Mining techniques – A Website usage perspective

***Ms.R.Sangeetha**
****Dr.K.Meenakshisundaram**

*Assistant Professor, Meenaakshi Ramasamy Arts &Science College, Thathanur.
Ariyalur (D.T.)-621804
**Associate Professor, Department of Computer Science, Erode Arts & Science College,
(Autonomous) Erode–638 009.

Abstract

Now a day the usage of website address is increasing every day. The Web can be regarded as a large repository of diversified in formation in the form of millions of websites distributed across the globe.So increasing number of websites in the Web has made it extremely difficult for users to find the right information that satisfies their current needs. In order to address this problem, many researchers explored Web Mining as a way of developing intelligent websites, which could present the information available in a website in a more meaningful way by relating it to a user's need.

Key words: web content, web mining, way post approach.

1. Introduction

Now days The World Wide Web or simply "Web" originally started as a research development for information sharing purposes. However, today it consists of a vast heterogeneous information source distributed over numerous websites all over the world. It has become a part of life for almost everyone around the globe who has

For example communication via email, messaging services (e.g., chat groups), up-to-date news, blogging, online shopping and weather reports are a few elements of the Web that have influenced many people's lives. In addition, more and more business organizations are making their presence on the Web to increase their business flexibility and earning potential to be able to compete on a more level playing field with other e-commerce businesses.

The Web, officially founded in 1989 by Sir Tim Berners-Lee, consisted of no more than a single web server (i.e., website) with a single page describing the project itself. Since then the growth of the Web continues to be exponential and it has not ceased to grow. (Ansari et.al.) It was reported that the Web had a staggering 74.4 million registered websites online in October 2005. On the other hand, a recent survey conducted in April 2008 noted that the Web had 165.7 million websites online with the month's most growth of 3.1 million sites coming from the United States and 1.1 million sites of these were from Google's blogger service. This suggests that the Web will continue to grow as long as there are people who keep developing it.

Web Mining is the application of data mining techniques on web resources to find patterns in the Web. Web content data consists of textual information contained in the documents of a website. Web structure data refers to the hyperlinks interconnecting the collection of documents in a website. Each document in a website may have a number of outgoing and incoming hyperlinks. A link between two documents in a website suggests that the documents may be related and may contain relevant information. Web structure mining applies data mining techniques on this network of hyperlinks structure to extract information that can be used for a variety of purposes, for example web crawling. The earliest approach to intelligent websites was customization, which can be traced back to the mid-1990s.

Customization involves changing the interface and contents of a website to match a user's need. This is done through a manual form filling method which requires the users to choose from a set of predefined interest categories or specify their interests.

Recommender systems are web applications that employ web usage mining techniques on web logs to come up with personalized recommendations on interesting information in a website. These recommendations could be as simple as suggesting popular links in a website or involve a more sophisticated real time approach, for example proposing links to documents that are related to a user's need as the user browses a website. All of the above mentioned intelligent web applications are primarily focused on filtering the information available in a website to suit a user's need.

1.1Objectives of the paper

Web usage mining techniques that could be applied to web logs to discover users' paths and identify way post from them which, can be used to suggest potential shortcuts between documents in a website. Listed below are a set of specific objectives that we will address:

1. *Identify a group of target documents based on users' access patterns and extract navigational paths terminating at these documents from users' sessions.*
2. *Group similar navigational paths together and find hyperlinks (i.e., way posts) that can act as potential shortcuts points by identifying documents that commonly co-occur in these paths.*
3. *Evaluate the shortcuts generated from the way posts on their usefulness in providing Shortened navigational paths.*

2.1Web Usage Mining

According to a recent survey, the number of web usage mining studies undertaken over the past few years has been rapidly increasing. Web usage mining is the application of data mining techniques on large web repositories to discover useful knowledge about user's behavioral patterns and website usage statistics that can be used for various website design tasks. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. This is probably because web logs are the easiest and cheapest way of gathering information about the users and a website. Other sources of usage data may include proxy servers and client side logs.

