Query Expansion using Artificial Relevance Feedback

Sandeep Joshi
Associate Professor
Department of Computer Science
SGI, Sikar, Rajasthan, India

Satpal Singh Kushwaha
Research Scholar
Department of Computer Science
SGI, Sikar, Rajasthan, India

ABSTRACT
World Wide Web is growing rapidly so with this rapid expansion in the size of web, Information extraction on web is achieving its importance day by day. The user's query[1] plays a crucial role in the information retrieval process. So for the better information retrieval[2] results several methods have been proposed which help the user in the query expansion task. Some methods use thesaurus for the query expansion purpose. Thesaurus is nothing but a list of synonyms. Latest techniques for query expansion are mining user logs and creating user profiles. In the proposed system we present query expansion using Artificial Relevance Feedback Mechanism. The proposed system provides a simple way of query expansion based on Artificial Relevance Feedback.

General Terms
Artificial Relevance Feedback, Query Expansion.

Keywords
Thesauri, clustering, lexical co-occurrence, relevance feedback.

1. INTRODUCTION
1960, Maron and Kuhn's[3] mentioned query alteration first time by signifying that terms closely associated to the original query terms can be added to the query for the retrieval of relevant documents.

Research by Spink et al. [4] has revealed that most search engine users normally devise very small queries of two to three words. Such small queries do not have many useful words and do not adequately illustrate the subject that the user wants to search on. In the same work it is recommended that web users tend to go more often from wide to narrow formulation in queries since the most common query alteration is to append terms. The aim of query expansion is to recommend potential terms to be added to the user's original query, so that the quality of the retrieved results can be enhanced.

We have proposed a system that uses the concept of artificial relevance feedback mechanism for query expansion. By this technique the query expansion terms are extracted directly from the internet. Experiments carried out with human users have shown that the terms proposed for expansion by our system, can considerably shorten the time of a web search. These experiments together with the results they produced are presented in Section 4.

2. QUERY EXPANSION
A number of techniques for performing query expansion have been developed. An broad review on the subject has been produced by Efthimiadis [5]. Early methods involved extracting terms from thesauri [6, 7] but as these proved to be labor demanding, researchers moved to methods like lexical co-occurrence [8] and clustering [9]. Lexical co-occurrence is the procedure of developing associations between words based upon their co-occurrence in documents. In clustering, documents that share a large number of terms are grouped together and candidate terms from each cluster are used for the expansion of the original query.

Most of the systems which uses clustering for the intention of query expansion reported quite unenthusiastic results on their performance. The connection of the method proposed here with the methods of lexical co-occurrence and clustering is that the source, which supply the candidate terms for the query expansion is the set of the extracted documents as compare to some knowledge composition, as is the case with the thesaurus based approaches. As a result, if the user select terms that do not acquiesce results from the probable domain, then these selected terms will not be helpful to the user for the query expansion purpose.

So for handling these type of conditions without make use of previous information is too not easy and will not be able to deal with in this effort. See Ref. [10] for a analysis of knowledge-based techniques for query expansion. In the proposed system it is supposed that the user provides an initial search string to the system, which is quite common and acquiesce results that have the required class of credentials as a subset.

The other method to perform query expansion, perhaps the most effective of all, is that of relevance feedback [11]. In this method the user submits a query, which capitulate an initial set of results. From this set she selects a number of documents supposed to be relevant. The system expands the query based upon the terms in the selected documents. Regardless of the major improvement in the quality of results this method produces, the research carried out by Spink, shows low use of this type of relevance feedback techniques provided in search engines [12]. The little use should not essentially be credited to the interactive nature of this technique. Occasionally when a user has already found a set of appropriate results, then they may not desire to expand the query more.

In the proposed system we propose a new type of relevance feedback technique that is known as Artificial Relevance feedback technique. In this approach we assume that the top returned documents are relevant document for the purpose of

extracting terms for query expansion. This method of terms extraction is known as Artificial Relevance feedback Mechanism. We have described this process in the next sections.
There are three approaches for the user query expansion[13]

a. Manual Query Expansion
b. Interactive Query Expansion
c. Automatic Query Expansion

Manual and interactive query expansion requires user's involvement in the query expansion process. In the proposed system we uses the approach of automatic query expansion which is the process of supplementing additional term or phrases to the original query to improve the retrieval effectiveness without users involvement.

Several methods for query expansion have been developed and used. Some methods get relevant terms from thesaurus[14], but the creation and the maintenance of the thesaurus is a laborious work. Other methods are using Latent semantic analysis where user query terms are semantically analyzed and relationship in between them is used to expand the user query but that approach also uses large dictionaries.

