CVS: A Correlation-Verification Based Smoothing Technique on Information Retrieval and Term Clustering

Christina Yip Chung
Verity, Inc.
892 Ross Dr, Sunnyvale, CA 94087
U.S.A.
cchung@verity.com

Bin Chen
Exelixis, Inc.
170 Harbor Way, S. San Francisco, CA 94083
U.S.A.
bchen@exelixis.com

ABSTRACT

As information volume in enterprise systems and in the Web grows rapidly, how to accurately retrieve information is an important research area. Several corpus based smoothing techniques have been proposed to address the data sparsity and synonym problems faced by information retrieval systems. Such smoothing techniques are often unable to discover and utilize the correlations among terms.

We propose CVS, a Correlation-Verification based Smoothing method, that considers co-occurrence information in smoothing. Strongly correlated terms in a document are identified by their co-occurrence frequencies in the document. To avoid missing correlated terms with low co-occurrence frequencies but specific to the theme of the document, the joint distributions of terms in the document are compared with those in the corpus for statistical significance.

A common approach to apply corpus based smoothing techniques to information retrieval is by refining the vector representations of documents. This paper investigates the effects of corpus based smoothing on information retrieval by query expansion using term clusters generated from a term clustering process. The results can also be viewed in light of the effects of smoothing on clustering.

Empirical studies show that our approach outperforms previous corpus based smoothing techniques. It improves retrieval effectiveness by 14.6%. The results demonstrate that corpus based smoothing can be used for query expansion by term clustering.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; G.1.1 [**Numerical Analysis**]: Interpolation—*Smoothing*

SIGKDD 2002, Edminton, Canada

Keywords

Text mining, smoothing, query expansion, term clustering, information retrieval

1. INTRODUCTION

Information volume in enterprise systems and in the Web is growing rapidly. Text mining research facilitates users to effectively mine valuable information from huge amount of data by scoring documents relevant to user queries ([1]). One of the widely used approaches employs a unigram model to represent documents. Other methods include the use of a vector space model, in which a document and a query are modelled as vectors of terms. The dot product of the vectors measures the relevancy of the document to the query.

These models suffer from the data sparsity problem — the dimension of terms is huge such that the vectors of documents and queries are sparse. The problem is escalated by the synonym problem where documents containing synonyms to terms in a query are often not assigned high scores by the dot product approach.

Smoothing techniques overcome these problems by associating feature terms in a given document with related terms that do not appear in the document. A *feature term* of a document is a term with non-zero occurrence frequency and is relevent to the theme of the document. A *related term* does not occur in the document but is related to its theme. The term *smoothing* refers to the adjustment of the maximum likelihood estimator of a language model ([26]). Consequently, a document can be represented more accurately with both the feature terms and related terms.

This paper proposes CVS, a Correlation-Verification based Smoothing method that is based on corpus statistics. CVS considers joint distributions of terms to identify significant correlations. We study the effects of CVS and several corpus based smoothing techniques on query expansion and term clustering. Several variants of CVS will also be described.

Most smoothing techniques rely on corpus statistics, relevance feedback and lexical references. Prior corpus based smoothing techniques refine document vectors with the distribution of individual terms in a corpus. CVS uses co-occurrence information of terms to smooth a document. Terms that co-occur frequently in documents are regarded as strongly correlated. To avoid missing terms with low co-occurrence frequencies in a document but related to its theme, the joint distributions of the terms in the document and the corpus are compared for statistical significance.

We also study how corpus based smoothing techniques can

be applied to enhance recall in information retrieval. Information retrieval can be improved by adjusting weights of related terms in document vectors for more accurate representation or by expanding a query to include related terms. Query expansion is a more transparent process because it allows users to view and modify terms in an expanded query. Smoothing techniques based on lexical references and relevance feedback typically lend themselves naturally to query expansion. Prior work on corpus based smoothing methods applies smoothing to refine document vectors. CVS combines corpus based smoothing method with query expansion.

