Modeling Term Associations for Ad-hoc Retrieval Performance within Language Modeling Framework

Xing Wei and W. Bruce Croft

Center for Intelligent Information Retrieval
University of Massachusetts Amherst
140 Governors Drive
Amherst, MA 01003
{xwei,croft}@cs.umass.edu

Abstract. Previous research has shown that using term associations could improve the effectiveness of information retrieval (IR) systems. However, most of the existing approaches focus on query reformulation. Document reformulation has just begun to be studied recently. In this paper, we study how to utilize term association measures to do document modeling, and what types of measures are effective in document language models. We propose a probabilistic term association measure, compare it to some traditional methods, such as the similarity co-efficient and window-based methods, in the language modeling (LM) framework, and show that significant improvements over query likelihood (QL) retrieval can be obtained. We also compare the method with state-of-the-art document modeling techniques based on latent mixture models.

Keywords: Information Retrieval, Language Model, Term/Word Associations/Relationships, Term/Word similarity, Document Model, Topic Model.

1 Introduction

Modeling term associations is important to Information Retrieval (IR) systems. It is well-known that ranking algorithms solely based on matching the literal words that are present in queries and documents will fail to retrieve much relevant information. For example, matching only the word “fruit” will miss the documents containing “apple” that are also relevant to “fruit”. For this reason, term associations, which are also called “term relationships” or “word similarity” in literature, have been introduced to add new terms to the query/document representations that are related to the original terms. There can be associations between two single terms (term-term association); or between two groups of terms (term group association).

There has been much research in IR to associate related terms for queries and/or documents. Manual techniques such as using hand-crafted thesauri and automatic techniques such as clustering all attempt to provide a solution, with varying degrees of success. Although manual processing can usually provide precise and useful information with relatively less noise, an automatic method is expected to be more effective due to many problems related with manual processing [15], such as labor intensiveness, inconsistencies and ambiguity. Most automatic approaches to modeling term associations are based on term co-occurrence or grammatical analysis. Grammatical analysis is provides very specific knowledge about term relationships, but it is not as robust as using term co-occurrence [12]. Accurate but limited knowledge that provides few related terms is unlikely to substantially improve the retrieval output. Term co-occurrence has been widely used in term-association studies based on the intuition that co-occurring words are more likely to be similar, such as in term-term association models (e.g., measuring term similarity with co-efficient of two term-document vectors, which was widely used in earlier work such as term clustering [15, 23, 21] and Hyperspace Analogue to Language (HAL) [4]) and term group association models (e.g., measuring document similarity with co-efficient of two document-term vectors in document clustering [16] and Latent Dirichlet Allocation [2]). After term associations are constructed by these methods, some post-processing techniques can be applied to further improve the associations such as in [6], or to make the results compatible with systems using clustering such as in [15, 16].

With the term associations derived from previous methods, texts are reformulated (i.e. usually expanded) to improve the retrieval effectiveness. Some reformulations are not as explicit as replacing query terms with new terms, but instead the reformulation process is implicit, such as in the spreading activation techniques [22, 7, 8], in which the expansion is actually acquired during the process of following links between nodes that represent terms or documents. Both query and document reformulation processes have been investigated.

Query reformulation has been extensively studied with many term-association models in various IR frameworks [10, 21, 14, 26, 19] (In the works that phrases are considered, such as [14], we view a phrase as one term in this paper). The well-known pseudo-relevance feedback process, which expands the initial query vocabulary by adding terms contained in previously retrieved documents, is one of the best query expansion techniques in terms of retrieval performance [19]. Most relevance feedback models do term group association to find terms related to the entire query, which contains more information than individual words and thus can produce better results [21, 14]. Some query expansion techniques based on term-term associations such as [1] do post-processing to generate associations with the entire query. These query-based expansion processes have to be done online, in that they require an extra search for each query, which negatively affects query response time. Also, the efficiency of an IR system depends heavily on the number of terms of the query submitted to the system; query expansion therefore has its disadvantages in spite of the generally good retrieval results.