2.2Web Content Mining

Documents in a website contain textual information which is highly unstructured and varies accordingly to the purpose of a website. Unlike a highly structured database, retrieving information from a collection of documents is a challenging task as the documents do not share a common content representation. Furthermore there is no standard range of attributes available that can be used to distinguish one document from another in a website. In order to extract information from documents, a suitable document representation model is required. One such approach is the bag-of-words (BOW), commonly used to represent a document as a large collection of words. In BOW, a document's text is broken up into individual tokens or terms. Words that best distinguish a document are then determined by ranking terms using a frequency-based method (e.g., Term Frequency Inverse Document Frequency).

Web content mining involves applying data mining techniques on text contained in documents to extract useful knowledge from them, such as the topic relations between documents. This particular knowledge is commonly used for document classification. Document classification involves creating a topic hierarchy based on all documents of a website by categorizing each document into a topic class it fits best. The topic hierarchy built can then be used for a variety of applications such as web browsing and web crawling. The majority of current studies rely on a manual classification technique which requires feedback from users on each document's topic category and its importance to their information needs. However, manual classification of documents is a tedious and time-consuming task. It requires a reasonable number of feedbacks to generate a topic hierarchy and many users do not have the time to tag every document they visit in a website. This suggests the need for a method to automatically discover a document's topic and its relevance to a user's need.

Godoy and Amandi (16) addressed the above issue with their Web Document Conceptual Clustering (Web DCC) technique. The technique utilises an incremental clustering approach that learns a topic hierarchy from the content of documents by combining the use of linear classifiers with k-Nearest Neighbour (kNN) algorithm. Initially, documents are preprocessed into BOW and are represented with weighted feature-vectors in space. Web DCC starts with a root node and creates new child nodes whenever it encounters a document that could not be classified under the existing nodes of the topic hierarchy. Each node in the hierarchy has a separate linear classifier and is associated to a topic category represented by a set of words that best describes it.

A feature selection method is used to choose the set of words that describes a category at each node. At each node, the linear classifier constructs a prototype instance which is used to compare and decide whether or not a document belongs to the current topic category. The advantage of this technique is that it is an incremental topic learning mechanism that avoids the need for explicit user feedback in document classification. However, Web DCC is computationally expensive as it needs to explore all documents in a website to generate the topic hierarchy.

2.3Web Structure Mining

A website consists of a collection of documents interconnected by hyperlinks. Each document may have a number of outgoing and incoming hyperlinks. A hyperlink from document A to document B may represent the continuity between units of information in a website.

A crawler starts with an initial list of URLs to visit and cycles through the list visiting each document in turn. While visiting a document it identifies all hyperlinks and adds them to the list of URLs to visit. This task is carried out recursively based on a set of selection policies for later processing by indexes. A search engine uses the information from crawlers to find documents that match a query. The efficiency of a search engine can be measured by the proportion of relevant documents that are actually retrieved (recall) and the proportion of documents retrieved that are actually relevant (precision). Taking into consideration that users' information needs vary significantly from one individual to another, a search engine tries to retrieve all relevant documents.

However, the Web spans over numerous websites and it is impossible to search the Web's entire document collection. Furthermore, a search engine can only retrieve a fraction of documents within a given time in response to a query. Therefore, it makes more sense to retrieve just a few selected important documents from a website, rather than its entire collection. This introduces the need for a way to identify and retrieve only the important documents of a website.

Page, Brin, Motwani and Winograd addressed the issue of distinguishing between important and non-important documents with their Page Rank method that computes a rank for every document in the Web. Page Rank is based on the citation count technique which uses the number of citations to a publication as an indicator of its importance. The more citations a publication receives the more important it becomes. Similarly, Page Rank uses the incoming hyperlink structure as citations to a document from other documents to approximate the overall importance of a document.

Later studies into web structure mining began employing hybrid methods which, included the use structure and content mining techniques to find interesting documents that are topically oriented to a user's need. Introduce a technique called Focussed Crawling to perform topic specific web crawls. Focussed crawlers tend to visit documents related to a specific topic rather than visiting the entire web.

