IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

Static Clustering of Web pages for Relevant Recommendation

Tina D Abreo[1], Anand Khandare[2], Prachi Janrao[3]

M.E Student, Computer Engineering Department, Thakur College of Engineering & Technology, Mumbai, India [1]

Asst. Professor, Computer Engineering Department, Thakur College of Engineering & Technology, Mumbai, India[2,3]

Abstract: Today the Web is the large repository of information. The existence of a profusion of information, in combination with the continuously updating the web information, make its retrieval a difficult process for the user. The search activity carried out by the user is sometimes not productive according to his query. The concept of understanding the query of the user and the system varies. This intern gives rise to the results which are irrelevant for the user. The web pages in the results which are retrieved may sometimes not be according to the domain specific meaning. This leads to occupying more time of the user in searching the required web page. By clustering the web pages according to the domain will give better search results for the requested query. This paper discusses the recommendation of web pages and clustering it statically, to check the efficiency of the end result.

Keywords: Semantic Web, Domain Knowledge, Recommendation, Clustering.

I. INTRODUCTION

The internet is enormous, diverse and of heterogeneous nature which is again dynamic. As people have more attention towards the internet, it is used frequently due to which there more and more research is being done. Day-to-day there is an addition of numerous web pages and also deletion of many. The Web is continuously developing a large traffic and complexity on the Websites present, also when the new web pages are added. The data present is also unstructured which sometimes minimizes the efficient information retrieval. Hence it is important to obtain the relevant information required. As a solution, there was a need to understand the semantics related to the information available on the Web. Web mining uses the techniques of traditional data mining which is used to discover and also extract the required information from web resources. We have three classes from which information can be discovered: a) Web Activity- an activity which tracks the server logs and web browser. b) Web graphs- which basically give the links between web pages, users, and other related data. c) Web content-simple, the data contained in a particular web page and on a web page [10]

We can compare the web mining with data mining on basic three differences observed: a) Scale - In data mining considering say, 2 million of records would be a large dataset, while in web mining 10 million would not be a big statistics. b) Access - data mining of an organization is private requires the access rights, whereas as in web mining data is public and rarely data access rights are required. c) Structure – In data mining, the data to be mined is stored in a data repository that is the database, hence it has some level of structure. And in Web mining, the processing is done on the data which is unstructured or semi-structured obtained from the web pages [10]

Web Mining involves the combinational features of semantic Web and Web mining, which improves the level of intelligence access to web information. The domain of Web mining is a highly researched topic of several web communities such as AI, Information Retrieval, Machine Learning and Natural Language Processing. As it is a combination of two research areas which are Data Mining and World Wide Web, it makes the new technologies and infrastructure components essential to building a web structure which handles its users efficiently. Recommendation plays an important role in any personalization system.

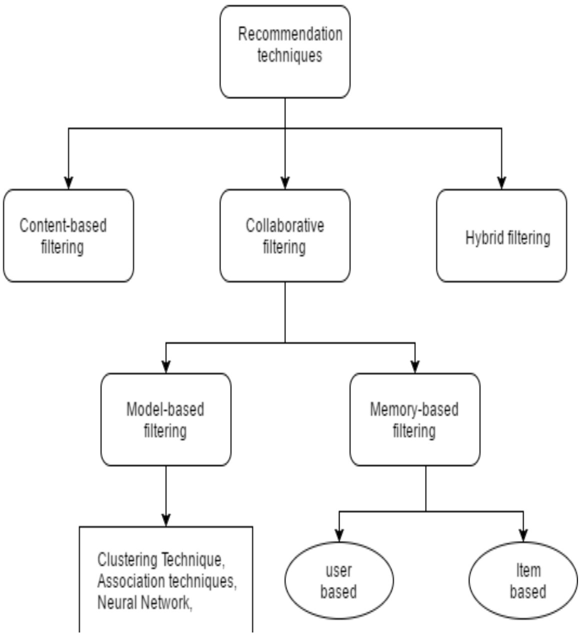


Fig. 1 Recommendation techniques [2,3]

DOI 10.17148/IJARCCE.2016.5988

Based on the reviews and survey of different objects over a period of time helps to get its semantic information, which is used to categorize or cluster in the same group. We have three general types of filtering used in recommendation which is: a) content-based filtering technique b) collaborative filtering technique c) Hybrid filtering technique.

Content-based technique emphasizes on the analysis of the attributes of objects in order to generate predictions. Content-based filtering is used when documents like web pages, publications, and news have to be recommended. In this, the keyword searched is used for filtering from a large data. [2] Collaborative Filtering technique takes into account the preferences of the user in a similar interest and predicts the result for next similar query searched by the user.

Two types we have are memory-based and model-based. In memory-based, the techniques work by considering the user reviews or item reviews and then by using the information of the pattern formed on basis of k nearest neighbors. Whereas, Model-based techniques used data mining techniques like clustering, association algorithms, etc. to give the efficient recommendation to the user.

