ISSN-2229-371X

Volume 2, No. 8, August 2011

Journal of Global Research in Computer Science

RESEARCH PAPER

Available Online at www.jgrcs.info

AN APPROACH TO MEASURE RESEARCH RELATED DOMAINS USING WEB-MINING TECHNIQUES

Nilesh Jain* Research Scholar, Mewar University, Chittorgarh(Raj.)
nileshjainmca@gmail.com
Dr Vijay Singh Rathore Professor & Director, Shree Karni Collge, Jaipur

Abstract— **as the time passes by, people are seeing more and more research activities in various fields including medicine, pure science, and technology and so on. The research works are published by the researchers in various journals and conferences, which are made publically available by the publishers. Though such research is spread all over the world, there should be a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research. Government and the other state organizations periodically need adaptation of certain new technology or process for improving or implementing certain policies. Such new requirements are generally met by calling for researchers to join in hand with the organization with their proposal. Filtering such proposals also takes a hectic schedule and thorough understanding of the researchers profile and to gauge his ability to complete the work. In order to solve this problem we emphasize on extracting meaningful information from the web through web mining techniques that helps understanding the region wise trends in research domain activities and further extract more meaningful information like patterns that suggest the progress in a particular area and prominent contributors in the area.**

Keywords:-webmining,AI,IR,KSOM

I. INTRODUCTION

Nowadays the World Wide Web has become an important medium for disseminating scientific publications. We all know that it is always helpful to know who the "experts" are working on a particular research area to conduct scientific research. Usually, many scientific publications are made available on the Web supported by individual researchers, research institutions and database (such as EI, SCI, SSCI, etc.) to share their research findings. However, it is apparent that information on expertise on research areas is hidden in the huge sea of information and cannot be extracted automatically from the Web. Hence research on finding expertise from Internet has become a meaningful task, attracting much interest from the academic community.

Web mining is the application of data mining techniques to extract knowledge from Web.

As the time passes by, people are seeing more and more research activities in various fields including medicine, pure science, and technology and so on. The research works are published by the researchers in various journals and conferences, which are made publically available by the publishers. Though such research is scattered all over the world, there is a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research. Government and the other state organizations periodically needs adaptation of certain new technology or process for improving or implementing certain policies. Such new requirements are generally met by calling for researchers to join in hand with the organization with their proposal. Filtering such proposals also takes a hectic schedule and thorough understanding of the researchers profile and to gauge his ability to complete the work. In order to solve this problem we emphasize on extracting meaningful information from the web through web mining techniques that helps understanding the region wise trends in research activities and further extract more meaningful information like patterns that suggest the progress in a particular area and prominent contributors in the area.

II. OBJECTIVE

The primary objective of this research is to build a citation-based indexing and retrieval system for scientific publications over the WWW. These publications often appear in some academic institution's Web sites in PostScript, PDF or HTML format. Currently, many intelligent systems for the retrieval of Web documents (in HTML format) have been developed, but not on Web scientific publications in PostScript or PDF format. Citation index can be used as a powerful search tool for scientific literature. This gives us the motivation to generate the citation indices of Web scientific publications and store them into a citation database. Through such citation indices, intelligent retrieval of Web scientific publications is possible.

To achieve this, we will be developing a citation-based indexing and retrieval system known as PubFinder. It consists of three major components, namely, Citation Indexing Agent, Web Citation Database and Intelligent Retrieval Agent. The Citation Indexing Agent automatically generates citation indices of Web scientific publications and stores them into the Web Citation Database. The Intelligent Retrieval Agent applies data mining techniques on the Web Citation Database to

support intelligent retrieval of Web publications. This project focuses on applying data mining techniques to the Web Citation Database for scientific publication retrieval. The Web Citation Database has been analyzed to investigate the possible knowledge that could be extracted. As most researchers are interested in scientific publications within certain research areas, identifying different research areas and authors from the same research area becomes one of the most important knowledge to be mined from the Web Citation Database.

Therefore, the mining tasks can be defined as document clustering and author clustering. One possible way to achieve this is through the use of clustering techniques.

