Mixing Evaluation Methods for Assessing the Utility of an Interactive InfoVis Technique

Markus Rester[1], Margit Pohl[1], Sylvia Wiltner[1],
Klaus Hinum[2], Silvia Miksch[3],
Christian Popow[4], and Susanne Ohmann[4]

[1] Institute of Design and Assessment of Technology, Vienna University of Technology, Austria
markus@igw.tuwien.ac.at
[2] Institute of Software Technology & Interactive Systems, Vienna Univ. of Technology, Austria
[3] Department of Information & Knowledge Engineering, Danube University of Krems, Austria
[4] Department of Child and Adolescent Psychiatry, Medical University of Vienna, Austria

Abstract. We describe the results of an empirical study comparing an interactive Information Visualization (InfoVis) technique called Gravi++ (GRAVI), Exploratory Data Analysis (EDA) and Machine Learning (ML). The application domain is the psychotherapeutic treatment of anorectic young women. The three techniques are supposed to support the therapists in finding the variables which influence success or failure in therapy.
To evaluate the utility of the three techniques we developed on the one hand a report system which helped subjects to formulate and document in a self-directed manner the insights they gained when using the three techniques. On the other hand, focus groups were held with the subjects. The combination of these very different evaluation methods prevents jumping to false conclusions and enables for an comprehensive assessment of the tested techniques.
The combined results indicate that the three techniques (EDA, ML, and GRAVI) are complementary and therefore should be used in conjunction.

Key words: Information Visualization, Evaluation, Utility, Focus Groups, Insight Reports, Methodology

1 Introduction

Several authors have pointed out the importance of evaluation studies of Information Visualization (InfoVis) techniques (see e.g. [1], [2], [3]). In the past few years usability studies concerning visualization techniques have become more frequent, and valuable information about the design of such systems has been gathered. Nevertheless, as [4] mentions, there is still too little systematic information about the specific strengths and weaknesses of the features of InfoVis techniques. Studies presenting data from practical experiences with InfoVis techniques can help to develop a more systematic framework to support the decision which InfoVis technique to use in a given context.

Medical data is a very interesting application area for Information Visualization. One of the reasons for this is the complex and time dependent character of these data. For such data, interesting InfoVis techniques have been developed in the past few years.

In the following, we will describe a study analyzing several different methods used to assess the therapeutic treatment of anorectic young women. During the therapy process a large amount of highly complex data is collected. Statistical methods are not suitable to analyze these data because of the small sample size, the high number of variables and the time dependent character of the data. The data results from extensive questionnaires the young women and their parents have to fill in several times before, during and after the therapy. These questionnaires treat questions like, e.g., the young women's propensity for depression, their social behavior or their attitude about eating. The therapists want to find patterns in the young women's behavior and try to isolate the specific factors influencing success or failure in the therapy (predictors). InfoVis techniques might be a valuable possibility to represent these data, but in accordance with the therapists we also chose two other potential techniques (Machine Learning and Exploratory Data Analysis).

Up till now, evaluation in Information Visualization was centered around two variables: time and error. This approach has been criticized recently [5]. For many applications, the measurement of time and errors is too narrow. Many visualization methods support extensive exploration processes and the formulation of hypotheses. For an exploration process, the measurement of time does not make sense, and in the context of the development of hypotheses, errors in a narrow sense do not occur. In an ill-structured domain with no clear-cut results like psychotherapy, for example, other approaches are necessary. Therefore, the concept of insights was introduced to make the results of the exploration processes based on InfoVis techniques more tangible [6]. Unluckily, there is no agreed upon definition of insights although cognitive psychology has dealt with this topic quite extensively (see e.g. [7]). Most authors define 'insight' in a quite pragmatic manner. In addition, there are no general frameworks for categorizing insights. [8] points out that a starting point might be using user tasks as, for example, finding clusters or extreme values. There are some general cognitive activities which often appear as insight categories, as, for example, finding detailed, factual information, identifying clusters, generalizations, identifying changes over time, etc. [6, 9]. We developed our own classification system, partly based on the generic categories described above and partly adapted to the specific task for which our visualization method was developed. Finding predictors plays an important part in the therapists work, therefore it is a central category of our analysis.

Developing a theoretical framework for the concept of insights and the definition of relevant categories of analysis will be an important area of future research.

