ISSN:0975-9646

Study on Improving User Navigation by Reorganizing web structure based on Link Mining

Deepshree A.vadeyar[#1], Yogish H.K[*2]

[#]Student of M.Tech, department of CSE
EWIT,Bangalore-560091
[*]Associate Professor
EWIT,Bangalore-560091

***Abstract*— Website design is easy task but, to navigate user efficiently is big challenge, one of the reason is user behaviour is keep changing and web developer or designer not think according to user's behaviour, so to improve user navigability by reorganizing website can be done by web transformation. We proposed web transformation by using link mining and we used categorization of link like 1 or 0 for present and absence, for clustering of links we considered each page as node and link as edges later clustering performed on weblog like user sessions and number clicks on page**.

***Keywords*— k-mean, links, weblogs, website design.**

I. INTRODUCTION

There are lakhs of user for website since it is large source of information, web site also contain many links and pages every user require different pages at same time or same user may access different pages at different time. As user increases over www we need to make web intelligent we concern here about intelligent website. To make web site intelligent we must know what is content of website, which are users and how website structured all this known as web mining.

Just as data mining aims at discovering valuable information that is hidden in conventional databases, the emerging field of Web mining aims at finding and extracting relevant information that is hidden in Web-related data. Web mining is a multi-disciplinary effort that draws techniques from fields like information retrieval, statistics, machine learning, natural language processing, and others. Web mining deals with three main areas: web content mining, web usage mining and web structure mining.

Web content mining, also known as text mining, is generally the second step in Web data mining. Content mining is the scanning and mining of text, pictures and graphs of a Web page to determine the relevance of the content to the search query. The main uses for this type of data mining are to gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information.

Web usage mining is extracting information according to user navigation and behavior pattern like time spent on pages, traversing path and number of click on page, the user access patterns called as web logs or profiles. Web usage mining is very important for the organization, since by using logs organization can perform prediction like: requirement of user, pages need to add, pages are useless, user interest etc.

Web structure mining can be defined as mining of links between pages, which is also called as hyperlinks which enable user to access web sites in form of URL and navigate user. In web structure mining developer uses the data from web usage and change structure of web site, pages which is most visited and user spent more time is linked to the start page.

Motivation for choosing web structure mining is: since web site is big source of information, but users mostly browsing useless page which irritates user and user lost interest from searching data over website. A primary cause of poor website design is that the web developers' understanding of how a website should be structured can be considerably different from those of the users; however, the measure of website effectiveness should be the satisfaction of the users rather than that of the developers. Thus, Web pages should be organized in a way that generally matches the user's model of how pages should be organized.

There are two ways to improve user navigability web personalization and web transformation.web personalization deals with user behavior and user profile, sessions and history of data also called as web logs which is created by user's activity on web site, but transformations approaches mainly focuses on developing methods to completely reorganize the link structure of a website. We adopted web transformation technique to facilitate user navigation.

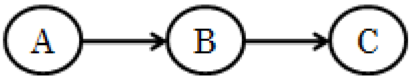

Fig. 1(a). Normal website structure.

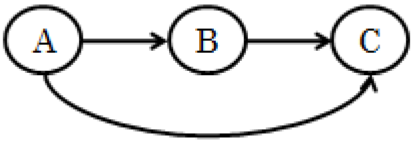

Fig. 1(b). Reorganize website structure.

In fig-1(a) it shows website is like graph and nodes A,B,C are pages and links between pages are edges through which user can navigate, if we found from weblogs that Page 'C' is access very frequently website is reorganized and new link is created from first page so user can access page 'C' in less clicks and time shown in Fig-1(b).

Reorganization of pages depends on parameters like: First consideration is, in –links and out-links of web pages. Second is the traversing path of user and finally user access pattern. All technique can be used with data mining techniques to facilitate user navigation, since web is huge and user always wants response as fast as possible so our goal is to reorganization website by changing its structure so user should reach target in less clicks and in less time.

Proposed strategies are links based clustering to change structure of website as we know clustering is one way of data reduction so lager web usage data is use as compact and time of reorganizing web site structure is reduced. In later section we provide details of some popular clustering algorithm and comparison between clustering algorithm with respect to time

II. RELATED WORK.

This paper is about survey of web structure mining and clustering techniques over web pages and hyperlinks, as structure mining is useful for organization if done according to user need, so to facilitate user we considered structure mining by performing data mining techniques on weblogs also known as part of web usage mining.

