Term Dependence:
Truncating the Bahadur Lazarsfeld Expansion

Information Processing & Management
30 (2) 1994, 293–303.

Robert M. Losee, Jr.
University of North Carolina
Chapel Hill, NC 27599-3360 U.S.A.

Phone: 919-962-7150
Fax: 919-962-8071
losee@ils.unc.edu

June 28, 1998

Abstract

The performance of probabilistic information retrieval systems is studied where differing statistical dependence assumptions are used when estimating the probabilities inherent in the retrieval model. Experimental results using the Bahadur Lazarsfeld expansion suggest that the greatest degree of performance increase is achieved by incorporating term dependence information in estimating $\Pr(d|rel)$. It is suggested that incorporating dependence in $\Pr(d|rel)$ to degree 3 be used; incorporating more dependence information results in relatively little increase in performance. Experiments examine the span of dependence in natural language text, the window of terms in which dependencies are computed and their effect on information retrieval performance. Results provide additional support for the notion of a window of ± 3 to ± 5 terms in width; terms in this window may be most useful when computing dependence.

1 Introduction

Those who study information retrieval often assume that the features or terms used in both queries and document representations are statistically independent. The assumption of statistical independence is obviously and openly understood to be wrong; it is made because of the great expense that is expected to be incurred if higher order dependencies are used in estimating probabilities. Some research incorporates dependence in a limited manner. If one wishes to study the incorporation of this increased information, the primary problem becomes determining how much dependence needs to be incorporated to obtain close to the best results that are obtainable, given the relatively high costs of incorporating a great deal of dependence information.

This research has been motivated by two concerns. The first was to what extent dependence based estimates are beneficial for estimating probabilities used in computing document weights for ranking documents in order of decreasing expected worth. If a particular form of dependence estimate does not result in a significant increase in system performance, then it need not be made.

A second concern is whether dependence between terms only needs to be computed for terms that are in close proximity in the query, that is, within 3 to 5 terms of one another, or whether much is gained by incorporating dependence in estimates for all terms in a query.

Readers familiar with the foundations of probabilistic models may wish to pass over the next two sections, which introduce these models.

2 Information Retrieval Models

Models of information retrieval systems usually suggest that documents be assigned a *retrieval status value* by which documents may be ranked, with the highest ranked document presented to the searcher first, followed by the presentation of documents of expected lower value [1, 15, 18]. One specific model of retrieval is the probabilistic model, which uses the following retrieval rule:

> A document should be retrieved if the expected cost of retrieving the document ($EC_{retrieve}$) is less than the expected cost of not retrieving the document ($EC_{not-retrieve}$).

More formally, a document should be retrieved if

$$EC_{retrieve} < EC_{not-retrieve}.$$

The expected costs of retrieving or not retrieving a document may be estimated, transforming the retrieval rule to

$$\Pr(rel|d)C_{retr,rel} + \Pr(\overline{rel}|d)C_{retr,\overline{rel}} \\ < \Pr(rel|d)C_{\overline{ret},rel} + \Pr(\overline{rel}|d)C_{\overline{ret},\overline{rel}}.$$

where $\Pr(Rel|d)$ represents the conditional probability that a document is relevant given that it has a set of characteristics d, with the n individual characteristics being $d_1, d_2, \ldots, d_n$. $C_{retr,rel}$ is the cost of retrieving a relevant document, while $C_{retr,\overline{rel}}$ represents the cost of retrieving a non-relevant document, with similar notation for not retrieving a document. The decision to retrieve a document may now be transformed to: Retrieve a document with characteristics d if and only if

$$\frac{\Pr(rel|d)}{\Pr(\overline{rel}|d)} > \frac{C_{retr,\overline{rel}} - C_{\overline{ret},\overline{rel}}}{C_{\overline{ret},rel} - C_{retr,rel}},$$

where the right hand side of this expression is a cost constant that must be exceeded if retrieval is to occur. Documents may then be ranked by the value of the left hand side of this formula [15, 14]. This value may be estimated as

$$\frac{\Pr(rel|d)}{\Pr(\overline{rel}|d)} = \frac{\Pr(d|rel)\ \Pr(rel)}{\Pr(d|\overline{rel})\ \Pr(\overline{rel})}.$$

.

