An Efficient Approach For Information Retrieval Based On Relevance Feedback Algorithm

Dipalee S. Hirde
Student of M.E, Department of Computer Science & Engineering,
HVPM's College of engineering & technology, Amravati.

Prof. R. R. Keole
Asst. Professor , Department of Information Technology & Engineering,
HVPM's College of engineering & technology, Amravati.

ABSTRACT: Information Retrieval (IR) is concerned with indexing and retrieving documents including information relevant to a user‘s information need. Relevance Feedback (RF) is a class of effective algorithms for improving Information Retrieval (IR) and it consists of gathering further data representing the user‘s information need and automatically creating a new query. Relevance Feedback consists in automatically formulating a new query according to the relevance judgments provided by the user after evaluating a set of retrieved documents. Finding relevant document is one of the hard tasks. we propose a class of RF algorithms inspired by quantum detection to re-weight the query terms and to re-rank the document retrieved by an IR system. Information retrieval (IR) is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Automated information retrieval systems are used to reduce what has been called "information overload". Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown and IR system return relevant document to the user. The process may then be iterated if the user wishes to refine the query.

KEYWORDS: Information retrieval, quantum mechanics, relevance feedback, quantum detection.

I. INTRODUCTION

Information retrieval (IR) has experienced huge growth in the past decade as increasing numbers and types of information systems are being developed for end-users. The incorporation of users into IR system evaluation and the study of users information search behaviours and interactions have been identified as important concerns for IR researchers .The proposition that IR systems are fundamentally interactive and should be evaluated from the perspective of users is not new. IR system return the information to the user need. First user enter a query into IR system. IR system return documents related to query. Suppose query is too long IR system can't understand which type of document user need. So IR system return only relevant and irrelevant documents. If user want only relevant document, Then RF (Relevance Feedback) algorithm is use, RF algorithm is a class of effective algorithm for improving IR system and modify the query by reducing the term found in irrelevant document. It consist of gathering further data representing the user information need and automatically creating new query. IR system return only relevant document that user need and top rank the document. Suppose user enter a query but there is no document related to query text. Then pattern matching algorithm is use, after using Pattern matching algorithm IR system match the query text with the content present inside the document and IR system return the relevant document that user want.

An IR system addresses the problems caused by query ambiguity by gathering additional evidence that can be used to automatically modify the query . Usually a query is expanded because the queries are short and it cannot exhaustively describe every aspect of the user's information need; however, some irrelevant documents may be retrieved or relevant documents may also be missed when a query is not short .The automatic procedure that modify the user's queries is known as Relevance Feedback (RF); some relevance assessments about the retrieved documents are collected and the query is expanded by the terms found in the relevant documents, reduced by the terms found in the irrelevant documents or reweighted using relevant or irrelevant documents.

Relevance feedback (RF) is the retrieval task where the system is given not only a user query, but also user feedback on some of the top ranked results. Feedback gives the retrieval system a chance to improve its results by exploiting the extra information through more elaborate techniques. This can be helpful in cases where the users want as many relevant results as possible. RF is one of the most useful Query Modification techniques in the field of Information Retrieval (IR). This method is put into practice when the user needs to improve the query formulated to the IR system, because the documents initially retrieved do not completely fulfill the user's information need. Relevance feedback works in the following way: a user submits a query representing his/her information need to the IR system, which then ranks the documents according to their corresponding degrees of relevance to the query (with the documents most closely matching the query ranked first). The user then inspects this list,1 and determines which documents are relevant and which are not relevant to his/her information need (the relevance judgments). Using this information, the IR system updates the initial query, modifying the importance of the terms it contains 2 (term reweighting), and adding new terms that are considered useful to retrieve more relevant documents (query expansion). This process is repeated until the user is completely satisfied with the set of retrieved relevant documents. Relevance feedback has been successfully applied in a great variety of IR models.

RF can be positive, negative or both. Positive RF only brings relevant documents into play and negative RF makes only use of irrelevant documents; any effective RF algorithms includes a "positive" component. Although positive feedback is a well established technique by now, negative feedback is still problematic and requires further investigation, yet some proposals have already been made such as grouping irrelevant documents before using them for reducing the query.

