many documents in the FT 96 collection are financial and business newsarticles, the predominant number of significant collocates refer to the first sense of the word *pyramid* – financial scheme or company.

6.3.3 Comparison of collocates with terms from engineered term networks

Comparison of statistically formed collocation lists with engineered term structures – thesauri and lexical networks – showed a rather insignificant overlap of terms. We compared MI- and Z-ranked collocation lists with WordNet term relations and INSPEC thesaurus entries. For a brief description of WordNet see section 3.1.1 earlier in this thesis. WordNet contains general lexicon of the language, INSPEC thesaurus, however, is limited to the physics domain. In our analysis we had to select only those terms which are present in both INSPEC and WordNet.

First we analysed a polysemantic word *pressure*. All top collocates of *pressure* in both MI and Z lists (table 6.9) are related to either of its two general lexicon senses:

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

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

[*] definitions are taken from WordNet

MI	Z
PWR 3.28301	**rate** 111.394
IDO 2.94352	**inflationary** 95.2335
blocker 2.76148	**inflation** 76.6912
inflationary 2.65797	**Bundesbank** 73.6852
VVER 2.62802	**under** 71.3450
EAP 2.61010	**ERM** 66.2900
UAC 2.57057	**dmark** 65.5720
VOC 2.55431	**currency** 60.1104
nonft 2.54556	**cut** 59.3700
micromachine 2.54556	**price** 54.8346

Table 6.9. Lists of top collocates for the term *pressure* sorted by MI and Z statistics

Sense 1	Sense 2
Synonyms	
force	imperativeness, insistence, press urgency
Coordinate terms (terms with the common hypernym)	
force duress lifeblood wheels	urgency hurry, haste criticality, criticalness, cruciality
Hypernyms	
force influence power, powerfulness, potency quality attribute abstraction	urgency necessity need, demand condition, status state

Table 6.10. Terms related to the above two senses of the term *pressure* in WordNet

There are no terms in common with either INSPEC, or WordNet. The most striking fact about the selected collocates (table 6.9) is that most of them belong to the subject domain of economy/finance, as was already noted for some other general terms in previous sections. Since INSPEC contains terms in the physics domain, the lack of any common terms with the collocation lists is self-explanatory. As for the lack of common terms with WordNet, one of the reasons is that WordNet contains a very restricted set of relations – synonymy, hyponymy, meronymy and antonymy (table 6.10). The majority of collocates do not fit into these categories. They are related to the node situationally through a wide range of lexical-semantic relations, which are often difficult or impossible to classify (cf. the attempt to classify collocates by [Halliday76] in section *2.4.1*). Moreover, engineered term networks do not contain proper nouns, which are common especially in MI lists.

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 the top 12 collocates 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 (table 6.11). This

highlights one more important difference between engineered term networks and collocates – each term in either INSPEC, or WordNet represents a concept, therefore complex concepts are represented by compound terms. Collocation lists built in this project, in contrast, always contain single words.

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

Table 6.11. Top terms in Z-ranked collocation list matching INSPEC and WordNet terms related to the term *fuel*

6.4 Retrieval experiments

6.4.1 Experimental design

The experiments were designed to evaluate separately query expansion with Z- and MI-ranked collocates. In section *6.2* the rationale was given for using a fixed number of top ranked collocates for query expansion, instead of setting a threshold on collocation scores. The general mechanism of query expansion with global collocates we used is to take the top *n* collocates of each query term ranked by one of the statistics, merge them together and remove duplicates. The original query terms were kept in the expanded query.

First it was decided to evaluate top 8 and 16 ranked collocates for each query term. If the results suggest any significant improvement, then other numbers of collocates could be considered.

Query expansion with global collocates was implemented using the Okapi system by creating databases of collocates in Okapi format. Four separate collocation databases were built:

top8-mi - top 8 MI-ranked collocates;
top16-mi - top 16 MI-ranked collocates;
top8-z - top 8 Z-ranked collocates;
top16-z - top 16 Z-ranked collocates.

Each database contains one record for each query term. Each record has three fields: dn (record number), kw (node) and co (collocates). The kw field of a record contains a node, and the co field contains a list of its collocates.