Hybrid Filtering techniques are the combinations of pure recommendation techniques which are content and collaborative filtering which results in a better system to avoid the issues faced by pure recommendation system [2]. The aim of hybrid techniques is to make use of combination of algorithms which will provide more accurate and effective result of recommendations than by using single algorithm as the disadvantages of one algorithm can be overcome by another algorithm [3]

II. RELATED WORK

There are various researches in the field of Mining and Semantics which has enriched the field. A survey on text categorization [4] has discussed the feature selection methods which reduce the dimensionality of the dataset by removing features that are considered irrelevant for the classification. It also discusses several approaches of text categorization, feature selection methods, and applications of text categorization.

DipikaSahu, YaminiChouhan in [5] reports the summary of various techniques of web mining approached from the following angle like Feature Extraction, Transformation and Representation and Data Mining Techniques in many application domains.

Also, the introduction to linking mining and block-link mining with its issues has led to a new research point in future. Web recommendation system is usually used for filtering information to predict the user's upcoming browsing activity and then accordingly recommending the web pages items to the users that are likely to be of their interest. The authors- R. Thiyagarajan, K. Thangavel, R. Rathipriya [6], have proposed a framework which takes into account the offline and online system. Offline system provides knowledge base by analyzing the logs of the server. Then the Online system then uses the filtered data to predict the next browsing pattern.

The main aim of the system is to group the user's behavior which is similar, to return the efficient result. K-means clustering algorithm is applied to form clusters and factor MSR is done to evaluate the outcome. The implementation of this algorithm shows that the prediction of user activity capturing is much accurate.

There can be cases when the overlapping clusters are obtained for that more clickstreams could be considered to have a precise recommendation. Knowledge extraction for the semantic web using web mining[7], explains that adding knowledge or semantics is not only helpful for users, but the machine should also understand investigation of the problem of extracting knowledge from a large number of web documents in order to develop ontology is done.

This research introduces web usage patterns as a novel source of semantics in ontology learning. The methodology proposed by the authors combines the web content mining with the web usage mining to extract the knowledge.

Hence, the perspective of both the web user's and web authors are captured in accordance with the content of the web, which leads to extraction of the set of conceptual relationships. Web structure mining can also be used in future for more enhancements in extracting knowledge. In Semantic-based document clustering: A Detailed Review [8], a survey of different text algorithm is done by the authors which would be useful for the future research work.

As the document clustering using data mining technique is an unsupervised paradigm, the reviewed algorithms could be helpful for clustering the documents according to the semantics of the document. The researchers in Recent Development in Text Clustering technologies [9] have discussed various ways of text clustering of documents according to semantics. They have described the issues of overlapping clusters. The traditional way of clustering text is as shown:

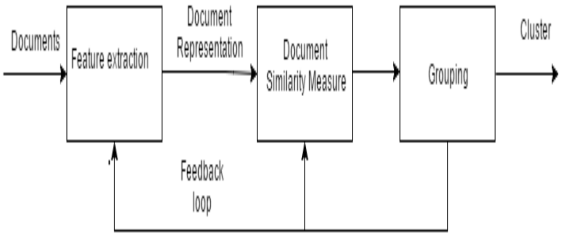


Fig. 1 Stages in Text Clustering [9]

DOI 10.17148/IJARCCE.2016.5988

III.METHODOLOGY USED

As there is the need for retrieval of web pages which would be relevant and useful for the user, there has to be a system which would understand the main semantic behind the every searched query. Here we will discuss the static way of clustering the web pages according to domain knowledge of the searched query. The proposed method of static clustering of web pages is being inspired by a k-means algorithm for text clustering according to their domain. The system will be taking a search query from the user or rather a keyword from the user. This keyword will be used to fetch the web pages from the server. Google API helps to retrieve the web pages which assure us that the retrieval is going to be accurate. As it takes into consideration factors like domain name, on-page factor, etc. into an account. Together with the keyword and the number of web pages to be recommended by the system the fetched web pages a brought down to the system. Further is an important part of the system where the essential processing is done to get the result according to the domain. In static clustering, we have predefined the cluster names according to the domain which is generalized. Following is the stepwise discussion of the method:

1) We enter the query to be searched which is our input also with the number of URL of web pages to be retrieved.

2) The web pages are retrieved and recommended to the user.

3) The recommended web pages are listed according to the factors like domain age, etc.

4) The clusters we have predefined are used to form the cluster of web pages which are being recommended to the user.

