URL Mining Using Agglomerative Clustering Algorithm

Chinmay R. Deshmukh, R .R. Shelke

Abstract: The tremendous growth of the web world incorporates application of data mining techniques to the web logs. Data Mining and World Wide Web encompasses an important and active area of research. Web log mining is analysis of web log files with web pages sequences. Web mining is broadly classified as web content mining, web usage mining and web structure mining. Web usage mining is a technique to discover usage patterns from Web data, in order to understand and better serve the needs of Web-based applications. URL mining refers to a subclass of Web mining that helps us to investigate the details of a Uniform Resource Locator. URL mining can be advantageous in the fields of security and protection. The paper introduces a technique for mining a collection of user transactions with an Internet search engine to discover clusters of similar queries and similar URLs. The information we exploit is a "clickthrough data": each record consist of a user's query to a search engine along with the URL which the user selected from among the candidates offered by search engine. By viewing this dataset as a bipartite graph, with the vertices on one side corresponding to queries and on the other side to URLs, one can apply an agglomerative clustering algorithm to the graph's vertices to identify related queries and URLs.

Index Terms: Agglomerative Clustering Algorithm, URL Mining, Re-ranking, Query Log analysis.

1. INTRODUCTION

World Wide Web (WWW) is an unstructured collection of pages and hyperlinks. People from different backgrounds and interests access and provide web pages. Application of data mining approaches on World Wide Web is referred as web mining. Web mining has attracted a lot of researchers due to huge amount of active data available on the World Wide Web. Broadly, web mining tasks include web usage mining, web content mining and web structure mining. Web content mining is a process of discovering information from millions of sources across the World Wide Web. User interaction on the web is recorded on a web logs. As each user interaction corresponds to a mouse click it is oftenly referred as click stream. Web usage mining is performing mining on web usage data or web logs. Extracting patterns from on line information, such as HTML files or E-mails is referred as web content mining. The intuition is that a hyperlink from document A to document B implies that the author of document A thinks document B contains worthwhile information. With the increasing size and popularity of Internet, there exist over a billion of static web pages and some commercial web engines serves tens of millions of queries per day. So, there is a strong need of automatic methods that can organize this data. One strategy for bringing the degree of order to a massive, unstructured dataset is to group similar items together.

- *Chinmay R. Deshmukh, M.E (Computer Science and Engg.), H.V.P.M COET, Amravati, Amravati, India.*
- *Prof. R .R. Shelke, Computer Science and Engineering, H.V.P.M COET, Amravati. Amravati, India.*

The clustering strategy introduced here follows from two related observations. First, the fact that users with same information need may phrase their query differently to a search engine – cheetahs and wild cats – but select the same URL from among those offered to them to fulfill the need that they are related. Second , the fact that after issuing the same query, users may visit two different URLs – "www.fundz.com" and "www.mutualfundsite.com" say , is evidence that URLs are similar and are clustered together.

2. CLUSTERING

Clustering methods can be used in order to find groups of documents with similar content. So, the result of clustering is typically a partition (also called) clustering P, a set of clusters P. Each cluster consists of a number of documents d. Objects — in our case documents — of a cluster should be similar and dissimilar to documents of other clusters. Usually the quality of clusters is considered better if the contents of the documents within one cluster are more similar and between the clusters more dissimilar. Clustering algorithms compute the clusters based on the attributes of the data and measures of (dis)similarity. In the following, we first introduce standard evaluation methods and present then details for agglomerative clustering approaches, k-means, etc. We will finish the clustering section with a short overview of other clustering approaches used. In this section, we present the agglomerative clustering for clustering click stream transactions using upper similarity approximation. The algorithm for clustering click stream transactions is given below:

Algorithm: Rough Agglomerative Clustering
Input: A set of n objects in a data set U = {x1, x2 ….xn}, Threshold θ, the number of clusters p (≤ n)
Output: Cluster scheme C
Step 1: Start
Step 2: Initially consider each object of U as a cluster of one member Ci = {xi} and C = {C1, C2, ., Cn }
Step 3: For each pair of clusters Ci and Cj calculate sim (Ci, Cj) = (Ci∩Cj)/ (Ci U Cj)
Step 4: For each cluster Ci, find out the similarity upper approximation Si, for a given threshold θ.

Step 5: If Si = Sj, form a new cluster Cij = Ci ∪ Cj, i.e. put xi and xj in the same cluster.
Step 6: Update C
Step 7: Repeat Steps 5 and 6 till there is no change in the number of clusters.
Step 8: Output C.
Step 9: Stop.

3. URL MINING

Web log mining is analysis of web log files with web pages sequences. Web mining is broadly classified as web content mining, web usage mining and web structure mining. Web content mining is a process of discovering information from millions of sources across the World Wide Web. User interaction on the web is recorded on a web logs. As each user interaction corresponds to a mouse click & it is often referred as "click stream". Click stream is a sequence of URLs browsed by a user within a particular website in one session. URL Mining is one of the subsequences of Web Mining in which we can identify the actual developer of URL, the date on which this URL was made, the purpose of URL and so on. The need of URL Mining resides in its Previous Work which included some shortcomings as follows:

- Query Log Analysis
- Clustering URL's which include

1. Text Free Pages : A distance function calculated based on the relationship between two web pages , for e.g. the relationship between two pages cant be found out if a webpage contains just a picture of Emu and the other webpage contains just appearance and behavior of Emu.
2. Pages with restricted access: URL's may be password protected or temporarily unavailable, making its cluster rarely or totally unusable.