Two indexes are created for each database: dn – index of the record number field (dn), and te – index of the node field (kw). Each record was, thus, indexed by a single node term in the kw field.

Okapi collocation databases serve as an intermediate layer in the existing querying technique. When the initial query is submitted, it is searched first against one of the

collocation databases. Each query term matches a single record by the indexable node term in the kw field. The contents of both kw and co fields of all found records is then merged together and, after removing duplicates, submitted to FT 96 Okapi database as the expanded query (fig.6.1). For the Perl script implementing this algorithm see Appendix D.1.2.

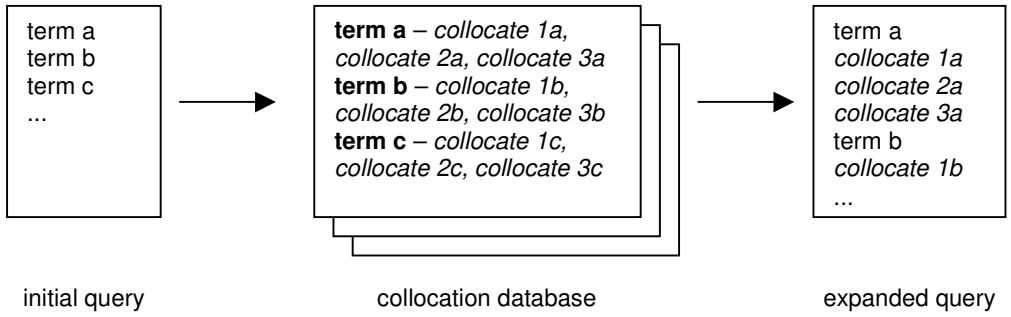

Figure 6.1. Query expansion with global collocates

The collocation database was searched using a simple unweighted search, since there is only one record corresponding to each query term. The expanded query was searched against FT 96 using weighted function bm2500 without relevance information, which instantiates Robertson-Sparck Jones' Combined Weight formula (see section *4.1.1*). The value for the tuning constant k_1 (section *4.1.1*), controlling the effect of within-document frequency (TF_i), was set to 1.2. Another tuning constant *b*, controlling document length normalisation, was set to 0.75. These values proved to work well in the previous Okapi runs.

First 1000 documents of each ranked set were evaluated using the trec_eval program and a file with TREC relevance judgements. Out of 50 topics (250-300), 6 topics (262, 270, 276, 279, 281, 296) have no relevant documents, and were ignored by trec_eval. For this reason all evaluation results are based on 44 topics.
We evaluated 4 query expansion runs:

top 8 MI – query expansion with top 8 MI collocates per query term;
top 8 Z – query expansion with top 8 Z collocates per query term;
top 16 MI – query expansion with top 16 MI collocates per query term.
top 16 Z – query expansion with top 16 Z collocates per query term.

The results of expanded query runs were compared to the results of Okapi unexpanded weighted search with the initial queries. The search was done also with bm2500 with the same settings for tuning constants.

6.4.2 Analysis of results

All query expansion runs – top 8 MI, top 8 Z, top 16 Z and top 16 MI – performed worse than Okapi run without query expansion, as the summary of trec_eval results shows in table 6.12.

	no expansion	top 8 MI	top 8 Z	Top 16 MI	top 16 Z
Retrieved	42686	43534	44000	44000	44000
Relevant	1583	1583	1583	1583	1583
Relevant retrieved at 1000	632	573	520	526	504
Average precision (non-interpolated) for all rel docs (averaged over queries)	0.1310	0.0432	0.0375	0.0344	0.0340

Table 6.12. Summary of retrieval results for query expansion with global collocates

The analysis of runs top 8 MI and top 8 Z by query shows that top 8 Z improved 7 queries, did not affect 3 and hurt 34 queries. top 8 MI improved 3, did not affect 2 and hurt 39 queries.