3. **Way posts Approach**

The objective is to introduce our approach to an adaptive website and to provide details of the work carried out in acquiring data from the web logs. In particular, we will highlight a research gap found in the area of adaptive websites and suggest our approach to address it. Following that, the chapter provides a discussion on the acquisition of data from web logs which is divided into three tasks, namely Web Logs Filtration, Crawler Logs Elimination and Sessionisation. This will give an account of the nature of each task and elaborates on the methods employed to carry them out. In order to establish our discussion, a detailed description of the web logs which will be used consistently through this and later chapters for evaluation is provided.

3.1Wayposts Handling

A complex website design leads to an increased number of links that the users need to click before finding the specific document they are looking for, thus making website navigation an extremely difficult task.As mentioned in the previous chapter, linking also commonly known as shortcutting is a form of organisation adaptation which is used to transform a website's design. Shortcutting aims to reduce the number of clicks required in a user's navigational path to reach a target document, assisting the user in navigating through a website and thereby also improving the website's organization. Target documents are desired documents of interest containing information that satisfy the users' current needs. A path can be thought as a user's journey from an initial document to a target document in a website. The sequence of documents visited along this journey indicate navigational route taken by a user to reach intended target document.

Longer paths tend to pose a higher risk for users to stray off-course from their intended route, often resulting in them missing their target document, since each document contains a large number of links from which a user needs to select one that will bring him closer to his target document. This is similar to travelling in a trunk road which has several intersections and each of these intersections leads to another set of intersections and so on to different locations. Choosing the right link at each intersection (i.e., documents users need to navigate) allows a user to reach his target document quickly. Alternatively, providing a shortcut to the target document greatly reduces the number of intersections in a path and ensures the users could reach their target document with fewer clicks.

Existing studies tend to minimize the number of clicks required to find a target document by providing a shortcut between the initial and target document in a path. This approach assumes the sequence of intermediate documents appearing in the path is insignificant to a user's information need and bypasses them. However, we believe that these documents may contain crucial information corresponding to a user's need that leads him to the target document. This work aims to explore this possibility and presents a way to reduce the number of clicks required to find a target document by identifying way posts, which can act as navigational shortcuts based on frequently travelled users' paths. Way posts are intermediate documents in a path that could act as a significant guide for users to find the target document. It is similar to landmarks on roads, which inform drivers that they are travelling on the correct route to their intended destination. Likewise, way posts act as signs in a website reassuring users that they are progressing toward their intended target document, giving them a sense of satisfaction throughout the browsing session.

4. Logs Handling

The web logs used for evaluations throughout this study is of the CLF type. These web logs are obtained from the in-house web server of the Robert Gordon University's School of Computing. The web server records users' activities on the School's website for every month over a period of six months interval. At every interval of the sixth month the old web logs are purged and replaced with newer ones. Our data is sampled from the month of October in the second half interval of the year 2004. The reason being that it is right after the start of the academic year 2003/2004 and probably has the most number of accesses made to the website by both external and internal users compared to any other month in that interval. External users refer to any client that are not directly associated with the University, while internal users are those attached to the institution such as the students and staff.

4.1Data Acquisition from Web Logs

In this section, this work describes the initial work carried out with respect to our approach. It focuses on data acquisition where raw web logs are converted into a format that is meaningful and easier for knowledge extraction. Data acquisition is made up of the following three tasks: Web Logs Filtration, Crawler Logs Elimination and Sessionisation. Web logs filtration involves identifying unwanted document accesses from the raw logs thereby leaving only document accesses that could be used to infer users' browsing behaviors. The crawler logs elimination task shares the same goal with the former task. It determines the document accesses that are made by automated programs called bots or crawlers and removes them, leaving only those document accesses that are associated to human users. Finally, the sessionisation task involves identifying and isolating the series of documents that was viewed by a user during his visit to the website.

Various off-the-shelf applications such as are readily available to mine web logs. However, we decided to develop one to better understand and appreciate the difficulties involved in the mining process. Furthermore, developing such an application provides a wider range of flexibility and customization to better suit the purpose of this study as opposed to commercially available applications. As such, we have implemented an application called Log Parser that analyses and extracts users' browsing behavior patterns from a collection of web logs. This tool provides a general framework to process raw web logs appearing in common log format and presents the data in a form ready for the application of web usage mining techniques for knowledge discovery.