Recent methods for query expansion are mining user logs[15] and construction of user profile[16] .Web based query expansion is the process where the information terms to the user query are retrieved from web directly by sending the user query to the general purpose search engine then the unique or distinct terms[17] with high information are retrieved from the returned results. Because we are using automatic[18] query expansion so the top three unique words with high frequency are automatically selected to expand the user query.

3. PROPOSED QUERY EXPANSION ALGORITHM

The proposed algorithm proceeds as follows:

a. The user submits the query to the proposed system.
b. Sending the user query to a general purpose search engine (AltaVista in our case).
c. Extraction of the distinct words from the results returned from search engine.
d. Formulation of the new queries using the top five terms (based on their occurring frequency) extracted in the last step.
e. Sending the new formulated user queries to a general purpose search engine (AltaVista in our case).
f. Extraction of the distinct words from the results returned from search engine.
g. Calculating the frequency of the extracted distinct words.
h. The unique words are then arranged in the decreasing order of their frequencies.
i. At last top three words are chooses for the query expansion.

The query expansion process consists of searching for the relevant terms by sending the user query and the sub-expanded user query to the general-purpose search engine. This procedure is fairly straightforward, and can be described as follows:

Getting relevant terms from the Web dynamically

As was mentioned earlier, the proposed system does not use special dictionaries for query expansion, but uses existing Web[19] documents to search for relevant terms. The user query is submitted to the general-purpose search engine twice: the first time by itself and the second time in combination with terms supposed to be appropriate to the query. Using the Web as a source for relevant terms makes the system more dynamic and up-to-date, reflecting which topics are more accepted at the present time.

Artificial Relevance Feedback Mechanism

In Relevance Feedback Mechanism the relevant terms are extracted from the relevant documents returned from the web. The judgment of the relevancy is done by applying some algorithms like ranking[20] algorithms which are used to rank the web pages extracted from the web. But in the proposed system we are assuming that the top returned web pages are relevant web pages to extract the query expansion terms without applying any ranking algorithms. Such type of automatic query refinement is called Automatic Relevance Feedback Mechanism and provides good results.

The key advantage of the query expansion process implemented in the proposed system is the capability to obtain terms relevant to a query from the Web directly. This makes the system to recognize terms covering large range of queries without maintaining a huge dictionary[21] or database of terms in a large variety of fields.

4. EXPERIMENTS

In order to assess the performance of the system, a set of experiments has been conducted. The query expansion algorithm is very complex and an explanation of its process would greatly help in understanding it. The outline of the algorithm has been described in the topic 3. In this experiment, the original query "travel" will be using the query expansion algorithm. All the examples are simplified for illustrative purpose. Assume the initial query is

"Query = travel"

Step 1. By using AltaVista a general-purpose search engine to get documents resulting from the original query. This step uses PHP to fetch result documents for the query "travel".

Step 2. In this step distinct terms are extracted from the web documents returned from AltaVista for the query "travel". For this purpose the whole documents text is not analyzed only heading, titles and snippets are checked to get relevant terms. The results can be shown as follows, which are stored in an array.

Array ([1] => fortravel [2] => trending [3] => earchesamtrak [4] => train [5] => traveltravel [6] => sitestravel [7] => trailerstravel [8] => dealsthailand [9] => travelsearch [10] => allyahoo [11] => travelexpediaorbitzkayakaol [12] => travelshopping [13] => sitesrelated [14] => conceptsrelated [15] => searchessouthwest [16] => cad9fa [17] => ee [18] => space [19] => td [20] => plain [21] => ls [22] => container [23] => min [24] => ri [25] => space [26] => visibility [27] => frm [28] => vidmsg [29] => cepx [30] => spnslist [31] => travel [32] => travel [33] => channel [34] => travel [35] => insurance [36] => resultsyahoo [37] => travel [38] => cheap [39] => flights [40] => hotels [41] => travel [42] => guides [43] => car [44] => rentalfind [45] => cheap [46] => flights [47] => hotels [48] => car [49] => rentals [50] => travel [51] => plan [52] => trip [53]

