ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Webpage Recommendation using Keyword Association Link Network and Prediction using CSB Model

Monika B. Bisane[1], R.V. Mante[2]

P.G. Student, Department of CSE, Government College of Engineering, Amravati, Maharashtra, India[1]

Assistant Professor, Department of CSE, Government College of Engineering, Amravati, Maharashtra, India[2]

ABSTRACT: Webpage recommendation plays an important role for understanding about the evolution of the events. But as the huge amount of webpages are generated over the internet day by day, it is difficult for user to understand or to get exact information about the event. This huge amount of webpages may contain uncertain data which may trouble the user while using internet. So In this paper, a method is proposed in which webpage recommendation can be done by constructing Semantic Pyramid. This semantic pyramid consisted of different levels i.e. Theme layer, Backbone layer and Tidbit layer. Here, the main idea is to consider web event as a system composed of different keywords and then by using Keyword Association Link network (KALN) web event representation is done. After that based on that KALN and by using TF for classification the semantic Pyramid is constructed. Thus the proper web link provided to the user. Once the recommendation is done the next step here is the prediction. This prediction of next link can be done when user makes request for various pages/links in his session by using CSB algorithm which works on frequent pattern matching from history of records.

KEYWORDS: Keyword Association Link network, Semantic Pyramid, Webpage recommendation, Web event

I. INTRODUCTION

Nowadays, people are using the internet in a huge amount. And there is also the huge growth in the webpages over the internet. To get the information regarding any event like earthquake, floods, and many more occur in the society or any topic which will attract the broad attention on the web people are using internet. But as there could be huge number of web links (webpages) are present related to that Web event, it is difficult for the user to understand the event or any topic. As time passes, hundreds and thousands of web pages, blogs, and post are generated. This huge amount of information makes it impossible for user to understand the event by reading that number of pages and also its takes too much time. That huge number of webpages contain certain and uncertain information. So it is challenging task for the webmasters to remove the uncertain data and to provide the appropriate web page to the user. So this paper proposed a method for webpage Recommendation so that user can get exact links i.e. the link that gives exact information regarding any web event or any topic.

Webpage Recommendation plays an important role in web system. Huge amount of work has been done on webpage recommendation. It is the technique in which number of link are provided related to the particular topics. And that links contains huge amount of information so that user can understand the evolution of the web event or topics easily. As on the World Wide Web huge amount of data is generated day by day. Due to this if user enter any query then it will give huge number of pages relate to that query so it is difficult and time-consuming for user to read all this pages. Also some of this pages may contain uncertain data i.e. the data which is not as per the query. This may create little bit problem for the user. So here, this disadvantage has been overcome.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Rather than Webpage Recommendation, Prediction is also one of the most important aspect on web system. Prediction can be define as, as user searches number of queries in his session. So based on his history record which link he has to provide first that work can be called as prediction. With the help of that prediction link are provide to the user as per his interest. So in this project, the work on prediction has been also done.
Paper is organized as follows. Section II describes related work on Webpage Recommendations. The flow diagram represents the step for webpage Recommendation and prediction system in Section III. Section IV presents experimental results. Finally, Section V presents conclusion.

