Data Visualization for Online Bill Submitting Users' Behavior Tracking

Mr. A. Mallikarjuna, Teaching Assistant
Dept. of Computer Science
Sri Venkateswara University,
Tirupati, Andhra Pradesh, India

Dr. G. Anjan Babu, Associate Professor
Dept. of Computer Science
Sri Venkateswara University,
Tirupati, Andhra Pradesh, India

Abstract

Online Billing is an important component of electronic commerce. E-commerce additionally includes enterprise resource planning systems (ERP), data mining and data warehousing. In this paper, we present an interactive online bill submission application that can be used to provide with greater capabilities to interpret and explore user behavior and manipulations made by the user. When the number of manipulations exceeds a minimum threshold value we can also conclude that there is some kind of fraud in the bill. This System uses two different types of visualization

Techniques:

Visualizations of sessions and visualization of different types of various behaviors by using a number of graphs like bar charts, pie charts, scatter plots etc. We have focused on the discovery of behavioral patterns of users and conceptual relationships among various fields. In particular, by analyzing the data visualization charts, we have detected several previously unknown strategies used by the user. At last, we have detected several correlations among fields, which gave us useful feedback on user's mindset.

Keywords

Data Visualization, Knowledge Discovery, Visual Data Mining, Interactive Data Mining, ERP.

1. Introduction

Online billing is being widely adopted in e-commerce applications as well as in other applications providing powerful tools to assess user's behavior. This paper deals with the detection and manipulation of fraudulent activity inside large organization's primitives.

This problem can be solved via the construction of a visualization system which incorporates the browsers events and knowledge derived from the data mining techniques on real time data.

In this paper, we propose a data exploration approach exploiting information visualization in order to involve analyzers in a visual data mining process aiming to detect manipulations, number of changes, relations between data fields, which can Potentially reveal previously unknown knowledge inherent in online bill submission, such as the price manipulations done by the users, correlation among data fields and many other aspects, including their impact on the total budget.

Here we propose a system that logs all the interactions of users with the system interface. It captures all the browsing and manipulated events by the users and uses this data to visualize charts to detect users mind set and frauds done by the user in bill submission.

Our proposed system is Web based and relies on the AJAX [17] technology which captures all the user's interactions with the user interface (running in the web browser). The system is composed of a user-interface framework for a stand-alone application that analyses the logs in order to extract information from them and to graphically represent it. To demonstrate the effectiveness of the approach and of the proposed system, we have used it in the context of various colleges & universities where national & international level seminars are organized by taking AICTE grants and then producing a fraudulent bill.

2. Related work

Our proposed framework has been taken from existing e-testing system, e-Workbook system and monitoring online test through data visualization [1], which are used to administer online tests for learners. In these existing systems the behavior of the students while writing the test, marks obtained by the student, final score & grade, difficulty level of the questions etc. were analyzed.

In e-testing it is important to administer tests composed of good quality question items. By the term "quality" we intend the potential of an item in effectively discriminating between strong and weak students and in obtaining tutor's desired difficulty level. Since preparing items is a difficult and time-consuming task, good items can be re- used for future tests. Among items with lower performances, instead, some should be discarded, while some can be modified and then re-used. A Web-based e-testing system is present which will

Detect defective question items and, when possible, provides the tutors with advice to improve their quality. The system detects defective items by firing rules. Rules are evaluated by a fuzzy logic inference engine. In monitoring the online tests through data visualization, the details regarding the user are recorded and the actions done by the user are stored in a log file along with the time of actions. We handle all the events like window closed, window minimized, modifying the answers, coming back to a previous question and many others are captured and are visualized using Data Visualization techniques. The behavior of the user is represented in form of various graphs like Bar charts, pie charts etc.

3. Information visualization for knowledge discovery

3.1. Information visualization

Information visualization focuses on data sets lacking inherent 2D or 3D semantics and therefore also lacking a standard mapping of the abstract data onto the physical screen space.

Interactive information visualization tools provide analysts with remarkable capabilities to support discovery of hidden knowledge. By combining powerful data mining methods with user- controlled interfaces, user's behaviors can be mined for Web based data.

3.2. Data visualization

Data Visualization provides a graphical representation of data, documents and structures which are useful for various purposes. Data visualization provides an overview of complex and large data sets, shows a summary of the data, and helps humans in the identification of possible patterns and structures in the data. Thus, the goal of data visualization is to simplify the representation of a given data set, minimizing the loss of information [6], [7].

3.3. Classification of visual data mining techniques

The techniques can be classified based on three Criteria (see figure 1) [6].

- The data to be visualized
- The visualization technique
- The interaction and distortion technique used.

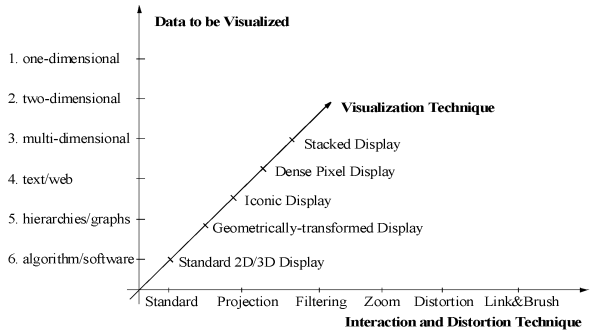

Fig 1: Classification of Information Visualization Techniques

The data type to be visualized [1] may be
• One-dimensional data, such as temporal data
• Two-dimensional data, such as geographical maps
• Multidimensional data, such as relational tables
• Text and hypertext, such as news articles and Web documents
• Hierarchies and graphs, such as telephone calls and Web documents
• Algorithms and software, such as debugging operations

The visualization technique used may be classified into
• Standard 2D/3D displays, such as bar charts and x-y plots as
• Geometrically transformed displays, such as landscapes and parallel coordinates
• Icon-based displays, such as needle icons and star icons
• Dense pixel displays, such as the recursive pattern and circle segments techniques
• Stacked displays, such as tree maps or dimensional stacking

The interaction and distortion technique used.
Interaction and distortion techniques allow users to directly interact with the visualizations. They may be classified into
• Interactive Projection
• Interactive Filtering
• Interactive Zooming
• Interactive Distortion
• Interactive Linking and Brushing
Visualization methods can be either geometric or symbolic. In a geometric visualization, data are represented by using lines, surfaces, or volumes and are usually obtained from a physical model or as a result of a simulation or a generic computation. Symbolic visualization represents non-numeric data using pixels, icons, arrays, or graphs.

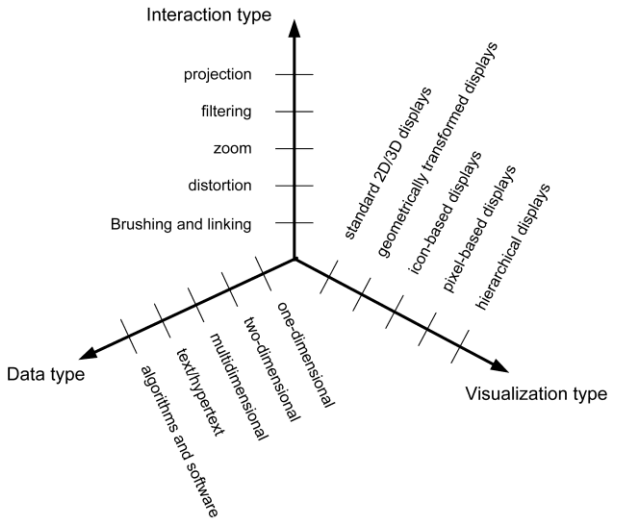

Fig 2: Three-dimensional Visualization space

A general classification of visualization methods is presented in [11] and [12]. It constructs a 3D visualization space by classifying the visualization methods according to three orthogonal criteria, the data type, the type of the visualization technique, and the interaction methods, as shown in Fig. 2.

3.4. Visual data mining

The vision of a visual data mining system stems from the following principles: simplicity, user autonomy, Reliability, reusability, availability and security. A complete visual data mining system must include security measures to protect the data, the newly discovered knowledge, and the user's identity because of various social issues [9]. A stronger visual data mining strategy lies in tightly coupling the visualizations and analytical processes into one data mining tool. Visual data mining has several advantages over the automatic data mining methods. It leads to a faster result with a higher degree of human confidence in the findings, because it is intuitive and requires less understanding of complex mathematical and computational background than automatic data mining.

The visual data mining process starts by forming the criteria about which visualizations to choose and which attributes to display. These criteria are formulated according to the exploration task. The user recognizes patterns in open visualizations and selects a subset of items s/he is interested in. The result of this selection is a restriction of the search space, which may show new patterns to the user, some of which s/he might not have been aware of before.

