

# Visual Data Mining with Information Visualization

Miss. Ankita V. Suramwar, Prof. Anup Gade

**Abstract:-** Data visualization is the graphical display of abstract information for two purposes: sense-making (also called data analysis) and communication. Important stories live in our data and data visualization is a powerful means to discover and understand these stories, and then to present them to others. In this paper, we propose a classification of information visualization and visual data mining techniques which is based on the data type to be visualized, the visualization technique and the interaction and distortion technique. We exemplify the classification using a few examples, most of them referring to techniques and systems presented in this special issue.

**Keywords:** Visual Data Mining, Data Visualization

\*\*\*\*\*

## I. INTRODUCTION

Never before in history have data been generated at such high volumes as it is today. Exploring and analyzing the vast volumes of data has become increasingly difficult. Information visualization and visual data mining can help to deal with the flood of information. The advantage of visual data exploration is that the user is directly involved in the data-mining process. There are a large number of information visualization techniques that have been developed over the last few years to support the exploration of large datasets. In this chapter, we provide an overview of information visualization and visual data-mining techniques and illustrate them using a few examples.

Data visualization is the graphical display of abstract information for two purposes: sense-making (also called data analysis) and communication. Important stories live in our data and data visualization is a powerful means to discover and understand these stories, and then to present them to others. The information is abstract in that it describes things that are not physical. Statistical information is abstract. Whether it concerns sales, incidences of disease, athletic performance, or anything else, even though it doesn't pertain to the physical world, we can still display it visually, but to do this we must find a way to give form to that which has none.

### Review of Literature:

#### A. A dimensionality reduction approach to support visual data mining: co-ranking-based evaluation, Communications (COMM), IEEE International Conference in June 2016

This paper brings into focus a visualization based approach to mining the EO data. This method aims to map the existing data correlations in the multidimensional information space to the spatial correlations revealed by the 3D space. The assessment of the results considers a single and global quality criterion, involving the number of the

intrusions and extrusion to reveal the performance of dimensionality reduction methods. Additionally, effective data mining involves the user into data exploration, making use of his knowledge to initiate and validate new hypothesis. The major issue when handling with this type of data is to reduce its complexity preserving the relevant information for understanding the structure and / or the semantic content of the data. Recently several visual data mining systems (VDMs) were developed. A VDM system aims to combine the traditional data mining algorithms with information visualization methods, allowing users to extract data models or patterns by directly interacting with data.

#### B. Wastewater treatment aeration process optimization: A data mining approach, Journal of Environmental Management in Dec 2016

Presently, a data-driven approach has been applied for aeration process modeling and optimization of one large scale wastewater in Midwest. More specifically, aeration process optimization is carried out with an aim to minimize energy usage without sacrificing water quality. More specifically, aeration process optimization is carried out with an aim to minimize energy usage without sacrificing water quality. Models developed by data mining algorithms are useful in developing a clear and concise relationship among input and output variables. Results indicate that a great deal of saving in energy can be made while keeping the water quality within limit. Limitation of the work is also discussed.

#### C. Identifying user habits through data mining on call data records, Engineering Applications of Artificial Intelligence in Sep 2016

In this, authors had tried to use various segmentation methods for recognition and classification of fruits. Color image processing and image segmentation are the methods used for fruit classification based on color.

Classification based on Size can process by Regional descriptor method. Boundary descriptor and feature extraction are used to classify based on a shape of the fruit. Automation of fruit recognition and classification is an interesting application of computer vision. According to author the traditional fruit classification methods are relied on manual process based on visual skill and these methods become inconsistent, time consuming and tedious. External form appearance is the only source for classification of fruits. Researches in this area uses the machine vision systems for improving quality of a product which frees the people from the conventional hand sorting of fruits.

**D. A Visual data mining techniques for classification of diabetic patients, Advance Computing Conference (IACC), IEEE 3rd International in May 2013, IEEE**

Clustering technique is quite often used by many researchers, it uses Expectation-Maximization (EM) algorithm for sampling. The study of classification of diabetic patients was main focus of this research work. The study of classification of diabetic patients was main focus of this research work. Diabetic patients were classified by data mining techniques for medical data obtained from Pima Indian Diabetes (PID) data set. This research was based on three techniques of EM Algorithm, h-means+ clustering and Genetic Algorithm (GA). These techniques were employed to form clusters with similar symptoms. Result analyses proved that h-means+ and double crossover genetics process based techniques were better on performance comparison scale. The simulation tests were performed on WEKA software tool for three models used to test classification. The hypothesis of similar patterns of diabetes case among PID and local hospital data was tested and found positive with correlation coefficient of 0.96 for two types of the data sets. About 35% of a total of 768 test samples were found with diabetes presence.

