International Journal of Advanced Computer Technology (IJACT)
ISSN:2319-7900

A study and Review paper on Generic Comparison of Information Retrieval Methods

Mr. Rajesh Doss
Assistant Professor, Department of Computer Science
National Defence Academy
Khadakwasla, Pune - 411023, India
extrajesh@gmail.com

Dr. K.David
Associate Professor, Department of Computer Science
Hans Rover College of Engineering & Technology
Perambalur, Trichy - Tamilnadu, India
jdbdavid@gmail.com

***Abstract* : Data today is present everywhere in different forms. The ultimate role falls in searching and downloading or utilizing the correct and absolute form of data from web resources. According to a recent survey, most of students community would like to learn their queries from the blogs or download the required content from the internet and gain their knowledge. The knowledge gained from various source provides more conflicts at certain levels. The information uploaded in various blogs, spaces and unauthorized formats are not true or only up to the level of the blog creator where there is no editor or universal standards. In order to overcome these conflicts various data mining techniques are to be proposed for retrieving the appropriate source of information from the web based databases or blogs and would be compiled into a framework through knowledge Management systems.**

Index Terms – Information Retrieval, Knowledge Management, Web based Data Mining.

I. INTRODUCTION

The role of Internet in student's life is increasing every day. Its growth rate is inevitable because of quicker access to gaining updated knowledge. The Internet provides an easy structure to spread information worldwide. Web technologies support different standards (like HTML, XML, Web Services) in order to learn and exchange or transfer information efficiently. Even though, there are various educational tools and Environments, students prefer to learn through internet web pages, web spaces and blogs. But the tremendous growth of information in blogs and other unauthorized web spaces leads to acquiring wrong information or unbiased information that later proceeds to spoil student's or learners career. Currently, the problems faced by Learners with respect to students (Future Generation) are;

- How to retrieve information in a faster manner?
- Are the retrieved data relevant to learner's requirement?
- How to check the accuracy of the retrieved data?

Above all, the problem of accuracy creates more serious issue among general learners. There are various technologies and methods available in terms of concepts separately to solve the above said problems. Hence, this paper enumerates the compiled study of different concepts & approaches for interlinking the concepts to frame a knowledge management model for deriving a permanent solution for Correctness or trueness in the information acquired from web.

2. Information Retrieval and web searching Process

The power of learning increases when we get the right information at the right time and also an appropriate format for a given goal. Finding the right information has been researched extensively in the IR-field over the last decades. The search process became more elaborate with the apparent rise of theWeb. It led to the introduction of search engines such as GOOGLE which not only indexes (hyper)text, but also images, PDF-documents and interactive databases such as Citeseer (Citeseer, 1997). In other words, search engines attempt to retrieval relevant resources, rather than documents alone[12].

The importance of the timing aspect is particularly obvious when investment decisions are involved, such as on the blogs related with technical queries or information. Getting some information late could have huge consequences both in personal and financial development. Implementing a strategy for getting information in time often depends on many things such as choosing the right partner/supplier: some news sites are 'faster' then others in picking up news. The third aspect mentioned deals with formats in the broad sense. It refers to "file format" (e.g. PDF, or HTML) as well as "structural format" (e.g. "abstract", Loosely defined, a profile is the collection of all characteristics of a searcher that are relevant for the retrieval process. The goal of this position paper is to investigate how profiles can be used to improve the IR-process and define the model for correctness of information.

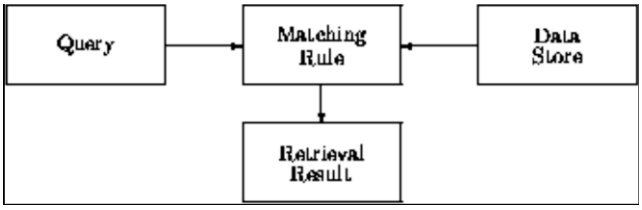


Figure 2.1- Basic Information retrieval (IR) Model

ISSN:2319-7900

The Figure above depicts the basic functions of any information retrieval (IR) system is relevance ranking: the (characterizations of) resources are ranked such that the resources that are "most relevant" are listed first, and the ones that are least relevant are listed last. In (Dhyani et al., 2002) an overview is given of metrics that are used to determine the relevancy of a Web-document with regard to a query. Furthermore, it is pointed out that relevancy involves more than topical relevance; other attributes of resources (such as its quality and price) are important as well[12].

2.1 Relevance

In (Gils et al., 2003) a conceptual model for information supply is presented. This model is based on the notion that similar information can be conveyed by multiple representations, leading to the notion that several representations (resources on the Web) can belong to a single information service (provide access to their underlying representations). Based on other's work, we define:

Definition 2.1 (representation format) It is enforced that each representation has exactly one type. Examples of these types are: PDF, HTML and Web service.

