International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868
Foundation of Computer Science FCS, New York, USA
Volume 11 – No. 6, November 2016 – www.ijais.org

A Novel Approach to Cluster Web Pages Dynamically based on Domain Knowledge

Tina D'abreo
M.E. Student
Thakur College of
Engineering and Technology

Anand Khandare
Asst.Professor,
Computer Dept.
Thakur College of
Engineering and Technology

Prachi Janrao
Asst.Professor,
Computer Dept.
Thakur College of
Engineering and Technology

ABSTRACT
Web Pages which are recommended by the normal web page recommendation system are listed and are not clustered. The web search is based on keyword. The search engine does not understand the meaning of the searched query as it does not have a background domain knowledge of the searched query. The earlier search engine designed clustered the web pages according to static clusters formed [2]. As static clustering, faced some drawbacks of mapping the Web pages, there was a need to find the solution for the same. This paper presents a solution to form the clusters dynamically considering the domains for efficient clustering.

General Terms
Data mining, Semantic-based Mining, Recommendation Systems, Clustering techniques.

Keywords
Semantics, Web Mining, Domain Knowledge, Clustering.

1. INTRODUCTION
Web is the backbone of information today. There has been a huge growth in the resources available online and is ever increasing. Managing the data available online and retrieving the useful and relevant information is difficult. Making correct web pages available to the user is yet not achieved to the great extent. Recommending the web pages to the user is important. The browsing patterns of the user if analyzed will help to recommend the web pages to the user. Web Usage mining mines the patterns from the user logs. The main goal of the recommendation system is to improve the web site usability by knowing the interest of the user.[13]

The web recommendation can be online and off-line with respect to web server activity. In Offline process the knowledge base is built by analyzing, such as server access log file or web logs which are captured from the server. Then these web logs are used in the online component for capturing the list of the web pages and recommend them to the user on his next visit.

Online search activity i the most important and valued one on the Internet and search engines are the gateways to access its information. The issue faced with search engines is that of relevance of the retrieved web pages if they are recommended according to the search domain of the user. So this paper discusses the method which maximizes the search efficiency by recommending the trusted web pages to the user.

2. RELATED WORK
Taoying Li and Yan Chen in [1] have presented a method of clustering web pages based on the search keyword to improvise the search results. Here, they have used a technique to match the degree of the Web page and the keyword to be searched which will show the web pages. The clustering of search result is done according to the matched degree.

In [2] the authors have proposed a novel methodology for the clustering the web pages according to the domain. The Recommendation is based on the factors which are calculated by the API so as to get correct ranking, which will give relevant clustered web pages. But the static clusters sometimes fail to map the web pages to the relevant cluster.

Banage T.G. S Kumara and Incheon Paik T.H.S Siriweera in [3] has proposed a cluster-based Web services recommendation approach to improving the performance of service recommendation. They have considered service semantic similarity recommendation. They have considered semantic similarity and association between services, to cluster the services. Initially candidate cluster and candidate service is selected, thereafter which a filtering process selects a service with better Qos values. This approach has increased performance by reducing search spaces.

In [4] the authors have determined semantic relationships between non-identical, but possibly semantically equivalent, words in multiple domain vocabularies, so as to capture relationships across information obtained in distinct domains. They have used WordNet ontology to measure the semantic relation between the textual words. The result has marked improvement in the precision of predicting the preferences for items to user in the domain.

N. Madurai Meenachi and M. Sai Baba in [5] a survey has been presented a review on the ontology usage in different fields or domains like on the usage of ontology in various domains like Medical, Agriculture, Education, Marine, Communication, Computer, Chemical, etc. Domains like medical, Education have a lot of efforts been put in to generate the ontology whereas in that of plant life there has to be yet more development. The survey supports the idea of enhancement of ontology which would let the semantic web evolve and lead to knowledge sharing based on the domain. This is a continuous process and it needs many experts to participate in its development. Also, the survey highlights the need to develop ontology in the domain of nuclear power.

In [6] the authors - Preetibala Deshmukh and Vikram Garg, have discussed a technique which can maximize the accuracy in low computation time which is EPRAD for effective page rank. By mining the KDD which contain links, the result has shown that the proposed algorithm is efficient as compared to existing PageRank algorithm.

