K-means clustering algorithm to improve website performance

Shubham Shandilya¹, Gaurav Kumar² and Sunil Kumar³
¹²³Department of Information Technology, IMT College of Engineering, Greater Noida

Abstract- Website navigation has become one of the most important design features across many domains, including finance, ecommerce, entertainment, education, government, and medical. A main reason is that the web developers understanding of how a website should be structured can be significantly different from that of the users. Various methods have been proposed to relink Web Pages to improve navigability using user navigation data. The proposed k-means clustering algorithm reorganizes accessing methodology that can avoid unpredictability in terms of reloading time. The proposed method takes less time when compared to the existing algorithms.

I. Introduction

There have also been significant changes in the way people use and access the web, and this has become different how sites are designed. Users understanding the content of a website often depends on users understanding how the website works. Navigation through a large web site for finding relevant information can be tedious and frustrating. Previous studies on website has focused on a variety of issues, such as understanding web structures [1], finding relevant pages of a given page [2], mining informative structure of a news website [3] and extracting template from web pages [4]. Our work, on the other hand is closely related to the literature that examines how to improve website navigability through the use of user navigation data. Various methods have made an effort to address this question and they can be generally classified into two categories [5]

1. Personalization, &
2. transformation.

To facilitate reconstituting pages is known as personalization. In other hand modifying the web site structure and navigate easy to all users is known as transformation. Now in this paper we are going to use transformation approach, this method mainly focus on reconstucting links structure of the web site. Since completely reconstructing a web site may lose some data’s, locations, and highly unpredictable, unanalysed. finally web site reconstruction can totally can change the structure of the web site, they cannot be frequently performed to improve the navigability. We address the question of how to improve the structure of a website rather than reorganize it substantially. Specifically, we develop a k-means cluster algorithm that facilitates user navigation on a website with minimal changes to its current structure. Our paper is about online shopping web site. where the content all dynamic. It is unstable.

II. Related Work

This method describes how to improve website navigation for the user without changing its current structure. so we are going to provide only minimal changes to the current structure so that the data cannot be changed totally, it remains same and its predictable and analysed.

In [6] 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 [7] 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 [8] 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
Web Personalization:
Web personalization is the process of “tailoring” web pages to the needs of specific users using the information of the users’ navigational behaviour and profile data. Perkowitz and Etzioni describe a method that automatically synthesizes index pages which contain links to pages pertaining to particular topics based on the co-occurrence frequency of pages in path traversal, to facilitate effective user navigation. The methods have been proposed by Mobasher et al. and Yan et al. these authors introduce some clusters from users profiles which is taken from weblogs and then dynamically create links for users who are classified into different methods based on their access patterns.

III. Web transformation
Web transformation, on the other side, involved changes the structure of a website to facilitate the navigation for a large group of users instead of personalizing pages for individual users. Fu et al. explain an approach to restructure web pages so as to provide users with their desired information in fewer clicks. However, this method study only local structures in a website rather than the site as a whole.

Maximal Forward Reference:
We use backtracks to identify the ways that a user has navigated, where a backtrack is defined as a user’s reopen to a previously browsed page. The surmise is that users will backtrack if they do not find the page where they expect it. Thus, a path is defined as a sequence of pages visited by a user without backtracking, a hypothesis that is parallel to the maximal forward reference defined by the Chen et al author. Basically, each backtracking point is the end of a path.

IV. Mini Sessions
Recall that a mini session is relevant only if its length is larger than the corresponding path. Consequently, only relevant mini sessions need to be considered for improvement and this leads to a large number of irrelevant mini sessions (denoted as TI) being eliminated from consideration in our MP model. A log file is an ordered set of web page requests made by users. The requests are stored in the order that the server receives them. If multiple users are browsing the site concurrently, their requests are inter mining in the log file.

V. Search & Advance search
The site opens up door for web users through the Home page. The Home page is structured in such direction that the layout is as user friendly as possible. There is a navigational menu is placed at the top of the page which links to various inner pages. There is a category drop down which is placed on left side for easy manipulation [9]. The centre area has been used for displaying latest products in the chorological order.

Advance Search
It has two search one search is used for to search the data with the help of keyword and the other search is used to get the data with the help of the dynamic data entered by the user. We have used the K mean clustering algorithm to search the data for the user.

Algorithm
K-means clustering algorithm
K-means is one of the simplest unsupervised learning algorithms that solve known clustering problem. The procedure follows a simple and easy way to classify a data set through a certain number of clusters (assume k clusters). The main idea is to introduce k centers, one for each cluster.
These centres all should be placed in a cunning way because they have different locations which causes different result. So, the better option is to place them as much as far away from each other.

\[
J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} ||x_i - v_j||^2
\]

where, ‘\(||x_i - v_j||\)’ is the Euclidean distance between \(x_i\) and \(v_j\).
‘\(c_i\)’ is the no of data points in \(i\)th cluster.
‘\(c\)’ is the no of cluster centres.

Steps for k-means clustering algorithm:
1. Mining Candidate Link Algorithm
2. Input: \(P_i\) – Users Profile data
3. Output: Links that can be used for redesign
4. Steps 1: We identify the usage pattern of users \(\lambda\) from \(P_i = \{P_1, P_2, \ldots, P_m\}\) set for user \(U_i\) to get link \(P_m\)
5. Steps 2: For every access link set obtain the set of candidate links \(\{C_1, C_2, \ldots, C_p\}\)
6. Steps 3: For all users and their all access link set obtain the set of candidate links
7. Steps 4: Obtain the Dice’s similarity coefficient for all candidate link set.
8. Steps 5: Apply KNN classifier
9. Steps 6: Then the links having problem for maximum number of users are selected for redesign the website.

VI. Result

The proposed k-clustering approach reduces execution time and memory requirement. The number of new links accessed was successfully predicted and logged to reduce the time required to fetch the page.

<table>
<thead>
<tr>
<th>Time Threshold</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of new links (mostly accessed)</td>
<td>5.469</td>
<td>1.256</td>
<td>0.372</td>
<td>5.216</td>
<td>0.946</td>
<td>0.382</td>
<td>4.962</td>
<td>0.843</td>
<td>0.376</td>
</tr>
<tr>
<td>No. of new links (few accessed)</td>
<td>4.391</td>
<td>0.902</td>
<td>0.278</td>
<td>3.708</td>
<td>0.837</td>
<td>0.305</td>
<td>2.727</td>
<td>0.703</td>
<td>0.314</td>
</tr>
<tr>
<td>No. of links to be imposed</td>
<td>2.453</td>
<td>0.750</td>
<td>0.467</td>
<td>2.014</td>
<td>0.812</td>
<td>0.481</td>
<td>2.089</td>
<td>0.751</td>
<td>0.468</td>
</tr>
</tbody>
</table>

From the table the results are statistically analysed and plotted in a bar chart as shown in figure1. The graph shows that the pages that are less frequently accessed are less than the pages that are more frequently accessed.
VII. Conclusion

In this paper, we have proposed a k-means algorithm to improve the navigation effectiveness of a website while minimizing changes to its current methodology. Our algorithm is particularly appropriate for online websites whose contents are relatively unstable. It improves a website not by changing but reorganizes its structure and hence is suitable for website maintenance on a progressive basis.

References