International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-3 Number-3 Issue-12 September-2013

A Literature Survey on Automatic Query Expansion for Effective Retrieval Task

Jagendra Singh[1], Aditi Sharan[2], Sifatullah Siddiqi[3]

Abstract

In this paper, we present a survey of important work done on automatic query expansion. Automatic query expansion is the process of automatically supplementing additional terms or phrases to the original query and is considered an extremely promising technique to improve the retrieval effectiveness. In this survey, we discussed a large number of recent approaches to automatic query expansion that include linguistic based, corpus-specific based, query-specific based and search log based approaches. Some of them use lexical resource such as WordNet and other use search log and web data for query expansion. The following questions are also addressed in this work. Why the query expansion is important for information retrieval? What are the main steps of automatic query expansion? What approaches of automatic query expansion are available and how do they compare? What are the critical issues and research directions of automatic query expansion?

Keywords

Information retrieval, Query expansion, WordNet, Query expansion approaches.

1. Introduction

Information retrieval (IR) means finding the relevant documents from a large dataset according to user query. Information retrieval is composed of basic components such as document indexing, searching and ranking. There are a large number of applications in which information retrieval used such as digital -- libraries, information filtering (recommender system), media search (news search, blog retrieval, image retrieval), search engines (enterprise search, mobile search, web search, federal search) and many others. The most critical issue for retrieval effectiveness is the term mismatch problem: generally, the indexers and the users do not use the same words for same concept. This is known as the *vocabulary problem* [1], compounded by synonymy (same word with different meanings, such as "java") and polysemy (different words with the same or similar meanings, such as "tv" and "television"). Synonymy, together with word inflections (such as with plural forms, "television" versus "televisions"), may result in a failure to retrieve relevant documents, with a decrease in *recall* (the ability of the system to retrieve all relevant documents). Polysemy may cause retrieval of erroneous or irrelevant documents, thus implying a decrease in *precision* (the ability of the system to retrieve only relevant documents).

One of the most natural and successful techniques to deal with term mismatch problem is to expand the original query(Query Expansion) with other words that best capture the actual user intent, or that simply produce a more useful query—a query that is more likely to retrieve relevant documents. Considering the above problem there is a need of some automatic techniques that can assist the user in formulating the query.

As the volume of data has dramatically increased while the number of searcher supplied query terms has remained very low. According to authors in [2], the average query length was 2.30 words, the same as that reported ten years before in [3]. While there has been a slight increase in the number of long queries (of five or more words), the most prevalent queries are still those of one, two, and three words. In this situation, the vocabulary problem has become even more serious because the paucity of query terms reduces the possibility of handling synonymy while the heterogeneity and size of data make the effects of Polysemy more severe. Thus the need and the scope of QE have increased.

2. Query Expansion

Query expansion is the process of adding some new terms to the original query to improve the retrieval performance.
There are four different ways of expanding the query:

Jagendra Singh, Ph.D Scholar, Jawaharlal Nehru University, New Delhi, India.
Aditi Sharan, Assistant Professor, Jawaharlal Nehru University, New Delhi, India.
Sifatullah Siddiqi, Ph.D Scholar, Jawaharlal Nehru University, New Delhi, India.

1. Manual (user chooses the expansion terms).
2. Interactive (system suggests the query expansion terms to the user for expanding query).
3. Automatic (whole process, is invisible for the user).
4. Hybrid (combination of two or more query expansion methods).

2.1 Automatic Query Expansion (AQE)

by very positive experimental findings obtained in laboratory settings. In fact, AQE has regained much popularity thanks to the evaluation results obtained at the Text REtrieval Conference series (TREC),
where most participants have made use of this technique, reporting noticeable improvements in retrieval performance. AQE is currently considered an extremely promising technique to improve the retrieval effectiveness of document ranking and there are signs that it is being adopted in commercial applications, especially for desktop and intranet searches. For instance, Google Enterprise, MySQL, and Lucene provide the user with an AQE facility that can be turned on or off. In contrast, it has not yet been regularly employed in the major operational Web information retrieval (IR) systems such as searching engines.

