New Event Detection for Web Recommendation using Web Mining

Nilam Pujari
Department of Computer Engineering
JSPM's RSCOE College of Engineering, Pune
Savitribai Phule, Pune University
Pune, Maharashtra, India, 411033

Vaishali Barkade
Department of Computer Engineering
JSPM's RSCOE College of Engineering, Pune
Savitribai Phule, Pune University
Pune, Maharashtra, India, 411033

***Abstract*:** In this paper, we propose techniques for the uncertainty analysis on the web events and it is applicable to webpage recommendations. This approach observes, summarize and track events from a collection of new web pages. Given a set of web services, we calculate the response-time for the web services. Here, used MLP (Multilayer Perceptron) model for improve performance of web service. The collected literature review of webpage recommendations it classified into two classes such as non-content based methods and content-based strategies. This proposes techniques identify the various underlying levels of linguistics uncertainty in terms of various web events, and it is used to the webpage recommendations. Our aim is to contemplate an internet event as a system composed of various keywords, and therefore the uncertainty of this keyword system is expounded to the uncertainty of the actual web event, we tend to establish the different levels of linguistics uncertainty, and construct a linguistics pyramid to precise the uncertainty hierarchy of an internet event.

Keywords: Social Event, Uncertainty Analysis, Web Events, Web Mining, MLP and Web Page Recommendation.

I. INTRODUCTION

In recent years, the popularity of the internet has grown to almost every degree young or old people, who use it for various purposes. People use the internet information in the field of interest, study related to work or study, get a good deal acquire, raise awareness about commodities and travel, its surroundings and the world latest news and so on each day, a large number of informative website, the web pages and web documents are already added to a huge collection. Any popular search engine returns thousands of links related to search queries. It has become difficult users will easily get the most relevant information from this plethora of relevant information available. Users spend a lot of time getting pages viewed appropriate information. If the intention and interest of the user to view the webpage once identified, it makes it easier to make the area of that information available at a higher priority. Personalized web browsing not only for the specific user's intent, but also for the future but there are also other advantages. Intuitive web pages for users to view thus, you can significantly reduce the acquisition time and load on the network. Another thing the benefit of knowing the user's intentions is for the purpose of E-Commerce. Only to users who can view related and targeted marketing ads increase in the number of customers.

Making the discovery and prediction that will enable the web mining process to be tailored to the user's interest personalize the web analyze user's web browsing patterns to web mining deals and the structure of the website or the content of the webpage. A lot of research has been done personalized to realize the director, for many to web other users browsing patterns. But as the number of web pages increases for every 1 second, personalization based only on web usage mining has the following disadvantages it does not take into account the context of the web page. So the web semantics are the context of a web page is an equally important concept to consider. But some researchers are exploring this area; there is still room for improvement.

The process of providing information related to a user's current page is called web personalization. This information is typically displayed on the current page in the form of a web page link. The idea behind web personalization is that the web page that the user is currently browsing is interested in the topic, and that the user is interested in more similar information. For example, in the case of E-commerce, the relevant information is made by other similar products that the user is viewing, or purchased by other users who purchased or viewed this product, this example also works in research or targeted-oriented web browsing. The important information needed to suggest these similar web pages is not only the current page, but also other users who visited other pages before and after this current page. The recommendation system is used more than ever in a wide range of services such as financial investment, healthcare and E-Commerce. In E-Commerce, the Web Recommendation System (WRS) is relying on the user's history and behavior to recommend future item purchase and page views, these data are not available to the user.

In this paper study about the related work done, in section II, the proposed approach modules description, mathematical modeling, algorithm and experimental setup in section III .and at final we provide a conclusion in section IV.

II. LITERATURE REVIEW

This paper [1] introduces the first large-scale TDT test collection for Turkish, and investigates the necessity and TT problems in this language. We present our test collection construction approach inspired by TDT research initiatives. They show that a simple word truncation stemming scheme can compete with lemmatize based stemming scheme in Turkish TDT with several similar scales. Findings of this techniques show that, contrary to previous observations on information retrieval in Turkey, it affects the effectiveness in stopping the required word. In addition, there is a higher effect in the easy-to-understand method in combination of Shin two different similar measures.

In this paper [2] they propose web page recommendation techniques in this they combines Semantic-Enriched Domain with Web usage knowledge from the websites. Two new models representing domain knowledge are proposed. The first model uses ontology to represent domain knowledge. Auto generated meaning for the second model domain terms representing the network are relevant to the home page. In this paper, they propose model called conceptual prediction model, and generate a semantic network of semantic Web knowledge. A lot of effective queries have been developed in the orientation in detail and these are known bases. Based on these queries, a set of recommended strategies has been proposed for generating candidate Web pages. These results were compared with the Web Usage Mining (WUM) method.