Figure-1 shows the data mining work of this project, which is listed as follows:

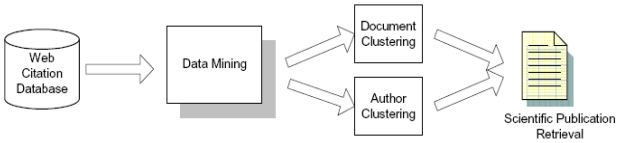

Fig-1 Data mining work

- *Mining for Document Clustering*. Data mining is applied to the Web Citation Database to group Web publications based on keyword similarities between them. Two kinds of neural network techniques, Kohonen's Self-Organizing Map (KSOM) (Kohonen, 1995) and Fuzzy Adaptive Resonance Theory (Fuzzy ART) (Carpenter *et al.*, 1991), are investigated.

- *Mining for Author Clustering*. Author Co-Citation Analysis (McCain, 1990) is incorporated into the data mining process to categorize authors into different research areas from the Web Citation Database. This is based on the assumption that if the frequency of two authors cited by the same publication is very high, these two authors may belong to the same or similar research field.

III. KOHONEN'S SELF-ORGANIZING MAPS

The Kohonen's Self-Organizing Maps (KSOM) (Kohonen, 1995) basically converts patterns of arbitrary dimensionality into the responses of one- or twodimensional arrays of neurons. The feature mapping can be thought of as a non-linear projection of the input pattern space on the neurons' array that represents features.

Learning within self-organizing feature maps results in finding the best matching neuron cells that also activate their spatial neighbors to react to the same input. After learning, each input causes a localized response having a position on the neurons' array that reflects the dominant feature characteristics of the input. The KSOM neural network training algorithm (Kohonen, 1995) is shown in Figure 2-2.

For each encoded input vector X, do step 1 to step 2 and repeat the same process for the whole training set for user-defined number of times:

1. Obtain the similarity measure between the input vector and the weight vectors of the output nodes, and compute the winner output node as the one with the shortest normalised Euclidean distance.

$$|X - W_{\infty}| = \min_i \{|X - W_i|\}$$

where W_i is the weight vector.

2. Update weight vectors as:

$$\Delta W_i(t) = \alpha(N_t, t)[x(t) - w_i(t)] \quad \text{for i} \in N_m(t)$$

where $N_m(t)$ denotes the current spatial neightbourhood, α is a positive-valued learning function, $0 < \alpha(N_t, t) < 1$. The function α can be respresented as:

$$\alpha(N_t, t) = \alpha(t) \exp\left(-\frac{|r_i - r_m|}{\sigma^2(t)}\right) \quad \text{for i} \in N_m(t)$$

where r_m and r_i are the position vectors of the winning cell and the winning neighbourhood nodes respectively, and $\alpha(t)$ and $\sigma(t)$ are suitable decreasing functions of learning time t.

Fig 2 Algorithm KSOM

IV. FUZZY ADAPTIVE RESONANCE THEORY

Basically, an Adaptive Resonance Theory (ART) network consists of two layers of units. The units contained in the first layer receive input from the outside world. Therefore, this layer is referred to as the feature representation field. The units contained in the second layer are used to represent the clusters of the input data. This layer is referred to as the category representation field. Weighted connections exist between every unit of these two layers.

The ART family consists of a series of models, including ART, ART1, ART2,

ART3, Fuzzy ART, etc. The first ART model was developed by Grossberg (Grossberg, 1986) to solve the problem of trade-off between continued learning and buffering of old memories (i.e. stability-plasticity dilemma). ART1 is the binary version of ART, which can stably learn to categorize binary inputs presented in an arbitrary order (Carpenter and Grossberg, 1987a). ART2 (Carpenter and Grossberg, 1987b), ART3 (Carpenter and Grossberg, 1990) and Fuzzy ART (Carpenter *et al.*, 1991) have been developed to handle multiple valued pattern vectors, that is, either binary or analog data. ART2 and ART3 may be computationally inefficient due to the need to iteratively normalize patterns (Carpenter and Grossberg, 1987b; Carpenter and Grossberg, 1990). The Fuzzy ART model is based on the fuzzy logic computations. It is capable to categorize arbitrary collections of arbitrarily complex analog input patterns effectively.