Document reformulation can be done offline without query inputs, thus being transparent to users and more efficient in terms of query response time. Offline processing, however, can be time-consuming and memory-expensive because it needs to process the associations of every term in every document of the entire collection, which is one of the reasons that document expansion was not popular until recent years. Two types of term associations have been applied to document reformulation: (i) Term group associations for document reformulation are usually based on documents. In the cluster-based document model [16], related documents are grouped and used to expand documents; in the LDA-based document model [24], documents are associated with related terms. Improvements have been obtained on several TREC collections with both of these two models, but they are both very expensive and difficult to apply to large collections, and parameter tuning for these models makes them even more expensive. (ii) Simple term-term association has significant advantages over term group association considering the offline efficiency of document reformulation. Cao et al. reformulate documents within the language modeling framework using term associations extracted both from a manually built thesaurus (WordNet) and from a co-occurrence based automatic technique, which considers term co-occurrence in a fixed-sized window. They achieve significant improvements over a baseline query likelihood system on some TREC collections [5], and obtain better results by further processing the original term-term associations with Markov chains [6]. The window-based approach, however, always requires an appropriate setting for the window size, and the improvements using only the automatic model are not as impressive.

Cao et al.'s work sheds light on the effectiveness of integrating term-term associations into the language modeling framework, which has been confirmed by a number of groups to be a theoretically attractive and potentially very effective probabilistic framework for studying information retrieval problems [20]. On the other hand, the assumption of the term independence ("bag of words") of the unigram language model is well known to be inappropriate for natural language. This has led many language model researchers to study term associations.

As a summary, we are interested in an automatic term-association method based on term co-occurrence in the language modeling framework, especially for dealing with document reformulation. Although term-association models have been studied for decades, none of the association processes have been performed within the language modeling framework, even that some integration processes of term associations are carried out with language models and some association processes like the window-based co-occurrence model are probabilistic methods. In this paper we study the traditional term co-occurrence based automatic term-association methods in the document reformulation task, and propose a new and simple method, which is based on the language modeling approach and thus fits within this framework naturally, to model term associations for retrieval operations.

2 Related Work

The history of examining term associations to improve retrieval effectiveness is almost as long as the history of IR itself. Since the binary term matching model, IR researchers have been trying to expand the matching of literal terms to include the matching of many other related words.

2.1 Hand-crafted thesauri

The earliest method of detecting and using term associations in IR was by building hand-crafted thesauri. This approach still attracts considerable interest from the IR community and open resources like WordNet and the open directory project[1] have been studied extensively [5, 6, 9].

Manual indexing has often been viewed as a gold standard and a thesaurus as a “correct” way of incorporating new words or phrases, but building and maintaining a thesaurus is very labor-intensive and it is very difficult to get people to agree on the semantic classifications involved. Inconsistencies and ambiguity in the use of these thesauri have produced poor results when they are used for retrieval experiments. Also, it is a fact that human beings tend to stick to obvious principles of classification. It is easy for human beings to group such words as “fruit” and “apple” together, but it may be difficult for them to find out that “boundary”, “layer” and “flow” are related by their combined use in aerodynamic contexts [15]. Therefore, an automatic, instead of a manual approach, is expected to be more effective for improving retrieval.

2.2 Similarity Coefficient

A variety of similarity coefficients have been developed and applied to measure term associations in IR environments, such as the Cosine metric, weighted and unweighted Tamimoto [15], etc. The coefficient used in Qiu & Frei’s Concept Based Query Expansion is one example [21]. They built a term-document matrix and computed the similarity between any two terms as follows;

$$SIM\ (t_i, t_j) = \sum_{k=1}^{n} d_{ik} \cdot d_{jk} \quad , \tag{1}$$

$$d_{ik} = \frac{(0.5 + 0.5 \times \frac{ff\,(d_k, t_i)}{\max\ ff\,(t_i)} iif\,(d_k))}{\sqrt{\sum_{j=1}^{n} ((0.5 + 0.5 \times \frac{ff\,(d_j, t_i)}{\max\ ff\,(t_i)} iif\,(d_j))^2}} \quad , \tag{2}$$

where $ff(d_k, t_i)$ is the frequency of term t_i in document d_k, $iff(d_k)=\log(m/|d_k|)$, m is the number of terms in the collections and $|d_k|$ is the number of different terms in document d_k. max $ff(t_i)$ is the maximum frequency of term t_i in all documents. The d_{ik}’s and d_{jk}‘s signify feature weights of the indexing features (documents). Then, the similarity between a term and a query is defined as the weighted sum of the similarity values between the term and individual terms in the query. To expand a query, terms with the highest similarity to the query are added and the weight of each added term takes its similarity value with the original query. Significant improvements in retrieval effectiveness were reported in their paper [21].