In the paper, A review on Web Recommendation System [7], the authors have explained the urge to have recommendation system to assist the user with their browsing activities. These

systems analyze the user requirements and provide relevant information. This paper has reviewed the various recommendation system, analyzed the problems and their related solutions. To improve the quality of recommendation a new system is introduced which uses KNN and genetic algorithm in mining process to analyze the static weblogs.

Sivakumar J and Ravichandran K.S in [8] have presented a survey report on domain semantic-based Web mining which has analyzed both the domains semantic web and web mining. The paper has discussed the techniques of Semantic Web, Web mining and semantic-based web mining and its application. Since adding semantics to mining provides traditional mining with background knowledge the information retrieval from knowledge management is benefited. Many areas of industry would be benefited from this research to make semantic-based web mining possible to be used in all the fields.

The paper [9] discusses the importance of Search Engines, its role, and it's working. The concept and overview of Search Engine Optimization and its types are also described. The optimization techniques of on page and off-page are important to be used to have top results. Different techniques are discussed from which white Hat SEO technique is best. White Hat SEO technique aims at updating the quality of web page so as to have genuine web links.

In [10] the authors- Ranjna Jain, Neelam Duhan, A.K Sharma, have explained the importance of search engines. Also with a lot of improvements in searching technique, yet the search engine displays the result based on keywords and does not understand the meaning of the keywords. The semantic web which is the next version of WWW is developed with the aim to reduce the issue face by representing structured form and discovering data semantically. The paper shows the survey of some Semantic Search Engines (SSE) where the main focus is on their architecture and the techniques the use to crawl, index the web pages, ranking, etc.

In [12] Carlos Cobos, Martha Mendonza and Elizabeth Leon have developed a algorithm for clustering the results of the web which is called IFCWR. Here, Fuzzy C-Means was used for clustering the web results. IFCWR had shown improvement in clustering quality and performance when compared with Suffix Tree Clustering an Lingo. The AMBIENT and MORESQUE datasets were compared, using precision, recall, fmeasure, SSLK.

In [14], the author of the paper, has proposed the method of structuring of web content. propose the structuring of the Web content as a hierarchical environment, taking into account the site content and structure, the HTML document structure and the term importance. Furthermore, we propose an effective partitional clustering algorithm for a Web site. The preliminary results prove the effectiveness of the new Web content representation and the accuracy of the Web clustering algorithm.

3. METHODOLOGY USED

As the web is massive net of web pages, it becomes difficult to get the result of the searched query in its particular domain. If the Web pages are clustered according to domain, it will be easy for the user to search in the related domain of searched query. This would save the time of the user by simply searching resulted web pages in the respective domain. The clusters were created statically which had a drawback of webpages not being mapped to the appropriate cluster.[2] Hence there was a need to develop a new approach to from the clusters of domain dynamically for every search query.

The process explained below is of creating the domain clusters dynamically, which solves the drawback of static clustering of web pages based on domain knowledge.

As the earlier system of static Clustering, the new search engine will take its input the query or a keyword from the user. As a normal search engine, the crawler will fetch web pages from the server. Google API helps to retrieve the web pages which accurately based on the factors like domain name, domain age, Google PageRank, Alexa rank, etc. The fetched web pages are then clustered according to the domains**.**

The clustering here is dynamic which means no predefined clusters. The number of the clusters is also not limited to a defined k clusters. This will give appropriate domain clusters which will result in the appropriate mapping of web pages.

Following is the stepwise discussion of the method:

Input: Search keyword or search query.

Processing*:*

- *Step1:* Dynamic retrieval of webpages with API, Ranked according to the factors considered with their priority.
- *Step2:* The domain clusters are formed dynamically by considering the web page title object and considering the nouns from the group of words.
- *Step3:* Mapping the retrieved web pages to the appropriate dynamically formed clusters taking into account title object and url object.

Output: Clustered webpages according to the domain.