3 Term Independence

If the n features are assumed to be statistically independent in the document, that is, $d = d_1 d_2 \cdots d_n$, and are independent in both the set of relevant and the set of non-relevant documents, then the weight for a document may be computed as

$$\frac{\Pr(rel|d)}{\Pr(\overline{rel}|d)} = \prod_{i=1}^{n} \frac{\Pr(D_i = d_i|rel)}{\Pr(D_i = d_i|\overline{rel})}\ \frac{\Pr(rel)}{\Pr(\overline{rel})}.$$

Removing the portion constant for all documents, documents may be ranked by

$$\prod_{i=1}^{n} \frac{\Pr(d_i|rel)}{\Pr(d_i|\overline{rel})}.$$

If features are assumed to be binary, that is, the probability that a feature has value d with probability p (q) in relevant (non-relevant) documents, respectively, the probability of the feature is estimated as

$$\Pr(d|rel, p) = p^d (1-p)^{(1-d)}$$

and

$$\Pr(d|\overline{rel}, q) = q^d (1-q)^{(1-d)}.$$

Therefore,

$$\prod_{i=1}^{n} \frac{\Pr(d_i|rel)}{\Pr(d_i|\overline{rel})} = \prod_{i=1}^{n} \left(\frac{p_i/(1-p_i)}{q_i/(1-q_i)} \right)^{d_i}$$

ignoring factors constant for all documents. This expression may be modified to

$$\sum_{i=1}^{n} d_i \log \frac{p_i/(1-p_i)}{q_i/(1-q_i)}$$

without effecting the ranking of documents.

4 Computing Term Dependence

Term dependencies exist when the relationships between terms in documents are such that the presence or absence of one term provides information about the probability of the presence or absence of another term. The dependencies may be computed using a number of different techniques, depending on the retrieval model that is used. These computational techniques vary the degree of term dependence computed as well as the accuracy and computational speed of the estimates. We describe these method here.

The most commonly used commercial retrieval model is the Boolean model. Models of term dependencies have been proposed and tested that assume knowledge of specific correlations between terms [4]. Another model has been proposed that assumes that when Boolean queries are placed in conjunctive normal form, most of the dependence exists between the disjunctions of terms [12]. The former work emphasizes the relationships in a Boolean expression between what are suspected to be the most highly related terms in the Boolean query, while the latter tries to incorporate into a retrieval model assuming independence those terms or hyperterms that are most likely to be statistically independent.

Probabilistic models, such as that developed earlier, require that probabilities be estimated. Another technique that has been proposed to incorporate term dependencies has been to use the maximum entropy technique [3, 8]. This method assigns values to probabilities in such a way that randomness is maximized wherever possible, that is, where there is no information to the contrary. Requiring relatively long periods of time to compute, parameter values derived from this technique cannot at present be estimated in real time for practical system use, although this may change as parallel hardware and software become more widely available at lower costs.

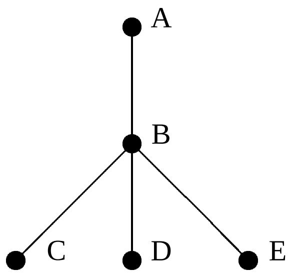

Figure 1: Maximum spanning tree.

Term dependencies may be computed using another method if one is willing to arbitrarily limit the dependencies considered to those expected to have the most effect on the results [17]. Chow and Liu suggest the construction of a tree such that the mutual information between an item and the item immediately above it are maximized (Figure 1) [2]. Given two points on the tree such that the *i*th point is directly and immediately above the *j*th point, a Maximum Spanning Tree (MST) may be defined as that tree maximizing the sum:

$$\sum_{i,j} I(Node_i, Node_j),$$

where $I(i,j)$ represents the expected mutual information provided by i about j,

$$I(i,j) = \sum_{i,j} \Pr(i,j) \log \frac{\Pr(i,j)}{\Pr(i)\Pr(j)}.$$

Consider five terms, A through E, with mutual informations such that their maximum spanning tree looks like Figure 1. If term independence were assumed, the probability that one would find a body of text with terms A, C, and E only, would be

$$\Pr(A,C,E) = \Pr(A=1)\Pr(C=1)\Pr(E=1)\Pr(B=0)\Pr(D=0).$$

If, however, the dependence information provided by the MST were used, one might more accurately calculate this probability as

$$\begin{aligned}\Pr(A,C,E) &= \Pr(A=1)\Pr(C=1|B=0)\Pr(E=1|B=0)\\ &\quad \times\Pr(B=0|A=1)\Pr(D=0|B=0).\end{aligned}$$

Thus the information that one node provides about another neighboring node is used; the probability of a particular node on the graph having a given value is conditioned by the probability that the node above it has a certain value.