2 Compared Techniques

An interactive InfoVis technique named Gravi++ (GRAVI) was developed to support the therapists and clinicians in exploring the multidimensional, abstract, and time dependent data [10]. GRAVI is based on a spring metaphor.The questions from the questionnaires are positioned on a circle. The icons representing the anorectic young women are arranged within this circle depending on the strength of attraction of the questions. The questions function, to a certain extent, like magnets. The final position of the patients' icons is a combination of the forces of all given answers on the questions (see

Fig. 1). GRAVI uses animation to deal with the time dependent data. The position of the patients' icons change over time. This allows analyzing and comparing the changing values. Various visualization options are available, like Star Glyphs and attraction rings to communicate the exact values of each answer or traces to show the paths of the patients' icons over all time steps.

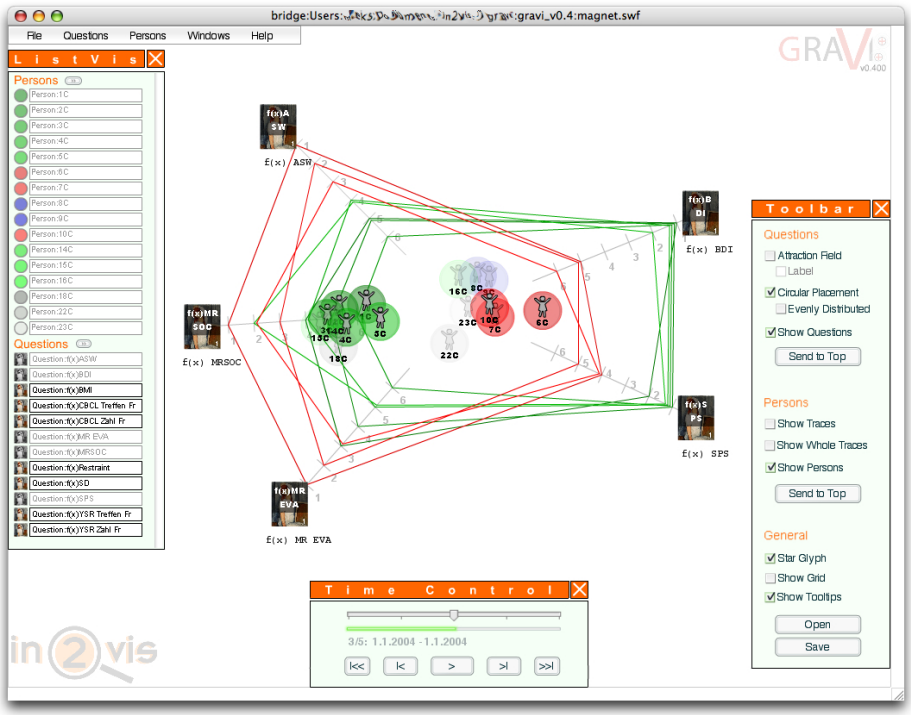

Fig. 1. GRAVI: Interactive InfoVis-Tool for Exploration of Multi-Dimensional Time Dependent Data (Typical Screenshot). Concept of Spring-Based Positioning Leads to Formation of Clusters.

We decided to compare GRAVI with the following techniques used so far for analyzing the data: Exploratory Data Analysis (EDA) and algorithms of Machine Learning (ML). In the case of EDA boxplots, histograms, scatterplots, and statistical measures were used (e.g., Fig. 2). The ML algorithms were: a C4.5 decision tree (e.g., Fig. 3) and a Support Vector Machine (SVM) trained by Sequential Minimal Optimization (SMO).

Exploratory Data Analysis (EDA) was developed by Tukey [11] and is based on statistics. It helps users to review and analyze data on a descriptive level. Tukey thought that the emphasis on statistical testing might be too narrow an approach. He, therefore, suggested EDA as a possibility to formulate hypotheses and assess assumptions. Subjects were given printouts of these techniques.

