# Modeling Semantic Question Context for Question Answering

**Protima Banerjee and Hyoil Han**

Drexel University, College of Information Science and Technology
3141 Chestnut Street
Philadelphia, PA 19104
pb66@drexel.edu and hyoil.han@acm.org

## Abstract

Within a Question Answering (QA) framework, Question Context plays a vital role. We define Question Context to be background knowledge that can be used to represent the user's information need more completely than the terms in the query alone. This paper proposes a novel approach that uses statistical language modeling techniques to develop a semantic Question Context which we then incorporate into the Information Retrieval (IR) stage of QA. Our approach proposes an Aspect-Based Relevance Language Model as basis of the Question Context Model. This model proposes that the sparse vocabulary of a query can be supplemented with semantic information from concepts (or aspects) related to query terms that already exist within the corpus. We incorporate the Aspect-Based Relevance Language Model into Question Context Model by first obtaining all of the latent concepts that exist in the corpus for a particular question topic. Then, we derive a likelihood of relevance that relates each Context Term (CT) associated with those aspects to the user's query. Context Terms from the aspects with the highest likelihood of relevance are then incorporated into the query language model based on their relevance score values. We use both query expansion and document model smoothing techniques and evaluate our approach. Our results are promising and show significant improvements in recall using the query expansion method.

## Introduction

In today's environment of information overload, Question Answering (QA) is a critically important research area. To make effective use of the massive amounts of readily available data, humans need efficient tools that will bypass irrelevant information to find the precise "nuggets" of data that answer their specific questions. It is the goal of automated QA systems to ensure that the user is presented with the right information in a timely manner.

Hirchsmann (Hirschmann, 2002) describes a high-level QA architecture, which we conceptualize as being broken down into two stages. The first stage can be thought of as primarily an Information Retrieval (IR) stage which consists of Candidate Document Selection. The second stage performs Intelligent Information Processing and Information Extraction, and encompasses Answer Extraction and Answer Selection. In addition, a generic QA Architecture includes models of the User and Question Context. These models enhance the capability of the system to return only those "nuggets" of information that are relevant to the user. The purpose of this paper is to propose an approach to modeling and incorporating the Question Context Model (QCM) into the first stage (the IR stage) of the Question Answering process.

This paper elaborates and expands on our earlier work (Banerjee, 2008) which described the theoretical basis of the Aspect-Based Relevance Language Model and some interesting early qualitative results. Here, we model Question Context from the Aspect-Based Relevance Language Model by first obtaining all of the latent concepts that exist in the corpus for a particular topic using Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999). Then, we calculate a relevance score that relates each term (called Context Term [CT]) associated with those concepts to the user's query by following the same assumptions as the Relevance-Based Language Model (Lavrenko, 2001). Words (or terms) from the most relevant concepts are then incorporated into the query based on their relevance score values.

Our approach is significant for the following reasons:

1. It relates the query as written by the user to one or more semantic concepts which are sense disambiguated and present in the corpus, rather than to individual words that are not disambiguated as would be the case in a "bag of words" model.
2. It uses the corpus itself as a knowledge source, rather than relying on externally available semantic information.

We use the words question and query interchangeably and the words aspect and concept interchangeably in this paper. The remainder of this paper presents related work, background, the proposed aspect-based relevance language model, our experimental setup, results and conclusions.

## Related Work

In recent TREC conferences, semantic approaches to concept recognition have been used to improve performance for Information Retrieval tasks. Caporaso (Caporaso, 2006) has developed semantic analysis methods, which were used to detect expression of particular conceptual or ontological entities using MeSH and the Gene Ontology as resources. The goal of their approach was to develop a broad retrieval of all documents that mentioned any concept related to the query. This approach is related to the idea of "document zoning," (Lavrenko, 2002) which divides the document into topic areas so that only certain sections are used for information retrieval purposes.

In the model proposed by Trieschnigg (Trieschnigg, 2006), a thesaurus is used to identify concepts in documents and topics. The use of a thesaurus such as the Unified Medical Language System (UMLS) allows identification of multi-word terms and the mapping of synonyms to a single concept. In the same line of research, Reeve (Reeve, 2008) used a language modeling approach to mapping a noun phrase in documents to concepts in UMLS.

