Information Retrieval on Document Streams using Relevance Feedback Algorithm

P. Praveena
Department of Computer Science and Engineering
Vignan's Institute of Engineering for Women
Visakhapatnam, India

M. Sailaja
Department of Computer Science and Engineering
Vignan's Institute of Engineering for Women
Visakhapatnam, India

V. Prathyusha
Department of Computer Science and Engineering
Vignan's Institute of Engineering for Women
Visakhapatnam, India

T. Pooja
Department of Computer Science and Engineering
Vignan's Institute of Engineering for Women
Visakhapatnam, India

G.Pavani Latha
Assistant Professor
Department of Computer Science and Engineering
Vignan's Institute of Engineering for Women
Visakhapatnam, India

---***---

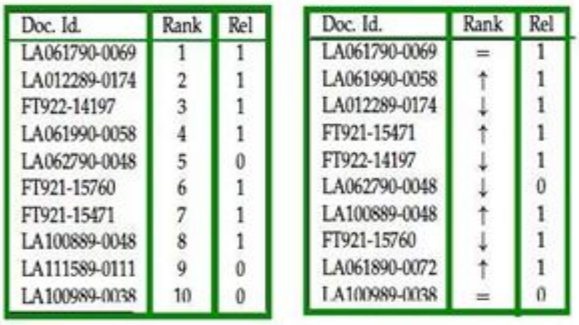
***Abstract*—Information Retrieval is the process of tracing and recovering of specific information from the stored data. Traditionally, the documents are retrieved based on the index (i.e., filename, folder name, sub-folder name). The data which is extracted based on these categories might not be accurate. Our main motive is to extract the information which is exactly matched with the user requirement by applying RF algorithm on the search technique. The term Relevance Feedback indicates relevance feedback in which the data can be extracted either based on the index or content. The model is very accurate in re-weighting query terms by projecting the given query vector on the subspace represented by the eigen vector.**

Keywords*— *Information Retrieval, Relevance Feedback,Eigen vector

I. INTRODUCTION

Information Retrieval is the process of gathering useful and important information from different information resources in order to achieve the output. The process of extracting the useful information from a document and in turn it searches for metadata that describes about the data. Information retrieval can be done physically or manually from large data sources, so it takes a lot of information overloading problems. To extract the information with high accuracy retrieval based on content wise is used to reduce the problem of overloading.

Till now all the search keywords try to match the content based on the filename or the folder name which is present inside the drive or sometime if the search is done based on the database tables. The search keywords try to verify the content is matched either from filename or category name but not based on the content[1].In the process of information retrieval the objects may vary from one type to other based on the requirements like text documents to images and images[2] to audio and audio to video or maps and so on. Often the documents are not stored in the information retrieval system; instead they are represented in the metadata.



Doc. Id.	Rank	Rel
LA061790-0069	1	1
LA012289-0174	2	1
FT922-14197	3	1
LA061990-0058	4	1
LA062790-0048	5	0
FT921-15760	6	1
FT921-15471	7	1
LA100889-0048	8	1
LA111589-0111	9	0
LA100989-0038	10	0

Doc. Id.	Rank	Rel
LA061790-0069	=	1
LA061990-0058	↑	1
LA012289-0174	↓	1
FT921-15471	↑	1
FT922-14197	↓	1
LA062790-0048	↓	0
LA100889-0048	↑	1
FT921-15760	↓	1
LA061890-0072	↑	1
LA100989-0038	=	0

figure1: Before and After RF

II. LITERATURE SURVEY

Many researchers have been carried out in order to prove the current RF Algorithm for extracting the data based on the content as well as the index.

Massimo Melucci, University of Padua, Italy [3] discussed about "Relevance Feedback Algorithms Inspired by Quantum Detection". R. Blanco and P. Boldi [4] discussed about "Extending bm25 with multiple query operators". C Carpineto and G. Romano [5] performed "A survey of automatic query expansion in information retrieval". Ingo Frommholz, Benjamin PiwowarskiMouniaLalmas, Keith van Rijsbergen [6] discussed about "Processing queries in session in a quantum-inspired IR framework". R. B. Griffiths [7] analyzed "Consistent Quantum Theory". This paper presents an elementary introduction to consistent quantum theory.

III. PROPOSED METHOD

The main purpose of relevance feedback is for improving the information retrieval on the document streams. Some sample documents with different names and different contents are considered and relevance feedback algorithm is applied to extract the most related files into the first preference.