## II. PROPOSED SYSTEM

Graphical visualization has been already used in different aspects of human activity, but the effectiveness and even applicability of methods can become a real problem with data volumes growth and data production speed. The described problem comes from the following points:

(1) The need of artificial preparation of data slices, for partial data visualization;

(2) Visual limitation to the number of perceived data factors.

We need to overview existing data visualization methods and provide approaches, which can solve these problems. These approaches must provide more perceptible and informative data representations to help the analyst in finding hidden relations in Big Data.

Most of data visualization methods usually does not appear from nothing, but they become a development of earlier existing methods.

At most, the analyst tools must meet the following requirements:

(1) Analyst should be able to use more than one data representation view at once;

(2) Active interaction between user and analyzable view;

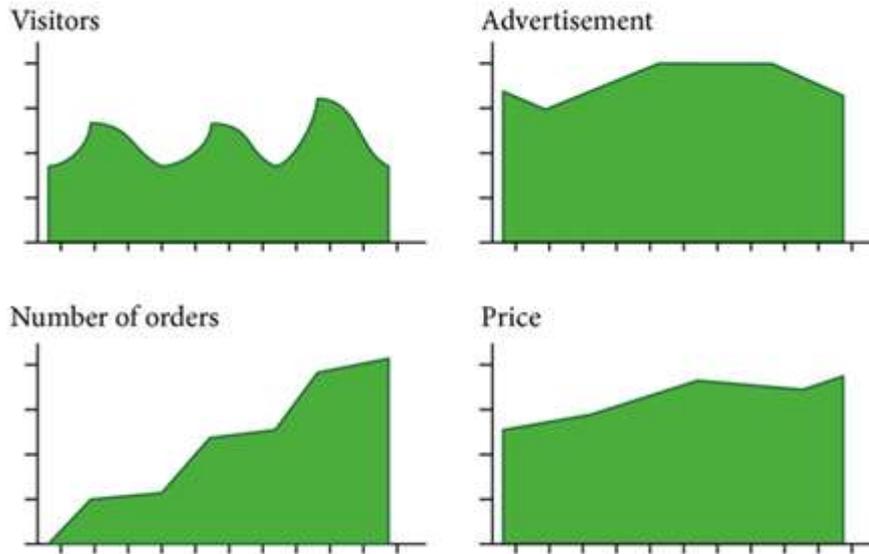
(3) Dynamical change of factors number during working process with view.

Below we will describe these requirements more clearly.

### More than one view per representation display

In order to reach a full data understanding, analyst usually uses simple approach, when he places different classical data views, which include only a limited set of factors so that he can easily find some relations between these views or in one concrete view [4, 5].

Despite the fact that there can be used completely every method of data visualization, often, we can see an approach, when the analyst uses just some similar or near to similar graphical objects. As an example, linear or dot diagrams (Figure 1). Of course, the analyst might be interested in comparing totally different visualizations of the same data, but the whole process of visual analysis, in that case, becomes much harder. Now, the researcher must compare not only similar graphical objects, but he also has to clearly distinguish different data and make a decision, based on different factors [6].

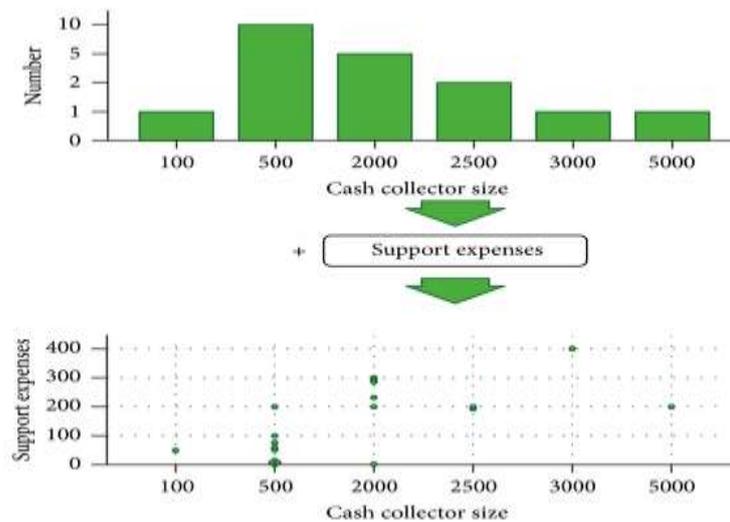


**Fig: Dynamical Changes in Number of Factors**

Perhaps the most fundamental operation in visual analysis is contained in a data visualization specification. An analyst has to indicate which data is to be shown and how it should be shown to ease the information perception.

