Enhanced Web Mining Using Advanced Clustering Algorithm and Pincer Search Algorithm for Pattern Discovery

Reshabh Singh¹, Gaurav Pandey², Mohit Paul³
M.Tech Student¹, Assistant Professor², ³
Department of Computer Science and Information Technology, SHIATS, India

Abstract:
During the last few decades, the digital information of our world grew up beyond limit and that way it need a huge database with light accessibility and availability[4]. So that we build massive collection of information on world wide web with the strength of web mining and web based application which is most popular and common way of gathering information any time anywhere. Web usage mining which is one of the applications of data mining help to provide pre-processing, pattern discovery and pattern analysis from web logs files. Many researchers researched to find out shortest and fastest pattern in web mining. The purpose of this paper is to optimize the effective result of pattern discovery by using pincer search association rule and advanced clustering algorithm of clustering.

Keywords: Web usages mining, Web logs files, Pincer search algorithm, Advanced clustering algorithm.

I. Introduction
1. Web Mining-Web mining is one of the important applications of the data mining techniques which provides access of relevant information from the web[1]. The web mining techniques can be divided into three areas

a) Web content mining- It manages discovery of different type of data file like audio, video, images, pdf etc. from the web. It is based on text context mining.

b) Web structure mining- It is concerned with discovery of the hyperlink structure on the web. It applied intra document structure and inter document structure with in the web logs. The web structure mining uses html and xml to tag web pages with the structure.

c) web usages mining- This is also known as web log mining. Web usage mining analyze to discover frequent pattern from stored web browsing data. It usages web logs, proxy server logs, user profiles, user queries.

![Web Mining Technique](image)

Figure 1. Web Mining Technique

2. Web Log Files- web logs files are those files which have information about website activity. The log files generated itself from the website server. When a user visit website, the information about it appended to current log files. The range of log files between 1KB to 100MB.

   a) Location of web log files in web servers- The web log files have three locations-
      - Web server log
      - Client Browser
      - Web proxy server

   b) Types of web log files:
      - Access log file- Access log file works on the gathering information from every click event, hits, access of the user.
      - Error log files- When any error takes place in the internal server than it aware to the client and user receives error 404.
      - Agent log- It held record about client, client browser, version of browser and operating system which is used.
      - Referrer log file- It stored the information about links which are using by client and the URL clicking by them.

   c) web log file format- it has three types-
      - W3c log file format
      - NCSA common log file format
      - IIS log file format

3. Process of Web Usage Mining-Web usage mining has generally three process from-

   A. Data pre-processing- web log data pre-processing refers to identify users, sessions, page views and change into reliable data form. The data pre-processing has various t steps to complete this phase and they are- cleaning of data, identification of user, path completion, authentication of user data, formatting, identifications of sessions and user identification transaction.
B. Pattern discovery- pattern discovery is one of most important key component of web mining. Pattern discovery is using clustering, association rules, sequential analysis, dependency modelling, classification and statistical analysis. Basically it is extract form of web data by using graphs, charts, tables and forms.

II. Background and Related Works
There are numerous work done in web usage mining to find out accurate and frequent set of data pattern. Shaily G. Langhnoja and MehulP.Barot[4] were working on the pattern discovery by using A Prieri association rule algorithm and DBSAS clustering algorithm. But these algorithms are much failure rate on huge data base so here I propose upper and better algorithm for using patterns discovery. Amarpreet Kaur Toor and Amarpreet Singh[6] was working on the ACA and they conclude result that the approach of ACA is much efficient than other traditional approach. Daol Tony Lin[10] was done lot of work to find the frequent sets which help of different approaches. ThabetSlimani[14] was proposed pattern discovery in the web usage mining.

III. Methods and Materials
Already preprocessed data is taken
Clustering- It is technique of grouping similar data items in different cluster. Data clustering helps to calculate the statistic, machine learning and data communities with diverse emphases.

Clustering algorithm- there are different approach in partitioning clustering like PAM, CLARA and CLARANS. In the hierarchical clustering, there are three algorithm cure, BRICH, DBSCAN. In the modern approach of clustering there are various algorithm and they are ACA, SOM, HAC, K-Mean algorithm.

• Advanced algorithm- In this paper we have used ACA for pattern discovery. The ACA approach says that set of two simple data sets to retain the labels of cluster and the all distance of all the data objects to the nearest cluster during each iteration that can be used in next iteration[6]. That way is calculated the distance between the current data object and the new data cluster center. if the computed distance is smaller than or equal to the distance to the old center, the data object stays in its cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k-1 clustering centers, saving the accessing time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers and find the nearest cluster center. It assigns this point to the nearest cluster center and then separately record the distance to its center. Because in each iteration some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

• Process of ACA- An ACA follows the process and it is-

Input- the number of required cluster k, Dataset S.
D= [d1, d2, . . . , dn] containing n data objects. di= { x1, x2, . . . , xm }/ Set of attributes of one data point
Output: A set of K clusters.

1. Draw multiple sub-samples {SI, S2, . . . , Sj} from the original dataset.
2. Repeat step 3 for m=1 to n.
3. Apply combined approach for sub sample. 4. In each set, take the middle point as the initial centroid.
5. Compute the distance between each data point to all the initial centroids.
6. For each data point find the closest centroid and assign to nearest cluster.
7. Choose minimum of minimum distance from cluster center criteria.
8. Now apply new calculation again on dataset S for K clusters.
9. Combine two nearest clusters into one cluster.
10. Recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

Time complexity- the time complexity of ACA is O(nk) which is faster than any other clustering algorithms. If the data points occurs in the same cluster than the time complexity is O(1), otherwise it is O(k). When the half data moves in the current cluster than the time complexity is O(nk/2).

