

Efficient Productivity For Recommendation Of Semantic Web Pages

B.Gayatri Devi¹, S.Soundarajan²

¹PG Scholar, Department of M.E. Computer Science and Engineering,
Velammal Institute of Technology, Chennai, India.

² Vice Principal, Velammal Institute Of Technology,
Chennai,India.

Abstract: The objective of this paper is to provide personalized information by learning the user's interests from traces of interaction with that user. The major concept of web page recommendation is to provide a related set of documents for the user based on the query typed in by the user. The Existing system consists of the ontology to represent the domain knowledge, The automatically generated semantic network for relations and the conception prediction model(CPM) for the automatic generation of the semantic network. The proposed system provides more accuracy for offline viewing of the web pages. This system proposes a efficient method for faster and precise recommendation of web-pages to the user according to the user's preference. The required web-pages are hosted into the IIS .The admin can crawl the web-pages that are needed by the user. The web- pages that are uploaded by the admin can be viewed by the user henceforth. The Proposed system consists of four Modules, An IIS module for uploading of static web-pages, Realm Ontology for specifying terms and relationships, an Semantic Linkage module for finding the association between terms and web-pages, and PM module used for solving the "new page" problem..

Keywords: Semantic Network, Recommendation System, WUM, IIS, web-crawler, Mining

I. INTRODUCTION

As the World Wide Web continues to grow at an tremendous rate, the complexity of many web sites increases with it. For the users of these web sites it becomes increasingly difficult and time consuming to find the information they are seeking. User interfaces could help users find the information that is in accordance with their interests by personalizing a web site. Some web sites present users with personalized information by letting them choose from a set of predefined topics of interest. Users however do not always know what they are interested in beforehand and their interests may change. The recommendation System or Recommender System is a subclass of information-filtering system that seeks to predict the 'rating 'or 'preference' the user would give an item. Web-based recommender system greatly helps the web user's in navigating the web locating relevant and useful information and receiving dynamic recommendations on web-sites on products or services

that match their interests. The major concept is to provide a related set of documents based on the user's input query.

The recommendation system is divided into types: (i) content based filtering and (ii) collaborative filtering. The content based system examine properties of the items that are to be recommended. This filtering is also called as cognitive filtering and recommends items based on a comparison between the content of the items and a user profile. The content of each item is represented as a set of descriptors or terms, typically the words that occur in a document. The user profile is represented with the same terms and built up by analyzing the content of items which have been seen by the user. Collaborative filtering methods are based on collecting and analyzing a large amount of information on users' behaviours, activities or preferences and predicting what users will like based on their similarity to other users. A key advantage of the collaborative filtering approach is that it does not rely on machine analyzable content and therefore it is capable of accurately recommending complex items such as movies without requiring an "understanding" of the item itself. Many algorithms have been used in measuring user similarity or item similarity in recommender systems. This system is based on the assumptions that people who agree in past will agree in future too and that they will like the similar kinds of items they like in the past. The concept of data mining is used for the extraction of the web pages. This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

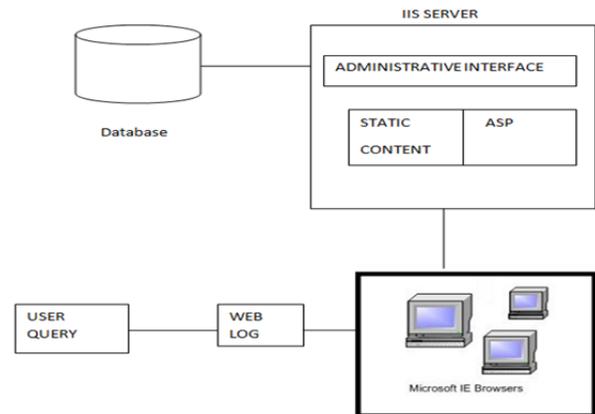
The popular algorithms that are used in recommendation system are as follows: (i) Pearson correlation and (ii) Clustering algorithms. The Similarity between two users (and their attributes, such as articles read from a collection of blogs) can be accurately calculated with the Pearson correlation. This algorithm measures the linear dependencies between two users or Variables as a function of their attributes. The Clustering algorithms are a form of

unsupervised learning that can find structure in a set of seemingly random (or unlabeled) data. In general, they work by identifying similarities among items, such as blog readers, by calculating their distance from other items in a featured space. The number of independent features defines the dimensionality of space. If the Items are “close” together, they can be joined as a cluster. Many clustering algorithms exist. The simplest one is k-means, which partitions items into k clusters.