Definition 2.2 (structural format) Not all representations that belong to a single information service have to convey the same amount of information. For example, one conveys the "full content" and another is merely an "abstract". These (types of) structural format are modeled as feature types in (Gils et al., 2003). In this article we refer to them as the structural format. Using these definitions we can introduce our notion of relevance. Apart from topical relevance, which is the 'traditional' way of measuring relevance, we define that other constraints must be met as well. Examples of such constraints are its format (as explained in the previous section), but also price, quality etcetera. It may very well be that a searcher is willing to pay a certain amount of money in order to get his hands on a high-quality resource! Hence, we define relevance as follows:

Definition 2.3 (Relevance) Resources are relevant with regard to a query if and only if this resource meets all the criteria that a searcher poses on it. These criteria can be formulated in either the query, or the user-profile. This definition resembles the notion of functional versus non-functional requirements in Software Engineering (Sommerville, 1989).

Web-based educational technologies allow educators to study how students learn (descriptive studies) and which learning strategies are most effective (causal/predictive studies). Since web-based educational systems are capable of collecting vast amounts of student profile data, data mining and knowledge discovery techniques can be applied to find interesting relationships between attributes of students, assessments, and the solution strategies adopted by students. The focus of the study is three-fold: 1) to introduce various approaches for faster searching process; 2) to use clustering ensembles to build an optimal framework for clustering web-based assessment resources; and 3) to propose a framework for the discovery of interesting association rules within a web-based educational system. Taken together and used within the online educational setting, the value of these tasks lies in improving student performance and the effective design of the correctness of learning process [15]

3. Data Mining

Data Mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help individuals or companies to focus on the most important and highly sensible information in their data warehouses. Data mining tools predict on future trends and behaviours, allowing businesses to make proactive, knowledge-driven decisions. Most companies already collect and refine massive quantities of data[14]. The application areas of DM as contained in recent literatures as corroborated in Jiawei (2003) include: medical treatment/disease symptoms identification, retail industry, telephone calling patterns, DNA sequences, natural disaster, web log click stream, financial data analysis, bioinformatics, melody track selection, content-based e-mail processing systems, analyses of data from specific experiments conducted over time, analysis of nation's census database, and so on. DM techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. There are three groups of DM users namely, Application users, Designers and Theorists. It is usually common that the theorists based on some principal assumptions usually formulate new ideas. The most commonly used techniques in
Data mining are:

1. Artificial Neural Networks: this is a nonlinear predictive model that learns through training and resembles biological neural networks in structure.
2. Decision trees: tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
3. Genetic Algorithms: They are optimization techniques that use process such as genetics combination, mutation, and natural selection in a design based on concepts of evolution. It tries to mimic the way nature works. It is an adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetics.
4. Rule Induction: the extraction of useful if-then rules from data based on statistical significance.
5. Regression Methods: this tries to identify the best linear pattern in order to predict the value of one characteristic we are studying in relation to another.

3.1 DM tasks

Some of the tasks solved by Data Mining are:

ISSN:2319-7900

1. Prediction: a task of learning a pattern from examples and using the developed model to predict future values of the target variable.
2. Classification: a task of finding a function that maps records into one of several discrete classes.
3. Detection of relations: a task of searching for the most influential independent variables for a selected target variable.
4. Explicit modelling: a task of finding explicit formulae describing dependencies between various variables.
5. Clustering a task of identifying groups of records that are similar between themselves but different from the rest
of the data.
6. Market Basket Analysis: processing transactional data in order to find those groups of products that are sold together well
7. Deviation Detection: a task of determining the most significant changes in some key measures of data from previous or expected values[14]

3.2 Benefits of DM techniques

A company or an organization or an individual encompassing data mining techniques can enjoy a number of benefits; these includes understanding users or Learners' behaviour, making a judgement on the effectiveness of the individual blogs or a company's web site- if there is one, and benchmarking marketing campaigns (Doherty, 2000 & Mena, 1999).

3.2.1 Understanding users or Learners 'behaviour

The benefits that fall under this category are summarized below:

1. Establishing the probability of users or Learners coming back to the company or their web site or web blogs
2. Calculating the number of new users or Learners coming to the company or their web site or web blogs
3. Identify patterns relating either to navigation routes that users or Learners follow or to what they buy or learners look upon.
4. Discover whom byes what and look for any cross-relationships between clients.

