An Effective Approach to Verbose Queries Using a Limited Dependencies Language Model

No Author Given

No Institute Given

Abstract. Intuitively, any 'bag of words' approach in IR should benefit from taking term dependencies into account. Unfortunately, for years the results of exploiting such dependencies have been mixed or inconclusive. To improve the situation, this paper shows how the natural language properties of the target documents can be used to transform and enrich the term dependencies to more useful statistics. This is done in three steps. The term co-occurrence statistics of queries and documents are each represented by a Markov chain. The paper proves that such a chain is ergodic, and therefore its asymptotic behavior is unique, stationary, and independent of the initial state. Next, the stationary distribution is taken to model queries and documents, rather than their initial distributions. Finally, ranking is achieved following the customary language modeling paradigm. The main contribution of this paper is to argue why the asymptotic behavior of the document model is a better representation then just the document's initial distribution. A secondary contribution is to investigate the practical application of this representation in case the queries become increasingly verbose. In the experiments (based on Lemur's search engine substrate) the default query model was replaced by the stable distribution of the query. Just modeling the query this way already resulted in significant improvements over a standard language model baseline. The results were on a par or better than more sophisticated algorithms that use fine-tuned parameters or extensive training. Moreover, the more verbose the query, the more effective the approach seems to become.

1 Introduction

Imagine (or perhaps recall) that you just came back from a well-deserved vacation in the South Pacific. When someone asks you about your vacation, you are happy to recount how it was. First you tell it to the people at home, then to your neighbors, then to your colleagues at work. At first there will be much variation in your story, but by and by all has been said, and the rendition of your experience becomes stable, only mentioning the essential parts. Or think of an event that lands as late breaking news on your paper's front page. As days go by, the story may reappear a few times, but eventually all has been said.

Now suppose a search engine would need to return the most relevant (as opposed to the most entertaining) story about your vacation. Should it be one from the earlier stages where it still meandered haphazardly along all that happened? Or one of the later more concise and orderly accounts?

Let us look at this phenomenon from the language modeling perspective to IR [11]. In this paradigm a text is viewed as a sample from a stochastic source that produces words according to some distribution. With the vacation story, you were the source, and your stories were different samples from that source. As the source is assumed to be stochastic, the words and their frequencies will change from one account to the next, as in the case of your stories.

Without a model of the underlying process, however, it would be difficult to reconstruct the distribution of the source from the samples alone. Therefore, language models can be distinguished by how they model the source and by how the distribution is derived from the samples. As current language models don't use an explicit representation of the meaning of documents, we can illustrate our approach with a simple abstract example. Assume a language of just the words a and b, and two documents $D_1 = [a\ a\ a\ a\ a\ b\ b\ b\ b\ b\ a]$ and $D_2 = [a\ b\ a\ b\ a\ b\ a\ b\ a\ b\ a\ b]$. Using $Q = [a\ b\ a\ b\ ]$ as the query (or topic), which document would be considered the most relevant for a given language model? In the multi-bernoulli model [11], D_1 and D_2 would get the same score, as all words in the query are also in the documents. The multinomial unigram model [14] also assigns the same score because the frequencies of a and b are the same in D_1 and D_2 and hence the $p(Q|D) = \prod_i p(q_i|D)$ are the same. If Q were extended with a word c that does not appear in the documents, so that smoothing [16] was called for, words would be discounted by the same amount, and again the documents would receive the same score. Basically, we are trying to estimate a relevance model (1) without further knowledge about the corpus, (2) under the assumption that the term occurrences are independent, and (3) in the absence of training data. These issues have received much attention lately. For example, several researchers have studied bigrams and trigrams [14] or even studied the optimal distance over which to consider dependencies in general [13, 10] or based on natural language constraints [6]. Metzler and Croft [10] in particular distinguished among full independence, sequential dependence, and full dependence. The terms mean what they suggest: in sequential dependence the ranking of a document depends only on the dependency of adjacent words, whereas in full dependence any clique of words is to be considered. In this paper we consider a fourth option, halfway between sequential and full dependence, namely when a word comes after another, but separated by words in between. For example, in D_1 and D_2 above, one can accumulate the distances from every a to every b to derive a probability that a is followed by b. In the example, this probability is much lower for D_1 than for D_2. Imagine that, as in the vacation story that was told over and over again, the sources of D_1 and D_2 would go on for a long time producing one new document after another according to their distributions. If we assume for concreteness a dependency of no more than five words, then (as we will see) in the long run a would appear about as often as b for D_2 but twice as often for D_1. This is obviously different from the word counts that would suggest a 50% probability for each. Moreover, the distribution in the long run seems to reflect the impression that D_2 is more like Q than is D_1. This paper will show how the term dependencies of a particular document predict the asymptotic behavior of its source, and with it the term distribution that would be observed if the source would continue to produce new documents.

