Interactive Sense Feedback for Difficult Queries

Alexander Kotov
akotov2@illinois.edu

ChengXiang Zhai
czhai@illinois.edu

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, 61801, USA

ABSTRACT

Ambiguity of query terms is a common cause of inaccurate retrieval results. Existing work has mostly focused on studying how to improve retrieval accuracy by automatically resolving word sense ambiguity. However, fully automatic sense identification and disambiguation is a very challenging task. In this work, we propose to involve a user in the process of disambiguation through interactive sense feedback and study the potential effectiveness of this novel feedback strategy. We propose several general methods to automatically identify the major senses of query terms based on global analysis of document collection and generate concise representations of the discovered senses to the users. This feedback strategy does not rely on initial retrieval results, and thus can be especially useful for improving the results of difficult queries. We evaluated the effectiveness of the proposed methods for sense identification and presentation through simulation experiments and user studies, which both indicate that sense feedback strategy is a promising alternative to the existing interactive feedback techniques such as relevance feedback and term feedback.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*search process, query formulation, relevance feedback*

General Terms

Algorithms, Experimentation

Keywords

Query Analysis, Query Reformulation, Interactive Feedback, Sense Disambiguation

1. INTRODUCTION

Ambiguity is a fundamental property of natural language, which negatively affects the quality of retrieval results by decreasing precision. Generally, an ambiguous query can be defined as any query which contains one or several polysemous terms. The difficulty of lexical ambiguity resolution (or sense disambiguation) varies greatly depending on several factors. When a query is sufficiently long, other terms in the query may serve as effective disambiguation clues due to the collocation effects [12]. In such cases, a search system may attempt to resolve ambiguity in an unsupervised way or by leveraging external resources, such as on-line dictionaries [13] or thesauri (e.g., WordNet [34] [30] [15]). Automatic disambiguation, however, proved to be very challenging, particularly because queries are usually very short and even humans cannot perform it with perfect accuracy.

The problem of ambiguity is exacerbated when a user's information need corresponds to a minority (non-popular) sense of an ambiguous query term in the collection. In such a case, the initial retrieval results would most likely be dominated by a large number of non-relevant documents covering the popular, but distracting senses of an ambiguous query term, while the relevant documents covering the non-popular sense that the user is interested in may be ranked so far down in the ranked list that even diversification of search results would not be very helpful. Clearly, for such difficult queries, any feedback techniques that rely on the assumption that there is some relevant information in the top ranked results (e.g., pseudo feedback, document-level relevance feedback, top results-based term feedback) would not work well either. Consequently, designing an effective feedback method for such difficult queries is a theoretically and practically important problem, particularly in those domains, where short and ambiguous queries prevail, such as Web search.

In this work, we propose *interactive sense feedback* (ISF), a new method for interactive query disambiguation and reformulation, which, unlike the previously proposed methods for interactive relevance feedback [21], such as explicit [9] and term feedback [10] [31], does not rely on the assumption that the initial retrieval results contain relevant documents. Because of its independence of the initial retrieval results, ISF can leverage user interaction both during the early stages of the search process or after it is complete.

At the high level, the proposed ISF is similar to query spelling correction, a popular and widely used feature of all major search engines. When a user submits a misspelled query, she may not be aware (at least immediately) of the reason the search results are of poor quality. A search system, however, can detect the problem, step in and try to improve the results by asking a user if she accidentally mis-

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

spelled the query. Similarly, when users submit ambiguous queries, they are likely to spend some time and effort perusing search results, not realizing that the sense of a polysemous term that they had in mind is not the most common sense in the collection being searched. Similar to spelling correction, along with presenting the initial search results, a search system can provide sense suggestions to narrow down the scope of the query. Ideally, sense suggestions can be presented as clarification questions (e.g., *"Did you mean <ambiguous query term> as <sense label>?"*), where the *sense label* can be either a single term or multiple terms.

Our approach is aiming to *only signal and reveal the ambiguity* of one or several query terms, leaving the final decision whether to disambiguate the query or not to the user. In some sense, our approach takes the best of both worlds: search systems can leverage the vastness of the data and their processing capabilities to infer the *collection-specific senses* of query terms and signal potential problems early on, while the users can leverage their intelligence and world knowledge to interpret the signals from the system and make the final decision. If the users are satisfied with search results, they may simply disregard sense suggestions. However, if the quality of search results is poor and a user can easily identify the desired sense of an ambiguous query term, she may indicate that sense and rely on the search system to update the results, according to the provided feedback.

We illustrate the idea of interactive sense feedback with the following example scenario. Suppose a user submits an ambiguous short query like "piracy" and is looking for documents about instances of copyright law violations as opposed to armed ship hijackings. In a collection of recent news documents, the intended sense of "piracy" corresponds to a minority sense, and one would expect the top-ranked retrieved documents to be non-relevant. Instead of having a user go through the search results and locate the relevant documents, a search system can instead find all the contexts, in which the query term occurred in the collection, indicate that the query term likely has two distinct collection-specific senses and ask the user *"Did you mean piracy as copyright infringement?"* or *"Did you mean piracy as ship hijacking?"*.

From the above discussion, it follows that interactive sense feedback needs to address the following two major problems. The first problem is designing an efficient algorithm for automatic off-line identification of discriminative senses of query terms through the *global* analysis of document collection. We emphasize the global analysis because a local analysis method such as pseudo-feedback cannot discover minority senses when the initial search results are poor, a scenario which we focus on in this work. The second problem is how to generate a representation of the discovered senses in such a way that each sense is easily interpretable and the best sense (i.e. the sense that results in the best retrieval performance) is easily identifiable by the users.

To solve the first problem, we propose and study several different algorithms for discovering the query term senses based on the global analysis of the collection. We compare these algorithms based on their upper bound retrieval performance and select the best performing one.

To solve the second problem, we propose several alternative methods for concise representation of the discovered senses and conducted a user study to evaluate the effectiveness of each method with the actual retrieval performance of user sense selections.

The rest of the paper is organized as follows. In the next section, we review the related work. In Section 3, we introduce the general idea and formally define the concept of interactive sense feedback. In Sections 4 and 5, we provide a detailed description of the methods used for sense detection and representation, respectively. In Section 6, we experimentally determine the potential and actual retrieval effectiveness of sense feedback and report the results of our user study. Finally, section 7 concludes the work and provides directions for future research.