A class of query expansion methods identify correlated terms by their joint distributions and expand a query directly with correlated terms. These methods do not consider the correlation of term pairs in light of the relationship with other term pairs in the corpus. Our approach augments these methods by selecting better terms for query expansion by clustering. As opposed to *document* clustering ([9, 19]), we consider *term* clustering ([22, 15, 23, 3]). Term clustering groups related terms into a hierarchy of clusters by minimizing intra-cluster distances and maximizing inter-cluster distances. User queries are expanded by terms in the clusters that contain the terms in the query. A hierarchy of clusters gives a user guidance in adjusting the degree of query expansion based on depths of clusters.

Empirical studies show that our approach outperforms two representative corpus based smoothing techniques. It improves retrieval effectiveness by 14.6% as compared to 6.8% and 10.8% by other methods. The studies also demonstrate that corpus based smoothing can be applied to improve information retrieval by term clustering and query expansion.

This paper is organized as follows, Section 2 introduces prior work; Section 3 discusses CVS; Section 4 applies corpus based smoothing techniques to query expansion by term clustering; Section 5 presents results of empirical studies; Section 6 comments on CVS method and discusses several variants; and finally, Section 7 concludes the paper.

2. PRIOR WORK

Various smoothing techniques have been proposed in the literature. One class of techniques use purely corpus statistics to refine term distributions in documents ([26, 6, 10]). Our approach belongs to this category, but differs from them by considering correlations among terms in addition to the distributions of individual terms. Another class of smoothing techniques uses lexical references to expand a query with additional, lexically related terms ([24, 16]). A drawback of this approach is that a lexical reference cannot capture idiosyncrasies of a corpus. Prior results of using lexical references for query expansion are not encouraging. The third class of techniques use relevance feedback to expand a query ([18, 25]). Relevance feedback is effective only if accurate relevance feedback information is available, which requires user intervention.

Most traditional work on comparing term distributions in two corpora applies statistical tests, such as χ^2 test, to compare the distributions of terms in the corpora ([12, 13, 8, 11, 20]). Kilgarriff ([11]) finds that χ^2 test identifies too many common terms as distributed differently in the corpora. He proposes to use Mann-Whitney rank test to unveil the statistical significance in term distributions. Such rank tests require sufficiently big corpora. But since most documents are short, [1] rank tests are almost inapplicable without major modifications.

A class of query expansion techniques directly expand a query with correlated terms ([2, 5, 4, 7]). A drawback is that the affinity between a pair of terms is not viewed in light of relationships among other term pairs in the corpus. For example, the term "insurance" is strongly correlated with the terms "business directory" and "instant insurance quote", but "business directory" may be correlated with terms not related to "insurance". Expanding "insurance" with "instant insurance quote" is better than with "business directory". This paper addresses this shortcoming by using term clusters obtained from term clustering to expand a query. The correlations among all term pairs are captured in the clustering process.

3. APPROACH

In this section, we describe two smoothing methods that are representative of corpus based smoothing techniques. We also formulate the proposed method, CVS.

3.1 Language Model

In this paper, we make the naive Bayes Assumption — a term's occurrence in a document is independent of any other term. We use the multinomial model — a document is represented by its terms and occurrence frequencies ([17]). Statistics of terms in a given document are used to decide whether the document is relevant to a query or whether it belongs to a certain cluster. Therefore, it is critical to calculate such statistics accurately. Smoothing can improve the accuracy on estimating such statistics.

Let $\mathbb{C}$ be a corpus of documents, d a document in the corpus, and $\mathbb{T}$ the set of terms selected to model documents in the corpus. Let $f(t|d)$ be the observed occurrence frequency of the term t in document d.

The conditional probability of having the term t in the document d, denoted as $p(t|d)$, is estimated by:

$$p(t|d) = \frac{f(t|d)}{\sum_{s \in d} f(s|d)}$$

The probability of having the term t in the corpus $\mathbb{C}$, denoted as $p(t|\mathbb{C})$, is given by:

$$p(t|\mathbb{C}) = \frac{\sum_{d \in \mathbb{C}} f(t|d)}{\sum_{x \in \mathbb{T}} \sum_{d \in \mathbb{C}} f(x|d)}$$

3.2 Prior Smoothing methods

We choose to study the Jelinek-Mercer method and the Dirichlet method ([26, 10]) because of their simplicity and yet being representative of various corpus based smoothing techniques.

