RESEARCH ARTICLE OPEN ACCESS

Web-Page Recommendation Using an Enhanced Incremental Sequence Mining Algorithm Along With Ontology

Gauri Sonawane [1], Prof. Dr. Mrs. S. A. Itkar [2]
Department Of Computer Science and Engineering
PES Modern College of Engineering
Savitribai Phule Pune University
India

ABSTRACT

Web page recommendation is a process to recommend appropriate web pages to the user according to the user interest.When user is on a webpage they should get a proper recommendation so that they gain relevant results. Appropriate knowledge discovery from Web usage data and correct representation of that knowledge for successful Web-page recommendation is important. The paper presents a technique to give better recommendations through semantic enhancement by combining the domain and Web usage data of a website. Here, domain ontologies are used to provide conceptual understanding of a particular domain and an incremental mining method,e.g. PLWAP for Update (PL4UP), can be utilized to update web access patterns which are frequent called as (FWAP),which are discovered from the Web usage data. Two important modules are presented.The First is Web Usage Mining which uses the user access sequences which comes from the web logs and gives the most frequent sequences. The second model utilizes ontology to represent the domain knowledge. Also the conceptual prediction model, called as Termnavnet is a navigation network of domain terms and frequent web access patterns is used for supporting Web-page prediction.

Keywords :- Ontology,PLWAP,PL4UP,Web-page recommendation, FWAP.

I. INTRODUCTION

Webpage recommendation is obtaining great interest in the web world were web-pages are present in an infinite quantity. Webpage recommendation has become popular as users get relevant links of various webpages and website without searching repeatedly. When a user makes a search for a website, a sequence of visited Web-pages during a session can be created. A session is the period from starting, to exiting the browser by the user. This sequence is arranged into a Web session S. The goal of a Webpage recommendation system is to efficiently forecast the other pages that will be visited from a given current Web-page of a website. There are a various problems in building an efficient recommender system, such as how to efficiently learn from available historical information and search important knowledge, how to model and use the discovered knowledge, and how to make efficient Webpage recommendations based on the discovered content.Research has been carried to resolve these problems over the past years. It has been said that the approaches based on tree structure and probabilistic models can effectively demonstrate Web access sequences (WAS) in the Web usage data[1].With the help of these approaches it is possible to construct the transition links between Webpage.If,given the current visited Webpage (say a state) and k previously visited pages (the previous k states), the Web-pages that will be visited in the next navigation step can be forecasted. The working of these approaches is based on the sizes of training datasets. Prediction accuracy depends on the dataset size that means bigger the size of dataset higher accuracy should be gained. However, these approaches make Webpage direction to completely base on the Web access sequences learnt from the Web usage data. Therefore, if a user is visiting a Webpage that is not in the Web access sequence, then these approaches are unable to offer any direction to this user. This issue is referred as ’new page problem’.This problem can be overcome with the help of ontology.

II. RELATED WORK

The work in [2] shows a better technique for webpage recommendation where three modules are proposed.One is the ontology and semantic network module which represents the domain knowledge of that website.Another is the module which works for generating and analyzing the access sequences by the users which uses the PLWAP Mine algorithm for generating frequently accessed patterns.This is the WUM module .The next one is the CPM module which works probabilistically and helps in webpage prediction.The modules are combined and a better recommender system is constructed.

Here [3] a better technique of sequential pattern mining is discussed and implemented.The technique is the WAP-tree mine algorithm.It shows the working of WAP-Tree i.e how the tree construction and mining is done.The main task is to assign the binary position code to every node so that rescanning of every node can be avoided and the task can be performed in less time.

Mining patterns which are sequential is a process to apply data mining techniques to a database to discover the relationship between ordered list of events.Here[4] the web usage mining is focused as an application to sequential mining technique.

Increasing information on the web is generating a continuous flow of data.Web pages create a data sequence which is continuous and which flow in the web log time to time,which cause the need to update the earlier sequence patterns generated by a mining process.Many algorithms are present which generate web access sequence patterns some of them are WAP-tree,PLWAP-tree etc.PLWAP algorithm gives good results in terms of time and memory but is not very good when an incremental update happens in the web sequences. This [5] work introduces two algorithms RePLWAP and PL4Up which updates sequential patterns avoiding the scan of entire database even if small itemsets become frequent.