2. RELATED WORK

Methods to improve the quality of retrieval results by reducing the negative impact of lexical ambiguity have been studied for many years and this research direction proved to be very challenging. Below we briefly overview three major lines of related previous work.

The first line is aimed at understanding the nature of lexical ambiguity in IR. This direction has been started by the work of Krovetz and Croft [12], who conducted a series of experiments in order to examine the quantitative aspects of lexical ambiguity in IR test collections and determined its influence on retrieval performance. They concluded that achieving benefits from disambiguation methods is dependent on how successful a sense-aware IR system is in discriminatingly applying them, however an effective and robust automatic disambiguation method is still an open problem in IR. Their results were later enhanced by a series of works by Sanderson [26, 23, 24, 22]. In [22], Sanderson concluded that improvements in IR effectiveness from using automatic disambiguation methods can be observed only if those methods can provide the accuracy close to that of humans and wrong disambiguation decisions can dramatically hurt the retrieval performance.

The second line aims at performing automatic sense disambiguation during retrieval by using external resources (such as machine-readable dictionaries [13], thesauri (e.g., WordNet) or supervised methods. Voorhees [34] conducted the first large scale study of a retrieval system which featured automatic word sense disambiguation based on WordNet and concluded that automatic sense disambiguation did not improve the performance. Mandala et al. [17] proposed a method to combine three different thesaurus types for query expansion: manually constructed (WordNet), automatically constructed based on document co-occurrence relations and automatically constructed based on head-modifier relations and found out that improvements in retrieval performance can be achieved by combining all three types of resources. Liu et al. [15] proposed several heuristics for disambiguating the query terms that used adjacent query terms and WordNet. Kim [11] proposed an approach for coarse-grained disambiguation of nouns by mapping them into 25 unique terms associated with the root synsets of each of the noun hierarchies in WordNet. Gonzalo et al. [8] showed that the performance of vector space retrieval model can be improved if WordNet synsets are chosen as the indexing space. In general, approaches relying on external resources share the common problems of coverage (a query term may have a specialized sense in a particular domain, which may not be covered by a generic lexical resource) and domain mismatch (some of the dictionary senses may not occur in the collection being searched). Automatic disambiguation has also been addressed in the context of vector space retrieval methods. Schütze et al. [27] proposed a method to learn the senses from a vector space representation of the term contexts during training and classify the senses during testing. They achieved the best experimental results by allowing a word to be tagged with up to three senses and combining term and sense ranking. Stokoe et al. [29] used state-of-the-art disambiguation algorithm based on supervised machine learning that was trained on an external corpus to perform retrieval experiments on the TREC WT10G data set and concluded that sense based vector space retrieval consistently outperformed traditional vector space models even if the accuracy of the disambiguation algorithm is below 90%. The query expansion method proposed by Qiu and Frei [20] for generalized vector-space retrieval models used global term co-occurrence data to select the best expansion terms by ranking them according to the vector-space based similarity score of a term to the entire query. Fonseca et al. [7] represented the query concepts as a set of related past queries from the search logs and proposed an interactive query expansion technique for web queries.

The third line aims at addressing the problem of ambiguity indirectly by improving the initial retrieval results through various types of relevance feedback. In the context of pseudo-relevance feedback (PRF), the problems of minority sense and query drift have been addressed through clustering [32] [14] [33]. In particular, Liu and Croft [16] proposed to cluster the initially retrieved documents and used the discovered clusters to smooth the document language model. Pu and He [19] went one step further and proposed to use independent component analysis as a dimensionality reduction technique and cluster the top retrieved documents in the latent semantic space. Xu et al. [36] used a combination of query-specific clustering and external resource (Wikipedia) for query expansion. However, both the traditional and clustering-based PRF can be effective only when there are some relevant documents in the top results, which is generally not the case for difficult queries. In the context of interactive term feedback, an alternative to the document-based relevance feedback, Anick and Tipireni [3] proposed a method for creating lexical hierarchies of expansion terms, based on the linguistically-aware processing of the document collection. A similar method, but based on using simple co-occurrence statistics has been proposed by Sanderson and Croft [25]. Carmel et al. [5] proposed to use lexical affinities to automatically select the expansion terms in such a way that the information gain of the retrieved document set is maximized. Tan et al. [31] proposed a method for interactive term feedback based on clustering the initial retrieval results. They reported that users were having difficulties in selecting the good expansion terms, primarily because term clustering generally lacks semantic coherence. We believe that term feedback has two major limitations. First, similar to PRF, it uses the initially retrieved documents for generating feedback terms, which makes it ineffective for difficult queries. Second, since term feedback does not take into account the relationships between individual terms, it cannot capture the semantics of the feedback terms well. Wang et al. [35] proposed the concept of negative feedback, when only negative signals are used for improving difficult queries, however, to the best of our knowledge, there has been no prior work on improving difficult queries through interactive relevance feedback. Therefore, the primary motivation behind interactive sense feedback is to overcome the limitations of existing relevance feedback methods for difficult queries, when poor search results are caused by query ambiguity.

3. INTERACTIVE SENSE FEEDBACK

3.1 General idea

Despite years of research, there is still no consensus within the AI and IR research communities about what kind of information is the most useful for sense disambiguation. Depending on the definition of a word sense, there are two major ways to approach word sense disambiguation. Within the first view, the sense of a word is defined as its intrinsic property, which corresponds to the high-level concepts denoted by the word lexeme. This view assumes that the correct and comprehensive specification of the word sense requires complete knowledge about the world and can only be provided in the form of a manually created knowledge base. The second view assumes that the sense of a word, rather than being its predefined property, is determined by various contextual clues, such as its syntactic role and the nearby context.

This work adopts the latter view and is based on the assumption that *the senses of a query term can be differentiated by grouping and analyzing all the contexts, in which it appears in the collection*. Consequently, a sense-aware retrieval model should consider not only individual terms, but also all the contexts, in which those terms appear in the collection. It uses two types of contexts: the *local context*, which corresponds to an individual co-occurrence of a term with other terms within a certain unit of text (such as a text window or the entire document) and the *global context*, which aggregates all the local contexts associated with a word. Such aggregation allows to eliminate noise and identify strong, collection-wide contextual co-occurrence relations of a given term with other terms in the vocabulary of a collection. The global context for a particular term can then be analyzed to identify the subsets of terms, which consistently appear in the global contexts of each other. We consider such subsets of terms as the *collection-specific senses* of a query term.