In this paper [3] they deeply analyze the semantics of web events such as evolutional semantics of events, they propose a method Keyword Level Association Link Network (KALN) at the keyword level. First, the first kiln is built, and then the traditional data mining technology by a predetermined time. Then, the hierarchical KALN consisting of the theme layer network, the backbone layer network and the tidbit layer network is constructed on the basis of the original KALN by the information entropy for distinguishing different semantic levels of web events. With a hierarchical KALN semantic analysis, humans can easily get Full understanding of web events.

As a simultaneous user increase, scalability is one of the major challenges in the design of the DVE system. One solution is a multi-server architecture that is a scalability issue. On the other hand, the quality of the partitioning of the load of the server, on the other, is the partitioning process itself. However, all of these methods ignore the effect of network latency between servers on the accuracy of the load balancing solution. As shown in this paper [4], changes in server load due to network latency affect the performance of the load balancing algorithm. In this paper they formally analyze this problem and discuss two efficient delay adjustment schemes to deal with the problem.

In this paper [5], they propose an on-line topic model which analyzes the time evolution of a topic in a document collection. The challenge is of course evolution and multiple times. For example, some words may be used consistently over a hundred years, while other words will emerge and disappear over a period of several days. Therefore, the proposed model assumes that the current topic-specific distribution for a word is generated based on the multi-scale word distribution of the previous epoch.

Taking into account the long and short time scale dependence yields a more robust model. We derive an efficient on-line reasoning method based on the probability EM algorithm which updates the model successively using newly obtained data.

In this paper [6], they propose an approach to make Web page recommendations based on Markov logic network. This system makes full use of the log of web users, cluster characteristics of web pages, and contents characteristics of web pages to recommend related web pages to web users.

In paper [7] developed a system based on the basic idea of the system developed by Ghassan Beydoun into the field of enterprise knowledge sharing: to mine the useful information from users surfing trails to generate FCA knowledge base and enable enterprise knowledge sharing with social navigation mechanism. Authors import enterprise ontology knowledge base in order to make up the deficiency of semantic expansion in the system designed by Ghassan Beydoun.

In paper [8] developed a method based on meta search, combined with article summarization technique, to assist an ESL learner in deciding keywords, and to decrease the number of repeated search in the process, thereby increase the learning and reading efficiency of the learner. The special features of our system are (1) Use Meta search and article summarization technology to construct an article search mechanism. (2) Apply a technique originally used for evaluating social networks, to article summarization.

III. PROPOSED APPROACH

A. Proposed System

Below figure shows the system architecture of the proposed system. In proposed system we are providing the input as a several “Web Pages”. After processing on the given pages system generates the output which is nothing but the recommended pages for the searched keywords.

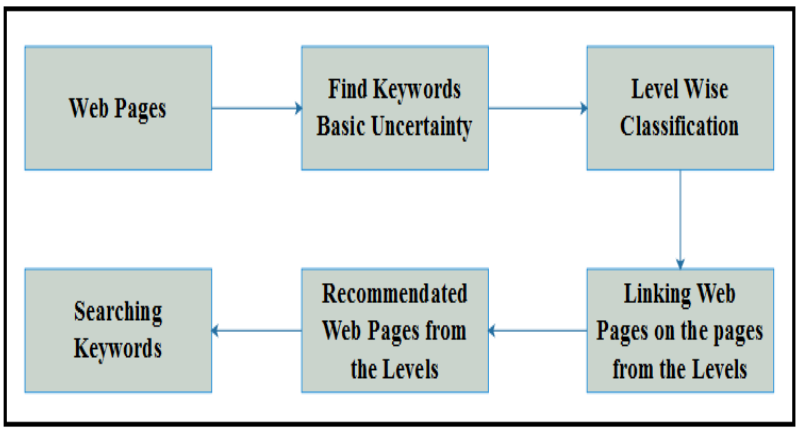

Figure 1.Proposed System Architecture

In this system architecture we have following modules:

I. Web Pages:
These frameworks that are identifying the various underlying levels of semantic uncertainty in terms of events and makes use of these recommendations for web pages. The purpose of this paper is to grasp web events as a system composed of different keywords, and to correlate the uncertainty of this keyword system with the uncertainty of web events. For example, the Libya war (2011) is a web event that contains 1000 of posts, web pages, and blogs.

