Repossession of recurrent web document with an advent of clustering algorithms in the Semantic Web mining

Mamta Sharma **Vijay Rana**

Arni University **S.B.B.S University**

Abstract

Everyday an immense sum of web documents, reports, e-mails, and web pages are generated from diverse sources, such as enterprises, governments, organizations, and individuals. This kind of unstructured data is generally not stored on relational or transaction database systems, but on web servers, files servers, or even personal workstations. Enormous enterprises often spend plenty of manpower on organizing these web documents into a logical formation for later use. They follow a systematic and automatic process in organizing these web documents without human intervention or grounding work. This paper takes on the challenge of developing an accurate, efficient, and scalable method for clustering web documents into a structure that facilitates browsing.

Keywords: web document; clustering; web mining; itemset; clusters;

1.1 INTRODUCTION

Data mining, also termed as Knowledge Discovery in Databases (KDD), is an explanation of data explosion. Merely stated, data mining is a mode of extracting appealing information, such as regulations, patterns, regularities, or constraints, from data in bulky databases [2]. The extracted knowledge should be non-trivial, previously unknown, implicit, and potentially useful in that it may serve as an important input for making decisions.

The key functionalities of data mining are relative mining, classification and prediction, and cluster analysis. These techniques apply to different types of data to solve different problems. Some applications of data mining are target marketing, customer relation management, market basket analysis, cross selling, market segmentation, forecasting, quality control, fraud detection, and intelligent query answering [4].

Web document clustering has been studied intensively because of its wide applicability in areas such as web mining, information retrieval [1], and topological analysis. Another catalyst for developing an effective web document clustering algorithm is the huge amount of unstructured data on the Internet. The majority of this information is in text format, for example, emails, news, web pages, reports, etc. Organizing them into a logical structure is a challenging task. More recently, clustering is employed for browsing a collection of web documents or organizing the query results returned by a search engine. It may also serve as a preprocessing step for other data mining algorithms such as web document classification. An ambitious goal of web document clustering is to automatically generate hierarchical clusters of web documents [5].

The nature of database technology and automated data collection tools leads to tremendous amounts of data stored in databases, data warehouses, and other information repositories. These large amounts of data are worthless unless they become knowledge - not to mention analyzing them is a trivial task either. This problem is called data explosion meaning rich in data, but starved in knowledge.

1.2 Clustering

Clustering is a process of partitioning a set of data objects into a set of meaningful subclasses, called clusters. Formally, there is given a collection of n objects each of which is described by a set of p attributes, clustering aims to derive a useful division of the n objects into a number of clusters [7]. A cluster is a collection of data objects that are similar to one another based on their attribute values, and thus can be treated collectively as one group. Clustering is useful in getting insight into the distribution of a data set.

A clustering algorithm attempts to find natural groups of data based on similarity of attributes.

1.3 Requirements of clustering in data mining.

- Scalability. Many clustering algorithms work fine on small data sets; however, some of them fail to handle large data set containing over ten thousands of data objects. An immediate solution to this problem is to perform clustering on a subset (or sample) of a given large data set, but it may lead to biased results.

- High dimensionality. A database can contain several dimensions or attributes. Most of the clustering algorithms work well on low-dimensional data, but may fail to cluster data objects in high-dimensional space, especially when the data objects are very sparse and highly skewed. In high dimensional data sets, natural clusters usually do not exist in the full dimensional space, but only in the subspace formed by a set of correlated dimensions. Locating clusters in the subspace can be challenging.

1.4 Clustering of global itemset

One typical example is web document clustering which is also the focus of this paper. Many clustering algorithms simply construct a new dimension for each distinct word in the web document set [8]. Due to the large corpus in English, the space usually contains over ten thou Consider the four web documents in table 3.1. After applying the Apriori algorithm to the web document vectors, we compute the global frequent items: “Registration”, “layer”, “patient” and “treatment”. Thus, each web document is represented by a feature vector which is supposed to be a vector of inverse web document frequencies (IDF). For the purpose of better understand, however, we use simply the frequency of an item, i.e., the number of occurrences of a word in a web document. For example, the feature vector of web document *File1* is (5, 0, 0, 0, 1) which represents the frequencies of the global frequent items “Registration”, “layer”, “patient” and “treatment” in web document *file1* respectively. The four feature vectors in table 3.1 form the vector model for our subsequent clustering operations.

	Web document name	**Feature Vector** **(Reg., layer, patient, treatment)**
1	**File1**	**5 0 0 1**
2	**File2**	**3 0 0 1**
3	**File3**	**1 1 4 2**
4	**File4**	**5 0 2 0**

Table 3.1 Web document Set

	Global support
{Patient}	**60%**
{Treatment}	**40%**
{Registration}	**100%**
{Patient, Reg.}	**50%**