The proposed RF Algorithm works in following way:

A. *Feature set generation:*

Initially, the user raises a query to the search engine which identifies the matched feature from the input query. Feature set is formed based on the input query. The input query is extracted with a set of distinct features then the related features are formed into state vector machine (SVM).

B. *State vector estimation:*

The state vector machine considers two parameters

- Where 0, 1 indicates true or false.
- If the input keyword and the words in the document have exact match then it is positive relevance feedback.
- If the input keyword and the words in the document are not matched then it is negative relevance feedback.

C. *Eigen Vector extraction:*

The Eigen values and Eigen vector features are mostly used in the analysis of the linear transformation. It is used for finding the matrix diagonalization based on the keywords in order to extract the relevance feedback. So, here we try to assume an eigen vector (v) of a linear transformation T is a non-zero vector that, when T is applied, it doesn't change direction. Applying T to the eigen vector only scales the vector by scalar value (λ) called an eigen value .This condition can be written in the equation:

$$\boxed{T(v) = \lambda v}$$

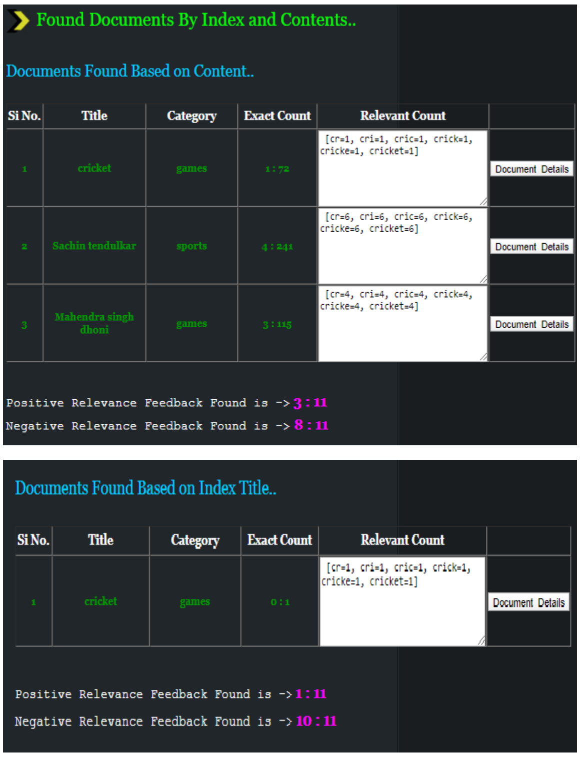
The above equation is referred to as the eigen value equation or eigen equation. Eigen matrix projects the frequency of the word count in the document.

D. *Query vector projection & Document Reranking:*

- High word frequency related to the input query, ranked top among PRF else NRF.
- On this basis, the documents are retrieved.
- The proposed algorithm is able to identify the positive relevance feedback among the set of documents based on the query keyword.

IV. IMPLEMENTATION AND RESULTS

The project application is designed and developed with java programming language, in which the front end of the application is done in HTML, JSP and CSS. The back end of the application is designed with MY-SQL database.



Found Documents By Index and Contents..

Documents Found Based on Content..

Si No.	Title	Category	Exact Count	Relevant Count	
1	cricket	games	1 : 72	[cr=1, cri=1, cric=1, crick=1, cricke=1, cricket=1]	Document Details
2	Sachin tendulkar	sports	4 : 241	[cr=6, cri=6, cric=6, crick=6, cricke=6, cricket=6]	Document Details
3	Mahendra singh dhoni	games	3 : 115	[cr=4, cri=4, cric=4, crick=4, cricke=4, cricket=4]	Document Details

```
Positive Relevance Feedback Found is -> 3 : 11
Negative Relevance Feedback Found is -> 8 : 11
```

Documents Found Based on Index Title..