In the last years, a huge number of AQE techniques have been presented using a variety of approaches that leverage on several data sources and employ sophisticated methods for finding new features correlated with the query terms. Today, there are firmer theoretical foundations and a better understanding of the utility and limitations of AQE such as: Which are the critical parameters affecting the method performance, what type of queries is AQE useful for, and so on? At the same time, the basic techniques are being increasingly used in conjunction with other mechanisms to increase their effectiveness, including method combination, more active selection of information sources, and discriminative policies of method application. These scientific advances have been corroborated

2.2 Important Steps in Automatic Query Expansion

According to survey, Automatic query expansion can be divided into four steps. Each step is discussed here in details.

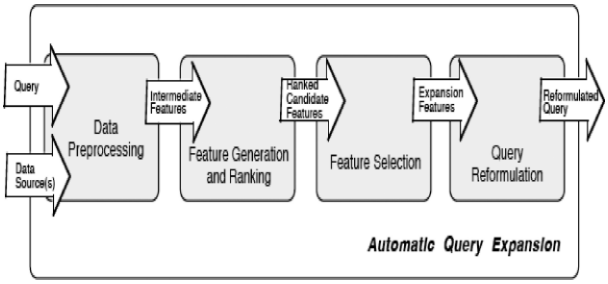

Figure 1: Steps of automatic query expansion.

2.2.1 Pre-processing of Raw Data

In this step, the data used for expanding the user query transforms into a new format that will be more effectively processed by next steps. This step does following task.

(1) Text extraction from documents such as HTML, PDF, MS Word documents.
(2) Tokenization (extraction of individual words).
(3) Stop word removal (removal of common words such as articles and prepositions).
(4) Word stemming (reduction of derivational words to their root form).
(5) Word weighting (assignment of a score that reflects the importance of the word, usually in each document).

2.2.2 Query Term Features Generation and its Ranking

In this section, the system generates and ranks the query term expansion features. Features generation is based on the query term properties. The ranking of features are important for most query expansion methods because only a small amount of the query term expansion features are selected to expand query. The input to in this stage is the user submit query and data source, the output is a set of expansion features with their associated scores.

2.2.3 Selection of Query Term Features

After ranking the query term features in above step, the top elements are selected for query expansion. The selection is made on an individual basis, without considering the mutual dependencies between the expansion features.

2.2.4 Query Formulation and Reformulation

This is the last step of query expansion, this step describes how to submit expanded query to information retrieval system for better results. This

usually assigns a weight to each feature describing the expanded query term reweighting.

A number of AQE techniques have been presented using a variety of approaches that leverage on several data sources and employ sophisticated methods for finding new features correlated with the query terms but no one technique is providing satisfactory results. Although in this work we mainly focus on the use of query expansion for improving information retrieval, but there are other many application of AQE such as Question Answering, Multimedia Information Retrieval, Information Filtering, Cross-Language Information Retrieval, Mobile Search, Expert Finding, Slot-based Document Retrieval, Federated Search etc.

3. Categorization of Automatic Query Expansion Techniques

Automatic query expansion techniques based on technique used can be classified into five main groups and there sub groups according to our survey, these groups are:

1. Query based techniques
    - Distribution difference based techniques
    - Model based techniques
    - Document summarization based techniques
2. Corpus based techniques
    - Concept term based techniques
    - Term clustering based techniques
3. Linguistic based techniques
    - Stemming based techniques
    - Ontology browsing based techniques
    - Syntactic parsing based techniques
4. Web data based techniques
    - Anchor Text based techniques
    - Wikipedia based techniques
5. Search log data based techniques
    - Related queries based techniques
    - Exploiting query documents relationship based techniques

3.1 Linguistic based Techniques

A user's query is considered to be an imprecise description of their information need. Automatic query expansion is the process of reformulating the original query with the goal of improving retrieval effectiveness. There are many efficient query expansion techniques that ignore information about the dependencies between words in natural language. However, recent approaches have shown significant improvements in retrieval effectiveness by using such dependency information between words over those techniques that ignore these dependencies information.