A new algorithm called Sequential Stream Mining Algorithm SSM is introduced [6].It is based on tree structure.It handles the issues of mining the frequent sequential patterns in data streams. The issues include the inability of various algorithms to carry multiple scans of a streamed dataset which degrades the performance in terms of result and the result may not be accurate as on demand. There is a very limited attention in the recent work towards stream sequential mining.The algorithm SSM is a better contribution for this.

Ontology is a name and definition of entities which tells about the type,interrelationships,properties of the entities that exist for a domain.It can be said as a model of Knowledge for a particular domain or group of domains.In terms of web,ontology can be used to define the user profiles in terms of information gathering.This [7] work creates an ontology of user profiles global base and local repository.Ontology can be constructed manually or it can be automatic.

III. SYSTEM FLOW

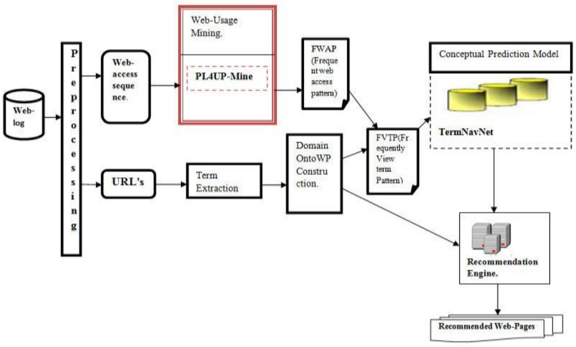

Fig.1 System Architectural Diagram.

The System works as follows:-

A. ***Weblog***

The raw data is in a web log,recording information about the surfing histories of the Web users.The weblogs contain the web access sequences and the set of URL's which are further separated in the preprocessing phase to pass the content to further modules.

B. ***Preprocessing***

The pre-processing is done on the weblogs. Web logs contains the information about the users tracking history in the form of access sequences and set of URL's. These Web access sequences and URL's are separated and passed to the further modules.

C. ***Web Usage Mining***

It is the technique used to discover and to analyze pattern which are usually the web access sequences which are frequently visited by the web users. Here, the frequent web access sequences of web pages are discovered so that a better recommendation can be performed. An advanced and popular incremental usage mining technique, named Pre-ordered linked web access pattern for update(PL4UP-Tree Mine) which is a technique used instead of PLWAP-mine in this system to discover the Web usage knowledge, which gives Frequent Web access patterns (FWAP),i.e. patterns of frequently visited Web-pages.

D. ***Domain Ontology Construction***

1) Collect the terms from the web log-.Collect the Web log file from the Web server. Run a pre-processing unit to analyse the Web log file and produce a list of URLs of Web-pages that were accessed by users.
2) Define the concepts.

3) Define taxonomic and non-taxonomic relationships between the concepts.

E. Conceptual Prediction Model

This gives the probability of each page linked with other page. According to this probability web-pages are recommended.

F. Recommendation Engine

It will actually recommend the web pages to the user.

IV. ALGORITHMIC ASPECT

This section tells about the algorithmic approach used i.e the algorithms used in the recommendation system.

1) PLWAP-Mine Algorithm[5]

This is the sequential pattern mining algorithm which uses the tree structure. It constructs the tree taking the WASD which is a database of web access sequences and a minimum support value as input and generates FWAP as output.

Input:WASD web access sequence database, minimum support.

Output: Complete set of frequent patterns.

begin:

i. The PLWAP algorithm computes frequent 1-items from the database transactions as F1 = {a:5, b:5, c:3}, listing each event with its occurrence. It generates frequent sequences from each transaction.

ii. Using the frequent sequences, it builds the PLWAP tree by inserting each sequence from Root to leaf node.

iii. Mining the PLWAP tree to generate frequent pattern,by following the header linkage of the first frequent item.

2) PL4UP-Mine Algorithm[5]

This algorithm is proposed to overcome some of the limitations of PLWAP-mine. It is used for mining sequence patterns incrementally. These patterns are based on PLWAP structure. The main idea is to avoid the scan of whole database when there is any updation in it.PL4UP algorithm initially builds a bigger PLWAP tree which is based on a lower tolerance support percentage t%, which is lower than the regular given minimum support percentage.

Algorithm PL4UP-Tree()

Input: original database (DB),Incremental database(db),minimum support percent λ ,tolerance support θ ,TFP^{DB},SFP^{DB}, patterns based on t and s, old candidate lists (C_1, F_t , S_1, PF, PS).