Some of the first types of IR interactions were associated with relevance feedback. Looking closely at this seemingly simple type of interaction, we see the difficulties inherent in Interactive Information Retrieval IIR studies. Assuming that users are provided with information needs, each user is likely to enter a different query, which will lead to different search results and different opportunities for relevance feedback. Each user, in turn, will provide different amounts of feedback, which will create new lists of search results. Furthermore, causes and consequences of these interactions cannot be observed easily since much of this exists in the user's head. The actions that are available for observation querying, saving a document, providing relevance feedback are surrogates of cognitive activities. From such observable behaviours we must infer cognitive.

II. LITERATURE REVIEW

Efthimiadis and Biron. [1] found in their experiments that standard RF techniques used in pseudo RF experiments performed only slightly poorer than experiments using RF from users and with no feedback. However, the actual performance varied according to the algorithm used to rank terms for query expansion.

J. H. Lee [2] , proposed an ad-hoc RF technique based on multiple RF techniques. The basic hypothesis is that, as different RF techniques may produce different modified queries, and different queries will retrieve different documents, then using a combination of RF techniques to modify queries will retrieve more of the relevant documents. An initial experiment was carried out treating the top 30 documents as relevant and using a vector-space retrieval function.

S.E. Robertson [3] found increased performance over no feedback, especially when using passages rather than the whole document, to select expansion terms

P. Vakkari, [4,5] examined long-running searches to examine how relevance assessments changed over time. In his study he demonstrated that not only did subjects chose different documents at different stages in their task, they also used different search tactics and strategies38. Vakkari provided support for Spink's observation that high numbers of partial assessments correlates with a lack of ability to discriminate relevant and non-relevant. This may occur at the start of a search, for example. He also found evidence to indicate that when a user has a good idea of what constitutes relevant material, he is less likely to make a high number of relevance assessments.

Xuanhui Wang [6] we conduct a systematic study of methods for negative relevance feedback. We compare a set of representative negative feedback methods, covering vector-space models and language models, as well as several special heuristics for negative feedback. Evaluating negative feedback methods requires a test set with sufficient difficult topics, but there are not many naturally difficult topics in the existing test collections. We use two sampling strategies to adapt a test collection with easy topics to evaluate negative feedback. Experiment results on several TREC collections show that language model based negative feedback methods are generally more effective than those based on vector-space models, and using multiple negative models is an effective heuristic for negative feedback. Our results also show that it is feasible to adapt test collections with easy topics for evaluating negative feedback methods through sampling.

Yuanhua [7] we propose a novel learning algorithm, Feedback Boost, based on the boosting framework to improve pseudo feedback. A major contribution of our work is to op- timize pseudo feedback based on a novel loss function that directly measures both robustness and effectiveness. we compare Feedback Boost with a well-performing learning to rank approach applied for pseudo feedback and observe that Feedback Boost works clearly better. These results show that the proposed Feedback Boost is more effective and robust than any of the existing method for pseudo feedback, including both traditional pseudo feed- back methods and new learning-based approaches.

Jun Miao [8] we study how to incorporate proximity in- formation into the Rocchio's model, and propose a proximity- based Rocchio's model, called PRoc, with three variants. In our PRoc models, a new concept (proximity-based term frequency, ptf) is introduced to model the proximity information in the pseudo relevant documents, which is then used in three kinds of proximity measures. Experimental results on TREC collections show that our proposed PRoc models are effective and generally superior to the state-of-the-art relevance feedback models with optimal parameters. A direct comparison with positional relevance model (PRM) on the GOV2 collection also indicates our proposed model is at least competitive to the most recent progress.

E.M. Voorhees [9], The study of difficult queries has attracted much attention recently, partly due to the launching of the ROBUST track in the TREC conference, which aims at studying the robustness of a retrieval model and developing effective methods for difficult queries [28].

G. Salton [10] We summaries the work on automatic RF techniques. It is clear from the vast majority of work on automatic query modification that it can prove an effective, practical solution for improving the quality of on-line searching and it has been demonstrated to work well under a number of conditions. In particular, it is a very useful technique for improving the performance of short queries or queries which provide poor initial rankings. The basic approach of reweighting and expanding queries, using terms drawn from the relevant documents, works well with the major contribution often coming from the expansion component of the query modification , although this may be collection dependent.

