Survey Paper on Information Retrieval Algorithms and Personalized Information Retrieval Concept

Parul Kalra Bhatia Tanya Mathur Tanaya Gupta

ABSTRACT
Personalized Information Retrieval systems (PIR) are of great need now a day. With growing size of database & requirement of precise data, PIR are of great importance. But this area is still being under research for the best methodology of searching. The PIR system instead of providing irrelevant data along with relevant one, provide us with just the possible relevant data matching our need & requirement.
In this paper a survey is done on different algorithms that are being worked on so far on PIR systems. Their drawbacks & new changes that can be inculcated. Different algorithms are being used to retrieve data in the PIR systems. Each algorithm was applied to the database & their results were noted. Then their drawbacks were noticed & some changes were made to overcome those.

Keywords
Personalized information retrieval (PIR), Information retrieval (IR), Page Ranks, Precision and Recall.

1. INTRODUCTION
With increasing data the need of its retrieval also arrived. To solve the need of retrieving the saved data a system was introduced called "information retrieval systems" (IRS). When given a query, the system gives results related to any word present in the Query. With time the system got evolved & undergone many changes. New methodologies were introduced to improve the system. But the biggest drawback of information retrieval system was that it gives thousands of results for a certain Query out of which only few are relevant and required by the user. This imprecision cause wastage of time and gives irrelevant data. The presence of that extra data can lead to skipping of the useful data.
To solve this problem IRS were planned to be personalized and hence called Personalized Information Retrieval systems. These systems are not much in use & are still under research. Many researchers are working on this system to find the best algorithm for searching a data. PIR systems will be so designed that they will provide precision & recall. In the following the user will get the data that will be precise & within his/her area of interest. He needs to specify the domain and purpose of his search. It gives search result as per the requirement of the user, sending query. During processing of a Query. The system will know the searching domain of the user and then provides as per his requirement. The data received is precise and can be recalled. This will help the user in getting best result for his query saving his time of searching & checking several non-required documents. In this paper a survey have been done on different works of different researchers. The algorithms frequently worked on & their consequences & results. Different algorithms that were implemented & worked on for retrieval of information in PIR.
The paper is divided in different sections with each section explaining different algorithm & their results with their negative & positive aspects.

2. TRADITIONAL INFORMATION RETRIEVAL ALGORITHMS.

2.1 Boolean algorithm
This is the basic model of information retrieval. Boolean model deals with using logical functions in the query to retrieve the required data. This is an early approach for data retrieval and is used as first model in finding information in the collection of data. This model is based on set theory and Boolean algebra; together they form a model for determining the data. Documents that are being searched in the database are sets of terms while Queries, given by the user are Boolean expressions on terms [1]. The terms are combined using AND and OR operators, where AND is intersection or logical product of any term and OR is union or logical sum of any terms. Combining terms with the OR operator will define a document set that is bigger than or equal to the document sets of any of the single terms. So, the query social OR political will produce the set of

Documents that are indexed with either the term social or the term political, or both, i.e. the union of both sets.[2]

The approach of Boolean model is as follows: Suppose, Document (D) =Logical conjunction of keywords. Query (Q) =Boolean expression of keywords and record, R (D, Q) = D ®Q

$$D = t1 \cup t2 \cup \ldots\ldots \cup tn$$

$$Q = (t1 \cup t2) \cup (t3 \cup \emptyset t4)$$

D ® Q, thus R(D,Q) = 1, [1]

This approach was simple and easy to implement. But the biggest drawback of this approach was that there was no concept of ranking and it gives only the exact match [1]. The users using the system or giving the query are not much familiar with Boolean terms and hence are not able to give the correct logical operators..

2.2 Ranking algorithm
Ranking algorithm was introduced to bring the concept of ranking. Since Boolean do not have ranking mechanism, it may skip important data, so there was a need of ranking. The result is ranked on the basis of occurrence of terms in the queries. This method eliminates the often-wrong Boolean syntax used by the end-users, and provides some results even though a term of the query is incorrect. It is not the term used in the data, it is misspelled. This methodology also works well for the complex queries that may be difficult for users to express using Boolean operators. For example, "human factors and/or system performance in medical databases" is difficult for end-users to express in Boolean logic because it contains many high - frequency or medium-frequency words without any clear necessary Boolean syntax. but the ranking model would do well with this query[3]. Page ranking

algorithms are used by the search engines to present the search results by considering the relevance, importance, the score of content and techniques of web mining to order them according to the user interest. Some ranking algorithms depend only on the link structure of the documents while some use a combination of both that is they use document content as well as the link structure to assign a rank value for a given document [4].