• Comparison of other algorithm with ACA-

<table>
<thead>
<tr>
<th>Parameter</th>
<th>SOM</th>
<th>k-means</th>
<th>HAC</th>
<th>ACA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error rate</td>
<td>0.8189</td>
<td>0.8456</td>
<td>0.8379</td>
<td>0.3672</td>
</tr>
<tr>
<td>Execution time</td>
<td>279</td>
<td>1281</td>
<td>1341</td>
<td>1000</td>
</tr>
<tr>
<td>Accessing time</td>
<td>fast</td>
<td>slow</td>
<td>slow</td>
<td>Very fast</td>
</tr>
<tr>
<td>Number of clusters</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>4</td>
</tr>
</tbody>
</table>

• Advantages of ACA-
1) ACA does not require re-iteration.
2) ACA can find multiple link effect.
3) ACA has a notion of voice.
4) ACA requires just one parameter and is mostly ordering the next iteration.

Association rule mining- the main task of association rule mining is find out discovering frequent item sets, frequent sequential patterns, frequent sequential rules and frequent association rules. It is prominent tool for knowledge discovery in web mining. Association rule mining (ARM) finds out complete relationship between frequent sets and how they merge with it and how they perform operations. There are various types of association rules- apriori, pincer search, f-p growth tree, border algorithm, incremental algorithm.

• Pincer Search algorithm- A bidirectional search have both approach of searching which are top-down and bottom-up approach. The pincer search algorithm works on this principle. It calculates to find the frequent item set in bottom-up approach. In the database it counts candidate maximal item set itself to see if any actual frequent.
The pincer search approach says that-
1. L0 := ∅; k := 1; Cj := {(i)} iεI
2. MFCS := {{1,2,...,n}}; MFS := ∅
3. while Ci≠∅
4. read database and count supports for \( C_k \) and \( MFCS \)
5. remove frequent itemsets from \( MFCS \) and add them to \( MFS \)
6. determine frequent set \( L_k \) and infrequent set \( S_k \)
7. use \( S_k \) to update \( MFCS \)
8. generate new candidate set \( C_{k+1} \) (join, recover, and prune)
9. \( k := k + 1 \)
10. return \( MFS \)

**Approach used for Pattern Discovery** - the using approach for web usage mining shown in the figure below. This follows three steps - (Proposed approach)
1. In the starting it is collecting web logs files and pre-processing operations takes place. After this operation web log files are stored in the database.
2. Now we are applying data mining operations like association rule mining and clustering for pattern discovery in the taken database. We are using combine operation of association rule and clustering. We are using Pincer search for association rule and Advanced clustering (ACA) for clustering.
3. In the final step association rule mining operation will be used to find client access patterns from this clustered group of database. That way the minimal frequent sets are find and it was much accurate than other approach and the execution time of the pattern discovery is much faster than the traditional approach.

**IV. Implementation and Results**
Web using mining is using web log file of website for building implementation. Here we are using a database of company is considered as input dataset. Data pre-processing and pattern discovery steps are discussed already. For applying the pincer search and ACA on the database we are using java platform and my sql server.

For pre-processing we are using data from the taken database and the snap of the pre-processing given here

**Figure 2. Pattern Discovery Approach**

**Figure 3. Selecting data files for cleansing**
The calss of item set is given here under on which association rule mining pincer search applies

**Figure 4. Data item sets for ARM**
the minimal frequent set after applying pincer search algorithm we find results which are given under snaps

**Figure 5. MFCF after applying Pincer search**
After applying association rule mining Pincer search algorithm on the clustered data we get result as shown below

**Figure 6. Comparison of clustered using in data base**
The following graph shows number of clustered and its capability under the database.

![Graph showing clustered data and capability](image)

**Figure 7. Clustering approach in database**

**V. CONCLUSIONS**

Web usage mining techniques are most required area of research nowdays. The clients of websites, how they are search websites is the goal of web usage mining. In this proposed approach, the goal is fulfilled by using pincer search association rule mining on the clustered data which is using advanced clustering algorithm (ACA) that will be applied clustering technique first and then we are using association rule techniques. Association rule mining may have some unfaithful result shows of irrelevant rules, so many rules follows accurate prediction of result. Clustered data access quick patterns reduce data set for pincer search association rule mining and improve result accuracy and producing most frequent sets of patterns discovery of web usages mining process impressive and much effective.

**VI. REFERENCES**


[5]. Amanpreet Kaur Toor and Amarpreet Singh, An Advanced Clustering Algorithm (ACA) for Clustering Large Data Set to Achieve High Dimensionality, Toor and Singh, J ComputSciSystBiol 2014, 7:4 http://dx.doi.org/10.4172/jcsb.1000146

[6]. Tan, Steinbach, Kumar, Data Mining Association Analysis: Basic Concepts and Algorithms, july 2015.


[9]. Feng Chen, Pan Deng, Jiafu Wan, Daqiang Zhang, Athanasios V. Vasilakos, S. and Xiaohui Ron, Data Mining for the Internet of Things: Literature Review and Challenges, Volume 2015, Article ID 431047.


[12]. Bamshad Mobasher, Chapter 12: Web Usage Mining.


[16]. Thabet Slimani, Efficient Analysis of Pattern and Association Rule Mining Approaches.

[17]. Nivedita Pandey, Prof. Love Verma, A comprehensive method for discovering the maximal frequent set.