In our research, the input to the ART network is document vector, which contains analog data. Therefore, only ART2, ART3, and Fuzzy ART could be considered. Properties of learning for Fuzzy ART have been reported in (Carpenter *et al.*, 1991). One of the important properties is the short training time. Hence, Fuzzy ART is chosen to be the ART network model used in the research. Fuzzy ART incorporates computations from fuzzy set theory into ART1 systems by replacing the non-fuzzy intersection operator ( ) that describes ART1 dynamics (Carpenter and Grossberg, 1987a) by the

fuzzy AND operator ( ) of the fuzzy set theory.

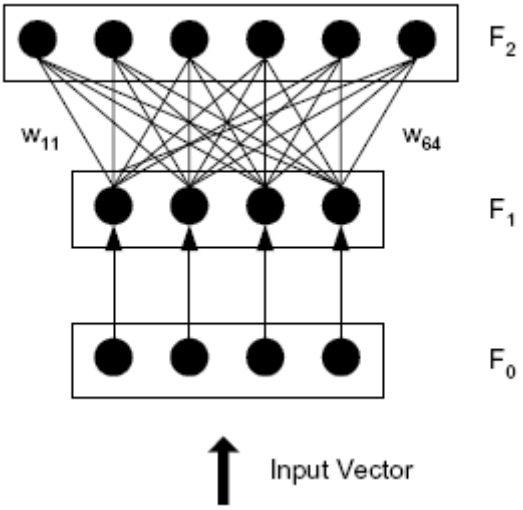

Fig 3 Architecture of Fuzzy ART neural network model.

Figure 3 illustrates the architecture of the Fuzzy ART neural network model. Each Fuzzy ART system includes a pre-processing field $_0F$, an input field $_1F$, and a category representation field $_2F$. $_0F$ modifies the current input vector, while $_1F$ receives both bottom-up input from $_0F$ and top-down input from $_2F$. If the original input vector is M-dimensional, then the $_1F$ field will have 2M nodes (as the original input vector needs to go through the complementary coding in the $_0F$ field), and the $_2F$ field will have N nodes, where N represents the maximum number of categories that the $_2F$ field can accomodate. Each of the N category nodes in the $_2F$ field have 2M connections with the $_1F$ field. Each node j in $_2F$ field has an associated vector W_j,distributed along the connections from that node to all the nodes in the $_1F$ field. W_j is called the weight vector for the jth cluster. Initially, before the learning occurs, all the weights in the vector W_j have the value 1 and each category node is said to be uncommitted. Only after a category node codes its first input, it becomes committed.

For each encoded input vector, do step 1 to step 4 and repeat the process for the whole training set until no change in the weights of the network.

1. Normalise the input vector to prevent category proliferation. The complemented coded $F_0 \rightarrow F_1$ input **I** is a 2M-dimensional vector:

   $$\mathbf{I} = (a, a^c) = (a_1, \ldots, a_M, a_1^c, \ldots, a_M^c)$$

   where $a_i^c = 1 - a_i$ for $i \in [1, M]$.

   A complemented coded input is automatically normalized, it is because

   $$|I| = |(a, a^c)| = \sum_{i=1}^{M} a_i + (M - \sum_{i=1}^{M} a_i) = M$$

2. For the input **I** and F_2 node j, the choice function T_j is defined by

   $$T_j(I) = \frac{|I \wedge w_j|}{\alpha + |w_j|}$$

   where the fuzzy intersection $\wedge$ is defined by $(P \wedge Q)_i = \min(p_i, q_i)$ and where the norm $|\ |$ is defined by $|P| = \sum_{i=1}^{M} |p_i|$.

   The system makes a category choice where at most one F_2 node can become active at a given time. The index J denotes the chosen category, where

   $$T_J = \max\{T_j : j = 1 \ldots N\}.$$

   The output vector **Y** of the field F_2 is set as $y_J = 1$ and $y_j = 0$ for $j \neq J$.