Any graphical visualization can be applied absolutely to any data, but there is always a topical question of whether the chosen method is correctly applied to the dataset, in order to get any useful information? Typically, for Big Data, the analyst cannot observe the whole dataset,

find anomalies in it, or find any relations from the first glance [6]. So, another one topical approach is a dynamical change in the number of factors. After the analyst has chosen one factor, he is willing to see a classical histogram, which shows the distribution of records number depending on record type. As an example below, in Figure 3, on top histogram, we can see dependency between the number of cash collector units currently in use by payment system and the volume of each cash collector.



**Fig: Visualization in terms of Graphs**

**Filtering**

The issue of value discernibility was always topical for visual analysis and it becomes more important in case of Big Data [6]. Even if we show only 60 unique values, not to

mention millions of them, on one diagram, it is very difficult to place a label for each one.

And even more, there can be totally different value ranges in one data set. Therefore, some values would be just dominated by others with higher amplitude levels. So, as a result, the perception of whole diagram would be complicated. For example, some organizations, which work 24 hours per day can have different customers flow, and showing the dependency of customers by hour, we will lose perception ability for a group of night hours, when values are near equal and have a much lower amplitude comparing to a day hours.

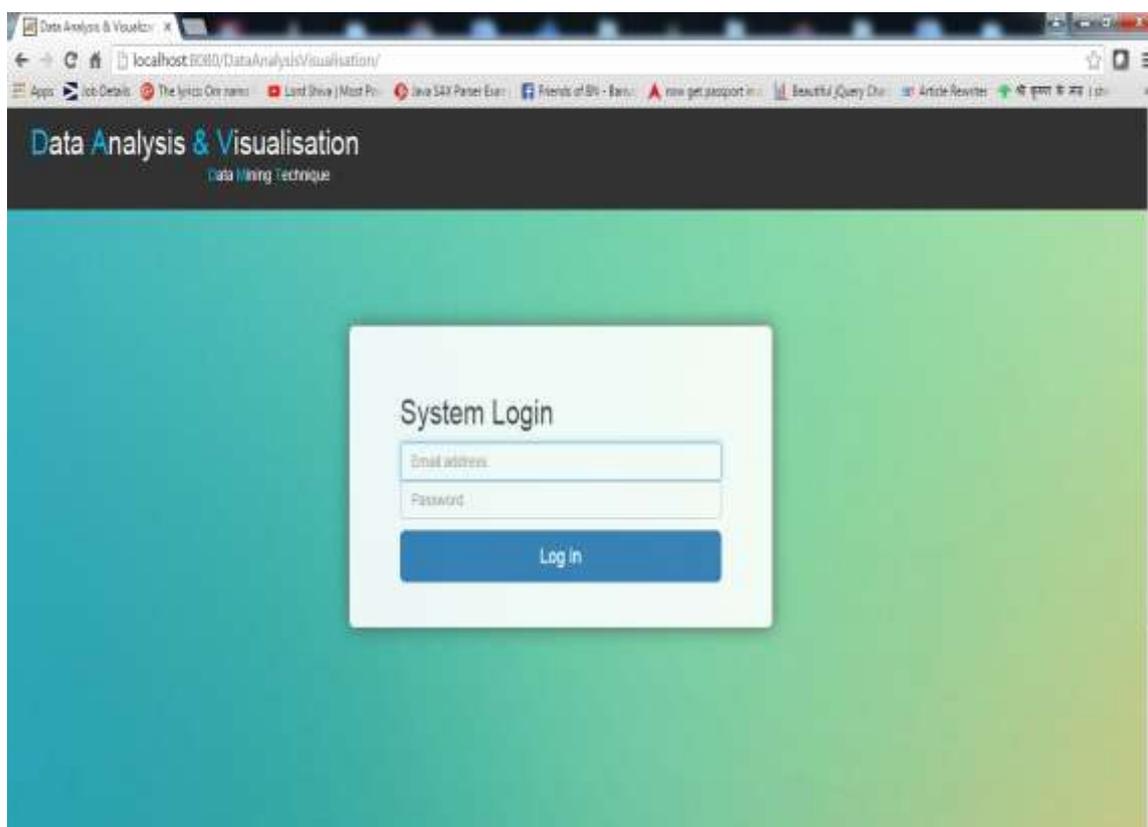
Analyst usually wants to see both whole data representation and a partial and more detailed data

representation lying in his area of interest. Moreover, the area of his interest is not static and can dynamically change during research process.

