Comparative Analysis of Web Usage Data Clustering Using Asymmetric Binary Variables and K-Means

Theint Theint Shwe
University of Computer Studies, Mandalay
theint2shwe@gmail.com

Abstract

World Wide Web overwhelms us with the immense amounts of widely distributed interconnected, rich and dynamic information. As a consequence of this, Web Usage Mining becomes one of the popular research areas. It involves the application of data mining techniques to discover usage patterns from the Web access logs data. Clustering is one of the important functions in Web Usage Mining to group the user access patterns which have the same access behavior. In this paper, we would like to propose a new approach, asymmetric binary variables (one type of Jaccard coefficient) to perform clustering. And then the performance of our proposed approach is compared with k-means clustering. The resulting clusters from these two methods are tested with two internal validation methods: Dunn Index and DB Index (Davies and Bouldin Index). Finally, we point out the strengths and weaknesses of each method. According to the analysis results, the findings of clustering upon these methods can be seen clearly.

1. Introduction

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering [1]. Clustering is often called unsupervised learning, because unlike supervised learning, class values denoting an a priori partition or grouping of the data are not given. The quality of clustering can be accessed based on a measure of dissimilarity of objects, which can be computed for various types of data, including interval-scaled variables, or combinations of these variable types [2].

In this paper, we would like to perform clustering using Web server access logs and compare the performance of two clustering methods. For k-means, we use Euclidean distance measure methods. Especially, we would like to point out the importance of the choice of initial cluster centers and the numbers of clusters that affect the applied clustering algorithm.The random choice of initial seeds changes the cluster members and clusters' quality.

By using asymmetric binary variables for clustering, the processing steps of the clustering can be reduced. Moreover, we don't need to define initial cluster centers and only need to define maximum distance values for organizing clusters. For this method, changing the maximum distance value can get the highest cluster quality.

The purpose of the research is to analyze the final results (clusters) how much they perform precisely from the different clustering methods and which results the highest quality for clusters.

The contribution of the research is to correctly characterize the users' behaviors by using the clustering algorithms. The system is intended to be applicable in Prefetched systems, Personalized Systems, Web Site Maintainers and Web Site Developers. Because of seeing the behavior of the users correctly, they have a chance to fulfill users' satisfaction.

2. Related Work

D. Banumathy et.al proposed a framework to improve the cluster quality from the k-means algorithm as two steps process. Their proposed algorithm is tested in medical domain and showed that refined initial starting points and post processing refinement of clusters indeed lead to improved solutions[3].

D. Qi et.al introduced a Self-Organizing Map (SOM) based approached to mining web log data. They used web log file for October, 2006 from the http://cs.lamar.edu as their test data[4].

C. I. Mary et.al presented Ant Colony Optimization (ACO) to improve clustering. Their important point is to improve the clustering quality after grouping. The quality of clusters was tested using two measures called Entropy and F-measure. They described that their method provides better results than conventional algorithm [5].

The paper [6] described the similarity-based clustering approach to group the communities of users for prefetching the predictive page. They used web logs as their data source.

P.S. Bradley et. al proposed a framework to improve the web sessions' cluster quality from k-means clustering using genetic algorithm (GA)[7].

3. Data Source for Clustering

Our proposed system uses the data source as web logs. The raw data from the web access log file is in the following form:

Table1. Sample web access log raw data

192.168.0.254 - - [17/Jul/2009:14:12:41 +0630] "GET http://www.google.com/images? HTTP/1.0" 302 520 TCP_MISS:DEFAULT_PARENT
192.168.0.254 - - [17/Jul/2009:14:12:41 +0630] "GET http://www.ieee.org/ HTTP/1.0" 302 520 TCP_MISS:DEFAULT_PARENT
192.168.0.252 - - [17/Jul/2009:14:12:41 +0630] "GET http://www.google.com/ HTTP/1.0" 302 520 TCP_MISS:DEFAULT_PARENT
192.168.0.254 - - [17/Jul/2009:14:12:41 +0630] "GET http://images.google.com/images? HTTP/1.0" 302 520 TCP_MISS:DEFAULT_PARENT
192.168.0.251 - - [17/Jul/2009:14:12:41 +0630] "GET http://www.ieee.org/ HTTP/1.0" 302 520 TCP_MISS:DEFAULT_PARENT

Table2 presents the portion of the preprocessed data used in our system.