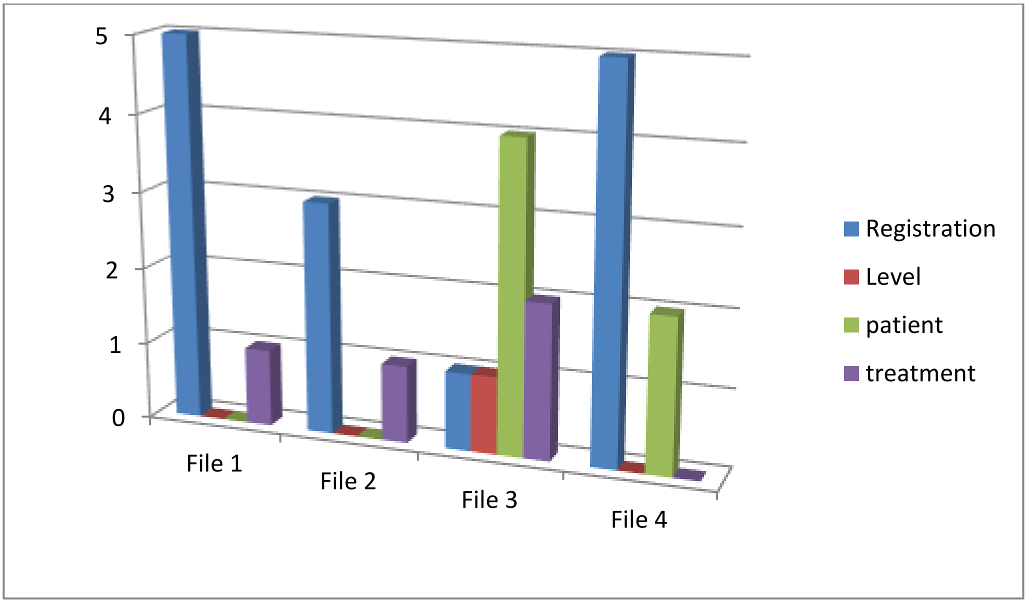
Table 3.2 Global Frequent Itemset

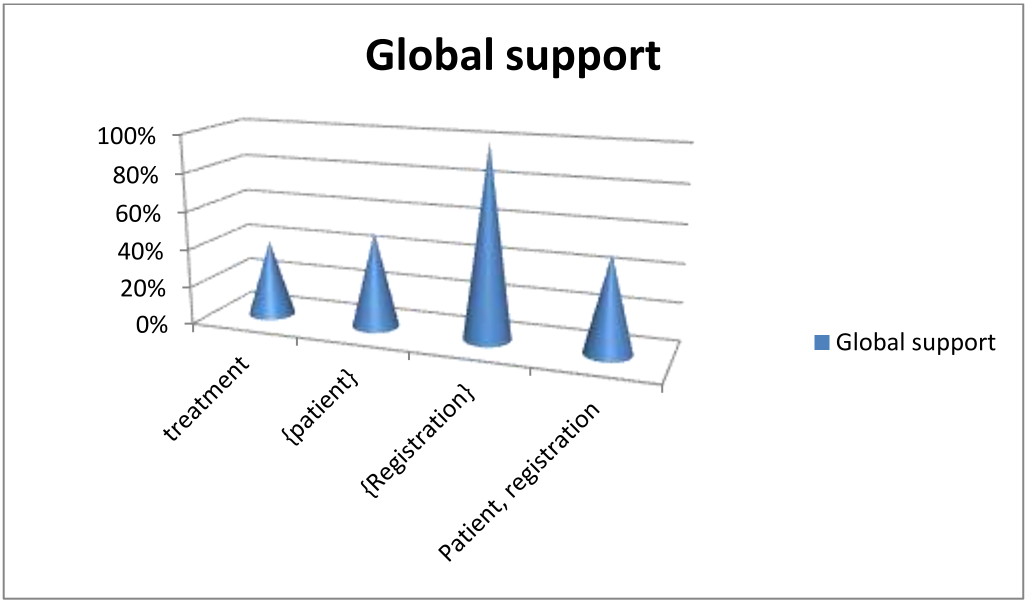
Minimum Global support = 35%

Table 3.2 specifies all the global frequent *k*-itemsets with their global supports. For example, the global support of the global frequent item {patient*}* is 50% because half of the web documents in the set contain the item "patient". In this example, an itemset is frequent only if its global support is larger than or equal to 35%.

1.5 Results

The graph for web documents in file 1, file 2, file3 and file 4 and global support of frequent itemsets can be constructed as below:





1.6 Conclusion

We use the low-dimensional feature vector, which is composed of global frequent items, in place of the original high-dimensional web document vector. This replacement drastically reduces the dimension of the web document vector space. Consequently, it greatly enhances the efficiency and scalability.

we purposes a new algorithm which requires only two scans of the web document set to cluster all the web documents: one scan for constructing initial clusters and one scan for making clusters disjoint, so efficiency will be enhanced.

References

[1] C. Aggarwal, S. Gates, and P. Yu. On the merits of building categorization systems by supervised clustering. In *Proceedings of (KDD) 99, 5th (ACM) International Conference on Knowledge Discovery and Data Mining*, pages 352–356, San Diego, US, 1999. ACM Press, New York, US.

[2] R. Agrawal and R. Srikant. Fast algorithm for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12-15 1994.

[3] R. Agrawal and R. Srikant. Mining sequential patterns. In *Proc. 1995 Int. Conf. Data Engineering*, pages 3–14, Taipei, Taiwan, March 1995.

[4] Yu HIRATE, Eigo IWAHASHI, and Hayato YAMANA, FP-Growth: An Efficient Algorithm for Mining Frequent Patterns without any Thresholds, *Graduate School of Science and Engineering, Waseda University, {hirate, eigo, yamana},* Sept. 2001

[5] F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In *Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD)'2002*, Edmonton, Alberta, Canada, 2002. http://www.cs.sfu.ca/˜ ester/publications.html.

[6] J. Hipp, U. Guntzer, and G. Nakhaeizadeh. Algorithms for association rule mining - a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64, July 2000.

[7] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. *KDD'99*, 1999.

.

[8] Miller. Princeton wordnet, 1990.

[9] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. *KDD Workshop on Text Mining'00*, 2000.

[10] H. Uchida, M. Zhu, and T. Della Senta. Unl: A gift for a millennium. The United Nations University, 2000.