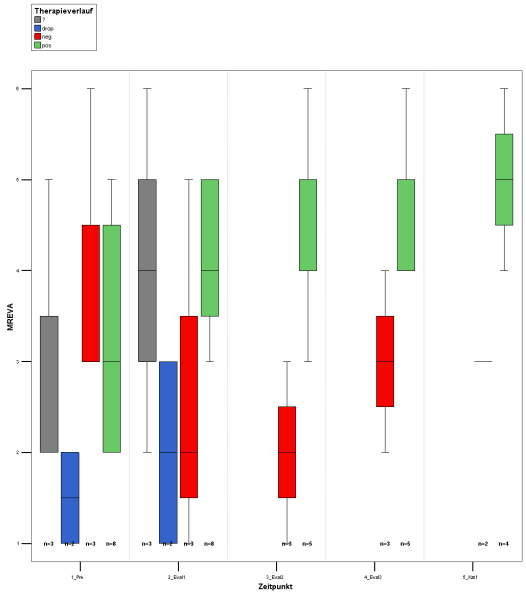

Fig. 2. Exploratory Data Analysis (EDA) Sample: Boxplots

Machine Learning is an area of AI concerned with the development of algorithms that enable computers to 'learn'. A Machine Learning technique learns from observed examples or data. In general, there are two types of machine learning algorithms: supervised and unsupervised. In case of supervised learning, a priori knowledge about the data is used and in case of unsupervised learning, no prior information is given regarding the data or the output. We utilized two supervised schemes using WEKA [12]: a Support Vector Machine with Sequential Minimal Optimization algorithm [13, 14] and a pruned C4.5 decision tree [15]. The output of these two techniques were again available to the subjects as printouts.

3 Evaluation Methods

An extensive evaluation of InfoVis has to take place on different stages. Important areas of interest can be: usability evaluation, insight study, case study, and transferability assessment (see [16] for details). For results of a usability evaluation of GRAVI see [17]. The used methods in the insight study were insight reports [16] and focus groups (cf. [18]).

A sample of 32 subjects participated in the study. They were computer science students and can therefore be described as domain novices. Therefore they received a comprehensive introduction to the domain (data, real users' tasks, etc.) and introductions to the three different techniques to use. The evaluation with insight reports took place in a laboratory setting and lasted for an overall of 155 minutes. There was equal time for the

J48 t=all

```
Instances:    80
Attributes:   13 (BMI, ASW, BDI, SPS, SD, Restraint, MREVA, MRSOC, YSR:ZahlFr,
YSR:TreffenFr, CBCL:ZahlFr, CBCL:TreffenFr, Therapieerfolg)

J48 pruned tree
------------------

ASW <= 3
|   CBCL:TreffenFr <= 1
|   |   BMI <= 3: neg (9.66/2.34)
|   |   BMI > 3: pos (3.69/1.37)
|   CBCL:TreffenFr > 1
|   |   ASW <= 2: neg (6.77/3.24)
|   |   ASW > 2: pos (11.03/5.62)
ASW > 3: pos (33.85/3.65)

Number of Leaves  :       5
Size of the tree :  9

=== Summary ===

Correctly Classified Instances          52               80      %
Incorrectly Classified Instances        13               20      %
Total Number of Instances               65
Ignored Class Unknown Instances         15

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
  0.975     0.4        0.796     0.975     0.876      0.887    pos
  0.867     0.06       0.813     0.867     0.839      0.918    neg
  0         0          0         0         0          0.694    drop

=== Confusion Matrix ===

  a  b  c   <-- classified as
 39  1  0 |  a = pos
  2 13  0 |  b = neg
  8  2  0 |  c = drop
```

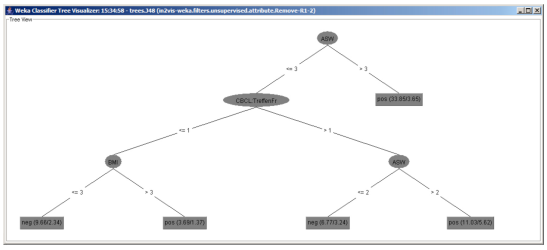

Fig. 3. Machine Learning (ML) Sample: C4.5 Decision Tree

three techniques (GRAVI, EDA, ML). Subjects were divided into three groups which used the three techniques in different order. Every technique was was once used in first, second, and third place (MEG, EGM, GME).

The subjects used a report system to formulate and document their findings during the exploration process in a self-directed manner. Whenever an insight occurred they had to generate a report with this system. The following data was collected: used material, description of finding, and confidence rating. The insight reports were later classified in the following categories: complexity of each insight, plausibility of an insight, and whether an assigned insight has been elaborated in more detail and if so, whether this elaboration was sound or not valid (see Fig. 4).

Focus groups can give interesting insights into the users' attitudes and experiences although they do not provide representative results [18]. [19] reports that focus groups are especially valuable for evaluating InfoVis techniques as they are able to uncover unexpected problems that cannot be perceived by other research methods. In this sense, they can be an interesting complementary approach to other more systematic methods.