Semantic smoothing (Croft, 2003) has been widely discussed as another approach for the incorporation of semantic information into a language modeling framework. A context-sensitive semantic smoothing model is proposed by Zhou (Zhou, 2006); this approach uses word co-occurrence information in conjunction with the UMLS thesaurus to identify topical matches between question and document language models. However, like the other approaches discussed in this section, this approach to semantic smoothing requires the existence of an external reference or knowledge source, making it applicable to only those domains for which such resources are available.

## Background

The theoretical basis of the Aspect-Based Relevance Language Model makes use of Probabilistic Latent Semantic Analysis (PLSA) and the Relevance-Based Language Model. The following sections describe the characteristics of both of those approaches that are important to note when understanding the contribution of our approach.

### The Aspect Model

The Aspect Model forms the foundation of the Probabilistic Latent Semantic Analysis (PLSA) method proposed by Hofmann (Hofmann, 1999). The underlying premise of the Aspect Model is that we can define words and documents in terms of "aspects" which are associated with a latent class variable.

The Aspect Model is based on two assumptions:

- Words ($w$) and documents ($d$) are independent of one another (bag of words assumption)
- Documents and words are both conditioned on a latent aspect ($z$) which may be thought of as a concept

The Aspect Model has several intuitively appealing features. First, by conditioning words and documents on a latent variable, the zero-frequency problem is addressed. Secondly, a priori knowledge is not required about the concepts within the corpus for the algorithm to work effectively. And finally, the usage of probabilistic methods defines a generative model of the data which are able to address common text processing issues such as synonymy and polysemy. Mathematically, the Aspect Model is represented as follows:

$$P(d,w) = \sum_{z \in Z} P(z)P(w \mid z)P(d \mid z)$$

PLSA uses the Expectation Maximization (Dempster, 1977) algorithm to estimate the probabilities $P(z)$, $P(w|z)$ and $P(d|z)$ for latent variable models for all aspects ($z$) in the set of all possible aspects ($Z$). It should be noted here that one property of PLSA which is particularly important for our efforts is the ability to disambiguate between multiple senses of the same word. At the conclusion of the algorithm, the appearance of the same word ($w$) with different $P(w|z)$ values for different z-categories accounts for the same physical word appearing in different senses.

### Relevance-Based Language Models

Conceptually, the Relevance-Based Language Model (Lavrenko, 2001) represents relevance to an information need as a probability distribution of words which is sampled by some process to manifest a query. Given a collection of documents and a user's query, there exists a set of documents that are relevant to that query in the user's judgment. In a typical retrieval environment, however, we do not know the full set of relevant documents to a query and furthermore, we may not even have any examples of documents which are relevant to the query. Lavrenko (Lavrenko, 2001) suggests a methodology that constructs a relevance model from a set of top ranked documents returned from a query. A relevance model is formally defined as the probability of observing a word $w$ in a set of relevant documents $R$, or $P(w|R)$. The query $q$ is also treated as a sample from $R$, although the sampling process that produces $q$ is not necessarily the same as the process that generates $w$. Lavrenko and Croft formally derive a process whereby $P(w|R)$ can be estimated via $P(w|M_D)$, where $M_D$ is the document model for a limited set of top-ranked documents returned from the query.

## The Proposed Aspect-Based Relevance Language Model

### Motivation

One of the fundamental problems associated with natural language Question Answering is the general sparseness of the queries presented by a user. For example, the question "What position does Warren Moon play?" from the TREC 2006 Question Answering Track (Voorhees, 2006) requires a fair amount of background information before it can be correctly "understood" by either a person or a system. To correctly answer the question, one must realize that Warren Moon is a person, that he plays football, and that in football players are assigned regular positions. In other words, there is a context associated with this question that plays a critical role in our understanding of the question and our understanding of what entails a reasonable answer. To this end, we propose that the sparse vocabulary of a query can be supplemented with semantic information from concepts (or aspects) related to query terms that already exists within the corpus. The problem, however, with this approach is that we have no way of understanding what concepts are most relevant to this particular query. In other words, we need to be able to determine which concepts are most relevant to this particular issuance of the query. We propose the Aspect-Based Relevance Language Model as a solution to this problem.

