ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Survey on Efficient Approach for Information Retrieval Using Relevance Feedback Algorithm

Dipalee S. Hirde, Prof. R. R. Keole

M.E Student, Department of Computer Science & Engineering, HVPM's College of Engineering & Technology, Amravati, India

Asst. Professor, Department of Information Technology & Engineering, HVPM's College of Engineering & Technology, Amravati, India

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) is concerned with indexing and retrieving documents including information relevant to a user's information need. Although the end user can express his information need using a variety of means, queries written in natural language are the most common means. However, a query can be very problematic because of the richness of natural language. Indeed, a query is usually ambiguous; a query may express two or more distinct information needs or one information need may be expressed by two or more distinct queries. Text Retrieval Conference(TREC) test collection from which the query is submitted to an IR system based on the Vector Space Model (VSM). This system would return both relevant documents and irrelevant documents. Finding relevant document is one of the hard tasks. 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". Many universities and public libraries use IR systems to provide access to books, journals and other documents. web search engine are the most visible IR application .An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy.

An object is an entity that is represented by information in a content collection or database. User queries are matched against the database information. However, as opposed to classical SQL queries of a database, in information

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

retrieval the results returned may or may not match the query, so results are typically ranked. This ranking of results is a key difference of information retrieval searching compared to database searching.

Depending on the application the data objects may be, for example, text documents, images, audio, mind maps or videos. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata. 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 to the user. The process may then be iterated if the user wishes to refine the query.

II. LITERATURE REVIEW

Buckley and Voorhees [1] introduced a new evaluation metric, which allows to overlook non-judged documents and does not require to consider them to be irrelevant (the metric is computed by analyzing the relative rankings of the relevant and irrelevant documents). Second, we compute the standard metrics such as Mean Average Precision (MAP), Normalized Discounted Cumulative Gain (NDCG) only for the documents for which we have judgments.

Ingo Frommholz [2] present how a geometrical retrieval framework inspired by quantum mechanics can be extended to support polyrepresentation. We show by example how different representations of a document can be modelled in a Hilbert space, similar to physical systems known from quantum mechanics. We further illustrate how these representations are combined by means of the tensor product to support polyrepresentation, and discuss the case that representations of documents are not independent from a user point of view.

M. Shanmugham [3] Present a class of RF algorithms inspired by the quantum detection has been proposed to re-weight query terms by projecting the query vector on the subspace represented by the eigenvector which is the optimal solution to the problem of finding the maximal distance between two quantum probability distributions. The complexity of the calculation of the eigenvector is limited by the small size of the matrix that represents the distance between two quantum probability distributions.

AblimitAji [4] evaluate an approach that relies on a novel source of such knowledge, namely, the revision history of a document. Many information retrieval models, notably statistical language models, assume a generative process of document creation, whereas the terms are chosen to be included in the document according to their importance to the chosen document topic(s), previously chosen terms, and other factors that vary by model. Yet these models only examine one (final) version of the document to be retrieved, effectively ignoring the actual document generation process, even when it is available.

Shuqin Liu [5] present a weighted coefficients of image retrieval algorithm based on relevance feedback are determined in advance, which is lack of flexibility. In order to obtain satisfactory retrieval results, this algorithm requires a large amount of feedback calculation and efficiency of the algorithm is low. Aiming at the faults of relevance feedback, the adaptive adjustment algorithm of weighted coefficients based on quantum particle swarm optimization is presented, which is composed of user feedback process and particle evolution process..

Diane Kelly [6] present a foundation around which others can discuss methods for studying IIR. This includes the creation of more detailed reviews of some of the topics discussed in this paper such as IIR history, measures and ethics. People have varying opinions about how IIR evaluation should be conducted. The content of this paper represents one such opinion that is informed heavily by the literature, the author's research experiences and an academic background that is rooted in the behavioral sciences.

Luis M. de Campos [7] present an approach for relevance feedback in the Bayesian Network Retrieval (BNR) model. Our proposal is based on the propagation of partial evidences in the Bayesian network, representing the new information obtained from the user's relevance judgments to compute the posterior relevance probabilities of the documents.

Pang et al. [8] introduced the relevance degree information into a spectral embedding framework, and proposed a novel Ranking Graph Embedding (RANGE) algorithm by modeling the global structure and the local relationships in and between different relevance degree sets, respectively.

Tian et al. [9] investigated the reranking problem from the probabilistic perspective and derived an optimal reranking function based on Bayesian analysis. In their methods, textual information is modeled as a likelihood to

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

reflect the disagreement between the reranked results and the text-based search results, and the visual information is modeled as a conditional prior to indicate the ranking score consistency among visually similar samples.

Liu et al. [10] proposed a novel unsupervised one-class learning method by jointly learning a large margin one-class classifier and a soft label assignment for targets and outliers. Extensive experiments have shown its effectiveness in outlier image removal and ISR. These approaches have been successfully employed in CBIR and ISR, however, the selection of appropriate kernel is still an open problem. Moreover, the idea of hypersphere in SVDD has not been employed.