The sections that follow show how the approach of asymptotic behavior relates to other language models, and how it accomplishes the following objectives:

- It shows that under very realistic, plausible, and elementary conditions the *source underlying a document is ergodic*, and therefore a stationary distribution to represent the source can be derived from just one document,
- It shows how documents can be ranked based on their underlying stationary distributions,
- It shows how an initial (ad hoc) distribution for a document can be established, based on a semantic approach called the *Hyperspace Analog to Language* (HAL).

2 The document source as an ergodic chain

One reason that language models use lower order dependencies is the (in)tractability of the Bayesian chain rule. Another is often simply a lack of knowledge about higher order dependencies. Yet, in practice, bigrams already give a reasonable improvement over unigrams [7]. In addition, [14] and others have shown that an interpolation of unigram and bigram models performs well.

The practical considerations aside, the question remains whether higher order dependencies would lead to better models, even if it is tempting to assume the affirmative. To begin answering the question, it is important to realize that the current approach to language modeling is applicable to any stochastic source and the languages they produce (human, machine, or perhaps of unknown origin). The models pay no heed to the fact that the documents to be modeled are produced by humans. Yet this throws out particular constraints that could make the methods more tractable. Some constraints can be borrowed from cognitive science, some follow directly from confining the languages under consideration to natural language:

- Many cognitive phenomena can be understood sufficiently well in terms of word-pairs. Pertinent examples can be found e.g. in the research on memory [13], work as mentioned above on the ‘semantic space’ [3], and results from old theories on ’spreading activation’ [1] to recent brain (ERP) studies [5]. This supports the view that the source underlying the document can be modeled as a (first order) Markov process.
- Words in a natural language corpus can be separated by any number of intermediate words. (Think of adding an extra adjective before a noun.) This means there cannot be any cycles in the process. Identifying words with the states of the process then means that the Markov chain is *aperiodic.*
- You can always get from one word to another by continuing to produce text (words can never be used up). Consequently, the Markov chain is *irreducible.*

The first point was already proposed by Shannon in his famous article [12], without the backup from cognitive science. The next two points, that the Markov process is both aperiodic and irreducible means that it is *ergodic*. An ergodic chain has the property that in the long run it reaches a stationary distribution (also called stationary kernel, or steady state), irrespective of the initial state.
It is easy to sample a document and generate a new one on the basis of its distribution; see the examples in [12], or any of the many sites on the web that offer programs to do this[1]. What we would like to compute however is the distribution of the source underlying the document. Or in the metaphor of the introduction, we would like to model the final stable and concise story as the most relevant to the query about the vacation. With little knowledge of the source, one could use a Gibbs sampler, i.e. generate a long series of documents and sample until the distribution seems to converge. The Gibbs sampler was proposed for example by Wei and Croft [15] to find a distribution for their LDA model. Besides the benefits of the Gibbs sample, there are several issues to overcome: (1) it is computationally demanding, (2) it is hard to know when the process has converged, and (3) The fixed point may not be unique and e.g. depend on the initial state. The observation above that the process we advance here is ergodic obviates all three issues at once. The final distribution of the Markov chain can easily be computed without sampling (it is the eigenvector with eigenvalue 1), and it is guaranteed to be unique.

Note, first, that the properties mentioned to derive this result are valid for natural languages in general. This means that the method may be used for languages other than English (and which are increasingly visible on the Web). Second, it also answers the question about the higher order dependencies, in that it is unlikely that these will contribute much to improving search results. With the answer comes an other question to the fore: how to compute the lower order dependencies given the documents. The next section offers a proposal, one we will use in an experiment further on, but it is by no means meant as the last word on finding initial distributions.

3 Deriving the initial distribution

In language modeling, the document source represents the author producing the document. As an author could produce different renderings of the same story, these renderings would be different samples of the source, and so the term distribution could differ from one document to the next.

