A Classification Scheme for Characterizing Visual Mining

Elaheh Mozaffari and Sudhir Mudur

Concordia University, Computer Science & Software Engineering,
1515 St. Catherine St. West, Montreal, Quebec, Canada
{e_mozafa,mudur}@cs.concordia.ca

Abstract. Visual mining refers to the cognitive process which integrates the human in analysis of information when using interactive visualization systems. This paper presents a classification scheme which provides user-centered representation of goals and actions that a user performs during the visual mining process. The classification scheme has been developed using content-analysis of published literature containing precise descriptions of different visual mining tasks in multiple fields of study. There were two stages in the development. First, we defined all the sub-processes of visual mining process. Then we used these sub-processes as a template to develop the initial coding scheme prior to utilizing specific data from each of the publications. As analysis proceeded, additional codes were developed and the initial coding scheme was refined. The results of the analysis were represented in the form of a classification scheme of the visual mining process. The naturalistic methods recommended by Lincoln and Guba have been applied to ensure that the content analysis is credible, transferable, dependable and confirmable.

Keywords: Visual mining, large dataset analysis, human information behaviour.

1 Introduction

In today's applications, data are becoming available in large quantities. Fields as diverse as bioinformatics, geophysics, astronomy, medicine, engineering, meteorology and particle physics are faced with the problems of making sense out of exponentially increasing volumes of available data [1]. Therefore, one of our greatest challenges is to take advantage of this flood of information and turn raw data into information that is more easy to grasp and analyze.

Over the years, a large number of interactive visualization systems have been developed, all claiming to help users analyze, understand and gain insight into the large quantity of available data through appropriate transformations of the raw data into visual representations. We refer to the human analytical process that uses such visually represented information as being the Visual Mining (VM) process. It concerns the cognitive process which integrates the human factor during the course of mining and analyzing information through the visual medium. It contributes to the visual discovery of patterns which form the knowledge required for informed decision making.

G. Salvendy, M.J. Smith (Eds.): Human Interface, Part II, HCII 2011, LNCS 6772, pp. 46–54, 2011.
© Springer-Verlag Berlin Heidelberg 2011

Purely from a technology perspective there are many studies which have focused on techniques and tools for building up visualization systems. Also, the importance of understanding users' workflows and practices has been recognized by many researches [2-4] . Jim Gray points out that without integration of users' workflows and interactions with the information, even the best system will fail to gain widespread use [5]. However, there are to date no reports on studies from the perspective of user behavior in visual mining of large data sets.

To understand users of large datasets while performing visual mining, studies about users' information behaviours are critical. Information behavior is defined as: "The totality of human behavior in relation to sources and channels of information, including both active and passive information-seeking, and information use" [6]. However, as previously mentioned, studies with specific focus on scientists's information behaviour (how they look for required information and actually use them) in the visual mining process are rare. The study that is reported in this paper answers this call, and aims to improve our understanding of information behaviors of users activities during the process of visual mining of large datasets.

The rest of the paper is organized as follows. In section 2 we review the related studies and identify the problem. In section 3 we describe and justify our choice for the methodology used for addressing this problem. Section 4 includes description and justification of our method of chosing visual mining case study samples and the analysis of user information behavior in visual mining in these samples. Results of this research are presented in section in section 5. Section 6 concludes this paper and discusses potential for future work.

2 Background

Information behaviour has been the focus of many researches in the few last decades in the field of library and information science. The highlights of studies on information behaviour include Wilson's (1981) model of information-seeking behaviour [7] , Dervin's (1983) sense-making theory [8] , Ellis's (1989 and 1993) behavioural model of information seeking strategies [9,10] , Kuhlthau's (1991) model of the stages of information seeking behaviour [11], Belkin's (1993) characterization scheme of information-seeking [12] and Wilson's (1997) problem solving model [13].