Table2. Sample preprocessed data

User	Site
192.168.0.254	http://www.google.com
192.168.0.25	http://www.ieee.org
192.168.0.252	http://www.google.com
192.168.0.254	http://www.google.com
192.168.0.21	http://www.ieee.org

Thereafter, the relational table was created by using these data objects, Users and Sites. The assumption was that IP address as User. Sites are also represented with alphabetic letters. If the user accesses the particular web site, assign 1 otherwise 0 at the respective place in the relational table. So, the data from the table2 was transformed into table3 as follow:

Table3. A relational table containing binary variables

Users	Web sites							
	A	B	C	D	E	F	G	H
U1	1	1	1	1	0	0	1	1
U2	1	1	1	1	1	1	0	0
U3	1	1	1	0	0	1	0	1
U4	1	1	1	1	0	0	1	1
U5	1	1	1	1	1	1	0	0
U6	1	0	1	1	1	1	1	1
U7	1	1	1	1	1	1	1	0
U8	1	0	1	1	1	1	1	1

4. Jaccard Coefficient for Clustering

4.1. Asymmetric binary varialbes

A binary variable is asymmetric if the outcomes of the states are not equally important, such as positive and negative outcomes of the disease test. Given two asymmetric binary variables, the agreement of two 1s (a positive match) is considered more important than that of two 0s (a negative match). The similarity based on such variables is called non-invariant similarity and the most well-known coefficient is Jaccard coefficient.

4.2. Creating contingency table

The binary variables can be thought of as having the same weight, and then we have 2 by 2 contingency table as follows:

Table4. A contingency table for binary variables

Object j				
		1	0	sum
Object i	1	q	r	q+r
	0	s	t	s+t
	sum	q+s	r+t	p

For this system, the number of negative match, t (object i doesn't access object j) is unimportant and ignored in the computation. We use the Jaccard coefficient of asymmetric binary variables as follow:

$$d(i,\ j) = \frac{r + s}{q + r + s} \qquad (1)$$

4.3. Processing steps of system using asymmetric binary variables

In the first step, we must create the contingency table for the asymmetric binary variables. Then, we use web server access logs as the primary data source for the system. Secondly, relational table is built upon the preprocessed web access log data. And then, distance values are calculated using asymmetric binary variables, one type of Jaccard coefficient using equation (1).Finally, we must define the maximum distance value to group data into one cluster. In this step, the distance values of users less than the maximum distance value exist in one cluster. Clustering results are then analyzed with two internal validation methods.

5. K-Means for Clustering

The k-means is one of the best known partitional clustering algorithms for Web Usage Mining. It is perhaps also the most widely used among all clustering algorithms due to its simplicity and efficiency. Given a set of data points and the required number of k clusters (k is specified by the user), this algorithm iteratively partitions the data into k-clusters based on some distance function.

5.1. Function of k-means algorithm

The k-means algorithm partitions the given data into k clusters. Each cluster has a cluster center, which is also called the cluster centroid.

At the beginning, the algorithm randomly selects k data points as the seed centroids. It then computes the distance between each seed centroid and every data point. Each data point is assigned to the centroid that is closest to it. A centroid and its data point therefore represent a cluster. Once all data points in the data are assigned, the centroid for each cluster is re-computed using the data points in the current cluster. This process repeats until a stopping criterion is met [8].

5.2. Euclidean Distance Function

$$m_j = \frac{1}{|c_j|} \sum_{x_i \in c_j} x_i , \qquad (3)$$

where $|c_j|$ is the number of data points in cluster $c_{j.}$ The distance from a data point x_i to a cluster mean (centroid) m_j is computed with:

$$dist(x_i, m_j) = \|x_i - m_j\|$$

$$= \sqrt{|x_{i1} - m_{j1}|^2 + |x_{i2} - m_{j2}|^2 + \ldots + |x_{ir} - m_{jr}|^2} \qquad (4)$$

5.3. Processing steps of system using k-means

The system uses the preprocessed web access logs data as input in the first step. And then, k-means algorithm is used to perform clustering. The distance measure used in k-means is Euclidean distance function. Finally, the resulting clusters are tested with two internal validation methods.

