Query Expansion techniques

Ashish Kankaria
Indian Institute of Technology Bombay, Mumbai
akankaria@cse.iitb.ac.in

1. NEED OF QUERY EXPANSION

The Information Retrieval system described above works very well if the user is able to convey his information need in form of query. But query is seldom complete. The query provided by the user is often unstructured and incomplete. An incomplete query hinders a search engine from satisfying the user's information need. In practice we need some representation which can correctly and more importantly completely express the user's information need.

Figure 1 explains the need of query expansion. Consider an input query "Sachin Tendulkar". As a search engine developer we would expect that user wants documents related to cricketer Sachin Tendulkar. Consider our corpus has 2 documents. First document is an informative page about Sachin Tendulkar which contains the query terms where as second document is an blog on Tendulkar which has various adjectives related to Sachin Tendulkar like master blaster or God of cricket but does not have query terms.
If we retrieve just based on occurrence of query terms, we would mark second document as irrelevant but actually the second document indeed talks about Sachin Tendulkar and it very well relevant to the user's information need. Thus such documents cannot be retrieved if query is not modified. Thus is it is intuitive that query needs to be expanded, but how do we expand query ?
Following are some of the simple techniques of query expansion,

- Finding synonyms of words, and searching for the synonyms as well
- Finding all the various morphological forms of words by stemming each word in the search query
- Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results
- Re-weighting the terms in the original query
- Creating a dictionary of expansion terms for each terms, and then looking up in the dictionary for expansion

2. EXTERNAL RESOURCE BASED QUERY EXPANSION

In these approaches, the query is expanded using some external resource like WordNet, lexical dictionaries or thesaurus.These dictionaries are built manually which contain mappings of the terms to their relevant terms. There techniques involve look up in such resources and adding the related terms to query. Following are some of external resource based query expansion techniques. [5]

2.1 Thesaurus based expansion

A thesaurus is a data structure that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words and generally lists them in alphabetical order. In this approach thesaurus is used to expand the query terms and all the connected words of query terms are added to query.

Thesaurus based system have been explored and put to use by many organizations. A well known example for such systems is Unified Medical Language System (UMLS) [2004] [3] used with MedLine for querying the bio medical research literature. They maintain a controlled vocabulary which is human controlled. The vocabulary contains similar terms for each bio medical concept. Use of a controlled vocabulary is common for domains which are well resourced

Qui and Frei **??** propose the use of similarity thesaurus for query expansion. The build the thesaurus automatically using the domain specific data available. A thesaurus based query expansion system works well only if we have a rich domain specific thesaurus.

2.2 WordNet based expansion

WordNet is a lexical database for multiple languages. The similar terms from multiple languages are connected via synsets (set of senses). WordNet can be used to fetch related term for a particular term in multiple languages and can help in satisfying user's information need.

Approaches based on the use of external resources like WordNet for query expansion, though extensively studied, have been eventually dropped. Voorhees et al. [2005] [20] use

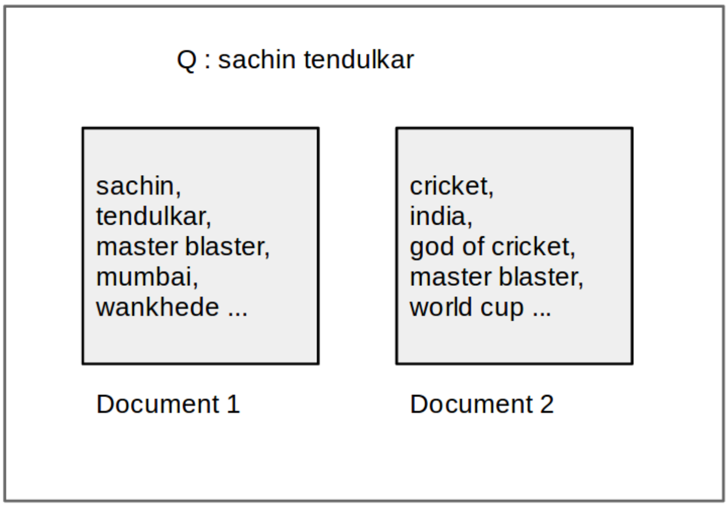

Figure 1: Need of Query expansion

WordNet for query expansion and report negative results. Synonyms terms were added to the query. He observed that this approach makes a little difference in retrieval effectiveness if the original query is well formed.