5 The Bahadur Lazarsfeld Expansion

Document probabilities may also be estimated based on the Bahadur Lazarsfeld Expansion (BLE). When the full expansion is used, an exact probability is calculated, while if the expansion is truncated, an estimate of the probability is computed. The expansion begins with the estimate of the independent probability, and then multiplies this by a correction factor. Truncating the Bahadur Lazarsfeld expansion reduces the accuracy of the correction factor. The correction factor consists of a series of individual factors consisting of the correlations between two or more terms and a factor, such that when the full expansion is used, the exact probability is computed. Computed as

$$\begin{aligned}\Pr(d) &= \prod_{i=1}^{t} p_i^{d_i}(1-p_i)^{(1-d_i)}\Bigg[1 + \\ &\sum_{i<j} \varrho_{ij}\frac{(d_i-p_i)(d_j-p_j)}{\sqrt{p_i p_j (1-p_i)(1-p_j)}} + \\ &\sum_{i<j<k} \varrho_{ijk}\frac{(d_i-p_i)(d_j-p_j)(d_k-p_k)}{\sqrt{p_i p_j p_k (1-p_i)(1-p_j)(1-p_k)}} + \cdots + \\ &\varrho_{12\ldots t}\frac{(d_1-p_1)(d_2-p_2)\ldots(d_t-p_t)}{\sqrt{p_1 p_2 \ldots p_t (1-p_1)(1-p_2)\ldots(1-p_t)}}\Bigg],\end{aligned}$$

the sum may be arbitrarily truncated so that one can include all dependence up to term pairs, three-way dependence (term triples), and so forth. The correlations are computed as

$$\varrho_{1,2,\ldots i} = \frac{E\left[(d_1-p_1)(d_2-p_2)\cdots(d_i-p_i)\right]}{\sqrt{p_1 p_2 \cdots p_i (1-p_1)(1-p_2)\cdots(1-p_i)}}$$

As Yu et al. (1983) discuss, truncation of the Bahadur Lazarsfeld expansion can result in improper probabilities, that is, estimates produce negative probabilities or probabilities over 1. Such estimates, besides being obviously “wrong,” can play havoc with calculations using these probabilities, resulting in system errors.

6 Experimental Techniques

The experiments to be conducted here use the Bahadur Lazarsfeld expansion with varying degrees of truncation to estimate probabilities. The documents are then ranked and the quality of the ranking analyzed.

These tests use the Cystic Fibrosis (CF) database developed at the University of North Carolina [16, 19]. The CF database contains 100 natural language queries, 1239 document abstracts, and exhaustive relevance judgements. The quality of the relevance judgements is felt to be high, making this an attractive database for experimentation. The abstracts were used as the document representations for these experiments when available. About 17 common words, including "CF," were removed from the database to improve processing speed for the first set of experiments described here.

Parameters have been estimated using "retrospective" techniques, that is, they are estimated before the retrieval process begins with full knowledge of the characteristics of relevant and non-relevant documents. They are not estimated as the documents are retrieved and knowledge is gained, simulating the operation of a production system. Because of the small size of many sets of relevant documents from which parameters must be estimated, it is often the case that parameters have the value 0 or 1. While these are inadmissible probabilities, they are used here to accurately reflect the values present; we chose not to pursue the problem of prior knowledge about parameter values [10]. System software was developed that made special provisions for these values.

Documents have been ranked by the Expected Precision (EP) of the document [9, 11]. Computed as

$$\Pr(rel|d) = \frac{\Pr(d|rel)\Pr(rel)}{\Pr(d)},$$

the probability a document is relevant is computed here using retrospective techniques, with advanced knowledge of all probabilities and correlations for the appropriate relevance class. We note that the expected precision is similar to the ranking formula traditionally used in information retrieval experiments, using a ratio of the probability that a feature is found in a relevant document to the probability that the feature is found in a non-relevant document, with the probability that the feature is found in non-relevant documents estimated by the probability that the feature is found in the database [5, 10].