Although many techniques in this area have been tested and some interesting results were obtained, most of the techniques have been used to do query expansion. Few studies on document modeling with term similarity coefficients have been conducted.

[1] http://www.dmoz.com/

2.3 Co-occurrence in Windows

Another important group of term association measures estimates the conditional probability of a term given another term. Van Rijsbergen [23] and Cao et al. [5] compute the conditional probability using co-occurrence samples. To compute the conditional probability of two terms by their co-occurrence in a window is a practical method for both its simplicity and effectiveness. A fixed-sized window is applied to measure the co-occurrence in [5] and a sliding-window method (Hyperspace Analogue to Language, HAL) is described in [4]. A typical computation of the co-occurrence probability (the strength of term association) is as follows:

$$P(t_j \mid t_i) = f(t_i, t_j) / \sum_k f(t_i, t_k) \ , \qquad \textbf{(3)}$$

where $f(t_i, t_j)$ is the frequency of co-occurrences of t_i and t_j.

2.3.1 Fixed-sized window

A fixed-sized window is often used to measure the co-occurrence of two terms. In this window-based method, two words are considered as co-occurring once when the distance between them is less than the window size. For instance, Xu & Croft developed a metric used for query expansion based on the fixed-sized window method and achieved excellent performance [25, 26]; Cao et al. applied fixed windows in document modeling in combination with WordNet [5] and obtained significant improvements on two TREC collections.

2.3.2 Sliding window

In addition to setting a threshold to judge the co-occurrence of terms as in the fixed-sized window method, the distance between two words are also taken into account in some term-association models, such as in [4, 11, 17, 1]. Sliding window method is one of the examples, which is also called HAL Space (Hyperspace Analogue to Language) [4, 17]. By moving a window across the text, an accumulated co-occurrence matrix for all terms is produced. Compared to the fixed-sized window method, the sliding window method takes accumulated co-occurrence in all possible fixed-sized windows and in this way, the strength of association between two words is inversely proportional to their distance. Some interesting results with the sliding window method are obtained in previous works, including query expansion tasks in the language modeling framework [1, 4, 17]. However, its effectiveness on document modeling tasks is still unknown.

In both the fixed-sized window and the sliding window methods, the size of the window is a parameter that needs to be determined.

2.4 Latent mixture models

Because of the success of statistical approaches to representing text, IR has the potential of benefiting from recent advances in the fields of statistical modeling and machine learning. Research in these fields has led to new mathematical models that effectively represent documents through latent mixture modeling techniques. Some of these models have also been studied in IR research with interesting results, such as the mixture of unigrams model [18] and (probabilistic) Latent Semantic Indexing ((p)LSI) [12]. The Latent Dirichlet Allocation (LDA) model [2], which possesses fully generative semantics and overcomes the drawbacks of previous latent mixture models such as pLSI, has quickly become one of the most popular probabilistic text modeling techniques in machine learning. LDA has recently been shown to outperform both the unigram document model and the cluster-based document model in the language modeling framework for IR [24].

However, latent mixture models are usually very expensive and difficult to apply on large collections. There is often no exact inference techniques for these models and approximation techniques have to be adopted to iteratively approach the solution. Parameter tuning for these complicated models makes them even more expensive. Furthermore, they require a new training process for each new collection; in contrast, term-term associations can often be used across collections.

3. Modeling Term Associations by Joint Probability

3.1 Term-Association Models

Previous research described in Section 1 and Section 2 has shown the effectiveness of modeling and integrating term associations into information retrieval processes. Especially, constructing term-term associations and integrating them into document models is an attractive way considering both of its online efficiency and large-collection feasibility. Also, the recently developed language modeling framework has opened up new ways of thinking about retrieval problems. Its solid theoretical setting and promising experimental results provide and motivate new directions of the construction and integration process of term associations. In this section, we present an approach in the language modeling framework to estimating the conditional probability of terms by joint probability through Bayesian rule, and the joint probability will be computed by unigram document models.

