"Guided cognition" and "Validated cognition" Approaches in Visual Data Mining

Michael H. Böhlen[♥], *Simeon J. Simoff*[♠] *and Arturas Mazeika*[♥]

[♥] Faculty of Computer Science
Free University of Bozen-Bolzano, Italy
{boehlen, arturas}@inf.unibz.it

[♠] Faculty of Information Technology
University of Technology Sydney, Australia
simeon@it.uts.edu.au

Abstract

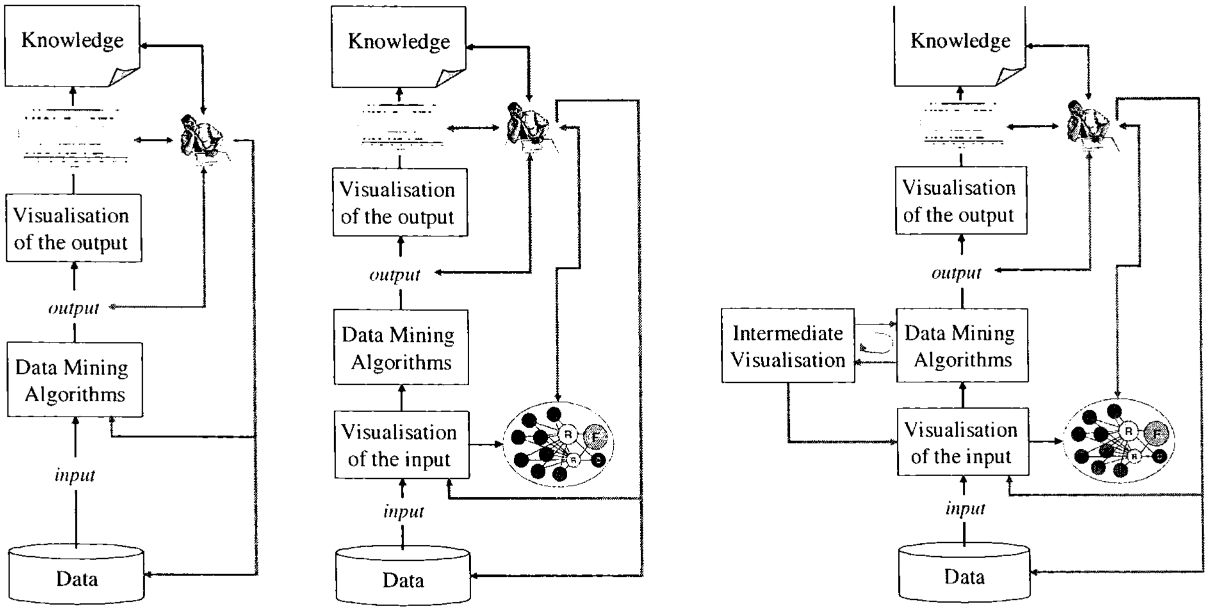
Visual data mining is a promising way of dealing with the complexity of integrated data sets of various granularity and sizes. It looks at having an access to the entire data set in its most granular level. This paper presents a view of the visual data mining as a "reflection-in-action" technique. We have illustrated two different types of this technique, namely "guided cognition" and "validated cognition". This work is motivated by the fact that visual, though very attractive, means also subjective, and non-experts are often left to utilise visualisation methods (as an understandable alternative to the highly complex statistical approaches) without the ability to understand their applicability and limitations.

1 Introduction

The rapid development of data collection and storage technologies has lead to a growth of data at a rate far exceeding that with which humans are able to analyse and make sense of it (Dunham, 2003). This data abundance and the developments in computing and display technology lead to the development of statistically meaningful visualisation techniques for presenting and interactively manipulating complex data in a consistent manner. As vision is by far the human's most powerful input information channel, many researchers in the area of data mining and computing science have worked on information visualisation methods that facilitate human understanding of data and data analysis results. Such understanding is constructed by humans by forming a mental model which captures only the essence of the phenomena in consideration. In terms of visual data mining this means that humans do not necessarily need detailed visualisation of the whole data set, but a considerably lower amount of information, generated out of that data set, which, when visualised, is sufficient in forming human perception and model of the patterns in the original data set. In other words, to be able to capture the essence of the data, visualisation techniques should "fit" human cognition, i.e. the human mental model. The consistent development of such visualisation techniques and the corresponding visual data mining models remain key research issues in the area. The first part of this paper addresses these issues. We look at visual data mining techniques that aim at guiding human cognition by abstracting and manipulating visual forms from the row data. The algorithms embed statistical procedures for composing the visual presentations in a way that ensures the "objectiveness" of observed patters. The second part of the paper looks at the ways of validating patterns that are discovered visually from raw data. As the ultimate goal of the knowledge discovery process is in gaining deeper understanding of the phenomenon, "seeing" data assists in understanding it. However, "seeing" is a subjective process and depends on the mapping of data attributes to the dimensions of the visual presentation. An issue in visual data mining is how to validate the observed pattern. For example, an obvious clustering pattern visible in 2D or 3D projections of the data points may not necessarily hold in other projections, or when considering all projections, and yet, it may hold the key to understanding the structure of the data. The second part of the paper is devoted to this issue.