The Jelinek-Mercer method adjusts the probability of a term in a document by a linear interpolation on the observed probabilities of the term in the document and in the corpus:

$$p'(t|d) = \beta\big((1-\lambda)p(t|d) + \lambda p(t|\mathbb{C})\big)$$

where λ is a smoothing parameter, β is a scaling factor to ensure that all probabilities sum to 1.[2]

[1] The average web page size is only 1-2KB.

[2] To ensure $\sum_{s \in d} p'(s|d) = 1$, we have $\beta = \frac{1}{\sum_{s \in d} p'(s|d)}$.

The Dirichlet method is a general formulation for the Laplace method. It adjusts the probability of a term in a document using the multinomial distribution to model a document:

$$p'(t|d) = \beta\big(\frac{f(t|d) + \lambda p(t|\mathbb{C})}{\sum_{t\in\mathbb{T}} f(t|d) + \lambda}\big)$$

where λ is a smoothing parameter, β is a scaling factor to ensure that all probabilities sum to 1.

3.3 CVS

Previous corpus based methods consider only the distributions of individual terms. We propose CVS, Correlation-Verification Based Smoothing, that considers the co-occurrence information as well as distributions of individual terms.

The probability of a term t in the document d can be adjusted by its correlation to terms observed in the document.

$$p'(t|d) = \beta\big(p(t|d) + \lambda \sum_{s\in d, IsCorrelate(s,t)} p(s|d)p(t|s,\mathbb{C})\big) \quad (1)$$

where λ is a smoothing parameter, β is a scaling factor to ensure that all probabilities sum to 1, the predicate $IsCorrelate(s,t)$ is true if the terms s,t are strongly correlated. The second component is the contribution by smoothing based on correlation information.

By Bayes rule, we can estimate $p(s,t|\mathbb{C})$ as:

$$p(t|s,\mathbb{C}) = \frac{p(s,t|\mathbb{C})}{p(s|\mathbb{C})} = \frac{\sum_{d\in\mathbb{C}} f(s,t|d)}{\sum_{y\in\mathbb{T}}\sum_{x\in\mathbb{T}}\sum_{d\in\mathbb{C}} f(x,y|d)} \frac{\sum_{x\in\mathbb{T}}\sum_{d\in\mathbb{C}} f(x|d)}{\sum_{d\in\mathbb{C}} f(s|d)}$$

where $f(s,t|d)$ is the co-occurrence frequency of the terms s, t in document d and is defined as $\min\{f(s|d), f(t|d)\}$.

To avoid *over-smoothing* (terms irrelevant to the theme of a document are assigned non-zero weights), a term pair is used to adjust a document vector only if its co-occurrence frequency in the entire corpus exceeds a threshold. While fixing the threshold to a predefined value by trial-and-error can efficiently sift some correlated terms, terms that are strongly correlated to the specific theme of a document but with relatively low co-occurrence frequencies can be missed. Therefore, we complement the approach with a statistical test of significance. Following Klas and Fuhr ([14]), we treat a corpus as a mega-document which is a concatenation of all documents. We use the same symbol $\mathbb{C}$ to represent the mega-document of the corpus $\mathbb{C}$. We employ χ^2 test to measure the statistical difference in joint distributions of terms s,t in the document d and $\mathbb{C}$ to identify relevant terms to the document with low frequencies. This provides a sound statistical framework to select a threshold for identifying strongly correlated term pairs in a particular document.

Let N_d and N_c be the total frequencies of all terms in the document d and $\mathbb{C}$ respectively. Only terms s,t with observed co-occurrence frequency exceeding their expected co-occurrence frequency (in the document d), $N_d\frac{f(s,t|\mathbb{C})}{N_c}$, are considered in the χ^2 test.