6. Internal Validation Methods

6.1. Dunn Index

$$D_{nc} = \min_{i=1,..,nc} \{ \min_{j=i+1,..,nc} \{ \frac{d(c_i, c_j)}{\max_{k=1,..,nc} diam(c_k)} \} \} \quad (5)$$

where d(c_i,c_j) is a dissimilarity function between cluster c_i and c_j defined as

$$d(c_i,c_j) = \min_{x\in c_i, y\in c_j} d(x,y) \quad (6)$$

and diam (C) is the diameter of cluster, which may be considered as a measure of dispersion of the clusters. The diameter of a cluster C can be defined as

$$diam(C) = \max_{x,y\in C} d(x,y) \quad (7)$$

It is clear that if the dataset contains compact and well separated clusters, the distance between clusters is expected to be large and the diameter of the clusters is expected to be small. Thus based on the Dunn Index definition, we may determine that large values of the index indicate the presence of compact and well- separated clusters.

6.2. DB Index

Given that K is the number of clusters, c_i and c_j are the closet clusters according to average distance d and diam is the diameter of a cluster, the DB index is defined as

$$DB = \frac{1}{K} \sum_{i=1}^{K} \max_{j\neq i} [\frac{diam(c_i) + diam(c_j)}{d(c_i, c_j)}] \quad (8)$$

It is also clear that DB index is the average similarity between each cluster and its most similar one. It is desirable for clusters to have the minimum possible similarity to each other; therefore, we seek clustering that minimizes DB.

7. System Evaluation

The data from table3 is used as the input data source of the system. When we use the Jaccard coefficient by changing the different maximum distance value to organize cluster, the evaluation results are as follow:

Table5. Evaluation results using asymmetric binary variables

No of clusters	Maximum distance value	Clusters	Dunn Index	DB Index

5	0	C1=U1,U4,U2=U2,U5
		C3=U3, C4=U6,U8
		C5=U7
			0.856	0

4	<=0.2	C1=U1,U4,C2=U2,U5, U7,U3=U3,C4=U6,U8	0.667	0.237
4	<=0.25	C1=U1,U4,C2=U2,U5, U7, C3=U3,C4=U6,U8	0.379	0.995

3	<=0.35	C1=U1,U4,U6,U7,U8,
		C2=U2,U5,U6,U7,U8,
		C3=U3
			0	can't define

According to the evaluation results of table5, we can see that maximum distance value 0 can give the best cluster quality. But, it may create empty cluster if the identical usage pattern doesn't exist among users.

On the other hand, the defined maximum distance value to form clusters become large, the probability of overlapping clusters may be higher. So, we can generally draw the conclusion that we can adjust the cluster quality by changing the different maximum distance value.

7.1. Advantages of asymmetric binary variables

- Simplest method to perform clustering
- requires few processing steps to get compact clusters
- can get good quality clusters by modifying different maximum acceptable distance value
- easy to implement with less complexity

7.2. Disadvantages of asymmetric binary variables

- can't perform clustering without the identification of maximum distance value

The evaluation results of the systems using k-means is as shown in table6.

Table6. Evaluation results using k-means

No of clusters	Initial Centroids	Clusters	Dunn Index	DB Index
4	U1,U3, U5,U7	C1=U1,U4, C2=U3, C3= U5,U2, C4=U7,U6,U8	0.5	0.328
4	U5,U6, U7,U8	C1=U5,U2, C2=U6,U3,C3=U7,U4, U1,C4= U8,U4,U1	0.4	1.498
4	U2,U4, U6,U8	C1=U2,U3,U5,U7,C2= U4,U1, 3=U6,U4=U8	0	0.885

3	U1,U3, U7	C1=U1,U4, C2=U3,
		C3=U7,U2,U5,U6,U8
			0	1.452

In accordance with the analysis results of table6, even if the same number of clusters, k, we choose, the quality of the cluster is not the same. The quality of the cluster also depends on the random initial centroids. So, we can find out that k-means clustering can give local optimum and initial seeds selection is so important. Moreover, cluster quality can be declined if the value of k is not large enough.