EXAMPLE: A simple illustration of ranking

USEFUL DATA HELP USER TASK RETRIEVAL ALGORITHM

QUERY: user useful in data retrieval algorithm

VECTOR: (1101011)

RECORD 1: contains user, useful, data, retrieval

VECTOR: (1101010)

RECORD 2: contains user, useful, help, algorithms

VECTOR: (1011001)

RECORD 3: contains useful, calculation, algorithm

VECTOR: (1000101)

SIMPLE MATCH

QUERY (1101011)

REC1 (1101010)

(1101010)=4

QUERY (1101011)

REC2 (1011001)

(1001001)=3

QUERY (1101011)

REC3 (1000101)

(1000001)=2

This model is of great use and is in use in information retrieval nowadays. Ranking model is convenient and user friendly and provides the data in chronological order. It has various approaches & methods. Next two models, that is, vector space and probabilistic uses ranking principle.

2.3 Vector based model

Vector space model is an algebraic document that uses vectors for representation. Documents and queries both are vectors.

$$d_j = (w_{1,j}, w_{2,j}, \dots\dots\dots, w_{t,j})$$

$$q = (w_{1,q}, w_{2,q}, \dots\dots\dots, w_{t,q})$$

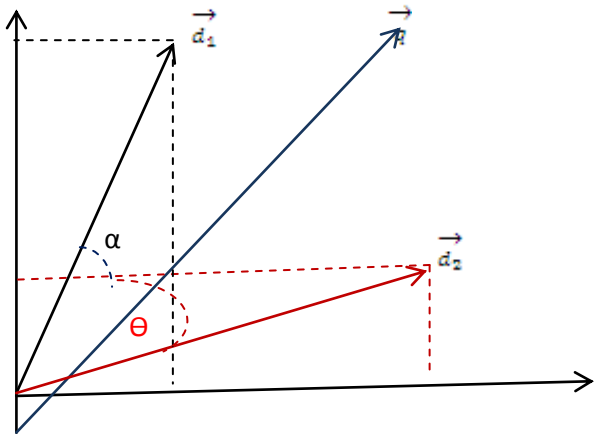

Fig1.Representation of different vectors

Where, d=document and q=query.

In this model distance of query vector and document vector is found out and their cosine gives an angle that determines the distance between them. Lesser is the cosine, ranking will be the higher. Where, sim(dj,q) is similarity between the documents.[5]This is a simple model based on linear algebra and no binary is being used. It allows continue measurement of distance between document and queries. The terms are weighted by importance giving partial matches. Due to smaller scalar product and large dimensionality in large documents, this model is not suited for them. Due to being semantic sensitive, false match may occur. Also it assumes terms to be independent.

$$\mathrm{sim}(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \cdot \|q\|} = \frac{\sum_{i=1}^{N} w_{i,j} w_{i,q}}{\sqrt{\sum_{i=1}^{N} w^2_{i,j}} \sqrt{\sum_{i=1}^{N} w^2_{i,q}}}$$

Probabilistic model

As the name suggest, this model is based on the probability theory of data. This model gives us the probability of retrieving the relevant document. Here the probability of retrieving the relevant data is matched with probability of retrieving the irrelevant data. This model is also based on ranking principle where, retrieved data is in ranked order.

It shows how optimum retrieval quality can be achieved. Optimum retrieval is defined With respect to Representations. The “Probability Ranking Principle" described in [Robertson 77] says that optimum retrieval is achieved when documents are ranked according to decreasing values of the probability of relevance (with respect to the current query).[6]

In this model, a matrix is developed comparing the relevant documents and irrelevant documents.

N = the number of documents in the collection

R = the number of relevant documents for query q

n = the number of documents having term t

r = the number of relevant documents having term t

t= any term in the query.[7]

r_i ***Relevant documents with t_i***	n_i-r_i ***Irrelevant documents with t_i***	n_i ***Documents with t_i***
R_i-r_i ***Relevant document without t_i***	N-R_i–n+r_i ***Irrelevant document without t_i***	N-n_i ***Document without ti***
R_i ***Relevant document***	N-R_i ***Irrelevant document***	N ***Samples***

$$p_i = \frac{r_i}{R_i} q_i = \frac{n_i - r_i}{N - R_i}$$

This is how probabilistic ranking is being done. Each document's probability-of-relevance estimate can be reported to the user in ranked Output. It would presumably be easier for most users to understand and base their stopping Behavior [i.e., when they stop looking at lower ranking documents] these models have achieved retrieval performance (measured by precision and recall) comparable to, non-probabilistic methods.