A large difference between the performance of all expanded runs and the unexpanded Okapi run suggested that fine tuning of variables such as the number of query expansion terms will not result in significant performance improvements over Okapi search with initial queries. Therefore, to decide whether further retrieval experiments with different variable values are necessary, it was decided to gain a deeper understanding of the potential of collocates to retrieve relevant documents. To achieve this we undertook a series of statistical analyses, described in the following section.

6.5 Statistical analysis

6.5.1 Presence of global collocates in relevant documents

We decided first to find out how many collocates occur in relevant documents. We did this by calculating for each collocate its term selection value in the probabilistic model – Offer Weight (OW), described in section *4.1.2*. Offer Weight below or equal to zero indicates that the term has no occurrence in relevant documents. To calculate OW scores of collocates, first their Relevance Weights (RW) have to be determined. We used the Okapi term weighting operation, which implements RW formula (section *4.1.1*). Weighting was done with complete relevance information, i.e. r_i = the number of all relevant documents with term *i* for a given query, *R* = total number of relevant documents for a given query. Relevance information was obtained from TREC relevance judgements. Offer Weight was then calculated by multiplying RW with r_i (the same as for RW).

Occurrence of collocates in relevant documents was calculated for top 30 MI, top 30 Z and top 100 Z collocates (table 6.13).

	Total number of collocates	Collocates with OW ≤ 0	Percentage of collocates present in relevant documents
Top 30 MI collocates	4080	3750	8%
Top 30 Z collocates	4080	2589	36.5%
Top 100 Z collocates	13261	9277	30%

Table 6.13. Presence of global collocates in relevant documents

We then conducted a correlation analysis between MI and OW on the top 100 MI collocates, and Z and OW on the top 100 Z collocates. It was then decided that a more suitable statistical method for the analysis of these variables would be regression analysis, described in the following section.

6.5.2 Regression analysis

Offer Weight is a standard term selection value in Okapi query expansion following relevance feedback. Previous experiments showed that Okapi relevance feedback achieves significant performance improvements over unexpanded runs. Based on these results we assumed that Offer Weight is the best available method in the probabilistic model to identify good relevance-discriminating terms for query expansion. The higher the Offer Weight, the stronger is the term's relevance discriminating capability. OW, however, is calculated when some relevance information is obtained. We carried out regression experiments in order to find out whether it is possible to predict OW from the statistical information on collocates available to us prior to relevance feedback.

We have the following variables characterising each collocate:

1. **MI** – MI score;
2. **Z** – Z score;
3. **JF** – joint frequency of the collocate and the node term (*f(x,y)* in MI and Z formulae);
4. **NOPOS** – number of postings (documents) containing the collocate in question.

Another variable – individual term frequency – was considered to give similar information as NOPOS, therefore was not used in the regression.

Multiple linear regression is a suitable technique to examine whether the above variables can be used to predict Offer Weight. The SPSS package was used for this purpose.

To explore the relationship between the independent variables above and the response variable OW, we first ran a regression on all four variables. All significant collocates (with Z > 1.6) of all query terms in 50 topics were used in the analysis. The summary statistics of this regression run are presented in table 6.14.

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.211(a)	.045	.045	7.9502
a Predictors: (Constant), NOPOS, Z, MI, JF				

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1915569.482	4	478892.370	7576.645	.000(a)
	Residual	41011024.834	648843	63.206		
	Total	42926594.316	648847			
a Predictors: (Constant), NOPOS, Z, MI, JF						
b Dependent Variable: OW						

Coefficients

Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	.181	.018	10.153	.000
	MI	-.309	.010	-30.805	.000
	Z	.152	.001	115.846	.000
	JF	1.787E-04	.000	26.518	.000
	NOPOS	1.090E-04	.000	49.193	.000
a Dependent Variable: OW					

Table 6.14. Summary for regression using MI, Z, JF and NOPOS to predict OW

The significance values in the table above show that all four regression coefficients are significantly different from 0. Z coefficient indicates that Z is the best predictor variable, while negative coefficient for MI indicates that it has negative correlation with OW. The positive intercept value shows that this four-variable model has some explanatory power.

Since Z appeared to be the best predictor variable, we re-ran the regression dropping all other variables in turn. The following regression models were tried:

1. Z, JF, NOPOS
2. Z, JF
3. Z, NOPOS
4. Z, MI
5. Z

The statistical characteristics of these models are given in Appendix A.3.