So focus groups with the same subjects were held a week after the laboratory setting. They lasted about 100 minutes each. Eight questions were discussed (e.g., ease

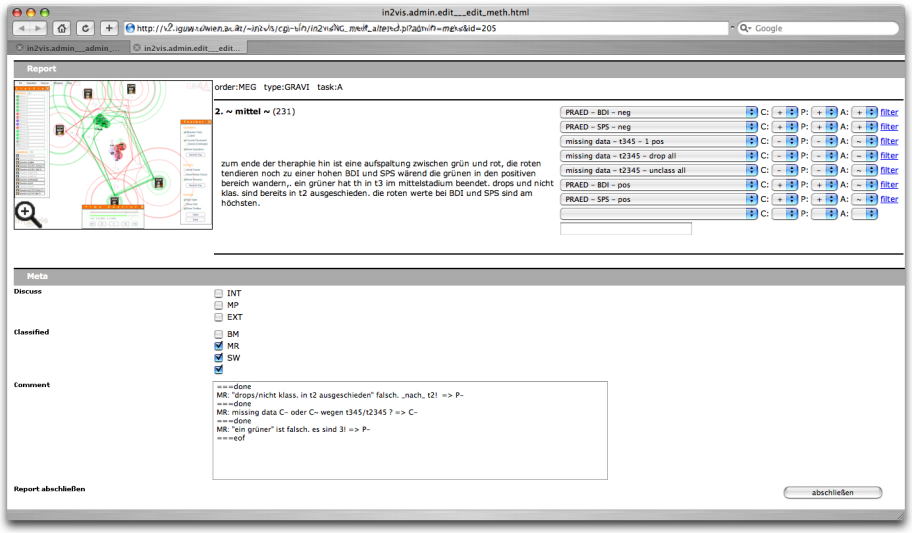

Fig. 4. Insight Report Documented by Subject with Classification and Categorization Options for Investigators

of use and utility, major strength and weakness, similarity and difference of insights gained with the different techniques, appropriateness of combined use). The value of this method is that it reveals subjective impressions on questions not asked before and gives a different perspective on as well as arguments for interpretation of the data collected in the experiment.

The discussion guideline consisted of eight questions. A set of the first four questions had to be discussed by subjects for each of the three used techniques (GRAVI, EDA, ML) separately. Afterward four more questions were addressed to them concerning all three techniques:

1. Appropriateness of the allowed time.
2. Ease of use and usefulness of the technique for gaining insights.
3. Overall confidence in insights gained with the technique.
4. Major strength and weakness of the technique.

5. Similarity and difference of gained insights using different techniques.
6. Assumed comprehension rates of the complex matter with each technique.
7. Appropriateness of combined use of the three techniques.
8. Order for best possible comprehension of the data.

4 Results

4.1 Insight Reports

The 32 subjects documented an overall of 876 reports. In the classification process we defined 805 different insights which were assigned 2166 times to the reports. Statistical

analysis of the collected data from this experiment was carried out. In depth details of these results are currently subject to reviewing and will be published in the near future.

To sum up, the results could lead to the conclusion that ML is not a recommendable technique. The subjects' confidence ratings were low, the complexity of the gained insights was low and few predictors were found. On the other hand GRAVI performed very well concerning insights with high domain value (finding predictors). Confidence ratings were also generally high. EDA lies somewhere in between. Histograms and scatterplots are well known. But the interpretation of boxplots and statistical measures require some familiarity with these techniques. EDA seems especially suited – or more precisely, it was utilized in particular – to analyze single values of individual patients in specific time steps. This may also be the reason why there were fewer wrong arguments with EDA. In contrast, there were many wrong arguments with ML.

4.2 Focus Groups

Appropriateness of the Allowed Time Concerning ML the subjects' statements clearly show a connection to the position of ML in the order of used techniques: in the case ML was the first technique all of the subjects stated that there was too little time for the tasks. If ML followed GRAVI the allowed time was rated appropriate. Using ML at last led to the assessment that there was too much time left. Many explanatory statements were as follows: subjects are not familiar with ML; ML is no suitable technique to start with if one is not a domain expert already; ML is complex and confusing; and there were no new insights that have not been already gained with the other two techniques.