III. PROPOSED WORK

We are going to propose a IR system using which the user can easily get the relevant document. When the user enter the query for search the document, then it directly compare within the data of the document file. So the relevant document will found by the system. We are also working to add feature, the system will recommend the keyword to the user for getting the best result or document. The basic procedure is:

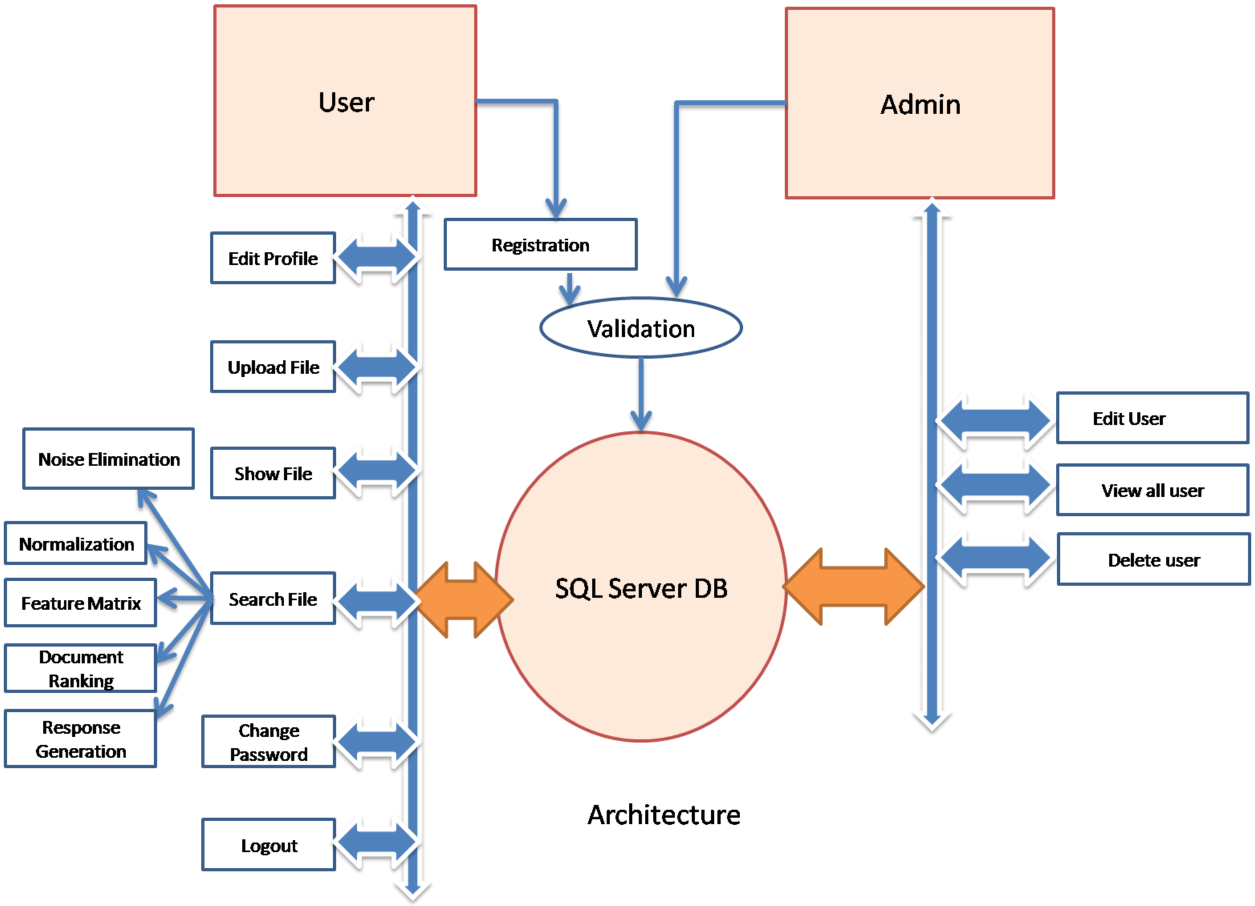


Fig No 1: Architecture Of Proposed System

1. The user issues a (short, simple) query
2. The system returns an initial set of retrieval results.
3. The user marks some returned documents as relevant or nonrelevant.
4. The system computes a better representation of the information need based on the user feedback.
5. The system displays a revised set of retrieval results.
6. It provides relevant documents only to user‘s information need.
7. Easy to retrieve the data.
8. It reduces the manual work.
9. Explicit Relevance Feedback also called as Term relevance feedback. The system will suggest the term which types of term the user should add in search.
10. Implicit Relevance Feedback will find out the frequently search document easily.