=> travel [54] => guides [55] => personalized [56] => recommendations [57] => articles [58] => deals [59] => cachedexpedia [60] => travel [61] => vacations [62] => cheap [63] => flights [64] => airline [65] => tickets [66] => plan [67] => trip [68] => expedia [69] => buy [70] => airline [71] => tickets [72] => read [73] => reviews [74] => reserve [75] => hotel [76] => deals [77] => vacations [78] => rental [79] => cars [80] => cruises [81] => great [82] => prices [83] => guaranteed [84] => cachedorbitz [85] => travel [86] => airline [87] => tickets [88] => cheap [89] => hotels [90] => car [91] => rentals [92] => book [93] => cheap [94] => airline [95] => tickets [96] => hotel [97] => reservations [98] => car [99] => rentals [100] => vacations [101] => travel [102] => deals [103] => orbitz [104] => cheapest [105] => airfare [106] => hotel [107] => deals [108] => cash [109] => refund [110] => cachedcheap [111] => flights [112] => cheap [113] => hotels [114] => price [115] => predictor [116] => bing [117] => travelfind [118] => cheap [119] => flights [120] => price [121] => predictor [122] => cheap [123] => hotels [124] => great [125] => travel [126] => deals [127] => receive [128] => fare [129] => predictions [130] => bing [131] => travel [132] => travel [133] => worldwide [134] => travel [135] => comparison [136] => shop [137] => prices [138] => airfares [139] => hotels [140] => cruises [141] => car [142] => rentals [143] => cachedkayak [144] => cheap [145] => flights [146] => hotels [147] => airline [148] => tickets [149] => cheap [150] => tickets [151] => book [152] => cheap [153] => flights [154] => hotels [155] => vacations [156] => rental [157] => cars [158] => hotel [159] => flight [160] => travel [161] => deals [162] => hundreds [163] => travel [164] => sites [165] => travel [166] => summer [167] => holiday [168] => ideas [169] => travel [170] => guides [171] => worldwide [172] => travel [173] => inspiration [174] => guides [175] => holiday [176] => features [177] => vacation [178] => packages [179] => msn [180] => cachedtravel [181] => guides [182] => information [183] => aol [184] => travelfind [185] => travel [186] => information [187] => vacation [188] => packages [189] => airline [190] => tickets [191] => hotels [192] => aol [193] => travel [194] => online [195] => destination [196] => vacation [197] => travel [198]

Fig 1: Extracted distinct terms from the returned web documents for query "travel"

[cheap] => 21 [flights] => 14 [tickets] => 13 [hotels] => 13 [vacations] => 12 [deals] => 12 [airline] => 11 [vacation] => 8 [packages] => 8 [hotel] => 5 [discount] => 5 [rentals] => 5 [car] => 5 [book] => 5 [airfare] => 4 [cruises] => 4 [rental] => 3 [cars] => 3 [great] => 2 [family] => 2 [expedia] => 2 [trips] => 2 [stop] => 2

Fig 2: Top 23 words with frequency greater than or equal to 2.

Step 3: Top five (based on occurring frequency) distinct words extracted from the results of step 2 are as follows:

"hotel, cheap, flight, ticket, vacation"

Step 4: Formulation of the five sub queries using words extracted in the **step 3**:

Table 1. Formulated five sub queries.

S.NO	Formulated Sub queries
1	travel hotel
2	travel cheap
3	travel flight
4	travel ticket
5	travel vacation

Step 5: The sub queries are again send to the general purpose search engine (AltaVista in our case) and distinct terms are extracted again which are shown in the following figure.

Array ([0] => inhelpmailyahoo [1] => direct [2] => resultswebfilter [3] => timeanytimepast [4] => daypast [5] => weekpast [6] => min [7] => ri [8] => space [9] => visibility [10] => frm [11] => vidmsg [12] => cepx [13] => spnslist [14] => city [15] => travel [16] => hotel [17] => baguio [18] => travel [19] => hotel [20] => reviews [21] => resultsexpedia [22] => travel [23] => vacations [24] => cheap [25] => flights [26] => airline [27] => tickets [28] => plan [29] => trip [30] => expedia [31] => buy [32] => airline [33] => tickets [34] => read [35] => reviews [36] => reserve [37] => hotel [38] => deals [39] => vacations [40] => rental [41] => cars [42] => cruises [43] => great [44] => prices [45] => guaranteed [46] => cachedmore [47] => hotel [48] => travel [49] => networkdiscount [50] => online [51] => hotel [52] => reservations [53] => guides [54] => interactive [55] => travel [56] => forums [57] => travel [58] => tips [59] => cachedtravelocity [60] => travel [61] => cheap [62] => airline [63] => tickets [64] => hotels [65] => flights [66] => book [67] => travel [68] => specials [69] => cheap [70] => airline [71] => tickets [72] => hotels [73] => cruises [74] => car [75] => rentals [76] => flights [77] => travelocity [78] => stop [79] => resource [80] => travel [81] => vacation [82] => cachedorbitz [83] => travel [84] => airline [85] => tickets [86] => cheap [87] => hotels [88] => car [89] => rentals [90] => book [91] => cheap [92] => airline [93] => tickets [94] => hotel [95] => reservations [96] => car [97] => rentals [98] => vacations [99] => travel [100] => deals [101] => orbitz [102] => cheapest [103] => airfare [104] => hotel [105] => deals [106] => cash [107] => refund [108] => cachedmore [109] =>

reviews [110] => hotels [111] => flights [112] => vacation [113] => rentals [114] => tripadvisortripadvisor [115] => unbiased [116] => hotel [117] => reviews [118] => photos [119] => travel [120] => advice [121] => hotels [122] => vacations [123] => compare [124] => prices [125] => cachedmore [126] => cheap [127] => discount [128] => travel [129] => deals [130] => flights [131] => hotels [132] => cheap [133] => discount [134] => travel [135] => deals [136] => flights [137] => hotels [138] => meet [139] => budget [140] => book [141] => travel [142] => deals [143] => flights [144] => hotels [145] => vacation [146] => packages [147] => cachedmore [148] => travel [149] => cheap [150] => flights [151] => hotels [152] => travel [153] => guides [154] => car [155] => rentalfind [156] => cheap [157] => flights [158] => hotels [159] => car [160] => rentals [161] => travel [162] => plan [163] => trip [164] => travel [165] => guides [166] => personalized [167] => recommendations [168] => articles [169] => deals [170] => cachedmore [171] => hotels [172] => cheap [173] => hotel [174] => deals [175] => bing [176] => travelsearch [177] => hotels [178] => cheap [179] => hotel [180] => deals [181] => stay [182] => rate [183] => indicator [184] => bing [185] => travel [186] => travel [187] => hotelshotels [188] => travel [189] => net [190] => hotel [191] => hotels [192] => discount [193] => hotels [194] => travel [195] => related [196] => services [197] => complete [198] => comprehensive [199] => date [200] => hotel [201] => directory [202] => cachedvacation [203] => packages [204] => cheap [205] => vacations [206] => deals [207] => cheap [208] => trips [209] => destinations [210] => worldwide [211] => great [212] => deals [213] => vacations [214] => holidays [215] => ll [216] => easy [217] => reserve [218] => flight [219] => hotel [220] => rental [221] => car [222] => vacations [223] => cachedpromotional [224] => resultsmake [225] => homepageeverything [226] => query [227] => legal [228] => site)

Fig 3: Extracted distinct terms from the returned web documents for five sub queries.

Step 6: From the words extracted from the **step 5** top three high occurrence words are chooses to expand our original query, which are as follows:

"travel, flight, ticket"

So our final expanded query looks like as follows:

a) travel cheap
b) travel flight
c) travel ticket

Some more experimental results for query expansion process are shown in the following table.

Table 2. Results of the query expansion algorithm

Original Query	Suggested Terms for Query Expansion		
	I Terms	II Terms	III Terms
food	recipes	cooking	list
map	driving	direction	road
law	legal	information	sociology

Comparison with existing Search Systems

In the following table we have shown comparison results of the query expansion process of the proposed system with the existing search systems. In the following table we have used some general purpose search engine as well as some Meta search engines. We have choosed top three distinct terms suggested for query expansion purpose to the user.

Table 3. Query Expansion results comparison of the Proposed System with the Existing Search system

Original Query	Search System	Suggested Terms for Query Expansion		
		I Term	II Term	III Term
food	Proposed System	recipes	cooking	list
	Google	Corporation of India	Food	Security bill
	Yahoo	recipes	Corporation of India	inflation
	Mamma.com	Network	Network recipes	lion
	Dogpile	Network	Network recipes	lion
	MetaCrawler	Network	Network recipes	lion
Map	Proposed System	Driving	directions	road
	Google	Of India	Myindia	Of Delhi
	Yahoo	Of India	Google Quest	Google
	Mamma.com	Quest	Quest Driving directions	Quest directions
	Dogpile	Quest	Quest Driving directions	Quest directions
	MetaCrawler	Quest	Quest Driving directions	Quest directions
java	Proposed System	Oracle	Database	programming
	Google	Download	Tutorial	Interview questions

	Yahoo	Download	Script	games
	Mamma.com	Downloads	Chat rooms	updates
	Dogpile	Downloads	Chat rooms	updates
	MetaCrawler	Downloads	Chat rooms	updates

5. CONCLUSION

In this paper we have presented an interactive method for query expansion based Artificial Relevance Feedback Mechanism. The method is based on the fact that documents contain some terms with high information content, which summarize their subject matter. These terms are extracted from the collection of the retrieved documents. The top ranked terms are presented to the user as a list of candidate terms for expansion.
Comparison results with the other Search Systems show that proposed system suggest suitable expansion terms. We have tested the system by submitting query 42 times. The experimental results show that the proposed system suggests appropriate terms for query expansion 83.33% of the time.