Algorithm-wise, sense feedback works as follows:

1. First, a document collection is preprocessed to construct a contextual similarity matrix, which includes all the terms in the vocabulary of a collection using one of the methods in Section 4.1; the *contextual similarity matrix* is a sparse matrix, in which the rows corresponds to the global contexts of each term in the vocabulary of a collection.

2. Given a query, the retrieval system first constructs a *query term similarity graph* for each query term, which includes all the terms appearing in the global context of the given query term and the contextual co-occurrence relations between them. Next the system identifies clusters of terms in the query term similarity graph. Each of those clusters is then converted into a language model, which takes into account the strength of relations between the terms in the contextual similarity matrix and represents the *collection-specific sense* of a query term.

3. For each of the identified senses, the system generates a concise representation using the method described in Section 5, which can be presented to a user. If a user recognizes the intended sense among those presented by the system, the language model of the original query is updated with the language model of the selected sense. The updated query language model can then be used to retrieve a new set of documents reflecting user feedback and focused on the specific sense of the initially ambiguous query term.

The interactive sense feedback approach has several advantages over the existing feedback methods. Firstly, sense feedback does not rely on the initial retrieval results and can be used either on-line or off-line. Secondly, only those senses that actually occur in the collection would be presented to the users. Finally, sense feedback does not rely on any external resources, and hence is completely general.

3.2 Formal definition

We study interactive sense feedback with the language modeling approach to IR, specifically the KL-divergence retrieval model [37], according to which the retrieval task involves estimating a query language model, Θ_q for a given term-based query q and document language models Θ_{D_i} for each document D_i in the document collection $C = \{D_1, \ldots, D_m\}$. The documents in the collection are scored and ranked according to the Kullback-Leibler divergence:

$$\mathrm{KL}(\Theta_q||\Theta_D) = \sum_{w \in V} p(w|\Theta_q) \log \frac{p(w|\Theta_q)}{p(w|\Theta_D)}$$

Within the KL-divergence retrieval model, relevance feedback is considered as the process of updating the query language model Θ_q, given the feedback obtained after the initial retrieval results are presented to the users. Such feedback may be explicitly provided by the user or implicitly derived from the retrieved results. According to this view, sense feedback can be treated as the process of updating Θ_q with the sense of an ambiguous query term identified by the user as relevant to her information need.

By following the language modeling approach, given a term-based query $q = \{q_1, \ldots, q_n\}$, a particular sense s of the query term q_i is represented as a *sense language model* $\hat{\Theta}^s_{q_i}$.

Definition 1. Sense Language Model $\hat{\Theta}^s_t$ *for a particular sense s of term $t \in V$ is a probability distribution $p(w|\hat{\Theta}^s_t)$ over a subset of words $S \subseteq V$, where V is a vocabulary of a particular document collection C.*

Given that a user selects a particular sense s for the query term q_i, the language model $\hat{\Theta}^s_{q_i}$ associated with the selected sense can be naturally used for updating the original query language model Θ_q through linear interpolation:

$$p(w|\tilde{\Theta}_q) = \alpha p(w|\Theta_q) + (1-\alpha)p(w|\hat{\Theta}^s_{q_i})$$

where α is the interpolation coefficient between the sense language model and the original query model.

Definition 2. Contextual Term Similarity Matrix *is a sparse matrix* $\mathbf{S}$ *of size $n \times n$ where $n = |V|$. Each row* $\mathbf{S}_i$ *corresponds to a word $w_i \in V$ and represents a probability distribution over all other words w in the vocabulary V, such that the probability mass would be concentrated on the terms, which are strongly semantically related to w_i. Each element* $\mathbf{S}_{ij}$ *of the matrix corresponds to a probability $p(w_j|w_i)$, which indicates the strength of semantic relatedness of the words w_i and w_j in a document collection C.*

Definition 3. Term Similarity Graph $G_{w_i} = (V_{w_i}, E_{w_i})$ *for a term w_i is a graph, in which $\forall j \in V_{w_i}, \mathbf{S}_{ij} \neq 0$ and $\forall u, v$, such that $(u,v) \in E_{w_i}$,* $\mathbf{S}_{uv} \neq 0$.

Having formally defined the concept of a sense, in the following sections we discuss the proposed approaches to sense detection and presentation in more detail.

4. SENSE DETECTION

Our sense detection method has two components: constructing the contextual similarity matrix and clustering the query term similarity graph. We discuss each below.

4.1 Contextual term similarity matrix construction

Constructing the contextual similarity matrix for a document collection requires a method to calculate the strength of relations between the terms in the vocabulary. In this work, we experiment with two such methods: mutual information (MI) and HAL scores.

4.1.1 Mutual Information

Given two words w and v, the mutual information between them is calculated by comparing the probability of observing w and v together with the probabilities of observing them independently, according to the following formula:

$$MI(w,v) = \sum_{X_w=0,1} \sum_{X_v=0,1} p(X_w, X_v) \log \frac{p(X_w, X_v)}{p(X_w)p(X_v)}$$

where X_w and X_v are binary variables indicating whether w or v are present or absent in a document. The probabilities are estimated as follows:

$$\begin{aligned}
p(X_w = 1) &= \frac{c(X_w = 1)}{N} \\
p(X_w = 0) &= 1 - p(X_w = 1) \\
p(X_v = 1) &= \frac{c(X_v = 1)}{N} \\
p(X_v = 0) &= 1 - p(X_v = 1) \\
p(X_w = 1, X_v = 1) &= \frac{c(X_w = 1, X_u = 1)}{N} \\
p(X_w = 1, X_v = 0) &= \frac{c(X_w = 1) - c(X_w = 1, X_v = 1)}{N} \\
p(X_w = 0, X_v = 1) &= \frac{c(X_v = 1) - c(X_w = 1, X_v = 1)}{N} \\
p(X_w = 0, X_v = 0) &= 1 - p(X_w = 1, X_v = 0) - \\
&\quad\; p(X_w = 0, X_v = 1) - p(X_w = 1, X_v = 1)
\end{aligned}$$

where $c(X_w = 1)$ and $c(X_v = 1)$ are the numbers of documents containing the words w and v, respectively, and $c(X_w = 1, X_v = 1)$ is the number of documents that contain both w and v. Mutual information measures the strength of association between the two words and can be considered as a measure of their semantic relatedness. The higher the mutual information between the two words, the more often they tend to occur in the same document, and hence, the more semantically related they are. For each term in the vocabulary of the collection, we identify the top k terms that have the highest mutual information scores with the given term and use them as a global similarity context of a term in the contextual similarity matrix of the collection.