About web usage mining, author in [1] explains about weblogs like who accessed order of page request, total time for page view. This paper includes several pre-processing like; **1: Data cleaning**-It is method of removing irrelevant items or logs like removing of file with .gif and .jpg extensions. **2: User identification**-It involves USER ID for each user to provide uniqueness even different users are on same IP.

3: Session identification- This is defines according to time i.e. time between page request and page close or time out. **4: Path completion-** It is defined as if some information or page is important and mostly accessed but not recorded in logs and not linked cause problem .**5: Formatting**- It is method of converting transactions or logs it to a format of data mining like removal of numeric value for determining association rules.

In[2]Author focus on requirements and issues of web structure mining and what are the parameters and data mining techniques can be applied, Author proposed here k-means Clustering algorithm and Apriori association rule mining algorithm, they also uses probabilistic clustering algorithm known as conceptual clustering algorithm like COBWEB .They also introduced the agglomerative and hierarchical clustering algorithm this all used to solve index page problem in Adaptive websites, they considered two parameters sequence of page views and links clicked during each sessions author used here two quality measures based on number of times user get correct page and efforts done by user.

Author here define problem based on contents, hyperlinks and title, algorithm uses by author based on frequent occurrence of pages in user logs.

We focused on structure mining and many work done on structure mining as in [3]Author proposed way of finding browsing efficiency ,here they first performed architecture of web site as graph of pages as nodes and links between pages as edges then on basis on logs, proxy server information and user cookies efficiency is calculated.

Author in [4] proposed 0-1 programming model for reorganizing websites based on cohesion between web pages. They uses two approaches grouping of similar session and grouping of pages with co-occurrence frequency on which they performed clustering and association rule mining for pages involved in the session.

They also used constraint as length of shortest path from home page to each page.

In[5]Author proposed structure mining based on number of links traversed in a session, here rather than directly changing structure they added more links between web pages which are more frequently browsed.

In[6] Author proposed reorganization by classification techniques based on type of file extension, number of links page, ratio of session on last page to the total session on web site and average time for which user on websites or user is login.

In [7] Author proposed some more parameter for website transformation and here author uses sessions which divided in to mini sessions and user traversing path, author also uses two threshold path threshold-length of pages from start page to target page in mini session and out-degree threshold-number of links allowed from pages at the time of reorganization

In this paper our second focus on data mining technique, survey includes many approaches on weblogs as in [8] author gives k-means clustering algorithm on co-citation of pages, here they considered common links shared between pages and similarity measures they consider is cosine similarity measures.

In [9] Author proposed weighted page rank algorithm based on rank assign on page which is most popular according to user behaviour, later on this k-means is performed.

They also focus on parameters like in-links, out-links on every page.

In[10]author suggested clustering on frequent item-sets and their frequent item-sets are user session and access patterns

In [11] Author proposed clustering and association rule mining separately. For association rule mining Apriori algorithm is used based on pre-processed logs and for clustering co-occurrence of pages is used.

Author in [12] proposed clustering based on page views by user and they proposed complete linkage clustering algorithm based on user transaction.

A. Clustering on web logs

Clustering or cluster analysis is defined as technique of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics etc. There are many clustering techniques, this paper contain comparison of three clustering technique. We will start with most basic k-means clustering algorithm, The term "k-means" was first used by James MacQueen in 1967 this algorithm form k-cluster for n-objects according to nearest mean. K-means initialize the cluster means by randomly generating k points in the data space. This is typically done by generating a value uniformly at random within the range for each dimension. Each iteration of K-means consists of two steps: i) cluster assignment, and ii) centroid update. One more imp aspect of k-mean is the sum of squared errors scoring function. Another algorithm we chosen is Connectivity based clustering, also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance, here set of objects represent as hierarchical tree, in this technique, a set of N items to be clustered according to similarity measures of single linkage(shortest distance),complete linkage (larger distance) and average linkage using median and mode, here we compared only single linkage. Basically categorized as Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy and Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In general, the merges and splits are determined in a greedy manner. The results of hierarchical clustering are usually presented in a dendrogram. Both this approach works on Distance matrix created by Euclid distance. One of most robust clustering algorithm towards noise is DBSCAN (Density Based Spatial Clustering of Application with noise, proposed by Martin Ester, Hans-Peter Kriegel, Jorge Sander and Xiaowei Xu in 1996,it is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. it is most basic clustering algorithm which provides cluster of arbitrary shaped. This algorithm depends on two parameter epsilon and minimum points(Min-pts) which define minimum number of objects or points can be placed in clusters. There are some other clustering algorithms like: Expectation Maximization (EM) given in 1977 by Arthur Dempster, in this approach objects are assigned to cluster according objects has a probability to belongs to cluster.