WordNet is also used by Smeaton et al.[1995] [18] along with POS tagging for query expansion. Each query term was expanded independently and equally. The interesting thing is they completely ignored the original query terms after expansion. As a results of this precision and recall dropped but they were able to retrieve documents which did not contain any of the query terms but are relevant to the query.

Figure 2 summarizes the types of query expansion techniques that have been explorer.

3. QUERY LOGS BASED EXPANSION

With the increase in usage of Web search engines, it is easy to collect and use user query logs. Query logs are maintained by each search engine in order to analyze the behavior of the user while interacting with search engine. These kind of approaches use these query logs to analyze the user's preference and adds corresponding terms to query. This method can fail when user wants to search something which is not at all related to earlier searches.

Cui et al. [2002] [10] developed a system which extracts the expansions terms based on user's behavior which is stored in form of query logs. They maintained a list of all the documents visited for a particular query. Probability of document being visited when a particular query word is present in a query is calculated to find the relevance of the document.

Yin et al. [2009] [23] considered query log as bipartite graph that connects the query nodes to the URL nodes by click edges. Given a query node q and a URL node u, there will be an edge (q, u) if u is among the clicked answers for query q. They have recorded more than 10 percent improvement over the baseline.

Random walk models are used to learn associations by combining evidence from various lexical sources like WordNet.

4. RELEVANCE FEEDBACK BASED EXPANSION

Relevance feedback based methods execute the initial query on collection and extract top k documents. Then the ranked document are used to improve the performance of retrieval. It is assumed that the initial retrieved documents are relevant and thus can be used to extract expansion terms. These models fail when the initial retrieval algorithm of search engine is poor. These models can be classified into following types :

4.1 Explicit feedback from user

This an interactive approach in which the initial retrieved documents are presented to user and the user is asked to select the relevant documents. These models are not much useful because users expect the system to be autonomous and retrieve the results for the user. User would ultimately get irritated by repeated interaction required from him for each search. These type of models can be used for testing search engines where developers are willing to interact with the system.

4.2 Implicit feedback

This is a type of model in which the user's feedback is inferred by the system. The feedback can be inferred from user's behavior like : The pages which user opens for reading, or pages on which user clicks once the results are displayed back to the user

4.3 Pseudo Relevance Feedback (PRF)

In Pseudo Relevance Feedback based models initial query is fired and top k results are obtained. Then important terms, mostly based on co-occurrence, from these documents are