Box 1

Given an n-word vocabulary, the HAL space is represented as a $n * n$ matrix constructed by moving a window of size w over the corpus ignoring punctuation,

[1] For example http://www.nightgarden.com/infosci.htm explains the procedure and links to a 'Shannonizer' where you can input text, or refer to a URL, to generate a text based on bi-grams

sentence, and paragraph boundaries. The strength of co-occurence decreases with the number of intervening words. Instead of an large-scale corpus, let us take just the sentence *The effects of spreading pollution on the population of Atlantic salmon.*

	the	effects	of	spreading	pollution	on	population	atlantic	salmon
the		1	2	3	4	5			
effects	5								
of	8	5		1	2	3	5		
spreading	3	4	5						
pollution	2	3	4	5					
on	1	2	3	4	5				
population	5		1	2	3	4			
atlantic	3		5		1	2	4		
salmon	2		4			1	3	5	

The table above shows the HAL matrix for a window size of 5. Take e.g. the entry for ‘population’. To find the distance to ‘pollution’, go backward starting at ’population’ with strength 5 (for ‘the’) counting down to 3 for ‘pollution’.

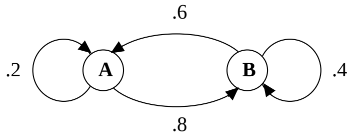
Fortunately, the ergodic chain has a property that is very useful here, namely that its asymptotic behavior is independent of the initial state. In other words, if one would continue to sample the source, then in the long run it would not matter what sample, i.e. what document, was observed first; the asymptotic behavior would be the same. What remains then, is to derive an initial distribution given the document. This is where language models differ greatly from one another. As we mentioned in the introduction, an important distinction lies in the degree of term dependency that is assumed. In this paper we follow the approach of Lund and Burgess [9] who computed co-occurrence statistics from a rich source of spontaneous conversations: Usenet newsgroups. They called the representation of these statistics the ‘Hyperspace Analog to Language’ or HAL. HAL is computed by sliding a window over the corpus and assigning weights to word pairs, inversely to the distance from each word to every other in the window. This results in a word by word matrix with the accumulated word distances in the cells. **Box 1** may clarify the construction further. (Note that the number in a cell is formally not a distance because the matrix is usually not symmetric.) If a word is connected to a second word via a small number, than it is more likely followed by that word than if the number had been high (e.g. the table shows that ’of’ is more likely to be followed by ’the’ than the other way around). Based on this observation, the HAL matrix is transformed into a transition probability matrix *pHAL* by normalizing the row vectors.

Box 2

For readers unfamiliar with the Markov approach, the essential steps in the algorithm are illustrated below. Assume a language of just the words a and b, whose dependencies are defined by the transition probabilities in matrix H. H defines a Markov chain, where state **A** ouputs a and state **B** outputs b.

H	a	b
a	.2	.6
b	.8	.4

$\Rightarrow$



For initial state s_0 (e.g. **A** if started with word a), the next state is given by $s_1 = s_0 * H$, where

$$H = \begin{pmatrix} .2 & .6 \\ .8 & .4 \end{pmatrix}$$

followed by $s_2 = s_1 * H = s_0 * H^2, ..., s_n = s_0 * H^n$ with $H^n = \frac{1}{.8+.6}\begin{pmatrix} .6 & .6 \\ .8 & .8 \end{pmatrix} + \frac{-0.4^n}{.8+.6}\begin{pmatrix} .8 & -.6 \\ -.8 & .6 \end{pmatrix}$ which converges to: $\lim_{n\to\infty} H^n = \begin{pmatrix} .4286 & .4286 \\ .5714 & .5714 \end{pmatrix}$ so the Markov chain becomes stationary with $P(a) = .4286$ and $P(b) = .5714$, independent of the initial state. (The formal derivation was only given to show the convergence. The stationary distribution can also be computed directly from the transition matrix.) In the same way these values can be obtained for the examples in the introduction. Computing the HAL matrix with window of size 4, the distributions converge to:
$D_1 = [a\ a\ a\ a\ a\ b\ b\ b\ b\ b\ b\ a]$, $P(a) = .36$ and $P(b) = .64$
$D_2 = [a\ b\ a\ b\ a\ b\ a\ b\ a\ b\ a\ b]$, $P(a) = .49$ and $P(b) = .51$
$Q = [a\ b\ a\ b\ ]$, $P(a) = .44$ and $P(b) = .56$
Computing the Kullback-Leibler divergence yields
$KL(Q||D_1) = .017$, and $KL(Q||D_2) = .007$, so D_1 diverges more from Q than D_2, and therefore D_2 is ranked as more relevant.