1.1 Visual data mining as a knowledge discovery methodology: scenarios and issues

Visual data mining is seen as an approach that integrates the exploration and pattern spotting abilities of the human mind with the processing power of computers to form a powerful knowledge discovery environment that is supposed to capitalise on the best of both worlds (Wong, 1999). The methodology is based on both functionality that characterises structures and displays data and human capabilities that perceive patterns, anomalies, relationships, and tendencies. The utilisation of visual communication between computer and the human for depicting novel and interpretable patterns in data has been emphasised in the definitions of visual data mining. (Ankerst, 2000) summarised three ways of embedding visualisation in the data mining process, which have been extended in (Noirhomme-Fraiture *et al.*, 2002). Here we revisit and extend these views, presented in Figure 1. In the first case, visualisation is applied only to the output (results) of the analytical data mining algorithms. The idea is illustrated in Figure 1a. In this case such patterns are visualised to assist interpretation (for example, depending on human cognitive style, associations may be easier for comprehension through a visualisation of association rules (Hofmann *et al.*, 2000), rather than through a list of rules sorted by confidence, support, lift and other measures of coverage and "interestingness"). Based on the results of the interpretation, the human may go back to the data mining algorithms and either rerun the one used before with a change in parameters, or use different algorithm and, if available, its corresponding visualisation. The human also has the choice to revisit the data.


Knowledge
Visualisation of the output
output
Data Mining Algorithms
input
Data
Knowledge
Visualisation of the output
output
Data Mining Algorithms
Visualisation of the input
input
Data
Knowledge
Visualisation of the output
output
Intermediate Visualisation
Data Mining Algorithms
Visualisation of the input
input
Data

a. Visualising results b. Visualising input data and results c. Visual data mining that includes intermediate visualisation.

Figure 1. Current paradigms in visual data mining.

The second scenario, presented in Figure 1b looks at including in the loop the visualisation of the raw data input. This is perhaps closer to an instant visual data mining process, as the human can have a full picture of the data space and perform investigation without any assumptions or preconceptions. Depending on the case the human may interpret the visual patterns and the analysis may be completed without running additional data mining algorithms. When for explorative analysis this step may be sufficient, predictive and classification tasks require backup with analytical methods, hence the rest of the steps shown in Figure 1b. On the other hand, though intuitive and attractive, displaying the raw data may not necessarily lead to successful results for number of reasons. Due to the limits of visual display, we usually are interacting with 2D or 3D projection(s) of the multidimensional data set. Such projections may not necessarily present a cross-section that can click our cognition, as the example, shown in Figure 2. This is an important issue in utilisation of raw data displays for visual data mining. In this paper we present and discuss the *guided cognition* approach towards the data exploration, arguing that the display of raw data should be equipped with algorithms that assist our perception system to depict forms and structure. Such procedures include *algorithms for abstracting forms from raw data* and *algorithms for manipulating feature values* to assist in depicting visual patterns. When the algorithms in the first approach use statistical characteristics to ensure that the form abstraction is accurately visualising the structure of the data, the second approach requires additional validation techniques. Therefore we have labelled this approach as *validated cognition*. Such approach relies on the integrated interactive mining procedures, where the visual data mining techniques are tightly integrated with other procedures that assist in validating the results. This scenario will require a number of intermediate visualisation steps, as illustrated in Figure 1c. This includes, but is not limited to the "visualisation of intermediate result" as described in (Ankerst, 2000), where the visualisation was restricted only to the output of the data mining algorithms. This scenario is closer the visual classification techniques, where a 2D visual representation is used to support an interactive approach to decision tree construction. This representation actually involves the human in the data mining process and the result is a decision tree constructed by the perception of the human (whether expert or not) or additionally with the aid of domain knowledge by an expert (Wong, 1999). Such approach fits the "guided cognition" methodology. Noted first in (Wong, 1999), and, unfortunately, still valid (though there has been some progress), is the fact that in the present state of the art visual data mining support analytical data mining techniques are loosely coupled with available visualisation schemata. The interaction between visualisation and data mining algorithms is in broader sense than visualisation of algorithmic decisions in the tightly integrated visualisation, discussed in (Ankerst *et al.*, 2002), however, the detailed discussion of this issue is beyond the scope of this paper.