5) The web page title and the URL object is mapped with the predefined clusters and categorized respectively

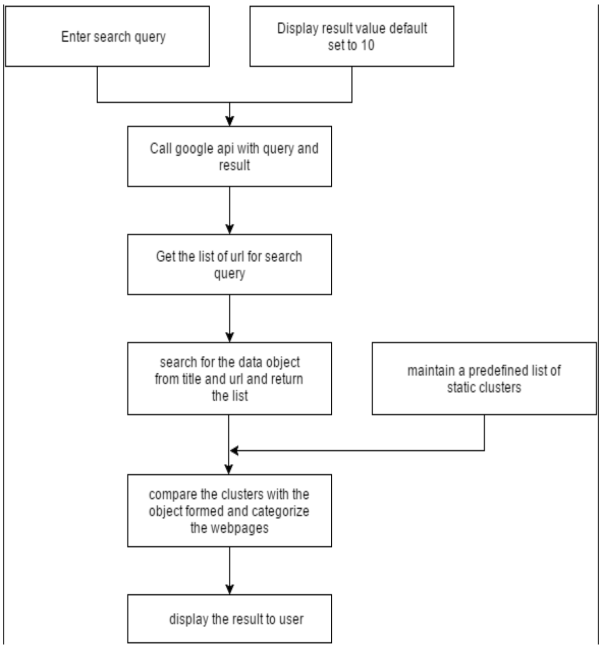

Enter search query
Display result value default set to 10
Call google api with query and result
Get the list of url for search query
search for the data object from title and url and return the list
maintain a predefined list of static clusters
compare the clusters with the object formed and categorize the webpages
display the result to user

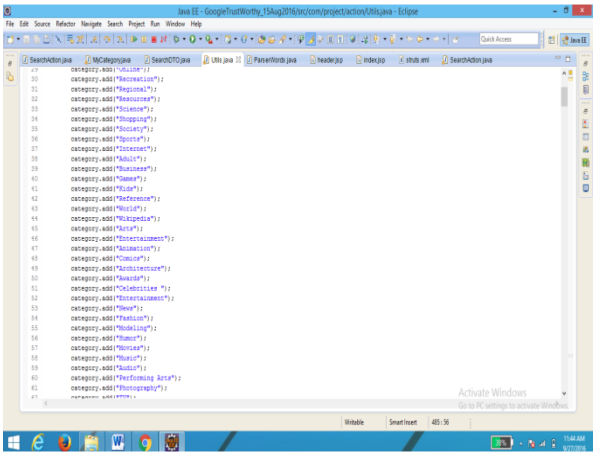
Fig. 3 Flow of methodology used

IV.RESULT

The above-mentioned method for static clustering the web pages after implementing showed the following results of web page recommendation.

Fig. 4 shows the predefined static clusters we make in our program.

Fig. 5 shows the UI of the user input to the system.


Java EE - GoogleTrustWorthy_15Aug2016/src/com/project/action/Utils.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
SearchAction.java
MyCategory.java
SearchDTO.java
Utils.java
ParserWords.java
header.jsp
index.jsp
struts.xml
SearchAction.java

```
        category.add("Recreation");
        category.add("Regional");
        category.add("Resources");
        category.add("Science");
        category.add("Shopping");
        category.add("Society");
        category.add("Sports");
        category.add("Internet");
        category.add("Adult");
        category.add("Business");
        category.add("Games");
        category.add("Kids");
        category.add("Reference");
        category.add("World");
        category.add("Wikipedia");
        category.add("Arts");
        category.add("Entertainment");
        category.add("Animation");
        category.add("Comics");
        category.add("Architecture");
        category.add("Awards");
        category.add("Celebrities ");
        category.add("Entertainment");
        category.add("News");
        category.add("Fashion");
        category.add("Modeling");
        category.add("Humor");
        category.add("Movies");
        category.add("Music");
        category.add("Audio");
        category.add("Performing Arts");
        category.add("Photography");
```

Fig.4 Predefined clusters/ categories

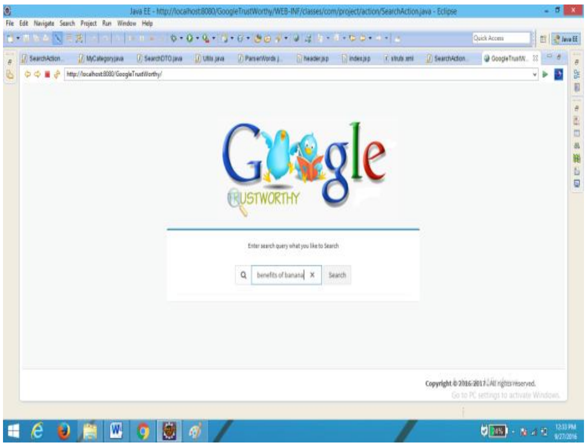
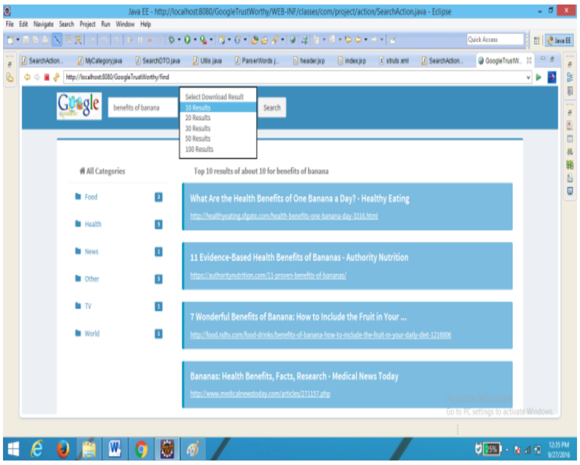