II. RELATED WORK

S Thorsten Brants et al [1], have proposed a method in which, new event detection is done. This system is the extension of the TF-IDF model. Generation of source specific models, similarity score normalization based on document-specific averages, similarity score normalization based on source-pair specific averages, term reweighting based on inverse event frequencies, and segmentation of the documents are the extension included in this model. Thus the new event detection is used.
F. Can, S. Kocberber et al [2], proposed a method where TDT has been done. As huge amount of information on internet are growing day by day. Here user prefer to track the topics that they are interested in. For that two steps are proposed NED(New Event Detection) and TT(Topic Tracking).This method finds the first stories of previously unseen new events and all related stories on a certain topic. The aim of that TDT is to organize the temporal ordered stories of a news stream as per the events.
G. Kumaran and J. Allan [3],, proposes a method in which for New event Detection they used text classification techniques and also they used named entities in a new way. The new multistage system is presented here for performing New Event Detection. This method is the extension of the baseline vector space system. And this extension is being more directive foe NED.
In the above three methods the main focus is the detection of topic. But the deep analysis of the topic is not given here.
A. Sun and M. Hu [4], proposes a method in which there is the study of the query guided event detection. Here. User's queries ,news article and blogs posts are group together. i.e They integrates queries, news article and blog posts through the idea of query profile.
C. C. Yang et al [5]. proposes a method for Discovering event evolution graphs from news corpora. The event timestamp, event content similarity, temporal proximity, and document distributional proximity are used to model the event evolution relationships between events in an incident. For the exact browsing and extraction of the proper information, the event evolution graph is constructed. In these model there is only focus on the keywords, the association rules between the keywords are neglected.
X. Luo, Z. Xu [6],, proposes a method where along with the keywords association rules of keywords are use for building the ALN of webpages. Here, for building association link network for semantic links on the web the discovery algorithm for associated resources is proposed. Then connection rich ALN is developed which is the effective tool for building the Association Link Network for Semantic Link on Web Resources.
Chao Wang *, Jie Lu, Guangquan Zhang in 2007 [7] proposed a method which points out the existence of key information in web pages and the significance of mining and using key information. It has also proposed KIM, a two-step method that can automatically mine key information from web pages. The method first extracts a list of candidate key information. It then uses an entropy evaluation to filter most of the noisy information in the list so that the key information can be discovered easily.
ThiThanh Sang Nguyen, Hai Yan Lu, and Jie Lu in 2014 [8] proposed the two new models for representation of domain knowledge of a website. The first model is ontology based which represents domain knowledge of the website. The second model is semantic network to represent domain terms, web-pages and relation among them. This study introduced the conceptual prediction model which integrates the domain knowledge and web usage knowledge, to support effective and better web-page recommendation. This paper provides better web-page recommendation by

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

combining domain and web-usage knowledge. But the problem is that Ontology is difficult to construct .So it would be better to incorporate more analysis of the contents of Web events.
There are different work proposed for the webpage recommendation or event detection .But In our paper it can be done by Using KALN in which both the keywords and its association rule are considered and after that by deep understanding of the web event the semantic pyramid is constructed and based on that SP proper webpage is recommended to the user.

III. PROPOSED SYSTEM

General concept is given as follows. In figure 1. Proposed System Architecture overall concept of paper is given. The main stages of the project are as follows:
1] Webpage collection
2] Keyword Extraction
3] Association
4] Keyword Weight
5] Pyramid Generation
6] Prediction
Each of this modules is explained as follows:

A. Webpage Collection

Firstly, the huge no of links and information in that links are collected from different sources.

B. Keyword extraction

In the keyword extraction method the main aim is to get keywords which plays an important role in describing any web event. When user enter any query to search then related to that query different links are retrieved from the collected webpages also the content of the webpages are obtained. This content will be stored in file. After that the next step is to clear the unwanted data which can be done by using Pre-processing. Pre-processing technique like remove the stop word and then store the needed word in file. Thus the important keyword of the web event are obtained.

C. Association

For the association, there is one word dictionary (i.e. dataset related word). This dictionary is compared with the file that we get by the keyword extraction. Then the match word are saved in another file.

D. Keyword Weight

After getting the words.TF of each word is calculated and the link, word and count are stored into the database.
Thus the KALN are constructed. KALN is the system which consists of the keywords and association relations between that keywords. This keyword set and the association relation set of keyword are extracted from the collected webpages.

E. Pyramid Generation

As KALN is generated after that the next step is to construct semantic pyramid by using that KALN. Here that pyramid is generated by using TF .The semantic pyramid consists of different level of semantic. That level are Theme layer, Tidbit layer, Backbone layer. The top most level is the Theme level. Middle level is the Backbone layer and the bottom layer is the Tidbit level.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

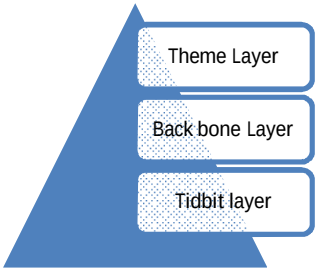

Fig 1. Semantic Pyramid

- **Theme layer**:-

The keywords came under this layer contains the keywords which have huge tf value. This layer is the core of the KALN. It explains what this KALN is talking about. It consists of more important keywords.