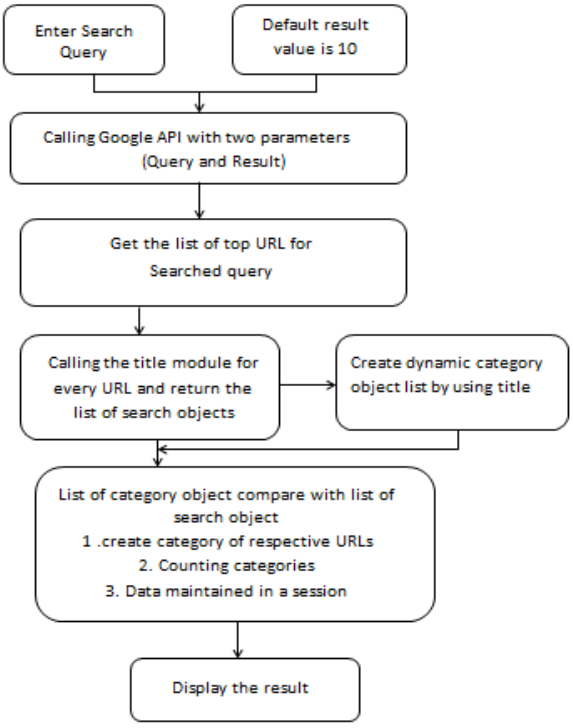

Figure 1. Flow of Dynamic Methodology Used

4. RESULTS

Following are the images of the system implemented using the above explained methodology.

Figure 2 shows the GUI for search query to be entered. Here, we have searched 'Christmas preparation ideas'

Figure 3 shows static clustering of web pages based on domain. To the left the clusters are visible.

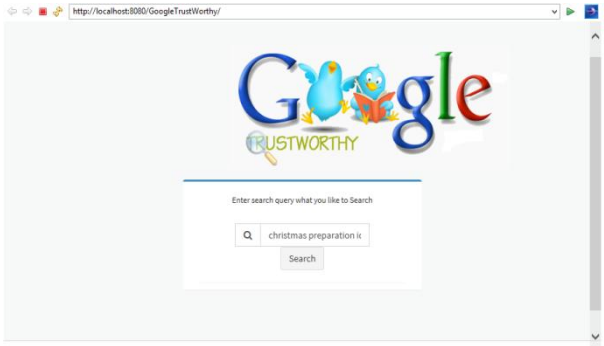

Figure2. Search Query

Figure 4 shows dynamic clustering of web pages based on domain. To the left the clusters are visible.

Figure 5 shows the result when we click the domain cluster health of all the clusters formed. The domain cluster Health has three web pages clustered together.

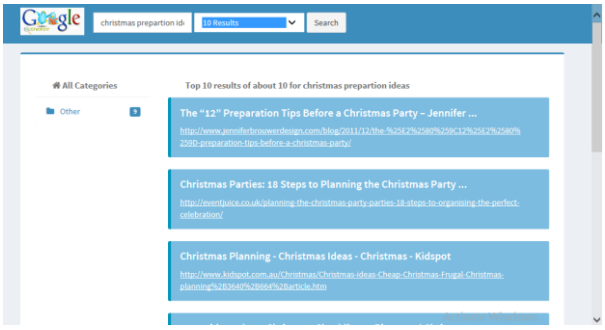

Figure 3. Result of Static clustering

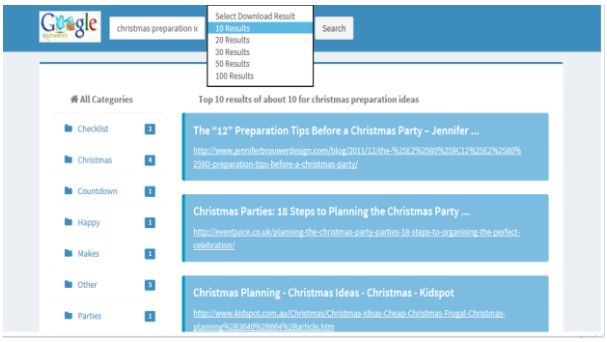

Figure 4. Result of Dynamic clustering- 10 results.