The filtering system and an overview map are used as an approach, solving these problems (Figure 4). Analyst can change the range on an overview map and see the detailed visualization of data in that range.

### Snapshots User Actions

#### User logins from this page



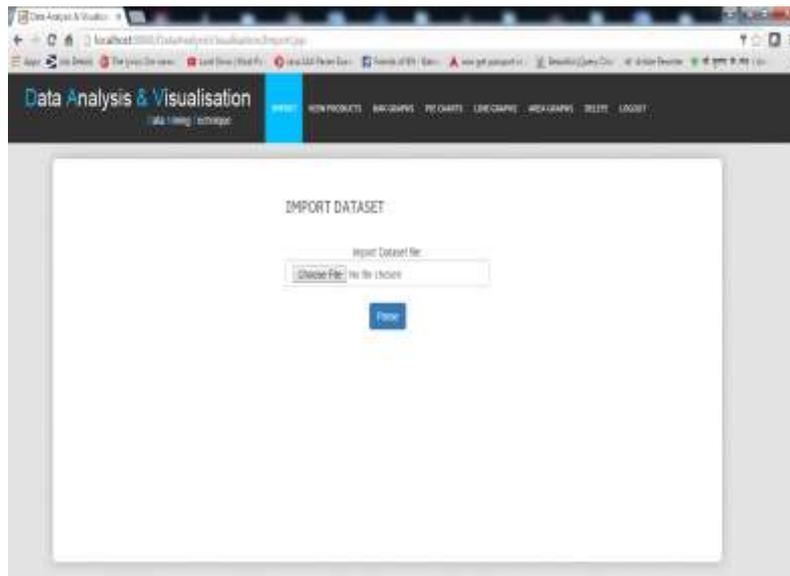
**Fig: Login Page**

The Login page is designed to authenticate the user before enter into the Analysis and Visualization system. The reason behind that is, most of the user's data is confidential and they dont want to share the data with anyone or access the data by anyone.

Some kind of data is highly confidential like SSN number, any passwords , any user or email ids.

To prevent this to access these from any third party, we have implemented the Authentication system,

### User uploads and parses the dataset



**Fig: File Import and Parse Dataset Page**

Here, we can import the dataset file that has been downloaded from any external resource or obtained from any source.

**Dataset:** A data set is a collection of related, discrete items of related data that may be accessed individually or in combination or managed as a whole entity.

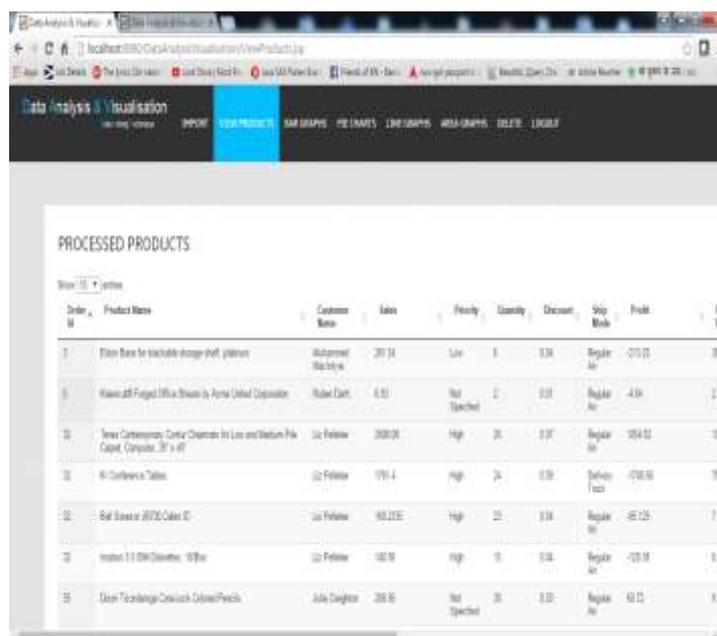
It is organized into some type of data structure. In a database, for example, a data set might contain a collection of business data (names, salaries, contact information, sales figures, and so forth). The database itself can be considered a data set, as can bodies of data within it related to a

particular type of information, such as sales data for a particular corporate department.