The whole process can then be repeated on the selected subset of data items. Alternatively, new visualizations can be added. The process continues until the user is satisfied with the result, which represents a solution to her/his initial problem. The user has full control over the exploration by interacting with the visualizations [13]. Visual data mining has been used in a number of scientific disciplines. Some recent examples include detecting telephone call frauds by a combination of directed graph drawings and bar plots, a classifier based on a parallel coordinate plot, and a visual mining approach by applying 3D parallel histograms to temporal medical data.

3.5. Combining automatic and visual data mining

The efficient extraction of hidden information requires skilled application of complex algorithms and visualization tools, which must be applied in an intelligent and thoughtful manner based on intermediate results and background knowledge. The whole KDD process is therefore difficult to automate, as it requires high-level intelligence. By merging automatic and visual mining, the flexibility, creativity, and knowledge of a person are combined with the storage capacity and computational power of the computer. A combination of both automatic and visual mining in one system permits a faster and more effective KDD process [14].

4. The approach

In this section, we describe the approach to discover knowledge related to user's activities during online bill submission, which can be used by analyst to produce new strategies to identify the user's behavior. In particular, we have devised a new symbolic data visualization strategy, which is used within a KDD process to graphically highlight behavioral patterns and other previously unknown aspects related to the user's activity in online bill submission. As we know from the literature, KDD refers to the process of discovering useful knowledge from data [14].

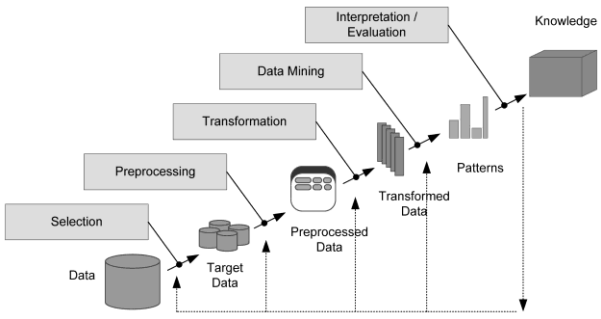

Fig 3: The steps of a KDD process

The Various steps in KDD Process are:

- Data preparation,
- Data selection,
- Data cleaning, and
- Interpretation of results.

These steps are essential to ensure that useful knowledge is extracted from data.

4.1. Data collection

In our approach, we have aimed at gathering data concerning mouse movements. In particular, the data schema is organized in online bill sessions. Each session is a page containing a sequence of text fields, with each text field denoting a category of bill. The following information is relevant for a page visit:

- Duration of the visit,
- Presence and duration of inactivity time intervals (no interactions) during the visit,
- Sequence of responses given by the user during the visit, and
- Estimation of the time spent by the user in filling the bill.

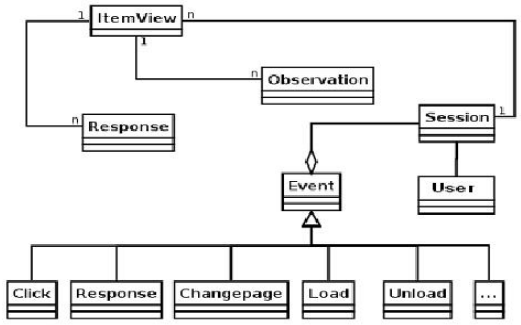

Fig 4: Conceptual schema of the data collected during an online bill session.

These data are organized according to the schema depicted in Fig. 4. In particular, for each online bill session, the user's activity is represented by a set of Item View elements, which in turn are associated to a set of Response and Observation objects. Each Item View represents the visit of the user at the web page containing a given category of bill. Usually, during a session, there is at least one Item View per item, but in some cases, the learner can skip or revise one or more bill types. Each Item View embodies a set of Observation elements, each of them representing the activity of item evaluation by the user. With the term evaluation, we mean the mouse activities of the user on the web page containing the complete bill and the set of categorical bills. For each observation, we store the type of user interactions (i.e., the kind of mouse movements) and the Duration of that observation. Finally, in each Item View, the user can give one or more responses, and for each of them, we store its correctness and time stamp.

The events captured during the user's interaction are the following:

- Events that occurred in the browser window (Open, close, resize, load, and unload) and
- Events that occurred in the browser client area (key pressing, scrolling, mouse movements, and clicks).

5. T he system

In this section, we describe the system implementing the proposed approach. The system is composed of an **Interface Framework** and a **Log Analyzer** application. The former, based on the AJAX technology [17], captures and logs all of the user's interactions with the online bill system interface (running in the Web browser). It can be instantiated in any bill submission system and is further composed of a client-side and a server-side module. The latter is a stand-alone application that analyzes the logs in order to extract information from them and to graphically represent it.