3. PERSONALISED INFORMATION RETRIEVAL

As we have discussed that Information Retrieval involves various algorithms and concepts to make retrieval of information efficient but when it comes to precision these algorithms might not be much fruitful. When we talk of searching a document or a word in it lets say apple what would one expect out of it? Fruit is the word that clicks our mind but there are few that have a different word apple Inc...So our retrieval precision depends on what the user want that is precision is high if he gets apple that he wanted that could be either of the two. Personalized information retrieval deals with such aspects where the user expectation is kept in mind. Major problem is how to build such system.

Study reveals that it can be done as follows:

1. Building a User Profile.

2. Using Evolutionary Algorithms.

3.1 Building a user profile

In this method a user profile was build, on the basis of which the system would generate the response. This profile would reflect the preferences and would depict the behaviour of the user.
"combine search technologies and knowledge about the query and user context into a single framework in order to provide the most appropriate answer.[10]
A user profile (or user model) is a stored knowledge about particular

user. Simple profile consists usually of keywords describing user's area of long time interest. Extended profile is replenished with information about the user such as name, location, mother tongue and so on. Advanced user profiles contain rather than set of keywords a list of queries characterizing user's behavior and habits [11].

It can be achieved by two methods:

1. Explicit Feedback
2. Implicit Feedback

Explicit Feedback taken from the user regularly. Feedback form depicting what a user wants and choices.

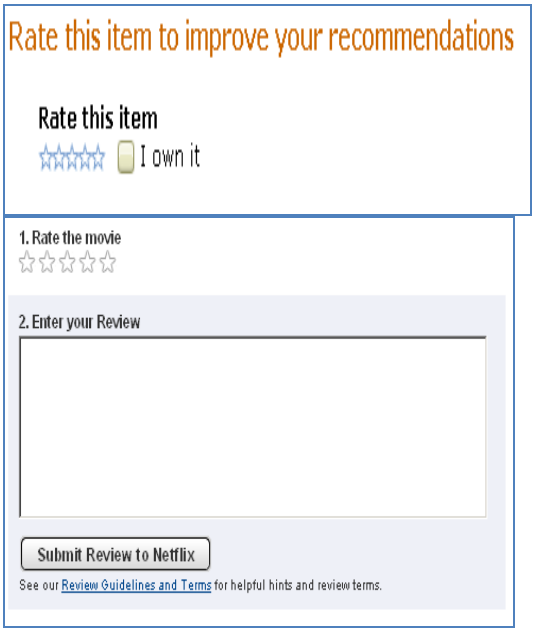

Fig2. A simple illustration of explicit feedback[16]

Limitations:
1. Users typically pose very short queries
2. This may be because

1. users have a difficult time articulating their information needs
2. traditional search interfaces encourage short queries .[16]

Implicit Feedback, it is users information, their needs and document preferences that can be unobtrusively obtained, by watching users' interactions and behaviors with systems

What are some examples?

- Examine: Select, View operation etc.
- Save: Email, Printing, Copy commands etc.
- Reference commands like Link and Cite
- Create: Type, Edit etc.

Why is it important?

It is generally believed that users are unwilling to engage in explicit relevance feedback It is unlikely that users can maintain their profiles over time Users generate large amounts of data each time the engage in online information-seeking activities and the things in which they are 'interested' is in this data somewhere.[16]

3.2 Evolutionary Algorithms

Evolutionary algorithms (EA) belongs to a family of iterative stochastic search and optimization methods based on mimicking successful optimization strategies observed in nature [12,13,14,15]. The essence of EAs lies in the emulation of Darwinian evolution utilizing the concepts of Mendel an inheritance for the use in computer science and applications [15]. Together with fuzzy sets, neural net-works and fractals, evolutionary algorithms are among the fundamental members of the class of soft computing methods. EA operate with population (also known as pool) of artificial individuals (referred often as items or chromosomes) encoding possible problem solutions. Objective function are used for evaluating encoded individuals, which assigns a value to each individual. The quality (ranking) of each individual is represented by fitness value, as solution of given problem. Competing individuals search the problem domain towards optimal solution [13].

