International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

A Novel Effective Algorithm for Improving Information Retrieval on Document Streams

Pachhigolla Usharani [#1] Medisetty Nidesh [#2] , Penmetsa Meher Pravallika [#3] ,Gorapalli Srinivasarao [#4] ,

[#1,#2,#3] B.Tech Student, Department of CSE ,
Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam, Visakhapatnam, AP, India
[#4] Assistant Professor, Department of CSE ,
Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam, Visakhapatnam, AP, India.

ABSTRACT

Information Retrieval (IR) is mainly used for extracting the most related information from a set of resources that are available. Now a day's information retrieval is done only based on index (i.e. filename, folder name or sub-folder name), hence the data which is extracted based on these categories is not always accurate. Hence there is no single mechanism which can extract the most related and exact information for the given search keyword. In this paper we mainly try to extract the information which is almost exactly matched with the user requirement by applying RF algorithm on the search technique. The term RF indicates relevance feedback in which the data can be extracted either based on index as well as content, so that the data user who try to search the files can get related files as top priority and those which are not exactly matched will be set as non priority files by the application and they will be send to the last level. Here we mainly inspired by the term quantum detection in order to extract the data based from the information resource. By conducting various experiments on finding relevance feedback based on quantum detection, the simulation results clearly tell that this model is very accurate in re-weight query terms by projecting the given query vector on the subspace represented by the eigenvector.

Key Words:

Information Retrieval, Quantum Detection, Relevance Feedback, Subspace, Query Projection.

1. Introduction

Information Retrieval is the process

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

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 describe about the data. As we all know that information retrieval is done manually or physically from large data sources, so it takes a lot of information overloading problems. Recently for extracting the information with high accuracy all are trying to extract the information based on content wise so that we can able to reduce a lot of overloading problem. Best example for information retrieval in the real time is Google search engine. In the first stage the data user try to assume that all the data which is to be searched is in the form of unstructured manner rather that in a structured manner, so there is a need to search the data and arrange the data in a structured manner. Here each and every search query is treated as an object ,which is an entity that is represented by information in a content collection or database.

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

a) Before RF

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

b) After RF

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

Figure 1. Denotes the main common similarities between before Relevance Feedback and after Relevance Feedback

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

II. Related Work

In this section we will mainly discuss about the related work that is carried out in order to prove the current RF Algorithm for extracting the data based on content as well as index. For this we try to study about Eigen Vector in this section in detail

Eigen Vector and Eigen Matrix

Eigen values and Eigen vectors feature are mostly used in the analysis of linear transformations. The word Eigen- is adopted from the German word eigen for “proper”, “characteristic”. Actually the word eigen came from the study of principle axes of the rotational motion of the rigid bodies. Most of the applications use this eigen vectors for performing their operation in a wide range of applications, some of the best applications that uses eigen vectors technique is facial recognition and diagonalization. Here in this proposed

application, Eigen matrix or eigen vector is used for finding the matrix diagonalization based on the keywords that was used in order to extract the relevance feedback. Here the positive relevance feedback as well as negative relevance feedback for the appropriate document based on the query input. 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 eigen vector by the 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 **eigenvalue equation** or **eigen equation.**

In general, λ may be any scalar.

For example, λ may be negative, in which case the eigen vector reverses direction as part of the scaling, or it may be zero or complex.

In the next section we will discuss about quantum detection in detail for elaborating the paper in depth for the relevance search based on index and content.

3. Novel Effective Algorithm for Improving Information Retrieval On Document Streams

In this section, we mainly describe the proposed Novel Effective Algorithm for Improving Information Retrieval on Document Streams. Now let us Discuss About that in detail as follows:

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

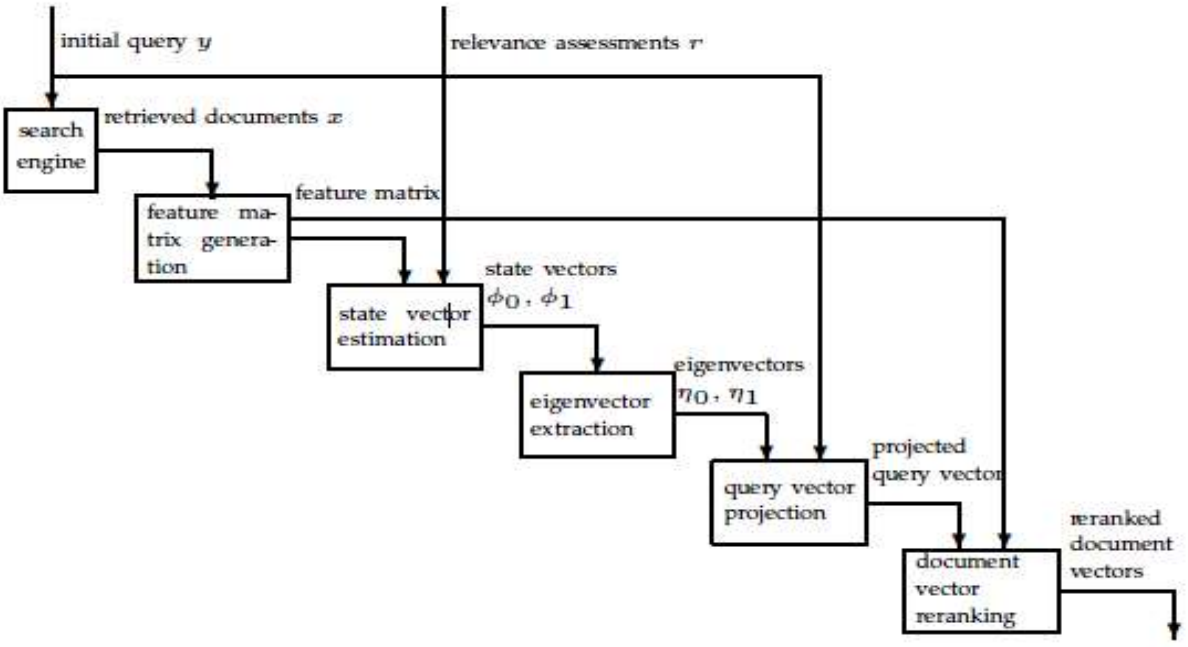


Figure. 2. Represents the Proposed Model For Extracting Relevance Feedback

From the above figure 2, we can clearly represent the proposed architecture flow diagram of our current thesis which will mainly discuss about the extraction of relevance data from a set of documents.Here we try to take some sample documents which are having different filenames and different contents present inside the database and we try to apply RF algorithm to extract the most related files into the first preference and then later with next updated files.We can demonstrate the figure 2 with following step wise explanation.

The proposed RF algorithm works in the following way:

STEP 1:

Initially, the user raises a query to the search engine which identifies the matched features from that input query.

STEP 2:

Feature set can be formed based on the input keyword.

STEP 3:

The input query is extracted with a set of distinct features then the related features are formed into state vector machine (SVM).

STEP 4:

Here SVM takes two parameters like

Where 0, 1 indicates true or false. If input keyword and the words in the documents have exact match then it is positive RF otherwise, negative RF.

STEP 5:

Eigen matrix projects the frequency of the word count in the document.

STEP 6:

High word count frequency related to the input query, ranked top among PRF else NRF.

STEP 7:

Finally all the documents are retrieved in the form of ranked order, based on PRF and NRF.

This proposed algorithm can able to identify the positive relevance feedback among a set of documents based on a query keyword.

4. Experimental Result

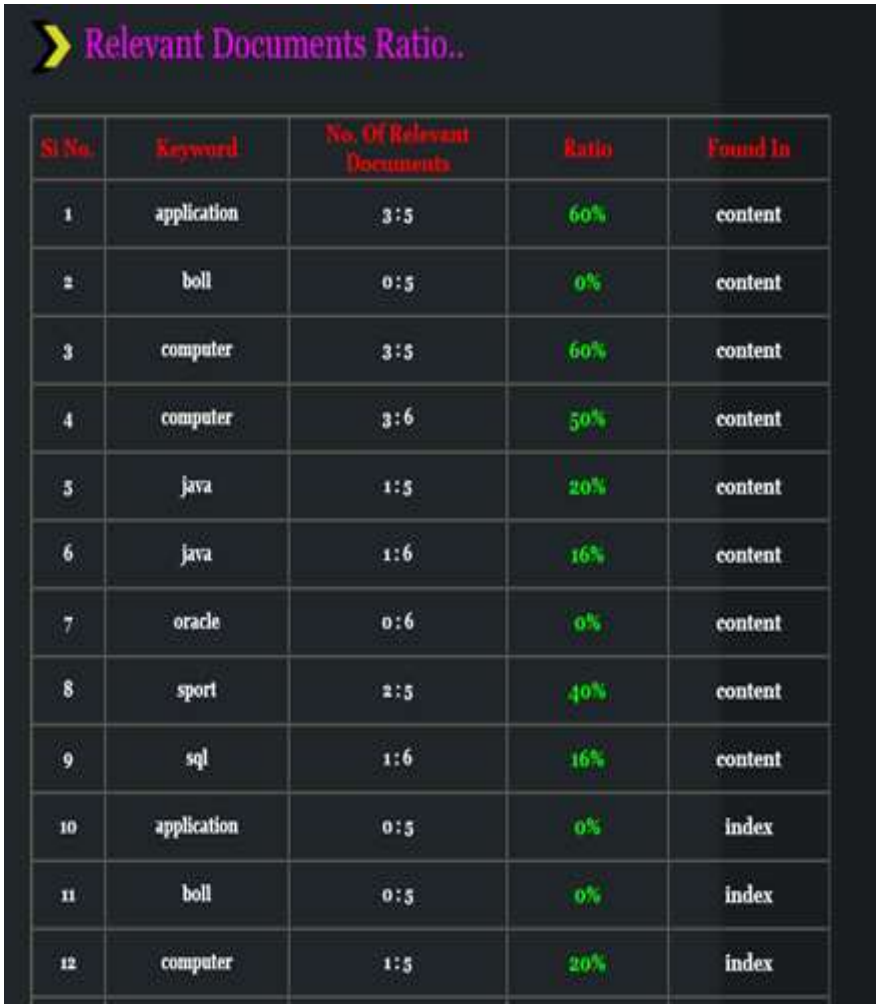
The proposed application is designed and developed with java programming language, in which the front end of the application is done with HTML, JSP and CSS.The back end of the application is designed with MY-SQL database .Here we try to take some set of documents into the database and try to apply our proposed RF algorithm on our current approach:

User try to search the document based on Keyword from a set of documents

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

From the above screen we can clearly identify that the data user try to a keyword as search input and based on that keyword the documents are retrieved based on content wise and index wise.If we look at the output clearly the document which is matched mostly with inner content is displayed first and later the document with the index matched. Here we can also check the percentage of matched as well as not matched documents.

