ENHANCED CLUSTERING APPROACH FOR WEB USAGE MINING

V Aruna[1] , Dr. Harsh Pratap Singh[2] , Dr. D. Sujatha[3]
1 .Research Scholar ,Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, MadhyaPradesh, India
2 Research Guide, Dept. of Computer Science & Engineering, Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal-Indore Road, MadhyaPradesh, India
3 Research Co-Guide, Dept. of Computer Science & Engineering Malla Reddy College of Engineering & Technology.

Abstract

Data mining technology has been considered as important means for extracting patterns and trends from huge amount of information. So, this approach is essentially used to extract the unknown pattern from the large set of information for business and also for real time applications. data processing may be a computational intelligence discipline which has emerged as a effective tool for data analysis, new KDD and good higher cognitive process. The raw and unidentified data from the huge volume of dataset may be classified initially in an unsupervised fashion by using cluster analysis.Clustering technique is the process of grouping objects together by which similar objects are combined into one group and disimilar objects into other group. Clustering technique may be employed in many applications for instance biological, financial applications and many more. one of these application types is Web clustering where differing kinds of objects will be clustered into different groups for various purposes. This paper deals with the various aspects of Web data processing and provides a summary about the different techniques employed in this field.

IndexTerms-Clustering , Web usagemining ,Objects

1. Introduction

1.1 Web Usage Mining

Web usage mining is one of the application of knowledge mining techniques to get interesting usage patterns from web usage data, so as to know and better serve the necessity of web-based applications. Usage data captures an origin of web users along with their browsing behavior at a web site. Web usage mining could also be classified further based on the type of usage data considered. (a) Web Server data In web server data there are user logs which are collected by the online server and typically include IP address, page reference and time interval .

(b) Application Server Data Commercial application servers have significant features to enable e-commerce applications.

(c) Application Level Data These sort of events are often defined in an application and logging are often turned off them generating histories of those events.

