A Review of Information Retrieval System using Relevance Feedback Algorithm

Sneha deep
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
sneha.shillu002@gmail.com

Vrajesh Chawra
Central College of Engineering and Management
Dept. of Computer Science and Engineering
Raipur, Chhattisgarh, India
chawra05@gmail.com

Abstract

In old days, people have become conscious about the consequences of archiving and finding information. With the arrival of computers, it became possible to store huge amount of information; and finding the useful information from document collections which is became a necessity. Out of this necessity in the 1950s, the field of Information Retrieval (IR) was born. The field of information retrieval has matured considerably over the last forty years. Several IR systems are used on an everyday by a wide variety of users. Information retrieval is become a very important research area in the field of computer science. Information retrieval (IR) is usually concerned with the searching and retrieving of knowledge-based information from database. In this paper, we represent different models and techniques for information retrieval. This paper reviews some of the technique used for Information retrieval.

Keywords— *Information retrieval, quantum mechanics, relevance feedback, quantum detection.*

I. INTRODUCTION

Current information retrieval system allows user to retrieve huge amount of electronically stored information objects. Information retrieval is concerned with retrieving and indexing documents including information which is relevant to a user's information need. Users can express their needed information using various ways, one of the most common means is queries written in natural languages. However, a query can be very challenging for the reason that the richness of natural language. Usually a query is unstructured as well as ambiguous therefore a query may express two or more than two different information needs and also retrieval of one information may be expressed by two or more different queries.

The user accessed the data by the information retrieval system, the data is usually in the form of unstructured or in the form of semi-structured. However a database system will be used to answer

"Mexico City has the worst air pollution in the world. Pertinent documents would contain the specific steps Mexican authorities have taken to combat this deplorable situation"

Most of time it happens that there are several searchers may not able to have a well-defined idea of what information they are searching for, users may not be able to convey their conceptual idea of what information they need into a suitable query and also whatever the information available for the information retrieval are unaware to the user. Previously, the user face difficulty to express their exactly same idea for the retrieval of information that they required, the researcher recognized that and provide the useful information which is present in the user's search. Therefore, searchers may not be able to convey their required need of information into their request, after that system generate the initially a set of several documents that contains useful information which is indicated by the user.

This helps to give rise to the Relevance Feedback (RF): the user indicating their required documents as relevant to their needs and pointing up this information to the information retrieval system. The system provides the information –helps in retrieving more similar documents to the relevant to the previous documents. The process of relevance feedback is usually presented as a activity cycle, an information retrieval system presents a uses with a set of retrieval documents, the user only concern with that are relevant and the system will uses this information to produce a improved version of the query. The improved query is then used to retrieve a new set of documents for the make known of user. This whole process is known as iteration of Relevance Feedback (RF).

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 include 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 [1]

The procedure by which an information retrieval system uses the information of relevance feedback given by the user is the main focus of this paper reviews some of the technique used for Information retrieval. The main focus of this paper is the procedure by which an information retrieval system uses the relevance information which is given by the user. This paper covers various features of Relevance Feedback: the representations used in RF, how these illustration lead to make a choice how to modify a query as well as the role of interface in RF.

It may be possible for an IR system for small collections of documents to access each document in turn and deciding whether or not it is likely to be so relevant to a user's query [2][3]. This becomes impractical for larger collections, especially in interactive systems. Therefore, it is very necessary to make the document collection into an easily make available the representation; documents that are most likely to be relevant that document one can targeted, here is the example, those documents that are having at least one word that come into view in the user's query [4].

This complete change from a text document to a representation of a text is called indexing the documents. Here is various types of indexing techniques in spite of this the majority rely on choosing a good document descriptors, there are many types of descriptors such as keywords, or terms, to represent the meaningful information from the content of documents. For IR a 'good' descriptor is a term that used to explain the information content of the document and it also one that can help to differentiate the document from the other documents in the collection of documents. Again a 'good' descriptor has a certain discriminatory as power. The power of a term such as a 'good' in discriminating documents can be used to differentiate between the relevant and non-relevant documents.

Once the text document has been tokenized it is very necessary to make decision which terms should be used to represent the documents. Therefore, we need to make choice which descriptors are so useful for the joint role of illustrating the document's content and the discriminating of the document from the other documents in the collection. The high frequency terms getting ones in a high proportion of the documents in the collection that will tend not to be very effective either in discriminating between documents or in representing documents.