4.1.2 Hyperspace Analog to Language

Hyperspace Analogue to Language (or HAL) [4] is a representational model for high dimensional concept spaces. It was created based on the studies of human cognition. Previous work [28] has demonstrated that HAL can be effectively applied to IR. Constructing the HAL space for an n-term

	the	eff	of	poll	on	pop
the	1	2	3	4	5	
eff	5					
of	4	5				
poll	4	5				
on	2	3	4	5		
pop	5	1	2	3	4	

Table 1: HAL space matrix for the sentence "the effects of pollution on the population"

vocabulary involves traversing a sliding window of width w over each word in the corpus, ignoring punctuation, sentence and paragraph boundaries. All words within a sliding window are considered as the local context of the term, over which the sliding window is centered. Each word in the local context receives a score according to its distance from the center of the sliding window (words that are closer to the center receive higher score). After traversing the entire corpus, an $n \times n$ HAL space matrix $\mathbf{H}$, which aggregates the local contexts for all the terms in the vocabulary is produced. In this matrix, the row vectors encode the preceding word order and the column vectors encode the posterior word order. An example HAL space matrix for the sentence "the effects of pollution on the population" constructed using the sliding window of size 10 (5 words before and after the center word) is shown in Table 1.

In the HAL-based approach, the global co-occurrence matrix is first produced by merging the row and column corresponding to each term in the HAL space matrix. Each term t corresponds to a row in the global co-occurrence matrix $\mathbf{H}_t = \{(t_1, c_1), \ldots, (t_m, c_m)\}$, representing the number of co-occurrences of the term t with all other terms in the vocabulary. After the merge, each row $\mathbf{H}_t$ in the global co-occurrence matrix is normalized to obtain the contextual term similarity matrix for the collection:

$$\mathbf{S}_{ti} = \frac{c_i}{\sum_{j=1}^{m} c_j}$$

Unlike mutual information, HAL uses the contextual windows of sizes smaller than the entire document to create the local contexts, which presumably would produce less noisy sets of semantically related terms.

4.2 Sense detection algorithm

Algorithm 1 is a high-level representation of a method to detect the senses of a given query term q_i.

Algorithm 1 Sense detection for a query term q_i

```
1. forall $j : \mathbf{S}_{ij} \neq 0$
       $V_{q_i} \leftarrow V_{q_i} \cup j$
2. forall $(u, v) : (u, v) \in V_{q_i} \times V_{q_i}$
       if $\mathbf{S}_{uv} \neq 0$
           $E_{q_i} \leftarrow E_{q_i} \cup ((u, v); \mathbf{S}_{uv})$
   $G_{q_i} \leftarrow G(V_{q_i}, E_{q_i})$
3. $C \leftarrow cluster(G_{q_i})$
   for $k = 1$ to $|C|$
       forall $t : t \in V_{C_k}$
4.             $p(t|\hat{\Theta}^k_{q_i}) = \frac{\sum_{v:(t,v) \in E_{C_k}} \mathbf{S}_{tv}}{\sum_{w \in V_{C_k}} \sum_{u:(w,u) \in E_{C_k}} \mathbf{S}_{wu}}$
```

The algorithm works as follows:

1. Given a query term q_i, a set of terms related to q_i from the contextual similarity matrix $\mathbf{S}$ forms a set of vertices of the term similarity graph G_{q_i};
2. For each pair of vertices in G_{q_i}, check if there exists a relation in $\mathbf{S}$ with non-zero weight between the terms corresponding to those vertices. If so, the strength of relation becomes the weight of the edge between those terms in G_q;
3. The dynamically constructed query term similarity graph G_q is clustered into a set of subgraphs using one of the graph clustering algorithms;
4. Each cluster (subgraph) C_k is converted into a sense language model $\hat{\Theta}^k_{q_i}$, by normalizing the sum of the weights of all edges adjacent to each term node in the cluster with the sum of the weights of all edges in the cluster.

The hypothesis that the query term similarity graphs contain inherent clustering structure is based on the observation that they are likely to be small world graphs. Small world graphs correspond to a subset of graphs, in which most pairs of nodes are connected with very short paths. Small world graphs are known to contain inherent community or cluster structure. In this work, we experiment with two methods for finding this structure: Clauset-Newman-Moore community clustering algorithm [6] and clustering by committee [18].

5. SENSE PRESENTATION

In the proposed sense feedback approach, a sense is represented as a sense language model. Although such a representation is effective for retrieval, it may not be suitable for presenting the senses to the users, since interpreting it may place a significant cognitive burden on them. Therefore, we need to generate a concise representation of a sense that can be easily interpreted. We explore two sense presentation methods: using the top k terms with the highest probability in the sense language model and selecting a small number of the most representative terms from the sense language model as a sense label. The latter approach uses a subgraph of the query term similarity graph, from which the sense language model was created to find a subset of terms that cover the subgraph in such a way that the sum of the weights of the vertices in the cover is maximized. This is known as the Dominating Set Problem, which is NP-complete.

Algorithm 2 Generate a set of labels L for a sense language model $\hat{\Theta}^s_q$

```
   $L \leftarrow \varnothing$
   $C \leftarrow \varnothing$
   $W \leftarrow \varnothing$
   forall $t : t \in \hat{\Theta}^s_q$
1.     $W_t \leftarrow W_t \cup \sum_{v:(t,v) \in E_{C_s}} \mathbf{S}_{tv}$
   $W \leftarrow sort(W)$
2. forall $t : t \in W_t$
       if $t \notin C$
3.         $L \leftarrow L \cup t$
           forall $v : (t, v) \in E_{C_s}$
               $C \leftarrow C \cup v$
```

Therefore, we employ a greedy Algorithm 2, which works as follows:

1. Sort the vertices according to their weights;
2. Traverse the sorted set of vertices W_t, each time selecting the remaining *uncovered* vertex with the highest weight and add this vertex to the set of sense labels L;

3. Add the selected vertex and all the vertices adjacent to it in the cluster subgraph to the set of covered vertices and select the next label, until all the vertices of the subgraph, which corresponds to the labeled sense, are covered.