3.1.1 Stemming based techniques

All linguistic techniques are based on global language properties such as morphological, lexical, syntactic and semantic word relationships to expand or reformulate query terms. They are typically based on dictionaries, thesauri, or other similar knowledge representation sources such as WordNet. Expansion features are usually generated independently for the full query and for the content of the database being searched; they are usually more sensitive to word sense ambiguity.

Using word stems is one of the simplest and earliest language-specific AQE techniques. The stemming algorithm can be applied either at indexing time (only the document word stems are stored and then they are matched to the query word stems), as in most systems [4] or at retrieval time (the original document words are stored and then they are matched to the morphological variants of query terms). The latter strategy may be more effective described by Bilotti et al. in [5], but it requires structured querying an ability that may not be present in all document retrieval systems.

3.1.2 Ontology browsing based techniques

Ontology browsing is another well known language-specific AQE technique [6]. Knowledge models such as ontologies and thesauri (the distinction between the two is blurred) provide a means for paraphrasing the user's query in context. Both domain-specific and domain-independent ontologies have been used [7], including the combination of multiple thesauri [8]. Most of the recent work has focused on the use of WordNet. As already remarked, WordNet is very appealing for supporting AQE, but its application may raise several practical issues; e.g., lack of proper nouns and collocations, no exact match between query and concepts, one query term mapping to several noun synsets. Furthermore, the use of WordNet suffers from the disambiguation. In particular, its use for query expansion is advantageous only if the query words are disambiguated almost exactly [9], while word sense disambiguation remains a hard problem according to auther Navigli et al. in [10].

There are several ways to circumvent these difficulties. To increase the coverage of single and multiword concepts, WordNet has been enriched with an automatically constructed thesaurus [10]. The disambiguation issue has been addressed in a more effective manner in some recent papers. In [11], the authors argue that instead of replacing a given query word with its synonyms, hyperonyms, and hyponyms, it might be better to extract the concepts that pertain to the same semantic domain of query, through other types of definitional information derivable from WordNet, such as gloss words and common nodes. The different types of information present in WordNet can also be combined, e.g., to assign terms in the same query into semantically similar groups, followed by conventional expansion of each group [12]. The auther Liu et al. in [13] and [14], classical Wordnet concepts, extracted by a sequential application of heuristic rules to pairs of query terms, are then integrated with other feature extraction methods.

3.1.3 Syntactic parsing based techniques

Syntactic analysis is the third approach for providing additional linguistic information to the original query. In this approach, the main objective is to extract relations between the query terms, which can be used to identify expansion features that appear in related relations. For example, it is possible to index the user query and the top-ranked snippets by relation paths induced from parse trees, and then learn the most relevant paths to the query [15]. The syntactic approach may be more useful for natural language queries to solve more general search tasks, the linguistic analysis can be more effectively integrated with statistical [16] or taxonomic information [17].

3.2 Corpus Based Techniques

The techniques in this category analyze the contents of a full database to identify features used in similar ways. Most early statistical approaches to AQE were corpus-specific and generated correlations between pairs of terms by exploiting term co-occurrence, either at the document level, or to better handle topic drift, in more restricted contexts such as paragraphs, sentences, or small neighbourhoods.

Concept terms [18] and term clustering [19, 20, 21] are two classical strategies, already discussed in the preceding sections. Interlinked Wikipedia articles, latent semantic indexing and mutual information based approaches are used to build an association thesaurus, that is described in [22, 23, 24, 25]. This AQE paradigm has also been recently extended with good results to multimedia documents [26]. Note that since global techniques are data-driven, they may not always have a simple linguistic interpretation.