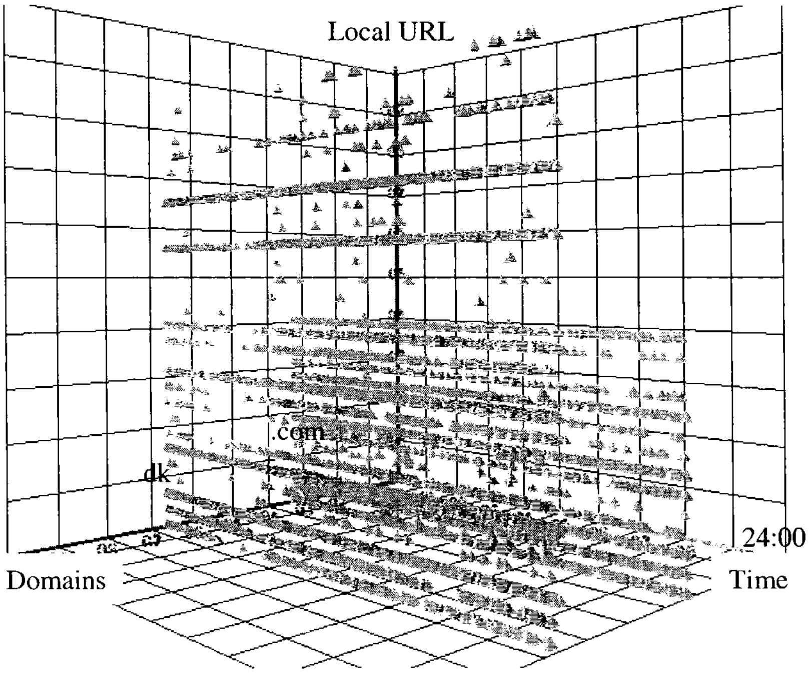


Figure 2. Raw data does not necessarily reveal its structure readily.

In both scenarios the data miner is in a role similar to Donald Schön's "reflective practitioner" (Schön, 1983; 1991), originally developed in the analysis of design processes. The overall process is summarised in . In his later work (Schön, 1991), he has presented a number of examples of disciplines that fit a process model with characteristics similar to the design process. What is relevant to both guided and validated cognition approaches described in this paper is Schön's view that designers put things together (in our case, the miner replaces the designer, and the things s/he put together are the visual pieces of information) and create new things (in our case the miner creates new models interacting with the visual fors) in a process involving large amount of variables as well as a range of obligations and limitations (in our case, the limitations are come from the dimensionality that the visualisation component of the visual data mining system can handle); that almost anything a designer does involves consequences that far exceed those expected (the techniques use to guide and validate the visual method may lead to results that far exceed those expected and that may change the line of the analysis of the data) and that a design process is a process which has no unique concrete solution (in our case we operate with numerous visual models, assisting in revealing the different aspects of the phenomena). Schön also states that he sees a designer as someone who changes an indefinite situation into a definite one through a reflexive conversation with the material of the situation. By analogy, visual miner changes the complexity of the problem through the reflexive investigative iterations [conversation] with the data, currently derived visual models and their validation. The reflective step is a revision of the reference framework taken in the previous step in terms of attributes selection, set of visual models and corresponding guiding techniques, and the set of validation techniques.

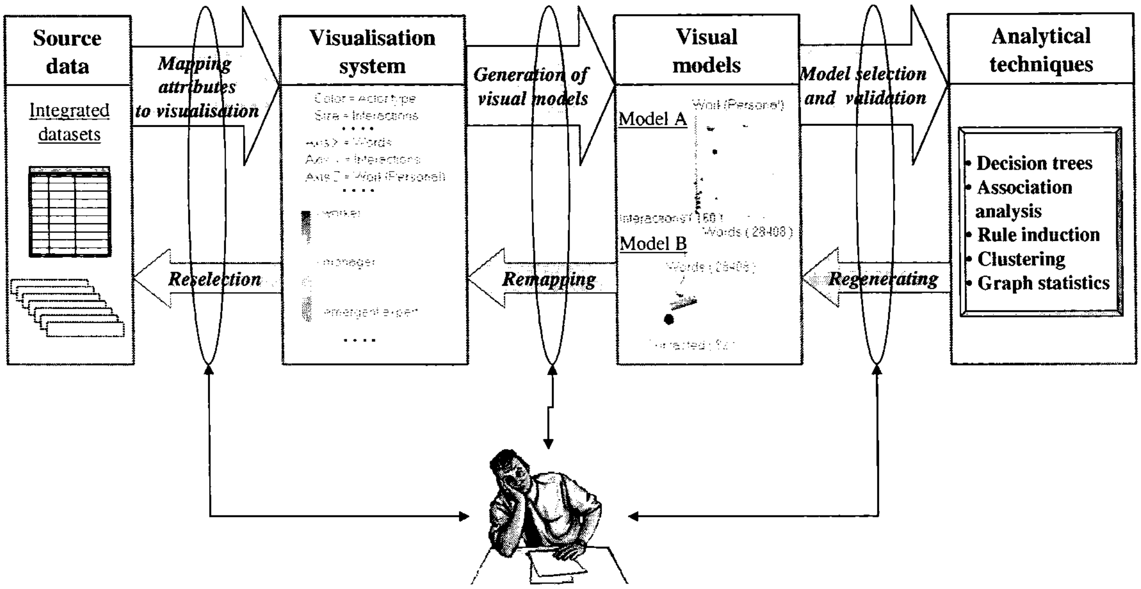


Figure 3. Visual data mining in terms of "reflection-in-action".

We illustrate the ideas behind the two approaches on portions of two case studies, one with large and one with small data sets.

2 The guided cognition approach

This section addresses the issue of stimulating ("guiding") our cognitive mechanisms to discover visual forms that lead to meaningful interpretation. We use the density surface technique (Bohlen *et al.*, 2003) and the change of density level to guide the data miner through the process. The data source is a three dimensional web log. The first dimension is the domain (we collapsed the URL to the last part, i.e., the country). There are hundreds of different countries, many of them with very few clicks only. In our plots we show the two most frequent domains: .com and .dk (the local domain). The second dimension is the date and time. For our analysis we picked a 24 hour segment starting at 00:00. The third dimension is the local (html) page that was requested. The organization is such that the general pages (departmental home page, high level departmental descriptions of teaching, research, etc) are toward the beginning of the axis whereas the more specialized pages (personal pages, specific research projects, courses, etc) are toward the end of the axis.

Figure 4 shows a scatter plot visualization of the raw data, which is still the most common type of visualization being used. In a scatter plot visualization each triple, e.g. (.dk, 08:53:12, www.cs.auc.dk/research/), is represented as a primitive visual object (a dot, a tetrahedron as in our case). Clearly the visualizations in Figure 4 provide limited information only. We see that some pages are requested very frequently. This is a very dominant pattern and yields a visualization with pronounced lines. While true it is also information that we most likely knew already before. It is also very difficult to recognize differences between the lines. While it is possible to tell apart pages that are hardly ever accessed from pages that are accessed often it is not possible to say if .com and .dk access the same pages most often. In general the interpretation of a 3D scatter plot benefits a lot from rotations. Rotation alleviates the problems of occlusion and corroborates the perception of the 3D space. The illustrations in Figure 2 and Figure 4 emphasise this problem.

Figure 5. complements the data with model information. In this case the model information is a density surface. Roughly, the density of a 3D data set is a 4D structure and we visualize a cut at a specific density level. These cuts are the surfaces displayed in Figure 5. A density surface is visualized using a very simplistic method: we draw points on the surface as primitive objects. This is a simplistic but effective method as illustrated in the sequence of steps in Figure 5.

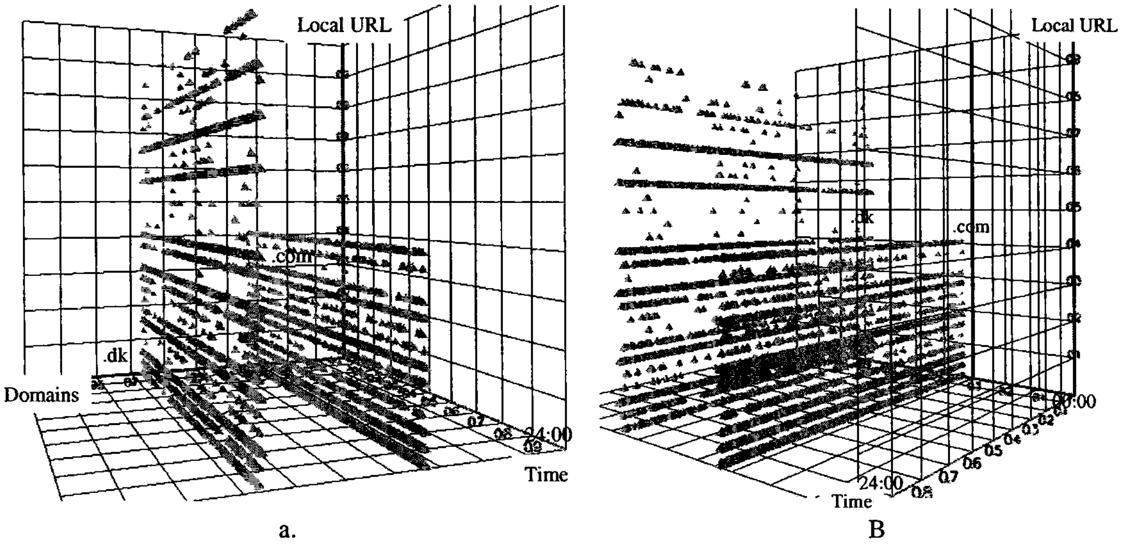
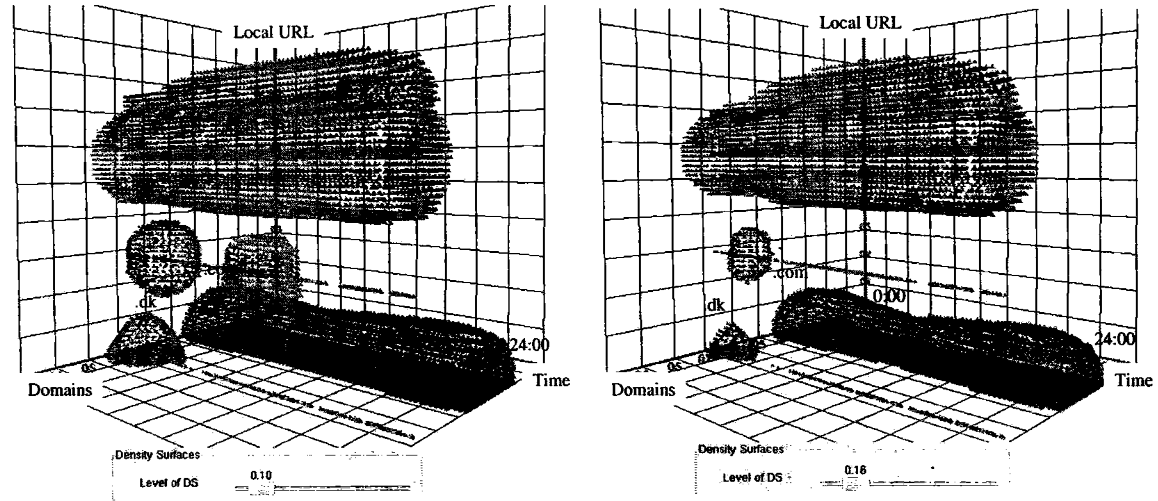