To get a sense of the association or closeness between two terms, *w* and *t*, we consider *P*(*w*|*t*), which is the probability of observing *w* when *t* is given. By Bayesian rule, we have

$$P(w \mid t) = P(wt) / P(t) \ , \qquad \textbf{(4)}$$

To estimate the join probability of observing the word *w* and the term *t*, instead of counting co-occurrence samples in windows, we assume that *w* and *t* are identical and independent samples from a unigram document model *D*. Then the total probability of observing *w* together with *t* is:

$$P(wt) = \sum_{D \in \Pi} P(D) P(wt \mid D) = \sum_{D \in \Pi} P(D) P(w \mid D) P(t \mid D) \ , \qquad \textbf{(5)}$$

where $\prod$ represent some finite universe of unigram document models. We choose to use uniform priors $P(D)$ and limit the universe $\prod$ to the collection we test on. Then,

$$P(w \mid t) = \frac{\sum_{D} P(w \mid D) P(t \mid D)}{\sum_{w} \sum_{D} P(w \mid D) P(t \mid D)} . \tag{6}$$

Thus, for each term t, there is a list of words w with the probability $P(w|t)$ representing the association of w and t. We can view this probability as the association/closeness between w and t.

3.2 Document Language Models with Term Associations

The basic approach for using language models for IR is the query likelihood model where each document is scored by the likelihood of its model generating a query Q.

$$P(Q \mid D) = \prod_{q \in Q} P(q \mid D) \text{ ,} \tag{7}$$

where D is a document model, Q is the query and q is a query term in Q. $P(Q|D)$ is the likelihood of the document model generating the query terms under the 'bag-of-words' assumption that terms are independent given the documents. And $P(q|D)$ is specified by the document model with Dirichlet smoothing [27],

$$P_U(w \mid D) = \frac{N_d}{N_d + \mu} P_{ML}(w \mid D) + (1 - \frac{N_d}{N_d + \mu}) P_{ML}(w \mid coll) \text{ ,} \tag{8}$$

where $P_{ML}(w|D)$ is the maximum likelihood estimate of word w in the document D, and $P_{ML}(w|coll)$ is the maximum likelihood estimate of word w in the entire collection. N_d is document length. μ is the Dirichlet prior, and in our experiments we used a fixed value with μ=1000.

In the original query likelihood model, documents are estimated by the independence assumption, which is not appropriate to natural language that is much more complicated than simple "bags of words". Modeling term associations is a straightforward way to integrate related words into text models. To integrate the association information into document models, we first compute the word distribution in documents through the probabilistic association measure (Eqn (9)), and then combine it with the original term model by linear combination:

$$P_T(w \mid D) = \sum_{t \in D} P(w \mid t) P(t \mid D) \text{ .} \tag{9}$$

It is similar to the retrieval methodology using translation models proposed by Berger and Lafferty to incorporate term associations into document language models [3]. With the translation model, the document model becomes

$$P_{TR}(w \mid D) = \sum_{t} tr(w \mid t) P(t \mid D) \text{ ,} \tag{10}$$

where *tr*(*w*|*t*) is the translation model for mapping a document term *t* to an arbitrary term *w*. The translation probability *tr*(*w*|*t*) describes the degree of link between a term *w* and the document term *t*. If we set *tr*(*w*|*t*) to be *P*(*w*|*t*), then Eqn (9) and Eqn (10) will be same.

The linear combination method is widely used in integrating related words into document models, such as in [16, 5, 24]. The final document model would be

$$\begin{aligned} P(w \mid D) &= \lambda P_U(w \mid D) + (1-\lambda) P_T(w \mid D) \\ &= \lambda\left(\frac{N_d}{N_d+\mu} P_{ML}(w \mid D) + (1 - \frac{N_d}{N_d+\mu}) P_{ML}(w \mid coll)\right) \\ &\quad + (1-\lambda)\sum_{t \in D} P(w \mid t) P(t \mid D) \end{aligned} \cdot \qquad (11)$$

where λ is the integration co-efficient. This is the only parameter to our model, and is also one of the parameters to the other models we compare to in Section 4.