Other algorithm is farthest first here it takes approximately same time as k-means but it forms group of object based on largest distance so objects consider in cluster are those objects which is far from centroid so here according to our parameter we will not get proper objects in cluster.

Related work shows that website is considered as graph with pages as nodes and links between pages as edges but we not focus on graphical clustering since here we get similarity on vertices or weights on edges but here we considered parameters which is not depends on edges and not act like weight.

There is one more algorithm which is considered as clustering algorithm but based on supervised method called as SOM(Self Organizing map),it gives best out comes as it will choose best among all but training time is more.

Last important factor in clustering is similarity measures it is used to define similarity between objects, roughly we can say it is used to assign objects in cluster and there are many similarity measures like Cosine, Euclid and Manhattan.

Cosine similarity is defined as smallest angle between two vectors and if angle is 90 degree we called objects or vector has no similarity, equation for object a and b we can write cosine similarity equation as.

$$\cos\theta = \frac{a^{T} \quad b}{|a| \quad |b|}$$

Euclid and Manhattan define as distance between two objects and it is calculated by below equation where and b are two objects.

$$d(a,b) = \sqrt{\sum_{i=1}^{n}\left(a-b\right)}^{2}$$

We can perform other similarity measures like simple matching and Jaccard coefficient which gives similarity between binary objects, there are some other similarity measures like Pearson Correlation coefficient for variance, and here our requirement is full fill by distance based measures since we used parameters session and number of clicks on links.

At last we used weka tool to perform clustering comparison on web usage data set collected from Depaul University, as this paper is about the study of web structure which is depend on web usage data, below table-1 shows comparison between the three major clustering algorithm according to time required to form model or cluster, here we chosen, number of cluster=10 for k-means and hierarchical, for DBSCAN we set two parameters MinpPts=50 and epsilon=2.

TABLE I. CLUSTER EVALUATION OF REAL DATA SET

Algorithm	K-means	Hierarchical	DBSCAN
Time in sec	0.55	42.38	2.23

III. CONCLUSION

Website reorganizes facilitate user to improve navigability, this paper surveys the broad areas of web site reorganization and link analysis on the basis of web logs and user session and data mining techniques applied on web data, which enables user to reach target in fewer clicks. This survey is beneficial for web developer to understand different aspect of website, for researcher to improve more in website and for commercial organization. Website reorganization is imp aspects as now days; it is vast source of information. From clustering comparison we can conclude that time taken to build model is less in k-means so k-means in fastest than any other algorithm and as web structure mining needs to decrease waiting time it is good to use.

ACKNOWLEDGMENT

We would like to express our gratitude to all those who gave their valuable times to guide and solving doubts. We also want to thanks Computer science and Engineering department of East West Institute of Technology, Bangalore for giving me permission to continue my research on this topic.

REFERENCES

[1] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava" Data Preparation for Mining World Wide Web Browsing Patterns".

[2] Mike Perkowitz and Oren Etzioni "Towards adaptive Web sites: Conceptual framework and case study" Artificial Intelligence 118 (2000) 245–275, 1999.

[3] Joy Shalom Sona, Asha Ambhaikar "Reconciling the Website Structure to Improve the Web Navigation Efficiency" June 2012.

[4] C.C. Lin, "Optimal Web Site Reorganization Considering Information Overload and Search Depth," European J. Operational Research.

[5] R. Gupta, A. Bagchi, and S. Sarkar, "Improving Linkage of Web Pages," INFORMS J. Computing, vol. 19, no. 1, pp. 127-136, 2007.

[6] Y. Fu, M.Y. Shih, M. Creado, and C. Ju, "Reorganizing Web Sites Based on User Access Patterns," Intelligent Systems in Accounting, Finance and Management.

[7] Min Chen and young U. Ryu "Facilitating Effective User Navigation through Website Structure Improvement" IEEE KDD vol no. 25. 2013.

[8] Yitong wang and Masaru Kitsuregawa "Link Based Clustering of Web Search Results" Institute of Industrial Science, The University of Tokyo.

[9] Amar Singh,navjot Kaur,"To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm"IJARCSE August 2013.

[10] Bamshad Mobasher,Robert Cooley, Jaideep Srivastava "Creating Adaptive Web Sites Through Usage-Based Clustering of URLs".

[11] Bamshad Mobasher,Robert Cooley, Jaideep Srivastava"Automatic Personalization Based on Web Usage Mining".

[12] Bamshad Mobasher, Honghua Dai, Tao Luo, Miki Nakagawa" Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization".