INFORMATION RETRIEVAL ANALYSIS ALGORITHMS IN DATA MINING WITH SOFT COMPUTING TECHNIQUES

S.Surya[1]
Ph.D Research Scholar,
Dr.P.Sumitra[2]
Assistant Professor,
PG and Research, Department of Computer Science and Applications,
Vivekanandha College of Arts and Sciences for Women (Autonomous),
Elayampalayam,Tiruchengode(DT),Namakkal(DT),TamilNadu,India.

ABSTRACT

This paper provides an in-depth research survey of intelligent information retrieval system from a huge database through a collection of web sites and web pages. Now a day there is an increase in challenge for complex domain in discovering information retrieval system. So there should be a high level of focus to be handled for possible ways in discovering the knowledge in web. The research area discusses on existing algorithms with advantages and disadvantages. Storing and fetching of the huge amount of data through the web server and a web client are the two main techniques involved in data mining process operation. The process operation with soft computing method gave a great solution in optimizing general search queries submitted by users.

Keywords: Data Mining, Soft Computing, Information retrieval System and Methodalogies.

1. INTRODUCTION

Computers are basically used for storing and accessing huge amounts of data or information both with online (internet) source or offline source. Now a days there are more a number of challenges faced by several IR systems.

IR – Information retrieval is a process of retrieving an actual set of information from an unstructured set of data or information in a huge database stored in a computer. According to developer IR is a problem oriented with respect to power and efficiency in converting required set of information. Under the research area, IR has three important roles:

a. Content Analysis: Document's capacity has been described by content analysis in a form that is related to the processing system.
b. Information Structures: Exploiting contact between documents to enhance the efficiency and effectiveness of recovery strategies.
c. Evaluation: the dimension of the effectiveness of retrieval.

Precision – Capability of retrieving top-graded documents that are widely relevant.

Recall – Capability in search for discovering all relative elements in the corpus.

The paper discusses the brief view of the information retrieval system concept and compare data mining and soft computing algorithm and related developed application. It also

describes the basic concepts in an information retrieval system, information retrieval application and algorithm and method of IRS through DM and soft computing.

2. INFORMATION RETRIEVAL SYSTEM

IR – Information retrieval is the main source of accessing information system resources which is related to the demand of information from a large collection of data. Either on content based indexing or full-text document, searches can be identified. IR is the process of discovering for information in a document, searching for script them, and also searching for metadata that describe information, and for databases of texts, images or sounds like as multimedia contents. Another process known as information overload is an automated process in IRS. This system is a software that keeps access to journals, documents and books. The IRS also stores the information and supervise the documents[1]. The IRS is classified into three major subsystems:

a. Representation of document
b. Representation of user's query
c. Match user's query

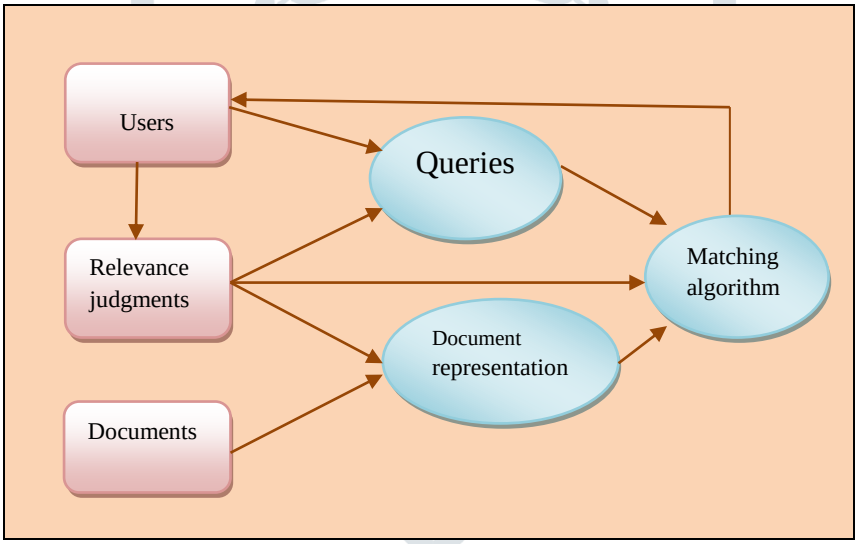

Figure1. Block diagram of IRS

A document collection considered of many documents whole information about verity topics of interests. Document contents are converted into a document representation automatically.

Relating queries in an easier way is accomplished with document representation. The primary responsibility in representation is how to select suitable index terms. Typically representation done by mining keywords that are desired as content finding plus systematizing in a useful format provided.

Queries convert the user's information demand into a specified form that accurately represents the user's identifies a required information that will be perfect in the process of matching. Basically query is used for the purpose of standardizing retrieval of information.

Kind's system performance restriction like precision and recall has been helped, to the effectiveness of the system in gathering users' information queries[1].

3. INFORMATION RETRIEVAL SYSTEM - APPLICATION

Information retrieval applications also require attributes like speed, accuracy, consistency and luxury use in recovering relative documents that content user queries.