The Core Algorithm

1. Crossover, Mutation, Reproduction
2. Fitness proportionate selection
3. Genetic Algorithms

Genetic Programming

Genetic Algorithms Genetic Algorithms (GA) introduced by John Holland and extended by David Goldberg are wide applied and highly successful EA variant. Basic workflow of original (standard) generational GA (GGA) is:

1. Define objective function
2. Encode initial population of possible solutions as fixed Length binary strings and evaluates chromosomes in initial Population using objective function.
3. Create new population (evolutionary search for better solutions)
a. Select suitable chromosomes for reproduction (parents)
b. Apply crossover operator on parents with respect to probability of crossover to produce new chromosomes (known as offspring)
c. Apply mutation operator on offspring chromosomes with respect to probability of mutation. Add newly constituted Chromosomes to new population.
d. Until the size of new population is smaller than size of current population go back to (a).
e. Replace current population by new population
4. Evaluate current population using objective function
5. Check termination criteria; if not satisfied go back to (3).[17]

4. CONCLUSION

The paper targets traditional and advanced algorithms that are generally used and researched upon. Information Retrieval Systems are used in every field and to personalize the retrieval, new algorithms are being worked upon. Different approaches are being researched upon for improving performance and efficiency of the Information Retrieval Systems. The Goal of the paper is to discuss these Algorithms in detail. This paper clearly explains and compares the algorithms and their limitations and why there is a need to focus on personalized information retrieval systems as retrieving information is a day to day phenomena and making it accurate and precise is what needs to be done.

5. REFERENCES

[1] "IR models: the Boolean model", http://www.csee.umbc.edu/~ian/irF02/lectures/06Models-Boolean.pdf.

[2] Djoerd Hiemstra, University of Twente "Information Retrieval Models".

[3] "Ranking Algorithm" http://orion.lcg.ufrj.br/Dr.Dobbs/books/book5/chap14.htm

[4] Ashutosh Kumar Singh, Ravi Kumar P "A Comparative Study of Page Ranking Algorithm for Information Retrieval."

[5] "IR models: Vector Space Model", http://www.csee.umbc.edu/~ian/irF02/lectures/07Models-VSM.pdf.

[6] Norbert Fuhr "Probabilistic Models in Information Retrieval". Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.

[7] Chapter 14:Ranking Algorithm, Donna Harman, National Institute of Standards and Technology

[8] Ed Greengrass "Information retrieval: A Survey".

[9] Djoerd Hiemstra and Arjen P. de Vries "Relating the new language models of information retrieval to the traditional retrieval models" University of Twente.

[10] M. Lalmas, A. MacFarlane, S. M. R¨uger, A. Tombros, T. Tsikrika, and A. Yavlinsky, editors. Advances in Information Retrieval, 28th European Conference on IR Research, ECIR 2006, London, Proceedings, volume 3936 of Lecture Notes in Computer Science. Springer,UK, April 10-12, 2006.

[11] Suhail S. J. Owais, Pavel Kromer, and V_aclav Sn_a_sel. Query Optimization by Genetic Algorithms. In DATESO, pages 125{137, 2005.

[12] Mehrdad Dianati, Insop Song, and Mark Treiber. An introduction to genetic algorithms and evolution strategies. Technical report, University of Waterloo,Ontario, N2L 3G1, Canada, July 2002.

[13] Gareth Jones. Genetic and evolutionary algorithms. In Paul von Rague, editor,Encyclopedia of Computational Chemistry. John Wiley and Sons, 1998.

[14] Melanie Mitchell. An Introduction to Genetic Algorithms. MIT Press, Cambridge,MA, 1996.

[15] Ulrich Bodenhofer. Genetic Algorithms: Theory and Applications. Lecture notes Fuzzy Logic Laboratorium Linz-Hagenberg, Winter 2003/2004.

[16] Diane Kelly,School of Information & Library Science, IPAM | 04 October 2007.

[17] Optimizing Information Retrieval Using Evolutionary Algorithms and Fuzzy Inference System V_aclav Sn_a_sel1, Ajith Abraham2, Suhail Owais3, Jan Plato_s1, and Pavel Kr□omer1 Department of Computer Science, Faculty of Electrical Engineering and computer Science, V_SB - Technical University of Ostrava, 17. listopadu 15, 708

[18] H. A. R. Townsend. Genetic Algorithms - A Tutorial, 2003