6. EXPERIMENTS

In this section, we present the results for experimental evaluation of sense feedback. We first describe our experimental setup and two experimental settings used to study the upper-bound and actual retrieval effectiveness of sense feedback. In the first setting, in order to determine the upper bound for potential retrieval effectiveness of sense feedback on several standard TREC datasets, we simulated the optimal user behavior by measuring the retrieval performance of all the senses discovered by each sense detection method and saving only the results of the optimal (best performing) sense. We also determined the optimal parameter settings for each sense detection method through simulation experiments and compared the upper-bound effectiveness of each method with the baselines. In the second setting, in order to find out whether the users can recognize the query term senses discovered by the best sense detection method and effectively use them to improve the quality of retrieval results, we conducted a user study by asking the users to pick one sense for each query based on different sense presentation methods. We then determined the best method for sense presentation and the actual performance of sense feedback, using different sense presentation methods.

6.1 Datasets and experimental setup

All experiments were conducted on three standard TREC collections: AP (Associated Press), which was used for various Ad Hoc tracks; ROBUST04, which was used for the 2004 Robust track [1] and AQUAINT, which was used for the 2005 HARD [2] and Robust tracks. Various statistics for the experimental datasets are summarized in Table 2.

Corpus	#Docs	Size(Mb)	#Topics	Avg. top.
AP88-89	164,597	507	100	3.5
ROBUST04	528,155	1910	250	2.65
AQUAINT	1,033,461	3042	50	2.56

Table 2: Statistics of the experimental datasets

The TREC topics 51-150 for the AP88-89 collection are long, sentence-like queries, including on average more than 3 query terms. The TREC topics 301-450 and 601-700 for the ROBUST04 collection are mostly 2-3 term queries with a small number of highly ambiguous one term queries (e.g, metabolism, robotics, tourism, creativity). The AQUAINT topics include 50 topics, known to be hard (i.e. resulting in very low retrieval performance) from the previous Robust tacks. All documents and queries have been preprocessed by stemming with the Porter stemmer and removing the stop words. For each of the test collections, we precomputed the contextual term matrices using both the mutual information and HAL scores as measures of similarity. We did not include very rare terms (the ones that occur less than 5 times in the entire collection) or very popular ones (the ones that occur in more than 10% of documents) in the contextual term similarity matrices. A maximum of 100 most contextually similar terms according to the particular similarity measure have been stored for each term in the contextual term similarity matrix. For the construction of the query term similarity graphs, we selected only those terms with a similarity value greater than 0.001.

6.2 Upper-bound performance

We experimentally determine the upper bound for retrieval performance of sense feedback and compare it with the baseline feedback method on all three test collections. As a baseline, we use the model-based feedback method proposed in [38] with the suggested parameter settings: mixture noise was set to 0.95 and feedback coefficient to 0.9. Note that since the proposed sense feedback method is meant to be a complementary, rather than a competing method to pseudo feedback (any pseudo feedback method can be easily combined with sense feedback), we only include pseudo feedback as a reference baseline and are not interested in comparing with the best performing pseudo feedback methods.

The upper bound for the retrieval performance of sense feedback is determined through simulation of a user who is always able to select the optimal sense for each query. Specifically, we first determine all possible senses for each query term and use each sense to update the initial query model, perform retrieval, and estimate its effectiveness using the relevance judgments. The sense that maximizes the average precision of the retrieved results is then chosen as the best sense for a given query. For model-based pseudo-relevance feedback we used the top 10 retrieved documents. For initial retrieval, we used the KL divergence retrieval method with a Dirichlet smoothing prior set to 2000. Before comparing different sense detection methods to the baseline in the following section we determine the optimal configuration for each of them. We use the AP88-89 dataset for parameter tuning.

6.2.1 Parameter setting

First, we set the interpolation coefficient α to 0.9 and empirically determined the optimal size of the sliding window, used for construction of the HAL-based contextual similarity matrix. Figure 1 shows the performance of Community Clustering (CC) and Clustering By Committee (CBC) with respect to MAP on the HAL-based contextual similarity matrix by varying the size of the sliding window used for its construction.

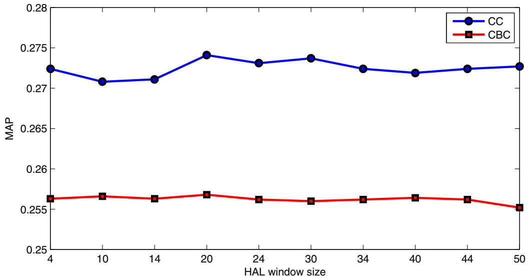

Figure 1: Performance of sense detection methods by varying the size of the HAL sliding window

Two important conclusions can be made based on the analysis of Figure 1. First, community clustering consistently outperforms clustering by committee for all sizes of the HAL window. Second, the optimal size of the HAL window for both sense detection methods is 20 (10 words to the left and 10 words to the right from the target word).

Next, we determine the optimal value of the interpolation coefficient α for different combinations of methods for construction of the contextual term similarity matrix and sense detection. In these experiments, we set the size of the HAL window to its optimal value of 20.

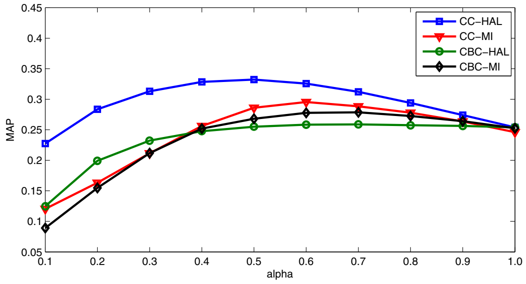

Figure 2: Performance of sense detection methods by varying the interpolation parameter α (the name of the sense detection method is before the hyphen and the similarity measure is after the hyphen).

From Figure 2, it follows that the combination of community clustering and HAL-based similarity matrix outperforms all other sense detection methods. The best configuration for each sense detection method is as follows: $w = 20$ and $\alpha = 0.5$ for CC-HAL; $w = 20$ and $\alpha = 0.7$ for CBC-HAL; $\alpha = 0.6$ for CC-MI and $\alpha = 0.7$ for CBC-MI. Having determined the optimal parameter setting for each sense detection method, in the next section we determine the best sense feedback method with respect to the upper-bound retrieval performance and compare it to the baselines.

6.2.2 *Upper-bound comparison of sense feedback*

The upper-bound performance of different combinations of methods for construction of contextual similarity matrix and sense detection on all three experimental datasets is summarized and compared with the baselines in Table 3. For these experiments, we used the best configuration for each sense detection method empirically determined in the previous section. All feedback methods are evaluated based on their ranking of the top 1000 documents with respect to the mean average (non-interpolated) precision (MAP), precision at top 5 and 20 documents (Pr@5 and Pr@20) and the total number of relevant documents retrieved (RR). We also report the retrieval performance of the initial KL-divergence based retrieval (KL), which is used for model-based pseudo-feedback (KL-PF). As explained earlier, we include pseudo feedback only as a reference baseline since sense feedback is meant to be complementary with pseudo feedback, and they can be easily combined.