Most of the field like Universities, Corporate sectors utilizes Information retrieval (IR) systems. IR systems are also utilized in general libraries for supporting access to EBooks, journals, and other documents[2]. Some of the applications that are commonly followed in an information retrieval system is described below:

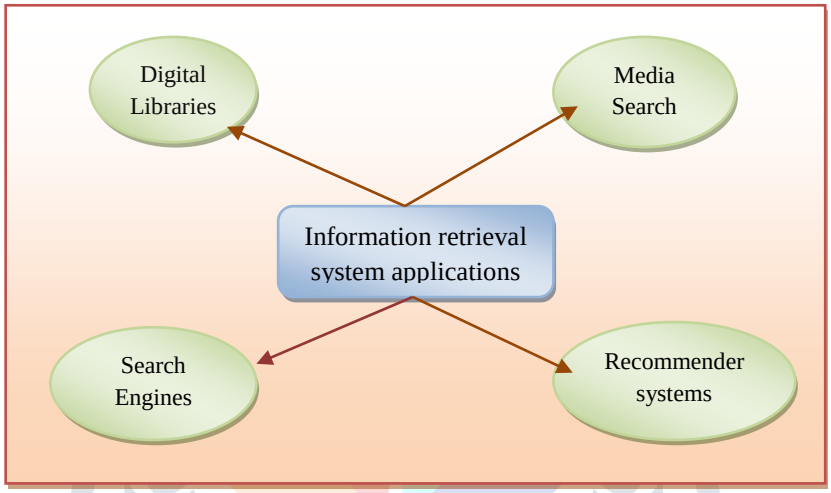

Figure 2. IRS Applications

3.1Digital Libraries

A group of data stored in a digital format that can be accessed by computer systems in a library is said to be digital library. With the help of the network all the digital content can be stored regionally or permit remotely. This digital library can be also mentioned as an information retrieval system[2].

3.2 Search Engines

The application which is used for information retrieval from a huge set of document and multimedia document is popularly known as search engines. Search engines are best known to be illustration, but there are searches continue like: Social search, Federated search, Enterprise search, Mobile search and Desktop search [2].

3.3 Media Search

Browsing the computer system, searching and retrieving images, figures from a huge database of digital images is known as an image retrieval system. Most historical and common functions of image find utilize some method of increasing metadata such as

highlight, keywords, or characterization to the images so that retrieval can be done the annotation words [2].

3.4 Recommender systems

Recommendation engines or Recommender systems work for a particular type of information filtering technique that effort to recommend information item like films, video on demand, music, eBooks, news, images, and web pages that are likely to be of wish to the user. These attributes can be from a user - information, user - social environment, data the mutual in filtering [2].

4. IRS ALGORITHM AND THROUGH VIEW OF DM AND SC

IR algorithms are several types to establish a line between each several applications. The IR system implements algorithms in three different types. Those types are categorized in the figure below.

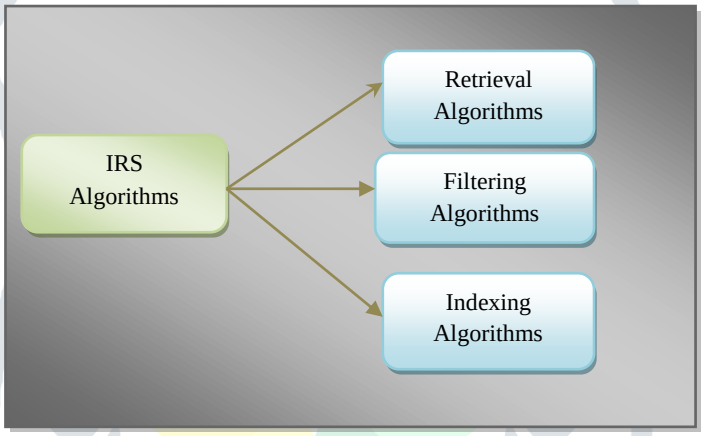

Figure 3. Block diagram of IRS algorithms

4.1 Retrieval Algorithms

To retrieve data from textual information is followed using the retrieval algorithm. Even IR can also be referred as retrieval algorithms. For analysis, there are two types of retrieval an algorithm followed, they are;

a. **Sequential scanning of the text**: In the case of searching string and some more searches memory can be excess in size of query but not with the extent of the database

b. **Indexed text**: Indexing is used for easy and speedy search within a large quantity of data. This is applicable to the usage of the Index.

The size of the index will generally equivalent to the size of the database in which the time of searching is sub linear in text size, inverted files and signature files are the examples for the same**.** Formally, we can explain a generic searching problem as follows: Given a string t, a regular expression, and information get by preprocessing the arranged and/or the text, the problem of finding whether t£∑*q£∑* (q for short) and get some or all of the following information:

1. The location where an instance (specifically the first, the longest) of q exists.

2. The number of instances of the design in the text.

As we deal with online information the query result should be effective and efficient in retrieval algorithm which is in turn a most significant role in the process. The demand in many fields has laid way for different retrieval algorithms in different approaches, by the way of hardware, parallel machines, and many more[2].