Java EE - http://localhost:8080/GoogleTrustWorthy/WEB-INF/classes/com/project/action/SearchAction.java - Eclipse
http://localhost:8080/GoogleTrustWorthy/
Google
TRUSTWORTHY
Enter search query what you like to Search
benefits of banana
Search
Copyright © 2016-2017. All rights reserved.

Fig. 5 User query to be search


Java EE - http://localhost:8080/GoogleTrustWorthy/WEB-INF/classes/com/project/action/SearchAction.java - Eclipse
http://localhost:8080/GoogleTrustWorthy/find
Google
benefits of banana
Select Download Result
10 Results
20 Results
30 Results
50 Results
100 Results
Search
All Categories
Food
Health
News
Other
TV
World

Top 10 results of about 10 for benefits of banana

What Are the Health Benefits of One Banana a Day? - Healthy Eating
http://healthyeating.sfgate.com/health-benefits-one-banana-day-3316.html

11 Evidence-Based Health Benefits of Bananas - Authority Nutrition
https://authoritynutrition.com/11-proven-benefits-of-bananas/

7 Wonderful Benefits of Banana: How to Include the Fruit in Your ...
http://food.ndtv.com/food-drinks/benefits-of-banana-how-to-include-the-fruit-in-your-daily-diet-1216006

Bananas: Health Benefits, Facts, Research - Medical News Today
http://www.medicalnewstoday.com/articles/271157.php

Fig. 6 Result of static clusters formed

DOI 10.17148/IJARCCE.2016.5988

The clusters are from the predefined static list. If a particular cluster is not pre- defined, then the web pages are would be mapped other nearby cluster. To solve this issue there is an urge to develop a method to dynamically form the clusters.

V. CONCLUSION

The recommendation systems have evolved significantly and will be evolving to provide the user with much efficient search results. This paper has discussed forming the clusters of web pages searched according to the domain related to the searched query. The clusters formed here are static which may sometimes lead to the incorrect mapping of the web pages. To solve this issue a new idea of having dynamic clusters is under research.

ACKNOWLEDGMENT

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

REFERENCES

[1] ThiThanh Sang Nguyen, Hai Yan Lu, and Jie, "Web-Page Recommendation Based on Web Usage and Domain Knowledge", IEEE Transaction on Knowledge and Data Engineering, Vol. 26, No. 10, October, 2014.
[2] Celine M R, DrSheetalRathi, "Recommendation of Recommender System", IJARCCE, VOL.5, Issue 8, August 2016.
[3] F.O.Isinkaye a, *, Y.O. Folajimi b, B.A. Ojokoh."Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal (2015) 16, 261–273
[4] Senthil Kumar B, BhavithaVerma B, "A Survey on Text Categorization", IJARCCE, Vol.5, Issue 8, August 2016.
[5] DipikaSahu, YaminiChouhan, "Comparative Study and Analysis of Techniques of Web Mining", IJARCCE, Vol.5, Issue 8, August 2016.
[6] R. Thiyagaraja, K. Thangavel, R. Rathipriya, "Usage Profile based Recommendation system", 2014 International Conference on Intelligent Computing Applications, IEEE.
[7] Jayatilaka,A.D.S, Wimalarathne, G.D.S.P ," Knowledge extraction for semantic web using web mining", 2011 International Conference on Advances in ICT for Emerging Regions, IEEE..
[8] Neepa Shah, SuneetaMahajan, "Semantic Based Document Clustering: A Detailed Review", IJCA, Volume 52 - No 5, August 2012.
[9] Saurabh Sharma, Vishal Gupta, "Recent Development of Text Clustering Techniques", IJCA, Volume 37 – No. 6, August 2012.
[10] http://www.scaleunlimited.com/about/web-mining/
[11] VidyaKannan, Dr G.N Srinivasan, "Yet Another Way of Ranking Web Documents Based on Semantics Similarity", IJARCCE, Vol.3, Issue 4,April 2014..
[12] K.Suneetha and M. Usha Rani, "Performance Analysis of web page recommendation algorithm based on weighted sequential patterns and markov model", IJCSI, Vol.10, Issue 1, January 2013

DOI 10.17148/IJARCCE.2016.5988