Retrieval performance is measured here by computing the Average Search Length (ASL), the average number of documents retrieved when retrieving a given relevant document. Assuming that the first document retrieved has rank 1, the second has rank 2, etc., average search length is the average rank for the set of relevant documents. This measure computes a mean rank for any set of documents with equal expected precision; this mean rank is then used as the rank for each relevant or non-relevant document in this set. The average search length was chosen for this work because it is a single number measure and may be easily understood.

In addition, average search length and expected precision were chosen as evaluation and ranking procedures for this work because of their analytic tractability, an area the author is pursuing in other research.

This method of analyzing retrieval performance is sensitive to the relatively few documents retrieved at the end of the retrieval process. Retrievals are thus also studied in terms of their fractional average search length (FASL), which is computed as with average search length except that instead of the rank of documents being retrieved being used in computations, the inverse of each of these ranks is used. This has the effect of providing a measure which weights the documents that are retrieved first more heavily than those retrieved at the end of the retrieval process. These latter documents are less important, in a sense, as they are less likely to be retrieved by a searcher who might give up before retrieving these difficult-to-find documents.

As an example, consider a retrieval set with two relevant documents at ranks 2 and 4. The average search length is thus 3. The average of the fractional values, $1/2$ and $1/4$, is $3/8$, which represents a lower rank (a higher decimal value) than would be found by inverting the average search length, i.e., $1/3$. Note that unlike average search length, where low values are "better" than high values (with 1 being the lowest possible average search length), fractional average search length is better the higher the value, with the best possible fractional average search length being 1 and the worst approaching 0.

A similar measure has been developed which computes the logarithm of the search length when computing the averages. As with fractional average search length, this measure weights documents retrieved early in the search more heavily than documents retrieved latter. The results obtained with this logarithmic measure are similar to those found with the fractional average search length and are thus not reported here.

7 Effects of Varying Degrees of Dependence

The results of a set of simulated searches using the 1239 documents and 100 queries in the CF database are reported in Table 1. Note that the variance of the average search lengths within any given column in Table 1 is much smaller than for any given row. This is because the degree of dependence for $\Pr(d|rel)$ has a greater degree of impact on the average search length than does the degree of dependence on $\Pr(d)$. One conclusion that can be drawn from this is that increasing the degree of dependence used in estimating probabilities does result, in most cases, in an increase in information retrieval performance. One can also conclude that the increase in performance is due primarily to the increased accuracy of the estimates

Average Search Length
for Varying Degrees of Dependence in BLE

		Degree of Dependence for $\Pr(d \vert rel)$				
		1	2	3	4	5
Degree	1	266.64	225.20	214.49	209.59	211.69
of	2	279.52	231.24	217.30	211.29	213.22
Dependence	3	284.20	232.81	220.16	213.55	215.56
for	4	282.06	227.87	214.79	207.35	209.19
$\Pr(d)$	5	282.25	227.08	213.37	206.28	208.13

Table 1: Average search length for retrieval with full set of database queries. Stop words are removed. 1 indicates term independence, 2 pairwise dependence, etc.

of $\Pr(d|rel)$. Note that $\Pr(d|rel)$ ranges from average search length's of 267 to 212 for a fixed dependence of degree 1 and ranges from 216 to 208 for a fixed dependence of degree 5 for estimating $\Pr(d)$.

This variation in retrieval performance may be studied by examining the percent increase in performance (or decrease in average search length) when a certain degree of dependence is incorporated. When pairwise dependence is used for both probabilities, a 15% decrease in average search length occurs. This is within the range of precision value increases reported by [20] for the two databases they analyzed. When term triples are used, a 21% decrease in average search length is found, similar to the 15% to 34% increase in precision found by Yu et al. Using a degree of dependence of 5 when estimating both probabilities only results in a 29% decrease in average search length. Note that a 24% decrease in average search length is obtained when using term triples in estimating $\Pr(d|rel)$ and assuming independence when estimating $\Pr(d)$.

These results suggest that practical retrieval systems gain very little by computing dependence information when estimating $\Pr(d)$ and gain little when computing $\Pr(d|rel)$ *beyond the third order of dependence* (e.g., beyond three way dependencies).

Table 2 shows the robustness of the average search length. The queries in the CF database are ordered by subject and thus the first group of queries can be treated as different than the second group, and so on [16]. The variance seen here may thus provide some indication of the difference encountered in actual academic searches. This table suggests that about 90% of the reduction in average search length can be obtained by estimating $\Pr(d)$ assuming independence and by estimating $\Pr(d|rel)$ assuming only third order dependence factors.