The studies presented above are however inadequate with regard to their suitability for representing user's information behavior in VM process. These studies are cannot completely model user information behavior in visual mining. Their sole goal is to describe the information-seeking activity, the causes and consequences of that activity, or the relationships among stages in information-seeking behavior [14] . For example the model proposed by Belkin et al. represents dimensions of information seeking behaviors in information retrieval system. Information seeking is one of the sub-process of VM process (as we shall explain in more details in section 3) therefore it can not completely describe VM process. In addition, the studies done with library patrons focus on the user's tasks that are perhaps learned behaviors due to their prior knowledge of how libraries work. They tend to ask questions that they know can be answered. Visualizations might support a different way of asking questions and getting answers [15].

This paper presents a study that began with the aim of extending our understanding of the interdisciplinary process of visual mining, and in doing so looked to strengthen and improve our understanding of user's information behavior in the visual mining process. The study has yielded a classification scheme for characterizing visual mining. Such a classification scheme has many different applications: support requirement analysis in system engineering of interactive visualization software, studying and assessing different tasks which typically occur in visual mining, improving the functionality and interface design of newer interactive visualization systems and by providing a system-independent taxonomy of tasks, it can be used for evaluating and classifying existing interactive visual mining systems based on what they support.

3 The Sub-processes of Visual Mining

Today, many different groups around the world are undertaking research on visualization for data mining and analysis in order to effectively understand and make use of the vast amount of data being produced. Different terms are used to describe this process, visual data mining [16, 17], visual exploration [18] and visual analytics [19] to name but a few. In addition, there appears to be some variation in understanding that people have of the process even under the same term. Niggemann [20] defined visual data mining as visual representation of data close to the mental model. Ankerst [21] considered visual data mining as a step in the knowledge discovery process which utilizes visualization as a communication channel between the computer and the user. Visual analytics is defined as an iterative process that involves information gathering, data preprocessing, knowledge representation, interaction and decision-making. In order to gain further insight, it integrates methods from knowledge discovery in databases, statistics and mathematics, together with human capabilities of perception and cognition [22] . From these definitions, one common theme that can be recognized is that they all rely on the human visual channel and take advantage of human cognition. They also emphasize on the three aspects: task, visualization and process.

From the above approaches it may be noted that in VM, data visualization takes place either before or after a data mining algorithm, or perhaps even independently. For the purposes of our research, however, we will focus on the human involvement in the visual data exploration process which utilizes human capabilities to perceive, relate and conclude. We consider VM as a hypothetical formation process that primarily uses the visual medium. Such visualization allows the user to interact directly with visually represented aspects of the data, analyze, gain insight and perhaps even formulate a new hypothesis. Later on, the user can evaluate the best possible hypotheses and make a judgment based upon it. In fact, this visual information exploration process helps to form the knowledge for informed decision making.

In order to identify analysis tasks and human information interactions in the VM process, we first look at how analysis works and then extend its sub-processes to the context of VM. The analytical process itself is both structured and disciplined. Usually analysts are asked to perform several different types of tasks such as assessing, forecasting and developing options. Assessing requires the analyst to describe their understanding of the present world around them and explain the past. Forecasting requires that they estimate future capabilities, threats, vulnerabilities, and opportunities. Finally, developing options in order to establish different optional reactions to potential events and assess their effectiveness and implications. The process begins with planning. The analyst must determine how to address the issue, what resources to use, and how to allocate time to various parts of the process. Then, they must gather relevant information and evidence in order to relate them to their existing knowledge. Next, they are required to generate multiple candidate explanations in the form of hypotheses, based on the evidence and assumptions, in order to reach a judgment about which hypothesis can be considered to be the most likely. Once conclusions have been reached, the analyst broadens their way of thinking to include other explanations that were not previously considered and provide a summary of the judgments they had made [19].

Building upon the above description of the analytical process, we defined the sub-processes of VM as follows:

1. The user initiates the VM process by planning how to address the issue, what resources to use and how to allocate time to various parts of the process to meet deadlines. The next step is to gather all relevant information by seeking information through searching, browsing, monitoring and generally being aware [23].
2. Searching refers to active attempts to answer questions, look for a specific item or develop understanding around a question or topic area.

Browsing is an active yet undirected form of behavior. For example, when performing physical acts such as 3D navigation tasks or scrolling/panning, the user has no special information need or interest, but becomes actively exposed to possible new information.