4. RE-RANKING

We consider another search task for the evaluation of our subtopic mining method, namely, searches result re-ranking. We chose to do so because we can easily use click-through data to conduct quantitative evaluation.

4.1 Our Method:-

When the search system provides the results to the user there might be multiple ways to present it if the query contains multiple subtopics, i.e., it is ambiguous or multi-faceted. These include simple ranking, clustering of URLs by subtopics, dynamic ranking of URLs belonging to the same topic, as well as re-ranking of URLs by subtopics.

Re-ranking is conducted in the following way:-

First the user will enter his general search query. After pressing "Search" button the appropriate results will be returned. The user can select any of the result from drop-down list and can open that link in separate web page after pressing the button "Open Result and Update User Profile". After his browsing is over, the user will enter another search input similar to the previous one, but this time he'll search by pressing button "Search based on User Profile". This time the results will be given but in re-ranked manner only. In this way, we can perform the task of "META SEARCH ENGINE". , which was the main focus of this paper. Also the user's profile can be seen via button "View and Manage User Profile".

Figure 1 shows an example UI in which the home page is composed of a Textbox for user to enter his search query. Also it includes 2 buttons at the top for search (mainly simple search and re-ranked search). At the bottom, we can see two more buttons indicating "Open Result and Update User Profile" and "View and Manage User Profile". Although re-ranking is a simple, interactive approach to search results presentation, there seems to be no study before on the approach to the best of our knowledge. Note that re-ranking can only be done for head queries, because it is based on log data mining, which is a shortcoming that any data mining method may suffer from.

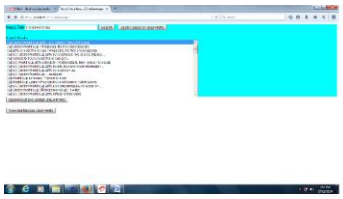

***Figure 1**: Home page of Our Meta search engine showing results for "Cricket world cup" input.*

Also figure 2 is given below as:-

***Figure 2:** Search result for query 'Cricket world cup'*

The second page of this project focuses on user's profile which will include a log of user search. One can get results based on User's profile on home page, that'll in turn show the search log in second page.

4.2 Evaluation:-

We tried to make a system based on concepts of content mining that'll yield results like a general search engine. The concept implemented is also well known as "META SEARCH ENGINE". The search log data contains the ranking results of URLs as well as the clicks on URLs in the searches. Next, we used the Microsoft's ASP.NET webpage development technology in order to build our own simple Meta Search Engine as a part of this project. Note that because of the one subtopic per search phenomenon, the clicked URLs in a search usually belong to the same subtopic.

5. QUERY LOG ANALYSIS

Query log analysis includes filtering methods and some crawling mechanisms that can be evaluated on the basis of Part Of speech tagging. There are two main approaches in retrieving the user's previous work related to a particular URL. This will include mining URLs for the evidence of pages visited by the user in the past session.

- Filtering methods remove words from the dictionary and thus from the documents. A standard filtering method is stop word filtering. The idea of stop word filtering is to remove words that bear little or no content information, like articles, conjunctions, prepositions, etc.
- Part-of-speech tagging (POS) determines the part of speech tag, e.g. noun, verb, adjective, etc. for each term. Text chunking aims at grouping adjacent words in a sentence. An example of a chunk is the noun phrase "the current account deficit". Word Sense Disambiguation (WSD) tries to resolve the ambiguity in the meaning of single words or phrases. An example is 'bank' which may have – among others – the senses 'financial institution' or the 'border of a river or lake'. Thus, instead of terms the specific meanings could be stored in the vector space representation. This leads to a bigger dictionary but considers the semantic of a term in the representation. Parsing produces a full parse tree of a sentence. From the parse, we can find the relation of each word in the sentence to all the others, and typically also its function in the sentence.

6. CONCLUSION

Clustering is the task of grouping similar objects into clusters. Clustering concepts are implemented here with the help of OOPS concepts. In this work, we presented a rough agglomerative clustering technique to cluster click-stream transactions based on Upper similarity approximation. We've experimented our approach on a click-stream dataset, which was collected from a Hungarian on-line news portal. Each click-stream transaction is of variable length. The presented method can yield results as per general search and also per user's preference after re-ranking. This is the model META SEARCH ENGINE that can yield results smoothly.

7. REFERENCES

[1] Jhoshi, A. and Krishnapuram, R., " Robust fuzzy clustering methods to support web mining, proceedings of the workshop on Data Mining and Knowledge Discovery, SIGMOD ' 98, Seattle, pp. 15/1 – 15/8, June 1998.

[2] Cooley, R., Web Usage Mining: Discovery and Applications of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.

[3] S, K., Radha Krishna, P.: Mining web data using clustering technique for web personalization, Int. Jour. of Computational Intelligence and Applications, 2(3) (2002) 255-265.