Because the average search length appears to be heavily influenced by those

Robustness of average search length

				Queries in CF Database		
				Quarter		
$\Pr(d\vert rel)$	$\Pr(d)$	$1-25$	$26-50$	$51-75$	$76-100$	$1-100$
1	*Indep*	243.97	294.69	294.30	231.79	266.64
2	*Indep*	209.43	263.29	226.99	199.81	225.20
2	2	208.63	269.46	232.69	212.35	231.24
3	*Indep*	205.18	258.50	208.78	184.74	214.49
3	3	202.43	263.87	209.89	203.02	220.16
4	*Indep*	201.48	253.73	204.66	177.84	209.59
4	4	199.30	252.53	199.47	177.46	207.35
5	*Indep*	202.07	253.98	205.22	184.73	211.69
5	5	196.29	250.86	200.24	184.18	208.13

Table 2: Average search length of searches from each quarter of the CF database. Left two columns represent the degrees of dependence, and where dependence of degree 1 is denoted by "Indep."

documents most difficult to retrieve, the fractional average search length was computed and is reported in Table 3. Fractional average search length minimizes the effect of a single document that is not retrieved till much later in a search and would probably **not** be retrieved in a practical search situation. When using term triples in estimating $\Pr(d|rel)$ and assuming independence in estimating $\Pr(d)$, the fractional average search length gain is 70% of that possible with dependence of degree 5 used in estimating both probabilities. This supports the notion that term triples and independence may be a satisfactory compromise position between retrieval performance and computation time when estimating $\Pr(d|rel)$ and $\Pr(d)$, respectively,

The fluctuations in average search length as dependence is varied may be explained in part by examining the effect of increasingly accurate parameter estimates on retrieval performance. When ranking documents by $\Pr(rel|d)$, one may understand the ranking process as ranking documents in decreasing order by two other factors, $\Pr(d|rel)$ and $1/\Pr(d)$. Ordering documents by $\Pr(rel|d)$ thus involves making a tradeoff between these two orderings.

As the degree of dependence is increased when estimating the two probabilities, differing tradeoffs are made, resulting in different ASLs. The estimate of $\Pr(d|rel)$, assuming independence, may tend to be a cruder estimate of the true parameter value than is the independence based estimate of $\Pr(d)$, because the lat-

Fractional average search length for Varying Degrees of Dependence

		Degree of Dependence for $\Pr(d\|rel)$					
		1	2	3	4		5
Degree	1	0.177	0.186	0.190	0.191	0.190	
of	2	.170	.187	.194	.195	.195	
Dependence	3	.160	.190	.195	.198	.197	
for	4	.146	.185	.193	.195	.194	
$\Pr(d)$	5	.141	.185	.192	.194	.194	

Table 3: Average search length with rank computed as 1/rank. 1 indicates term independence, 2 pairwise dependence, etc.

ter is computed from a much larger data set. On the other hand, the estimate of $\Pr(d|rel)$ usually converges to its exact value more rapidly than does $\Pr(d)$ because of the smaller number of relevant documents and thus the smaller number of relationships between terms used in computing $\Pr(d|rel)$.

Other fluctuations in performance are due to the high degree of sensitivity of the average search length to these ordering variations. Examining the ASLs for individual queries reveals that several queries have ASLs that move from being one or two digits values to being three digit numbers as a degree of dependence is added, and then move back again when another degree or two of dependence is added. In particular, the appearance that performance drops slightly when second and third order dependencies are used in estimating $\Pr(d)$ may be in part an artifact of this data set.

8 Detailed Study of a Query

A more detailed study of a query may help provide a deeper level of understanding of the magnitudes of values encountered in computing document profile probabilities with dependence information incorporated. This examination includes all words, including "stopwords," in the analysis. Query 18 in the CF database,

> Is dietary supplementation with bile salts of therapeutic benefit to cf patients?

is used as our example.