Figure 4. Simple operations with the visual perspective, such as rotation rarely can help in uncovering the patterns encoded in the raw data.

Figure 5 reveals information that we could not see from looking at the scatter plots. It becomes clear that .dk and .com users access quite different pages. Users from .com (the dark surfaces) most frequently access the departmental home page and possibly a few other pages from there. Users from .dk (the light surfaces) access the home page much less. Instead they directly proceed to the pages about teaching and research. By varying the density level from 0.16 to 0.08 we progress our understanding of the distribution of the clicks.

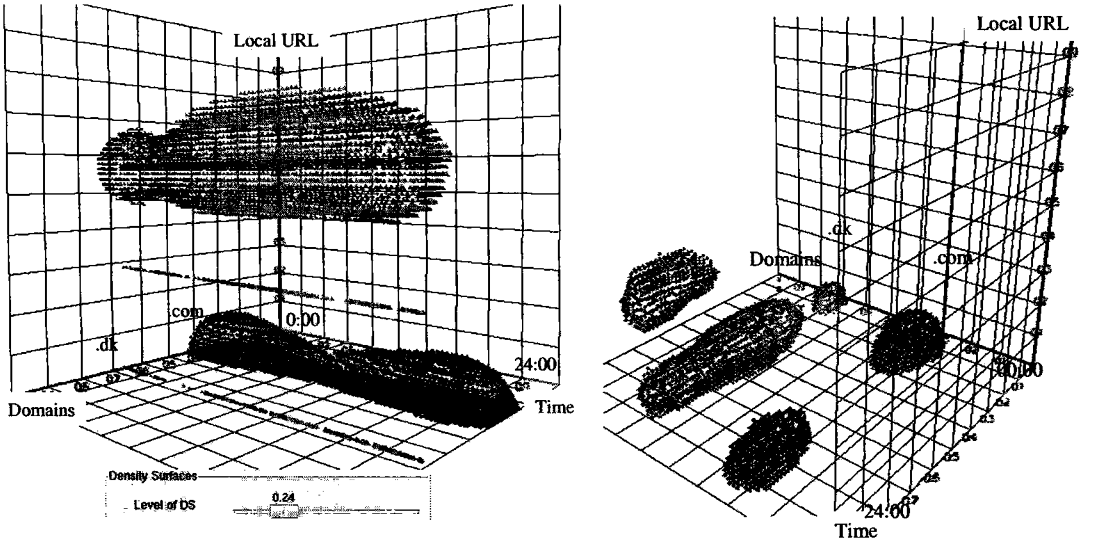


a. Patterns of sequential access emerged at the level of density around 0.10.

b. Further increase of the density level leads to clearer isolation of emerging patterns

Figure 5. "Switching on" the density surface technique reveals some patterns

Figure 6 further explores the data set by varying the density levels. We get a precise overall understanding of the distribution of the accessed pages. Throughout the investigation rotation is used to view the surfaces from different perspectives. There is no single perspective that works best in all cases.



a. Further increase of the density level leads to isolation of the final destinations

b. "Rotation" becomes again a useful operation

Figure 6. Detailed visual investigation of the original data set.

The above discussed examples clearly demonstrate the need for embedding statistical techniques that enhance the raw data with model information to guide visual analysis. The next case study illustrates an opposite approach where the visual analysis is done without embedded statistical technique, over the raw data. The hypothesis formulated this way are then validated (hence the name "validated cognition") using an analytical technique.