5.1. T he interface framework

The purpose of the Interface Framework is to gather all of the user actions while browsing web page of the bill and to store raw information in a set of log files in XML format. The framework is composed of a client-side and a server-side module. The former module is responsible for "being aware" of the behavior of the user while s/he is browsing the web page and for sending information related to the captured events to the server-side module. The latter receives the data from the client and creates and stores log files to the disk. Despite the required interactivity level, due to the availability of AJAX, it has been possible to implement the client-side module of our framework without developing plug-ins or external modules for Web browsers. JavaScript has been used on the client to capture user interactions, and the communication between the client and the server has been implemented through AJAX method calls. The client-side scripts are added to the online bill system web page. The event data is gathered on the Web browser and sent to the server at regular intervals. It is worth noting that the presence of the JavaScript modules for capturing events does not prevent other scripts loaded in the page to run properly. The server-side module has been

Implemented as a Java Servlet, which receives the data from the client and keeps them in an XML document that is written to the disk when the user submits the bill. The Interface Framework can be instantiated in the online bill system and then enabled through the configuration.

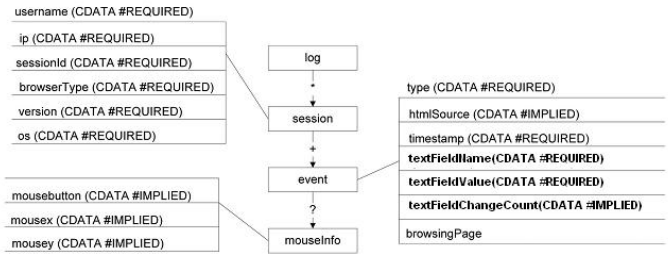

Fig 5: The information Model for log data.

The information Model used for the log data is quite simple and is shown in Fig. 5. The information is organized based on single session for every user. At this level, the username (if available),IP address of the user, session identifier, and agent information (browser type, version, and operating system) are logged. A session element contains a list of event elements. The data about user interactions are the following:

- Event type,
- Html source object involved in the event (if present),
- Mouse information (pressed button and coordinates),
- Timing information (time stamp of the event), and
- More information specific of the event, i.e., for a "response" event.

An important concern in logging is the log size. If an experiment involves a large set of users and the bill is composed of many fields, log files can reach big sizes. A configuration system, including the following configuration settings, has been conceived in order to reduce log sizes:

- List of events to capture,
- Subset of attributes for each event,
- Sections of the web pages (divs or table cells) to be monitored as event sources,
- Time interval between two data transmissions from the client to the server, and
- Sensitivity for mouse movements (short movements are not captured).

On the client side, everything can be done in the Web browser. The JavaScript modules for event capturing are dynamically generated on the server according to the configuration settings, are Downloaded, and run in the browser interpreter. Data are sent to the server through an AJAX request. On the server side, a module called Request Handler receives the data and sends them to a module called Interface Handler, which organizes the XML document in memory and flushes it to the disk every time a user submits the bill. The architecture of the framework is graphically represented in Fig. 6.

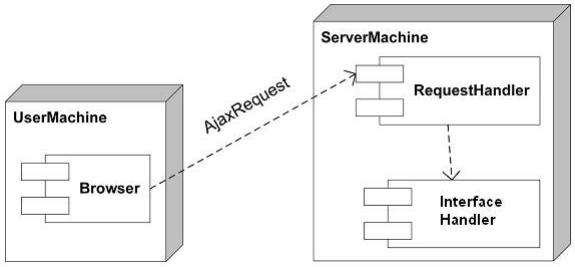

Fig 6: The Interface framework architecture.

5.2. The log analyzer application

The data analysis and test visualization operations are performed by a Web-based stand-alone application, optionally hosted on the same server of the Interface Framework, which takes as input the XML log files. The Log Analyzer Application is composed of two modules: the Query Engine and the Chart Generator. The Query Engine module performs queries on the log files in order to obtain the desired data as an instance of the data model shown in Fig. 5. Once the data in the format shown in Fig. 5 are obtained, these are given as input to the Chart Generator module. This module has been implemented through a Java Servlet, which takes the identifier of the user whose behaviour is going to be analyzed as a parameter. The module dynamically generates the chart and returns it in the PNG format. To produce the charts, the Java Applet has been used. The choice of dynamically constructing charts reduces the space necessary for the storage of the images. These charts are very much helpful in comparing the mind-sets of various users during bill submission.