Output: updated frequent patterns for updated database, U (TFP’ and SFP’),
updated candidate lists ($C_1{}'$,$F_t{}'$, S_1', PF', PS').

```
begin
      i. Update all candidate lists as follows:
    $C_1{}' = C_1 \cup C^{db}$; s'= $\lambda$ of $|DB| + |db|$; t'= $\theta$ of $|DB| + |db|$
    $F_t{}'$= element in $C_1$' with support $\geq$ t'
    $S_1$'= element in $C_{1'}$ with support < s'
    PF = elements in $S_1$' with support $\geq$ t'
    PS' = elements in $S_{1'}$ with support < t'
    $F_t^{db} = C_1^{db} \cap F_t$
    $S_1^{db} = C_1^{db} \cap S_1{}'$
    $PF^{db} = C_1^{db} \cap PF'$
    $PS^{db} = C_1^{db} \cap PS'$
    ii. If
     $PS^{db} \cap F_1{}' = \acute{\O}$
         then
     -Construct small $PL4UP^{db}t$ on tolerance support t, using
        $Ft^{db}$.
    -Mine to obtain the frequent patterns, $SFP^{db}$ and $TFP^{db}$
    -Combine the two tolerance frequent pattern to obtain
        TFP' and SFP' as:
      $TFP' = TFP_t^{DB} \cup TFP^{db}$
      $SFP'$ = patterns inTFP' with support $\geq \lambda$
        end
```

V. MATHEMATICAL MODEL

Let s be the system having input, functions and output.

S = { I, F,O }

Where,
I is a set of all inputs given to the System,
O is a set of all outputs given by the System,
F is a set of all functions in the System.

- I ={Web-log}

  The web-log is pre-processed and a set of sequences and URL's are extracted.

  i. $I_1 = \{a_1,a_2,a_3....a_n\}$(web access sequence)

  ii. $I_2 = \{u_1,u_2,u_3....u_n\}$(i.e. the set of URL's)

- I={I_1 U I_2}
- F = {F_1, F_2, F_3}

  1.F_1={S,E}(Frequent web access pattern discovery from the web sequences)

  S represents number of sequences,

  S={$s_1, s_2, ..., s_m$}.

  E={$e_1, e_2 \ . . . \ e_k$}

[New sequence coming to the list of sequences is considered as an event ,E represents number of events.]

2.F_2={T,D,X_j}(Related term discovery through ontology.)

-T is a set of domain terms extracted from web page title.

-D is a set of the Web-pages, each page d_j has a sequence of domain terms X_j .

T={1≤ i ≤ p}

D ={d_j : 1≤ j≤ q}

X_j ={t_1,t_2 . . . t_n, t_k } [1]

3. F_3={F_1 U F_2}{TermNavNet based on CPM(Conceptual Prediction Model)}

$$P_{s,x} = \left[\frac{\partial_{s,x}}{\sum_{y=1}^{n} \partial_{s,x}} \right]$$

. [1]

This is the first-order transition probability from the starting state S to next state.

$$P_{x,y} = \left[\frac{\partial_{x,y}}{\partial_x} \right]$$

[1]

This is the probability from state x to y.

$$P_{x,E} = \left[\frac{\partial_{x,E}}{\partial_x} \right]$$

[1]

This is the probability from state x to the end state E.

- O={Recommended Web pages}

VI. EXPERIMENTAL RESULTS

This section presents the performance of the algorithms. The result is obtained by using the dataset from website(http://kdd.ics.uci.edu/databases/msweb/msweb.html) The following results are of both the algorithms.

The graph shows the result of PLWAP-mine algorithm and PL4UP-mine algorithm in terms of utilized time using different threshold values and different updated dataset sizes.

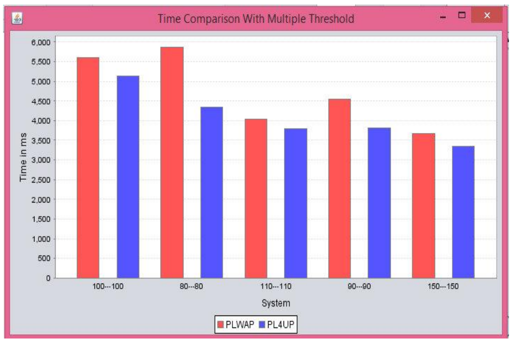

Fig 2. Execution Time in Seconds used by PLWAP and PL4UP for different threshold values.

