International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Survey on k-Means: An Approach for Improved Web Site Structure

Pramod B. Dhamdhere[1], Saba Siraj[#2]

[1]IOKCOE, Pune, Maharashtra, India

[2]Assistant Professor IOKCOE, Pune, Maharashtra, India

Abstract: ***Development of websites to facilitate effective user navigation is the challenging task observed these days. Because the way web developers think and design the system is quite different from that of the user. Different methods have been projected to re-link WebPages in order to recover navigability using user direction-finding data. The fully reorganized emerging structure can be highly impulsive, and the cost of disorienting users after the changes remains unanalyzed. The proposed system presents architecture to cluster the usage statistics of all the users to re-link WebPages. The re-ordering or reforming will mostly be based on clusters generated. Hence an optimal selection of clusters is significant step in implementation of the system. Hence system uses an enhanced K means clustering algorithm where in the number of clusters (optimal) can be routinely designed and clusters are generated consequently. The system also develops a arithmetical programming model to recover the user navigation on a website. The system is imagined the deliver the functionality of a test bench website for data collection and then reorder it based on statistics collected to present the effectiveness of our model.***

Keywords: K-means Clustering, User navigation, Relinking.

1. Introduction

Nowadays there has been increasing investments in website design but it is still exposed, however, that finding required information in a website is quite difficult and designing effective websites is cumbersome task. Palmer indicated that poor website design has been a key element in a number of high profile site letdowns. McKinney et al. also discover that users having difficulty in pinpointing the targets are probably to leave a website even if its information is of good quality.

Earlier studies on website has concentrated on a diversity of issues, such as understanding web structures, locating related pages of a given page, mining useful structure of a news website, and removing template from web pages. This work is related to the literature that observes how to recover website navigability through the use of user navigation data. Different works have made an effort to address this question and they can be usually categorized into two types: to help a particular user by animatedly reconstructing pages based on his contour and traversal paths, often denoted as personalization, and to adapt the site structure to simplify the navigation for all users, often stated as transformation.

A principal cause of poor website design is that the web developers' understanding of how a website should be organized can be considerably diverse from those of the users. Such variances result in cases where users cannot certainly trace the preferred information in a website. This problem is hard to escape because when forming a website, web developers don't have a perfect understanding of users' likings and can only form pages based on their own verdicts. However, the degree of website effectiveness should be the approval of the users rather than that of the developers. Thus, WebPages should be structured in a way that generally matches the user's model of how pages should be organized.

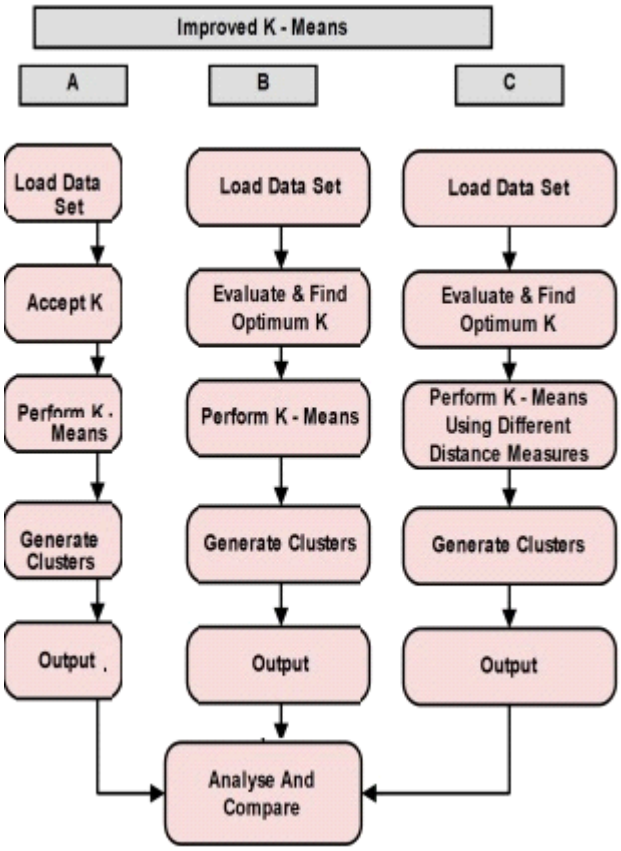

Figure.1. Improved K-Means Algorithm

Distinguishing the shortcomings of website reorganization tactics, proposed system addresses the question of how to recover the organization of a website rather than reorganize it substantially. Specifically, we develop a mathematical programming (MP) model that simplifies user navigation on a website with slight changes to its present structure. Our model is mostly suitable for informational websites whose matters are static and quite stable over time. Examples of informational websites are universities, hospitals, tourist