- **Backbone Layer**:

This is the second layer in semantic pyramid. And here keywords having tf in middle range are come under this category. It tells us about the main semantic of the web event.

- **Tidbit Layer**:

This tidbit layer consists of the keywords which contains the keywords having low tf value.
Thus Semantic Pyramid (SP) is constructed which consists of Theme .Backbone and Tidbit layer. From this layer each layer having specific semantic level of KALN.

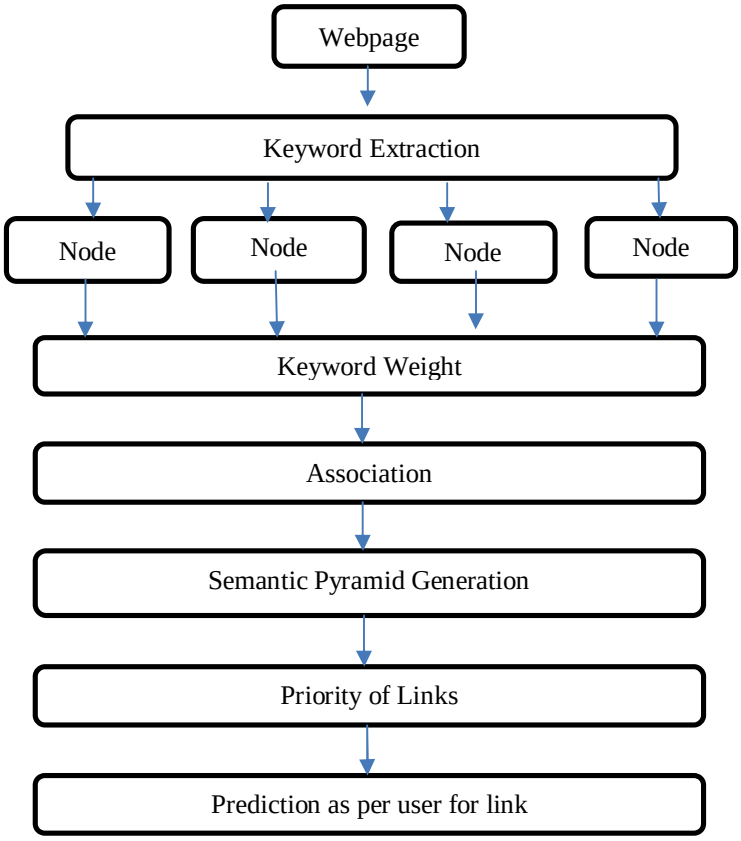

Fig. 2. Workflow diagram for web page Recommendation

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

F. Priority of links

The priority of the link is done on the basis of the TF which was calculated above.

G. Prediction

Here the prediction can be done by using CSB algorithm. The unprocessed file have been read and data is cleaned thus pre-processed file of respective dataset is obtained. Then the next step is find the unique user and that unique user data respective data set is written to one file. After that the next step is to identify all user sessions that data in stored in one file which contain session with page id. After that all the session are group together as per user. Then based on that web access sequence is generated. Then using CSB the prediction of the link has been done.

Thus the whole proposed system works. And Webpage recommendation can be done using KALN and Semantic Pyramid. After that prediction is also done.

IV. EXPERIMENTAL RESULTS

Figure shows the result of webpage Recommendation and prediction system. Figure 3 is the Home page for the webpage recommendation Here user have to login his account. New user have to create there account first. In Figure 4 there is a page where user can enter any query according to his interest. After clicking on LINKRETRIVE button and PARAGRAPHRETRIVE buttondifferent links are retrieved and paragraph are retrieved related to that links. Then PREPROCESSALLDATA button is used for pre-processing the data due to which important keywords are obtained. TFIDF button is used for keyword weighting. And using Keyword Association Link Network and Tf the appropriate links are obtained. Thusappropriate links are obtained as shown in fig 5. Then fig 6 shows the prediction of links as per user's historic session. Here

