OPTIMIZING IR SYSTEM WITH RF ALGORITHMS USING QUANTUM MECHANICS

[1]V. Kamakshamma, [2]Nasreen Md

[1]Asst. Professor in Dept. of CSE, PBR Visvodaya Institute of Technology & Science, Kavali, Andhra Pradesh, India.

[2]M.Tech Student, Dept. of CS, PBR Visvodaya Institute of Technology & Science, Kavali, Andhra Pradesh, India.

Abstract – Information retrieving is deals with indexing documents including useful data relevant to user's data need. A class of algorithms for improving information retrieving and which consists of collecting other data representing the user's information need and automatically creating a new query.

This paper presents a class of new Relevance Feedback algorithms inspired by quantum detection to re-weight the query keywords and re-rank the relevant documents retrieved by an Information retrieving system. This paper utilizes the query vector on subspace spanned by the eigenvector which increases the distance between distribution probability of relevance and non-relevance respectively. This paper explains the how the algorithms classifying and retrieving the useful information.

Index terms: Relevance Feedback, indexing, information retrieving, classifying, collecting data.

I. INTRODUCTION

Data retrieving is concerned with indexing and recovers documents excluding information related to a user's information need. Although the end user can express his in order to need using a variety of means queries written in language are the most common means. The query is submitted to an data recovery system base on the vector space model [1]. This system would return to both related documents, and un-related documents the number of related documents top ten is quite high; there are some unrelated documents [2]. For example 14551F00La1 is ir-relevant .an information retrieval system addressing the problems caused by query ambiguity by gathering further evidence that mechanically modify the query.

The automatic process that transforms the user' queries is known as state of being relevant feedback. Some relevance assessment about the retrieved documents are collected and the query expanded by the terms found in the related documents ,reduced by the terms found in ir-related documents or re-weighted using related and ir-related and documents. Relevance feedback is both positive and negative or both [3]. Positive relevance feedback brings related documents and negative related feedback brings only ir-related documents. State of being relevant algorithm includes only a positive component, negative feedback is still problematic and requires further investigation, and some proposals have already been made

such as grouping ir-related s documents before using them for reducing the query [4]. Feedback may be explicit when the user explicitly tells the system what the related documents and un related documents is called pseudo. When the system decides the related documents and the unrelated documents are hidden [5].

When the system monitor's the user behavior and decides what the related documents and unrelated documents are according to the user's action.

II. BACKGROUND WORK

A. *Relevance Feedback*

The Relevance Feedback algorithm 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. Suppose y is the query vector, $x_1, \dots, x_R$ and R relevant document vectors in $\mathbb{R}^k$, and $x_{R+1}, \dots, x_N$ are $N - R$ non-relevant vectors in $\mathbb{R}^k$. The RF computes the following new query vector

$$y^* = \underbrace{\underset{Original\ query}{\tilde{y}} + \underset{Positive\ RF}{\tilde{y}^+} - \underset{negative\ RF}{\tilde{y}^-}}_{Modified\ Query} \quad (1)$$

B. *Vector Space Model*

The VSM for IR represents both documents and queries as vectors of the k – dimensional real space $\mathbb{R}^k$. this vector space is defined by k basis vectors corresponding to the terms extracted from a document collection.

The state of the art is given pivoted normalization which defines the following weight:

$$\frac{k_1 tf}{tf + k_1\left(1 - b + b\frac{doclen}{avdoclen}\right)} \log \frac{N - df + 0.5}{df + 0.5} \quad (2)$$

Where tf is the frequency of the term in the document, df is the number of documents indexed by the term, N is the number of documents in the collection, $doclen$ is the document length, and $avdoclen$ is the average document length; the parameters b = 0.75, and k_1 = 1.2 are constants for each term and document and their values.

The retrieval function is the inner product between a document vector x and a query y, and it is defined as

$$x'y \quad x \in \mathbb{R}^k \quad y \in \mathbb{R}^k \quad (3)$$

C. *Best Match No.25*

BM25 basically multiplies the Inverse Document Frequency (IDF) by a saturation component, thus obtaining the following weight:

$$IDF \frac{(k_1+1)tf(k_3+1)qtf}{(k+tf)(k_3+qtf)} \quad (4)$$

Where IDF = $\log \frac{N-df+0.5}{df+0.5}$, $k = k_1\left(1 - b + b\frac{doclen}{avdoclen}\right)$, k_3 is between 7 and 1000, qtf is the frequency of the term in the query.

The RF algorithm that is implemented in BM25 model consists of modifying the IDE component of the term weight. The Iterative process of RF begins with situation in which no relevance data are available and the term weight is (4). At the next step, the IDF is replaced by $\log \frac{r+0.5}{R-r+0.5} \frac{N-df-R+r+0.5}{df-r+0.5}$ where r is the number of training relevant documents indexed by the term.

D. Quantum Probability

A probability space is given by some observables and by a probability is the theory of probability developed within Quantum Mechanics. In QM, a probability space can be represented as vectors, matrices and operators between them. An account of quantum probability in the context of IR is provided in [7] and [8].

III. PROPOSED METHOD

This section illustrates the RF algorithms inspired by the principles of quantum detection. In summary, these algorithms build query vectors as the optimal detectors of quantum signal detection system. These optimal detectors have to decide the relevance state of a document on the basis of the available data. These algorithms project the original query vector on a special subspace which is given by the principles of quantum detection. The vector that results from projection is matched against the vectors of the documents by the inner product function expressed by (3).

The general RF algorithm inspired by the principles of quantum detection is shown in Fig. 1.