Monitoring is a direct and passive behavior. The user does not feel such a pressing need to engage in an active effort to gather any information but may be motivated to take note of any expected or unexpected information. Also, when the user has a question in mind, and may not be specifically acting to find an answer, they would take note of any relevant information that appears.

Being aware is a passive undirected behavior and is similar to browsing except that the user could locate information or data unexpectedly.

3. The next step in the VM process is to relate the findings with the knowledge that is hidden in the expert's mind.
4. Based on the findings, the user then generates multiple candidate explanations in the form of hypotheses. By applying analytical reasoning the user can use their prerogative to either confirm or reject any hypothesis and formulate a judgment about which is the most relevant.
5. Once conclusions have been reached, the user will be engaged in the act of broadening their thinking to include other possible explanations that were not previously considered. Then the user summarizes analytical judgment either as assessment, estimation or evaluation of options depending on the goal.
6. As the concluding step, the user usually creates a product to include the analytical judgment in the form of reports, presentations or whatever other form of communication is deemed appropriate.

4 Method

In this section, we describe our methodology for characterizing visual mining process. But first, we provide the justification for choice of our methodology.

Surveys and interviews are the most common research methods for studying users' information behavior [24]. As we know, in a typical user study or survey, user's motivation, knowledge and expertise considerably influence user performance and thus the final conclusions. Of course, using domain experts provides more realistic results [25]. However, it is not easy to employ enough participants (domain experts) for interviews and surveys in this type of study, nor is it possible to have access to them for any extended period of time because most of the experts are distributed across different external institutions. Therefore we turned to scientific publications which, in general, clearly record the behavior of experts while being engaged in the visual mining process and equally importantly are also peer reviewed. We adopted the qualitative direct-content analysis approach [26], to reveal the visual mining behavior of scientists from such publications. Qualitative content analysis is an unobtrusive method which uses nonliving form of data, generally categorized as texts. And it is well established that one kind of text that can be used for qualitative data inquiry in content analysis is official publications [27, 28]. The advantages of working with prior published works are:

1. The data are stable and non-changing,
2. The data exists independent of research. This is because the data is not influenced through researcher's interaction as is the case with interviews. They already exist in the world regardless of the research currently being done [29] and
3. They provide information about procedures and past decisions that could not be observed by researchers [27].

We obtained the information on work practices through analysis of end-results of researches as they were described in published scientific literature. The naturalistic methods recommended by Lincoln and Guba were applied to ensure that the content analysis was (to the extent possible) credible, transferable, dependable and confirmable.

5 Case Study Samples

For our content we initially chose sixty one published papers primarily concerned with reports on effective use of visualization for analysis and mining of large datasets. The chosen papers were from four different domains, namely, medicine, bioinformatics, epidemiology and geoscience. Each paper was studied and those which did not report actual case studies by experts were excluded from further consideration. The final numbers of papers which contained cases studies that described interaction with visual information in each domain are given in Table 1. Every one of these papers was analyzed and used in the information interaction coding process described next.

Table 1. Numbers of papers analyzed in each domain for qualitative content analysis

Context	Number of papers
Bioinformatics	8
Medical	9
Geoscience	11
Epidemiology	3

6 Analysis of Case Studies

The Sub-Processes of Visual Mining explained in section 3 were used to develop the initial coding scheme prior to data analysis [26]. As analysis proceeded, additional codes were developed and the initial coding scheme was revised and refined.