The expected co-occurrence frequency of the terms s,t in the document d can be computed as:

$$E(s,t|d) = \frac{N_d(f(s,t|d) + f(s,t|\mathbb{C}))}{N_d + N_c}$$

Similarly, the expected co-occurrence frequency of the terms s,t in the mega-document $\mathbb{C}$ is:

$$E(s,t|\mathbb{C}) = \frac{N_c(f(s,t|d) + f(s,t|\mathbb{C}))}{N_d + N_c}$$

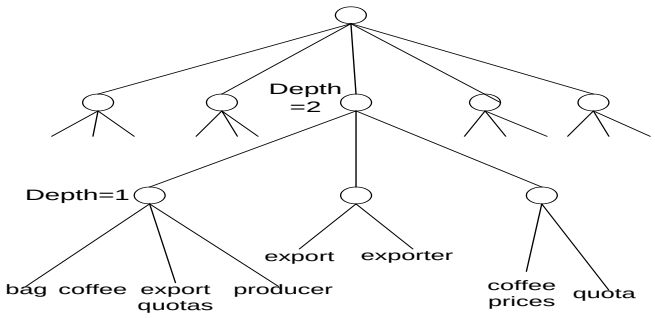

Figure 1: A cluster hierarchy on Reuters corpus generated by term clustering. An example of clusters of depth 1 is {export, exporter}.

Because text documents are dichotomous, while the chi-square values are based on a continuous distribution, Yate's continuity correction (subtracting 0.5 from the observed and expected frequencies before squaring) is applied. Yate's correction may be too conservative in lowering the chi-square values. However, since data with lower confidence could introduce noise, we believe that it is better to be conservative.

$$\chi^2 = \frac{(f(s,t|d) - E(s,t|d) - 0.5)^2}{E(s,t|d)} + \frac{(f(s,t|\mathbb{C}) - E(s,t|\mathbb{C}) - 0.5)^2}{E(s,t|\mathbb{C})} + \frac{(f(x,y|d) - E(x,y|d) - 0.5)^2}{E(x,y|d)} + \frac{(f(x,y|\mathbb{C}) - E(x,y|\mathbb{C}) - 0.5)^2}{E(x,y|\mathbb{C})}$$

where $x,y \neq s,t$. Define the null hypothesis as the two distributions being the same. Given a desired confidence level, the critical value can be looked up from the critical value table of χ^2 test with degree of freedom equal to 1. A term pair with χ^2 value higher than the critical value is regarded as strongly correlated in the document.

4. APPLYING SMOOTHING TO QUERY EXPANSION BY TERM CLUSTERING

In this section, we describe how we can apply corpus based smoothing techniques to improve the effectiveness of information retrieval by query expansion and term clustering.

A similarity measure is first used to measure the correlation between a pair of terms. Typical statistical similarity measures can be used, one of which is the mutual information:

$$sim(s,t) = \sum_{d\in\mathbb{C}} p(s,t|d)\log\frac{p(s,t|d)}{p(s|d)p(t|d)}$$

where $sim(s,t)$ is the similarity measure between the terms s,t. A term clustering application takes inputs the similarity measures between two terms and generates a hierarchy of clusters in which similar terms are grouped into the same cluster.

The methodology of the experiment is as follows: (1) Top k nouns / noun phrases are extracted from documents in the Reuters collection ([21]) (2) Different smoothing techniques are used to refine vector representations of documents. (3) Top k terms are clustered into a hierarchy using an agglomerative clustering method. (4) A pre-defined topic in the

Reuters collection is mapped to a cluster of depth *Depth* in the cluster hierarchy if the cluster consists of terms of the topic description. A topic may correspond to more than one cluster. (5)A query for a topic consists of terms in the topic description and terms in the clusters to which the topics are mapped. (6) Relevant documents for each topic query are retrieved. (7) Different smoothing techniques are compared using the corresponding best smoothing parameters which are manually fine tuned.

5. EXPERIMENT RESULTS

Top 2000 terms are selected for clustering. Relaxing the condition to select the top 6000 terms for clustering does not significantly increase the number of terms that can be clustered. Many of the 135 topics enlisted are not suitable for evaluation (e.g., the 27 currency codes) or do not have enough sample documents (e.g., the 78 commodity codes). Around 35 topics can be mapped to clusters in cluster hierarchies. Topics with less than 30 documents are eliminated. For fair comparison, only topics that are common to all evaluation cases are used. This results in 13, 15, 15 topics used for evaluation for clusters at depth 1, 2, 3, respectively. Nouns / noun phrases can be extracted from 7858 out of the 9603 documents.

