IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 11, November 2017

Relevance Feedback Algorithm Influenced By Quantum Detection

Shanmugha Priya B[1], Sharumathi R[2], Sruthi S[3]

Department of Information Technology, Sri Krishna College of Engineering and Technology

(An Autonomous Institution Affiliated to Anna University, Chennai.), Coimbatore, Tamil Nadu, India[1-3]

Abstrac: 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. In this paper, 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. 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.

Keywords: Information retrieval, Relevance Feedback, query vector.

I. INTRODUCTION

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. Consider topic 329 which is provided with the Text Retrieval Conference (TREC) test collection1 from which the query This system would return both relevant documents and irrelevant documents. An IR system addresses the problems caused by query ambiguity by gathering additional evidence that can be used to automatically modify the query . Usually a query is expanded because the queries are short and it cannot exhaustively describe every aspect of the user's information need; however, some irrelevant documents may be retrieved or relevant documents may also be missed when a query is not short. 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. RF has a long history: It was proposed in the 1960s ; it was implemented in the SMART system in the 1970s in the context of the VSM; it was investigated at the theoretical level; it eventually attracted interest from other researchers because of the consistent effectiveness improvements observed in many experiments. 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 includes 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 . 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, it is called pseudo when the system decides what the relevant documents and the irrelevant documents are (e.g., the top-ranked documents are considered as relevant documents), or it is implicit when the system monitors the user's behaviour and decides what the relevant documents and the irrelevant documents. are according to the user's actions (e.g., a document that is saved in the user's local disk is likely to be relevant). Although the potential can be large, pseudo RF can be unstable since it may work with some queries and it may not work with others, and therefore a system should learn how and when to apply it or not or to exploit some evidence such as term proximity. Query expansion is not the only means for refining the representation of an information need. An IR system might only re-weight the query terms and apply again the retrieval function using the re-weighted query. Either way, the IR system can re-rank the documents retrieved at the first run. Even supposing query expansion is in general more effective , term re-weighting can be still crucial because: it does not require disk accesses to the posting lists of the added query terms; it does not introduce noisy terms in the modified query; it can increase recall since the relevant documents ranked at the very lower positions of the retrieved document list can be moved to the highest ranks and made accessible to users; it is also crucial to increase precision since the non-relevant documents placed at the highest ranks after the first run might be moved to the lowest ranks. One major application is contextual search ; a contextual IR system may re-weight the query terms and then re-rank the documents

DOI 10.17148/IJARCCE.2017.61144

retrieved in the first run to fit the user's information needs according to some variables observed from the context such as the end user's reading level or the document's complexity. In this paper, we propose to replace the vector-space RF algorithms based on the VSM and the probabilistic algorithms based on the BM25 with algorithms inspired by quantum (signal) detection. We first propose to define signal detection in terms of quantum probability. As quantum probability generalizes classical probability, there are quantum probability distributions that cannot be defined within the usual theory of probability, thus allowing us to find optimal solutions which otherwise could not be found. The use of vectors and matrices in quantum probability allows us to seamlessly integrate our proposal in the VSM. Then, we define the optimal detectors from the setting prepared in terms of quantum probability distributions. The optimal detectors are the eigenvectors – which cannot be found by the theory based on classical probability – of a special matrix prepared from the quantum probability distributions. Finally, we project the query vectors on the eigenvectors found by the quantum probability distributions; the seamless integration of vector spaces and probability within a single quantum probabilistic framework allowed us to define a diverse set of algorithms. We also report on experiments to demonstrate the effectiveness of the RF algorithms inspired by quantum detection.

II. RELATED WORK

Xiang Sean Zhou* , Thomas S. Huang Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana Champaign, Initially developed in document retrieval (Salton 1989), relevance feedback was transformed and introduced into content-based multimedia retrieval, mainly content-based image retrieval (CBIR), during early and mid 1990ís (Kurita and Kato 1993; Picard et al. 1996; Rui et al. 1998). Interestingly, it appears to have attracted more attention in the new field than the previousóa variety of solutions has been proposed within a short period and it remains an active research topic. The reasons can be that more ambiguities arise when interpreting images than words, which makes user interaction more of a necessity; and in addition, judging a document takes time while an image reveals its content almost instantly to a human observer, which makes the feedback process faster and more sensible for the end user.

Tefko Saracevic School of Communication, Information and Library Studies, Rutgers University. Relevance is a, if not even the, key notion in information science in general and information retrieval in particular. This two-part critical review traces and synthesizes the scholarship on relevance over the past 30 years or so and provides an updated framework within which the still widely dissonant ideas and works about relevance might be interpreted and related. It is a continuation and update of a similar review that appeared in 1975 under the same title, considered here as being Part I. The present review is organized in two parts: Part II addresses the questions related to nature and manifestations of relevance, and Part III addresses questions related to relevance behavior and effects. In Part II, the nature of relevance is discussed in terms of meaning ascribed to relevance, theories used or proposed, and models that have been developed. The manifestations of relevance are classified as to several kinds of relevance that form an interdependent system of relevancies. In Part III, relevance behavior and effects are synthesized using experimental and observational works that incorporated data. In both parts, each section concludes with a summary that in effect provides an interpretation and synthesis of contemporary thinking on the topic treated or suggests hypotheses for future research. Analyses of some of the major trends that shape relevance work are offered in conclusions.

III. RESEARCH METHODOLOGIES

EXISTING SYSTEM

An IR system addresses the problems caused by query ambiguity by gathering additional evidence that can be used to automatically modify the query. Usually a query is expanded because the queries are short and it cannot exhaustively describe every aspect of the user's information need; however, some irrelevant documents may be retrieved or relevant documents may also be missed when a query is not short as shown in the previous example. 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.