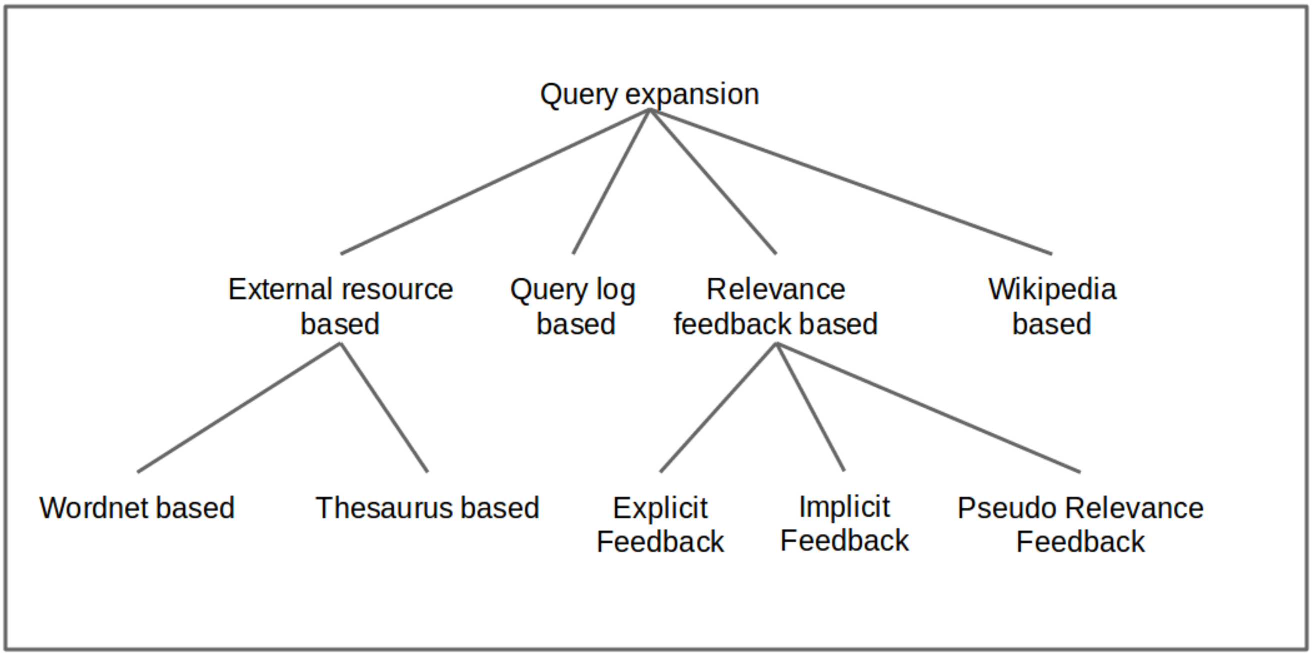

Figure 2: Query expansion techniques

extracted and added to query. Then this expanded query is re fired to retrieve final set of documents which are made available to the user. The relevancy of expansion terms depends upon the initial retrieved documents.

Pseudo relevance feedback captures the important terms only based on co-occurrence. But only co-occurrence is not enough for correctness of results. We need to consider semantic and lexical properties of word. Also as the feedback is independent of the user there is a chance of topic drift. Figure 3 explains a typical work flow of how Pseudo Relevance Feedback based information retrieval systems work. Pseudo Relevance Feedback also called as Blind Feedback automates the manual part of relevance feedback and has the advantage that assessors are not required. [24]

Pseudo Relevance Feedback has been successfully applied in various IR frameworks and has been proved to improve precision and recall of search engines.

Croft and Harper [1979] [8] first suggested this technique for estimating the probabilities within the probabilistic model. However, they also highlighted one fundamental problem - topic drift which many of Pseudo Relevance Feedback based system based especially if the initial retrieval is poor. Topic drift is caused as a result of adding terms which have no association with the topic of relevance of the query. Topic drift can be formally defined as:

"Tendency of a search to drift away from the original subject of discussion (and thus, from the query), or the results of that tendency". Several approaches have been proposed to improve *PRF* by -

- refining relevant document set [16, 17]
- refining the expansion terms from PRF [4]
- using selective query expansion [5, 9]
- varying the importance of documents [19]

Lv and Zhai [2010] [14] proposed a positional relevance model where the terms in the document which are nearer to the query terms are assigned more weight. This works on basic intuition that nearer the word to query term, more relevant it is.

Chinnakotla [2010] [6] suggest use of an assisting language to improve the performance retrieval of search engine. They translate the query to an assisting language and perform Pseudo Relevance Feedback twice, once for in query language and other in assisting language. Then they merge the expansion terms obtained from both the Pseudo Relevance Feedback instances using translation and retrieve the documents for expanded query. The assisting language is chosen such that the mono lingual performance of the assisting language should be high. Multilingual Pseudo Relevance Feedback was performed using English assisting language for French, German, Hungarian and Finnish languages. The results obtained by Chinnakotla [2010] [6, 7] are very promising and intuit us with a thought to extend these concepts for Indian languages.

