**IOSR Journal of Engineering (IOSRJEN)** **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 54-56

# Discovery of Web Usage Pattern using Latent Semantic Indexing Algorithm

V. Sangeetha
*Assistant Professor Department of Information Technology, Sankara College of science and commerce, Coimbatore, Coimbatore District*

***Abstract:*** *Web clustering is one of the familiar used techniques in the context of Web mining, which is to aggregate Web objects, such as Web pages or users session, into a number of object groups by measuring the mutual vector distance. Clustering can be performed upon these two types of Web objects, which results in clustering Web users or Web pages. The resulting Web user session groups are considered as representatives of user navigational behavior patterns, while Web page clusters are used for generating task-oriented functionality aggregations of Web organizations. The mined usage knowledge in terms of Web usage patterns and page aggregates can be utilized to improve Web site structure designs*
***Keywords:*** *Web clustering, web mining, vector distance, web objects.*

## I. Introduction

Web user transactions in various usage groups use standard clustering algorithms, such as *k*-means clustering algorithm.

An algorithm called Page Gather was proposed by Perkowith and Etzioni to discover significant page segments, which were used to help Web designers to add an additional index page that do not exist before to facilitate Web users to locate their interested contents , by using a Clique clustering algorithm. In the context of clustering, computational costs is a major concerned issue suffering researchers due to the particular characteristics of Web data, For example, the problem of the high dimension and the sparsity nature of Web data. It is difficult, to apply a standard clustering algorithm on the Web usage data with millions of user sessions to derive a collection of Web pages, which results in a tough computational task. The reason to use pages as dimensions, the user sessions must be treated as dimensions and clustering is performed on this very high-dimensional space. To address the issues, dimensionality reduction techniques and alternative clustering algorithms are explored. Amongst these, *Latent Semantic Analysis* is considered as an efficient dimensionality reduction algorithm with the latent semantic analysis capability, that is, the capability of discovering the hidden knowledge from Web data by taking the semantic property of data into consideration.

**Latent Semantic Indexing Algorithm**

The LSI algorithm and its related mathematical background, especially the knowledge of linear algebra in terms of Singular Value Decomposition operation, which forms the foundation of LSI algorithm. Upon the transformed semantic space, we propose a novel similarity function to measure the distance between two user sessions, which would be used in Web clustering.

**Web Usage Data Model**

The Web usage data is collected and stored in Web server logs of websites, and is preprocessed for data analysis after performing data- cleaning, page- identification, and user -identification for constructing the co-occurrence observation. Here we use refined usage data instead of the raw data. We first review the usage data model and introduce the concept of the session page matrix for Web usage mining. , the whole user session data can be formed as a Web usage data matrix represented by a session-page matrix $SP_{m\times n} = \{a_{ij}\}$

The entry value in the session-page matrix, $a_{ij}$ is usually determined by the number of hits or the amount time spent by specific user on the corresponding page. Generally, in order to eliminate the influence caused by the relative amount difference of visiting time duration or hit number, a normalization manipulation across page space in the same user session is performed. Once the usage matrix is constructed, we may applying conventional clustering algorithms on the user session data to classify user sessions into various groups, within which the classified sessions share the similar access interest. It is intuitive to perform clustering algorithms directly on each row vector of the usage matrix to determine the relative "close" session cluster by using a similar ity-based measure, such as the commonly adopted cosine similarity from Information Retrieval. In this work, we propose an algorithm, named *Latent Usage Information* to group user sessions semantically by taking

the latent semantic information into account. For better understanding LUI algorithm, we first discuss some theoretical backgrounds of the SVD algorithm.

**Singular Value Decomposition Algorithm**

The SVD Algorithm is a method used to decompose a matrix into three other matrices.

$A=USV^T$

Where A is an M X N Matrix, U is an M X N orthogonal Matrix, S is an N X N diagonal Matrix, V is an N XN orthogonal matrix. Real-matrix

$A = a_{ij}{}_{m \times n}$ , without loss of generality.

