Enhancing Visualization with Vizster in Knowledge Extraction Scenario

S. A. Ghadling
PG Student, DYPCOE, Akurdi, Department of Computer Engineering, University of Pune, India
M. A. Potey
HOD, DYPCOE, Akurdi, Department of Computer Engineering University of Pune, Maharashtra, India

Abstract:
To integrate visualization and data mining techniques like knowledge extraction toward a mixed-initiative knowledge discovery, it is needed to survey and reflect on the various ways which takes the best of human and machine capabilities. It categorizes the observed techniques in classes following a bottom-up bibliographic research approach, highlighting current trends, gaps and potential future directions for research work. In particular, it looks at strengths and weaknesses of data mining and information visualization (InfoVis), and for which purposes researchers in InfoVis use data mining techniques and also investigates how researchers in data mining employ InfoVis techniques. This information is further used to analyze the discovery process by comparing the analysis steps from the perspective of data mining and information visualization. The comparison sheds light on new perspectives on how mining and visualization can best employ human and machine skills.

Keywords: Knowledge Extraction, Visualization, Visual Data Mining, Visual Analytics

1. Introduction

Knowledge Discovery in Databases (KDD) is an automatic analysis, exploratory and modeling of large size data repositories. The KDD is an organized process of identifying novel, valid, useful, and understandable patterns from large and complex data sets. Data Mining (DM) is the core step of the KDD process that involves the inferring of algorithms that first explore the data, then develop the model and discover previously unknown patterns. This model is used for understanding phenomena from the data prediction and analysis. While information visualization (InfoVis) targets the visual representation of large-scale data collections to help people understand and analyze information, data mining, on the other hand, aims at extracting hidden patterns and models from data, automatically or semi-automatically. InfoVis can be seen as a human-centered approach to knowledge discovery in its most extreme representation, whereas data mining is a generally purely machine-driven approach that uses computational tools to extract models or patterns out of data.

Interactive Machine Learning[1][2] is an area of research where the integration of human and machine capabilities is advocated, beyond scope of visual data analysis, to build better computational models out of data. It suggests and promotes an approach where the user can interactively influence the decisions taken by learning algorithms and make refinements where needed. Visual analytics is an outgrowth of InfoVis and focuses on analytical reasoning facilitated by interactive visual interfaces. Often, it is presented as being the combination of InfoVis techniques with data mining capabilities to make it more powerful and interactive. Visual analytics is more than just visualization and can rather be seen as an integrated approach combining visualization, data analysis and human factors. Previously, it is not clear how this human- machine integration should happen.

The analysis permitted to categorize the observed techniques in classes. For each class a description of the main observed patterns areprovided followed by a discussion of potential extensions which are deem feasible and important to realize. The analysis is then followed by a comparison of the analytical processes as they happen in data mining and in visualization. This comparison, together with the knowledge gained in the literature review, permits to clarify some commonalities and differences between the automatic and visual approaches. This kind of reasoning can help framing the problem of automatic and interactive analysis and better understand the role of human and machine.

2. Related Work

The analysis is started with a literature review in order to ground the reasoning on observed facts and limit the degree of subjectivity. It is followed by mixed approach in which bottom-up and top-down analysis have been mixed to let the data speak for themselves and suggest new ideas or use the literature to investigate the assumptions or formulated hypotheses. Various dimensions are used in order to classify visualization: whether it is interactive or not, the knowledge discovery step it supports, the major mining and visualization techniques used, etc.

2.1. Pure Visualization (VIS)
It contains techniques based exclusively on visualizationwithout any type of algorithmic support. Since the focus is on how visualization and mining can cooperate in knowledge discovery, the VIS category of pure visualization techniques is not taken into account.

2.2. Computationally Enhanced Visualization (V++)
It contains techniques which are fundamentally visual but contain some form of automatic computation to support the visualization. This category pertains two techniques in which visualization is the primary data analysis means and automatic computation (that is the '++' in the name) provides additional features to make the tool more effective. In other words, when the '++' part is removed the technique becomes a "pure" visualization technique.

2.3. Visually Enhanced Mining (M++)
It contains techniques in which automatic data mining algorithms are the primary data analysis means and visualization provides support in understanding and validating the result. This category pertains two techniques in which data mining is the primary data analysis means and visualization (that is the '++' in the name) provides an advanced interactive interface to present the results. In other words, when the '++' part is removed it becomes a 'pure' data mining technique.