attractions, federal agencies, and sports organizations. Our model, however, may not be appropriate for websites that purely use dynamic pages or have volatile contents. This is because a steady state might never be reached in user access patterns in such websites, so it may not be possible to use the weblog data to improve the site structure. The relevancy of web page can be attained by considering the amount of in-links and out-links existing in a particular web page. When the web page has more number of out-links to a pertinent page, then that page can be treated as a central page. From this central page, all remaining web pages are compared for similarity and the most similar pages are grouped together. The grouping of most similar pages together is known as clustering. Clustering can be done based on different algorithms such as hierarchical, k-means, partitioning, etc. The simplest unsupervised learning algorithm that solve clustering problem is K- Means algorithm. It is a simple and easy way to classify a given data set through a certain number of clusters.

An important kind of search engine that provide results based on hypertext links between sites can be termed as Link Based search engine. Rather than providing results based on keywords or the preferences of human editors, sites are ranked based on the quality and quantity of other web sites linked to them. In this system, user submits a query to the meta-search engine. The meta-search engine searches for the relevant results of user's query. From the set of results retrieved from web search engine, they are formed as a meta-directory tree. This tree structure helps the user to retrieve information with high relevancy. The number of outward links in a page, i.e., the out degree, is an important factor in modeling web structure. Prior studies typically model it as hard constraints so that pages in the new structure cannot have more links than a specified out-degree threshold, because having too many links in a page can cause information overload to users and is considered undesirable. For instance, Lin uses 6, 8, and 10 as the out-degree threshold in experiments. This modeling approach, however, enforces severe restrictions on the new structure, as it prohibits pages from having more links than a specified threshold, even if adding these links may greatly facilitate user navigation. Our model formulates the out-degree as a cost term in the objective function to penalize pages that have more links than the threshold, so a page's out-degree may exceed the threshold if the cost of adding such links can be justified.

2. Related Work

Web personalization is the process of "tailoring" web pages to the needs of specific users using the information of the users' navigational behavior and profile data [8]. Perkowitz and Etzioni [3] describe an approach that automatically synthesizes index pages which contain links to pages pertaining to particular topics based on the co-occurrence frequency of pages in user traversals, to facilitate user navigation.

Web transformation, on the other hand, involves changing the structure of a website to facilitate the navigation for a large set of users [9] instead of personalizing pages for individual users. Fu et al. [5] describe an approach to reorganize web pages so as to provide users with their desired information in fewer clicks. However, this approach considers only local structures in a website rather than the site as a whole, so the new structure may not be necessarily optimal. Gupta et al. [7] propose a heuristic method based on simulated annealing to relink web pages to improve navigability. This method makes use of the aggregate user preference data and can be used to improve the link structure in websites for both wired and wireless devices. However, this approach does not yield optimal solutions and takes relatively a long time (10 to 15 hours) to run even for a small website.

Document clustering [3] algorithm is more efficient in performing the clustering by considering each document as initial centroid and then merges those documents into a cluster by considering the relevancy of contents, until all documents in a cluster have similar feature. The most common document clustering techniques are of two types such as: Agglomerative Hierarchical clustering and K-Means clustering. The K-means algorithm is considered as one of the most commonly used algorithms for classification of numeric data in data mining [6]. Lot of researches and studies are going on to address two of the major limitations of K-means algorithm- One to select efficiently the initial centroids and second to remove the need of giving the number of clusters required as input to the algorithm.