So, to find the document source distribution for a document requires only two steps:

1. Compute the ad-hoc distribution, in our case pHAL,
2. Compute the stable distribution (epi-HAL).

epi-HAL, for ‘ergodic process interpretation of HAL’, is easy to compute in several ways, which follow from the ergodic property. For example, one can compute the eigenvector of pHAL that belongs to the eigenvalue of 1. Doing this for all documents produces a source representation for each document. The same can be done for the query, which would represent the searcher. To rank the documents in order of relevance to the searcher, the documents are not compared to the query directly (as in the vector space model) but the sources are compared. Researchers in the language modeling community use the Kullback-Leibler (KL) divergence to compare distributions, and so will we. The algorithm is explained in **Box 2** using a very simple language for clarity.

The main goal of this paper is to explain and more formally justify our approach, which is what we did in the sections so far. Note that a longer query corresponds to a larger sample from the source, so one would expect that longer queries would automatically be more effective. In light of an observation recently published by Bendersky and Croft [2], this needs empirical verification. Therefore, the next section will add a more practical justification by showing that even a straightforward and simple implementation of our approach can already compete with a closely related but much more sophisticated language model.

4 Implementation and Evaluation

There certainly are other language models that use a Markov approach. Besides [15] mentioned earlier, notably Cao, Nie, and Bai [4] use the Markov chain for a similar reason as we do, namely to find a stable distribution to represent the document. But there are a number of choices made in [4] that we do not depend on: we do not use WordNet (for semantic relationships), there are several parameters we do not have to set, and we don't use training for optimization. Furthermore, although the authors of [4] make use of a stationary distribution, there are several issues with their approach: (1) it is computationally demanding, (2) it is hard to know when the process has converged, and (3) there is no indication, let alone a proof, that the algorithm has only one fixed point. So, e.g. depending on the initial state, their stationary distribution may or may not be the one sought after. The observation above that the process we advance is ergodic, obviates all three issues at once: The final distribution of the Markov chain can easily be computed without sampling (it is the eigenvector with eigenvalue 1), it converges very fast, and it is guaranteed to be unique.

We will now turn to an experimental evaluation of our ergodic process interpretation of HAL (epi-HAL). The experiment is comparable to that reported for the relevance model of Lavrenko & Croft [8], following a pseudo-relevance feedback paradigm. We first compute a document ranking in response to a query Q. The top n documents are used to derive a distribution M_{epi}^{n} by computing the epi-HAL over this collection. Similarly, M_{epi}^{Q} is computed for the query. These are used in turn to define a mixture model (cf. equation (15) in [8]).

$$\Pr(w|Q) = \lambda \Pr(w|M_{\text{epi}}^{Q}) + (1-\lambda) \Pr(w|M_{\text{epi}}^{n}) \tag{1}$$

The documents are re-ranked using the KL-divergence, and we use the standard baseline unigram LM in the Lemur toolkit. We set the number of feedback documents, n, to 30. For query extension we used 300 terms. Others use different values here, and such differences are to be expected as the distributions are calculated differently, and there is no better way known than to establish these numbers empirically. With these numbers (or another choice) the query model $M_Q = \Pr(w|Q)$ can be computed. Subsequently, documents are re-ranked via $KL(M_Q||M_D)$, where M_D corresponds to a document language model. In our case, M_D is delivered by the baseline language model. We noted earlier that we expect our approach to work better with longer queries, because a longer query means a larger, and hence more representative, sample from the source. (Note that one could see pseudo-relevance feedback as an attempt to make the query longer.) Such longer, or more verbose, queries also seem more representative of the way humans communicate their information needs, compared to typing in a few query words. Bendersky and Croft in their recent paper [2] simulate increasing verbosity by using TREC topics and take the *description* field as a more verbose version of the *title* field. If our intuition (and theirs) were correct one would expect better results for the description than for the title. They found, however, precision to go down substantially for the description. Bendersky and Croft's intuition is that the focus on the key concepts gets blurred as it were by the verbosity surrounding it. We think this intuition leads to two questions, or rather, predictions:

- Assuming the explanation is valid, what would this predict if the description and title were taken together as the new query? Such a query could become less effective then the description, because it is more verbose. Alternatively, it could become more effective because someway the key concept becomes more prominent. Or, combining the two arguments, a safer guess might be that it lands between the efficacy of description and title in isolation. So this has to be investigated empirically.
- Given that the HAL representation captures the semantic relationships between words in the corpus [3, 9], the cohesion between key concepts would be enhanced by the co-occurance of words expressing the concepts. In turn, that would increase the weight of certain words by increasing their value in the joint probability distribution (the query model). And so it would predict a higher effectiveness of title and description together, than either in isolation. (Note that Bendersky and Croft propose to enhance the focus on the key concept using a learning algorithm to weight the words in the query. A different approach that might lead to the same result.)

We will see how these predictions fare for various combinations of title and description.

4.1 Experimental Results

Besides the title and description from the TREC topics, we also added the narrative, as it is even more verbose than the description. We shall first present the results for the now classical AP corpus, and present some initial results with the ROBUST04 collection that Bendersky and Croft used. The results of AP8889 are in Table 1. We used topics 101-150 of AP8889 because it has an exclusion clause in the narrative. For example topic 102, describing Laser research for SDI,

Table 1. Comparing precision for various degrees of verbosity and different language models for AP8889 topics 101-150. *title*, *desc*, and *narr* stand for the corresponding TREC fields. $narr_{-rc}$ stands for narratives with the topic 101-150 exclusion clauses removed. 'Baseline' is from Lemur's default simple language model, 'Relevance model' follows [8], and 'epi-HAL' is the model proposed in the current paper.

Topics 101-150		$< title >$	$< desc >$	$< title, desc >$	$< title, narr >$	$< title, narr_{-rc} >$
Baseline	MAP	23.6	22.7	28.8	31.7	31.9
	prec@5	41.2	44.4	48.8	50.8	50.0
Relevance model	MAP	29.5	29.0	32.3	32.8	33.0
	prec@5	43.6	44.0	42.8	48.8	46.4
Stable Distribution (epi-HAL)	MAP	32.3	32.4	35.7	39.5	39.3
	prec@5	46.0	46.4	46.2	60.0	58.2

ends with "However, a document clearly focused on use of low-power lasers in consumer products, surgical instruments, or industrial cutting tools is NOT relevant." We used two versions of the narrative, one with, and one without the exclusionary clauses. This way we could get an indication of the effect of verbosity: with the exclusion clause intact, the query is obviously more verbose, but more off focus. We used the Lemur search engine toolkit for the computations. The following models were used: the baseline language model provided by Lemur, the relevance model proposed by Lavrenko and Croft [8], and the stable distribution approach we advance in the current paper. The results for the stable distribution was also computed in Lemur, using its smoothing model, but taking the stable distribution as query model. For the AP corpus and the given topics, the precision goes up with increasing verbosity. The baseline precisions breaks down going from title only to description only, as was observed previously by Bendersky and Croft. Both the relevance model and the epi-HAL model appear to be less sensitive to this break down. And as both are feedback models, perhaps it is the feedback that dampens the effect. For every model, however, when title and description are combined, the precision rebounds completely, and surpasses the precision over either in separation. So verbosity cannot be the sole ground for the lack of precision of the description by itself. Table 2 offers a preliminary comparison of epi-HAL with the best performing published results of a state-of-the-art model by Bendersky and Croft [2]. This variation adopts a machine

Table 2. ROBUST04 results, comparing mean average precision (MAP) for title, description, and their combination, for baseline and epi-HAL. Number of documents: 528,155, topics 301-450 and 601-700

ROBUST04	$< title >$	$< desc >$	$< title, desc >$
Baseline	25.7	24.8	28.7
epi-HAL	31.1	31.0	33.1
Bendersky and Croft	25.28	26.2	-

learning approach to identify which noun phrases in the description are key and use the key concepts to boost retrieval of verbose queries. No results were reported for this model on both title and description as Bendersky and Croft did not run the model on the combination of both. The MAP of 26.2 reported for are those for the KeyConcept[2]<desc> variation of the model.