- In Existing System, Usually a query is expanded because the queries are short and it cannot exhaustively describe every aspect of the user's information need.
- Some irrelevant documents may be retrieved or relevant documents may also be missed when a query is not short.

PROPOSED SYSTEM

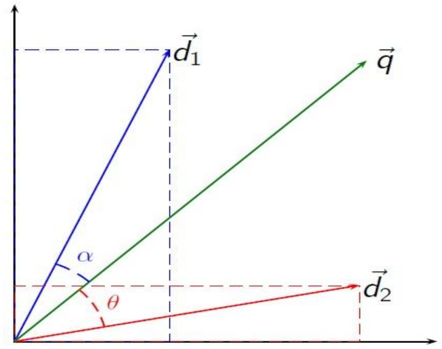
We propose a class of RF algorithms inspired by quantum detection to re-weight the query terms and to re-rank the document retrieved by an IR system. These algorithms project the query vector on a subspace spanned by the eigenvector which maximizes the distance between the distribution of quantum probability of relevance and the distribution of quantum probability of non-relevance.

In this section, we illustrate the main technical background of the framework proposed in this paper.

DOI 10.17148/IJARCCE.2017.61144

Vector Space Model:

Vector space model or **term vector model** is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings. Its first use was in the SMART Information Retrieval System.



Relevance Feedback:

Relevance feedback is a feature of some information retrieval systems. The idea behind relevance feedback is to take the results that are initially returned from a given query, to gather user feedback, and to use information about whether or not those results are relevant to perform a new query. We can usefully distinguish between three types of feedback: explicit feedback, implicit feedback, and blind or "pseudo" feedback.

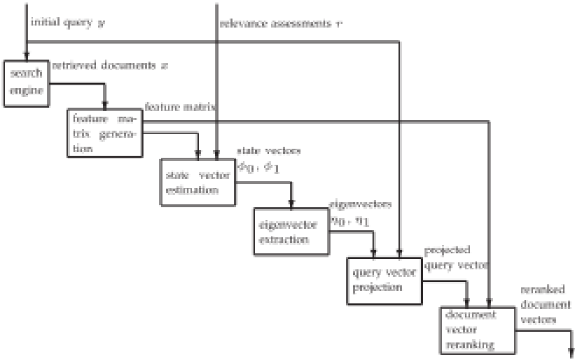
$$y^* = \underbrace{\overbrace{y}^{\text{original query}} + \overbrace{y^+}^{\text{positive RF}} - \overbrace{y^-}^{\text{negative RF}}}_{\text{modified query}},$$

Quantum Probability:

Quantum probability was aimed to clarify the mathematical foundations of quantum theory and its statistical interpretation.

A significant recent application to physics is the dynamical solution of the quantum measurement problem, by giving constructive models of quantum observation processes which resolve many famous paradoxes of quantum mechanics.

Some recent advances are based on quantum filtering and feedback control theory as applications of quantum stochastic calculus.


initial query y
relevance assessments r
search engine
retrieved documents x
feature matrix generation
feature matrix
state vector estimation
state vectors ϕ_0, ϕ_1
eigenvector extraction
eigenvectors η_0, η_1
query vector projection
projected query vector
document vector reranking
reranked document vectors

IV. CONCLUSION

In this paper, a class of RF algorithms inspired by 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. RF is then viewed as a signal detection technique – relevance is the document state to be detected and the queries are the detectors. First, the documents retrieved by an IR system to answer the original query are used to extract a feature matrix. Second, some relevance assessments are obtained according to whether RF is explicit or pseudo. The quantum probability distributions can be estimated and the optimal solution of a distance between two quantum probability distributions can be calculated. The eigenvector that results from this optimisation problem can be utilized to project the query vector. Third, the retrieved documents can be re-ranked to answer the modified query. The query term reweighting is different from the re-weighting performed by the classical RF algorithms since each query term variation depends on the other query term variations, thus capturing a kind of term dependence which is not captured by other RF algorithms.

DOI 10.17148/IJARCCE.2017.61144

V. FUTURE ENHANCEMENTS

This paper 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.

REFERENCES

[1] M. Melucci, Introduction to Information Retrieval and Quantum Mechanics. New York, NY, USA: Springer, 2015.

[2] M. Lalmas and I. Ruthven, “A survey on the use of relevance feedback for information access systems,” Knowl. Eng. Rev., vol. 18, no. 1, pp. 95–145, 2003.

[3] R. B. Griffiths, Consistent Quantum Theory. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[4] M. Gupta and M. Bendersky, “Information retrieval with verbose queries,” Found. Trends Inf. Retrieval, vol. 9, nos. 3/4, pp. 91–208, 2015.

[5] D. Harman, “Relevance feedback revisited,” in Proc. 15th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, Copenaghen, Denamrk, 1992, pp. 1–10.

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

[7] G. Salton, A. Wong, and C. Yang, “A vector space model for automatic indexing,” Commun. ACM, vol. 18, no. 11, pp. 613–620, 1975.

[8] A. Singhal, C. Buckley, and M. Mitra, “Pivoted document length normalization,” in Proc. 19th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 1996, pp. 21–29.

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

[10] 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.

[11] R. W. White and R. A. Roth, Exploratory Search: Beyond the QueryResponse Paradigm. San Rafael, CA, USA: Morgan & Claypool, 2009. [12] 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.

[13] C. Yu, W. Luk, and T. Cheung, “A statistical model for relevance feedback in information retrieval,” J. ACM, vol. 23, no. 2, pp. 273– 286, 1976.

DOI 10.17148/IJARCCE.2017.61144