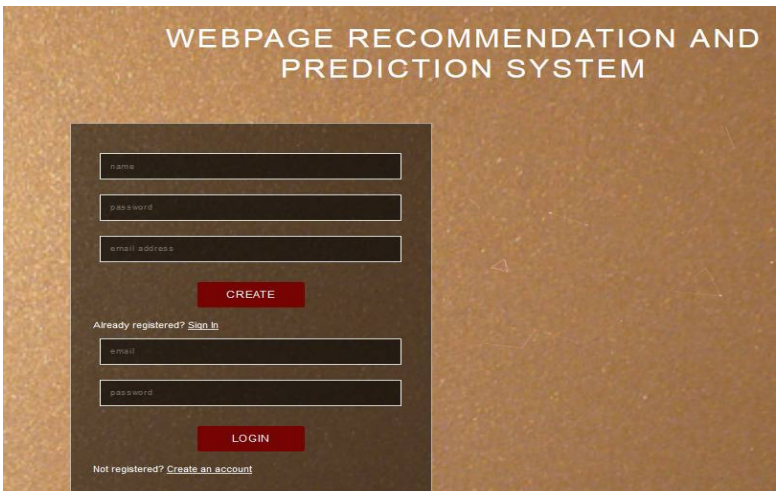

Fig 3: Webpage Recommendation Home Page

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

Fig 4: Run Application

Webpage Recommendations. Here user have to login his account. New user have to create there account first. In Figure 4 there is a page where user can enter any query according to his interest. Then different links and information in that links are retrieved. All that information is pre-processed. And using Keyword Association Link Network and Tf the appropriate links are obtained as shown in fig 5.

Fig 5: Appropriate Links Recommendations with the keywords as per its importance

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

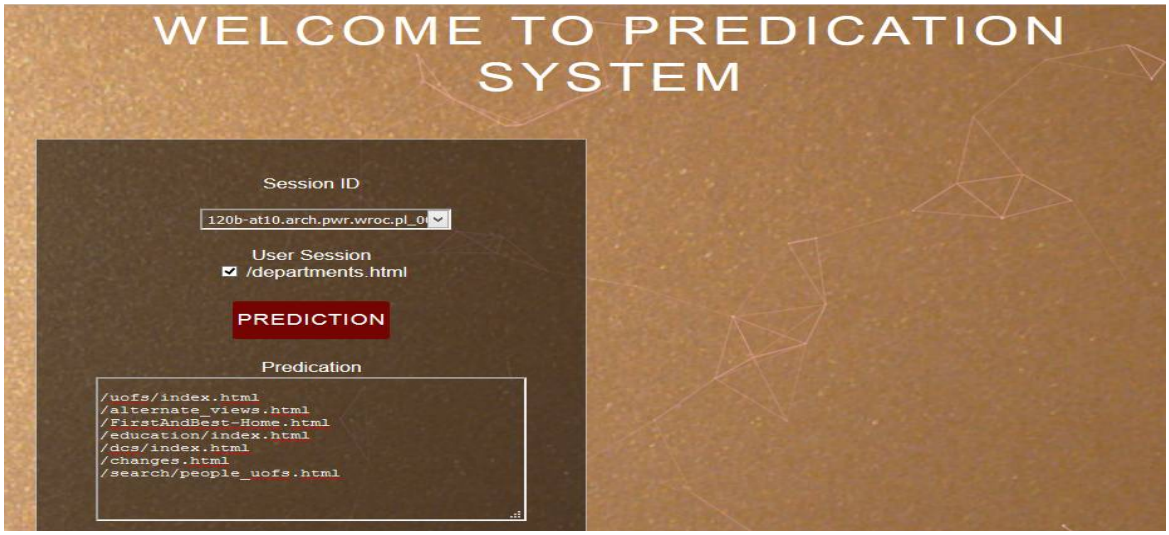

Fig 6: Prediction as per user for link

In fig 6 Session ID block contains different user's session ID. When one of the session id is selected it will provides different his different session. Then by clicking on one of that session it is providing the different links which may user browses in past. Thus prediction of different links are done.

Mathematical Explanation for TF.