K. Collins-Thompson [11] present traditionally, the search engines have ignored the reading difficulty of documents and the reading proficiency of users in computing a document ranking. This is one reason why Web search engines do a poor job of serving an important segment of the population: children. While there are many important problems in interface design, content altering, and results presentation related to addressing children's search needs.

Y. Lv, C. Zhai [12] present the pseudo-relevance feedback has proven effective for improving the average retrieval performance. Unfortunately, many experiments have shown that although pseudo-relevance feedback helps many queries, it also often hurts many other queries, limiting its usefulness in real retrieval applications. Thus an important, yet difficult challenge is to improve the overall effectiveness of pseudo-relevance feedback without sacrificing the performance of individual queries too much.

Claudio Carpineto [13] present the relative ineffectiveness of information retrieval systems is largely caused by the inaccuracy with which a query formed by a few keywords models the actual user information need. One well known method to over- come this limitation is automatic query expansion (AQE), This survey presents a unified view of a large number of recent approaches to AQE that leverage various data sources and employ very different principles and techniques.

J. Kamps [14] present, during retrieval, our system initially operates just like a regular information retrieval system: given a query, our system will retrieve potential list of documents from the latest revisions index for scoring, and passes this initial list to the RHA module. In the future, we plan to further optimize the retrieval efficiency by precomputing the RHA term weights for all documents, instead of performing RHA analysis and re-ranking at retrieval time.

Massimo Melucci [15] present 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. Focuses on explicit RF and on pseudo RF. Implicit RF is based on observations (e.g., click-through data) that are proxies of relevance. The main problem with proxies is that they are not necessarily reliable indicators of relevance and thus should be considered noisy. How quantum detection can help "absorbe" noise can also be investigated in the future work.

III. PROBLEM STATEMENT

An IR system addresses the problems caused by query ambiguity by gathering additional evidence that can be used to automatically modify the query. Besides negativeness and positiveness, the RF algorithms can be classified according to the way the relevance assessments are collected. Feedback may be explicit when the user explicitly tells the system what the relevant documents and the irrelevant documents are top-ranked documents are considered as relevant documents, or it is implicit when the system monitors the user‘s behavior and decides what the relevant documents and the irrelevant documents are according to the user‘s actions.

- Irrelevant documents will be occurred in IR system.
- Ambiguity of automatically modify in query.
- Query not expanded for user‘s information need.
- Lag of information to the user‘s need.
- On explicit RF and on pseudo RF. Implicit RF is based on observations that are proxies of relevance.
- The main problem with proxies is that they are not necessarily reliable indicators of relevance and thus should be consider noisy.
- How quantum detection can help absorbed noise can also investigate in the future work. In general, RF and in methods inspired by quantum detection and integrate the retrieval functionalities of modern IR systems within a single learning to rank frame word.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

- These systems do no rely on only non-retrieval technology.

The vector space model has the following limitations:

1. Long documents are poorly represented because they have poor similarity values (a small scalar product and a large dimensionality)
2. Search keywords must precisely match document terms; word substrings might result in a "false positive match"
3. Semantic sensitivity; documents with similar context but different term vocabulary won't be associated, resulting in a "false negative match".

Quantum detection has following Limitations:

1. The order in which the terms appear in the document is lost in the vector space representation.
2. Theoretically assumed terms are statistically independent.
3. Weighting is intuitive but not very formal.

IV. MOTIVATION

Information Retrieval is the process of obtaining relevant information from a collection of informational resources. It does not return information that is restricted to a single object collection but matches several objects which vary in the degree of relevancy to the query. So, we have to think about what concepts IR systems use to model this data so that they can return all the documents that are relevant to the query term and ranked based on certain importance measures. These concepts include dimensionality reduction, data modeling, ranking measures, clustering etc. The tools that IR systems provide would help you get your results faster. So, while computing the results and their relevance, programmers use these concepts to design their system, think of what data structures and procedures are to be used which would increase speed of the searches and better handling of data.

V. AIM & OBJECTIVES

Aim

To maintain the collection of documents according to different user search. Find the document according to content of the documents.

Objectives

- To find query-document or document-document similarity. The reduction is not really substantial.
- To measure the performance relevance judgements more accurately and more quickly. Users can identify more relevant documents for each query, while at the same time make fewer mistakes.
- Implement the concept of relevant document suggestion.

VI. PROPOSED WORK

We are going to propose a IR system using which the user can easily get the relevance 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:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