Atreya [2013] [2] suggests a very promising approach which takes into account the structure of the documents while assigning priorities to the expansion terms. The intuition behind the idea is that a term that occurs in title section of a document is more important for that document than the

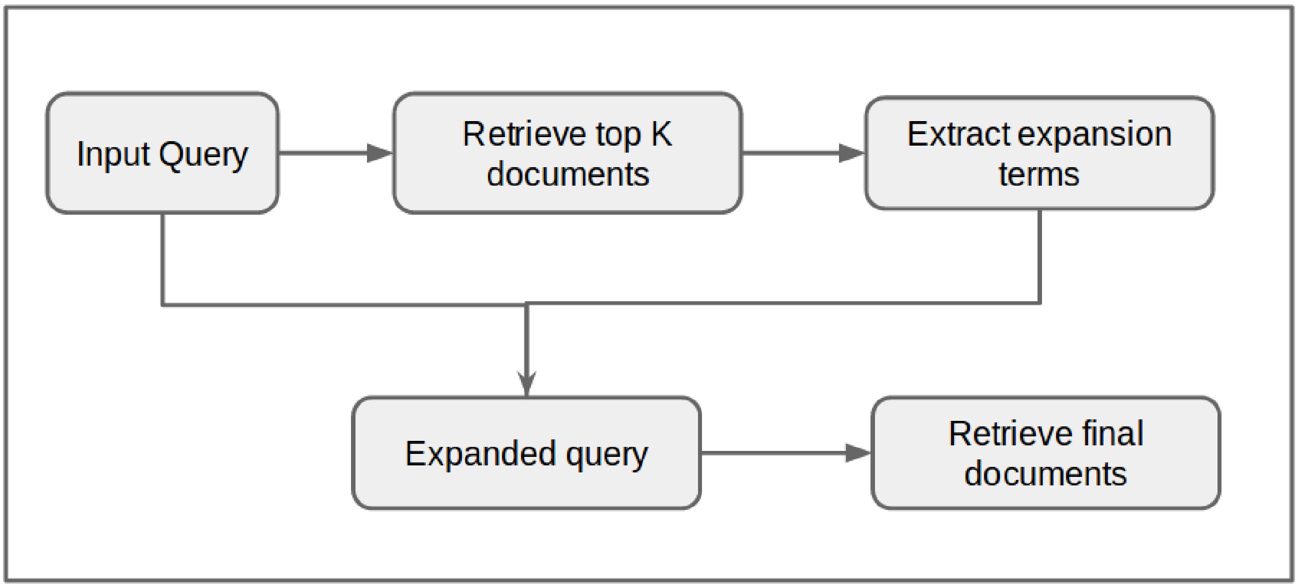

Figure 3: Architecture of Pseudo Relevance Feedback based system

term which occurs in the body. Title more compactly represents the entire document and thus it is very less probable that title will have a word which is unrelated to the document. The results were performed on Wikipedia document collection which divides the document into 4 components viz. Title, Body, Infobox and Categories. The results thus obtained are very promising and improve the precision and recall values by few percentage.

5. QUERY EXPANSION USING WIKIPEDIA

Wikipedia is the biggest encyclopedia available freely on the web. Wikipedia content is well structured and correct. Li et al. Many works suggest use of Wikipedia as source for query expansion[1, 11, 20, 21, 22]. [2007] [13] proposed query expansion using Wikipedia by utilizing the category assignments of its articles. The base query is run against a Wikipedia collection and each category is assigned a weight proportional to the number of top-ranked articles assigned to it. Articles are then re-ranked based on the sum of the weights of the categories to which each belongs.

Milne et al. [2007b] [15] a search interface which offers a domain-independent knowledge-based information retrieval is developed using Wikipedia structure for query expansion. They build there thesaurus based on Wikipedia articles. But due to large size of Wikipedia corpus irrelevant results were also retrieved.

Kaptein and Kamps [2009] [12] used Wikipedia link and category information to expand the query. Category information is used by calculating distances between document categories and target categories. Observation was that category information has more value than link information.