4.2Web Log Filtration

Given the sheer volume and diversity of document requests in the web logs, it is important to distinguish between requests that carry vital information with regards to a user's browsing behavior and those that do not. The web log filtration task aims to do just that by filtering out requests that are inadequate for usage mining from the web logs. For the purpose of this study, ~~we~~ decided to filter out all requests of the following kinds: request for graphical contents, requests for non-HTML files, unsuccessful requests and requests made by bots or crawlers.

Most requests for graphical contents and non-HTML files usually are part of an actual request made to the server to retrieve a HTML document. For example, the entirety of a HTML document could be made of text combined with graphical contents and other non-HTML files at times. Since these graphical contents and other non-HTML files need to be retrieved from the web server to properly display the document, they appear as individual requests in the web logs. However, the request that could be used to elicit information about a user is the one that retrieves the main document rather than its supporting files. Hence the need arises to eliminate such requests related to supporting role from the web logs.

Unwanted graphical and non-HTML requests are identified by referring to the retrieved file's extension in a log's request field. A log's request field records the name of the file that was retrieved and the method that was used to get ~~it1~~. For example, a HTML document request usually has a filename that ends with a ".html", ".shtml", ".asp" or ".php" extension, while graphical contents and non-HTML file requests end with extensions such as ".jpg", ".bmp", ".gz", etc. Based on this observation, a list was created to remove requests related to graphical contents and non-HTML files. Table 3.1 shows the list of file extensions used by the Log Parser application to remove unwanted graphical and non- HTML requests from the web logs.

Consequently, requests made by bots or crawlers cannot be detected using straight forward methods mentioned above. The task requires a more elaborate measure to identify and eliminate such requests. The following section will describe the difference between a user's request and of a crawler before going on to discuss the method employed to eliminate crawler requests.

4.3Crawler Logs Elimination

Crawler, also referred as but, is an automated program that browses the Web in a systematic manner. It is mostly employed by search engines to create a copy of all the visited documents, which is then indexed in order to provide fast searches. A web server does not distinguish between requests made by a crawler or a genuine user as all requests for a document are considered coming from an individual user. As such, the web logs are infused with repeated automated requests that are not made by a human user. If this data is used in the mining process it may lead to biased outcomes and therefore requests belonging to crawlers need to be eliminated from the web logs.

Eliminating requests made by crawlers are more difficult as they need to be explicitly identified. The DNS/IP Check and the Robots Exclusion Check may be the simplest form of crawler detection methods available. The DNS/IP check method involves detecting a crawler's requests by matching the originating IP address from the remote host field of a log with those of known crawlers. For this, a text file containing 199 distinct known crawler DNS/IP addresses was created. There are public databases like the Web Robots website3 that maintains a database of known crawler DNS/IP addresses which, could be used to create the text file. By using this file the Log Parser application is able to cross reference it with the originating IP address of each request and remove those associated to crawlers. However, note that the publicly available crawlers list is not exhaustive and it is becoming extremely difficult to keep up with the continuous evolution of the crawlers.

Sessionisation is the final task in the process of acquiring data from web logs. The sessionisation task involves in identifying and reconstructing the navigational route or routes taken by a user during his visit to the website. In this study, a session can be described as a group of documents requested by a user while browsing, which makes up his navigational route through the website. A user's visit to a website may consist of one or more number of sessions and the same user could also make multiple numbers of visits in a day. In order to simplify the process of identifying a user's session, we decided to consider all visits made by the user in a single day as one individual visit.

5. Conclusions

This project has investigated the idea of adaptive websites and presented an approach to adaptive websites using way posts, which serve as navigational shortcuts that could be used to improve a website's organisation.

1. Identify a group of target documents based on users' access patterns and extract navigational paths terminating at those documents from users' sessions. We have highlighted the challenges associated to finding a set of target documents from a website and indicated the importance of identifying such target documents correctly which, is critical for extracting navigational paths from a collection of users' sessions. In accordance to this, we presented the End Document method which identifies a set of potential target documents based on users' access patterns. The method uses the frequency with which a document appears as an end document in users' sessions and the number of its incoming visits to determine the candidacy of document as a target document. We have evaluated this method and shown that it is reasonably effective in finding target documents. To this extent, we have identified a set of viable target documents using this method from the data acquired of the RGU School of Computing web logs and extracted navigational paths leading to these target documents from users' sessions.