II. Find Keywords basis Uncertainty:
In web events contains different keywords. The uncertainty is a measure of a web event keyword system and keywords are the main semantic atoms of a web event, it can also be called semantic uncertainty.

III. Level Wise Classification:
Here, firstly web events are represented as KALN (Keyword Association Link Network) for the preserving semantics of web events in given time. In second steps it identifies the hierarchical uncertainty of KALN and then here construct the SP (Semantic Pyramid). Here, SP shows that the hierarchical structure of uncertainty of web events on the web sites. And it web events are classify according to semantic hierarchical levels.

IV. Web Page Recommendation:
By grasping the semantic uncertainty of the event, it was possible to improve the satisfaction of the web recommendation. However, these different levels of semantic uncertainty are overlooked in traditional hit rate-based or clustering-based web page recommendation techniques. There is a lot of work in the recommendations of web pages, but the uncertainty of the meaning of web events is rarely taken into account. web page recommendation divided into two categories firstly non-content based method and second one is content based method.

B. Algorithm

Algorithm 1: Naive Bayes

Process:

Step 1: Convert the data set into a frequency table
Step 2: Create Likelihood table by finding the probabilities like Overcast probability = 0.29 and probability of playing is 0.64.
Step 3: Now, use Naive Bayesian equation to calculate the posterior probability for each class. The class with the highest posterior probability is the outcome of prediction.

$$P(x\backslash c) = \frac{P(x\backslash c)P(c)}{P(x)}$$

Algorithm 2: Apriori Algorithm

Step 1: Start
Step 2: Join step:C_k is generated by joining L_{k-1} with itself.
Step 3: prune Step: any (k-1) itemset that is not frequent cannot be a subset of a frequent k-item set
Step 4: Pseudo code: C_k : Candidate itemset of size k
Step 5: L1= {Frequent items};
Step 6: for (k=1; L_k! =∅; $k++$) do begin
Step 7: C_{k+1} =Candidate generated from L_k
Step 8: for each transaction tin database do increment the count of all candidates in C_{k+1}
Step 9: That are contained in t L_{k-1} = Candidate in C_{k+1} with min_support
Step 10: return$\cup_k L_k$;
Step 11: end

Algorithm 3: C 4.5 Algorithms:

Process:

Step 1: Check for the below base cases:

i. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class.
ii. None of the features provide any information gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class.
iii. Instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the expected value.

Step 2: For each attribute a, find the normalized information gain ratio from splitting on a.
Step 3: Let a best be the attribute with the highest normalized information gain.
Step 4: Create a decision node that splits on a best.
Step 5: Recur on the sublists obtained by splitting on a best, and add those nodes as children of node.

Algorithm 4: TF-IDF Algorithm:

Step 1: Start
Step 2: class mapper
Step 3: Method map ((docld, N), term, o)
Step 4: for each element Ɛ(term,o)
Step 5: write (term,(decld,o,N))
Step 6: class reducer
Step 7: method reduce (term,(docld,o,N))
Step 8: N=0
Step 9: for each element Ɛ(decld, o,N)
Step 10: n=n+1
Step 11: tf=o/N
Step 12: idf=log(|D|/(1+n))
Step 13: return (decld,(term, tf×idf)
Step 14: End

C. Mathematical Model

Let S be a system, such that,
S = {Input, Process, Output}
Input: I = Web Pages
Process:
P = {P1, P2, P3, P4, P5}
Where, P represent the total number of steps perform in system to get output.

1. P1 = Finding Keyword Basis Uncertainty
   P1 = {p11, p12}
   Where P1 is the set of finding keyword basis uncertainty
2. P2 = Level Wise Classification
   P2 = {F1, F2... Fn}
   Where, P2 represent level wise classification and F1, F2 ...Fn are number of classes.
3. P3 = Linking web pages on the basis of Keywords
   P3 = {P31, P32...P3n}
   Where, P3 represent the set of linking web pages and P31, P32...P3n are number of web pages.
4. P4= Recommended web pages from the levels
   P4 = {P41, P42...P4n}
   Where, P4 represent the set of recommended web pages and P41, P42...P4n are number of pages.
5. P5 = Searching Keywords
   P5 = {P51, P52 ... P5n}
   Where, P5 represent the set of techniques of searching keywords and P51, P52...P5n are number of keywords.

Output:
Web Page Recommendation T = {T1, T2 ... Tn}
Where, T represents the set of recommended web pages.