The results are shown in Figures 2, 3, 4 and 5. The overall improvements in average precision for the Dirichlet method, the Jelinek-Mercer method and the CVS method are 10.8%, 6.8% and 14.6% respectively. The results demonstrate the use of corpus based smoothing techniques in improving retrieval effectiveness. The results can also be viewed in light of the improvements of clustering by smoothing.

Since clustering is unsupervised learning, one should expect better performance using supervised learning methods. If training data are not always available, clustering can be used for query expansion.

Depth	NoSmooth	Dirichlet	Jelinek-Mercer	CVS
1	0.67	0.73	0.72	0.74
2	0.47	0.50	0.51	0.60
3	0.45	0.54	0.48	0.49
Avg	0.53	0.59	0.57	0.61
% Change		10.8%	6.8%	14.6%

Figure 2: Effects of smoothing on retrieval effectiveness by query expansion and clustering. This table shows average precision with query expanded by clusters of various depths in cluster hierarchies refined by different smoothing methods. The CVS **method outperforms the Dirichlet method and the Jelinek-Mercer method.**

Precision degrades as one uses clusters of depth greater than 1 to expand a query. The average precision for clusters of depth 2 and 3 are more similar than those for clusters of depth 1. A cluster contains approximately 4 children clusters in the cluster hierarchies. This observation implies that 4 terms out of 2000 terms are appropriate to augment the description for a topic.

There is a benefit of expanding a query with deeper clusters that are not reflected in the results: While most topics cannot reach 100% recall level, expanding terms with clusters of depth 2 and 3 can give 100% recall level for most topics.

Smoothing improves the retrieval performance of base clusters (clusters of depth 1) by addressing the data sparsity problem. The similarity measure is refined, which results in more overlapping between distributions of terms. The improvement in base clusters is propagated to clusters higher up in the hierarchy.

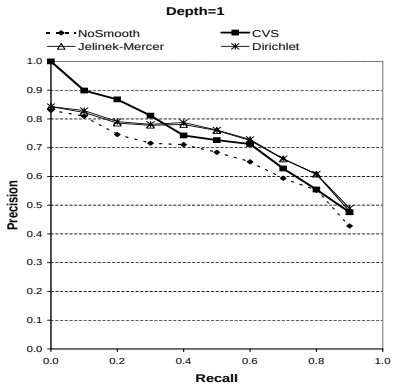

Figure 3: Recall-Precision curve with queries expanded with clusters of depth 1.

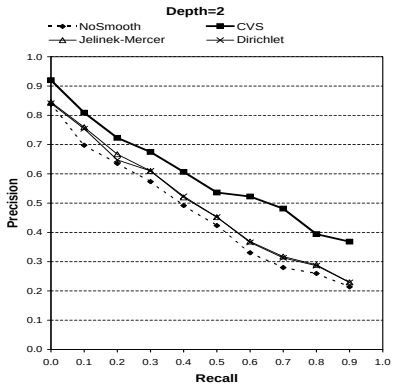

Figure 4: Recall-Precision curve with queries expanded with clusters of depth 2.

Another interesting observation relates to the structure of the cluster hierarchy. Without smoothing, only 826 terms out of 2000 terms are successfully clustered. Only CVS significantly improves the number of terms clustered. The smoothing parameter λ is fine tuned to optimize retrieval effectiveness. There is a potential risk of over-smoothing. We conjecture that one can be more liberal in setting the smoothing parameter for a better smoothing method. This significant improvement in CVS further confirms the superiority of CVS over the other methods.

6. CVS VARIANTS

In this section, we discuss the shortcomings and variants of CVS. Fast CVS improves the speed whereas Voting CVS and Iterative CVS concern the quality of smoothing.

6.1 Fast CVS

CVS requires $O(\mid \mathbb{T} \mid^2)$ memory overhead for storing the co-occurrence information between term pairs as compared with $O(\mid \mathbb{T} \mid)$ required by the Dirichlet method and the

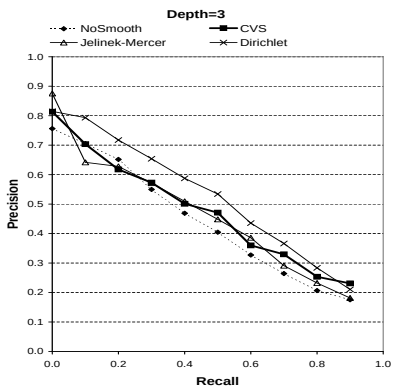

Figure 5: Recall-Precision curve with queries expanded with clusters of depth 3.