3.3 Query Based Techniques

Query based techniques take advantage of the local context provided by the query. They can be more effective than corpus based techniques because the latter might be based on features that are frequent in the collection but irrelevant for the query at hand. Query based techniques typically make use of top-ranked documents. The most commonly used methods are analysis of feature distribution difference and model based AQE. Both were discussed in depth in the preceding sections.

A different vein of research on query specific-techniques is based on pre-processing top retrieved documents for filtering out irrelevant features prior to the utilization of a term-ranking function. Besides using just Web snippets, several methods for finding more compact and informative document representations have been proposed, such as passage extraction [27] and text summarization [28]. In Chang et al. [29], the document summaries go through a further process of clustering and classification with the aim of finding an even more reduced set of orthogonal features describing each document (termed *query concepts*). In this case, clustering is used to extract intra document rather than cross-document contextual information.

3.4 Search Log Dataset Based Techniques

The fourth main AQE paradigm is based on analysis of search logs. The idea is to mine query associations that have been implicitly suggested by Web users, thus bypassing the need to generate such associations in the first place by content analysis. Search logs typically contain user queries, followed by the URLs of Web pages that are clicked by the user in the corresponding search results page. One advantage of using search logs is that they may encode implicit relevance feedback, as opposed to strict retrieval feedback. On the other hand, implicit measures are generally thought to be only relatively accurate [30] for an assessment of the reliability of this assumption) and their effectiveness may not be equally good for all types of users and search tasks [31]. Other problems with their use for AQE are caused by noise, incompleteness, sparseness, and the volatility of Web pages and query [32]. Also, the availability of large-scale search logs is an issue.

There are two main AQE techniques based on search logs. The first is to treat the individual queries as documents and extract features from those related to the original user query, with or without making use of their associated retrieval results [33, 34, 35]. The second technique, more widely used, consists of exploiting the relation of queries and retrieval results to provide additional or greater context in finding expansion features. Examples of the latter approach include using top results from past queries [36], finding queries associated with the same documents [37] or user clicks [38], and extracting terms directly from clicked results [39, 40].

3.5 Web Data Based Techniques

A common Web data source for AQE is represented by anchor texts. Anchor texts and real user search queries are very similar because most anchor texts are succinct descriptions of the destination page. However, in the absence of any implicit user feedback, it is difficult to find the anchor texts that are similar to the query because classical ranking techniques do not work well on very short texts. In [41], anchor texts are ranked using several criteria that best relate to the specific nature of the data, such as the number of occurrences of an anchor text (taking into account whether it points to a different site or to the same site) and the number of terms and characters in it. Each anchor text is then assigned a combined rank based on a median aggregation of its individual ranks. At query time, the highest-ranked anchor texts that have a non-empty intersection with the query are selected as refinement features.

Another interesting method, based on Wikipedia documents and hyperlinks, is proposed in [42]. The initial set of candidates associated with a query is restricted by considering only those anchor texts that point to a short set of top ranked documents from a larger set of top-ranked documents, followed by scoring each proportional to its frequency and inversely proportional to the rank of the documents it links to. Specific categories of Wikipedia articles are used in [43]. Other types of Web data that can be employed for AQE include FAQs [40] and the Open Directory Project Web pages [44].
The survey published in the literature has been summarized in Table I.

Table I: Classification of Several Main AQE Methods

Methods	Data Used	Feature Extraction method	Feature Selection method
Billerbeck et al. (2003)	Query logs	Query association	Robertson selection value
Graupman et al. (2005)	Corpus	Web table mining	Association rule
Collins et al. (2005)	Wordnet+corpus	Term associat-ion network	Markov chain
Song et al. (2006)	Corpus + top rank docs	Keyphrase extraction	Info gain + term weight
Riezler et al. (2007)	FAQ traning data	Phrase in FAQ answers	Mutual information
He et al. (2007)	Anchor text + corpus	Top rank docs + terms in anchor text	DFR on fields
Metzler et al. (2007)	Corpus + top rank docs	Markov random field	Maximum likelihood
Lee at al. (2008)	Top rank docs + corpus	Clustering of top ranks docs	Relevence model
Arguello et al. (2008)	wikipedia	Anchor text in top ranks wiki docs	Docs rank + link frequency
Cao et al. (2010)	Top rank docs + corpus	All term in top rank docs	Term classificatio n
Xu et al. (2011)	Wikipedia	All term in top ranked article	Relevence model
Roi et al. (2012)	Wordnet + corpus	Graph based term weighting	Term weigh using graph property