E.g: Term frequency

Suppose we have a set of English text documents and wish to determine which document is most relevant to the query "the brown cow". A simple way to start out is by eliminating documents that do not contain all three words "the", "brown", and "cow", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its *term frequency*.

In the case of the **term frequency** tf(*t*,*d*), the simplest choice is to use the *raw count* of a term in a document, i.e. the number of times that term *t* occurs in document *d*. If we denote the raw count by $f_{t,d}$, then the simplest tf scheme is tf(*t*,*d*) = $f_{t,d}$. Other possibilities include

- Term frequency adjusted for document length : $f_{t,d}$ / (number of words in d)
- eg.The food is good and service also good but parking is bad.

Step 1: Separate with (space)
The
Food
is
Good
and
service

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

also
good
but
parking
is
Bad

Step2) Remove the stop word
Food
good
service
good
parking
bad

Step3) Count the repeated count of each word in document
Food- 1
Good- 2
Service -1
Parking- 1

V. CONCLUSION

This paper aims to provide appropriate webpage to the user by using Keyword Association Link Network (KLAN) and Semantic Pyramid. Semantic pyramid consists of different layer are used for the recommendation of the links. And then based on the history of the user the prediction of the can be done. Thus different level of information is provided to people with different requirements.

REFERENCES

[1] T. Brants, F. Chen, and A. Farahat, "A system for new event detection," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, Toronto, ON, Canada, 2003, pp. 330–337.

[2] F. Can, S. Kocberber, O. Baglioglu, S. Kardas, H. C. Ocalan, and E. Uyar, "New event detection and topic tracking in Turkish," Journal of the American Society for Information Science and Technology, vol. 61, Apr. 2010, pp. 802–819, doi:10.1002/asi.v61:4.Eftychios A. Pnevmatikakis, Petros Maragos "An Inpainting System For Automatic Image Structure-Texture Restoration With Text Removal", IEEE trans. 978-1-4244-1764, 2008

[3] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," Proceedings of international ACM SIGIR conference on Research and development in information retrieval(SIGIR 04), ACM Press, Jul. 2004, pp.297-304, doi:10.1145/1008992.1009044

[4] A. Sun and M. Hu, "Query-Guided Event Detection From News andBlog Streams," Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on , vol. 41, no. 5, Sept. 2011 ,pp.834-839, doi:10.1109/TSMCA.2011.2157129.Xiaoqing Liu andJagathSamarabandu, "Multiscale Edge-Based Text Extraction From Complex Images",IEEE Trans., 1424403677, 2006

[5] C. C. Yang , X. Shi, and C. P. Wei, "Discovering event evolution graphs from news corpora," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 39, no. 4 Jul. 2009, pp.850-863, doi:10.1109/TSMCA.2009.2015885

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 5, May 2017

[6] X. Luo, Z. Xu, J. Yu, and X. Chen, "Building Association Link Network for Semantic Link on Web Resources," IEEE Transactions on Automation Science Engineering,vol.8,no.3,July2011,pp.482-494, 10.1109/TASE.2010.2094608.

[7] C. Wang, J. Lu, and G. Zhang, "Mining key information of Web pages: A method and its application," *Expert Syst. Appl.*, vol. 33, no. 2,pp. 425–433, Aug. 2007

[8] ThiThanh Sang Nguyen, Hai Yan Lu, and Jie Lu, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE Trans. On knowledge and data engineering, vol.26, no. 10, October 2014.

[9] J.Xuan*et al.*,"Building hierarchical keyword level association link networks for Web events semantic analysis," in *Proc. IEEE 9th Int.Conf. Depend. Auton. Secure Comput.*, Sydney, NSW, Australia, 2011, pp. 987–994

[10] JunyuXuan, *et al* "Uncertainty Analysis for the Keyword System of Web Events"IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, VOL. 46, NO. 6, JUNE 2016

[11] B.Zhou,S. C.Hui,andA.C. M.Fong,"Efficientsequentialaccesspatternminingforwebrecommendations",Int.J.Knowl.-BasedIntell. Eng.Syst.,vol. 10, no.2, pp. 155–168, Mar. 2006.