	Num. of clusters	Num. of terms
NoSmooth	361	826
Dirichlet	375	823
Jelinek-Mercer	372	841
CVS	442	1001

Figure 6: CVS **improves the number of terms clustered significantly by addressing the data sparsity problem, which results in better similarity measure.**

Jelinek-Mercer method. For 2000 terms, 2000 * 2000 * 4 bytes / 2 = 8MB is required to store the co-occurrence information which is reasonable given that most computer systems are equipped with 256MB main memory.

All methods scan the corpus twice, once for gathering corpus statistics, and once for smoothing each document. CVS is more computationally expensive than the other two methods in smoothing a document because the weight of a term in the corpus is adjusted by co-occurrence information with all terms in the document.

Typically smoothing can be applied once to all documents as a back-end process and hence time is not a critical issue. But if there is a hard constraint on time, CVS can be refined to trade in effectiveness for efficiency. The joint probability between a term pair computed from the original distributions in the corpus can be directly modified by smoothing. This in turn adjusts the similarity matrix (e.g., mutual information) used for clustering. The joint probability of the terms s, t can be adjusted by:

$$p'(s,t|\mathbb{C}) = \beta\big(p(s,t|\mathbb{C}) + \lambda\big(\sum_{t\in\mathbb{T}, IsCorrelate(s,t)} p(t|\mathbb{C})p(s|t,\mathbb{C}) + \sum_{s\in\mathbb{T}, IsCorrelate(s,t)} p(s|\mathbb{C})p(t|s,\mathbb{C})\big)\big)$$

6.2 Voting CVS

In a smoothing algorithm that checks the co-occurrence frequencies for strong correlation, a term t not in a document is assigned a weight which reflects its correlation with all terms in the document observed over the entire corpus. This may not be desirable because a term in the document is likely to be related to terms of different themes. For example, consider smoothing a document on car insurance. "insurance" may be strongly correlated with "pre-existing condition" in documents on medical insurance in the corpus. Smoothing may then add "pre-existing condition" to the document incorrectly, which introduces "noise". One solution is to enforce voting. Only when enough terms in the document are strongly correlated with a term do we add this term to the document.

6.3 Iterative CVS

We conjecture that if the smoothed probabilities are input to formulae (1), we may be able to get even more accurate estimators. Such iterations can continue until an optimal stage is achieved. Simulated annealing, genetic algorithm and other optimization approaches can help refine the iterations so that (1) can converge faster and local optimality can be avoided. We call this approach Iterative CVS. We evaluated Iterative CVS in some preliminary experiment with the number of iterations ranging from 1 to 100. Except in a few cases, Iterative CVS almost always under-performed CVS. We conjecture that some promoted related terms may be irrelevant to a document. Iterative CVS carries and amplifies this error from one iteration to the next. One solution is to limit the number of related terms as well as their weights. Further research is necessary to conclude whether iterations and optimization algorithms can be combined with smoothing.

7. CONCLUSION

We introduced CVS, a corpus based smoothing technique that considers co-occurrence information of strongly correlated terms. We illustrated a sound statistical framework to determine whether terms with low co-occurrence frequencies are specific to the theme of a document. We demonstrated how corpus based smoothing techniques can be applied to information retrieval by query expansion and term clustering. We also discussed the shortcomings of CVS and how they can be addressed by some variants. Empirical studies showed that corpus based smoothing can improve retrieval effectiveness and the quality of hierarchical clustering. CVS improves retrieval effectiveness by 14.6% which outperforms previous corpus based smoothing techniques.

There are several questions that remain to be answered. Further research is needed to differentiate noise from correlated terms with low co-occurrence frequency. The effects of smoothing on terms other than noun / noun phrases remain open. Intuition suggests that verbs could be good candidates to link related nouns together. The correlation information used in the smoothing method proposed is based purely on corpus statistics. The framework, however, is general enough to capture correlation information from domain knowledge such as relevance feedback and lexical references. Given the computation overhead of smoothing, it is important to address the issue of handling data streams effectively.