4. Research Challenges

In this section, we discuss three challenges of AQE that pose obstacles for a widespread adoption of AQE in a wider range of operational search systems, parameter setting, efficiency and transparency.

4.1 Setting of Parameters

All AQE techniques rely on several parameters. The number and type of parameters depend on the type of query expansion. For instance, for a typical pseudo relevance feedback method it is necessary to choose the number of pseudo-relevant documents, the number of expansion terms, and the balance coefficient for query reformulation. The retrieval

performance of the overall method is usually markedly dependent on the parameter setting.

4.2 Efficiency

Efficient is very important parameter to execute queries for IR systems such as web search engines that need to deliver real-time results to a very large numbers of users. Faster AQE techniques would allow researchers to carry out more experiments and interactive studies to better understand the applicability and limitations of this methodology.

There are three possible ways to address this issue:

(a) Limit expansion features to a few important items and then rank the expanded query in a standard way.
(b) Allow for a possibly large number of expansion features, but prune features and documents that are unlikely to lead to an improved result when ranking the expanded query.
(c) Use efficient index structure (for applications when it is possible) that support nearly full document ranking against nearly full expanded queries.

4.3 Transparency

Transparency is probably another critical issue, although it has not received much attention so far. AQE acts like a black box employing hidden features that may considerably complicate the interpretation of the logic used by the system to deliver results. For instance, some Web users may be unsatisfied finding documents (even relevant ones) that do not contain the terms in their query. This happens sometimes, using AQE. For example, a document may be returned because the anchor texts pointing to it contain the query terms, or because a query term is subsumed by a more general term in the document, according to a given ontology. When users obtain a result set that they find inadequate, they have no explanation for why certain results not containing the original query terms were ranked high so they have no easy way to improve the query.

5. Research Scope

Most current research effort aims at improving the retrieval effectiveness and robustness of AQE. In this section we focus on three relatively well-established topics: selective AQE, evidence combination, and active feedback.

5.1 Selective AQE

Selective AQE aims to improve query expansion with Decision mechanisms based on the characteristics of queries. Based on the observation that some queries are hurt by expansion, one simple strategy is to disable AQE if the query can be predicted to perform poorly.

Rather than just disabling AQE when its application is deemed harmful, it may be more convenient to apply different expansion strategies according to the type of query. An interesting form of query-dependent AQE is presented in [43] using Wikipedia pages. Queries are classified in three types: (1) entity queries, if they match the title of an entity or redirect page, (2) ambiguous queries, if they match the title of a disambiguation page, and (3) broader queries in all other cases. For each type of query, a different method of AQE is then carried out. Another approach that exploits a similar idea is described in [27]. Queries are classified as either navigational or informational, making use of anchor-link distribution.

5.2 Evidence Combination

Several combination methods have been proposed. Two approaches, already mentioned, consist of selecting the most common terms of those produced by multiple term ranking functions [45], or classifying as relevant or non-relevant the terms produced by the same term-ranking function with different document samples [46]. In [47], the focus is on improving the quality of query term reweighting, rather than choosing the best terms, by taking a linear combination of the term frequencies in three document fields (title, anchor texts, body). All these combination methods were applied as an improvement of pseudo-relevance feedback.

Linguistically-oriented AQE techniques can also greatly benefit from a combined approach due to data scarcity, general-purpose resources are limited in coverage and depth, but they can complement co-occurrence relations when the latter evidence is not available or reliable. In [13], WordNet concepts (synonyms and hyponyms) are combined by heuristic rules with other expansion features extracted using global and local statistical methods.