II. LITERATURE SURVEY

Massimo Melucci et. al. [5] focused on 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. These algorithms project the query vector on a subspace spanned by the eigenvector which maximizes the distance between the distribution of quantum probability of relevance and the distribution of quantum probability of non-relevance. The experiments showed that the RF algorithms inspired by quantum detection can outperform the state-of-the-art algorithms.

Kevyn Collins-Thompson et. al. [6] focused on models and algorithms to address the three key problems in improving relevance for search using reading difficulty: estimating user proficiency, estimating result difficulty, and re-ranking based on the difference between user and result reading level profiles. Author evaluate our methods on a large volume of Web query traffic and provide a large-scale log analysis that highlights the importance of finding results at an appropriate reading level for the user.

Donna Harman et. al. [7] focused on experiments, using the Crartfield 1400 collection~ showed the importance of query expansion in addition to query reweighing, and showed that adding as few as 20 well-selected terms could result in performance improvements of over 100%. Additionally it was shown that performing multiple iterations of feedback is highly effective.

Ingo Frommholz et. al. [8] focused on an approach inspired by quantum mechanics to represent queries and their reformulations as density operators. Differently constructed densities can potentially be applied for different types of query reformulation. To do so, we propose and discuss indicators that can hint us to the type of query reformulation we are dealing with.

Yuanhua Lv et.al. [9] Focused on an approach to adaptively predict the optimal balance coefficient for each query and each collection. We propose three heuristics to characterize the balance between query and feedback information. Taking these three heuristics as a road map, we explore a number of features and combine them using a regression approach to predict the balance coefficient. Our experiments show that the proposed adaptive relevance feedback is more robust and effective than the regular fixed-coefficient feedback.

Table I. shows various existing methods and its limitations

S. No.	Ref. No.	Method Used	Data Source	Approach	Strength	Limitation
1	5	RF algorithms	Text Document	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	Author showed that the RF algorithms inspired by quantum detection can outperform the state-of-the-art algorithms.	Does not remove noise.
2	6	Re-ranking Algorithm	Text Document	Proposed models and algorithms to address the three key problems in improving relevance for search using reading difficulty: estimating user proficiency, estimating result difficulty, and re-ranking based on the diff erence between user and result reading level profiles.	Author evaluate method on a large volume of Web query traffic and provide a large-scale log analysis that highlights the importance of finding results at an appropriate reading level for the user	Not efficient for all types of query.
3	7	Probabilisti c retrieval model	Crartfield 1400 collection	Showed the importance of query expansion in addition to query reweighing, and showed that adding as few as 20 well-selected terms could result in performance improvements of over 100%.	Author shown that performing multiple iterations of feedback is highly effective.	Collections of much longer documents have problems with feedback.
4	8	Quantum Mechanic	Text Document	Propose an approach inspired by quantum mechanics to represent queries and their reformulations as density operators.	Author propose and discuss indicators that can hint us to the type of query reformulation we are dealing with.	-
5	9	Adaptive Relevance Feedback Model	Text Document	Proposed a learning approach to adaptively predict the optimal balance coefficient for each query and each collection and also propose three heuristics to characterize the balance between query and feedback information.	The proposed adaptive relevance feedback is more robust and effective than the regular fixed-coefficient feedback.	Not so effective and robust features.