II. RELATED WORK

Many models and algorithms have been proposed for the mining of the needed information from large database. One such method is a Semantic Web Mining that aims at combining the two areas that have been gaining popularity-Semantic Web and Web Mining. Web Mining is used for finding the insights meaning about the Web resources and their usage. The formalizations of the semantics of Web resources and navigation behaviour are being used increasingly. This fits exactly with the aims of the Semantic Web-the that which enriches the WWW by machine process able information and which supports the user in his tasks[1].Another method was developed known as dynamic clustering-based method to increase the accuracy of Markov model in representing a collection of user web navigation sessions. The method makes use of the state cloning concept to redundant the states in a way that separates in-links whose corresponding second-order probabilities diverge. Additionally, a clustering techniques used to determine a efficient way to assign in-links with similar second-order probabilities to the same clone[2].Another methodology used is to incorporate the conceptual characteristic of a web-site into recommendation model based on usage is the web [3].The algorithm proposed for the recommender system is web crawler, a simple crawler which uses breadth first algorithm.[4]this system also uses relevancycalculation algorithm and discards the irrelevant Web pages. The ontology concept can be used in various areas like delivery of courseware content[5],distance learning[6].Another approach as proposed that uses the concept based approach for adding the semantics to the process of mining[7].A new process was developed for finding out the patterns that are sequential in the web usage data for the prediction of the next occurring web page[8].New algorithms have been proposed for the mining of the web sequential patterns ,The RePL4UP(Revised PLWAP For Update)scans only the additional information .The PL4UP(PLWAP for Update) constructs a bigger tree that includes all the sequences in the database with tolerance support[9].An algorithm known as longest common subsequence has been initiated for classifying the activities of the user and the predicting the user’s next movement[10].The retrieval of the information is the major concept. The traditional Information retrieval assumes the basic unit is document. In the internet these units are known as the web pages. The information retrieval means that finding out the relevant set of documents based on the input query by the user[11].

III. PROPOSED SYSTEM



The query entered by the user is first checked in the web log to see if the user is present or not .This process is to check if the user is new or needs to be registered. If the user is old then the recommendation is given to the user based on the previous searches done by the user. The IIS is an service that is provided by the windows operating system that allows to upload the web-pages. The web-pages are static and are saved in asp format. Then the uploaded pages are linked to the recommendation system. Using the realm ontology module the gist of the web page uploaded is understood. The next module is the semantic Linkage module, This module is used to find out the relationship between the words and is used during the searches if the searches are made by the related words and not the actual one. The last module is known as the PM(Prediction Model), This model/module is used for the incorporation of the knowledge on a specific domain and the network that holds the relations between the various keywords, web-pages and the domain terms. The main advantage of the PM is that it eliminates the new page problem, if the user is searching the web-page for the first time ,then the previous traditional recommendation systems cannot offer any recommendations to the new user. Therefore the usage of the Predicion Module is to provide the recommendations to the user even if the user is querying for the first time and irrespective of the face that user’s preference is not present in the web access sequence, The first time user is able to view the recommendation without any trouble. The administrator has the rights to add the web-page to the recommendation system, view the log-in and log-out time by the registered user’s, and add the new web page to the recommendation system according to the user’s preference or request given by the user. The key extraction algorithm is proposed for the better, faster And accurate recommendations of web pages to the user

The related keywords are stored in tabular format. When the user searches a keyword, the tabular column is crawled first for better efficiency. Therefore the search results are fast and accurate according to the user’s needs.

1)IIS(Internet Information Service):

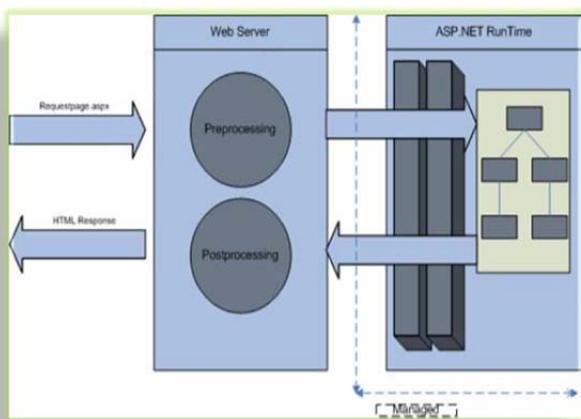
The Internet Information Service(formerly known as Internet information Server) is an extensible web server created by Microsoft for use with Windows NT family.IIS supports HTTP, HTTPS, FTP, FTPS, SMTP and NNTP. It has been an integral part of the Windows NT family since Windows NT 4.0, though it may be absent from some editions (e.g. Windows XP Home edition).IIS is not turned on by default when Windows is installed. The IIS Manager is accessed through the Microsoft Management Console or Administrative Tools in the Control Panel. The Microsoft Internet Information Server is designed to deliver high speed and secure information publishing, while also serving as a platform for developers and independent software vendors to extend the Internet's standard communication capabilities.