5.3 Active Feedback

In query-specific AQE techniques, treating the top documents as relevant is not the best strategy. For example, if the top documents have very similar contents, then their cumulative benefit will not be very different from any one of them. The main approach for choosing more informative feedback

documents is to emphasize their diversity. Several techniques have been proposed, such as re-ranking documents based on independent query concepts [48], using cluster centroids or ranking gaps [49], skipping redundant documents [50], estimating uncertainty associated with a feedback model [46], and choosing documents that appear in multiple overlapping clusters [51]. Diversity is always combined, explicitly or implicitly, with relevance. In [52], a more comprehensive framework is presented, which integrates relevance, diversity, and density, where density is measured as the average distance of a document from all other documents.

In [53], the authors present a machine learning approach to active feedback, analogous to that used in [54] to select relevant expansion terms. They classify the top-retrieved documents as good or bad, using various features such as the distribution of query terms in the document and the proximity between the expansion terms and the original query terms in the document. To train the classifier, they use top-retrieved documents labelled as good or bad depending on whether they improve or hurt retrieval performance when used as feedback documents.

6. Conclusion and Future Work

In general, automatic query expansion methods improve the performance of information retrieval system. But there is no proper solution for the vocabulary problem in information retrieval. Nowadays a large number of techniques are available (e.g. Linguistic, query specific, corpus specific, search log and based on web data) that fulfil the different requirements such as query type, computational efficiency, external data availability and characteristics of underlying ranking system. A number of experiments based on automatic query expansion have been done on benchmark dataset. This confirmed a remarkable improvement in retrieval efficiency with gain in recall and precision.

In spite of good results, AQE suffers from drawbacks such as topic drift etc. Domain specific resource reduces the risk of topic drift. But this approach can only be used when we know the domain of a query. The key aspects that need to be improved are parameter setting, robustness retrieval performance, the computational efficiency to execute large query.

In recent years, increase in use of search engines opened a new dimension of using terms of query logs. Such logs provide a good idea about the user need. Wikipedia is also explored as an external resource for query expansion. At present the most promising trends are the development of AQE methods that take into account such as term dependency, the utilization of search query and the injection of interactive facilities into basic AQE. This article will hope fully help to make AQE better know and mode widely accepted.

References

[1] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval", ACM Computing Surveys, Vol 44, Article 1, 2012.

[2] D. Crabtree, P. Andreae and X. Gao, "The vocabulary problem in human-system communication", In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 191–200, 2007.

[3] R. LAU, P. BRUZA AND D. SONG, "Belief revision for adaptive information retrieval", In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 130–137, 2004.

[4] D. A. Hull, "Stemming algorithms: a case study for detailed evaluation", J. Amer. Soc. Info. Science 47, 1, pp. 70–84, 1998.

[5] M. BilottI, B. Katz and J. Lin, "What works better for question answering: Stemming or morphological query expansion", In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop, 2004.

[6] R. Navigli and P. Velardi, "An analysis of ontology-based query expansion strategies", In Proceedings of the ECML/PKDD Workshop on Adaptive Text Extraction and Mining, 2003.

[7] J. Bhogal, A. Macfarlane and P. Smith, "A review of ontology based query expansion", Info. Process. Manage, pp. 866–886, 2007.

[8] R. Mandala, T. Takenobu and T. Hozumi, "The use of wordnet in information retrieval", In Proceedings of the ACL Workshop on the Usage of WordNet in Information Retrieval, Association for Computational Linguistics, pp. 31–37, 1998.

[9] J. Gonzalo, F. Verdejo, I. Chugur and J. Cigarr, "Indexing with wordnet synsets can improve text retrieval", In Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems, Association for Computational Linguistics, pp. 647–678, 1998.

[10] R. Navigli, "Word sense disambiguation: A survey", ACM Comput. Surv. 41, 2, pp. 1–69, 2009.

[11] R. Navigli and P. Velardi, "Structural semantic interconnections: A knowledge-based approach to word sense disambiguation", IEEE Trans.