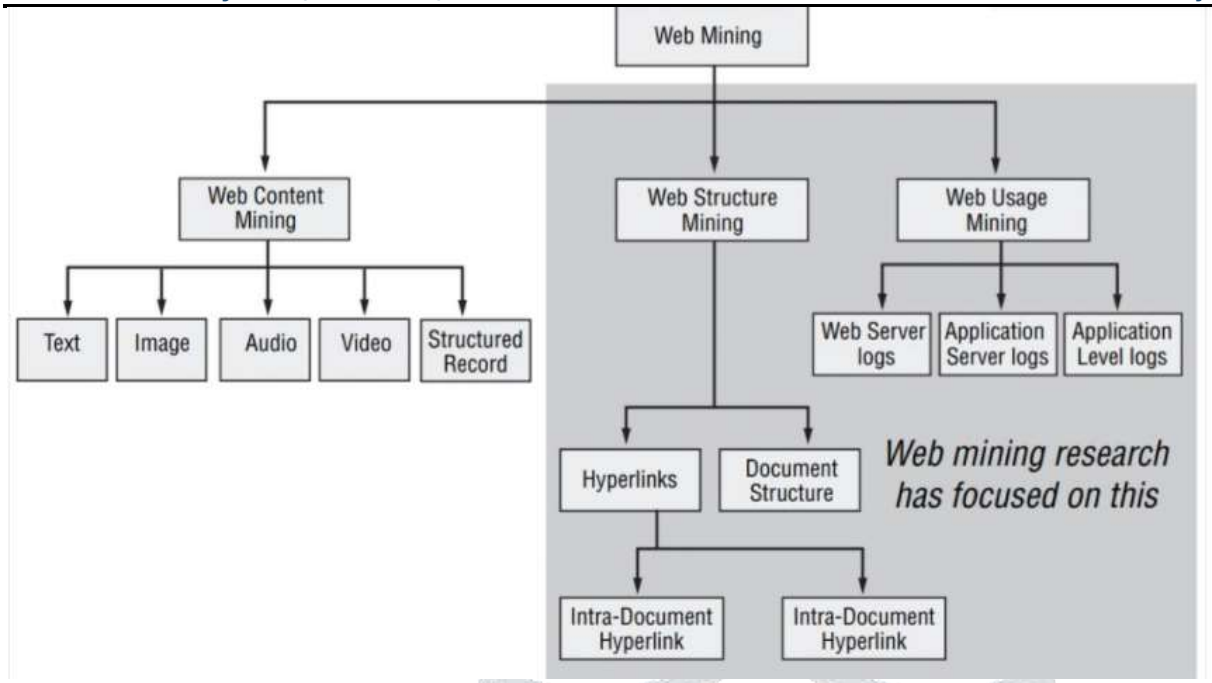

Figure 1: Web mining Taxonomy

Web usage mining consists of three main steps:

(i) preprocessing,

(ii) pattern discovery and

(iii) pattern analysis .

Figure 2 shows the block diagram of the process of Web usage mining.

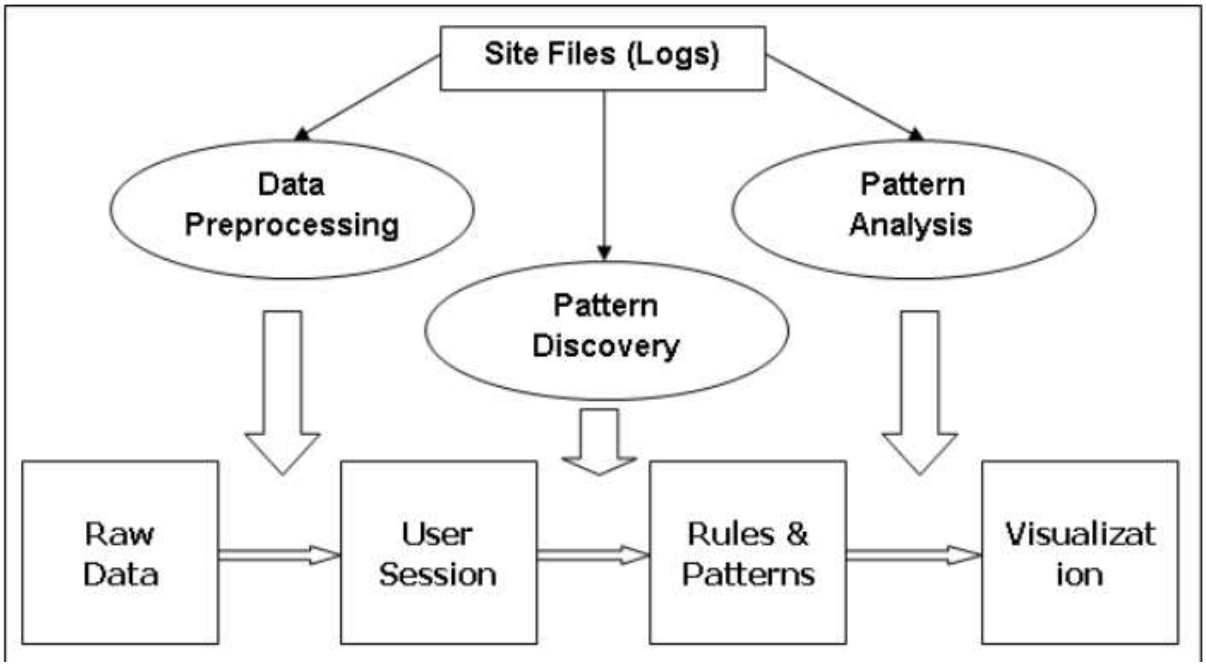

Figure 2. Process of Web usage mining

In the preprocessing phase the data required is collected from the various places like (client side, server side, proxy servers). After identifying the users, the click-streams of every user has to be split into sessions. generally the timeout for determining a session is about to 30 minute . The pattern discovery phase means applying data processing techniques on the preprocessed log data. It can be done by means of frequent pattern mining, association rule mining or clustering. This paper deals only with the task of clustering web usage log. In web usage mining there are two sorts of clusters to be discovered:usage clusters and page clusters.