IV. RESULTS AND DISCUSSION

A. *Experimental Setup*

The system is built using Java framework on Windows platform. The Net bean IDE is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

B. *Dataset*

1. Dataset 1: In this dataset contains webpages of Japan earthquake in the 2011 from 09-03- 2011 to 20-04-2011 and it contains 6884 number of webpage. These webpages are collected from and www.Baidu.com and www.Google.com.
2. Dataset 2: It is also webpages of Japan earthquake in 2011 from 09-03-2011 to 20-04-2011, but these webpages are collected from the forums, news websites, and blogs. The 3059 numbers are collected from the news websites, 4533 from the blogs, and 3059 from the forums.
3. Dataset 3: It collected from the China web events. In this contains 50 hot web events in China which is collected from the source like Forums News Websites, and Blogs of the www.Google.com.hk. The length of this Web events is 30 days, and 202 673 total number of Webpages.

C. *Expected Result*

In this section discussed the experimental result of the proposed system.

Table 1 shows the time comparison between the Naive Bayes and C4.5 algorithm. From the time comparison, it is conclude that the time required for the C4.5 is less than the time required for Naive Bayes algorithm.

Table 1: Time Comparison

System	Time Required in ms
Naive Bayes	900 ms
C 4.5	700 ms

Figure 2 shows the time comparison between the Naive Bayes and C4.5 algorithm. From the graphs it is conclude that the time required for the C4.5 is less than the time required for Naive Bayes algorithm.

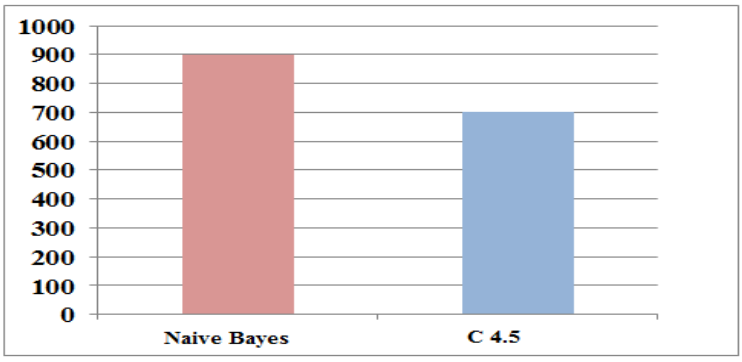

Figure 2: Time Graph

V.CONCLUSION AND FUTURE SCOPE

For the preserving uncertainty semantics in Web events we proposed content based Web event recommendation techniques. This proposed probabilistic technique used to the prediction of response time in the web service. In our model we have assumed that WS is deployed on a cluster of web servers and sometime the delay or crash during WS invocation is because the bad node in sever clustering responds to users requests. With the help of MLP we have predicted the probabilistic behavior of these web servers.

In future work, to minimize uncertainty of keywords our goal is to provide prediction while searching uncertain keywords. So MLP is best prediction algorithm to predict result.

REFERENCES

1. F. Can et al., "New event detection and topic tracking in Turkish," J. Amer. Soc. Inf. Sci. Technol., vol. 61, no. 4, pp. 802-819, Apr. 2010.
2. T. Nguyen, H. Lu, and J. Lu, "Web-page recommendation based on Web usage and domain knowledge," IEEE Trans. Knowl. Data Eng., vol. 26, no. 10, pp. 2574-2587, Oct. 2014.
3. 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 Computer Sydney, NSW, Australia, 2011, pp. 987-994.
4. Y. Deng and R. W. H. Lau, "On delay adjustment for dynamic load balancing in distributed virtual environments," IEEE Trans. Vis. Computer Graph. vol. 18, no. 4, pp. 529ˆa537, Apr. 2012.
5. T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Sequential modeling of topic dynamics with multiple timescales a ACM Trans. Knowledge Disc. Data, vol. 5, no. 4, Feb. 2012, Art. ID 19.
6. Wang Ping, "Web page recommendation based on Markov Logic Network," Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on, Chengdu, 2010, pp. 254-257.
7. D. Huang and H. Li, "The Research of Web Page Recommendation Model Based on FCA and Enterprise Ontology," Computational Intelligence and Industrial Application, 2008. PACIIA '08. Pacific-Asia Workshop on, Wuhan, 2008, pp. 232-236.
8. C. C. Chi, C. H. Kuo and C. C. Peng, "The Designing of a Web Page Recommendation System for ESL," Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007), Niigata, 2007, pp. 730-734.