**Parsing:** Parsing is the process of analyzing text made of a sequence of tokens to determine its grammatical structure with respect to a given (more or less) formal grammar. The parser then builds a data structure based on the tokens.

So here, we are importing the dataset file and loading it into the buffer. The parser we have written parses the file and convert it into the structured format.

### Displays the data loaded with feature available of Sorting and Searching



**Fig: View Parsed Data Page**

The data that is loaded and parsed by the system can be displayed on this page. This is the structured data that is generated from the parser. A table is an arrangement of data in rows and columns, or possibly in a more complex structure. Tables are widely used in communication,

research, and data analysis. So we are displaying the data in tabular format to make it more readable, searchable and sortable.

### Generate reports based on the parameter options

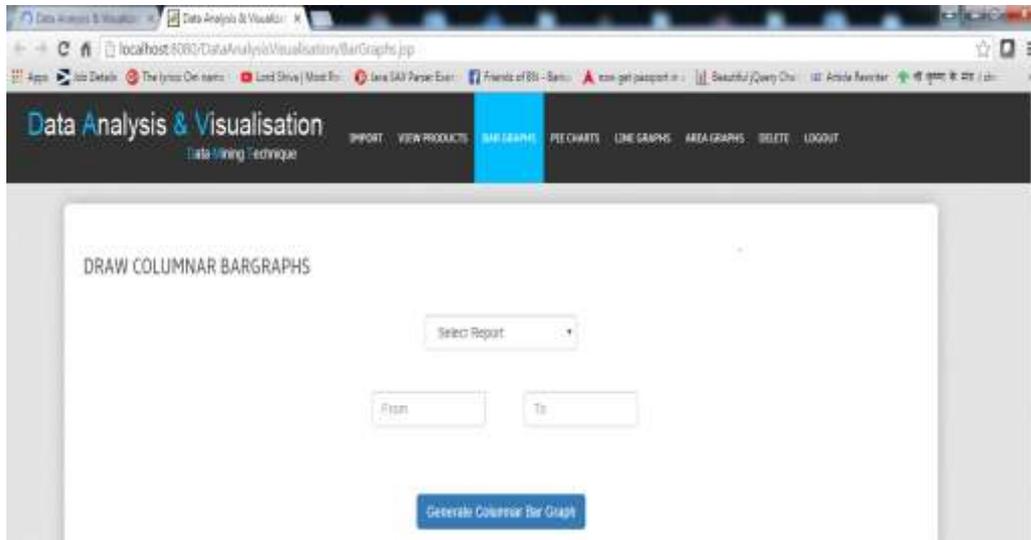


Fig: Generate Graphs Page

A graph or chart or diagram is a diagrammatical illustration of a set of data. It can be placed within articles, a table or chart, and a graph and a chart are quite similar and considered the same. Just like any other image, graphs must be accurate and convey information efficiently. They should be viewable at different computer screen resolutions. Ideally, graphs will also be aesthetically pleasing.

On this page, we are generating the graphs and charts on the basis of the visual query. We are selecting the type of graph, type of report, and date range to generate the graphs and charts.

Graphs have two axes, the lines that run across the bottom and up the side. The line along the bottom is called the horizontal or x-axis, and the line up the side is called the vertical or y-axis. So basically we are generating the set of result set and put the points to the X and Y axis.

### III. CONCLUSION

Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports. Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments. A picture is worth a thousand words – especially when you're trying to find relationships and understand your data, which could include thousands or even millions of variables. Data visualization is a vast topic

and consist of many sub-parts which are a subject in itself, we in our chapters have tried to paint a clear picture of what you need to know and what people will be looking of you in a visualization project. It is structured to provide all the key aspect of Data visualization in most simple and clear fashion. So you can start the journey in Data visualization world. People who want to get into data visualization are developers who want to work in analytics and visualization project, web, mobile app and software designer, design thinker, graduate students and university students.

So we have arrived at the end of our incursion into the field of data visualization. In the above chapters, we have presented a number of the most important theoretical and practical ingredients involved in the design of visualization methods and applications. As we have seen, designing an efficient and effective data visualization application is a complex process. This process involves representing the data of interest, processing the data to extract relevant information for the problem at hand, designing a mapping of this information to a visual representation, rendering this representation, and combining all this functionality in an easy-to-use application.