Coding of the data took place in multiple iterations. (1) Initial coding of each paper began with manual annotation of paper by reading case studies line by line, to highlight each relevant concept of human interaction and label it. Subsequent iterations of reading and coding of each paper in a constant comparison with previous paper and coding allowed emergence of categories and themes. We used NVivo 9 software that helps work with unstructured information such as documents, surveys, audio, video and pictures in order to assist in better decision-making [30]. NVivo 9 allowed us to code relevant concepts of VM in the articles and assign them to nodes which can be as hierarchical (tree nodes) or un-hierarchical (free nodes) as required. The relevant concepts of visualization were first coded as free nodes. Then, after coding a few articles and comparing them with previous ones, were either modified to tree nodes, renamed or deleted as required. Coding with NVivo 9 was convenient since it allowed adding, renaming, deleting or merging of codes as required but it did not, however, automate the coding process. (2) The consistent coding was addressed by including several iterations of coding around a period of a year. (3) Peer debriefing technique was used to confirm the interpretations and coding decisions. Peer-debriefer, a disinterested observer, analyzed the research materials and questioned the data, meaning, and interpretation. She was a colleague and had a PhD in computer Science, was not involved in the study. She had knowledge about qualitative research and phenomenon under investigation. The interactions between researcher and the peer-debriefer also included in the audit trail. She also acted as the auditor. (4) The coding changes were maintained by creating static models in NVivo 9 for future reference. In addition ideas, discussions, interpretations and decisions were recorded in the memos in NVivo 9 to keep tracking of the development of analysis. These allowed an audit trail to be maintained. (5) An external auditor examined the audit trail. (6) The dynamic models illustrating code relationships were used to visualize explore and present connections and patterns in the data. (7) At the end, member checking which is most important action in a naturalistic inquiry [31] was conducted to test the result of analysis with a geographers and a research fellow in biomedical engineering. They confirmed the results and verified the interpretations.

7 Results

The above mentioned process led us to formulate a set of criteria which characterizes the VM process. Table 2 presents these criteria and their possible values as the task model for visual mining.

Table 2. Classification Scheme of VM

Criteria	Values
Goal	assess, estimate, develop options
Information seeking	searching, browsing, monitoring, being aware
Retrieval	pattern, hypothesis, judgment

In the resulting classification scheme, the user's goal of visual mining requires an understanding of the current situation and explaining the past (assess), estimating future capabilities (estimate) and developing different possible options (develop options). In order to accomplish these goals, the user must gather relevant information and evidence through active or passive information-seeking activities which, as already described, are classified as searching, browsing, monitoring and being aware. The retrieved item(s) during these activities can be a pattern, hypothesis or final analytical judgment.

Finally, in order to further validate the classification scheme, typical real-world visual mining tasks were extracted and listed from the reviewed literature. All extracted tasks were re-described using the VM classification scheme in order to validate the model. Finally to ensure that further refinement is not needed, visual interaction tasks were extracted from ten new papers all containing reports of visualization case studies. All these tasks were comprehensively described by the VM task model. This process was repeated again with an additional five papers. Since no changes were required in the classification scheme, we concluded that our final classification scheme was stable and no further refinements were needed.

8 Conclusion and Future Work

To understand users of large datasets while performing VM, studies about users' information behaviours are critical. However, studies that focus on scientists's information behaviour in the visual mining process are rare. To his end, this paper has presented a summary of a study concerned with human interactions with visually represented information which aimed to improve our understanding of information behaviors of users of large datasets. By carrying out a trustworthy qualitative content analysis procedure using published papers reporting visual information interaction tasks, we have derived that user behaviours in this context can be differentiated along a small set of three criteria. These three criteria were represented in the form of a classification scheme of the visual mining process. This classification scheme allows to describe real world visual mining tasks which play an important role in analysis of large datasets.

In our future work we plan to use these criteria in modelling user behavior through behavioral strategies, validating these strategies against known case studies in different domains and applying it in comparative evaluation of visualization systems and in the design of newer systems.