Several documents having high rankings are shown in Table 4. Documents with an "R" on the left of the document number are relevant documents for the query, while those preceded by an "N" are non-relevant. The "Y" indicates that a

Query:	Is	diet	supl	with	bile	salt	of	ther	benf	to	cf	patn
$\Pr(d_i\|rel)$	1	0	0	1	1	0	1	0	0	1	0.5	0.5
$\Pr(d_i)$	.32	.011	.0056	.54	.015	.0008	.62	.012	.010	.51	.19	.40
R129	Y			Y	Y		Y			Y		
R205	Y			Y	Y		Y			Y	Y	Y
$= N355$												
$= N424$												
N266	Y			Y	Y		Y			Y	Y	
N657	Y			Y	Y		Y			Y		Y
$= N660$												
$= N725$												

Table 4: Query 18 and characteristics.

particular document contains the term in question. Note that only those terms in the query are used and that those documents which have identical profiles (when only considering terms in the query) are grouped together. Near the top of this table are the probabilities that the term occurs given that the document is relevant and an unconditional probability that the term occurs.

The factor $\Pr(d|rel)$ may be computed for R129 rather easily (as in Table 5) by noting that the only non-unit contributions to $\Pr(R129|rel)$ are supplied by "cf" and "patients," each of which has a probability of .5. The probability, $\Pr(R129|rel) = 1 \times .5^0(1-.5)^{1-0} \times .5^0(1-.5)^{1-0} = .5 \times .5 = .25$, the value computed assuming independence of features.

The two terms "cf" and "patients" have a correlation over the set of relevant documents of 1. Correlations between all other term pairs over the set of relevant documents are zero. The probability $\Pr(R129|rel)$ may be computed as the independent probability, .25, times the correction supplied by the Bahadur Lazarsfeld expansion expansion. This correction is 1 plus the sum of (the correlations times

$$(d_i - p_i)(d_j - p_j)/\sqrt{p_i p_j (1-p_i)(1-p_j)})$$

. The correction, given the two factors each with probability of .5, becomes

$$1 + ((0-.5)(0-.5))\,/\sqrt{(.5 \times .5 \times (1-.5)(1-.5))} = 1 + (.25/.25) = 2.$$

Multiplying the independent probability, .25, by the correction factor, 2, provides the probability assuming pairwise dependence, .5.

The rankings of these documents for varying degrees of dependence is indicated in Table 6. As the degree of dependence is increased, the ranking continues

Expected Precision for 8 Documents with
Varying Degrees of Dependence

Doc.	Degree	$\Pr(d \mid rel)$	$\Pr(d)$	$\Pr(rel \mid d)$
R129	1	.25	.000406	.9932
	2	.5	.00173	.466
	3	.5	.00159	.509
R205/N355/N424	1	.25	.0000626	6.45
	2	.5	.000864	.934
	3	.5	.000614	1.314
N266	1	.25	.0000931	4.33
	2	.0	.000752	.0
	3	.0	.000630	.0
N657/N660/N725	1	.25	.000273	1.48
	2	.0	.00237	.0
	3	.0	.00121	.0

Table 5: $\Pr(rel \mid d)$ is computed as $\Pr(rel)$, $.001614$, times $\Pr(d \mid rel)$ divided by $\Pr(d)$. The first two columns of probabilities are computed assuming the degree of dependence (for both probabilities) indicated in the "Degree" column.

Dependence	*ASL*	*Ranked Documents*
1	5	(R205, N355, N424), N266, (N657, N660, N725), R129
2	3	(R205, N355, N424), R129
3	3	(R205, N355, N424), R129
4	3	R129, (R205, N355, N424)
5	3	R129, (R205, N355, N424)

Table 6: Document rankings for varying degrees of dependence. Documents with equal ranking are grouped in parentheses.

to improve, although this does not improve the average search length once pairwise dependence is incorporated. In some circumstances, which are not present in this query, increasing the dependence results in groups of documents with identical profiles being moved far back in the document ranking. This may have the effect of increasing the average search length.

9 The Span of Dependence

The preceding performance figures were based on retrieval using probabilistic estimates of relevance assuming varying degrees of dependence between all terms in the query. However, the *span of dependence* may be limited so that dependence is only computed between terms within a certain proximity. More precisely, we use the *span* to represent the maximum number of intervening terms that may occur between two terms if the dependence between them is to be computed. Unlike experimental results reported above, these experiments used the full CF database queries with no stopwords removed. Thus the span of dependence as reported here represents the span between all terms, both common "stopwords" and non-"stopwords."

Haas and He note that "most lexical relationships between words probably appear in a window size ranging from ± 3 to ± 5 words" for text in the English language [6]. This work attempts to determine for the CF database whether knowledge about lexical relationships, as indicated by a dependency between terms, results in a significant difference in retrieval performance for different spans of dependence [7].