### REFERENCES

- [1] Yu-RuLin, Nan Cao, David Gotz, "Visual Mining for Data with Uncertain Multi-labels via Triangle Map", Data

- Mining (ICDM), 2014 IEEE International Conference on 29 January 2015
- [2] C. M. Velu , K. R. Kashwan, “Visual data mining techniques for classification of diabetic patients” , Advance Computing Conference (IACC), 2013 IEEE 3rd International, 13 May 2013
- [3] Laura P. R. Rivera , Sergio V. Chapa Vergara, AmilcarMenesesViveros, “Visual data mining over a video wall” , Electrical Communications and Computers (CONIELECOMP), 2012 22nd International Conference on , 26 April 2012
- [4] V. Vijayakumar, R. Nedunchezian,“A study on video data mining” International Journal of Multimedia Information Retrieval October 2012, Volume 1, Issue 3, pp 153–172 Year: August 2012.
- [5] Ko Fujimura Shigeru Fujimura, TatsushiMatsubayashi, Takeshi Yamada and Hidenori Okuda, "Topigraphy: Visualization for Large-scale Tag Clouds" in WWW 2008 / Poster Paper, pp. 1087-1088, 2008.
- [6] Jiyang Chen , Tong Zheng , William Thorne,“Visual Data Mining of Web Navigational Data Purchase or Sign In ”, Information Visualization, 2007. IV '07. 11th International Conference ,July 2007.
- [7] Alfredo Cuzzocrea, DavoodZall ,“ Parallel Coordinates Technique in Visual Data Mining ”, Information Visualisation (IV), 2013 17th International Conference December 2013.
- [8] D. Shukla, KapilVerma, JayantDubey, SharadGangele, "Cyber crime Based Curve Fitting Analysis in Internet Traffic Sharing in Computer Network", International Journal of Computer Application (IJCA), vol. 46, no. 22, pp. 41-51, 2012 c.
- [9] GangeleSharad, VermaKapil, D. Shukla, "Bounded Area Estimation of Internet Traffic Share Curve", International Journal of Computer Science and Business Informatics (IJCSBI), vol. 10, no. 1, pp. 54-67, 2014.
- [10] D. Shukla, KapilVerma, SharadGangele, "Iso-Failure in Web Browsing using Markov Chain Model and Curve Fitting Analysis", International Journal of Modern Engineering Research(IJMER), vol. 02, no. 02, pp. 512-517, 2012 a.
- [11] D. Shukla, KapilVerma, SharadGangele, "Least Square Curve Fitting Applications under Rest State Environment in Internet Traffic Sharing in Computer Network", International Journal of Computer Science and Telecommunications (IJCST), vol. 03, no. 05, pp. 43-51, 2012 b.
- [12] R. Agrawal et al., "Fast Algorithm for Mining Association Rules", Proceedings of International Conference on Very Large Data Bases, pp. 487-499, 1994.
- [13] M. S. Chen, J. S. Park, P. S. Yu, "Efficient Data Mining for Path Traversal Patterns in a Web Environment", IEEE Transaction on Knowledge and Data Engineering, vol. 10, no. 2, pp. 209-221, 1998.
- [14] M. S. Chen, X. M. Huang, I. Y. Lin, "Capturing User Access Patterns in the Web for Data Mining", Proceedings of IEEE International Conference on Tools with Artificial Intelligence, pp. 345-348, 1999. Show Context
- [15] S. Y. Chen, X. Liu, "Data Mining from 1994 to 2004: an Application-Orientated Review", International Journal of Business Intelligence and Data Mining, vol. 1, no. 1, pp. 4-21, 2005.
- [16] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", Proceedings of IEEE International Conference on Tools with Artificial Intelligence, pp. 558-567, 1997.
- [17] M. EL-Sayed, C. Ruiz, E. A. Rundensteiner, "FS-Miner: Efficient and Incremental Mining of Frequent Sequence Patterns in Web Logs", Proceedings of 6th ACM International Workshop on Web Information and Data Management, pp. 128-135, 2004.
- [18] I. Hamburg, M. Hersh, M. Gavota, M. Lazea, "Open Web-Based Learning Environments and Knowledge Forums to Support People with Special Needs", International Journal of Interactive Technology and Smart Education, vol. 1, no. 3, pp. 205-216, 2005.
- [19] J. Han, J. Pei, Y. Yin, R. Mao, "Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach", Data Mining and Knowledge Discovery, vol. 8, no. 1, pp. 53-87, 2004.
- [20] J. Han, J. Pei, H. Lu, S. Nishio, S. Tang, D. Yang, "H-Mine: Hyper-Structure Mining of Frequent Patterns in Large Databases", Proceedings of IEEE International Conference on Data Mining, pp. 441-448, 2001.