Relevance Feedback Algorithm

Relevance feedback (RF) is the retrieval task where the system is given not only a user query, but also user feedback on some of the top ranked results. Feedback gives the retrieval system a chance to improve its results by exploiting the extra information through more elaborate techniques. This can be helpful in cases where the users want as many relevant results as possible. RF is one of the most useful Query Modification techniques in the field of Information Retrieval (IR). This method is put into practice when the user needs to improve the query formulated to the IR system, because the documents initially retrieved do not completely fulfill the user’s information need. Relevance feedback works in the following way: a user submits a query representing his/her information need to the IR system, which then ranks the documents according to their corresponding degrees of relevance to the query (with the documents most closely matching the query ranked first).

Relevance feedback works in the following way: a user submits a query representing his/her information need to the IR system, which then ranks the documents according to their corresponding degrees of relevance to the query (with the documents most closely matching the query ranked first). The user then inspects this list,1 and determines which documents are relevant and which are not relevant to his/her information need (the relevance judgments). Using this information, the IR system updates the initial query, modifying the importance of the terms it contains 2 (term reweighting), and adding new terms that are considered useful to retrieve more relevant documents (query expansion). This process is repeated until the user is completely satisfied with the set of retrieved relevant documents. Relevance feedback has been successfully applied in a great variety of IR models.

Algorithm.

Start

```
1. Take input from user {i1…….in}
2. The query processing starts
3. shaping the modified vector
4. creating associated weights (a, b, c)
5. values for b and c should be incremented or decremented proportionally to the set of documents classified.
6. Information retrieval process
   The list of documents {d1….dn}
   While(d1==i1)
   {
           Result= Dr
   }
   Else
   {
           Result=Dnr
   }
7. Assigning ranking to the documents {d1………dn} according to relevance
8. Generate the response for the current information retrieval process.
```

where,

i1…in= input query from user.

d1…..dn=processed documents.

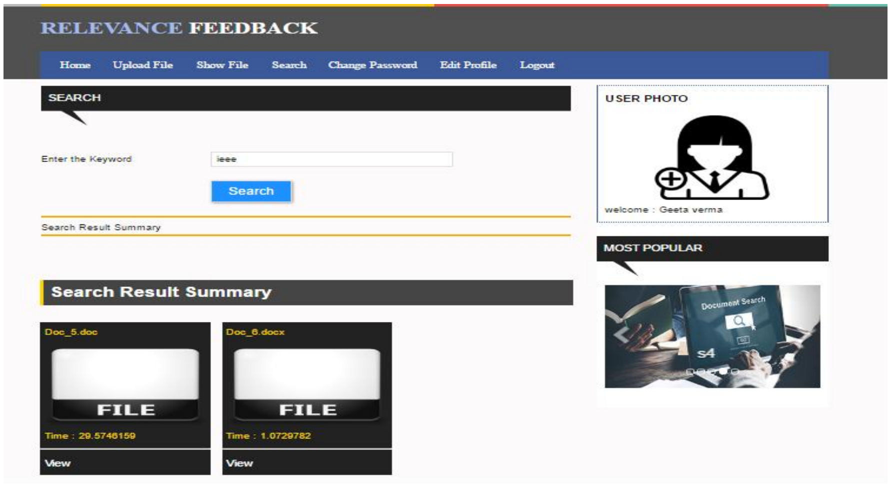
a,b,c= associates weight of document

Dr= presented to be sets of vectors containing the coordinates of related documents

Dnr= presented to be sets of vectors containing the coordinates of non-related documents

VI. RESULTS AND SIMULATION

In Fig 2: shows relevant document that user want. Relevant document is one whose total word count is 1 means word relevance feedback is found only in one document that is Doc_5 and display processing time require for each document.