### Question Context Model (QCM)

We define the Question Context Model (QCM) as a distribution of concepts (i.e., aspects), $P(z|R)$, according to their relevancy to the user's information need. We assume that $R$ is assigned probabilities $P(z|R)$ where $z$ is a latent aspect of an information need, as defined by PLSA. Thus, relevance to an information need is described not in terms of words, but in terms of the latent aspects (or concepts) associated with the information need. Conceptually, this alternative representation of the information need includes context information (via the inclusion of aspects) that extends beyond knowledge related to the specific words themselves.

However, in this case, as with the Relevance-Based Language Model (Lavrenko, 2001), $R$ and $P(z|R)$ are unknown. However, the query, which is composed of terms $q_1,\ldots,q_N$ is known. (Lavrenko, 2001) assumes that we can approximate a relevance model $R$, by viewing the query terms $q_1,\ldots,q_N$ as a random sampling of words from $R$. Using the same assumption, we can approximate:

$$P(z \mid R) \approx P(z \mid q_1,...,q_N)$$

Thus, we can view the probability of a latent concept existing within $R$ as being approximated by the probability of the existence of that concept given the words we have sampled from the relevance model so far. Using Bayes Rule, we arrive at the following equation:

$$P(z \mid q_1,...,q_N) = \frac{P(z)\sum_i P(q_i \mid z)}{P(q_1,...,q_N)}$$

Using the conditional independence assumption stipulated by the Aspect Model:

$$P(z \mid q_1,...,q_N) = \frac{P(z)P(q_1,...,q_N \mid z)}{P(q_1,...,q_N)}$$

Since $P(q_1,\ldots,q_N)$ will be the same for all $z$ we can effectively disregard this for ranking purposes.

This gives us the following equation:

$$P(z \mid R) \approx P(z \mid q_1,...,q_N) \cong P(z)\sum_i P(q_i \mid z)$$

The PLSA algorithm provides the methodology for the calculation of $P(z)$ and $P(q_i|z)$. Thus, we can obtain a distribution of concepts (i.e., aspects), $P(z|R)$, according to their likelihood of relevance to the user's information need and we define the $P(z|R)$ as the Question Context Model (QCM). Aspects are ranked, based on $P(z|R)$ and only top-$k$ ranked aspects are included as a part of the QCM. The terms associated with each aspect are called *Context Terms* (CTs).

### Two Sides to the "Zero-Frequency" Problem

To incorporate the Question Context Model (QCM) into a Language Modeling IR framework (i.e., the first stage of QA), we consider that there are essentially two sides to the "zero-frequency" problem (Witten, 1991): (1) A query input to a QA system rarely contains all of the words that would be relevant to the information need that the query represents; and (2) A document rarely contains all of the words that are related to the information content of the query. To examine the effectiveness of the proposed QCM, we need to understand its effectiveness when applied to both facets of the "zero-frequency" problem: context-based query expansion and context-based document smoothing.

### Incorporation into the QA Framework

#### Approach 1: Context-Based Query Expansion

First, we consider using a query expansion approach. In this case, the Aspect-Based Relevance Language Model, $P(z|R)$, provides a way to quantify the relevance of an aspect to a query. PLSA also provides a posterior probability $P(w|z)$, which quantifies the relationship between an aspect and an individual Context Term associated with the aspect. We define *Candidate Context Terms* (CCTs) by first ranking CTs based on a posterior

probability $P(w|z)$, and then selecting the top-ranked CTs. We can define a weight, $\alpha(w)$, that defines the relevance of an individual Context Term ($w$) to a query as follows:

$$\alpha(w) = P(z \mid R)P(w \mid z)$$

Here, we approach query expansion by weighting each CCT that is added to the query by this weight $\alpha(w)$. The final query submitted to the IR system is a composite of the original query with a term weight of 1 with the CCTs weighted with $\alpha(w)$.