Pattern Anal. Mach. Intell. 27, 7, pp. 1075–1086, 2005.

[12] Z. Gong, C. W. Cheang, "Multi-term web query expansion using wordnet", In Proceedings of the 17th International Conference on Database and Expert Systems Applications (DEXA'06), Springer, pp. 379–388, 2006.

[13] S. Liu, C. Yu and W. Meng, "An effective approach to document retrieval via utilizing wordnet and recognizing phrases", In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 266–272, 2004.

[14] M. Song, I. Song, X. Hu and R. B. Allen, "Integration of association rules and ontologies for semantic query expansion", Data Knowl. Engin. 63, 1, pp. 63–75, 2007.

[15] R. Sun and T. S. Chua, "Mining dependency relations for query expansion in passage retrieval", In Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 382–389, 2006.

[16] L. Fitzpatrick and M. Dent, "Automatic feedback using past queries: Social searching", In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 306–313, 1997.

[17] W. G. Furnas, T. K. Landauer, M. Gomez and S. Dumais, "The vocabulary problem in human-system communication", Comm. ACM 30, 11, 964–971, 1887.

[18] S. Gauch, J. Wang and S. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases", ACM Trans. Info. Syst. 17, 3, pp. 250–269, 1999.

[19] H. Bast, D. Majumdar and I. Weber, "Efficient interactive query expansion with complete search", In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 857–860, 2007.

[20] C. Crouch and B. Yang, "Experiments in automatic statistical thesaurus construction", In Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 77–88, 1992.

[21] O. Pedersen,"A co-occurrence based thesaurus and two applications to information retrieval", Info. Process. Manage. *33*, 3, pp. 307–318, 1997.

[22] S. Gauch, J. Wang and S. Rachakonda, "A corpus analysis approach for automatic query expansion and its extension to multiple databases", ACM Trans. Info. Syst. 17, 3, pp. 250–269,1999.

[23] J. Hu, W. Deng and J. Guo, "Improving retrieval performance by global analysis", In Proceedings ofthe 18th International Conference on Pattern Recognition. IEEE Computer Society, pp. 703–706, 2006.

[24] L. Park and K. Ramamohanarao, "Query expansion using a collection dependent probabilisticlatent semantic thesaurus", In Proceedings of the 11th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, pp. 224–235, 2007.

[25] D. N. Milne, I. Witten and D. M. Nichols, "A knowledge-based search engine powered by wikipedia", In Proceedings of the 16th ACM Conference on Information and Knowledge Management, ACM Press, pp. 445–454, 2007.

[26] A. Natsev, A. Haubold, L. Xie and R. Yan, "Semantic concept-based query expansion and re-ranking for multimedia retrieval", In Proceedings of the 15th International Conference on Multimedia, ACM Press, pp. 991–1000, 2007.

[27] J. Xu and W. Croft, "Query expansion using local and global document analysis", In Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 4–11, 1996.

[28] A. M. Lam and G. Jones, "Applying summarization techniques for term selection in relevance feedback", In Proceedings of the 24th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, pp. 1–9, 2001.

[29] Y. Chang, I. Ounis and M. Kim, "Query reformulation using automatically generated query concepts from a document space", Info. Process. Manage. 42, 2, pp. 453–468, 2006.

[30] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski and G. Gay, "Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search", ACM Trans. Info. Syst. 25, 2007.

[31] R. White, I. Ruthven and J. Jose, "A study of factors affecting the utility of implicit relevance feedback", In Proceedings of the 28th Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 35–42, 2005.

[32] G. Xue, H. Zeng, Z. Chen and W. Fan, "Optimizing web search using web click-through data", In Proceedings of the 13th ACM International Conference on Information and Knowledge Management, ACM Press, pp. 118–126, 2004.

[33] C. Huang, L. Chien and Y. Oyang, "Relevant term suggestion in interactive web searchbased on contextual information in query session logs", J. Amer. Soc. Info. Science Technol. 54, 7, pp. 638–649, 2003.