3 The validated cognition approach

The problem of "validated cognition" perhaps dates back to the works of John W. Tukey. Tukey was looking at methods for robust analysis in the presence of violation of initial assumptions, including robust visualisations methods. He emphasised that seeing may be believing or disbelieving, but above all, data analysis involves visual, as well as statistical, understanding. Perhaps the most famous and certainly one of the oldest visual explanations in mathematics is the visual proof of the Pythagorean Theorem. This proof is unusual in its brevity and its complete appropriateness to the problem. Thus, a visual reasoning approach to data mining promises to overcome some of the difficulties experienced in the comprehension of the information encoded in data sets and the models derived by pure analytical methods. These issues have been extensively discussed in a three consecutive workshops on visual data mining (see) What remained untouched was the issue of validating the results of our perception of visualised forms.

The validated cognition approach will be described on a case study of a company, whose managers wanted to grasp a better and valid picture of their social capital, in particular, of employees who emerged as proactive experts during the development of projects, and to understand the attributes that describe them. The data source includes the email interactions between company employees, record of documents flow, time-stamped consecutive versions of these documents, plus demographic data. For the illustration of the methodology discussed in this paper, we limit our investigation to the email data and will be using the email messages only in a half a year window, observing the communication of a group of employees that work on a particular project.

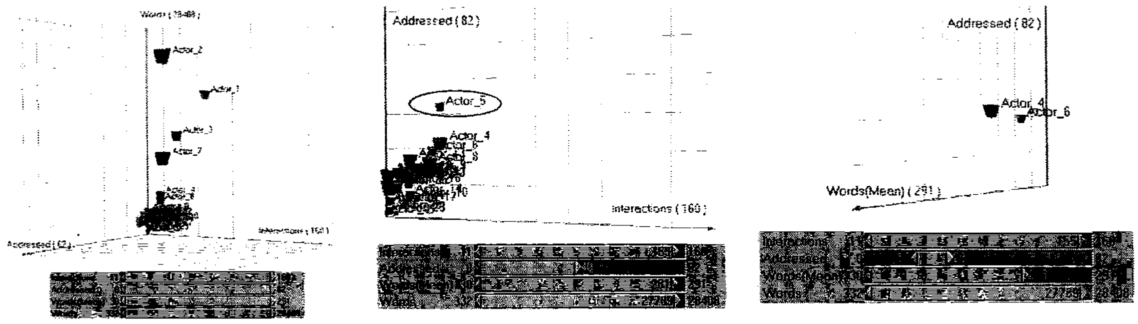
3.1 Goal of the study

We studied a distributed project group in a global company. The main aim of the study was to develop methods for identifying the so-called "invisible social capital" in an organisation, in particular, the goal of the study was to identify team members that emerged out of the group during the project run, based on their knowledge, expertise and attitude, and to identify the attributes that characterise their communication behaviour. To ensure that the project runs in a distributed team environment, project members were recruited from company sections in different regions of the globe, so that no project members were co-located. Team members were asked to communicate via e-mail, both for individual and group communications (group messages were sent to a common listserv account). Fax and phone were excluded as communication means. As part of the company some team members knew each other, but they had no history of working with each other as a team. For the illustration of our approach, we use only the email communication data set collected over a 3 month period. It includes 2954 messages, out of which, 1489 email messages were exchanged between individuals, the rest were sent to the project list (the content-based analysis of the group portion is beyond the scope of this paper). The assumption is that proactive participants will have more intensive communication exchanges in terms of individual message frequencies (both generating and being addressed) and by the amount of information that they communicate.

In this example, we illustrate a two stage analysis. The first stage is a visual analysis of the individual actors based on the attributes used to describe them. For the purpose of this paper we use four attributes: and three other attributes: "Interaction" (all the outgoing messages for particular individual), Addressed (all messages addressed to particular individual), "Words" (the total amount of words in all outgoing messages) and "Words (mean)" (the average amount of words in a message) We explore the structure of the data set and identify potential candidates for proactive team members. As we expect their behaviour to distinguish them from the rest of the group, we look for outliers or/and small isolated clusters. On the second stage, the results of our selection are refined by running a network analysis of the communication network established between project participants.

3.2 Visual data mining

The visualization of the structure of the data set revealed that a large group of participants had a relatively low intensity of individual communication and they cluster closer to the origin along the attributes "Interaction", Addressed and "Words". They were removed from the visualisation. The data for the remaining 30 project members is shown in Figure 7a[1]. Each pyramid represents a data point (in general, a projection of a data point), where the coordinates of the centre in our case are defined by the values of the above mentioned attributes. Project participants are labeled as "actors". Actor 1-3 and 7 are candidates for consideration, as they are out of the main clusters of project members. Then these points are filtered out. Actors 4 and 8 need further clarification. The attributes were remapped to the visual features and one dimension switched off. As a result, actor 5 emerged and was included in the candidate set, as shown in Figure 7b. Figure 7c illustrates how actors 4 and 6 were depicted. In a similar way, using filtering, remapping of attributes and navigation operations, actors 9 and 13 were identified. The final candidate list includes 1-6, 9 and 13.



a. b. c.