In this paper we try several association measures to model *P*(*w*|*t*) in Eqn (11), including the similarity co-efficient, the fixed-sized window method, the sliding window method, and the joint probability method we propose. In the similarity co-efficient method, we normalize its co-efficient to be consistent with the probabilistic application as following:

$$P(t_j \mid t_i) = SIM(t_i, t_j) / \sum_k SIM(t_i, t_k) \cdot \qquad (12)$$

4. Experiments and Results

4.1 Data

We conduct experiments on five data sets taken from TREC: the Associated Press Newswire (AP) 1988-90 with queries 51-150, Wall Street Journal (WSJ) 1987-92 with queries 51-100 and 151-200, Financial Times (FT) 1991-94 with queries 301-400, San Jose Mercury News (SJMN) 1991 with queries 51-150, and LA Times (LA) with queries 301-400. Queries are taken from the "title" field of TREC topics. Queries that have no relevant documents in the judged pool for a specific collection have been removed from the query set for that collection.

4.2 Parameters

There are several parameters that need to be decided in our experiments. For the retrieval experiments, the proportion of the term-association part in the linear combination must be specified (λ in (11)). For the similarity measure, the window sizes need to be determined. We use the AP collection as our training collection to estimate the parameters. The WSJ, FT, SJMN, and LA collections are used for testing whether the parameters optimized on AP can be used consistently on other collections. At the current stage of our work, the parameters are selected through exhaustive search or manually hill-climbing search. All parameter values are tuned based on mean average precision (MAP).

4.3 Experimental Results

In all experiments, both the queries and documents are stemmed, and stopwords are removed.

4.3.1 Other Term-Associating Methods

We test the effectiveness of some traditional term-term associating methods that we discussed in Section 2 in language document models, and present the retrieval results in Table 1.

Similarity co-efficient: With the parameter setting λ=0.8, which was obtained by training on the AP collection, we run experiments with the similarity co-efficient based document models (SCDM) on other collections. Some improvements, including significant improvements on one of the five collections, are achieved over query likelihood retrieval by integrating the similarity co-efficient into document models.

Fixed-sized window: With λ=0.7 and window size W=30, which were obtained by training on the AP collection, we run experiments with the fixed-sized window based document models (FWDM) on other collections. Significant improvements on two of the five collections are obtained over query likelihood retrieval.

Sliding window: Retrieval results of the document models based on the sliding window method, with λ=0.6 and W=50, are shown in Table 1. Significant improvements on two of the five collections over the query likelihood retrieval are achieved. Table 1 also shows that the sliding window performs better than the fixed-sized window, which was adopted in [5] and [6] as an automatic term associating method to be integrated into language document models.

Table 1. Comparison of query likelihood retrieval (QL) and retrieval with document models based on similarity coefficient (SCDM), fixed-sized window method (FWDM), or sliding window method (SWDM). The evaluation measure is average precision. %chg denotes the percentage change in average precision. Stars indicate statistically significant differences with a 95% confidence according to the Wilcoxon test.

Collection	QL	SCDM	%chg over QL	FWDM	%chg over QL	SWDM	%chg over QL	%chg over FWDM
AP	0.2161	0.232	+7.62*	0.2381	+10.15*	0.2375	+9.88*	-0.25
FT	0.2558	0.2652	+3.68	0.2640	+3.22	0.2690	+5.14	+1.86*
SJMN	0.1985	0.2068	+4.18	0.2118	+6.67*	0.2142	+7.86*	+1.12
LA	0.2290	0.2305	+0.62	0.2362	+3.12	0.2485	+8.48	+5.20*
WSJ	0.2908	0.2866	-1.44*	0.2827	-2.79	0.2905	-0.10	+2.76*

4.3.4 Term Associations by joint probability

We test document models based on the term-associating method by joint probability (JPDM) that we present, and show the retrieval results in Table 2. λ=0.6 for these experiments, and we process only the top 400 related terms of each term. On four of the five collections JPDM retrieval achieves significant improvements over query likelihood retrieval. On the WSJ collection, no improvements are achieved with λ=0.6, and then we especially tuned λ for it and obtained improvement with λ=0.2 as shown in the last line of Table 2.