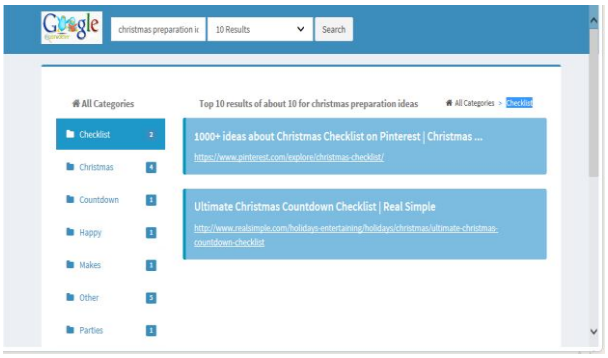

Figure 5. Result of Dynamic Clustering of domain Checklist formed

In static clustering we see that, since no predefined cluster is matched with any of the data objects formed from title and URL of web page, the retrieved web pages are mapped to a general cluster defined 'others'.

The number of clusters formed differs, as also some clusters in dynamic a new which are not formed in static clustering.

5. CONCLUSION

This paper has discussed the improved methodology for dynamic cluster formation. In [2], the author has discussed static clustering of Web pages, which had drawbacks as mentioned in section 4.1 of this paper. Dynamic clustering helps to form the clusters dynamically at any instant of the search query requested. Since of the formation of clusters is dynamic the mapping of web pages to the cluster will be accurate depending on its domain. This will provide the user with the optimized results which will lead to an efficient search in the domain of his interest.

6. ACKNOWLEDGEMENT

I would like to thank my guide Mr. Anand Khandare and my co-guide Ms. Prachi Janrao for their constant support and help in this research work.

7. REFERENCES

[1] Taoying Li and Yan Chen," Web Page Clustering Based on Searching Keywords", 2010 IEEE International Conference on Intelligent Computation Technology and Automation.

[2] Tina D'abreo, Anand Khandare and Prachi Janrao, "Static Clustering of Web pages for Relevant Recommendation", IJARCCE Vol.5, Issue 9, September 2016.

[3] B. T. G. S. Kumara, I. Paik, T. H. A. S. Siriweera and K. R. C. Koswatte, "Cluster-Based Web Service Recommendation," 2016 IEEE International Conference on Services Computing (SCC), San Francisco, CA, 2016, pp. 348-355.

[4] Anil Kumar and Nitesh Kumar and Muzammil Hussain, Santanu Chaudhury and Sumeet Agarwal, "Semantic clustering-based cross-domain recommendation", 2014 IEEE Computational Intelligence and Data Mining.

[5] N. Madurai Meenachi and M. Sai Baba, "A Survey on usage of Ontology in Different Domains", IJAIS, Volume 4– No.2, September 2012.

[6] Preetibala Deshmukh and Vikram Garg, "An Enhanced Page Rank Algorithm over Domain", IJCA, Volume 139 – No.1, April 2016.

[7] Animesh Shrivastav and Anand Singh Rajawat, "A Review on Web Recommendation System", IJCA Volume 83 – No.17, December 2013

[8] Sivakumar J and Ravichandran K.S, "A Review on Semantic-Based Web Mining and its Applications", IJET Vol 5 No 1 Feb-Mar 2013.

[9] Ayush Jain,"The Role and Importance of Search Engine and Search Engine Optimization", IJETTCS, Volume 2, Issue 3, May – June 2013.

[10] Ranjna Jain, Neelam Duhan, A.K Sharma, "Comparative Study on Semantic Search Engines", IJCA Volume 131 – No.14, December2015.

[11] Adam Schenker, Mark Last, Horst Bunke, and Abraham Kandel,"Clustering of Web Documents Using A Graph Model", available at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.30

[12] Carlos Cobos, Martha Mendonza and Elizabeth Leon, "Clustering of Web Search Results based on an Iterative Fuzzy C-means Algorithm and Bayesian Information Criterion",

[13] R.Thiyagarajan, K. Thangavel and R. Rathipriya, "Recommendation of Web pages using weighted k-means clustering", IJCA, Volume 86 – No 14, January 2014.

[14] Ioan Agavriloaei, Adrian Alexandrescu and Mitică Craus, "Improving Web Clustering through a New Modeling for Web Documents",IEEE, System Theory, Control, and Computing (ICSTCC), 2011 15th International Conference.