Figure 7. Visual exploration of the data set – looking for outliers.

[1] For the visual analysis in this case we used the off-the-shelf product Miner3D - a visual data mining tool by Dimension 5 (http://www.miner3d.com).

3.3 Visual data mining

The aim of the network analysis is to refine and validate the candidate list based on the differences in the position of different people in the project team as the project unfolds. We analysed the interactions between individuals during the project. The data set for this analysis included only the records in an email's "To:" and "From:" fields. The "Subject:" field and the actual content of the email messages at this stage were ignored. A large group of participants had a relatively low intensity of individual communication (the threshold was set to 10 messages), which left 30 project members in consideration. The network structure derived from the interaction between these project members is shown in Figure 8[2]. The network is non-uniform and directed, as indicated by the arrows at the ends of the edges, i.e. each edge may have one (from A to B or from B to A) or two links (from A to B and from B to A). Network parameters are presented in Table 1. The diameter of a network is the length of the longest geodesic path to be found in that network [geodesic path is the shortest path between two nodes]. In our case, the network is of the "small world" type (Barabasi, 2002), i.e. participants are not far from each other, in terms of message passing. This is confirmed by the low value of average eccentricity of nodes in the network (eccentricity summarises how far a node is from the node most distant from it in the network). The density of our network [the ratio of the total number of edges to the number of all possible edges in that network] shows that it seems to be of "scale free" type, where some hubs form the key nodes, with a number of "weakly" connected nodes. The cohesion index [the ratio of the number of mutual connections in the network to the maximum possible number of such connections] shows a relatively low amount of bidirectional interactions.

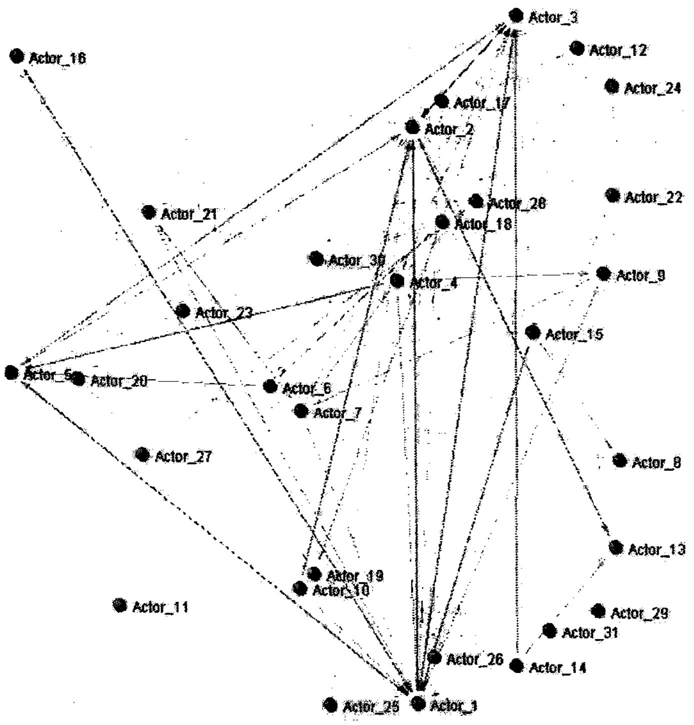


Figure 8. Structure of the interactions between project members

Table 1. Network parameters

Nodes	Edges	Diameter	Eccentricity (average)	Density	Cohesion
31	197	4	3	0.21	0.11

[2] We use Agna 2.1.1 and UCINET 6 to conduct the network analysis.

Consequently, we look at the sociometric statistics of the individual nodes, to select potential emergent authorities. These statistics for the thirty project members are shown in Table 2 (the table includes the emission degree of each node [the sum of all values corresponding to the edges originating in that node], the reception degree of each node [the sum of all values corresponding to the edges incident (directed) to that node] and the sociometric status of each node [the sum of its reception and emission degrees, relative to the number of all other nodes]. Setting up a reasonable threshold of 10 for the first two statistics and 0.5 for the sociometric status[3] we identify the nodes of interest that include actors 1-5, 8, 9, 13 and 21. This confirmed the initial list of actors (except, actor 6) generated during the visual stage. Actor 6 is a border case (as seem to be actor 7), and will need further separate investigation. We got in actor 8, which was originally spotted as a possible candidate, but was not identified through further visual manipulations. The final list, refined by content analysis of communication messages, included all actors, whose rows are shaded with grey in Table 2. The details of this analysis are beyond the scope of this paper.