6. ACKNOWLEDGMENTS

It gives us immense pleasure to express our sincere gratitude towards Dr. S. C. Jain, who is working as a professor at University College of Engineering at Rajasthan Technical University, Kota (Rajasthan), India for his invaluable guidance. In spite of his hectic schedule he was always approachable and spared his time to attend our problems and give us appropriate advice.

7. REFERENCES

[1] Lu Y., Meng W., Shu L., Yu C., and Liu K. Evaluation of result merging strategies for metasearch engines. WISE Conference, New York, NY, 2005, pp. 53–66.

[2] CROFT, W. B. 1983. "Experiments with Representation in a Document Retrieval System." Information Technology: Research and Development, 2(1), 1-21.

[3] MARON, M. E., and J. L. KUHNS. 1960. "On Relevance, Probabilistic Indexing and Information Retrieval." *Association for Computing Machinery*, 7(3), 216-44.

[4] A. Spink, D.Wolfram, B.J. Jansen, T. Saracevic, Searching the web: the public and their queries, J. Am. Soc. Inform. Sci. Technol. 52 (3) (2001) 226–234.

[5] E.N. Efthimiadis, Query expansion, Annu. Rev. Inform.
Syst. Technol. 31 (1996) 121–187.

[6] S. Gauch, J.B. Smith, Search improvement via automatic
query reformulation, ACM Trans. Inform. Syst. 9 (3) (1991) 249–280.

[7] E.M. Voorhees, Using WordNet to disambiguate word senses for text retrieval, in: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1993, pp. 171–180.

[8] O. Vechtomova, S. Robertson, S. Jones, Query expansion with long-span collocates, Inform. Retriev. 6 (2) (2003) 251–273.

[9] K.S. Jones, Automatic Keyword Classification for Information Retrieval, Butterworth, London, UK, 1971.

[10] R.C. Bodner, F. Song, Knowledge-based approaches to query expansion in information retrieval, in: Proceedings of the Canadian Conference on AI, 1996, pp. 146–158.

[11] S.E. Robertson, C.L. Thompson, M.J. Macaskill, J.D. Bovey, Weighting, ranking and relevance feedback in a front-end system, J. Inform. Sci. 12 (1–2) (1986) 71–75.

[12] Wen-Chen Hu, "An overview of the World Wide Web search technologies," In the proceedings of 5th World Multi-conference on System, Cybernetics and Informatics, SCI2001, Orlando, Florida, July 22-25, 2001.

[13] HARMAN, D. 1988. "Towards Interactive Query Expansion." Paper presented at ACM Conference on Research and Development in Information Retrieval, Grenoble, France.

[14] Harter, Stephen P, "Online Information Retrieval: Concepts, Principles, and Techniques",Orlando: Academic Press, 1986.

[15] S. Gauch, J.B. Smith, Search improvement via automatic query reformulation, ACM Trans. Inform. Syst. 9 (3) (1991) 249–280.

[16]. M. Mitra, A. Singhal, C. Buckley, Improving automatic query expansion, in: Proceedings of the 21st Annual InternationalACM SIGIR Conference on Research and Development in Information Retrieval, ACM Press, 1998, pp. 206–214.

[17] C. Carpineto, G. Romano, and V. Giannini, "Improving retrieval feedback with multiple term-ranking function combination". TOIS 20(3), 2002, pp. 259-290.

[18] SPARCK JONES, K., and E. O. BARBER. 1971. "What Makes an Automatic Keyword Classification Effective." J. American Society for Information Science, 22(3), 166-75.

[19] Rahardjo, B. and Yap, R. Automatic Information Extraction from Web Pages, SIGIR, 2001, 430-431.

[120] Kamps, J. (2004). Improving retrieval effectiveness by reranking documents based on controlled vocabulary. ECIR 2004, 283–295.

[21] Keliang JIA "Query Expansion Based on Word Sense Disambiguation in Chinese Question Answering System" Journal of Computational Information Systems6:1(2010) 181-187