Comparison of the squared multiple correlation (R Square) values for all models showed that the initial four variable model (MI, Z, JF, NOPOS) has the best explanatory power. It was then decided to conduct a retrieval experiment with a four-variable regression equation as a collocate-ranking function. The following regression equation was used to calculate predicted Offer Weights (*PREDICTED OW*) for collocates in the analysis:

$$PREDICTED\ OW = .181 - .309 * MI + .152 * Z + 1.787E\text{-}04 * JF + 1.090E\text{-}04 * NOPOS$$

Collocates of each query term were then ranked by their PREDICTED OW scores. Since top 8 collocates per query term showed better results in the initial query expansion runs, than top 30, we also used top 8 PREDICTED OW-ranked collocates.

The algorithm for query expansion and the values for bm2500 tuning constants were the same as in the previous runs. The summary of results for this run is given in table 6.15.

	PREDICTED OW
Retrieved	44000
Relevant	1583
Relevant retrieved	554
Average precision (non-interpolated) for all rel docs (averaged over queries)	0.0364

Table 6.15. Summary of retrieval results for PREDICTED OW run

The recall/precision values in the table show that PREDICTED OW performs similarly to top 8 MI and top 8 Z runs.

Following these results, we conducted two more regression analyses. The first regression was run to predict OW on only those terms whose JF ≥ 30. Four explanatory variables were used as before. The second regression was run to predict RW on the data consisting of both collocates and original query terms. The explanatory variables included four variables as before plus another variable QT (query term) – a binary variable with values: 1 – the term is a query term and 0 – the term is a collocate. The characteristics of these models are given in Appendix A.3. All these models had weaker explanatory power than the initial four-variable model to predict OW. Retrieval run on the scores predicted by one of them also proved to be worse than PREDICTED OW run.

6.6 Concluding remarks

The hypothesis examined in this set of experiments was:

Hypothesis 1: Expansion of the initial query with statistically significant global collocates of query terms results in significant performance improvement over the initial query evaluated under the same conditions.

The experiment results achieved do not support this hypothesis. Retrieval performance of the top ranked collocates was significantly worse than no query expansion. Regression on the statistical parameters of collocates to predict their value as relevance-discriminating terms showed some explanatory power, although retrieval

results of collocates ranked by the best regression equation were very similar to using Z or MI alone.

Regression analysis also suggested that Z overall is a better statistic for predicting collocates' relevance discriminating power, than MI. The fact that the top 30 MI ranked terms only occur in 8% of relevant documents, in contrast to 36.5% for top 30 Z collocates, also suggests that Z is more suitable for ranking collocates in query expansion.

One conclusion was that the use of global collocates alone without any other information did not look promising. As mentioned earlier in this section, in view of such big difference in performance between expanded and unexpanded queries, fine tuning of different parameters, like the number of query expansion terms or ranking function, was not expected to lead to significant improvements.

A possible reason why global collocates performed badly was the fact that collocates come from the many different contexts in which query terms occur throughout the collection. Since the number of non-relevant documents is much larger than relevant, only a small proportion of these contexts occur in relevant documents and have anything to do with the query topic. Even occurrences of the same sense of the query term can be used to describe a wide variety of topics. If, for example, a term has very broad lexical meaning and occurs in a wide range of topics, then its significant collocates can relate to any of these topics, not necessarily to the one meant in the query. And, of course, the problem is aggravated further if the query term is polysemantic.

Many of the query terms we used have fairly broad lexical meanings. To a certain extend they narrow each other's meaning through combined use in the query. If many collocates of the query terms happen to be from unrelated topics, their addition simply introduces more noise to the query and degrades its performance.

It was realised that some solution is needed to reduce the number of collocates from unwanted contexts. One direction to follow is to extract collocates not from the corpus, but from a limited number of documents, for which there exists some relevance information, implied (in the case of blind feedback) or asserted via a relevance judgement. The next set of experiments, described in the following chapter, was designed to explore this research direction.