**User Session in Latent Semantic space**

User sessions are obtained with matrix $U_K$, $\sum_K$ and $V_K$. and map into the K dimensional latent semantic space. Given session $s_i$, it is represented as a coordinate vector in the pages is determined as $s_i = \{a_{i1}, a_{i2}, \ldots, a_{in}\}$ .The coordinate vector $s_i$ in the $k$ dimensional latent semantic subspace is re-parameterized as

$s_i' = s_i V_k \sum_k = (t_{i1}, t_{i2}, \ldots, t_{ik})$

**Latent Usage Information Algorithm**

*The* latent Usage Information (LUI) for clustering Web sessions and generating user profiles based on the discovered clusters. This algorithm consists of two steps, the first step is a clustering algorithm, which is to cluster the converted latent usage data into a number of session groups, the next step is about generating a set of user profiles, which are derived from calculating the centroids of the discovered session clusters.

**Clustering user session Algorithm**

The K clustering algorithm, named MK-means clustering,

To partition the usage data matrix with latent k dimension.

A converted usage matrix SP and a similarity threshold ε

A set of user session clusters SCL =SCLi and corresponding centroids Cid =Cidi .

Choose the first user session $s_i'$ as the initial cluster $SCL_1$ and the centroid of this cluster $SCL_1 = \{s_i'\}$ and $Cid_1 = s_i'$.

For each session $s_i'$ , calculate the similarity between $s_i'$ and the centroids of other existing clusters $sim\{s_i', Cid_j\}$ and the steps should be repeated till all the session are processed.

**Building an User Profile**

Each user session is represented as a weight-based page vector. It is reasonable to derive the centroid of the cluster obtained by clustering algorithm as a user profile. we compute the mean vector to represent an centroid. each session cluster $SCL_i \in SCL$ , the mean page vector of all sessions in the cluster (i.e. centroid), is determined by the ratio of the sum of page weights in $SCL_i$ to the number of sessions in the cluster. In order to eliminate the impact of difference in visiting time or click numbers of each session, the weights are normalized while calculating the centroid of cluster. That is, the maximum weight in the constructed user profile is tuned to be 1, whereas other page weights are divided by the maximum weight.

**Experimental Results with Data Sets**

The data set is a commonly-used data source provided to test and compare knowledge discovery methods for the data mining purpose. Data pre-processing is needed to perform on the raw data set since there are some short user sessions existing in the dataset, which mean they are of less contribution for data mining. Support filtering technique is used to eliminate these user sessions, leaving the only sessions with at least four pages. After data preparation, we have setup a data set including 9308 user sessions and 69 pages, where each session consists of 11.88 pages in average.

## II. Conclusion

The LSI-based approach, named LUI, for grouping Web transactions and generating user profiles. The relationships among the co-occurrence observations into a usage data model in the form of a session-page matrix. Then, a dimensionality reduction algorithm based on the SVD algorithm has been employed on the usage matrix to capture the latent usage information for partitioning user sessions. Based on the decomposed latent usage information, we propose a *k*-means clustering algorithm to generate user session clusters. The

experimental results have shown that the proposed approach is capable of effectively discovering user access patterns and revealing the underlying relationships among user visiting records.

## References

[1]. Zhang, Y., J.X. Yu, and J. Hou, *Web Communities: Analysis and Construction*. 2006, Berlin Heidelberg: Springer.

[2]. Ghani, R. and A. Fano. *Building Recommender Systems Using a Knowledge Base of Product Semantics.* in *Proceedings of the Workshop on Recommendation and Personalization in E-Commerce, at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web Based Systems (AH2002).* 2002, p. 11-19, Malaga, Spain.

[3]. Chakrabarti, S., et al. *The Structure of Broad Topics on the Web*. in *Proceeding of 11th International World Wide Web Conference*. 2002, p. 251 - 262, Honolulu, Hawaii, USA.

[4]. Büchner, A.G. and M.D. Mulvenna, *Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining.* SIGMOD Record, 1998. 27(4): p. 54-61.

[5]. Chang, G., et al., eds. *Mining the World Wide Web: An Information Search Approach*. The Information Retrieval. Vol. 10. 2001, KAP.