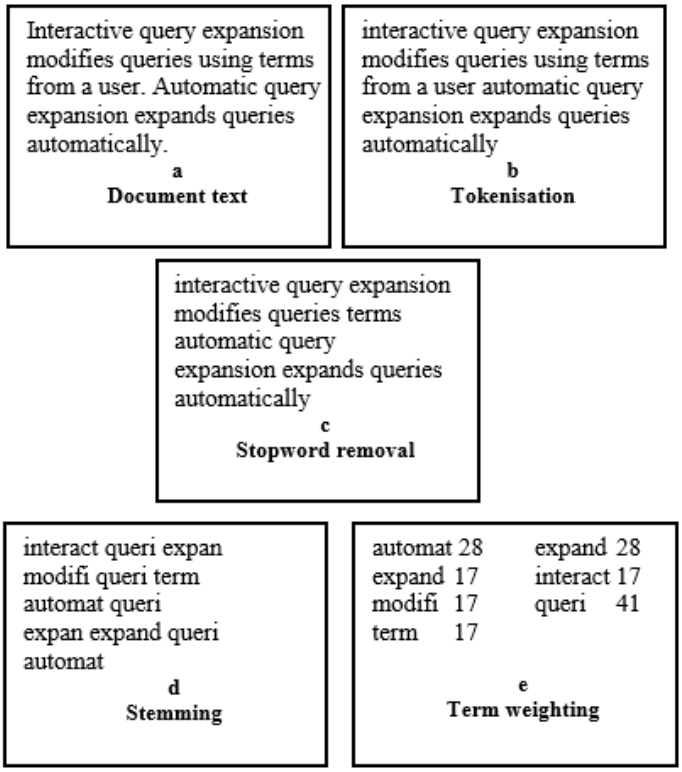

Fig. 1. Indexing a document

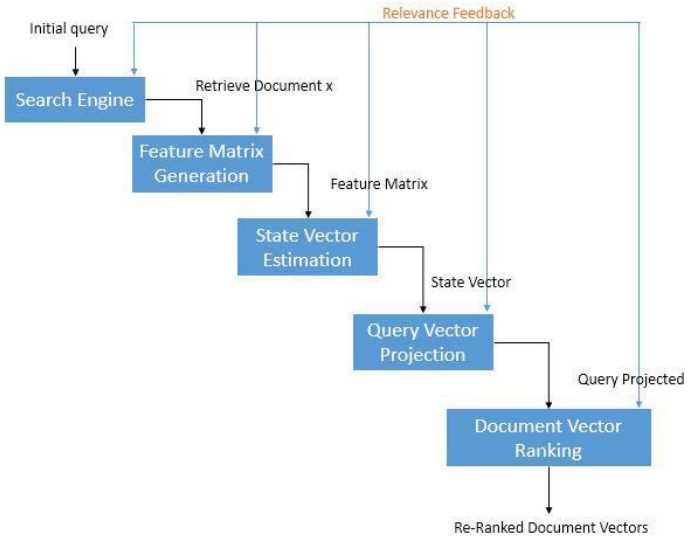

Fig. 2. Methodology for Relevance Feedback Algorithm

The RF algorithm is also known as Rocchio's algorithm and it is designed to compute the new query vector using a linear combination of the original vectors, the relevant document vectors and the non-relevant document vectors, where the labels of relevance are collected in a training set.

The RF algorithm empirically expressed by:

$$y^* = \underbrace{\overbrace{y}^{\text{original query}} + \overbrace{y^+}^{\text{positive RF}} - \overbrace{y^-}^{\text{negative RF}}}_{\text{modified query}},$$

III. TOOLS USED

There are many tools available for processing data and extracting user information from collection of datasets. Various tools can be used are described below.

a. Statistical Package of R: Used to train the datasets.
b. SOM-PAK package of Python: The SOM PAK program package contains all programs necessary for the correct application of the Self Organizing Map algorithm in the visualization of complex experimental data.
c. NumPy is the fundamental package for scientific computing with Python

IV. CONCLUSION

To conclude that, information retrieval is a mechanism of searching and retrieving the knowledge which is based on information from collection of text documents. Users can express their needed information using various ways, one of the most common means is queries written in natural languages.

However, a query can be very challenging for the reason that the richness of natural language. This review has deal with the basics of the information retrieval. At first we are defining the information retrieval system with their basic measurements. After all this we concern with the traditional information retrieval models and also discuss about the indexing techniques.

REFERENCES

[1] C. J. Van Rijsbergen, The Geometry of Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[2] X. Wang, H. Fang, and C. Zhai, "A study of methods for negative relevance feedback," in Proc.

31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2008, pp. 219–226.

[3] R. W. White and R. A. Roth, Exploratory Search: Beyond the QueryResponse Paradigm. San Rafael, CA, USA: Morgan & Claypool,2009.

[4] S. Wong and V. Raghavan, "Vector space model of information retrieval: A reevaluation," in Proc. 7th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1984, pp. 167–185.

[5] Yuanhua Lv, ChengXiang Zhai, "Adaptive Relevance Feedback in Information Retrieval", ACM November 2–6, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[6] Kevyn Collins-Thompson, Sebastian de la chica and David Sontag," Personalizing Web Search Results by Reading Level", 2011 ACM.

[7] Donna Harman, "Relevance Feedback Revisited", 2010 ACM.

[8] Ingo Frommholz, Benjamin Piwowarski , Mounia Lalmas and Keith van Rijsbergen. "Processing Queries in Session in a Quantum-inspired IR Framework", 2011 ACM.

[9] Yuanhua Lv, ChengXiang Zhai, "Adaptive Relevance Feedback in Information Retrieval", ACM November 2–6, 2009.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.