**Approach 2: Context-Based Document Smoothing**

When considering the flip side of the "zero frequency" problem, we propose a linear interpolation approach to document smoothing. The intuition behind our approach is that a document which contains CCTs should have a higher likelihood of generating the query. We propose to interpolate a document model $P_M(w|d)$ with information from the Question Context Model (QCM). Mathematically, we represent the interpolation approach as follows:

$$P(w \mid d) = (1 - P_{avg}(z \mid R)) * P_M(w \mid d) + P_{avg}(z \mid R) * P_{norm}(w \mid QCM)$$

Here, $P_{avg}(z|R)$ denotes the average relevance of the aspects that are included as a part of the Question Context Model. In the simple case where only the top-ranked aspect is included as a part of the Question Context, this reduces to the highest $P(z|R)$ value for the query.

$$P_{norm}(w \mid QCM) = \frac{\sum_{\substack{w \in CCT, \\ z \in QCM}} P(w \mid z)}{\sum_{\substack{w' \in CT, \\ z \in QCM}} P(w' \mid z)}$$

The term $P_{norm}(w|QCM)$ denotes a normalized $P(w|z)$ measure for each word in the Question Context Model: Since the Question Context Model contains only CCTs in the top-$k$ ranked aspects, it does not include every aspect that is part of the original PLSA model, nor every word that was associated with those aspects. Therefore, we must normalize the $P(w|z)$ values that were extracted from PLSA if our model is to remain a probability distribution. We normalize the $P(w|z)$ values by summing $P(w|z)$ for all aspects in the QCM which include the Context Term (CT) w, and then dividing this by the sum $P(w'|z)$ of all words w' which are included as Candidate Context Terms (CCT) for aspects that are included as a part of the QCM.

## Experimental Setup

The experimental methodology we used is shown as a block diagram in Figure 1. To validate our approach, we have used a random set of 100 factoid questions from the Text Retrieval Conference (TREC) 2006 Question Answering Track question set. Each TREC factoid question is associated with a question topic. In the TREC 2006 question set, there are 75 topics, each with approximately 5 factoid questions associated with it, for a total of 403 factoid questions. These questions were processed against the TREC AQUAINT dataset (Voorhees, 2005) which contains approximately 1 million newswire articles. The first step of our experimentation was to index and pre-process the AQUAINT dataset using some standard techniques such as stemming and stopword removal.

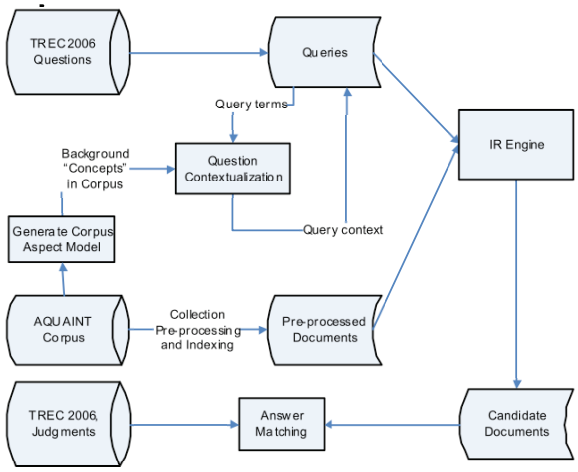


**Figure 1: Experiment Methodology**

We then used the following strategy to perform Question Contextualization:

- For each topic in the TREC 2006 Question Answering Track question set, we obtained a set of 30 top-ranked documents using Indri (Strohman, 2005) to serve as the training documents for PLSA. We determined empirically that 30 documents yielded the most effective training model for our purposes.
- For each topic in the question set, we determined a set of corpus-specific concepts (i.e., aspects) by running PLSA against the candidate documents collected. Our PLSA model was trained using 50 z-categories. In addition, we used part-of-speech tagging to limit the words to include only nouns (including proper nouns). We used the Lemur implementation (Ogilvie, 2001) of PLSA in our experiments.
- We then calculated the distribution that approximates $P(z|R)$ for each question related to that topic by using the Aspect-Based Relevance Model approach proposed in this paper.
- For each question, we ranked the aspects (or concepts) by $P(z|R)$ and their associated CTs by $P(w|z)$. We considered only those aspects which had the highest values for $P(z|R)$ as potential candidate concepts for inclusion into the QCM.

- Once we have obtained a set of top-ranked concepts (or aspects), we can consider the words within those concepts with the highest posterior probabilities as CCTs to be included as a part of the Question Context.

Once created, the Question Context Model is then incorporated into the QA framework using both the Approach 1 and Approach 2 methodologies described earlier.

