ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Web Page Prediction System Based on Web Logs and Web Domain Using Cluster Technique

Vasim Mujawar

M.E, Department of Computer Engineering, DPCOE, Wagholi, Pune, India

ABSTRACT: In intelligent web system web recommendation plays important role. In web mining, for web recommendation system the knowledge discovery and representation of information is an important and crucial task. Here in this paper new method is introduce to efficiently provide better Web-page recommendation generations through semantic-enhancement by integrating the domain and Web usage knowledge of a website. By the help of knowledge discovery user profile is created to block suspicious user that are harmful for websites or server. This model uses semantic web network to represent relations between domain, Web-pages & websites. Other model, the conceptual model, is proposed to auto generate a semantic web network of the semantic Web usage knowledge, which is the integrated with domain knowledge and Web usage knowledge.

KEYWORDS: Web page recommendation, web usage mining, web mining, knowledge representation, domain ontology

I. INTRODUCTION

The popularity of web page recommendation is increasing widely day by day. As the web users use website, a sequence of accessed or browsed Web-pages during a current session (the period from starting, to existing the browser by the user) can be generated. This sequence is organized into a Web session S = d1d2 . . . dk, where di (i = [1 . . . k]) is the page ID of the ith visited Web-page by the user.

The objective of web recommender system is to effectively predict web page or pages that are visited from a given web-page of a website. The performance of these approaches depends on the sizes of present datasets. The bigger the training dataset size is, the higher the prediction accuracy is. In this, system approach is based on web access sequence learnt from web usage data. Therefore, the predicted pages are limited within the discovered Web access sequences, i.e., if a user is visiting a Web-page that is not in the discovered Web access sequence, then these approaches cannot offer any recommendations to this user. This problem is called as new-page problem. The semantic-enhanced approaches are effective to overcome this problem.

Domain knowledge provides tremendous advantages in Web-page recommender available systems. To represent the semantics of Web-pages of website domain ontology is used.

II. MOTIVATION

A web-page prediction plays an important role in intelligent Web systems. To effectively overcome the new-page problem, useful knowledge discovery from web usage data and satisfactory knowledge representation for effective web-page recommendations are crucial and challenging. This work proposes a method to efficiently provide better Web-page recommendation through semantic-enhancement by integrating the domain and Web usage knowledge of a website

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

III. LITERATURE SURVEY

Web-Page Recommendation Based on Web Usage and Domain Knowledge [1]:
Web-page recommendation plays an important role in intelligent Web systems. Useful knowledge discovery from Web usage data and satisfactory knowledge representation for effective Web-page recommendations are crucial and challenging. This paper proposes a novel method to efficiently provide better Web-page recommendation through semantic enhancement by integrating the domain and Web usage knowledge of a website

A New Clustering and Pre-processing for Web Log Mining [2]:
Ontology based learning and domain knowledge extraction is used to perform better enhancement in web page recommendation system. These two aspects achieve by Dempster-Shafer theory

Aggregation of Similarity Measures in Schema Matching based on Generalized Mean [3]:
In applying sequence learning models to Web-page recommendation, association rules and probabilistic models have been commonly used. Some models, such as Sequential modelling, have shown their significant effectiveness in recommendation generation

Web usage mining, in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data [4]:
In this system Web access sequence is efficiently represented. Concept new page problem is arrived.