The results point in the same direction as the AP experiment: the baseline shows the precision collapse for description only, the feedback dampens the effect, precision recovers when title and description are combined, and for our approach the precision increases with verbosity. The epi-HAL largely outperformed the baseline by 21%, 25% and 15% respectively on the use of titles, descriptions, and titles plus descriptions, and outperformed the Bendersky and Croft model by 21% and 18% on the use of titles and descriptions respectively.

Note that these data are still preliminary for a detailed and more conclusive comparison with the learning approach of Bendersky and Croft, and are only cautiously indicative. However, as the descriptions of the query topics of the ROBUST04 collection are more verbose and grammatically complex than those of the W10g and GOV2 collections, we put forward the hypothesis that the encouraging performance of the epi-HAL model is due to the ergodic process having more description to process and hence stabilize to a more effective query representation. If so, this suggests performance improvements will be less pronounced on the W10g and GOV2 collections where the query topics are less verbose. Further experimentation is needed to bear this out.

5 Conclusions and Future Work

We derived a relatively simple language model, epi-HAL, that deviates in several respects from other language models proposed to date. Epi-HAL is based on the observation that texts are produced by humans. From this observation it follows that (1) there must be semantic dependencies underlying the documents, and (2) that the documents must obey surface constraints inherent to natural language. To represent the former, this paper derived the underlying semantics from the Hyperspace Analog to Language (HAL) a theory presuming that words that appear close together in text, will also be close in meaning. The surface constraints were represented by using an ergodic Markov chain.

We believe that current language models are overly general in that they do not incorporate these properties of natural language, the very fabric of the documents they purport to model. We compared a straightforward implementation of the proposed model with a sophisticated relevance model. Evaluation on TREC corpora showed that epi-HAL easily outperformed the relevance model for AP8889 and provided some initial encouraging results on the ROBUST04 collection. The epi-HAL model shows increased precision for more verbose queries, and therefore in the long run may respond more appropriately to the verbose inquiries humans typically engage in when communicating with one another.

The results of the experiments encourages us to pursue several avenues in future work. First, instead of modeling only the query by its stable distribution, the same can be done for the document model. Second a more elaborate and detailed experiment with larger corpora will be conducted. And finally, because the proposed model itself is relatively simple, its performance can be further improved via optimization of parameter settings as applied in current, much more sophisticated models.

References

1. John Anderson. *The Architecture of Cognition.* Harvard University Press, Cambridge, MA, USA, 1983.
2. Michael Bendersky and W. Bruce Croft. Discovering key concepts in verbose queries. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 491–498, New York, NY, USA, 2008. ACM.
3. C. Burgess, K. Livesay, and K Lund. Explorations in context space: Words, sentences, discourse. *Discourse Processes*, 25:211 – 257, 1998.
4. G. Cao, J. Y. Nie, and J. Bai. Using markov chains to exploit word relationships in information retrieval. In *the 8th Conference on Large-Scale Semantic Access to Content (RIAO07)*, 2007.
5. Dorothee Chwilla and Herman Kolk. Accessing world knowledge: Evidence from n400 and reaction time priming. *Cognitive Brain Research*, 25:589–606, 2005.
6. Jianfeng Gao, Jian-Yun Nie, Guangyuan Wu, and Guihong Cao. Dependence language model for information retrieval. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 170–177. ACM Press, 2004.
7. John Lafferty and Chengxiang Zhai. Document language models, query models, and risk minimization for IR. In *Proceedings of the 24th Conference on Research and Development in Information Retrieval*, pages 111–119, 2001.
8. V. Lavrenko and W. Bruce Croft. Relevance-based language models. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, 2001.
9. Kevin Lund and Curt Burgess. Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2):203–208, 1996.
10. Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA, 2005. ACM Press.
11. Jay M. Ponte and W. Bruce Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.
12. Claude Elwood Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.
13. R. M. Shiffrin and M. Steyvers. The effectiveness of retrieval from memory. In M. Oaksford and N. Chater, editors, *Rational models of cognition*, pages 73–95. Oxford University Press, 1998.
14. Fei Song and W. Bruce Croft. A general language model for information retrieval. In *Proceedings of the 22nd Conference on Research and Development in Information Retrieval*, pages 279–280, 1999.

15. Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM Press.
16. Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342, New York, NY, USA, 2001. ACM Press.