Figure 2 indicates the average search lengths found for varying spans of dependence for pairwise dependence (represented by a "2") and for the 3-way dependence (represented by a "3") for the estimation of $\Pr(d|rel)$, assuming independence for the estimation of $\Pr(d)$. These results suggest that there is usually a decrease in the average search length as the span of dependence is increased and thus more accurate estimates of the probabilities are obtained. The majority of the decrease appears to be with a span of dependence being in the range from 2 or 3 to 5 . From a pragmatic standpoint, there appears to be a point of diminishing returns as the span is increased, as the accuracy of dependence estimates increases at a slow rate.

10 Summary and Conclusions

The time it takes to compute a probability assuming a degree of dependence t, given n terms in the query, is roughly proportional to n^t for larger n and smaller

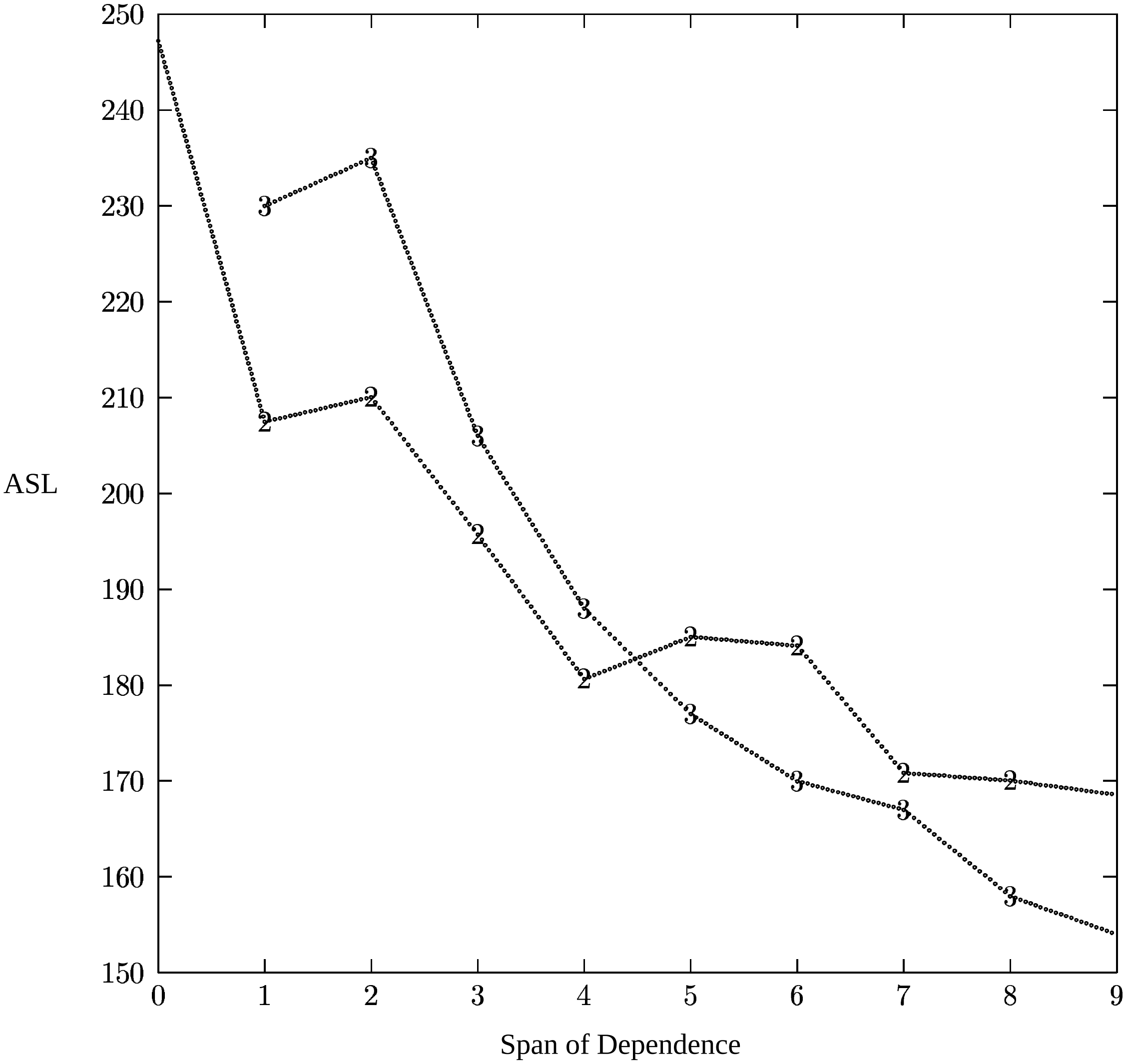

Figure 2: Average Search Length where the dependence is limited to pairwise dependence (represented by “2”) or to term triples (“3”) when estimating $\Pr(d|rel)$. Independence is assumed for $\Pr(d)$.