Table 2. Distribution of sociometric statistics of individual project members

Node	Emission	Reception	Status
Actor_1	46.0	61.0	3.67
Actor_2	28.0	52.0	2.67
Actor_3	33.0	40.0	2.43
Actor_4	18.0	16.0	1.13
Actor_5	23.0	30.0	1.77
Actor_6	7.0	16.0	0.77
Actor_7	19.0	7.0	0.87
Actor_8	14.0	11.0	0.83
Actor_9	14.0	11.0	0.83
Actor_10	15.0	3.0	0.60
Actor_11	5.0	1.0	0.20
Actor_12	4.0	6.0	0.33
Actor_13	11.0	14.0	0.83
Actor_14	16.0	5.0	0.70
Actor_15	14.0	7.0	0.70
Actor_16	9.0	7.0	0.53
Actor_17	8.0	3.0	0.37
Actor_18	8.0	9.0	0.57
Actor_19	13.0	8.0	0.70
Actor_20	3.0	6.0	0.30
Actor_21	11.0	11.0	0.74
Actor_22	6.0	6.0	0.40
Actor_23	7.0	0.0	0.23
Actor_24	2.0	2.0	0.13
Actor_25	4.0	0.0	0.13
Actor_26	1.0	5.0	0.20
Actor_27	5.0	2.0	0.23
Actor_28	5.0	8.0	0.43
Actor_29	2.0	7.0	0.30
Actor_30	1.0	5.0	0.20

The above presented approach showed that for smaller data sets, choosing an appropriate mapping of the attributes to the features of a visual mining system can allow to formulate some hypothesis about a phenomenon encoded in the data. However, these hypothesis need to be validated and refined separately with an analytical technique. The approach is well-suited for mining of integrated data sets, where for example, communication data is only part of the data set.

4 Summary, research problems and future work

Many data analysis tools include variety of visualisation (e.g. pie charts, bar charts in 2D and 3D versions, etc.). These charts usually provide an alternative form of displaying the obvious. They do not provide data manipulation and analysis capabilities, and are oriented towards the reduction of a set of data to individual statistics. Visual data mining looks at having an access to the entire data set in its most granular level. We discussed the specifics of visual data mining methodology and presented a view of the visual data mining as a "reflection-in-action" technique. We have illustrated two different types of this technique, namely "guided cognition" and "validated cognition". This work is motivated by the fact that visual, though very attractive, means also subjective, and non-experts are often left to utilise visualisation methods (as an understandable alternative to the highly complex statistical approaches) without the ability to understand their applicability and limitations. It is questionable to what extent exploring raw

[3] For more thorough analysis, the

data (especially large amounts in 3D systems) actually helps the mining as perceptual faculty is overloaded and confused, rather than stimulated. An important issue is the identification of particular ranges, beyond which only a visual mining technique with embedded guidance can help. Another issue encountered in this work is the lack of effective interactive summarisation models that can reduce the computational load in real-time reflective techniques.

The future work in this direction will be looking at:

- guiding techniques (density surfaces is only on of these techniques)
- automatic selection of guiding techniques, i.e. the criteria for selection of different techniques;
- automatic selection of validation techniques

Acknowledgements

This research has been supported by the Australian Research Council. We would like to acknowledge the software support provided by Miner3D (http://miner3d.com/).

References

Ankerst, M. (2000) Visual Data Mining, In *Fakultät für Mathematik und Informatik*, Ludwig-Maximilians-Universität, München.

Ankerst, M., Grinstein, G. and Keim, D. (2002) Visual Data Mining: Background, Techniques, and Drug Discovery Applications, In *Tutorial Notes, ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining (KDD 2002)*, Edmonton, Canada.

Barabasi, A.-L. (2002) *Linked: The New Science of Networks*, Perseus Publishing.

Bohlen, M. H., Juozapavicius, A., Kondratas, E., Mazeika, A. and Struk, A. (2003) A Triangular Reconstruction of Density Surfaces, In *Proceedings 3rd International Workshop on Visual Data Mining*(Eds, Simoff, S. J., Noirhomme-Fraiture, M., Böhlen, M. H. and Ankerst, M.), University of Technology Sydney, 19th November, 2003, Melbourne, Florida, USA, pp. 45-58.

Dunham, M. H. (2003) *Data Mining: Introductory and Advanced Topics*, Prentice Hall, Pearson Education Inc., Upper Saddle River, New Jersey.

Hofmann, H., Siebes, A. and Wilhelm, A. (2000) Visualizing association rules with interactive mosaic plots, In *Proceedings of ACM SIGKDD Int. Conf. On Knowledge Discovery and Data Mining (KDD 2000)*, ACM Press, Boston, MA.

Noirhomme-Fraiture, M., Schöller, O., Demoulin, C. and Simoff, S. J. (2002) Sonification of time dependent data, In *Proceedings of the 2nd International Workshop on Visual Data Mining - VDM@ECML/PKDD'2002*(Eds, Simoff, S. J., Noirhomme-Fraiture, M. and Böhlen, M. H.), University of Helsinki Press, Helsinki, Finland, pp. 113-125.

Schön, D. (1983) *The Reflective Practitioner*, Basic Books, New York.

Schön, D. (1991) *Educating The Reflective Practitioner*, Jossey Bass, San Francisco.

Wong, P. C. (1999) Visual Data Mining, *IEEE Computer Graphics and Applications*, **September/October**, 1-3.