3. Resonance occurs if the match function of the chosen category meets the vigilance threshold, i.e.

   $$\frac{|I \wedge w_j|}{I} \geq \rho$$

   then the weight vector w_J is adjusted according to the equation:

   $$w_J^{(new)} = \beta(I \wedge w_J^{(old)}) + (1 - \beta) w_J^{(old)}$$

   Otherwise, mismatch reset occurs, where the value of the choice function T_J is set to 0. The search process continues until a chosen category meets the vigilance criteria.

V. FRAMEWORK

The proposed framework model will be implemented in a simulated environment to achieve the objective of the system so that the model can be referred to the real issues to implement the mining techniques. This phase uses the architectural document from the design phase and the requirement document from the analysis phase, to obtain and use the model. The implementation phase deals with issues of quality, performance, maintenance etc.

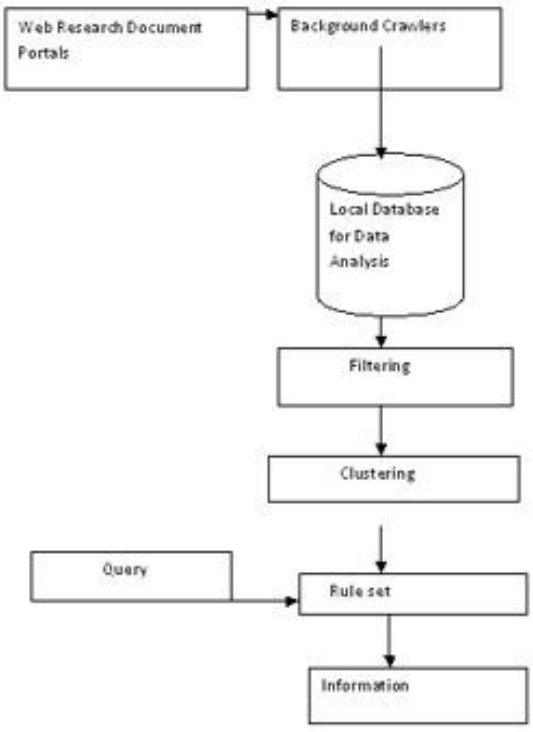

Fig 4-Framework for Research Domains

A. System Overview of Pubfinder

Figure 5 shows the system overview of the Pubfinder system. Citation Indexing Agent generates the Web Citation Database. It first downloads scientific publications from the Web. There are two ways to do this. The first method is similar to CiteSeer in that Citation Indexing Agent uses Web search engines (like AltaVista,Excite, HotBot, etc.) to search for pages that contain keywords such as "paper", "postscript", "publications", etc. Another way is to download the papers from the Web sites that are specified by the users. The downloaded papers are then parsed to extract the citations. The citation indices are generated subsequently and stored in the Web Citation Database. Intelligent Retrieval Agent applies data mining techniques to the Web Citation Database to discover the hidden relationships among the research publications and explore the useful knowledge that will help to improve the efficiency and effectiveness of the retrieval.