8. ACKNOWLEDGMENTS

The authors would like to thank Jianchang Mao for his support of this study and for providing valuable feedback on an early draft of the manuscript.

9. REFERENCES

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[2] C. Carpineto, R. de Mori, and G. Romano. Information term selection for automatic query expansion. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 308–314. National Institute of Standards and Technology (NIST), 1998. http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

[3] R. Fowler, W. Fowler, and B. Wilson. Integrating query, thesaurus, and documents through a common visual representation. In *International Conference on Research and Development in Information Retrieval (SIGIR 1991)*, pages 142–151, 1991.

[4] M. Franz and S. Roukos. A method for scoring correlated features in query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*, pages 337–338. ACM, August 24–28 1998.

[5] S. Gauch and J. Wang. A corpus analysis approach for automatic query expansion. In *Proceedings of the Sixth International Conference on Information and Knowledge Management (CIKM'97)*, pages 278–284, Las Vegas, Nevada, November 10–14 1997. ACM.

[6] I. J. Good. The population frequencies of species and the estimation of population parameters. In *Biometrika*, number 40 in 3,4, pages 237–264, 1953.

[7] K. Hoashi, K. Matsumoto, N. Inoue, and K. Hashimoto. Trec-7 experiments: Query expansion method based on word contribution. In *The Seventh Text REtrieval Conference (TREC-7)*, pages 373–381. National Institute of Standards and Technology (NIST), 1998. http://trec.nist.gov/pubs/trec7/t7_proceedings.html.

[8] K. Hofland and S. Johansson. Word frequencies in british and american english. In *The Norwegian Computing Center for the Humanities*, pages 43–53, Norway, 1982.

[9] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. WebSOM - self-organizing maps of document collections. In *Proceedings of Workshop on Self-Organizing Maps (WSOM97)*, pages 310–315, Espoo. Finland, 1997.

[10] F. Jelinek and R. Mercer. Interpolated estimation of markov source parameters from sparse data. In *Pattern Recogition in Practice*, pages 381–402, North Holland, Amsterdam, 1980.

[11] A. Kilgarriff. Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved lob-brown comparison. In *ALLC-ACH Conference*, 1996. http://www.hit.uib.no/allc/kilgarny.pdf.

[12] A. Kilgarriff. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Proceedings of 5th ACL workshop on very large corpora*, Beijing and Hongkong, August 1997.

[13] A. Kilgarriff and T. Rose. Measures for corpus similarity and homogeneity. In *Proceedings of 3rd conference on empirical methods in natural language processing*, pages 46–52, 1998.

[14] C. P. Klas and N. Fuhr. A new effective approach for categorizing web documents. In *Proceedings of the 22th BCS-IRSG Colloquium on IR Research*, 2000.

[15] D. Lawrie, W. B. Croft, and A. Rosenberg. Finding topic words for hierarchical summarization. In *International Conference on Research and Development in Information Retrieval (SIGIR 2001)*, pages 349–357, 2001.

[16] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pages 191–197, Berkeley, CA, USA, August 15–19 1999. ACM.

[17] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on Learning for Text Categorization*, pages 41–48, Madison, WI, 1998.

[18] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR1998)*, pages 206–214, Melbourne, Austrailia, August 24–28 1998.

[19] A. Rauber. LabelSOM: On the labeling of self-organizing maps. http://www.ifs.tuwien.ac.at/ andi, July 10–16 1999.

[20] P. Rayson and R. Garside. Comparing corpora using frequency profiling. In *proceedings of the workshop on Comparing Corpora*, pages 1–6, 2000.

[21] Reuters Research and Standards Group. Retuers corpus. http://about.reuters.com/researchandstandards/corpus/.

[22] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *International Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pages 206–213, 1999.

[23] A. E. Smith. Machine mapping of document collections: the leximancer. In *Proceedings of the 5th Australasian Document Computing Symposium*, Sunshine Coast, Australia, December 1 2000.

[24] E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. (SIGIR'94)*, pages 61–69, Dublin, Ireland, July 3–6 1994. ACM/Springer.

[25] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'96)*, pages 4–11, August 18–22 1996.

[26] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc iformation retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2001)*, pages 334–342, New Orleans, Louisiana, USA, September 9–13 2001.