In general the time allowed while using EDA was predominantly rated appropriate. Once more only when used as the first technique the subjects would have needed more time for the tasks. The familiarity with EDA was pointed out by the subjects. Only the statistical measures from EDA were criticized as difficult to interpret.

The ratings for GRAVI are similar to ML, though not as pronounced, and follow the position of GRAVI: if GRAVI is used as first technique, subjects would have needed more time to get familiar with both technique and domain. For GRAVI in second position we have a trend towards too much time available for the tasks. Used as last technique subjects rated the allowed time appropriate.

Ease of Use and Usefulness of the Technique for Gaining Insights ML had the lowest scores regarding the usefulness for gaining insights. 55% of all statements made by the subjects belong to the lowest category on this scale. Once again, unfamiliarity with and complexity of ML led to high level of uncertainty.

The assessment of EDA was twofold: scatterplots and histograms scored very well, whereas boxplots and statistical measures were not rated as useful. The former were favored for their simplicity and for being visualizations. The latter were criticized for being complicated and in the case of statistical measures for not being a visualization additionally.

Also for GRAVI the subjects appreciated some elements as well as disapproved of others. The interactivity of this technique in general and its powerful capability to handle the time dependent data in particular were rated as very useful. Different visual details, like poor visibility of missing data, were mentioned to hinder usefulness.

Overall Confidence in Insights Gained with the Technique ML had an even worse assessment in the focus groups compared to the ratings given in the lab setting: 65.6% of the statements rated ML in the category "low confidence". This high ratio is most likely due to peer pressure in one of the three groups where all of the 12 participants rated ML unanimously (low confidence).

Interestingly, EDA scored better than GRAVI in the focus groups. One possible explanation for this may be that EDA received a lot of high ratings in the focus group of those who used EDA at last. So we have probably on the one hand a form of learning effect leading to more domain expertise which also affects the confidence in observations. On the other hand EDA was the only technique the subjects were rather familiar with. So it is the more noteworthy that GRAVI did only receive a few more ratings in the "low confidence" category than EDA.

Major Strength and Weakness of the Technique Although the subjects could not make much use of ML they believe that for experts of ML this technique allows for very concise and valid insights. There was a strong appreciation of the automaticity of calculations and a high level of faith in the correctness of the results. The latter was also raised by the often positively mentioned confusion matrix, which is an self-evaluation on correctness provided by the ML algorithms. The visualization of the decision tree was rated a plus whereas the formula of SMO was mentioned to be confusing.

The mentioned strengths of EDA were: visual elements (scatterplots, histograms), simplicity, familiarity, and clarity of displayed data. The lack of interactivity, the impossibility of comparison of patients and/or groups of patients, and problems with the exploration of time dependency are the downside of EDA.

GRAVI impressed by its interactivity, many options to visualize data in different ways, the handling of time dependent data, its simplicity, and its intuitive interface. Subjects saw the major weaknesses of GRAVI in the fact that visualizing much data rapidly leads to glutted displays. Also the need for check and re-check of possible insights with different constellations is important. Otherwise false conclusions could easily be drawn.

Similarity and Difference of Gained Insights Using Different Techniques The subjects reported by majority that they found the same insights with the three different techniques. Almost 2/3 of the made statements went in this category. Nevertheless the detail of insights varied.

Assumed Comprehension Rates of the Complex Matter with Each Technique ML showed to contribute very little to the comprehension of the provided data. This is in clear accordance with the former statements of the subjects. EDA and GRAVI on the other hand could be utilized well by subjects.

Appropriateness of Combined Use of the Three Techniques 45% of statements put on record that the combined use of ML, EDA, and GRAVI makes perfect sense because all three techniques offer different views on the data and therefore facilitate a deeper understanding and extensive exploration.

Other 45% of statements pleaded for omission of ML due to its marginal contribution in comprehension of the data for the subjects who were not familiar with this complex technique.

Order for Best Possible Comprehension of the Data There were almost as many preferred orders in using the three techniques as there were subjects. But there are also some major similarities in the statements: ML is not suitable as the first technique but more useful to recheck insights gained with other techniques. GRAVI and also parts of EDA (simple visual parts: histograms and scatterplots) are viable techniques for first exploration of data. Another interesting outcome in the discussion was that the different techniques should not be used sequentially like in the laboratory setting but simultaneously. The already mentioned different views they provide on the data could add much more value in this way.