For the purposes of evaluating the effectiveness of our theoretical model, we use the well-known metric of recall (Manning, 2007). We consider the documents which are relevant to a given question to be only those documents which contain the correct answers as given by the TREC 2006 judgments.

## Results and Discussion

We used the Ponte query expansion approach which is based on a language modeling approach for our query expansion methodology. This method is appropriate to use as a baseline as it employs a query expansion approach based on the top-*N* retrieved documents. Language modeling including Ponte expansion has been used as a baseline by other query modeling approaches (Balog, 2008).

Figure 2 shows the results of our query expansion methodology using a Question Context that includes the top 20 words from the top-5 aspects. Our results show the following:

- A 9.4% improvement in recall performance when the Question Context includes the top-1 aspect, at 50 documents retrieved
- A 9.4% improvement in recall performance when the Question Context includes the top-1 aspect, at 100 documents retrieved
- A 9.3% improvement in recall performance when the Question Context includes the top-5 aspects, at 100 documents retrieved

Using the Wilcoxon signed rank test, we determined that these results are statistically significant ($p < 0.05$) from the baseline approach.

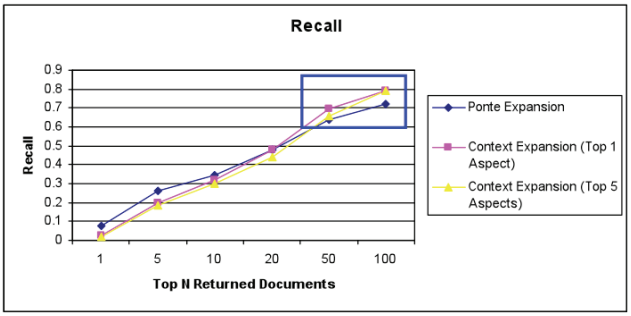


**Figure 2: Recall of Context Based Query Expansion vs. Ponte Query Expansion**

These results show that our methodology shows significant improvement in recall. Improvements in recall are critically important for Information Retrieval applications that feed a Question Answering system. In effect, by increasing the recall of relevant (answer-bearing) documents we are providing better opportunities for the second stage of QA system to find the correct answer. It should be noted that using our system, we have a recall of 80% at 100 documents returned when the Question Context contains the top-1 aspect. This means that 80% of the answer-bearing documents are returned in the top-100 results (i.e., documents) that are submitted to the second stage of QA for processing.

For the Query Context-based document smoothing approach, we compared our smoothing approach against the baseline Kullback-Leibler (KL) divergence approach. This is appropriate as this approach is typically used as the baseline for smoothing approaches (Murdock, 2006). The results of this methodology are shown in Figure 3. The results of this approach were less conclusive. Using the Wilcoxon signed rank test, no significant difference in behavior ($p < 0.05$) was observed between Context-based smoothing and the baseline approach.

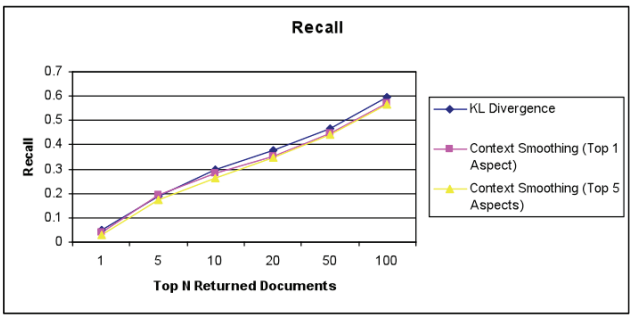


**Figure 3: Recall of Context-Based Smoothing vs. KL-Divergence**

To better understand the results of our methodology for both query expansion and document smoothing, it is useful to consider some examples. When the question "What does LPGA stand for?" is processed, the first two words included in the Question Context are "golf tour." Similarly, when the question "What tobacco company sponsors the Winston Cup Series?" is processed, the first three words to be added to the Question Context are "car," "NASCAR," and "racing." Intuitively, these words enhance the description of the query's information need. Thus, it seems reasonable that a query expansion technique which includes Candidate Context Terms (CCTs) will have a higher likelihood of finding relevant documents than by the use of the terms in the query alone. When document smoothing is considered, on the other hand, we find that many documents that include the term "LPGA" also include the term "golf" and "tour." Our model adds minimal value because the likelihood that a document that includes the term "LPGA" would generate the query is similar to the likelihood that a document that includes the

terms “LPGA,” “golf” and “tour” would generate the query. We believe that this is a characteristic of newswire data, which is typically contains articles focused around a tight topic area. Our future efforts should include investigation into domains which do not share this characteristic.