IIS at root level, is responsible for processing requests received on specific ports. For it, a service called the World Wide Web Publishing Service runs on the system.The internal processing of IIS is given below.

a)internal working of IIS:

There are various way to host a web application on IIS. Visual Studio has some inbuilt features to host and create a virtual directory on IIS directly. For deploying an application, we create a Virtual directory and the convert it into an application. Actually Virtual Directory is nothing more than a configuration entry in IIS for sharing a physical path for access through Web Server

When the page is Requested in the browser,The IIS server preprocesses the request and sends the request to the ASP.NET run time. The ASP(Active Server Pages) runtime then searches the web page that has been requested.After Postprocessing of the response the web page is given to the user as desired.The web pages that are requested are to be uploaded in the IIS beforehand



b) updating static web-pages on IIS:

The steps required for the updation of the static(asp.net) page to the IIS for web hosting are as follows:

- Enable the FTP(File Transfer Protocol) authentication on the IIS.When this option is enabled,the FTP membership based authentication is done and it provides more security and deployment options to the administrator.
- In the application pool,go to site, click on default site and go to add application .Give alias name(ex-

pages) and linj the path to the place where the domain is present.

- To give permission for all the user's right click the page(alias name for newly created application) and go to edit permission. Under security tab, click on edit and add the users to everyone and click ok.

2) realm onto module:

The most imoportant concept used in the ontology is the user's session.The session contains the informations about the navigation of the web-pages visited by the user during a specific time, the main keyeord of the web-site that can be extracted from the title of the page.By collecting the above details the domain knowledge is gained from the previously visited web-pages.This gained knowledge is used in the realm ontology for the web page recommendation in effective manner.An onto is used for finding out the relationship between the terms (can be explicit or formal).

a)components of the realm onto module:

The main components of the realm onto module are as follows:

- i. The concepts (or domain terms)
- ii. Relations between the concepts
- iii. The main features of the terms and the relations

The ontology implemenation is done using OWL/RDF (these are logic based language).Therefore information can be shared amongst the web servers.Reusing of the of the domain knowledge is done.The information that is represented as ontological of the discovered knoweldge is integrated for better recommendation.

Based on the ontology purpose,they can be classified into various degree of foramlity which can be in the form of conceptual data models,taxonomies etc.Here the conceptual data model approach is being used.This approach uses two aspects.one is that web page is an group of objects that are represented by the HTML tags.these tags are present in meta data .This is helpful in the core terms of the web-page,its structure etc can be understood.The other aspect is that the TITLE tag of the web page contains the important keyword that is short and precise for the page .therefore these tags are given hogher weightage. Giving the accurate title for the page is important as it defines the identity of the web page during the searching or crawling.

b)construction :

the realm onto module can be constructed in three steps.

- Collection of the related terms: To collect the terms ,(i)check the log file from the web server of a web site,(ii)run a pre-processing unit to the log file to find out the list of URL's and sites accessed by user's,(iii)crawl the web pages.
- Definition of concepts the The domain concepts for a web-site are defined based on the extracted terms .There is a possibility that the terms extracted share same features therefore the terms can be instances of a concept.
- Define taxonomic and non-taxonomic relationship.

The taxonomic relationship can be developed using three approaches:

- ✓ Top-down approach
The process starts from generalized to specialized concepts of a domain.
- ✓ Bottom-up approach
Starts from specific concepts and groups them into generalized concepts.
- ✓ Hybrid approach
Its combination of both the top-down and bottom-up approaches.

The non-taxonomic relationship can be defined as: 'provides' relation, the 'has' relation, the 'hasPage' relation etc.

c) crawling algorithm:

A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. Other terms for Web crawlers are ants, automatic indexers, bots, and worms or Web spider, Web robot, or Web scutter.

This process is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches.

Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also, crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for spam). A Web crawler is one type of bot[5], or software agent. In general, it starts with a list of URLs to visit, called the seeds. As the crawler visits these URLs, it identifies all the hyperlinks in the page and adds them to the list of URLs to visit, called the crawl frontier. URLs from the frontier are recursively visited according to a set of policies.