2. Group similar navigational paths together and find hyperlinks (i.e., wayposts) that can act as potential shortcuts points by identifying documents that commonly co occur in these paths. One drawback of existing shortcutting methods is that they bypass the intermediate documents of a user's path and provide a shortcut between the initial and target document of the path. We have presented our approach which believes that these intermediate documents may contain information corresponding to a user's need that leads him to the target document. With regards to this, we have presented a way of identifying way posts (i.e., intermediate documents) which can act as navigational shortcuts to target documents from a collection of users' paths. The method relies on grouping similar users' paths together and finding documents that co-occur in them. Documents which frequently co-occur in multiple users' paths which are similar indicate user's interest in them, thereby making them a suitable candidate as way posts. To this extent, we have implemented a k-means algorithm and used it to generate clusters of similar paths from the collection of users' paths found

in the earlier objective. Then, for each cluster generated we determined the documents which co-occur in all path members of the cluster and identified them as wayposts which, can act as potential navigational shortcuts. We have shown that the clusters generated by the algorithm are reasonably good from which potential wayposts could be identified for shortcutting.

3. Evaluating our approach by measuring the usefulness of the shortcuts generated from way posts. Since the objectives of our work do not include the incorporation of the suggested shortcuts to a website, it is not possible to gauge the actual usefulness of these shortcuts. Instead, we have presented a qualitative evaluation of some of the shortcuts generated and discussed scenarios in which they would be useful. Based on this, we have shown that the shortcuts generated using our approach could shorten users' navigational paths to their target documents and reduce the possible loss of information that may be contained within the intermediate documents of the users' paths.

Bibliography

1. Agarwal, C. C. On leveraging user access patterns for topic specific crawling, Data Mining and Knowledge Discovery 9(2): 123–145. (2004).
2. Altingovde, I. S. and Ulusoy, Exploiting interclass rules for focussed crawling, IEEE Intelligent System 19(6): 66–73. . (2004).
3. Ansari, S., Kohavi, R., Mason, L. and Zheng, Z. Integrating e-commerce and data mining: Architecture and challenges, Proceedings of the 2001 IEEE International Conference on Data Mining, IEEE Computer Society, pp. 27–34. (2001). BBC Web enjoys year of biggest growth, Online. Accessed on 05/2008. *http://news.bbc.co.uk/2/hi/technology/4325918.stm (2005).
4. Berendt, B., Mobasher, B., Nakagawa, M. and Spiliopoulou, MThe impact of site structure and user environment on session reconstruction in web usage analysis, Proceedings of the 4th WebKDD 2002 Workshop, at the ACMSIGKDD Conference on Knowledge Discovery in Databases, pp. 115–129. . (2002).
5. Brickell, J., Dhillon, I. S. and Modha, D. S. Adaptive website design using caching algorithm, Advances in Web Mining and Web Usage Analysis, Vol. 4811/2007, Springer, pp. 1–20. (2007).
6. Catledge, L. D. and Pitkow, J. E Characterizing browsing strategies in the World Wide Web, Computer Networks and ISDN Systems 27(6): 1065–1073. . (1995).
7. Chakrabarti, S., Dom, B. and van den Berg, M. Focussed crawling: A new approach to topic specific web resource discovery, Proceedings of the 8th World Wide Web Conference, Elsevier, Toronto, pp. 545–562. (1999).
8. Cooley, R., Mobasher, B. and Srivastava, J. Data preparation for mining World Wide Web browsing patterns, Knowledge and Information System 1(1): 5–32. (1999).
9. Cooley, R., Tan, P.-N. And Srivastava, J. Discovery of interesting usage patterns from web data, Web Usage Analysis and User Profiling, Vol. 1836/2000, Springer, San Diego, pp. 163–182. (2000).
10. Crescenzi, V., Merialdo, P. and Missier, P. Clustering Web pages based on their structures, Data and Knowledge Engineering 54: 279–299. (2005).