Data mining for web personalization [5]:
In this system traditional approach is used, that further uses sequence learning model

Integrating semantic knowledge with web usage mining for personalization, in Web Mining: Applications and Techniques [7]: The use of domain knowledge can provide advantage. Domain ontology is used**.**

IV. RELATED WORK

In this, two system are introduces earlier first is Traditional and second is semantic approach, traditional approaches use traditional rule for website recommendation for significant effectiveness for the knowledge of WebPages. But in semantic approaches it gives effective solutions over traditional approaches, in traditional approach markow-model and tree based structure are basically introduced. In Semantic approaches the integration of WebPages are introduces with some new rules like domain ontology. Domain ontology meaning means work integration with specific domain under web network. In this system, an ontology is built with the concepts data extracted from the documents, so that the documents must be clustered based on the similarity measure of the ontology concepts over web network. In order to produce semantically enhanced navigational patterns for web logs in cluster. Subsequently, the system can make recommendations, depending on the system input semantically matched with the produced navigational patterns over the clustering semantic ontology domain. So, here we continue with semantic approach due to different integration methods in the ontology domain

V. PROPOSED WORK

This work is presenting a new method to provide better web-page recommendations through semantic enhancement by three new knowledge representation models using cluster technique. Two new models have been proposed for representation of domain knowledge of a website. One is an ontology based model. A conceptual predicated model is also proposed to integrate the Web usage and domain knowledge to form a weighted semantic network of frequently viewed terms. Here we uses cluster technique for web log mining. Its having some clustering algorithm and it differ from traditional cluster technique. The Greedy Cluster algorithm is one of the best solution for web log mining.

Proposed algorithm (Greedy Clustering) The greedy clustering technique\method is widely used in many algorithms for web log as an efficient and effective way to approach a goal. In this technique representatives of the clusters are done repeatedly, the proposed algorithm is like…

Input: K- number of clusters; S- a simple set of users,

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Output: M- set of cluster representatives

begin

M = {*} //select random users m1 into the common profile set

M = {m1} for each user profile x ——————————————S – M, calculate the distance

between x and m1

Dist(x) = -ln (sim (x, m1))

For i =2 to K

begin //choose representative m1 to be far from previous representatives

Let m1€ S-M, such that dist (m1) = max (dist(x) | x€ S- M)

M = M*{mt} // Update the similarity of each point to the closest representatives

for each x + S-M
dist (x) = min (dist(x), -ln (sim(x, m1)))

end

return M // M will contain a set of distinct cluster representatives
end. So the greedy clustering is also easy to define log definition and its information

A. Collection of Accessed Web-pages:
This process firstly pre-processes Web logs to extract the URLs of Web-pages that have been visited by users at the given website, and then the URLs are crawled to fetch the metadata of Web-pages, i.e. the titles of Web-pages based on the TITLE tag on the HTML documents of Web-pages.

B. Extraction of Domain Terms:
This process extracts the domain terms from the titles of Web-pages retrieved in the first process (1). A term extraction algorithm is designed to extract terms from the Web-page titles. With this algorithm, tokens are firstly extracted, and then domain terms are generated based on these tokens. The results of this process are domain term sequences, each of which is a list of terms in the order as they appear in the titles.

C. Construction of a Semantic Network of Web-pages:
Based on the term sequences obtained from Process (2), a semantic knowledge representation model is built according to a collocation map (Park, Han & Choi 1995) and the Markov models (Borges & Levene 2005), in which occurrence weights of terms and associations between terms are taken into account to assess the frequencies of terms and collocations in the domain. The schema of this model is designed to represent the domain terms, Web-pages, and the relationships between them which can be populated to form a semantic network of Web-pages, referred to as TermNetWP. This network is the domain knowledge base of this website.

D. Implement an automatic construction of TermNetWP
The TermNetWP is implemented in OWL to enable the domain term network to be reused and shared by other parts of a Web-page recommender system. The algorithm to automatically construct a TermNetWP is as shown below:

Algorithm to Automatically construct a TermNavNet WP
Input: TSC (Term sequence collection)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Output: G (TermNetWP)

Process:

Let TSC = {PageID, X= t1t2… tm, URL}

Initialize G

Let R= root or the start node of G

Let E= the end node of G

For eachPageID and each sequence Xin TSC {

Initialize a WPaect identified as PageID

geobjFor eachterm ti □X{

If node tiis not found in G, then

- Initialize an Instanceobject Ias a node of G

- Set I.Name= ti

Else

- Set I= the Instanceobject named tiin G

Increase I.iOccurby 1

If (i==0) then

- Initialize an OutLink R-ti if not found

- Increase R-ti.iWeightby 1

- Set R-ti.fromInstance= R

- Set R-ti.toInstance= I

If (i>0& i<m) then

- Get preI= the Instanceobject with name ti-1

- Initialize an OutLink ti-1-ti if not found

- Increase ti-1-ti.iWeight by 1

- Set ti-1-ti.toInstance = I

- Set ti-1-ti.fromInstance = preI

If (i==m) then

- Initialize an OutLink ti-E if not found

- Increase ti-E.iWeight by 1

- Set ti-E.toInstance = E

- Set ti-E.fromInstance = I

-Set I.hasWPage = PageID

Add term ti into PageID.Keywords

}

}

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

The input data is a term sequence collection (TSC), in which each record consists of:

1) The PageID of a Web-page d □ D;

2) A sequence of terms X = t1t2 . . . tm □ TS,m > 0, extracted from the title of the Web-page; and

3) The URL of the Web-page.

TermNetWP can be used effectively not only to model the term sequences in connection with Web-pages, but also to present the co-occurrence relations of terms in the term sequences based on the following features: (i) it allows a term node to have multiple in-links and/or out-links so we can easily describe the relationships among terms/nodes in the semantic network, i.e. one node might have previous or next nodes; and (ii) it includes the Web-pages whose titles contain the linked terms so that the meaning of Web-pages can be found through these terms by software agents/systems. More importantly, TermNetWP enables reasoning of relationships between terms and Web-pages within a specific domain.

Mathematical Model

Let TSC= {PageID, X= t1,t2……tm,URL}

Let R=root or start node of G,

Let E= the end node of G,

Let perI= The instance object with name ti-1,

Equation:

$I = \prod_{ti \in x}^{ti \neq G} ti$ …….. equation 1;

$I + 1 = \prod^{ti=G} ti$……..equation 2;

If equation 1 is true then,

If i==0

Then

R=R- ti

I= R - ti

ti= ti+1

If equation 1 is true then,

If i> 0 and i< m

preI = ti-1-ti , ti=ti+1

I= ti-1-ti

If equation 1 is true then,

If i==m,

E=ti-1-E , ti= ti+1

I= ti-E

PageIDi= Ii

PageIDi[t] = ti, $ti \in x$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

$\prod_{x \epsilon TSC}$ Page ID = $\prod_{i=0}^{i=m}$ PageIDi

}

IV. RESULTS OF PRACTICAL WORK

The work done results are as shown in figures given below.

Figure 1 shows the starting index page it also serves as home page. It gives nothing as an output when that user logs on first time. So it has nothing related to his domain