t. Because the time necessary to compute increasing degrees of term dependence grows so rapidly, it is desirable to keep the degree of dependence t as small as possible. The experimental results discussed here suggest that computing the probability for the numerator of the weighting formula ($\Pr(d|rel)$) benefits far more from the incorporation of dependence information than does computing the probability for the denominator ($\Pr(d)$) assuming term dependence.

Computing the probability for the numerator results in relatively little additional increase in performance once three-way dependence is incorporated, given the sharp increase in the amount of work necessary to compute higher order dependencies. The results suggest that term triples might be profitably used in computing the probabilities for $\Pr(d|rel)$ while independence best might be assumed when estimating $\Pr(d)$.

Experiments studying the span of dependence suggests that most of the improvement in information retrieval system performance occurs when dependence information is used from terms with less than or equal to 5 terms intervening. Thus, dependence may be limited to these terms, decreasing the computation time needed when ranking documents.

References

[1] Abraham Bookstein. Information retrieval: A sequential learning process. *Journal of the American Society for Information Science*, 34(4):331–342, September 1983.

[2] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, May 1968.

[3] William S. Cooper and P. Huizinga. The maximum entropy principle and its application to the design of probabilistic retrieval systems. *Information Technology: Research and Development*, 1(2):99–112, 1982.

[4] W. Bruce Croft. Boolean queries and term dependencies in probabilistic retrieval models. *Journal of the American Society for Information Science*, 37(2):71–77, March 1986.

[5] W. Bruce Croft and D.J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of Documentation*, 35(4):285–295, December 1979.

[6] Stephanie W. Haas and Shaoyi He. Toward the automatic identification of sublanguage vocabulary. *Information Processing and Management*, 29(6):721–732, 1993.

[7] Stephanie W. Haas and Robert M. Losee. Looking in text windows: Their size and composition. *Information Processing and Management*, 30(5):619–629, 1994.

[8] Paul B. Kantor. Maximum entropy and the optimal design of automated information retrieval systems. *Information Technology: Research and Development*, 3(2):88–94, April 1984.

[9] Robert M. Losee. Predicting document retrieval system performance using an expected precision measure. *Information Processing and Management*, 23(6):529–537, 1987.

[10] Robert M. Losee. Parameter estimation for probabilistic document retrieval models. *Journal of the American Society for Information Science*, 39(1):8–16, January 1988.

[11] Robert M. Losee. An analytic measure predicting information retrieval system performance. *Information Processing and Management*, 27(1):1–13, 1991.

[12] Robert M. Losee and Abraham Bookstein. Integrating Boolean queries in conjunctive normal form with probabilistic retrieval models. *Information Processing and Management*, 24(3):315–321, 1988.

[13] M. Phillips. *Aspects of Text Structure*. Elsevier, Amsterdam, 1985.

[14] Stephen E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.

[15] Stephen E. Robertson, C. J. Van Rijsbergen, and M.F. Porter. Probabilistic models of indexing and searching. In Robert Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, *Information Retrieval Research*, pages 35–56, London, 1981. Butterworths.

[16] William M. Shaw, Jr., Judith B. Wood, Robert E. Wood, and Helen R. Tibbo. The cystic fibrosis database: Content and research opportunities. *Library and Information Science Research*, 13:347–366, 1991.

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

[18] C.J. Van Rijsbergen. *Information Retrieval*. Butterworths, London, second edition, 1979.

[19] Judith B. Wood, Robert E. Wood, and W. M. Shaw. The cystic fibrosis database. Technical Report 8902, University of North Carolina, School of Information and Library Science, Chapel Hill, N.C., September 1989.

[20] Clement T. Yu, Chris Buckley, K. Lam, and Gerard Salton. A generalized term dependence model in information retrieval. *Information Technology: Research and Development*, 2(4):129–154, 1983.