International Journal of Research
Available at https://journals.pen2print.org/index.php/ijr/
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019



Relevant Documents Ratio..

Sl No.	Keyword	No. Of Relevant Documents	Ratio	Found In
1	application	3:5	60%	content
2	boll	0:5	0%	content
3	computer	3:5	60%	content
4	computer	3:6	50%	content
5	java	1:5	20%	content
6	java	1:6	16%	content
7	oracle	0:6	0%	content
8	sport	2:5	40%	content
9	sql	1:6	16%	content
10	application	0:5	0%	index
11	boll	0:5	0%	index
12	computer	1:5	20%	index

From the above window we can clearly identify the documents which are matched with index as well as content separately and also we can see the percentage of both the keywords individually from a set of pre-defined documents which we choose as input. Also we can see the total file count with the ratio of matched and non-matched documents content. Here for example if there are 5 documents inside the database and if one is matched out of five documents then we treat that as 20 percent matched and 80 percent we treat as not matched.

5. Conclusion

In this paper, we finally developed a new class of algorithm for information retrieval not only based on index but also based on keyword, which is known as

Relevance Feedback (RF) algorithm. Here this RF Algorithms 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 we try to re-rank the documents retrieved from an Information retrieval system. By conducting various experiments on our proposed model like finding relevance feedback based on quantum detection with the principles like communication channel with probabilities, our simulation results clearly tell that our model is very accurate in re-weight query terms by projecting the given query vector on the subspace represented by the eigenvector.

6. References

[1] J. 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] Jansen, B. J. and Rieh, S. (2010), "The Seventeen Theoretical Constructs of Information Searching and Information Retrieval". Journal of the American Society for Information Sciences and Technology. 61(8), 1517-1534.

[4] G. Salton and M. McGill. "Introduction to Modern Information Retrieval". McGraw-Hill, New York, NY, 1983.

[5] Goodrum, Abby A. (2000). "Image Information Retrieval: An Overview of Current Research". Informing Science. **3** (2).

[6] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data". Englewood Cliffs, NJ: Prentice-Hall, 1988.

[7] Foote, Jonathan (1999). "An overview of audio information retrieval". Multimedia Systems. Springer.

[8] Beel, Jöran; Gipp, Bela; Stiller, Jan-Olaf (2009). "Information Retrieval On Mind Maps - What Could It Be Good For?". Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09). Washington, DC: IEEE.

[9] G. Salton and C. Buckley. "Term weighting approaches in automatic text retrieval". Information Processing and Management, 24(5):513–523, 1988.

[10] D. Zellh¨ofer, I. Frommholz, I. Schmitt, M. Lalmas, and K. van Rijsbergen Towards "Quantum-based DB+IR processing Based on the Principle of

Polyrepresentation". In Proceedings of ECIR, pages 729–732. Springer-Verlag, 2011.

[11] 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.

[12] "Unstructured Data and the 80 Percent Rule". Breakthrough Analysis. Retrieved 2015-02-23.

[13] D. Blei and J. Lafferty, "Correlated topic models," Adv. Neural Inf. Process. Syst., vol. 18, pp. 147–154, 2006.

[14] R. Hughes. "The Structure and Interpretation of Quantum Mechanics". Harvard University Press, Cambridge, MA, USA, 1989.

7 .About the Authors

PACHHIGOLLA USHARANI is currently pursuing her 4 Years B.Tech in Department of Computer Science and Engineering, at Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam ,Visakhapatnam, AP, India. Her area of interest includes Web Designing and Development.

MEDISETTY NIDESH is currently pursuing his 4 Years B.Tech in Department of Computer Science and Engineering, at Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam,Visakhapatnam, AP, India. His area of interest includes Web Designing and Development.

PENMETSA MEHER PRAVALLIKA is currently pursuing her 4 Years B.Tech in Department of Computer Science and Engineering, at Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam ,Visakhapatnam, AP, India. Her area of interest includes Web Designing and Development.

e-ISSN: 2348-6848
p-ISSN: 2348-795X
Volume 06 Issue 03
March 2019

GORAPALLI SRINIVASARAO is currently working as Assistant Professor in Department of Computer Science and Engineering, at Nadimpalli Satyanarayana Raju Institute of Technology, Sontyam ,Visakhapatnam, AP, India. She has more than 4 years of teaching experience in engineering colleges. Her area of interest includes Computer Networks.