From the analysis of Table 3, we can make the following conclusions:

1. The combination of community clustering and HAL-based construction of contextual similarity matrix outperforms all other methods and the baselines both in terms of MAP and Pr@N, indicating the potential of using the identify senses of the query terms to improve retrieval;

2. Community clustering generally outperforms clustering by committee both in combination with mutual information-based similarity matrix construction and HAL-based similarity matrix construction;

3. Sense feedback is effective for both short AQUAINT and ROBUST04 queries and longer AP queries.

Next, we compared the upper-bound effectiveness of sense feedback to the baselines in case of difficult queries (i.e., queries whose initial retrieval results have a MAP value less than 0.1). The results are presented in Table 4. As follows from Table 4, sense feedback effectively improves performance of difficult queries and outperforms both baselines, particularly improving the ranking of the top results, as indicated by significant improvement in Pr@5 and Pr@10, whereas for the AQUAINT dataset, pseudo-feedback decreased the retrieval performance.

		KL	KL-PF	SF
AP88-89	MAP	0.0346	0.0744	0.0876*
	Pr@5	0.1118	0.1529	0.25
	Pr@10	0.0824	0.1412	0.2031
ROBUST04	MAP	0.04	0.067	0.073*†
	Pr@5	0.1567	0.1675	0.3054
	Pr@10	0.1527	0.1554	0.2608
AQUAINT	MAP	0.0473	0.0371	0.0888*†
	Pr@5	0.125	0.075	0.2875
	Pr@10	0.1188	0.0813	0.2375

Table 4: Comparison of the upper-bound performance of sense feedback with KL-divergence retrieval model (KL) and model-based pseudo-relevance feedback (KL-PF) on difficult topics. * indicates statistically significant difference relative to KL (95% confidence level), according to the Wilcoxon signed-rank test. † indicates statistically significant difference relative to KL-PF (95% confidence level), according to the Wilcoxon signed-rank test.

The absolute numbers of difficult and normal topics improved by pseudo-feedback and sense feedback on different datasets are shown in Table 5.

	T	D	N	KL-PF		SF	
				D+	N+	D+	N+
AP	99	34	65	19	44	31	37
ROBUST04	249	74	175	37	89	68	153
AQUAINT	50	16	34	4	26	12	29

Table 5: Number of difficult (D) and normal (N) topics improved by pseudo-feedback (KL-PF) and sense feedback (SF) in different datasets.

As follows from Table 5, sense feedback improved the retrieval performance of a significantly larger number of both difficult and normal queries than pseudo-feedback in each dataset.

6.3 User study

Although it is clear from the simulation experiments that automatically identified senses have the potential to improve the quality of retrieval, the next important question is whether the users can recognize and select the optimal sense from retrieval perspective. In order to answer this question, we conducted a user study. For the user study we selected the AQUAINT topics, since those topics were used in 2005 TREC HARD track, which was created to explore

		KL	KL-PF	CC-MI	CC-HAL	CBC-MI	CBC-HAL
AP88-89	MAP	0.2492	0.3066	0.2955	0.3323	0.2786	0.2588
	RR	6833	7767	7058	7588	7141	6794
	Pr@5	0.4121	0.4444	0.5089	0.5771	0.4708	0.4371
	Pr@20	0.3652	0.4096	0.4417	0.4818	0.4042	0.3820
ROBUST04	MAP	0.2462	0.2569	0.2538	0.3002	0.2477	0.2571
	RR	10227	11386	9401	10842	9993	10387
	Pr@5	0.4659	0.4426	0.5159	0.5871	0.4840	0.4851
	Pr@20	0.3490	0.3454	0.3737	0.4116	0.3634	0.3657
AQUAINT	MAP	0.1942	0.2189	0.2237	0.2286	0.2060	0.2004
	RR	4107	4142	4166	4166	4153	4155
	Pr@5	0.496	0.488	0.5833	0.6120	0.5224	0.5
	Pr@20	0.394	0.427	0.4573	0.456	0.3959	0.389

Table 3: Comparison of the upper-bound performance of sense feedback methods with the baselines on all topics and all collections.

the methods for improving the accuracy of retrieval systems through "highly focused, short-duration interaction with the searcher". In the study, we asked the six participants to assume that they are typing a provided TREC query into the search engine box and the search engine asks to clarify the meaning of a query by first selecting a query term and one of its senses that best fits the description of the query and makes the entire query less ambiguous.

We used the best performing combination of community clustering and HAL scores to generate the candidate senses of the query terms for the user study and presented the discovered senses using one-term labels, two-term labels, three-term labels, top-three terms from the sense language model and the top 10 words from the sense language model. We then compared query term and sense selections made by the users with the query term and sense selections resulting in the best upper-bound retrieval performance determined through simulation. Table 6 shows the accuracy of sense selection by the users as the fraction (in percentages) of the users, who select *both* the optimal term and the optimal sense for feedback (in boldface) and the optimal term only (in parenthesis), regardless of whether the selected sense of the term is optimal.

	LAB1	LAB2	LAB3	SLM3	SLM10
User 1	**18**(56)%	**18**(60)%	**20**(64)%	**36**(62)%	**30**(60)%
User 2	**24**(54)%	**18**(50)%	**12**(46)%	**20**(42)%	**24**(54)%
User 3	**28**(58)%	**20**(50)%	**22**(46)%	**26**(48)%	**22**(50)%
User 4	**18**(48)%	**18**(50)%	**18**(52)%	**20**(48)%	**28**(54)%
User 5	**26**(64)%	**22**(60)%	**24**(58)%	**24**(56)%	**16**(50)%
User 6	**22**(62)%	**26**(64)%	**26**(60)%	**28**(64)%	**30**(62)%

Table 6: Percentage of users selecting the optimal sense of the optimal term for sense feedback (in boldface) and the optimal term, but suboptimal sense (in parenthesis).