4.2 Filtering Algorithms

In this algorithm the input is kept as document and text that are filtered are stored in the output form. This process is said to be a kind's transformation under information retrieval, to quote an example, decreasing the text size and standardizing the same will make the search easy. The common filtering/processing operations as discussed below,

a. Literal that are used commonly will be removed with a stop words list.
b. Uppercase letters convert to lowercase letters.
c. Removal of special symbols and converting various sizes into a common size.
d. Numbers and dates convert to a standard format.
e. Word stemming (removing prefixes and/or suffixes).
f. Extracting keywords automatically.
g. Counting word rate.

In spite of above said information the filtering operations followed with filtering algorithms can also have some sort of disadvantage. Same query, prior to advising information in the database, should get filtered with the documents given plus the process is impossible for finding special symbols, common words and upper cases, nor it is not possible to analyze a text piece that has been mapped to the like internal form.

4.3 Indexing Algorithms

Indexing is a process of outlining the data structure which makes fast searching of the document or text, as we noticed previously. There are many categories of indices, based on various retrieval approaches. Almost all types of indices are common on some kind of tree or hashing techniques implemented [2].

4.4 A topic relevant to Information Retrieval is Data Mining algorithms

The Process of extracting knowledge from data is known as data mining. Data mining is an important technique used to convert this data into information. Ina broad variety of writing practices like marketing, fraud detection, surveillance and scientific discovery, the application is used commonly. An important portion of the process is the verification and validation of knowledge on samples of data or information [3]. Data mining involves the following kinds of tasks in IRS:

a. Classification – organized the data into predefined classes. For example, an email program may effort to classify an email as legitimate or spam. Common algorithms include as
    - Decision tree learning,

- Nearest neighbor,
- Naive Bayesian classification and
- Neural networks.

b. Clustering – Clustering algorithm will group same and similar items together, this process is done once after the classification.

c. Regression - Attempts to discover a method which models the data with the least error.

4.5 A topic relevant to Information Retrieval is soft computing algorithms

By using regular IRS, retrieval of meaningful information with the internet is getting delayed for a few tasks. The distinct soft computing techniques available for information retrieval systems make better efficiency in obtaining knowledge relevant to a user's query. The ultimate aim of the IRS is to fetch relevant information concerning a user's query. Soft Computing is a field in CS/IT manage with the mixture of procedure that are fashioned to model and give solutions to real world problems mathematically. Soft Computing aims to exploit the uncertainty, imprecision and exact reasoning so as to achieve low-cost result that are robust and tractable. Applying soft computing method is giving positive solution in raising the efficiency of IR systems [4]. With the effective use of fuzzy logic the main role of an IR system is focused and implemented. The primary responsibility of applying fuzzy set method to IR is:

a. How to construct the Boolean model, the query language and documents.

b. How to construct the associative structure like fuzzy clustering or fuzzy thesaurus.

4.5.1 Boolean Model

This model is the most basic information retrieval method and a disapproved one too. This method takes the query like an unambiguous description of document collection. Boolean model colleague a document and its collection of keywords as a part of query divided by ***AND, OR, NOT***. The retrieval function in the end examined if the document is related or not [4].

4.5.2 Vector space model

In the Vector Space Model, documents and query are presented an angle and vector in between two vectors are processed using the same cosine function as:

$$Sim(dj,q) = dj.q / |dj|. \quad (1)$$

This method has invented a ***term weight*** scheme called if-IDF weighting. The new scheme has weights described as below:

a. Term frequency (tf): This factor resolves quantity of times a term has appeared in the document / query.

b. Inverse document frequency (IDF): This factor determines the inverse of the quantity / number of documents which holds like query / document term. These vector space models are used to find term weight for query processing efficiency way by giving mathematical solutions. Soft computing techniques are used in IRS with fuzzy logic that implemented exact weight of user's query term.

CONCLUSION

The paper discusses on the IR, visible algorithm and target implemented soft computing with data mining multidisciplinary filed. It has enabled easier and quick information discovery. The task of finding information, some statistical techniques have indeed done to be the most effective ones so far. Techniques implemented in various fields with different areas have led to a way for new technologies that are in turn used by people on a continuous basis, for example news clipping, web search engines, junk email filters. Moving forward, the field is getting many constraints that user focuses on the information hidden world. With a tremendous growth of information in the offline or online source, information retrieval will play important role in the mere future.

REFERENCES

1) Introduction to Information Retrieval – Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze.
2) Information Retrieval Data Structures & Algorithms - William B. Frakes and Ricardo Baeza-Yates
3) Data Mining: Concepts and Techniques - Jiawei Han & Micheline Kamber.
4) Namrata Nagpal," Applying Soft Computing Techniques in Information Retrieval", International Journal of Advanced Engineering, Management and Science (IJAEMS), [Vol-4, Issue-5, May- 2018], ISSN: 2454-1311.