The aim of clustering users is to determine groups of users having similar browsing behaviour. The users are often clustered on the basis of several information, the users are often requested filling out a form regarding their interests, for instance when registering on the online portal. The clustering of the users are often accomplished supported the forms. On the opposite hand, the clustering are often made on the basis of the knowledge gained from the log data collected during the user was navigating through the portal. differing types of user data are often collected using these methods, for instance (i) characteristics of the user (age, gender, etc.), (ii) preferences and interests of the user, (iii) user's behavior pattern.

The aim of clustering sites is to divide groups of pages that have similar content. This information is often useful for search engines or for applications that make use of dynamic index pages.

The last step of the entire web usage mining process is to perform a analysis on the patterns found during the pattern discovery step. The irrelevant patterns need to be filtered out, and therefore the resulted patterns or clusters need to be validated. Some visualization techniques can help this process for the user.

1.2 Requirements of Clustering

The following are important requirements of cluster analysis in data mining:

1)Scalability: Some of the clustering algorithms work well on small data sets containing fewer than 200 data objects. However, a large database contains large no. of data objects.

2)Minimal requirements for domain knowledge of determine input parameters: The clustering process results are often sensitive to input parameters sometimes

3)Ability to handle differing types of attributes: Many algorithms are designed to cluster interval-based data. However, applications can also require clustering differing types of knowledge .

4)Ability to handle noisy data: Most real-world databases contain outliners or missing, unknown, errors in data.

5)Discovery of clusters with arbitrary shape: it's needed to develop algorithms for detection clusters of arbitrary shape.

6)Insensitivity to the order of input records: Some clustering algorithms are sensitive to the order of input data; for example, may generated dramatically different clusters. Development of algorithms that are insensitive to the order of input .

7)Constraint-based clustering: Real-world applications may have to perform clustering under various sorts of constraints. Suppose that you have to choose the locations for a given number of new automatic cashdispensing machines (ATMs) in a city.

8)High dimensionality: A database or a knowledge warehouse can contain various dimensions and attributes. Many clustering algorithms handling low-dimensional data, involving only two-three dimensions. Human eyes can judge the standard of clustering for up to 3 dimensions easily.

9)Interpretability and usability: Users expect clustering results should be interpretable, effective, and usable.

2.Classifying the Different Web Clustering Algorithms

There exists an excellent type of Web usage clustering algorithms that may be categorized regarding several aspects. during this paper the aspects for classifying the various algorithms are the following: (i) the kind of the objects to be clustered

(ii) the aim of clustering,

(iii) The clustering algorithm used,

(iv)The kind of the clusters discovered (cluster overlap handling) and

(v)Similarity measure.

This paper deals with the various aspects of classifying and the process is described very well, and also the most vital and also the best known web usage clustering algorithms are categorized on the basis of the proposed aspects.

2.1 Object type

One aspect of categorizing an Web usage clustering algorithm is that the form of the objects to be clustered. As mentioned within the previous section the target of the clustering will be various.

• Web pages: the foremost common task is to cluster web content on the basis of the navigation behavior of the users.

• website sequences: the idea of the clustering algorithm may be not only the frequency of visiting an online page, but also the frequency of visiting a sequence of websites. during this case the order of the pages also plays a very important role.

• User rating results: if the users have the chance rating the various documents and web content, then the pages are often clustered on the basis of this information. this sort of clustering is simpler than clustering from log data because during this case the questions are often founded in such a way that the resulting answer vector suits the data probe for.

• Registration information: Another style of user feature collection is that the registration information during which the user supply many information about himself and about his interests. From the primary three form of objects the task of online page clustering is accomplished, while from the last kind of objects the users are often clustered.