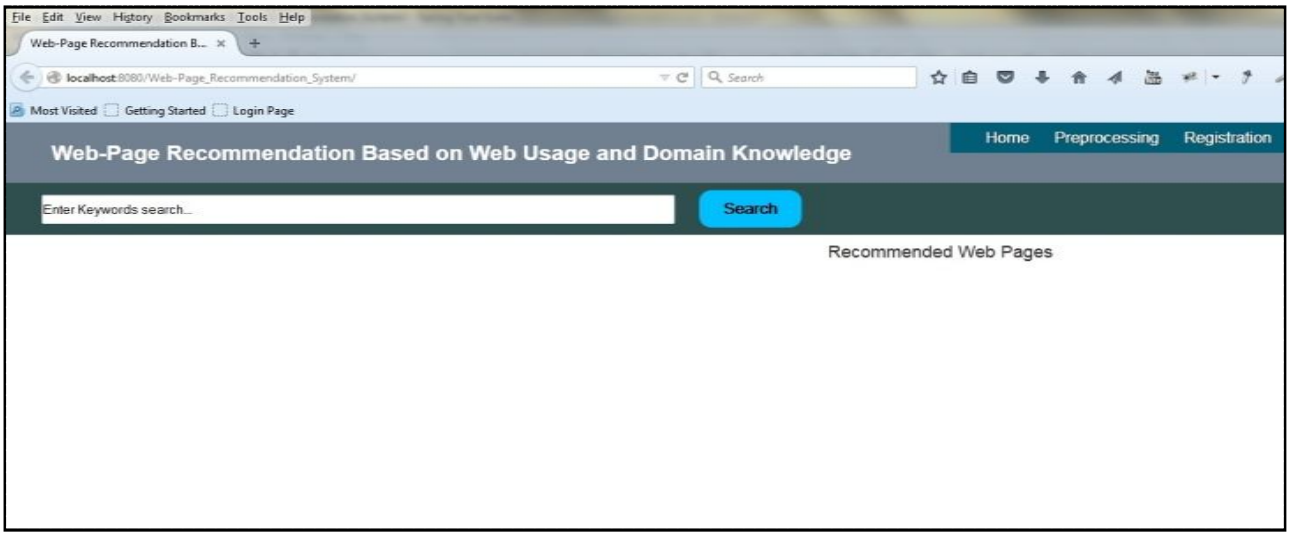

Figure: 1 Index page

Figure 2 shows the result of search query and recommendation page for it. Even though in database there is nothing related to its domain, still it will not give "New page problem" as stated in literature.

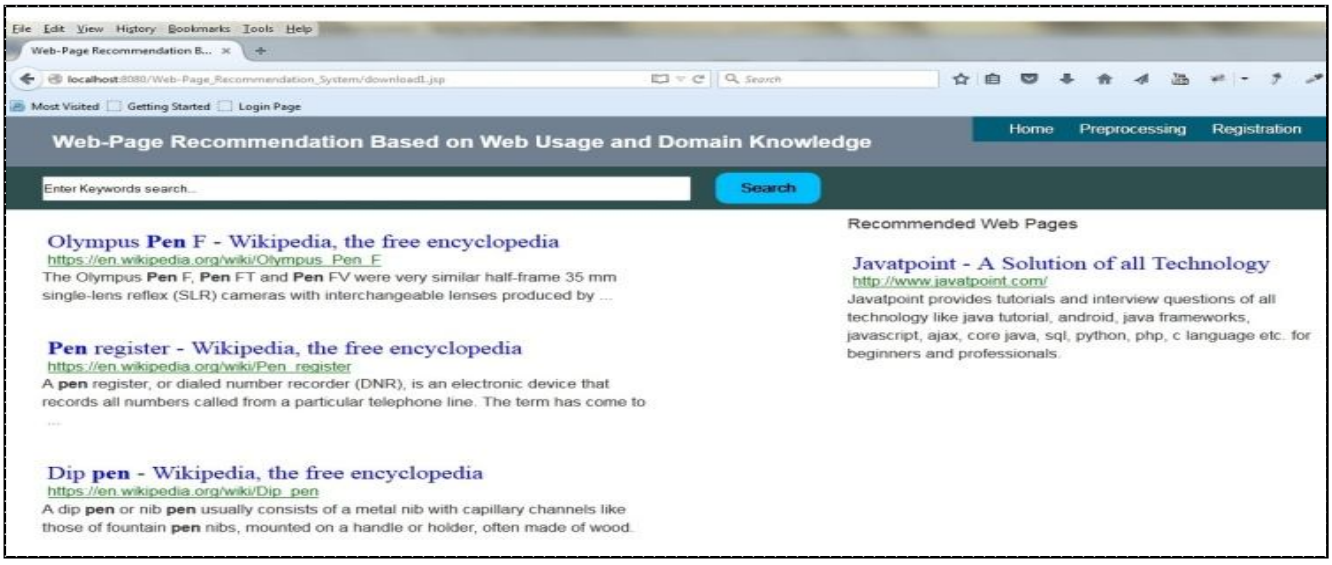

Figure: 2 Search result page

Figure 3 shows the page for browsing the log file.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

Figure: 3 preprocessing page without log file

Figure 4 shows the page for browsing the log file. After that based on entries in log file, system creates user domain.