4. Knowledge Management

Knowledge Management is the deliberate and systematic coordination of an organization's people, technology, processes, and organizational structure in order to add value through reuse and innovation. This coordination is achieved through creating, sharing, and applying Knowledge as well as through feeding the valuable lessons learned and best practices into corporate memory in order to foster continued organizational learning to leverage Knowledge to organizations advantage.

Need:

- Knowledge Management as a business strategy
- Transfer of best practices
- Customer focused knowledge
- Personal responsibility for knowledge
- Intellectual asset management
- Innovation and knowledge creation

The technology offer better perspectives of knowledge management

- Industrialization beginning in 1800
- Transpiration Technologies in 1850
- Communication Technologies in 1900
- Computerization Technologies in 1950
- Virtualization Technologies in early 1980
- Personalization and profiling technologies in 2000

With the above technologies and computers the knowledge management has come to mean the systematic, deliberate leveraging of knowledge assets. Technology enables valuable knowledge to be remembered via organizational learning and corporate memory and also enable value of knowledge to be widely disseminated to all users.Knowledge as asset – knowledge becomes increasingly more valuable than more traditional physical or tangible assets. Intellectual capital is often made visible difference between book value and market value.

Some critical challenges are to manage content effectively, facilitate collaborate, help knowledge workers connect and find experts and help the organization to learn and make decisions based on complete, valid and well interpreted data, information and knowledge We can view existing knowledge management applications from the perspective of tools implemented ([2]). From a wide range of existing tools we can pick several of them that could be purposefully deployed in e-learning systems as well[3]

4.1. Content management

We suggest giving a knowledge worker a wider opportunity to become a knowledge producer rather than pure knowledge consumer. Learning resources designers will no longer be the only knowledge providers to and learning resources will no longer be the only type of knowledge source in the knowledge repository. When proceeding through a learning resource, the knowledge worker masters its content by creation of her own knowledge. Having gone through a learning activity, a knowledge worker gains new knowledge related to the task covered by the learning activity as well as knowledge related to the (sometimes painful) process of mastering the topic knowledge. This process knowledge may

ISSN:2319-7900

be shared and reused by other knowledge workers of similar skill and competence profile undergoing the same learning process. In the current practice the resource content designer reflects the mastering process knowledge either by simply attaching it to the resource, e.g. in the form of FAQs, or by qualitative modification of the learning resource content and its organization

4.2. Advanced collaboration support

Besides adding mastering process knowledge that is eventually closely related to the learning material, a possibility to insert a knowledge piece expressing the knowledge worker's capturing, classifying or mapping of the newly acquired knowledge may be beneficial for other users. Similarly, it would be inspiring and stimulating to enable the users to contribute to the knowledge repository with a knowledge that was synthesized when collaborating on a problem solution or task performing. We also think that grouping of knowledge workers could be widely supported in e-learning systems encouraging thus cooperative learning activities that are not necessarily triggered by a specific learning task execution attached to a concrete learning resource. In a similar manner, managing learning resources as encapsulated knowledge resources could be revised ([3]). Freely browse able knowledge resources can be better decomposed into their atomic knowledge pieces and then combined in knowledge creation, dissemination, sharing and reuse.

4.3. User profiling

E-learning environment may become a large knowledge repository where a knowledge worker's orientation can get complicated. For a better convenience a push technology could be used based on the knowledge worker's interests and needs profile and knowledge resources marking. User profile may also capture the knowledge worker skills and competencies. They can be assessed with respect to expectations and be used for pushing special offers on knowledge resource collections ([4]). When enhancing the content, context and structure of learning resources with formal semantics, a flexible and customized knowledge mastering process can be created ([5]).

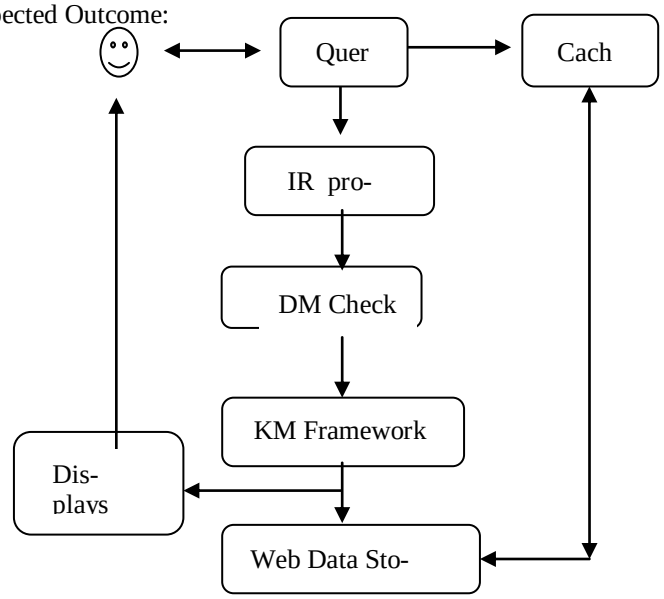
4.4. Data mining

Some knowledge management systems include a data mining tool. Adding data mining functionality to e-learning system may help to detect unknown patterns in user learning behavior, learning resources usage and knowledge mastering process bottlenecks.