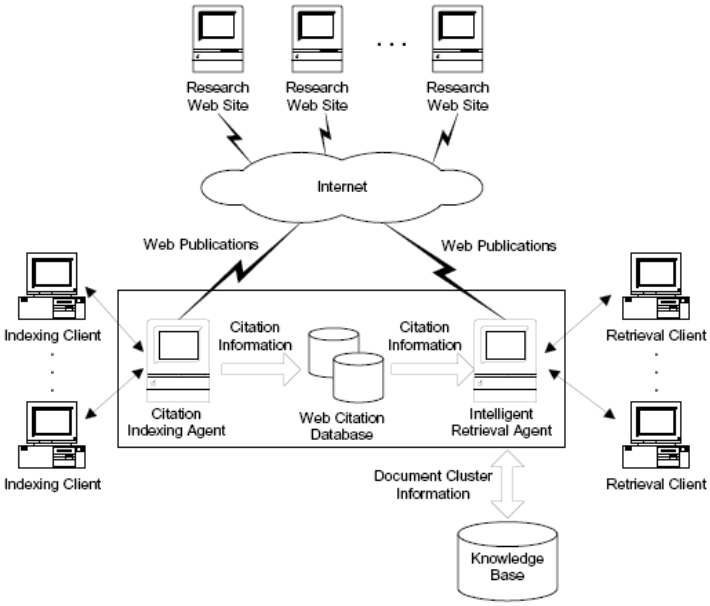

Fig 5 Overview of Pubfinder

VI. CONCLUSION

This research tackles by proposing and developing a framework for scientific publication indexing and retrieval system known as PubFinder. It consists of three major components: Citation Indexing Agent, Web Citation Database and Intelligent Retrieval Agent. The Citation Indexing Agent searches through the Internet to locate the possible Web sites that contain scientific publications, and then browse through these Web sites to parse the scientific literature. The citation information is then extracted and stored in the Web Citation Database. The synopsis focuses on discussing the Intelligent Retrieval Agent, which applies data mining techniques for document clustering and author clustering.

VII. RELATED WORK:

The clustering and rule set extractions are associated with knowledge discovery over the web. Therefore we focus on some of the key publications towards this area.

[1] States that The World-Wide Web provides every Internet citizen with access to an abundance of information, but it becomes increasingly difficult to identify the relevant pieces of information. Research in web mining tries to address this problem by applying techniques from data mining and machine learning to Web data and documents. The Web Mining is an application of Data Mining. Without the Internet, life would have been almost impossible. The data available on the web is so voluminous and heterogeneous that it becomes an essential factor to mine this available data to make it presentable, useful, and pertinent to a particular problem. Web mining deals with extracting these interesting patterns and developing useful abstracts from diversified sources. The present paper deals with a preliminary discussion of WEB mining, few key computer science contributions in the field of web mining, the prominent successful applications and outlines some promising areas of future research.

[2] Further elaborates that Web mining is the use of data mining technologies to automatically interact and discover information from Web documents, which can be in structured, unstructured or semi-structured form. The Web has become a major vehicle in performing research and education related activities for researches and students. There is tremendous amount of information and knowledge existing on the Web and waiting to be discovered, shared and utilized. We present an enterprise Web framework regarding semantic Web and mining in training institute, which can be used to not only improve the quality of Web mining results but also enhances the functions and services and the interoperability of educational information systems and standards in the educational field. Mining the educational information on the Web we are using new Semantic Web Mining technologies, such as Resource description Framework (RDF) and Web Ontology Language (OWL). For online educational institute Web site two important ontology's would need to be built one ontology describing all the educational services provided, with the relation between each other and the other ontology describing the Web site. Thus semantic Web ontology help build better Web mining analysis in educational institute and Web mining in-turns helps contract basis more powerful ontology in education.

[3] Extends the concept to education and explains that with the development of Internet, distance education platforms are very popular in today's society, but there are quite a few problems existent, such as the inadequate utilization of network teaching resources and the lack of individuation of the existed distance education platforms. To solve these problems have become the key of designing a good distance education platform. Web mining refers to the process of extracting useful data and information from Web sites or Web pages. In this paper we mainly discuss how to make use of Web mining technology to improve distance education platforms. We will introduce Web mining and its application in distance education platforms and propose a model of Web mining process in

distance education platforms.

[4] Present a preliminary discussion about Web mining, including its definition, the relationship between information mining and information retrieval on the Web, and the taxonomy and the function of Web mining. In addition, a prototype system called WebTMS (Web Text Mining System) has been designed.WebTMS is a multi-agent system which combines text mining and multi-dimensional document analysis to help users mine HTML documents on the Web effectively

[5] Paper introduces the web mining technology and the application of web mining in the long-distance education platform, points out the process of web mining, discusses the key techniques of personalized long-distance education platform applying web-mining technology. The study process of student are analyzed, the structures of teacher model and the structures of student model are given. The method in this paper can improve the quality of the long-distance education platform by way of practice.