Figure: 4 preprocessing page with log file

VI. CONCLUSION AND FUTURE WORK

Ontology based learning and domain knowledge extraction is used to perform better enhancement in web page recommendation system. A number of Web-page recommendation strategies have been proposed to predict next Web-page requests of users through querying the knowledge bases. The experimental results are promising and are indicative of the usefulness of the proposed models.

REFERENCES

[1] 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 engg.vol 26.no 10 October
[2] B.Uma Maheswari, Dr. P.Sumathi, A New Clustering and Preprocessing for Web Log Mining, in 2014 World Congress on Computing and Communication Technologies References Papers

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 8, August 2015

[3] Faten A. Elshwimy, Alsayed Algergawy, Amany Sarhan, Elsayed A. Sallam, Aggregation of Similarity Measures in Schema Matching based on Generalized Mean, in ICDE Workshops 2014 , pp. 74-79, 2014.
[4] B. Liu, B. Mobasher, and O. Nasraoui, Web usage mining, in Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, B. Liu, Ed. Berlin, Germany: Springer-Verlag, 2011, pp. 527603.
[5] B. Mobasher, Data mining for web personalization, in The Adaptive Web, vol. 4321, P. Brusilovsky, A. Kobsa, and W. Nejd l, Eds. Berlin, Germany: Springer-Verlag, 2007, pp. 90135
[6] G. Stumme, A. Hotho, and B. Berendt, Usage mining for and on the Seman-tic Web, in Data Mining: Next Generation Challenges and Future Directions. Menlo Park, CA, USA: AAAI/MIT Press, 2004, pp. 461480.
[7] H. Dai and B. Mobasher, Integrating semantic knowledge with web usage mining for personalization, in Web Mining: Applications and Techniques, A.Scime, Ed. Hershey, PA, USA: IGI Global, 2005, pp. 205232.
[8] S. A. Rios and J. D. Velasquez, Semantic Web usage mining by a concept-based approach for online web site enhancements, in Proc. WI-IAT08, Sydney, NSW, in Australia, pp. 234241.
[9] S. Salin and P. Senkul, Using semantic information for web usage mining based recommendation, in Proc. 24th ISCIS, Guzelyurt, Turkey, 2009, pp. 236241.
[10]A. Bose, K. Beemanapalli, J. Srivastava, and S. Sahar, Incorporating concept hierarchies into usage mining based recommendations, in Proc. 8th WebKDD, Philadelphia, PA, USA, 2006, pp. 110126.
[11] N. R. Mabroukeh and C. I. Ezeife, Semantic-rich Markov models for Web prefetching, in Proc. ICDMW, Miami, FL, USA, 2009, pp. 465470.
[12] M. OMahony, N. Hurley, N. Kushmerick, and G. Silvestre, Collaborative rec-ommendation: A robustness analysis, ACM Trans. Internet Technol., vol. 4, no. 4, pp. 344377, Nov. 2004.
[13] G. Stumme, A. Hotho, and B. Berendt, Semantic Web mining: State of the art and future directions, J. Web Semant.
[14] B. Zhou, S. C. Hui, and A. C. M. Fong, CS-Mine: Ancient WAP-tree mining for Web access patterns, in Proc. Advanced Web Technologies and Applications

BIOGRAPHY

Vasim Mujawar is a Research Scholar (M.E.C.E. Student) in the Computer Engineering Department, Dhole Patil College of Engineering, Wagholi (Pune). He received Bachelor of Engineering degree in 2011 from JJMCOE, Shivaji University Kolhapur, India. His research interests are Web User clustering, Web usage mining, etc.