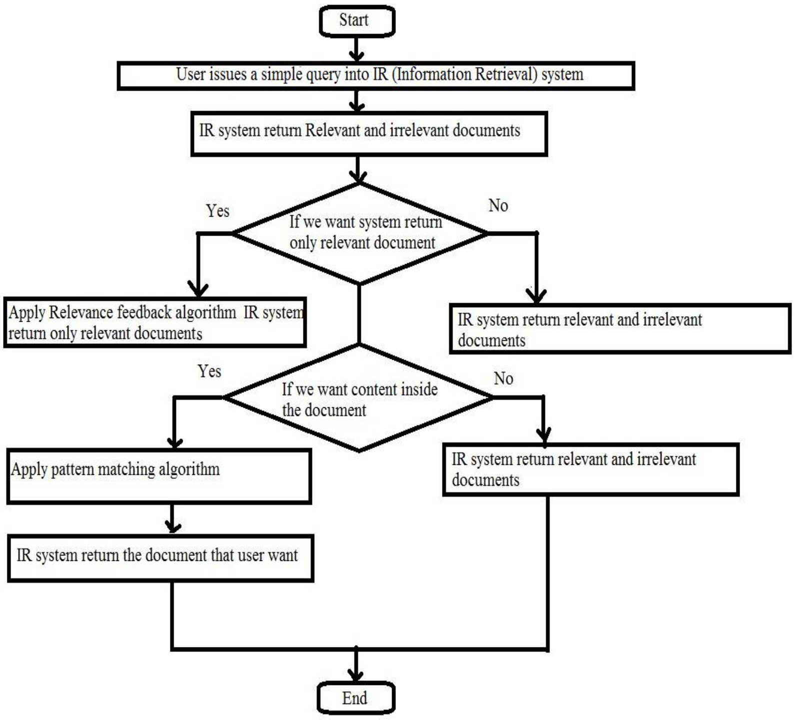

Fig No 1: Data flow diagram

1. The user issues a simple query into IR system.
2. The IR 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. The system displays a revised set of retrieval results.
5. If we want only relevant documents. Then we have to apply RF (Relevance Feedback) algorithm. After applying RF algorithm IR system return only Relevant documents by reducing the term found in irrelevant documents. Otherwise IR system return irrelevant document.
6. Sometimes user enter content inside the document as a query into IR system. IR system return relevant and irrelevant documents. But we want only document that user want. So pattern matching algorithm is used. Then IR system match the content that user enter with the contents inside the document and return relevant document.
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.

Pattern matching is the act of checking a given sequence of tokens for the presence of the constituents of some pattern. The pattern matching include,
1. User enter a query into IR system which represent data inside the document.
2. IR system extract document from the database .
3. IR system return relevant document that user want.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

VII. 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] C. Buckley and E. M. Voorhees,"Retrieval evaluation with incomplete information".In SIGIR'04, 2004.

[2] Ingo Frommholz, Birger Larsen, Benjamin Piwowarski,MouniaLalmas, Peter Ingwersen ,Keith van Rijsbergen, "Supporting Polyrepresentation in a Quantum-inspired Geometrical Retrieval Framework".IIiX 2010, August 18– 21, 2010.

[3] M. Shanmugham1, P. Logaiyan2, S. Annapoorana3, "Bearing Stimulated algorithms inspired by quantum detection", International Journal of Current Trends in Engineering &Research (IJCTER) e-ISSN 2455–1392 Volume 2 Issue 7, July 2016.

[4] AblimitAji, Yu Wang ,Eugene Agichtein, EvgeniyGabrilovich,"Using the Past To Score the Present: Extending Term Weighting Models Through Revision History Analysis".CIKM'10, October 26–30, 2010.

[5] Shuqin Liu,JinyePeng,"A Novel Image Retrieval Algorithm Based on Adaptive Weight Adjustment and Relevance Feedback ",JOURNAL OF COMPUTERS, VOL. 9, NO.11, NOVEMBER 2014.

[6] Diane Kelly,"Methods for Evaluating Interactive Information Retrieval Systems with Users ", Foundations and Trendsin Information Retrieval Vol. 3, Nos. 1–2 (2009) 1–224 c 2009.

[7] Luis M. de Campos, Juan M. Ferna´ndez-Luna ,Juan F. Huete"Implementing Relevance Feedback in the Bayesian Network Retrieval Model",JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 54(4):302–313, 2003.

[8] Y. Pang, Z. Ji, P. Jing, and X. Li, "Ranking graph embedding for learning to rerank", IEEE Trans. Neural Netw. Learn. Syst., vol. 24, no. 8, pp. 1292–1303, 2013.

[9] X. Tian, L. Yang, J. Wang, X. Wu, and X. Hua, "Bayesian visual reranking," IEEE Trans. Multimedia, vol. 13, no. 4, pp. 639–652, 2011.

[10] W. Liu, G. Hua, and J. Smith, "Unsupervised One-Class Learning for Automatic Outlier Removal",in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2014, pp. 3826–3833

[11] K. Collins-Thompson, P. N. Bennett, R. W. White, S. de la Chica, and D. Sontag, "Personalizing web search results by reading level", in Proc. 20th ACM Int. Conf. Inf. Knowl. Manage., 2011, pp. 403–412.

[12] Y. Lv, C. Zhai, and W. Chen,"A boosting approach to improving pseudo-relevance feedback", in Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2011, pp.165–174.

[13] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval", ACM Comput.Surv., vol. 44, no. 1, pp. 1–50, Jan. 2012.

[14] J. Kamps, S. Geva, and A. Trotman."Analysis of the inex 2009 ad hoc track results".In INEX, 2009.

[15] Massimo Melucci, "Relevance Feedback Algorithms Inspired By Quantum Detection",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 28, NO. 4, APRIL 2016.

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.