[34] R. Jones, B. Rey, O. Madani and W. Greiner, "Generating query substitutions", In Proceedings

of the 15th International Conference on World Wide Web, ACM Press, pp. 387–396, 2006.
[35] G. Xue, H. Zeng, Z. Chen and W. Fan, "Optimizing web search using web click-through data", In Proceedings of the 13th ACM International Conference on Information and Knowledge Management, ACM Press, pp. 118–126, 2004.
[36] M. Mitra, A. Singhal and C. Buckley, "Improving automatic query expansion", In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 206–214, 1998.
[37] L. Fitzpatrick and M. Dent, "Automatic feedback using past queries: Social searching", In Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 306–313, 1997.
[38] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log", In Proceedingsof the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, pp. 407–416, 2000.
[39] H. Cui, J. R. Wen and J. Y. Nie, "Query expansion by mining user logs", IEEE Trans. Knowl. Data Engin. 15, 4, pp. 829–839, 2003.
[40] S. Riezler, A. Vasserman, V. Mittal and Y. Liu, "Statistical machine translation for query expansion in answer retrieval", In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, pp. 464–471, 2007.
[41] R. Kraft and J. Zien, "Mining anchor text for query refinement", In Proceedings of the 13th International Conference on World Wide Web, ACM Press, 666–674, 2004.
[42] J. Arguello, J. L. Elsas, J. Callan and J. G. Carbonell, "Document representation and query expansion models for blog recommendation", In Proceedings of the 2nd International Conference on Weblogs and Social Media, AAAI Press, pp. 10–18, 2008.
[43] Y. Xu, G. Jones and B. Wang, "Query dependent pseudo-relevance feedback based on Wikipedia", In Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 59–66, 2009.
[44] J. Bai, J. Y. Nie, G. Cao and H. Bouchard, "Using query contexts in information retrieval", In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 15–22, 2007.
[45] C. Carpineto, G. Romano and V. Giannini, "Improving retrieval feedback with multiple term-ranking function combination", ACM Trans. Info. Syst. 20, 3, pp. 259–290, 2002.
[46] K. Collins and J. Callan, "Estimation and use of uncertainty in pseudo-relevance feedback", In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 303–310, 2007.
[47] B. He and I. Ounis, "Combining fields for query expansion and adaptive query expansion", Info. Process. Manage. 43, pp. 1294–1307, 2007.
[48] M. Mitra, A. Singhal and C. Buckley, "Improving automatic query expansion", In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 206–214, 1998.
[49] X. Shen and C. Zhai, "Active feedback in ad hoc information retrieval", In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 59–66, 2005.
[50] T. Sakai, M. Manabe and M. Koyama, "Flexible pseudo-relevance feedback via selective sampling", ACM Trans. Info. Syst. 4, 2, pp. 111–35, 2005.
[51] K. S. Lee, W.B. Croft and J. Allan, "A cluster-based resampling method for pseudo-relevance feedback", In Proceedings of the 31th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, pp. 235–242, 2008.
[52] Z. XU and R. Akella, "Incorporating diversity and density in active learning for relevance feedback", In Proceedings of the 29th European Conference on IR Research , Springer, pp. 246–257, 2007.
[53] B. He and I. Ounis, "Finding good feedback documents", In Proceedings of the 18th Conference on Information and Knowledge Management (CIKM'09), ACM Press, pp. 2011–2014, 2009.
[54] G. Cao, J. Gao, J. K. Nie and S. Robertson, "Selecting good expansion terms for pseudo relevance feedback", In Proceedings of the 31st Annual International ACMSIGIR Conference on Research and Development in Information Retrieval, ACM Press, pp. 243–250, 2008.

Jagendra Singh received the M.C.A degree from JNU (Jawaharlal Nehru University), New Delhi in 2007, M.Tech(Computer Sceience) from JNU in 2010 and persuing Ph.D from JNU New Delhi.