In previous experiments, we build term associations for each collection respectively. To test the easy applicability of the term-associating method we present, we also run experiments with the term associations constructed only from the AP collection (JPDM-ap), or all of the five collections (JPDM-all). Results of JPDM-ap are presented in Table 2 and JPDM-all in Table 3.

JPDM-all achieves the best performance among JPDM, JPDM-all and JPDM-ap. This shows that more training data lead to higher performance, because more data can imply more knowledge about the term associations. At the same time, term associations trained only on the AP collection are also effective on other collections. So, the term associations built by joint probability do not have to be trained on the specific collection of experiments.

Table 5 shows the comparison of JPDM-all and LDA-based document models (LBDM) [24]. The LBDM achieves better performance than the term association model we propose. However, based on our experiments, the term association modeling is much faster than the LDA model estimation. Also, we have shown that it is very easy and effective to apply the term associations trained on other collections, which is impossible for the LDA model training.

Table 2. Comparison of query likelihood retrieval (QL) and retrievals with JPDM and JPDM-ap.

Collection	QL	JPDM	%chg over QL	JPDM-ap	%chg over QL	%chg over JPDM
AP	0.2161	0.2400	+11.03*	0.2400	+11.03*	0
FT	0.2558	0.2754	+7.66*	0.2636	+3.05	-4.28
SJMN	0.1985	0.2180	+9.80*	0.2139	+7.74*	-1.88
LA	0.2290	0.2516	+9.85*	0.2426	+5.91	-3.59
WSJ	0.2908	0.2870	-1.32	0.2884	-0.83	+0.49
WSJ (λ=0.2)	0.2908	0.2971	+2.15	N/A	N/A	N/A

Table 3. Comparison of query likelihood retrieval (QL) and retrievals with LBDM, JPDM, and JPDM-all.

Collection	QL	LBDM	JPDM-all	%chg over QL	%chg over JPDM	%chg over LBDM
AP	0.2161	0.2629	0.2422	+12.05*	+0.92*	-7.91*
FT	0.2558	0.2795	0.2842	+11.10	+3.20	+1.68
SJMN	0.1985	0.2279	0.2186	+10.10*	+0.27*	-4.06*
LA	0.2290	0.2563	0.2547	+11.21*	+1.24	-0.63
WSJ	0.2908	0.3244	0.2910	+0.07	+1.41*	-10.30*

5. Conclusions and Future Work

We have proposed a probabilistic term association model in the language modeling framework, which measures term associations through their joint probability, and a document retrieval model that integrates term associations into document models through linear combination. We did experiments and compared the model we proposed with other popular term-association methods on ad-hoc retrieval tasks.

The experimental results showed that modeling term associations through joint probability was effective in the language modeling framework. Document models that include term associations outperformed the query likelihood model, and term associations constructed by joint probability achieved better performance than other term-association models, such as window co-occurrence methods, in the language modeling framework. Comparing the two window co-occurrence methods, the sliding window method performs better than the fixed-sized window method on the retrieval tasks. We also showed that term associations trained on other collections were effective in our model, and more training data leads to better performance.

Although the retrieval with term-associating model did not obtain improvements over the LDA-based document models [24], the results are interesting and encouraging considering the cost of LDA training.

For future work, we plan to investigate whether several association measures can be combined in one document modeling. We will also combine the term-association based document models with latent mixture model based document models and test the effectiveness of this combination. In addition, studying post-processing with the probabilistic term associations obtained from this paper would also be interesting.

6. ACKNOWLEDGMENTS

This work was supported in part by the Center for Intelligent Information Retrieval and in part by NSF grant #CNS-0454018. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

REFERENCES

1. Bai, J., Song, D., Bruza, P., Nie, J.-Y. and Cao, G.: Query expansion using term relationships in language models for information retrieval. In Fourteenth International Conference on Information and Knowledge Management (CIKM 2005).
2. Blei, D. M., Ng, A. Y., and Jordan, M. J.: Latent Dirichlet allocation. In Journal of Machine Learning Research, 3, 993-1022 (2003).
3. Berger, A. and Lafferty, J.: Information retrieval as statistical translation. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 222-229, August 15-19, 1999, Berkeley, California, United States.
4. Burgess, C., Livesay, K., and Lund, K., Explorations in Context Space: Words, Sentences, Discourse. Discourse Processes, 25(2&3), 211-257 (1998).