Then, the Log Analyzer has been used for analyzing the logs in order to extract information from them and to graphically represent it in order to trigger a visual data mining process where the analyst plays a central role. In the case of the mentioned experiments, the visual analysis of charts enables the analyst to infer interesting conclusions about both the strategies the users used to complete bill submission and the correlation between various categorical bills. In

the former case, the objective was not only to understand the user's strategies but will also check for the validity of the bills. On the other hand, categorical bill correlation analysis aims to improve the final correctness of the bill.

6. Experimental results

In order to demonstrate the effectiveness of the approach and of the proposed system, we have used them in the context of the AICTE bill submission for seminars/workshops, where the users actually spend fewer budgets than the required, but submit a fraudulent bill by making changes to various amounts required for successful conduction of the event. In our experiment, the User-Interface module is enabled and then the user fills the bill by modifying various fields multiple times and when the user submits the form, then the logger is enabled, and generates an approximately 4-Mbyte-sized XML log file. The logging activity produced no visible system performance degradation. Then, the Log Analyzer has been used for analyzing the logs in order to extract information from them and to graphically represent it in order to trigger a visual data mining process where the analyst plays a central role.

In the case of the mentioned experiments, the visual analysis of charts enabled the analyst to infer interesting conclusions about both the strategies the users used to submits the bill and the correlation between various categorical bills. Correlation analysis between various categorical bills helps to improve the correctness of the bill.

6.1. Strategies for executing tests

Some of the experiments found in literature, aiming to analyze users behavior during online bill submission, have regarded the monitoring of the users habit to verify the filled values and to change them with other more possible values. These experiments try to answer the following two questions:

- Is it better for users to trust their first impression or to go back for evaluating the fields again and, eventually, for changing the filled values?
- Are the honest and dishonest users changing values more frequently?

In our experiment, the text field values are stored every time modification is done. Multiple users are correlated based on the time taken by them to submit the same bill. From the Session time, the analyst can predict that there may be more modifications done by the users. In particular, the frequency of several habits are recorded and then correlated to the user's behaviour, such as the habit of making changes frequently to the field values. We have found no study in the literature that considers the user behaviour as a whole. By using our approach, we have replicated several experiments performed for traditional bill submission in an online fashion. In particular, we have analyzed the effects of response changing with respect to various categorical bills. More importantly, we have tried to put these aspects together, obtaining a more general analysis of the strategies used by the users to submit the bills. As for response changing, we have obtained the average change tendency of multiple users and their correlation with the session time.

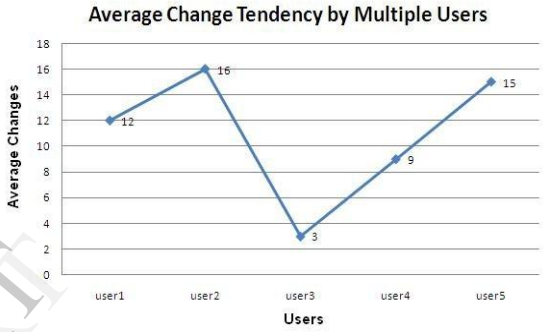

Fig 7: Average Changes done by multiple users

from the analysis of the charts obtained through our experiment, the following strategies have been discovered:

- **Idle Phase:** It is the total time that the user remains idle while filling an online bill.
- **Active Phase:** It is the total time taken by a single user to actively fill the bill online.
- **Passive Phase:** It is the total time taken by the single user to recheck the filled values.

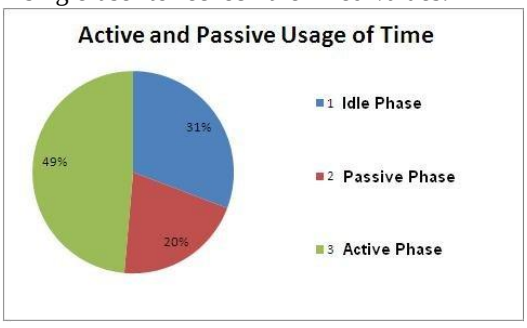

Fig 8: Strategies Usage

By visually analyzing the data of our experiment (Fig. 8), it came out that the most frequently adopted strategy is Active-phase (49%), followed by the Passive-phase (20%) and by the Idle- Phase(31%).

