Clickstream Pattern Mining with Data Mining

Divya Patel
PG Student
Kalol Institute of Technology & Research Center Kalol, Gujarat, India

Abstract— A system is one of the best web usage mining Application which reduces the difficulties faced by the users to meet their requirements. Web usage mining is the process of obtaining useful knowledge from the server logs. In this work, we can predict the user's navigational behavior and user next activities. We are using the bi-clustering approach with greedy search to overcome the problems of grouping available in traditional clustering approach. The practical implementation of this algorithm shows the more accurate prediction of user's inspiration on web. An attempt has been made through this paper to provide a holistic view as to what clickstream data analysis is, how mining techniques are applied on such data to generate useful information and what kind of applications exploit it to get useful information. That is how the log mining can be apply in now a days. The knowledge gained by the analysis is applied to target marketing and in the designing of web portals.

Key words: Web Usage Mining, Recommender System, Bi-Clustering, Web Log, Greedy Search

I. Introduction

Data mining is the extraction of hidden predictive information from large database. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining is actually part of the knowledge discovery process. The Knowledge Discovery process comprises of a few steps leading from raw data collections to some form of new knowledge. Classification is a form of data analysis that can be used to extract model describing important data classes or to predict future data trends. Now a days many types of classification algorithms are available like K nearest neghbor, Apriori, Baysian, data mining, Support vectore machine, K-means. So the people who are not expert in characteristics of data and their distribution, do not know which data classification method should be used to obtain good classification results for their given dataset. For this reason, choosing suitable classifier for given dataset is an important task.

Data mining: Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems[2]. It is an essential process where intelligent methods are applied to extract data patterns. So the overall goal of the data mining process is to extract information from a data sets and transform it into an understandable structure for further use.

II. Literature survey

hossam.faris et al. [1] User behavior identify based on user's historical records. Using clickstream data mining we can identify user's navigational behavior. KNN algorithm Used to find the pattern to enhance business growth. Identify if a future interaction can be associated with a certain user. Related work of that is mainly focused on authenticating and profiling the users derived into groups. Proposed Approach is learning the behavior of former user; therefore a set of previous interactions between user and website are required. In Future they wants to Data collected from server side logs. Plan to implement a data-collector browser plug-in. Utilize verification tools to report theft ID and Unusual user behavior. Chetan Kaushal et al. [2] Comparative Study of Recent Sequential Pattern Mining Algorithms on web clickstream Data. Many algorithm used inside and compare to each other working process, Result have been assess & compared to find an algo which has better working on real-life datasets. Sequential pattern algorithm used. Related work of that Sequential Pattern mining algorithm used. Compared the performance of algorithm running time & maximum memory usage (MB), used to web graph structure. Proposed approach is Algorithm apply on three real-life dataset. Algorithm implemented in java Eclipse & compared them in same running time. Maximum memory used, when algorithm is running. Sequential Pattern mining. Compared the performance of algorithm running time & maximum memory usage (MB) used to web graph structure. In Future they wants to Algorithm evaluated on three real-life dataset and present the result. Sequence pattern mining Algorithm take lease amount of memory on real-life dataset. More Constraints & pruning techniques can be included in this methods. Saloni Aggarwal et al.[3] To provide a holistic view as to what clickstream data analysis is, how mining techniques are applied on such data to generate useful information and what kind of applications exploit it to get useful information. Related work is various log mining models used to analyze the clickstream data. Similarity based on frequency & similar data of different users put in sequence. Using related rule mining to generate frequent item set. Mostly used in e-commerce web page mining. Outcome is highlights the different techniques applied to perform clickstream data analysis and what types of information can be retrieved using each method. Also it throws light on the types of mining methods employed in data mining on general basis so as to build a foundation for understanding web mining as a concept. Zhang et al.[4] This paper contains three models to summarize this kind of data streams, which are the batch model, the Evolving Objects (EO) model and the Dynamic Data Stream (DDS) model. Through creating, updating and deleting user profiles, these models summarize the behaviors of each user as a profile object. Based upon these models, clustering algorithms are employed to discover interesting user groups from the profile objects. We have evaluated all the proposed models on a large real-world data set, showing that the DDS model summarizes the data streams with evolving tuples more efficiently and effectively, and provides better basis for clustering users than the other two models. Experimental results showed that the DDS model out performs the batch and EO models in both clustering quality and efficiency and Towards the classification and frequent pattern mining over bi-Clustering models in future. Tany

Bhattacharya et al.[5] It prevent the user from wasting time in separating what use need encourage useful Exploring. To create protocols for Ontology which defines types & efficiency of the recommender system. Various types of recommender system present. Web personalization is an upcoming concept which is help the users to get whole new experience of exploring & solving their solution. In future they can be then tested and applied on the websites related to acquiring financial and personal details of the user.

III. Methods and Algorithms

A. Bi-clustering

1	"123"	2
4		1.2
2	3	1

Bi-clustering is two way clustering. It's for data analysis on web structure. Web analysis for enhance of business. We can predict the use next activities. To finding subgroups of rows & columns which are as similar as possible to each other. Dissimilar as possible to rest. It's used for data which is generate from logs. Different bi-clustering algorithm is below.

- Bi-cluster with all value constant
- Bi-cluster with rows value constant
- Bi-cluster with columns value constant
- Bi-cluster with both values constant

B. Preprocessing algorithm

Focuses on one of the most meaningful issues within the famous Knowledge Discovery from Data process. Data will likely have inconsistencies (not steady), errors, out of range values, impossible data combinations, missing values, data is not suitable to start a DM process. Most used Data Pre-processing Algorithm.(1).Noise Filtering (2).Normalization being done after Noise Filtering. Statistics methods used. Like Mean, median, mode. Normalization being done after Noise Filtering. Statistics methods used. Like Mean, median, mode. Data matrix is perform by a Bi-clustering. In our case this data matrix is are derived in two categories. So the rows of a data matrix will be different users which are visited the website pages and the columns will be the pages visited by all users. To generate these data matrix from the clickstream data we need to pre-process the clickstream data. We can generate the user access matrix from clickstream data using following equation.

$$a_{ij} = \begin{cases} \text{Hits}(U_i, P_j), \text{ if } P_j \text{ is visited by } U_i \\ 0\text{ , otherwise} \end{cases}$$

Where Hits(ui , pj) , is the count/ frequency of user ui accesses the pages pj during giving time period.

C. Greedy search

A greedy algorithm repeatedly executes a search procedure which tries to maximize the bi-cluster based on examining local conditions, with the hope that the outcome will lead to a desired outcome for the global problem. ACV and MSR are used as merit function to grow the bi-cluster. With ACV it Insert/Remove the user/pages to/from the bi-cluster if it increases ACV of the bi-cluster. Our objective function is to maximize ACV of a bi-cluster. With MSR it Insert/Remove the user/pages to/from the bi-cluster if it decreases MSR of the bi-cluster. Our objective function is to minimize MSR of a bi-cluster. The greedy approach is easy to implement and mostly time efficient. The main objective for this work. High volume bi clustering with high ACV and low MSR. ACV: The function of f(I, J) is used to extract optimal bi-clustering.

$$F\ (I,\ J) = \begin{cases} |I|*|J|,\ \text{if ACV (bicluster)} \geq \delta \\ 0,\ \text{Otherwise} \end{cases}$$

|I| and |J| is number of coloms and rows.
MSR: The function of f (I, J) is used to extract optimal bi-clustering.

$$F\ (I,\ J) = \begin{cases} |I|*|J|,\ \text{if MSR (bicluster)} \leq \delta \\ 0,\ \text{Otherwise} \end{cases}$$

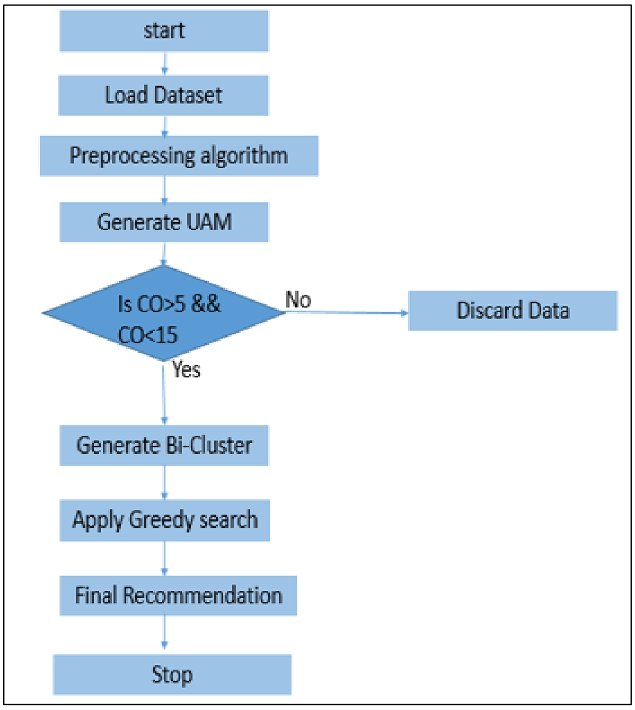
D. K-nearest neighbor algorithm

KNN a method for cluster analysis in data mining. K-means algorithm which is observe the matric and find nearest value in matrices. This algorithm like as a K-nearest neighbor algorithm. K-means to classify new data into the existing clusters. KNN is a data mining used to cluster observations into groups of related observation without any prior knowledge of those relationship.

IV. Proposed System Flow

Here represent the proposed flow architecture below. Proposed algorithm is that:

- Step 1. Evolutionary Bi-clustering Algorithm
- Step 2. Then Load data set in system.
- Step 3. Pre-process data and generate user matrix A.
- Step 4. Generate initial Bi-cluster using two way K-means clustering from user matrix A.
- Step 5. Improve the quality and quantity of the initial bi-clusters using Greedy method.
- Step 6. Search procedure with two bi-cluster evolution function ACV & MSR.
- Step 7. Evaluate the fitness of individuals.
- Step 8. Generate Recommendation for website.
- Step 9. Stop



V. Experimental Results and Analysis

A. *Data Set*

A real dataset is used for this experiment. The data set is taken from the UCI dataset repository (http://kdd.ics.uci.edu/) that consists of Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Pacific Standard Time). Visits are recorded at the level of URL category and are recorded in time order. Each sequence in the dataset corresponds to page views of a user during that twenty-four hour period. Each event in the sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail that is, at the level of URL, but rather, they are recorded at the level of page category (as determined by a site administrator). The categories are "front page", "news", "tech", "local", “opinion", "on-air", "misc", "weather", "health", "living", "business", "sports", "summary", "bbs" (bulletin board service), "travel", "msn-news", and "msn-sports". Any page requests served via a caching mechanism were not recorded in the server logs and, hence, not present in the data. This dataset is slightly changed flattering to our experiment, if the user visit only the “front page” then 1 is recorded on the first position of the matrix and other 16 column (category) are filled by 0 [7].

– The details of the data set are provided in Table I
– Dataset Use in the experiment

Dataset	MSNBC
Total number of users	989818
Average Number of visit per user	5.7
Number of URL per each categories	10-5000

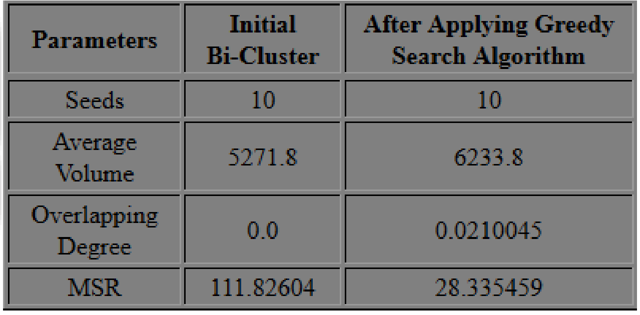
We have shown results of the MSNBC dataset. The user access matrix is generated from the first equation. In the next bi-clustering step Ku User clusters and Kp Page clusters are generated from user access matrix and initial Bi-clusters Ku * Kp are generated. These bi-clusters are enlarged and refined using Greedy search procedure. In this step the volume of bi-clusters is higher than initial bi-clusters. The Enlarged and refined bi-clusters are set as initial population to the Genetic Algorithm. It will generate optimal bi-clusters. The measure R is used to evaluate the overlapping degree between bi-clusters. It calculates the amount of overlapping among bi-clusters. The degree of overlapping of bi-clusters is defined as follows:

$$R = \frac{1}{|U| * |P|} \sum_{i=1}^{|U|} \sum_{j=1}^{|P|} \mathrm{Tv}$$

Where,

$$Tij = \frac{1}{(N-1)} * \left(\sum_{k=1}^{N} W_k(a_{ij}) - 1 \right)$$

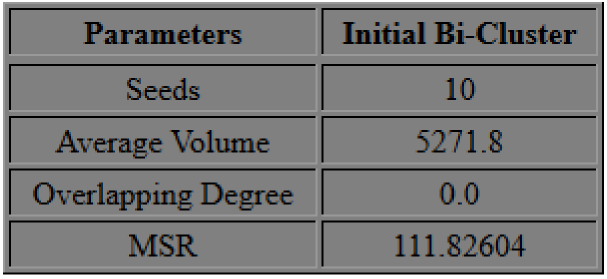
Where, N is the total number of bi-clusters, |U| represents the total number of users, |P| represents the total number of pages in the data matrix A. The value of wk(aij) is either 0 or 1. If the element (point) aij in A is present in the kth bi-cluster, then wk(aij) = 1, otherwise 0. If R index value is higher, then degree of overlapping of the generated bi-clusters would be high. The range of R index is 0 ≤ R ≤ 1. The results generated after each step are shown in the following table:



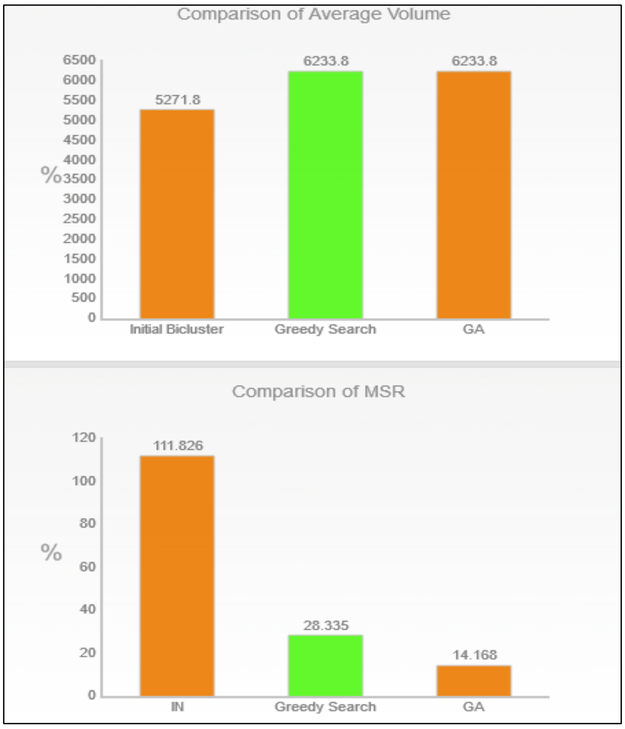
Parameters	Initial Bi-Cluster	After Applying Greedy Search Algorithm
Seeds	10	10
Average Volume	5271.8	6233.8
Overlapping Degree	0.0	0.0210045
MSR	111.82604	28.335459

Table 2: Results of Proposed System

Recommendation quality = ACV*100 So, our proposed system will gives the best quality and accurate results as compared to other recommendation functions like cosine similarity and hamming similarity.



Parameters	Initial Bi-Cluster
Seeds	10
Average Volume	5271.8
Overlapping Degree	0.0
MSR	111.82604



The Average volume of bi-cluster is increasing after each step. As the value of ACV is increasing the value of MSR will be decrease. A high ACV and Low MSR value indicates that the bi-cluster is strongly coherent

VI. Conclusion

As we depicted above that I had completed some Research gaps in some area that we can solve by using my system function.

So this is how we can get better solution in clickstream pattern mining. And further work of clickstream pattern mining will be done in DP 2.

References

[1] E Kohavi, Ron E Provost, Foster R , Book Section Applications of Data Mining to Electronic Commerce , Data Mining and Knowledge Discovery journal 5(1/2), p. 5-10, 2001 RaviBhusan, Dr. RajenderNath, Automatic Recommendation of Web Pages For Online Users Using Web Usage Mining, IEEE, International Conferenceon Computer Science(ICCS),2012,Page(s):371-374,ISBN:978-1-4673-2647-6.

[2] Lu Chen, Qiang Su, Discovering User's Interest at E-Commerce Site Using Clickstream Data, 2013 10th International Conference on Service Systems and Service Management (ICSSSM) IEEE 2013, pp.124-129.

[3] Agrawal, R., Ramakrishnan, S.: Mining sequential patterns. In: Proc. 11th Intern. Conf. Data Engineering, pp. 314. IEEE (1995).

[4] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang. Multidimensional regression analysis of time-series data streams. In VLDB, pages 323334, 2002.

[5] Daz-Avils, V.E. (2005). Semantic Peer-To-Peer recommender Systems (Tesis doctoral),<http://citeseer.ist.psu.edu/semanticpeer-to-peer1.pdf>[Retrieved 23/02/2008].

[6] June 2015 | IJIRT | Volume 2 Issue 1 | ISSN: 2349-6002 IJIRT 102424 INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH IN TECHNOLOGY 184 Web Page Recommendation System using Bi-clustering with Greedy Search and Genetic Algorithm.