References

1. Mann, B., Williams, R., Atkinson, M., Brodlie, K., Williams, C.: Scientific Data Mining, Integration and Visualization. In: Integration, and Visualization Report of the workshop held at the e-Science Institute (2002), `http://www.nesc.ac.uk/talks/sdmiv/report.pdf`
2. Atkinson, M., De Roure, D.: Data-intensive Reseach: Making best use of research data. e-Science Institute (2009)
3. Van de Sompel, H., Lagoze, C.: All Aboard: Toward a Machine-Friendly Scholarly Communication System. In: Hey, A.J.G., Tansley, S., Tolle, K. (eds.) The Fourth Paradigm: Data-intensive Scientific Discovery, Microsoft Research: Redmong, pp. 193–199 (2009)
4. Borgman, C.L.: Scholarship in the Digital Age: Information, Infrastructure, and the Internet. MIT Press, Cambridge, MA (2007)
5. Gray, J.: Scientific Data Management in the Coming Decade. SIGMOD 34(4), 34–41 (2005)
6. Wilson, T.D.: Human information behavior. Informing Science 3, 49–55 (2000)
7. Wilson, T.D.: On user studies and information needs. Journal of Documentation 37(1), 3–15 (1981)
8. Dervin, B.: An overview of sense-making research: concepts, methods and results to date. International Communications Association Annual Meeting, Dallas, Texas (1983)
9. Ellis, D.: A behavioural approach to information retrieval design. Journal of Documentation 46, 318–338 (1989)
10. Ellis, D., Cox, D., Hall, K.: A comparison of the information seeking patterns of researchers in the physical and social sciences. Journal of Documentation 49, 356–369 (1993)
11. Kuhlthau, C.C.: Inside the search process: information seeking from the user's perspective. Journal of the American Society for Information Science 42, 361–371 (1991)
12. Belkin, N.J., Marchetti, P.G., Cool, C.: Braque: Design of an Interface to Support User Interaction in Information Retrieval. Information Processing and Management 29, 325–344 (1993)
13. Wilson, P.: Information behavior: An inter-disciplinary perspective. In: Vakkari, P., Savolainen, R., Dervin, B. (eds.) Information Seeking in Context, pp. 39–50. Taylor Graham, London (1997)
14. Wilson, T.D.: Models in information behaviour research. Journal of Documentation, 55, 249–270 (1999)
15. Morse, E. L.: Evaluation of Visual Information Browsing Displays. PhD Thesis, University of Pittsburgh (1999)

16. Keim, D.A.: Information Visualization and Visual Data Mining. IEEE Transaction on Visualization and Computer Graphics 8, 1–8 (2002)
17. Simoff, S.J., Michael, H., Böhlen, M.H., Mazeika, A.: Visual Data Mining - Theory, Techniques and Tools for Visual Analytics. Springer, Heidelberg (2008)
18. Tominski, C. Event-Based Visualization for User-Centered Visual Analysis. Ph.D. thesis, University of Rostock, Rostock, Germany (2006)
19. Thomas, J.J., Cook, K.A.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE press, New York (2005)
20. Niggemann, O.: Visual Data Mining of Graph-Based Data. Ph.D. Thesis, University of Paderborn (2001)
21. Ankerst, M.: Visual Data Mining. Dissertation (Ph.D. thesis). Faculty of Mathematics and Computer Science, University of Munich (2000)
22. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual Analytics: Scope and Challenges. LNCS. Springer, Heidelberg (2008)
23. Bates, M.J.: Toward an Integrated Model of Information Seeking and Searching. In: Fourth international Conference on Information Needs, Seeking and Use in Different Contexts, vol. 3, pp. 1–15 (2002)
24. McKechine, L.E.F., Baker, L., Greenwood, M., Julien, H.: Research method trends in human information literature. New Review of Information Behaviour Research, 3, 113–125 (2002)
25. Plaisant, C.: The challenge of information visualization evaluation. In: Proc. of the Conference on Advanced Visual Interfaces (AVI). ACM, NY (2004)
26. Kyngas, H., Vanhanen, L.: Content analysis (Finnish). Hoitotiede 11, 3–12 (1999)
27. Patton, M.Q.: Qualitative research and evaluation methods, 3rd edn. Sage Publications, Thousand Oaks (2002)
28. Bhowmick, T., Griffin, A.L., MacEachren, A.M., Kluhsmann, B., Lengerich, E.: Informing Geospatial Toolset Design: Understanding the Process of Cancer Data Exploration and Analysis. Health & Place 14, 576–607 (2008)
29. Hesse-Biber, S.N., Leavy, P.: The practice of qualitative research. Sage publications, Thousand Oaks (2006)
30. QSR International, http://www.qsrinternational.com/news_whats-new_detail.aspx?view=367
31. Lincoln, Y.S., Guba, E.G.: Naturalistic inquiry. Sage Publications, Inc., Beverly Hills (1985)