6.2. Detection of correlation among text field values

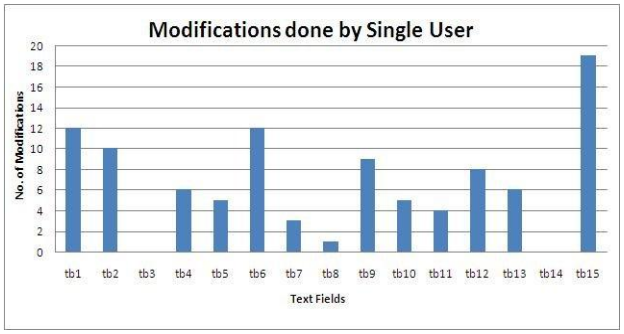

Fig 9: Modification done by Single User

By visually inspecting the charts, we can also be assisted in the detection of correlated text fields, where the chart (fig. 9) indicates the modifications done by a single user. The occurrence of such a pattern says that while the user is browsing the current text field s/he could infer the right amount by a previous text field value and makes a modification.

6.3. Detection of cheating

As proven by several studies in the education field, many users cheat while submitting online bills [19], [20]. Cheating detection in online bill submission is not an easy task. Most of the techniques employed so far have been based on the comparison of the results obtained in the bill. These techniques cannot give the certainty of the guilt, since a high similarity of two bills can be due to coincidence. Furthermore, as in all fraud detection systems, the task is complicated by several technological and methodological problems [18]. It could be useful to gain information on the users' behavior during the bill submission. Analysis on these data can be integrated to results comparison in order to have a more comprehensive data set as input for a data mining process aiming at detecting cheating. The tracking fraudulent bills information, available through the charts of our system, could be useful to prove that the user has cheated, by manipulating various categorical bills.

7. Conclusion

We have presented an approach and a system to let the analyst monitor users' strategies during online billing. The approach exploits data visualization to draw the data characterizing the users' billing strategy, in order to trigger the analyst's attention and to let him/her discover previously unknown behavioral patterns of the users' and conceptual relationships among filled text values. In this way, the analyst is provided With a powerful tool that lets him/her review the whole assessment process and evaluate possible improvements. We have extensively used the implemented system experimentally to evaluate online billing strategies. This lets us discover several relevant patterns regarding the fraudulent bills, the characteristics of used strategies, and the impact on the final bill. This approach can be further extended to provide security to many e- commerce applications in future. Finally, this can be explored the problem of cheating detection, since we believe that an approach based on logging and visualization can be promising and effective.

References

[1] D. Hand. H. Mannila. And P. Smyth, Principles of Data Mining, Adaptive Computation and Machine Learning Series, A Bradford Book, MIT Press, 2001.

[2] U. Fayyad and G. Grinstein, "Introduction," Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2002.

[3] G. Grinstein and M. Ward, "Introduction to Data Visualization," Information Visualization in Data Mining and Knowledge Discovery, Morgan Kaufmann, 2002

[4] D.A. Keim and M. Ward. "Visualization," Intelligent Data Analysis, M. Berthold and D.J. Hand, eds., second ed. Springer, 2003.

[5] D.A. Keim, Visual Exploration of Large Data Sets, second ed. Springer. 2003.

[6] I. Kopanakis and B. Theodoulidis. "Visual Data Mining Modeling Techniques for the Visualization of Mining Outcomes, " J. Visua Languages and Computing, no. 14, pp. 543-589. 2003.

[7] P. Buono and M. Costabile, "Visualizing Association Rules in a Frame work for Visual Data Mining," From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments, Essays Dedicated to Erich J. Neuhold on the Occasion of His 65th Birthday, pp. 221-231, Springer, 2005.

[8] Gennaro Costagliola, Massimiliano Giordano and Giuseppe Polese, "Monitoring Online Tests through Data Visualization " IEEE transactions on knowledge and data engineering, vol. 21,no.6, june 2009.

[9] H. Shao, H. Zhao, and G-R Chang, "Applying Data Mining to Detect Fraud Behavior in Customs Declaration," Proc. Int'l Conf. Machine Learning and Cybernetics (ICMLC 02), vol. 3, pp. 1241-1244, 2010.

[10] M.C. Chen, J.R. Anderson, and M.H. Sohn, "What Can a Mouse Cursor Tell Us More?: Correlation of Eye/Mouse Movements on Web Browsing," Proc. CHI 01 Extended Abstracts on Human Factors in Computing System. PP. 281-282. 2011.