## Conclusions and Future Work

In conclusion, we have formally presented a novel approach that uses statistical language modeling methods to create a Question Context Model (QCM). We then incorporated the QCM into the IR stage of QA using two approaches, one of which shows significant improvements in recall. This improvement in recall is critical to Question Answering, as it provides more opportunities for the second stage of the Question Answering system to extract and collect the correct answers. Our goals for the future include extending our usage of the Question Context Model from the first stage of QA into the second stage and applying our methods to domain specific corpuses, such as TREC Genomics.

## References


Balog, K., Weerkamp, W., and De Rijke, M. 2008. "A Few Examples Go A Long Way: Constructing Query Models from Elaborate Query Formulations," in Proceedings of ACM SIGIR 2008,,pp. 371-378.

Banerjee, P. and Han, H. 2008. "Incorporation of Corpus-Specific Semantic Information into Question Answering Context," CIKM 2008 - Ontologies and Information Systems for the Semantic Web Workshop, Napa Valley, CA.

Caporaso, J. G., Baumgartner Jr., W. A., Kim, H., Lu, Z., Johnson, H. L., Medvedeva, O., Lindemann, A., Fox, L. M., White, E. K., Cohen, K. B., and Hunter, L. 2006. "Concept Recognition, Information Retrieval, and Machine Learning in Genomics Question-Answering," in Online proceedings of 2006 Text Retrieval Conference.

Croft, W. B., ed. 2003. Language modeling for information retrieval: Kluwer Academic Publishers.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm," Journal of the Royal Statistical Society, vol. 39, pp. 1-38.

Hirschman, L. and Gaizauskas, R. 2002. "Natural language question answering: the view from here," Natural Language Engineering, vol. 7, pp. 275-300.

Hofmann, T. 1999. "Probabilistic latent semantic indexing," in Proceedings of ACM SIGIR 1999.

Lavrenko, V. and Croft, W. B. 2001. "Relevance based language models," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 120-127.

Lavrenko, V. 2002. "Localized smoothing of multinomial language models," CIIR Technical Report IR-222.

Manning, C. D., Raghavan, P., and Schutze, H. 2007. Introduction to Information Retrieval: Cambridge University Press.

Murdock, V. 2006. "Aspects of Sentence Retrieval," Ph.D. Thesis, Center for Intelligent Information Retreival, Amherst, Massachusetts: University of Massachusetts.

Ogilvie, P. and Callan, J. P. 2001. "Experiments Using the Lemur Toolkit," in Online Proceedings of the 2001 Text Retrieval Conference (TREC).

Reeve, L., Han, H, and Brooks, A. 2008. “Online Biomedical Concept Annotation Using Language Modeling,” (short paper) in the 2008 IEEE Conference on Bioinformatics and Biomedicine (BIBM), pp. 453-456.

Strohman, T., Metzler, D., Turtle, H., and Croft, W. B. 2005. "Indri: A language model-based search engine for complex queries," presented as a poster at the International Conference on Intelligence Analysis, McLean, VA.

Trieschnigg, D., Kraaij, W., and Schuemie, M. 2006. "Concept Based Document Retrieval for Genomics Literature," in Online Proceedings of the 2006 Text Retrieval Conference (TREC).

Witten, I. H., and Bell, T. C. 1991. “The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression.” IEEE Transactions on Information Theory, 37(4), 1085-1094.

Voorhees, E. M. and Harman, D. K. 2005. TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing): The MIT Press.

Voorhees, E. M. 2006. "Overview of the TREC 2006 Question Answering Track," in Online Proceedings of 2006 Text Retrieval Conference (TREC).

Zhou, X., Hu, X., Zhang, X., Lin, X., and Song, I. Y. 2006. "Context-sensitive semantic smoothing for the language modeling approach to genomic IR," in Proceedings of ACM SIGIR 2006, pp. 170-177.