3. Existing System

In the example shown in Fig. 2, the user has traversed three paths before reaching the target. An intuitive solution to help this user reach the target faster is to introduce more links. There are many ways to add extra links. If a link is added from D to K, the user can directly reach K via D, and hence reach the target in the first path. Thus, adding this link "saves" the user two paths. Similarly, establishing a link from B to K enables the user to reach the target in the second path. Hence, this saves him one path. We could also insert a link from E to K, and this is considered the same as linking B to K. This is because both B and E are pages visited in the second path, so linking either one to K saves only one path. Yet, another possibility is to link C to F, a non target page. In this case, we assume that the user does not follow the new link, because it does not directly connect a page to the target. While many links can be added to improve navigability, our objective is to achieve the specified goal for user navigation with minimal changes to a website. We measure the changes by the number of new links added to the current site structure. There are several reasons that we should insert minimal links. First, minimizing changes to the current structure can avoid disorienting familiar users. Second, adding unnecessary links can lead to pages having too many links, which increases users' cognitive loads and makes it difficult for them to read and comprehend. Third, since our model improves site structures on a regular basis, the number of new links should be kept at minimum such that the links in a website in the whole course of maintenance do not expand in a chaotic manner.

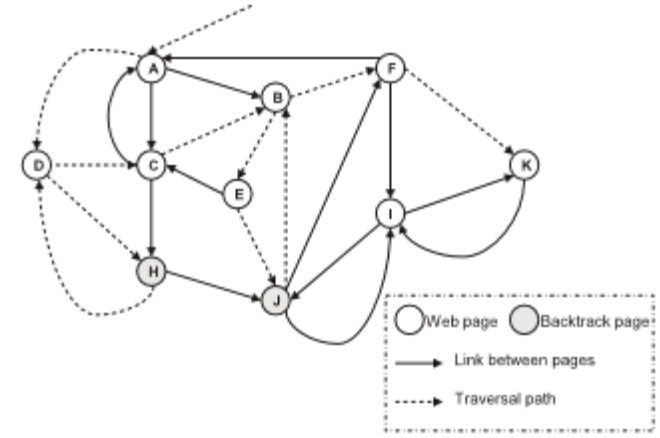

Fig3 . Example of a mini session.