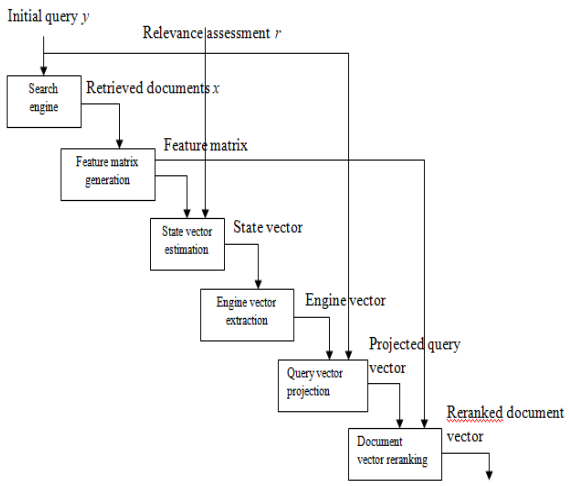

Fig. 1: General RF algorithm

A. Quantum Detection

Detection consists of identifying the information concealed in the data which are transmitted by the source places on one side, through a channel to the detector placed on the other side. The data are only a representation of the "true" information that one die wants to transmit.

A coder is placed between the source and the channel for encoding the signal into a particle which is assigned with a state vector. Each signal of a fixed finite alphabet is assigned a prior probability of emission and the coder does not intervene on the source, therefore each state has its own prior probability equal to the prior probability of the signal.

B. Optimal Detection

Detection of consists of defining the subsets of observable values that correspond to the states in which the signal can be sent through the channel and of minimizing the probability of error or maximize the probability of correct decision.

C. Detection of Relevance States

It is possible to correspond the theory of quantum detection too the theory of document retrieval – basically, relevance is a document state, a document can be viewed as a particle or signal, and query terms are viewed as observables.

In IR, the information that is relevant to the users information need is transmitted by a system to the user's information need is transmitted by a system to the user by means of a document which is only a representation of the information fulfilling the user's need. The values observed from a document can

serve to decide about the state of the document.

IV. IMPLEMENTATION

The algorithm that implemented in this paper step by step procedure is shown in below.

- Input:
  - No. of Points : n
  - No. of Dimensions: d
  - Point Set : P
  - Query Point: q
  - Family of Hash Family
    - Close points to same bucket
    - Faraway points to different buckets
    - Choose a Random function and hash *P*
    - Only store non-empty buckets
  - Hash q table
  - Test Every Point in q's bucket for Approximate Nearest Neighbor
  - If q's bucket may be Empty then
    - Use no. of hash tables
    - If any ann is found
    - Else
      - Poor resolution too many candidates
      - Stop after reaching the limit, small probability.
  - Want to find a hash function

$$if\ u \in B(q,r) then\ Pr[h(u) = h(q)] \geq \alpha$$
$$if\ u \notin B(q,R) then\ Pr[h(u) = h(q)] \leq \beta$$
$$r < R, \alpha \gg \beta$$

h is randomly picked from family

V. CONCLUSION

The paper presents RF algorithm inspired by quantum detection to re-weight query terms by projecting the query vector on the subspace represented by the eigenvector which is optimal solutions to the problem of finding the maximal distance between two quantum probability distributions. RF is then viewed as a signal detection technique – relevance is the document state to be detected and the queries are the detectors.

VI. FUTURE ENHANCEMENT

This paper mainly focuses 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 considered noisy. How quantum detection can help "absorb" noise can also be investigated in the near future.

REFERENCES

[1] Aji, Y. Wang, E. Agichtein, and E. Gabrilovich, ―Using the past to score the present: Extending term weighting models through revision history analysis,‖ in Proc. 19th ACM Conf. Inf. Knowl. Manage,2010, pp. 629–638.

[2] R. Blanco and P. Boldi, "Extending BM25 with multiple query operators," in Proc. 35th Int. ACM SIGI Conf. Res. Develop .Inf Retrieval 2012, pp. 921–930.

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

[4] K. Collins-Thompson, P. N. Bennett, R. W White S.Iachica. And D. Sontag, "Personalizing web results by level," in Proc. ACM Int. Conf. Inf Knowl. Manage.,2011, pp. 403–412.

[5] I. Frommholz, B. Larsen, B. Piwowarski, M. Lalmas,P. Ingwersen, and K.van Rijsbergen,"Supporting poly representation in a quantum- inspired geometrical retrieval

framework," in Proc. 3rd Symp. Interaction Context ins, 2010, pp. 115–124.

[6] I. Frommholz, B. Piwowarski, M. Lalmas, and K van Rijsbergen, "Processing queries in session in quantum Inspired IR framework," in Proc. Eur. Conf. Inf. Retrieval, 2011,pp. 751–754.

[7] M. Melucci. Introduction to Information Retrieval and Quantum Mechanics. Springer, in print.

[8] M. Melucci and C. J. van Rijsbergen. Quantum Mechanics and Information Retrieval, chapter 6, pages 125–155. Springer, Berlin, Germany, 2011.

AUTHORS

V. Kamakshamma was received her M.Tech degree in Computer science and Engineering from Visvodaya Engineering College, affiliated to JNTU, Anantapur. She was dedicated to teaching field since 12 years. She guided 10 P.G and 50 U.G students. At present she is working as Assistant Professor in PBR Visvodaya Institute of Science, Kavali, A.P, India.

Nasreen Md has received her B.Tech in University College of Engineering, Osmania University, Hyderabad in 2013 and pursuing M.Tech in Computer Science in PBR VITS, Kavali, A.P affiliated to JNTU, Anatapur in (2015-2017).