6. WHY PSEUDO RELEVANCE FEEDBACK ?

- Pseudo relevance feedback based methods are independent of any external lexical resource and thus a unknown word does not hinder the retrieval effectiveness of an Information Retrieval system.
- Large number of experiments and case studies give an indication that Pseudo relevance feedback systems work far better than the traditional external resource based system.
- External resources like WordNet, thesaurus may not be available for some languages in which case Pseudo Relevance feedback would seem the only option for query expansion.
- Pseudo Relevance Feedback extracts lexically and semantically related terms to query and thus documents which talk about query but do not contain query terms can be retrieved.
- Especially if the search engine is built for a resource scarce language like Hindi, Marathi, etc. it is possible that some of the user queries would result in very few documents retrieved.
- It is important the we provide documents which are somewhat relevant to user query. Pseudo relevance comes in handy for such queries.

For above mentioned reasons, we select Pseudo Relevance Feedback as our baseline and expand upon it.

7. REFERENCES

[1] B. Al-Shboul and S.-H. Myaeng. Query phrase expansion using wikipedia in patent class search. In *AIRS*, pages 115–126, 2011.

[2] A. Atreya, Y. Kakde, P. Bhattacharyya, and G. Ramakrishnan. Structure cognizant pseudo relevance feedback. In *Proceedings of IJCNLP 2013, Nagoya, Japan*, pages 982–986, 2013.

[3] O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[4] G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250. ACM, 2008.

[5] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.

[6] M. K. Chinnakotla, K. Raman, and P. Bhattacharyya. Multilingual prf: english lends a helping hand. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 659–666. ACM, 2010.

[7] M. K. Chinnakotla, K. Raman, and P. Bhattacharyya. Multilingual pseudo-relevance feedback: performance study of assisting languages. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1346–1356. Association for Computational Linguistics, 2010.

[8] W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance information. *Journal of documentation*, 35(4):285–295, 1979.

[9] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. A framework for selective query expansion. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 236–237. ACM, 2004.

[10] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *Proceedings of the 11th international conference on World Wide Web*, pages 325–332. ACM, 2002.

[11] S. Ganesh and V. Verma. Exploiting structure and content of wikipedia for query expansion in the context. In *International Conference RANLP*, pages 103–106, 2009.

[12] R. Kaptein and J. Kamps. Finding entities in wikipedia using links and categories. In *Advances in Focused Retrieval*, pages 273–279. Springer, 2009.

[13] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using wikipedia asexternal corpus. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 797–798. ACM, 2007.

[14] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 579–586. ACM, 2010.

[15] D. N. Milne, I. H. Witten, and D. M. Nichols. A knowledge-based search engine powered by wikipedia. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 445–454. ACM, 2007.

[16] 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*, pages 206–214. ACM, 1998.

[17] T. Sakai, T. Manabe, and M. Koyama. Flexible pseudo-relevance feedback via selective sampling. *ACM Transactions on Asian Language Information Processing (TALIP)*, 4(2):111–135, 2005.

[18] A. F. Smeaton, F. Kelledy, and R. O'Donnell. Trec-4 experiments at dublin city university: Thresholding posting lists, query expansion with wordnet and pos tagging of spanish. *Harman [6]*, pages 373–389, 1995.

[19] T. Tao and C. Zhai. Regularized estimation of mixture models for robust pseudo-relevance feedback. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 162–169. ACM, 2006.

[20] E. M. Voorhees. The trec robust retrieval track. In *ACM SIGIR Forum*, volume 39, pages 11–20. ACM, 2005.

[21] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 59–66. ACM, 2009.

[22] Y. Xu, G. J. Jones, and B. Wang. Query dependent pseudo-relevance feedback based on wikipedia. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '09, pages 59–66, New York, NY, USA, 2009. ACM.

[23] Z. Yin, M. Shokouhi, and N. Craswell. Query expansion using external evidence. In *Advances in Information Retrieval*, pages 362–374. Springer, 2009.

[24] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410. ACM, 2001.