2.4. Integrated Visualization and Mining (VM)
It contains techniques in which visualization and mining are integrated in a way that it's not possible to distinguish a predominant role of any of the two in the process. This category combines visualization and mining approaches. None of them predominate the other and ideally they are combined in a synergic way.

3. Implementation Details
The goal of the system is to extract the knowledge from the database and visualizing it using visualization tool by enhancing the results with respect to some features. Keel tool provides us various options like import, export, visualize, edit and partitioning. Import option converts any type of data into the keel format. Export option converts data from keel format to any other requested format. Edit option provides us facility to edit the data according to specifications. Partitioning data allows us to partition the data as per the requirement. Visualize option provides us facility to compare any two attributes and visualize it in the form of graph. These results can be saved into various formats so that it can be used for further processing.

The goal of Vizster was to build a visualization system that end-users of social networking services could use to facilitate discovery and increased awareness of their online community [3]. We wanted to support the exploratory and playful aspects of Friendster while giving users easier access to search and group patterns. While users regularly explored the network on Friendster, the linear format limited such explorations. This led us to develop richer network views and exploratory tools, while maintaining a local orientation. It is also learned that the use of imagery was indispensable for identifying people and establishing a presentation of self, and so must play a central role in the visualization.

Vizster presents social networks using a familiar node-link representation, where nodes represent members of the system and links represent the articulated "friendship" links between them. In this view, network members are presented using both their self-provided name and, if available, a representative photograph or image. The networks are presented as egocentric networks: networks consisting of an individual and their immediate friends. Users can expand the display by selecting nodes to make visible others immediate friends as well[3].

3.1. System Architecture
The Fig.1 shows the basic architecture of the system. Data from the database will be pre-processed and given for knowledge extraction tool. Extracted knowledge will get post- processed and submitted for visualization process, which can be further enhanced by some features.
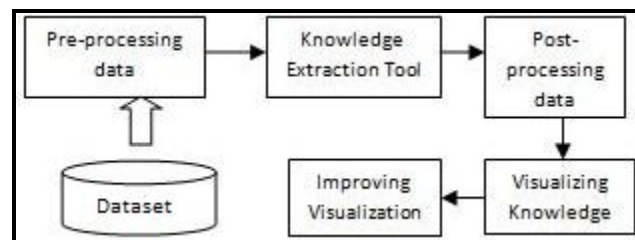

Figure 1: System Architecture

4. Results and Discussion

4.1. Datasets Used
Knowledge is extracted from friedman dataset using keel tool and given for visualization which is done by vizstertool. The results are shown as in figures. The experiment is also performed on some other datasets likepima, clevelanddatasets.

4.2. Result sets

The Fig.2 shows the extracted knowledge from Keel tool, where single attribute is available on each axis. Only one attribute will be the nominal and other is not nominal. Here, interested-in attribute is shown on X-axis and uid attribute is shown on Y-axis.
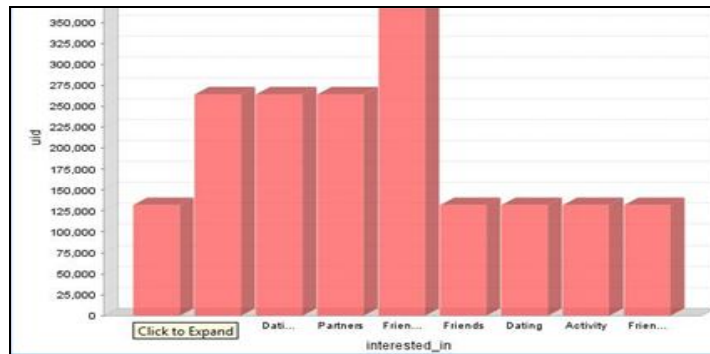

Figure 2: Extracted knowledge from database considering interested-in attribute on X-axis and uid attribute on Y-axis

Fig.2. Extracted knowledge from database considering interested-in attribute on X-axis and uid attribute on Y-axis