4.5. Help-desk

In some e-learning applications a knowledge worker may simultaneously proceed through several learning resources. A typical example is an e-learning system implemented in higher education where a whole range of e-courses may be offered. In order to help the knowledge worker, i.e. the student, in overcoming problems he encounters during the learning process, a unified help-desk may be created to supplement direct consultations with tutors.

Expected Outcome:



The above picture depicts the combination or interlinking of different concepts and approaches required for faster and correct information. As per the flow of process when the user enters the query for searching the previous cache memory along with its web storage is compared with the keywords. If it is found the query displays the outcome of query. There are more possibilities of searching new concepts by learners. In this case, the query is generated and through information retrieval process and the obtained data pass through the DM techniques for checking a good or correctness of queried data. A new frame work is setup through KM methods for combining the retrieved data after predicting measures. The finally outcome is stored in the web storage for further access and process. However the above flow is to be tested with reality tools as a next step.

5. CONCLUSION

Data today is present everywhere in different forms. The ultimate role falls in searching and downloading or utilizing the correct and absolute form of data from web resources. According to a recent survey, most of students community would like to learn their queries from the blogs or download the required content from the internet and gain their knowledge. The knowledge gained from various source provides more conflicts at certain levels. The information uploaded in various blogs, spaces and unauthorized formats are not true or only up to the level of the blog creator where there is no editor or universal standards. In order to overcome these conflicts various data mining techniques are to be proposed for retrieving the appropriate source of information from the web based databases or blogs and would be

ISSN:2319-7900

compiled into a framework through knowledge Management systems. Hence, this paper enumerates the different concepts applied in different approaches are compiled as study in interlinking the concepts like Information Retrieval, searching methods and framing a knowledge management for deriving as permanent solution for Correctness or trueness in the information acquired from web The finally outcome is stored in the web storage for further access and process. However the above flow is to be tested with reality tools as a next step.

REFERENCES

1. Emilio ferrara, Pasquale de meo and giacomo fiumara, Robert Baumgartner, "Web data extraction, applications and Techniques: a survey", Acm computing surveys, vol. V, no. N, july 2012, pages 1{54
2. Shian-Hua Lin, Jan-Ming Ho, "Discovering Informative Content Blocks from Web Documents" , *SIGKDD '02*, July 23-26, 2002, Edmonton, Alberta, Canada.
3. P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar, *Member IAENG,"*Content Based Ranking for Search Engines", Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, IMECS 2012, March 14 - 16, 2012, Hong Kong
4. Neha Gupta, Dr. Saba Hilal, " A Heuristic Approach for Web Content Extraction", International Journal of Computer Applications (0975 – 8887), Volume 15– No.5, February 2011
5. G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma "Signed Approach for Mining Web Content Outliers" ,World Academy of Science, Engineering and Technology 56 '09
6. P. Sivakumar, R. M. S Parvathi, "An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining", European Journal of Scientific Research ISSN 1450-216X Vol.50 No.3 (2011), pp.340-351
7. G. Poonkuzhali, K.Thiagarajan, K.Sarukesi and G.V.Uma "Elimination of Redundant Links in Web Pages" ,World Academy of Science, Engineering and Technology 52 2009
8. Jaroslav Pokorny, Jozef Smizansky, "page content rank: an approach to the web content mining", Charles University, Faculty of Mathematics and Physics , Malostranské nám. 25, 118 00 Praha 1, Czech Republic
9. Fergus Toolan "Web Mining" Intelligent Information Retrieval, Group University College, Dublin
10. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008
11. W. Bruce Croft, Donald Metzler, and Trevor Strohman, *Search Engines: Information Retrieval in Practice*, Addison Wesley, 2009
12. Peter Brusilovsky, Jae-wook Ahn and Edie *Rasmussen "Teaching Information Retrieval With Web-based Interactive Visualization"* 25 July 2010
13. Prahaladrao.M "Knowledge Management", Defence Electronics research laboratory Hyderabad.
14. C. Romero, s Ventura," Educational Data mining: as urvey 1995-2005", Expert Systems with Applications 33 (2007) 135–146
15. Steve Dale ,"Engaging the Social Web for Personal Knowledge Management (PKM)",March 5, 2012