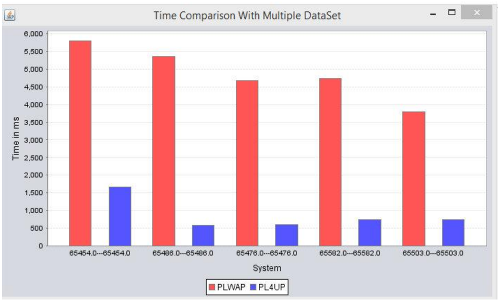

Fig 3. Execution Time in Seconds used by PLWAP and PL4UP for different dataset sizes.

The graphs shows the working of both the algorithm in terms of execution time. PL4UP uses less time even if the updated dataset sizes varies, Because it only constructs the tree for updated sequences and avoid rescaning of the entire database. The experiment is done by taking five different sizes of the datasets and minimum support threshold values.The experiments show that PL4UP takes less time to execute than PLWAP-Mine. Hence,for a proper recommendation considering the updation in database PL4UP can be used as a web usage mining technique for generating frequent patterns.

VII. CONCLUSION

Web Page Recommendation is to recommend the pages that web users are interested in. This can be done using web usage mining techniques, such as PLWAP-Mine and PL4UP-Mine.PLWAP-Mine has a limitation that if the database gets updated and there are too many small frequent item sets generated then PLWAP will not perform good in terms of execution time. To utilize the advantages and also to overcome the limitations of PLWAP.PL4UP-mine technique can be proved as more advantageous. If ontology of the website is used along with the technique then recommendation can be more Precise.

REFERENCES

[1] Thi Thanh Sang Nguyen,"Semantic-Enhanced Web-page Recommender System,"*University of technology Sydney*,December 2012.

[2] Thi Thanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge,"*IEEE transaction on knowledge and data engineering ,October* 2014.

[3] Yi Lu,C.I.Ezeife,"Position Coded Pre-order Linked WAP- Tree for WebLog Sequential Pattern Mining," *Springer* 2003.

[4] Yi Lu,C.I.Ezeife,"Mining Web Log Sequential Patterns with Position Coded Pre- Order Linked WAP-Tree"*Data Mining and Knowledge Discovery Springer 2005*.

[5] Yi Lu,C.I.Ezeife,"Fast increamental mining of web sequential patterns with PLWAP tree,"*Data Mining and Knowledge Discovery Springer 2009*.

[6] Monwar Mostafa,C.I.Ezeife,"A PLWAP based algorithm for mining frequent sequential patterns,"*Springer*.

[7] Xiaohui Tao,Yuefeng Li,"A Personalized Ontology Model for Web Information Gathering,"*IEEE Transaction on Knowledge and Data Engineering Vol 23 April 2011*.

[8] G.Stumme,A.Hotho,B.Berend,"Semantic Web mining:State of the art and future irections," *Elsevier 2006*.

[9] Chhavi,Rana,"Trends in Web Mining for Personalization,"*University Institute of Engineering and Technology,India,March 2014*.

[10] Bettina Berend, Andreas Hotho, Gerd Stumme, "Semantic Web Mining and the Representation, Analysis, and Evolution of Web Space," *Institute of Information Systems, Humboldt University Berlin, D-10178 Berlin, Germany*, 2005.

[11] Nazneen Tarannum, S. H. Rizvi, Ranjit R. Keole, "A Preliminary Review of Web-Page Recommendation in Information Retrieval Using Domain Knowledge and Web Usage Mining," *Computer Science & Information Technology*,January 2015.

[12] Sangeetha B, Saranya A, Revathi K, "An Intelligent Web System by Integrating Domain and Web Usage Knowledge,"*Computer Science and Engineering India*, March 2015.

[13] Nizar R. Mabroukeh and C. I. Ezeife, "Semantic-rich Markov Models for Web Prefetching," *School of Computer Science University of Windsor Windsor*, 2009.

[14] NazneenTarannumS.H.Rizvi, Prof. R.R. Keole, "Web-Page Recommendation In Information Retrieval Using Domain Knowledge And Web Usage Mining," May 2015.

[15] Dirk Thorleuchter, Dirk Van den Poel, "Extraction of Ideas from Microsystems Technology,"2012.

[16] Namita Ganjewar,"Effective knowledge representation integrated with web usage mining for web page recommendation," Mar 2015.