The Fig.3 shows the visualization of the extracted knowledge using Vizster tool. As per the selected attribute, nodes can be highlighted with different colors with respect to values assigned to the attribute.
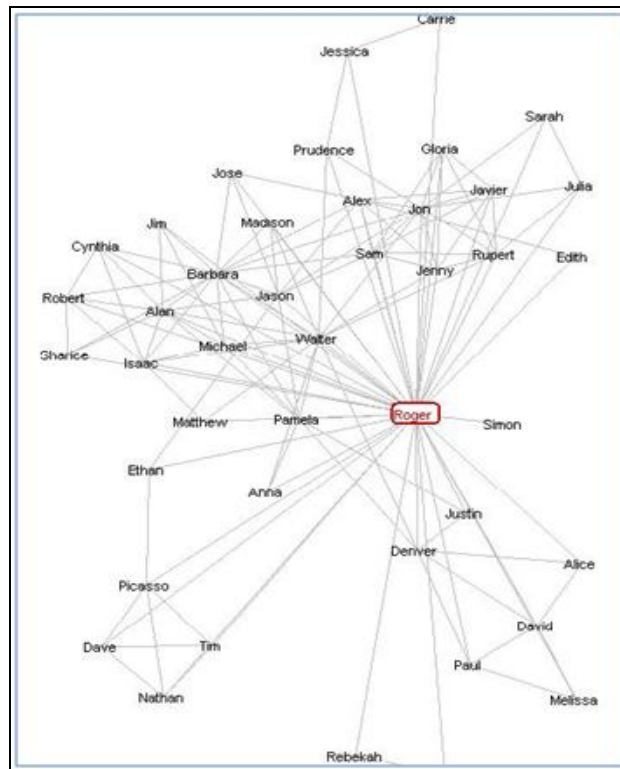

Figure 3: Visualization of the all data using visualization tool

The Fig.4 shows the visualization of the extracted knowledge using Vizster tool. Different communities can be shown as per the community status and common features between different nodes.
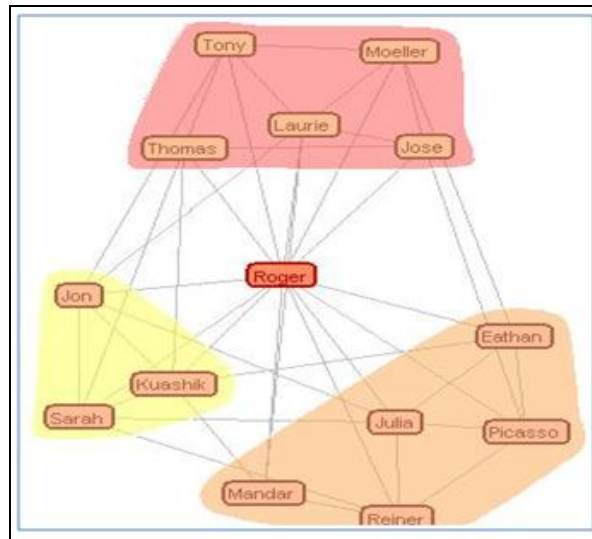
Figure 4: Visualization of the extracted knowledge using visualization tool

4.3. Comparative Analysis

Existing system was not able to convert the dataset from keel format to the vizster format as source and target nodes were not getting created. As compared to previous system, this system provides us facility to create source and target nodes as per extracted knowledge. The Fig.5 shows the extracted knowledge from Keel tool, where dataset doesn't contain Null attribute. Only specified attributes are shown in the graph.
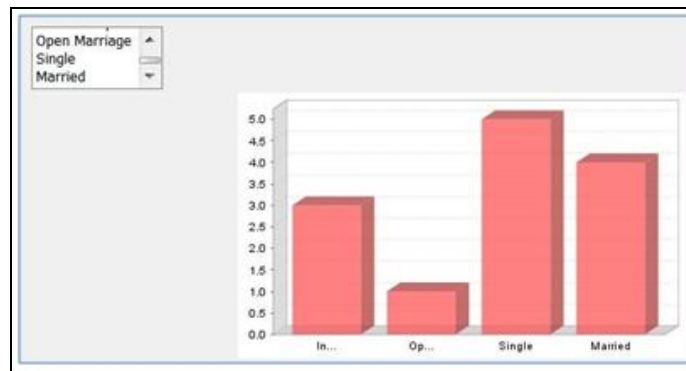

Figure 5: Extracted knowledge from database without considering null attribute

The Fig.6 shows the extracted knowledge from Keel tool, where dataset may contain few Null attributes. So, this graph contains null attribute as separate entity, which has its own count.
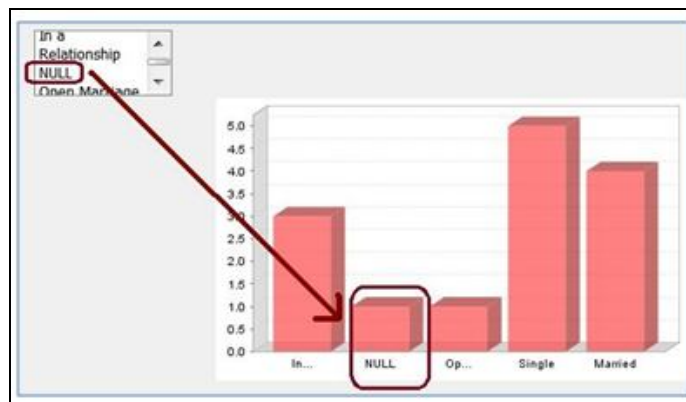

Figure 6: Extracted knowledge from database considering null attribute

5. Conclusion and Future Work

In this stage, knowledge extraction tool like Keel and design of Vizster are described. Keel is the tool which is used to import from any other format to keel format, export from keel format to any other format, visualize in graphical form, edit the data along with making partition. The output of keel is given to Vizster system for visualization. Vizster is a system for end-user visual exploration of online social networks. Grounded in an ethnographic investigation of the Friendster social networking service, the resulting Vizster system is an experiment in applying the fruits of information visualization within an end-user context more characteristic of play than analysis. As such, Vizster contributes a number of network analysis techniques designed to both reveal information and maintain a playful social environment. These features include connectivity highlighting and linkage views for users to explore greater network contexts, X-ray mode and profile search for the visual analysis of member profile data, and community analysis for investigating higher-level community structures. Observations of usage found that users find these features both useful and enjoyable for exploring their business.

In the next stage, knowledge will be visualized with enhancedVizster feature which includes the multiple attribute selection for highlighting the nodes for extracted knowledge. Keel will be able to handle the data with missing values as extra enhancement.

6. References

1. J. A. Fails and J. Olsen, "Interactive machine learning", IUI '03: Proceedings of the 8th international conference on intelligent user interfaces, New York, NY, USA: ACM, 2003, pp. 39-45.
2. M. Ware, E. Frank, G. Holmes, M. Hall, and I. H. Witten, "Interactive machine learning: letting users build classifiers", International Journal of Human Computer Studies, vol. 55, 2001, pp. 281-292.
3. J. Heer and D. Boyd, "Vizster: Visualizing online social networks", Proceedings of the 2005 IEEE Symposium on Information Visualization, 2005, pp. 33-40.
4. S. T. Teoh and K. Ma, "PaintingClass: interactive construction, visualization and exploration of decision trees", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C.: ACM, 2003, pp. 667-672.
5. M. Ankerst, C. Elsen, M. Ester, and H. Kriegel, "Visual classification: an interactive approach to decision tree construction", Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 1999, pp. 392-396.
6. D. Yang, Z. Xie, E. A. Rundensteiner, and M. O. Ward, "Managing discoveries in the visual analytics process", SIGKDD Explor. Newsl.,vol. 9, 2007, pp. 22-29.
7. G. Ellis and A. Dix, "Density control through random sampling: an architectural perspective", Information Visualization, IV 2002. 2002, pp. 82-90.
8. R.A. Amar, "Knowledge Precepts for Design and Evaluation of Information Visualizations", IEEE Transactions on Visualization and Computer Graphics, vol. 11, 2005, pp. 432-442.
9. C. Plaisant, J. Fekete, and G. Grinstein, "Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository", Visualization and Computer Graphics, IEEE Transactions on, vol. 14, 2008, pp. 120-134.
10. M. Ankerst, M. Ester, and H. Kriegel, "Towards an effective cooperation of the user and the computer for classification", Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2000, pp. 179-188.
11. Ghoniem, M., J.-D. Fekete, P. Castagliola. "A Comparison of the Readability of Graphs Using Node-Link and Matrix-Based Representations",InfoVis 2004, Austin, TX, 17-24.
12. Heer, J., S.K. Card, J.A. Landay. "Prefuse: A Toolkit for Interactive Information Visualization", CHI 2005, Portland, OR, 421-430.
13. Nardi, B., S. Whittaker, E. Isaacs, M. Creech, J. Johnson, J. Hainsworth. "ContactMap: Integrating Communication and Information through Visualizing Personal Social Networks", Communications of the ACM, April 2002.
14. Newman, M.E.J. "Fast algorithm for detecting community structure in networks", Physical Review E, 69, 2004, 066-133.
15. Viegas, F., d. boyd, D. Nguyen, J. Potter, and J. Donath. "Digital Artifactsfor Remembering and Storytelling: PostHistory and Social Network Fragments", HICSS-37, Big Island, HI, 2002.
16. Shneiderman, B. "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations", 1996 IEEE Conference on Visual Languages, Boulder, CO, 336-343