Figure 2: Checking Navigation for Website structure

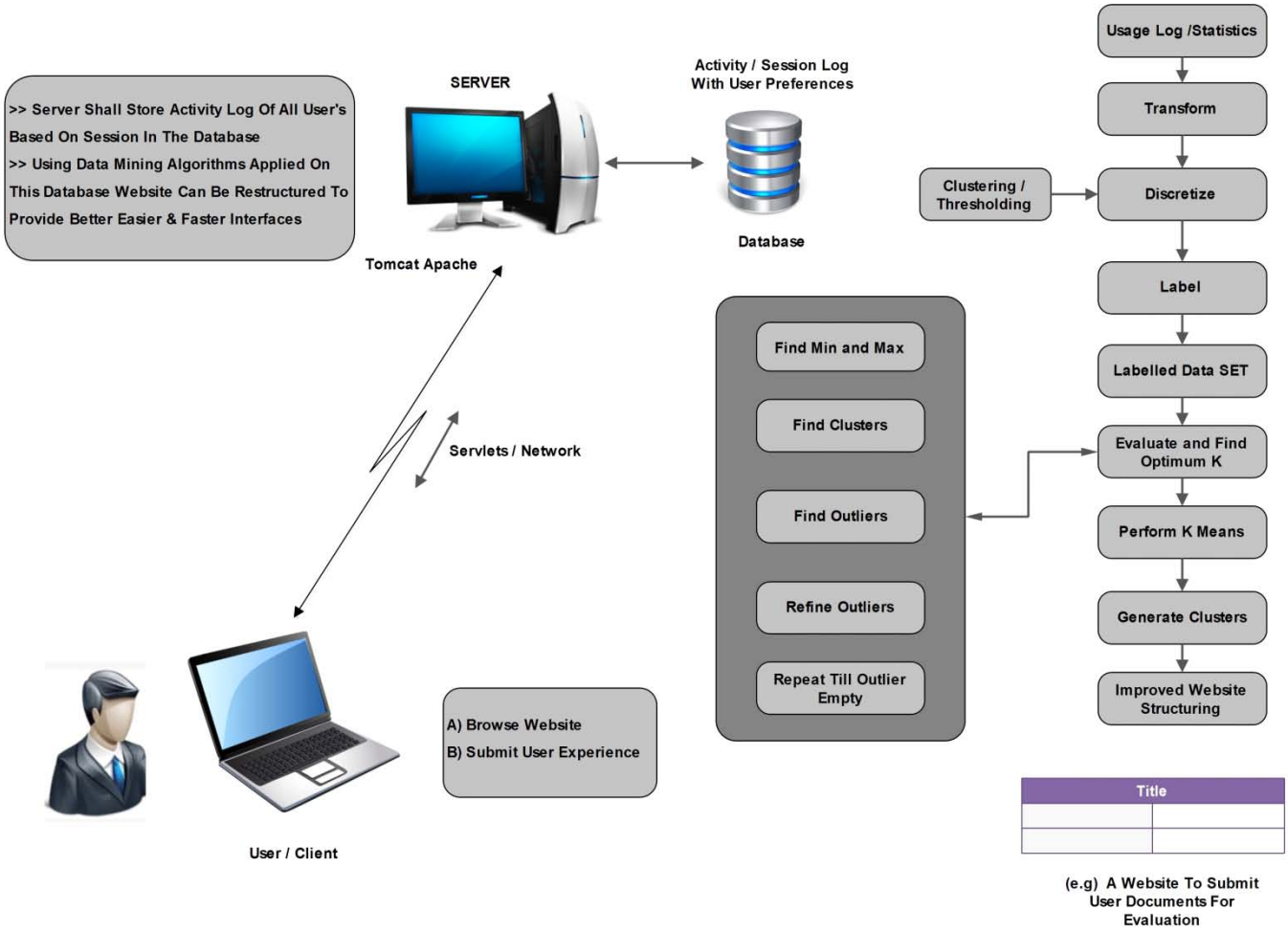

Figure 3: I-clustering using K means for improving Website Structure

The diagram shown above shows the proposed system. The system consists of end users/client, a server for storing the data.

- USER: The end user is the actual user who searches for the relevant data on the web using browser installed in his system.
- Web Server: It tracks the search request made by the user. Also server maintains the activity session log with the user preferences and stores it in the database. Using Improved k means clustering algorithm it improves the navigation of the website so that it provides better, easier and fast interface.

4. Proposed System

When User want Surfing on Internet that time user did not get actual information he want He/she has to spend a lot of time on that particular web site .In this Paper we propose a new Data mining algorithm k means i.e. improved k-means algorithm. This Improved K-means Work on at database of web server .this algorithm take input as session log with preferences And then transform these input into the number of clusters. The cluster is depends on the no of input so the total no link and the relinking of that all pervious links of particular website. With the help of relinking and linking we find the priority of that particular link and these link come on the very front page of web site. This way we can reduce time complexity over web.

5. Conclusion

The Proposed system is efficient system for navigating website structure. Using the data mining algorithm i.e. Improved K means algorithm it helps to restructure the link of website. It provides the easier, fastest and better interface to retrieve information from the website. The major benefit of this system is that it performs the process of retrieving

information in minimum span of time so that we can say it is time efficient.

References

[1] J. Palmer, "Web Site Usability, Design, and Performance Metrics," Information Systems Research, vol. 13, no. 2, pp. 151-167, 2002.
[2] T. Nakayama, H. Kato, and Y. Yamane, "Discovering the Gap between Web Site Designers' Expectations and Users' Behavior," Computer Networks, vol. 33, pp. 811-822, 2000.
[3] M. Perkowitz and O. Etzioni, "Towards Adaptive Web Sites: Conceptual Framework and Case Study," Artificial Intelligence, vol. 118, pp. 245-275, 2000.
[4] J. Hou and Y. Zhang, "Effectively Finding Relevant Web Pages from Linkage Information," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 4, pp. 940-951, July/Aug. 2003.
[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] M. Eirinaki and M. Vazirgiannis, "Web Mining for Web Personalization," ACM Trans. Internet Technology, vol. 3, no. 1, pp. 1-27, 2003
[7] C.C. Lin and L. Tseng, "Website Reorganization Using an Ant Colony System," Expert Systems with Applications, vol. 37, no. 12, pp. 7598-7605, 2010.
[8] 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, vol. 11, no. 1, pp. 39-53, 2002.
[9] "Facilitating Effective User Navigation through Website Structure Improvement", Min Chen and Young U. Ryu , Knowledge and Data Engineering, Vol. 25, No. 3, March 2013