7.3. Advantages of k-means

- Simple and easy to calculate

7.4. Distadvantages of k-means

- High time complexity for large data sets
- Can get only local optimum
- Increase time complexity for the number of distance calculation with the increase of the dimensionality of data

8. Comparison of two methods

- Asymmetric binary variables don't need to define the number of clusters and random initial center points but k-means must be defined this by user.
- Asymmetric binary variables can reduce calculation time than k-means because it doesn't need to compute a center point iteratively until the stopping criteria is met.
- By using asymmetric binary variables, we can generally draw a conclusion that maximum acceptable distance value approach to zero can give the higher quality clusters. But we can't say definitely like that in k-means.
- Asymmetric binary variables only depend on the identification of acceptable maximum distance value but k-means depend on the number of clusters and centroids.
- Asymmetric binary variables only need to adjust the maximum acceptable distance value to improve the quality of clusters but k-means requires one type of optimization algorithms to reduce mis-clustering and to improve quality of clusters.
- Asymmetric binary variables can occur overlapping clusters only when maximum acceptable distance value is too large and k-means may occur when the number of k is not large enough and the choice of initial seeds.

8.1. Computational Complexity of two methods

The computational complexity of k-means algorithm is O (nkt), where n is totaling number of objects, k is the number of clusters and t is the number of iterations. Normally k<=n and t<=n. The method often terminates local optimum.

The time complexity of asymmetric binary variables is the sum of distance calculation time plus identification of maximum acceptable distance value to perform clustering. The time complexity for distance calculation is

$$t(n) = \sum_{i=1}^{n-1} n - i \quad = O\left(\frac{n^2}{2}\right)$$

The time complexity of grouping data using maximum acceptable distance value takes constant time. When we ignore the constant, the time complexity for clustering using asymmetric binary variables is O(n^2).

9. Conclusion

The purpose of the paper is to perform the clustering with the highest quality as much as we can. And the desired property is that we would like to perform clustering with fewer processing steps and intended to reduce time complexity. According to our analysis results, asymmetric binary variables are more appropriate than k-means for quality of clusters. In the future, we will try to fulfill the remaining weaknesses of our proposed method.

References

[1] J. Han, M. Kamber, Data Mining Concepts and techniques, ISBN 1- 55860489-8.

[2] B. Liu, Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, ISBN-10 3-540-37881 -2 Springer Berlin Heidelberg, New York.

[3] D. Banumathy, S.Mithyakalyani, An Improved K-Means Clustering with Ant-Colony Optimization ,1st International Conference on Intelligent Electrical System, (NCIES'09), 24-25 April 2009, Maha College of Engineering , Salam, Inida.

[4] D. Qi, C. C. Li. Self-Organizing Map based Web Pages Clustering using Web Logs, Computer Science Department, Lamar University and School of Information Technology, Illinois State University.

[5] C. I. Mary, S. V. K. Raja, Refinement of Clusters from K-Means with Ant Colony Optimization ,Journal of Theoretical and Applied Information Technology, 2005-2009, JATIT.

[6] Y. Y. Win, C. N. Win, K. H. S. Hla, Web user Clustering for Predictive Prefetching, 4th International Conference on Computer Applications, Yangon, Myanmar, February 23-24, 2006.

[7] P.S. Bradley, U. M. Fayyad, Refining Initial Points for K-Means Clustering, 15th International Conference on Machine Learning, (ICML98) ,pp.91-99,Morgan Ksugmann, San Franciso.

[8] S. Chakrabarti, Mining the Web, Discovering Knowledge from Hypertext Data, ISBN-13:978-1-55860-754-5, Indian Institute of Technology, Bombay.