5. Cao, G., Nie, J.-Y., and Bai, J.: Integrating word relationships into language models. In Proceedings of SIGIR 2005, 298-305.
6. Cao, G., Nie, J.-Y., and Bai, J.: Constructing Better Document and Query Models with Markov Chains. In Proceedings of the ACM 15th Conference on Information and Knowledge Management (CIKM), November 2006, Arlington, USA.
7. Croft, W.B., Lucia, T.J., Cringean, J., and Willett, P.: Retrieving Documents By Plausible Inference: An Experimental Study. Information Processing and Management, 25, 599-614 (1989).
8. Croft, W.B. and Thompson, R.: I3R : A New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science, 38(6), 389-404, (1987).
9. Croft, W.B. and Wei, X.: Context-Based Topic Models for Query Modification. CIIR Technical Report, IR-424 (2005).
10. Fang, H. and Zhai, C.: Semantic Term Matching in Axiomatic Approaches to Information Retrieval. In Proceedings ACM SIGIR 2006, 115-122.
11. Gao, J.F., Nie, J.-Y., Zhang, J., Xun, E., Zhou, M. and Huang, C.: Improving Query Translation for CLIR using Statistical Models. In Proceedings of the 24th ACM SIGIR Conference on Research and Development in IR, pp. 96-104 (2001).
12. Manning, C.D., Raghavan, P., and Schütze, H.: Introduction to Information Retrieval, Cambridge University Press (2007).
13. Hofmann, T.: Probabilistic latent semantic indexing. In Proceedings of SIGIR 1999, Berkeley, CA, USA.
14. Jing, Y. and Croft, W.B.: An Association Thesaurus for Information Retrieval, In Proceedings RIAO-94, 146-160 (1994).
15. Jones, K. S.: Automatic Keyword Classification for Information Retrieval. London: Butterworths (1971).
16. Liu, X. and Croft, W.B. Liu, X., and Croft, W. B.: Cluster-based retrieval using language models, in Proceedings of SIGIR 2004, 186-193.
17. Lund, K. and Burgess, C.: Producing High-dimensional Semantic Spaces from Lexical Co-occurrence. Behavior Research Methods, Instruments,& Computers, 28(2), 203-208 (1996).
18. McCallum, A.: Multi-label text classification with a mixture model trained by EM. In AAAI workshop on Text Learning (1999).
19. Lavrenko, V. and Croft, W.B.: Relevance-based language models. In Research and Development in Information Retrieval, 120-127 (2001).
20. Ponte, J. and Croft, W.B. : A language modeling approach to information retrieval. In Proceedings of ACM SIGIR 1998 275-281.
21. Qui, Y. and Frei, H., Concept based query expansion, In Proceedings of ACM SIGIR 1993, 160-169.
22. Salton G. and Buckley, C.: On the Use of Spreading Activation Methods in Automatic Information Retrieval. In Technical Report 88-907, Department of Computer Science, Cornell University.
23. Van Rijsbergen, C. J.: Automatic Classification. In: Information Retrieval. 2nd edn. Chapter 3. London: Butterworths. (1979). http://citeseer.ist.psu.edu/vanrijsbergen79 information.html
24. Wei, X. and Croft, W.B. LDA-based Document Models for Ad-hoc Retrieval. In Proceedings of SIGIR 2006, 178-185.
25. Xu, J.: Solving the Word Mismatch Problem Through Automatic Text Analysis. Ph.D. Dissertation. Department of Computer Science, University of Massachusetts (1997).
26. Xu, J. and Croft, W.B.: Query expansion using local and global document analysis. In Proceedings of the 1996 ACM SIGIR Conference on Research and Development in Information Retrieval.
27. Zhai, C. and Lafferty, J.: A study of smoothing methods for language models applied to ad hoc information retrieval. In Proceedings of ACM SIGIR 2001, 334-342.