As follows from Table 6, on average for most labeling methods the users were able to select the best term for sense feedback at least for half of the queries in the study, which indicates that users are able to identify the potentially ambiguous query terms that can benefit most from sense feedback. The percentages of times that the users can select *both* the best term for sense feedback *and* the best sense of that term are less, achieving the maximum of 36%. The following interesting conclusions can also be made from the analysis of Table 6:

1. Users do not tend to select the best sense more often when they observe more terms both in the label and the sense language model. One-term label is often sufficient to recognize the best sense and adding more terms can potentially mislead and confuse the user. The best result of 36% correctly identified best sense selections for one of the users is achieved when the top-3 terms in the sense language model are presented as a sense label.

2. 3-term labeling and choosing the top 3 terms from the sense language model perform comparably, which suggests that terms with the highest probability are generally the most representative for a sense, and vertices corresponding to them cover most of the sense subgraph.

In order to determine the practical utility of interactive sense feedback, we generated and evaluated the retrieval results based on the actual user sense selections. First we tuned α, the parameter for interpolating the sense language model into the original language model. Using sense selections of users for the best sense representation method (we used top 10 terms with the highest weights in the sense language model for parameter tuning and evaluation, since it is the best sense representation method, according to Table 6), we varied the value of the interpolation coefficient α and plotted the resulting performance on all AQUAINT queries with respect to MAP in Figure 3.

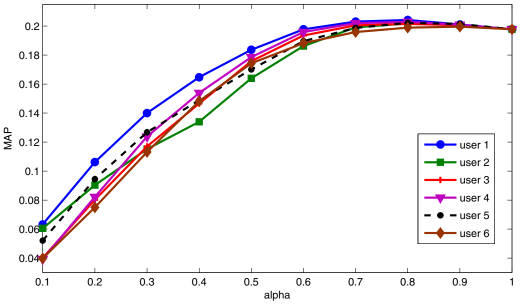

Figure 3: Retrieval performance of user sense selections for all queries in terms of MAP, depending on the value of interpolation parameter α.

From Figure 3 it follows that when $\alpha = 0.8$ sense feedback is consistently most effective in terms of MAP for all the users. Setting α to its optimal value, we determined the retrieval performance of user sense selections on difficult topics, according to different sense presentation methods. The results are presented in Table 7.

KL MAP=0.0473					
KL-PF MAP=0.0371					
	LAB1	LAB2	LAB3	SLM3	SLM10
User 1	0.0543	0.0518	0.0520	0.0564	0.0548
User 2	0.0516	0.0509	0.0515	0.0544	0.0536
User 3	0.0533	0.0547	0.0545	0.0550	0.0562
User 4	0.0506	0.0506	0.0507	0.0507	0.0516
User 5	0.0519	0.0529	0.0517	0.0522	0.0518
User 6	0.0526	0.0518	0.0524	0.056	0.0534

Table 7: Retrieval performance of user sense selections on difficult topics with respect to MAP, depending on the sense presentation method. Performance of the baselines is shown in the first two rows of the table.

As follows from Table 7, although the user sense selections do not achieve the upper bound performance, we can conclude that interactive sense feedback can effectively improve the retrieval performance of difficult queries.

6.4 Examples of discovered senses

To gain some insight at what the senses presented to the user look like, in Tables 8 and 9, we show some sample senses discovered by using the community clustering algorithm in combination with the HAL scores for the query term "stealth" of the AP topic #132 "stealth aircraft" and for the query term "cancer" of the AQUAINT topic # 310 "radio waves and brain cancer". Inferring the meaning behind each sense from the top representative terms is not hard, but sometimes requires certain background knowledge. For example Sense 2 of the query term "stealth" clearly corresponds to aircrafts with low radar visibility.

In case of the term "cancer", senses are less distinguishable, but nevertheless correspond to semantically coherent aspects of the query topic. For example, sense 1 most likely corresponds to cancer research, sense 2 is about different types of cancer, sense 3 is about cancer treatment and sense 4 is likely to correspond to cancer statistics in the US.

6.5 Error analysis

It is important to note that most TREC queries consist of at least 2-3 terms and are generally not highly ambiguous. Therefore, several collection-based senses of a query term may have comparable retrieval performance to the best sense and users often select these senses instead of the best performing sense. For example, for the query #625 "arrests bombing WTC" the best sense is the sense labeled as "police" for the query term "bombing". However, all the users who participated in the study selected the sense labeled as "arrest" for the query term "WTC". Similarly, for the query #639 "consumer on-line shopping" most users selected the sense labeled as "web" for the query term "consumer", whereas the best sense is the sense labeled "online" for the query term "shopping".

7. DISCUSSION AND FUTURE WORK

In this work, we presented a novel idea of interactive sense feedback. Sense feedback can automatically discover collection specific senses of query terms, present those senses to the users and update the queries based on user sense selections. Because the senses are discovered from the entire collection, this feedback strategy is not biased to focus on the popular senses covered in the top-ranked results, and thus is especially useful for improving performance for difficult queries.

We experimentally determined the upper bound for the retrieval performance of all possible combinations of several different methods for automatic sense discovery and measuring the strength of semantic relatedness between the terms. Experimental results show that the combination of Community Clustering and Hyperspace Analog to Language (HAL) results in the best overall retrieval performance and can also significantly improve the retrieval accuracy for difficult queries. We also proposed different presentation methods for the discovered senses and evaluated the effectiveness of user sense selections when senses are concisely represented. The results show that users are able to select the best senses in most cases, leading to improvement of average retrieval accuracy for difficult queries. Therefore, sense feedback has all the potential to be used as an alternative or supplemental technique to the existing interactive feedback methods, such as term, relevance and pseudo-feedback, particularly for difficult queries.

Our work can be extended in several ways. First, we can explore other methods for automatic sense detection and compare them with the ones proposed in this work. Second, we can investigate alternative ways of effectively presenting senses to the users. Finally, it would be very interesting to experiment with sense feedback on real ambiguous Web-style queries and incorporate sense feedback into search engine infrastructure as a complimentary strategy to search results diversification. We envision that sense feedback will show its full real potential in this case.

8. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Numbers IIS-0713581, CNS-0834709, and CNS 1028381, by a Sloan Research Fellowship, and by a Microsoft gift grant.

9. REFERENCES

[1] J. Allan. Overview of the trec 2004 robust retrieval track. In *Proceedings of TREC 13*, 2004.

[2] J. Allan. Hard track overview in trec 2005 - high accuracy retrieval from documents. In *Proceedings of TREC 14*, 2005.