3) semantic linkage module:

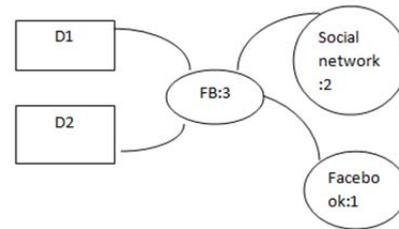
The semantics network of a web site is a knowledge map that depicts the domain terms, Web-pages, and Relations including the collocations of domain terms, and their associations between terms and web pages. The domain terms are gathered from the titles of the web pages assuming that the title of the web page is an effective one. Then the relations between the terms are extracted by using the following two techniques: (i) the collocation of the terms and (ii) the associations between the terms and web-pages. Additionally the domain terms and the co-occurrence relations are weighted to approximately find out how much the terms are semantically related to each other. Based on these relations, the strength of relations between the web-pages can be inferred. Using this, queries can be generated for finding out the relationship between the terms and web-pages, which is for finding the relevant pages for a web-page. This semantic network is referred to as SemNet.

a) construction :

To automatically construct SemNet, the following 4 steps are done:

- Collect titles of previous visited web-pages
- Collect the web-log file and run pre-processing unit to produce list of URL's. Crawl the list for titles.

- Extract the term sequences from web-pages
- Apply the key-extraction algorithm and organise the terms in order they appear.
- Build SemNet
- Each node represents a term and the before and after links determine the sequence of relation between the terms. The rectangle represents the web-pages (as shown in the diagram).
- Implement automatic construction of SemNet.
- The language to implement the SemNet is OWL. This code can be reused by other parts of the recommendation system.



4) Prediction Module:

For finding out the semantic Web usage knowledge that is efficient for Web-page recommendation, a prediction model (PM) is used to automatically generate a weighted semantic network. This generates the weighted semantic network of the terms that have been viewed frequently. The weight is for the transition between two terms based on Frequently viewed term pattern.

For the prediction the Markov model is used which is efficient to depict the collected navigation records. This module has two kinds of elements: (i) state node and (ii) the relation between the nodes. One node is the current state (the term viewed currently) and may contain the previous state or the next state nodes. Additionally there are two more states: (i) start state and (ii) end state.

IV. CONCLUSION

The offline recommendation system developed provides efficient and accurate results that satisfy the user's needs. The recommendations provided by the system are according to the user's preference and this system narrows down the related web-pages according to the user's satisfaction. The administrator can check the user's session details and if necessary can add a new web-page to the domain. The proposed key-extraction algorithm adds an efficiency to the semantic Linkage module thus providing a better recommendation to the system by finding out the correct web-page even when the user searches the system with related keyword. The system provides a better precision and is found to satisfy the user needs in a better method. The recommendation system proposed provides a better recommendation by predicting the next web-pages in accurate and according to user's needs.

REFERENCES:

- [1] J. Borges and M. Levene. (2004). "A Dynamic Clustering- Based Markov Model for Web Usage Mining," Technical Report .
- [2] G. Stumme, A. Hotho, and B. Berendt), "Usage mining for and on the Semantic Web," in *Data Mining: Next Generation Challenges and Future Directions*. Menlo Park, CA, USA: AAAI/MIT Press, pp. 461–480.
- [3] J. Borges and M. Levene., "Incorporating Concept Hierarchies Into Usage Mining Based Recommendations.", in *Proc. 8th WebKDD*, Philadelphia, PA, USA, pp. 110–126.
- [4] Raman Kumar Goyal, Vikas Gupta, Vipul Sharma, Pardeep Mittal "Ontology Based Web Retrieval", WEBKDD'06, Philadelphia, Pennsylvania, USA.
- [5] S. Boyce and C. Pahl, "Developing Domain Ontologies For Course Content", , "Educ. Technol. Soc.", vol. 10, no. 3, pp. 275–288
- [6] Sinéad Boyce and Claus Pahl, "On The Development Of Domain Ontology For Distance Learning Course", *Educational Technology & Society*, 10 (3), 275-288.
- [7] S. A. Rios and J. D. Velasquez (2008), "Semantic Web usage mining by a concept-based approach for off-line web site enhancements," in *Proc. WI-IAT'08*, Sydney, NSW, Australia, pp. 234–241.
- [8] S.T.T.Nguyen (2009), "Efficient Web usage mining process for sequential patterns," in *Proc. IIWAS*, Kuala Lumpur, Malaysia, , pp. 465–469.
- [9] C. Ezeife and Y. Liu (2009), "Fast incremental mining of Web sequential patterns with PLWAP tree," *Data Min. Knowl. Disc.*, vol. 19, no. 3, pp. 376–416.
- [10] Mehrdad Jalali, Norwati Mustapha, Ali Mamat, Md. Nasir B Sulaiman, "A Recommender System for Online Personalization in the WUM Applications", *Proceedings of the World Congress on Engineering and Computer Science 2009 Vol II WCECS 2009*, October 20-22, 2009, San Francisco, USA.
- [11] B. Liu, "Information retrieval and Web search," in *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, B. Liu, Ed. Berlin, Germany: Springer, 2011, pp. 183–236.