[6] Finds that Web based educations have become the new growth point of education development and primary developing trend of modern education technology. To meet the personalized needs of web based education; an improved association mining rule was proposed in the paper. First, data cube from database was established. Then, frequent item-set that satisfies the minimum support on data cube was mined out. Furthermore, association rules of frequent item-set were generated. Finally, redundant association rules through the relative method in statistics were wiped off. The algorithm had two advantages, and the first was that the execution time was short while searching for the frequent item-set; the second was that the precision of the rules was high. The algorithm was also used in personality mining system based on web based education model. The result manifested that the algorithm was effective.

[7] Finds that personalized distance education system should be able to find the individual differences of learners and construct personalized learning environment to meet their individual needs. Web mining technology user in this paper aims to provide for each user personalized interfaces and services according as their individual characteristics. In this paper, a functional architecture of a personalized learning system based on Web mining is proposed, and providing learner with a personalized learning environment is a focus. The process of Web mining in this architecture is given and mainly includes four steps: (1) the collection of data, (2) the pretreatment of data, (3) the analysis of data and (4) the determination and generation of personalized output. Finally, we apply the proposed architecture to a typical education system and get satisfactory results.

The papers relevant to distance education are chosen to extend the concept to research and research oriented knowledge that helps not only the researchers to identify the key areas of work and the authors to look forward for but at the same time supports other organizations to easily identify the areas that they can look forward for for certain research goals.

REFERENCES

[1] Ramakrishna, M.T.; Gowdar, L.K.; Havanur, M.S.; Swamy, B.P.M.; Web Mining: Key Accomplishments, Applications and Future Directions, Data Storage and Data Engineering (DSDE), 2010 International Conference on

[2] Nayak, A.; Agarwal, J.; Yadav, V.K.; Pasha, S.;, Enterprise Architecture for Semantic Web Mining in Education

[3] Youtian Qu Lili Zhong Huilai Zou Chaonan Wang, Research about the Application of Web Mining in Distance Education Platform, Scalable Computing and Communications; Eighth International Conference on Embedded Computing, 2009. SCALCOM-EMBEDDEDCOM'09. International Conference on, 25-27 Sept. 2009

[4] Wang Jicheng Huang Yuan Wu Gangshan Zhang Fuyan, Web mining: knowledge discovery on the Web, This paper appears in: Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on,1999

[5] Wang Jian Li Zhuo-Ling, Research and Realization of Long-Distance Education Platform Based on Web Mining, Computational Intelligence and Software Engineering, 2009. CiSE 2009. International Conference on, 11-13 Dec. 2009

[6] Jin Huang Aiqin Zhu Qi Luo,Personality mining method in web based education system using data mining, Grey Systems and Intelligent Services, 2007. GSIS 2007. IEEE International Conference on,Nov-2007

[7] Yuewu Dong Jiangtao Li, Personalized distance education system based on Web mining, Educational and Information Technology (ICEIT), 2010 International Conference on, 17-19 Sept. 2010

[8] Agrawal R. and Srikant R. (2000). Privacypreserving data mining, In Proc. of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, 439-450.

[9] Berners-Lee J, Hendler J, Lassila O (2001) The Semantic Web. Scientific American, vol.184, pp34-43.

[10] Berendt B., Bamshad M, Spiliopoulou M., and Wiltshire J. (2001). Measuring the accuracy of sessionizers for web usage analysis, In *Workshop on Web Mining*, at the First SIAM International Conference on Data Mining, 7-14.

[11] Chen L, Sycara K (1998) A Personal Agent for Browsing and Searching. In Proceedings of the 2nd International Conference on Autonomous Agents, Minneapolis/St. Paul, May 9-13, pp132-139.

[12] *Kohonen* Self Organizing Map(KSOM), *rguha.net/writing/pres/ksom.pdf*

[13] *Fuzzy Adaptive Resonance Theory* with Group Learning *nlab.ee.tokushima-u.ac.jp/nishio/Pub-Data/CONF/C257.pdf*

[14] http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5380149