[3] P. G. Anick and S. Tipirneni. The paraphrase search assistant: Terminological feedback for iterative information seeking. In *Proceedings of ACM SIGIR'99*, pages 153–161, 1999.

[4] C. Burgess, K. Livesay, and K. Lund. Explorations in context space: Words, sentences and discourse. *Discourse Processes*, 25:211–257, 1998.

[5] D. Carmel, E. Farchi, Y. Petruschka, and A. Soffer. Automatic query refinement using lexical affinities with maximal information gain. In *Proceedings of ACM SIGIR'02*, pages 283–290, 2002.

[6] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, 70(6):066111, 2004.

[7] B. M. Fonseca, P. Golgher, B. Pôssas, B. Ribero-Neto, and N. Ziviani. Concept-based interactive query expansion. In *Proceedings of CIKM'05*, pages 696–703, 2005.

[8] J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. Indexing with wordnet synsets can improve text retrieval. In *Proceedings of ACL/COLING Workshop*

Sense 1		Sense 2		Sense 3		Sense 4	
w	$p(w\|s)$	w	$p(w\|s)$	w	$p(w\|s)$	w	$p(w\|s)$
budget	0.05	117a	0.068	missil	0.33	technolog	0.187
senat	0.05	us	0.05	midgetman	0.22	research	0.15625
fiscal	0.045	plane	0.047	mx	0.17	advanc	0.15625
cut	0.0421	fighter	0.0463	trident	0.14	new	0.06
chenei	0.0391	f	0.0461	nuclear	0.12	make	0.06

Table 8: Examples of senses discovered for the term "stealth" in the query "stealth aircraft"

Sense 1		Sense 2		Sense 3		Sense 4	
w	$p(w\|s)$	w	$p(w\|s)$	w	$p(w\|s)$	w	$p(w\|s)$
research	0.065	diseas	0.076	treatment	0.062	us	0.1847
new	0.057	caus	0.058	chemotherapi	0.06	women	0.1086
studi	0.050	liver	0.051	doctor	0.06	men	0.086
scientist	0.048	lung	0.049	tumor	0.052	breast	0.0976
dr	0.0448	drug	0.049	patient	0.05	ovarian	0.068

Table 9: Examples of senses discovered for the term "cancer" in the query "radio waves and brain cancer"

on Usage of WordNet for Natural Language Processing, 1998.

[9] M. Iwayama. Relevance feedback with a small number of relevance judgments: Incremental relevance feedback vs. document clustering. In *Proceedings of ACM SIGIR'00*, pages 10–16, 2000.

[10] D. Kelly and X. Fu. Elicitation of term relevance feedback: An investigation of term source and context. In *Proceedings of ACM SIGIR'06*, pages 453–460, 2006.

[11] S.-B. Kim, H.-C. Seo, and H.-C. Rim. Information retrieval using word senses: Root sense tagging approach. In *Proceedings of ACM SIGIR'04*, pages 258–265, 2004.

[12] R. Krovetz and W. B. Croft. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2):115–141, 1992.

[13] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of SIGDOC'86*, pages 24–26, 1986.

[14] A. Leuski. Evaluating document clustering for interactive information retrieval. In *Proceedings of CIKM'01*, pages 33–40, 2001.

[15] S. Liu, F. Liu, C. Yu, and W. Meng. An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In *Proceedings of ACM SIGIR'04*, pages 266–272, 2004.

[16] X. Liu and W. B. Croft. Cluster-based retrieval using language models. In *Proceedings of ACM SIGIR'04*, pages 186–193, 2004.

[17] R. Mandala, T. Tokunaga, and H. Tanaka. Combining multiple evidence from different types of thesaurus for query expansion. In *Proceedings of ACM SIGIR'99*, pages 191–197, 1999.

[18] P. Pantel and D. Lin. Discovering word senses from text. In *Proceedings of KDD'02*, pages 613–619, 2002.

[19] Q. Pu and D. He. Pseudo relevance feedback using semantic clustering in relevance language model. In *Proceedings of ACM CIKM'09*, pages 1931–1934, 2009.

[20] Y. Qiu and H.-P. Frei. Concept based query expansion. In *Proceedings of ACM SIGIR'93*, pages 160–169, 1993.

[21] I. Ruthven and M. Lalmas. A survey on the use of relevance feedback for information access systems. *Knowledge Engineering Review*, 18(2):95–145, 2003.

[22] M. Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of ACM SIGIR'94*, pages 49–57, 1994.

[23] M. Sanderson. Retrieving with good sense. *Information Retrieval*, 2(1):49–69, 2000.

[24] M. Sanderson. Ambiguous queries: Test collections need more sense. In *Proceedings of ACM SIGIR'08*, pages 499–506, 2008.

[25] M. Sanderson and W. B. Croft. Deriving concept hierarchies from text. In *Proceedings of ACM SIGIR'99*, pages 206–213, 1999.

[26] M. Sanderson and C. van Rijsbergen. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems*, 17(4):440–465, 1999.

[27] H. Schütze and J. O. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

[28] D. Song and P. Bruza. Towards context sensitive information inference. *JASIST*, 54(4):321–334, 2003.

[29] C. Stokoe, M. P. Oakes, and J. Tait. Word sense disambiguation in information retrieval revisited. In *Proceedings of ACM SIGIR'03*, pages 159–166, 2003.

[30] M. Sussna. Word sense disambiguation for free-text indexing using a massive semantic network. In *Proceedings of CIKM'93*, pages 67–74, 1993.

[31] B. Tan, A. Velivelli, H. Fang, and C. Zhai. Term feedback for information retrieval with language models. In *Proceedings of ACM SIGIR'07*, pages 263–270, 2007.

[32] A. Tombros, R. Villa, and C. J. van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management*, 38(4):559–582, 2002.

[33] R. Udupa, A. Bhole, and P. Bhattacharyya. ä term is known by the company it keeps̈: On selecting a good expansion set in pseudo-relevance feedback. In *Proceedings of ICTIR'09*, pages 104–115, 2009.

[34] E. M. Voorhees. Using wordnet to disambiguate word senses for text retrieval. In *Proceedings of ACM SIGIR'93*, pages 171–180, 1993.

[35] X. Wang, H. Fang, and C. Zhai. A study of methods for negative relevance feedback. In *Proceedings of ACM SIGIR'08*, pages 219–226, 2008.

[36] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of ACM SIGIR'09*, pages 59–66, 2009.

[37] C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pages 111–119, 2001.

[38] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of CIKM'01*, pages 403–410, 2001.