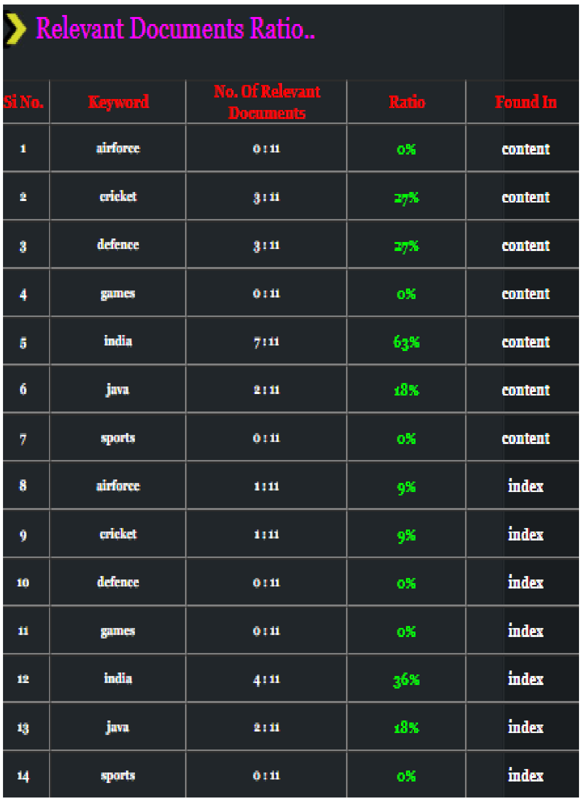
Si No.	Title	Category	Exact Count	Relevant Count	
1	cricket	games	0 : 1	[cr=1, cri=1, cric=1, crick=1, cricke=1, cricket=1]	Document Details

```
Positive Relevance Feedback Found is -> 1 : 11
Negative Relevance Feedback Found is -> 10 : 11
```

figure 2:User trying to search documents based on keywords

From the above screen we can clearly identify that the user can search the data using the keywords and based on that the documents are retrieved based on the content and index.

The document which is matched mostly with inner content is displayed first and later the document with the index matched.We can also check the percentage of matched as well as not matched documents.



Relevant Documents Ratio..

Si No.	Keyword	No. Of Relevant Documents	Ratio	Found In
1	airforce	0 : 11	0%	content
2	cricket	3 : 11	27%	content
3	defence	3 : 11	27%	content
4	games	0 : 11	0%	content
5	india	7 : 11	63%	content
6	java	2 : 11	18%	content
7	sports	0 : 11	0%	content
8	airforce	1 : 11	9%	index
9	cricket	1 : 11	9%	index
10	defence	0 : 11	0%	index
11	games	0 : 11	0%	index
12	india	4 : 11	36%	index
13	java	2 : 11	18%	index
14	sports	0 : 11	0%	index

figure 3: Documents matched with index and content

From the above screen,we clearly identify the documents which are matched with index as well as content seperately and also we can see the percentage of both the keywords individually from a set of predefined documents which we choose as input.

V. Conclusion

In this paper, we developed an algorithm for information retrieval not only based on index but also based on keyword, which is known as Relevance Feedback (RF) algorithm. RF algorithm is mainly inspired by quantum detection principle in order to re-weight each and every object that is matched with our query keyword and finally re-rank the documents retrieved. Our model is very accurate in re-weighting query terms by projecting the given query vector on the subspace represented by the eigen vector.

REFERENCES

[1] Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, "The SMART Retrieval System: Experiments in Automatic Document Processing", chapter 14, pages 313–323. Prentice-Hall, Englewood Cliffs, NJ, 1971

[2] I. Frommholz, B. Larsen, B. Piwowarski, M. Lalmas, P. Ingwersen, and K. van Rijsbergen. "Supporting polyrepresentation in a quantum-inspired geometrical retrieval framework". In Proceedings of IIiX, pages 115–124, 2010.

[3] Massimo Melucci, University of Padua, Italy, "Relevance Feedback Algorithms Inspired by Quantum Detection" IEEE Transactions on Knowledge and Data Engineering,Volume:28,Issue:4,April 1 2016

[4] R. Blanco and P. Boldi. "Extending bm25 with multiple query operators", In Proceedings of SIGIR, pages 921-930,2012.

[5] C. Carpineto and G. Romano. "A survey of automatic query expansion in information retrieval", ACM Comput. Surv. 44(1):1– 50, Jan. 2012.

[6] I. Frommholz, B. Piwowarski, M. Lalmas, and K. van Rijsbergen. "Processing queries in session in a quantum-inspired IR framework", In Proceedings of ECIR, pages 751–754, 2011.

[7] R. B. Griffiths. "Consistent Quantum Theory", Cambridge University Press, Cambridge, UK, 2002.