2.2 Clustering Purpose

It plays a key role what the aim of the clustering algorithm is. There exist many of purposes from which we just mention some frequently used. the foremost frequently used goal is to create a dynamic portal with page recommendation. For this reason the pages, documents or perhaps user rating results may be clustered. Another important aspect is to possess a portal with personalized pages or with user profiles. For this reason a user model should be created regarding their navigational behavior. Page categorizing or page indexing makes the navigation the users easier because during this case pages that are similar in an exceedingly given manner are enumerated on the subject of one another.

2.3Clustering algorithm

Because each clustering method performs differently for various purposes, the net usage clustering algorithms uses different basic clustering algorithms regarding the tasks they're accomplishing.. Moreover, beside the essential algorithms, some new approaches are utilized in Web usage mining in addition. Markov models and Fuzzy clustering algorithms are frequently employed in this field. Semantic Latence Analysis (LSA) is another approach which will be used for Web usage mining.

2.4 Cluster overlap handling

It is a motivating question how the boundary of a cluster is defined. In other cases, however, it's enabled to possess objects that belong to over one clusters at the identical time. during this case the algorithm discovers overlapping or fuzzy clusters.

2.5 Similarity measure

Because of the non-numerical feature of the net usage clustering problem, it's a very important aspect how the similarity of the objects to be clustered is defined. In some cases the non numerical attributes are omitted, or transformed to numerical values, and Eucledian or Minkowski distance is employed. In other cases new metrics are introduced so as to induce a stronger clustering result

Conclusion

This paper deals with the problem of extracting hidden data from huge amount of log information, namely, with web usage usage mining. The main target of this paper is to perform cluster among the various mining processes. when describing the task of cluster, the foremost common cluster strategies were enumerated basedon their basic approach. Web log file is given as input and then data cleaning is performed to eliminate irrelevant data items. The cleaned web log will be now used for pattern discovery. Clustering techniques are used for discovering different useful usage patterns.

References

[1] Q. Yang and H. H. Zhang, "Web-log mining for predictive web caching.," IEEE Trans. Knowl. Data Eng., vol. 15, no. 4, pp. 1050–1053, 2003.

2] Kosala and Blockeel, "Web mining research: A survey," SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, vol. 2, 2000.

[3] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining," in Data Warehousing and Knowledge Discovery, pp. 303– 312, 1999.

[4] J. Borges and M. Levene, "Data mining of user navigation patterns," in WEBKDD, pp. 92–111, 1999.

[5] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data mining and the web: Past, present and future," in ACM CIKM'99 2nd Workshop on Web Information and Data Management (WIDM'99), Kansas City, Missouri, USA, November 5-6, 1999 (C. Shahabi, ed.), pp. 43–47, ACM, 1999.

[6] S. Chakrabarti, "Data mining for hypertext: A tutorial survey.," SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, vol. 1, no. 2, pp. 1–11, 2000.

[7] M. Balabanovic and Y. Shoham, "Learning information retrieval agents: Experiments with automated web browsing," in Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources, pp. 13–18, 1995.

[8] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data, (New York, NY, USA), pp. 307–318, ACM Press, 1998.

[9] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a graph: Measurements, models and methods," Lecture Notes in Computer Science, vol. 1627, pp. 1–18, 1999.

[10] J. Hou and Y. Zhang, "Effectively finding relevant web pages from linkage information.," IEEE Trans. Knowl. Data Eng., vol. 15, no. 4, pp. 940–951, 2003.

[11] H. Han and R. Elmasri, "Learning rules for conceptual structure on the web," J. Intell. Inf. Syst., vol. 22, no. 3, pp. 237–256, 2004.

[12] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," ACM Trans. Inter. Tech., vol. 3, no. 1, pp. 1–27, 2003.

[13] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications, (London, UK), pp. 396–407, Springer-Verlag, 2000.

[14] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," SIGKDD Explorations, vol. 1, no. 2, pp. 12–23, 2000. [15] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper," 2002.