Enter the Keyword: ieee

Search Result Summary

welcome : Geeta verma

Search Result Summary

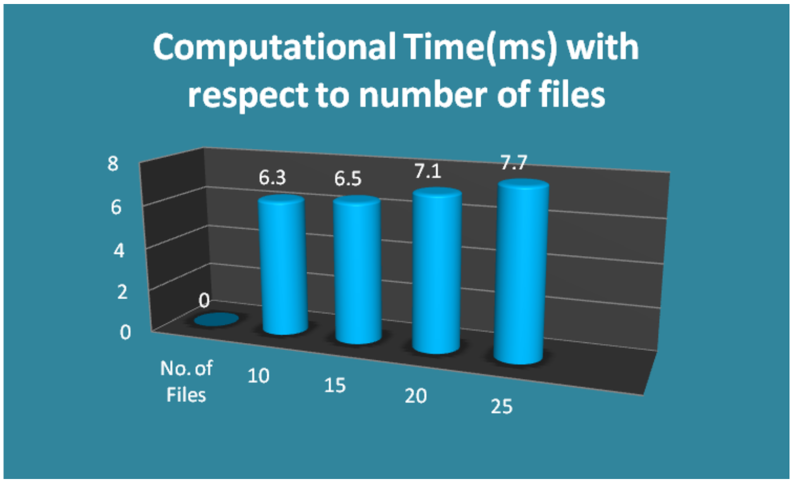
Doc_5.doc
Time : 29.5746159

Doc_6.docx
Time : 1.0729782

Fig 2:IR system display result.

Performance measure

No. of Files	Average Response time for file processing(MS) in Project
10	6.3
15	6.5
20	7.1
25	7.7



In the above graph the x axis indicates the number of files on which the relevance feedback algorithm performed. The y axis indicated the total computational time in ms to perform operation.

V. CONCLUSION

Relevance feedback can go through one or more iterations of this sort. The process exploits the idea that it may be difficult to formulate a good query when you don’t know the collection well, but it is easy to judge particular documents, and so it makes sense to engage in iterative query refinement of this sort. In such a scenario, relevance feedback can also be effective in tracking a user's evolving information need: seeing some documents may lead users to refine their understanding of the information they are seeking. The user submit a query into IR system. IR system return both relevant and irrelevant documents so the automatic procedure that modify the user's queries is known as RF; some relevance assessments about the retrieved documents are collected and the query is expanded by the terms found in the relevant documents, reduced by the terms found in the irrelevant documents or reweighted using relevant or irrelevant documents.

REFERENCES

[1] E. Efthimiadis and P. Biron. "*UCLA-Okapi at TREC-2: query expansion experiments*".Proceedings of the Second Text Retrieval Conference (TREC-2). NIST Special Publications 500-215. D. K. Harman (ed). pp 279-290. 1994.
[2] J. H. Lee. Combining the evidence of different relevance feedback methods for information retrieval. Information Processing and Management. 34. 6. pp 681-691. 1998.
[3] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu and M. Gatford. Okapi at TREC-3. "*Proceedings of the Third Text Retrieval Conference (TREC-3).*" NIST special publication 500-225. (D. K. Harman ed). pp 109-126. 1995.
[4] P. Vakkari."*Cognition and changes of search terms and tactics during task performance.*"Proceedings of RIAO Conference on Content-Based Multimedia Information Access. Paris. pp 894- 907. 2001.
[5] P. Vakkari. "*Relevance and contributing information types of searched documents in task performance.*" Proceedings of the twenty-third annual international ACM SIGIR Conference on Research and development in information retrieval. pp 2-9. Athens. 2000.
[6] Xuanhui Wang , Hui Fang , ChengXiang Zhai ,"*A Study of Methods for Negative Relevance Feedback. SIGIR'08, July 20–24, 2008.*
[7] Yuanhua , Cheng Xiang Zhai ,Wan Chen .*"A Boosting Approach to Improving Pseudo-Relevance Feedback." SIGIR'11, July 24–28, 2011.*
[8] Jun Miao, Jimmy Xiangji Huang, Zheng Ye, "*Proximity-based Rocchio's Model for Relevance Feedback.*" SIGIR'12, August 12–16, 2012.
[9] E.M. Voorhees. "*Draft: Overview of the trec2005 robust retrieval track*". In Notebook of TREC2005, 2005.
[10] E.M. Voorhees, "*Overview of the trec 2004 robust retrieval track*". In TREC2004,2005.
[11] G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. Journal of the American Society for Information Science. 41. 4. pp 288-297. 1990.

BIOGRAPHY

Dipalee S. Hirde is a Student of M.E Computer Science & Engineering Department, H.V.P.M'S College of Engineering & Technology, Amravati, Maharashtra , India. She received Bachelor of Engineering Degree in 2015 from SGBAU Amravati, Maharashtra, India. Her research interests are Education technology and Data Mining.

Prof. R. R. Keole is a Asst. Professor in Department of Information Technology & Engineering, H.V.P.M'S College of Engineering & Technology, Amravati , Maharashtra , India.