5 Conclusion

The use of diverse evaluation methods enables different views on the technology under investigation. Whereas insight reports can reveal strengths and weaknesses in form of summative tests followed by statistical analysis, focus groups often give reasons and additional subjective opinions of subjects and therefore also ensure correct interpretation of the former.

The outcome of insight reports could lead to the conclusion that ML is not a recommendable technique because of low confidence ratings, low complexity of the gained insights, and small number of found predictors. On the other hand GRAVI performed very well. There were many insights with high domain value (predictors) and with high confidence ratings. EDA seems especially suited to analyze single values of individual patients in specific time steps.

The outcome of focus groups shows that GRAVI is useful for gaining insights with a high confidence rating, because of its flexibility through interactivity, the ability to explore more dimensions simultaneously, and the straightforward navigation within the time dependent data. Moreover, subjects rated GRAVI an appropriate visualization tool. ML should be omitted unless there is enough expertise with this technique. If so, it still can and probably will be a powerful technique to gain insight. EDA rapidly leads to insights (although rather basic ones) due to the general familiarity with this technique.

Combining these results we see, that all three techniques offer different views on the data and therefore a combined use will likely lead to more insight and comprehension.

Acknowledgments The project “Interactive Information Visualization: Exploring and Supporting Human Reasoning Processes” is financed by the Vienna Science and Technology Fund [Grant WWTF CI038]. Thanks to Bernhard Meyer for the collaboration in the classification process.

References

1. Chen, C.: Empirical evaluation of information visualizations: an introduction. Int. J. Human-Computer Studies **53**(5) (2000) 631–635
2. Plaisant, C.: The challenge of information visualization evaluation. In Costabile, M.F., ed.: Proceedings of the working conference on Advanced visual interfaces, ACM Press (2004) 109–116
3. Tory, M., Möller, T.: Human factors in visualization research. Visualization and Computer Graphics, IEEE Transactions on **10**(1) (2004) 72–84
4. Spence, R.: Information Visualization. ACM Press (2001)
5. Stasko, J.: Evaluating information visualizations: Issues and opportunities (position statement). In Bertini, E., Plaisant, C., Santucci, G., eds.: BEyond time and errors: novel evaLuation methods for Information Visualization – Proceedings of BELIV'06, Venice, Italy (2006) 5–7
6. Saraiya, P., North, C., Duca, K.: An insight-based methodology for evaluating bioinformatics visualizations. Visualization and Computer Graphics, IEEE Transactions on **11**(4) (2005) 443–456
7. Eysenck, M.W., Keane, M.T.: Cognitive Psychology. A Student's Handbook. Psychology Press, Taylor & Francis Group, London, New York (2005)
8. North, C.: Toward measuring visualization insight. Computer Graphics and Applications, IEEE **26**(3) (2006) 6–9
9. Lanzenberger, M.: The Interactive Stardinates – An Information Visualization Technique Applied in a Multiple View System. PhD thesis, Vienna University of Technology, Vienna, Austria (September 2003)
10. Hinum, K., Miksch, S., Aigner, W., Ohmann, S., Popow, C., Pohl, M., Rester, M.: Gravi++: Interactive information visualization to explore highly structured temporal data. Journal of Universal Comp. Science **11**(11) (2005) 1792–1805
11. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading, Mass. (1998)
12. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd edn. Morgan Kaufmann, San Francisco, CA. (2005)
13. Platt, J.: Fast training of support vector machines using sequential minimal optimization. In Schoelkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning. MIT Press (1998) 185–210
14. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computing **13**(3) (2001) 637–649
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA. (1993)
16. Rester, M., Pohl, M., Hinum, K., Miksch, S., Popow, C., Ohmann, S., Banovic, S.: Methods for the evaluation of an interactive infovis tool supporting exploratory reasoning processes. In: BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors, New York, NY, ACM Press (2006) 32–37
17. Rester, M., Pohl, M., Hinum, K., Miksch, S., Ohmann, S., Popow, C., Banovic, S.: Assessing the usability of an interactive information visualization method as the first step of a sustainable evaluation. In: Proc. Empowering Software Quality: How can Usability Engineering reach these goals?, Austrian Computer Society (2005) 31–44
18. Kuniavsky, M.: User Experience: A Practitioner's Guide for User Research. Morgan Kaufmann (2003)
19. Mazza, R.: Evaluating information visualization applications with focus groups: the coursevis experience. In: BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors, New York, NY, USA, ACM Press (2006) 1–6