Exploring New Ways of Utilizing Automated Clustering and Machine Learning Techniques in Information Visualization

Johann Schrammel

University of Salzburg, 5020 Salzburg
schrammel@cure.at
Advisor: Prof. Manfred Tscheligi
manfred.tscheligi@sbg.ac.at

Research Area: Information visualization, human-computer interaction.

Research Topic. The main research topic of the thesis is to explore the possibilities of automated clustering and machine learning techniques for developing new approaches in information visualization.

Research Problem. The main goal of information visualization is to present data to the users in a way that optimizes intelligibility of the data and support the detection of relevant patterns in the data, where the application context defines what qualifies as 'relevant'. Many different approaches typically tailored to a specific problem have been developed within the past years. At the same time the application of mathematical methods for data analysis and identification of patterns has substantially increased, and is typically referred to as data mining. Different visualization techniques are used in data mining, however the systematic and dynamic integration of data mining techniques with visualization approaches is only in its beginning.

The main research problem tackled in the thesis is how to best integrate approaches from traditional data mining and information visualization domains to achieve more usable and helpful analysis tools for large amounts of data. The basic visualization approach explored is tag clouds, and especially concepts in displaying them according to a semantic structure as proposed e.g. by [2,3,1] are further developed. The overall research problem can be divided into sub-problems which need to be addressed in order to develop actually useful approaches:

- Develop new approaches and concepts for integrating data mining and information visualization techniques
- Develop a detailed understanding of the human perception and analysis process involved when interacting with these systems
- Develop a understanding on the impact of errors, artifacts and imperfection in the data analysis on experienced quality and usefulness
- Define quality benchmarks that are needed to be met to provide real value and benefit to the user when using such systems
- Evaluate the developed concepts and prototypes in realistic scenarios, to be able to estimate their value and applicability in real-world contexts.

P. Campos et al. (Eds.): INTERACT 2011, Part IV, LNCS 6949, pp. 394–397, 2011.

Relevance of research. We think research on this topic is highly important. The information society is producing more and more data, and better means to make sense of the available data are urgently needed. Also, new visualization approaches can help to communicate complex affairs in an intelligible manner, which is relevant in ever more context. Examples that illustrate this increasing need to understand complex data even in everyday situations are e.g. home control units in smart grid applications or interfaces for the specification of security, firewall and privacy settings.

The research hypothesis (claim). The overall research hypothesis is that utilizing automatically calculated relations between data elements and visualizing these identified relations in a proper way can help to develop a deeper understanding of the data. We also expect such approaches to support the user in performing different types of typical tasks such as data exploration or specific searches more efficiently.

Within the thesis work I plan to research this overall hypothesis and the related sub-aspects in a series of prototypes which explore the possibilities and potentials of different approaches. As of now the following design approaches already have been explored (a and b) respectively are planned (c to f):

a) Semantic layout of data elements: data elements are placed according to a calculated semantic layout based on the similarity of elements
b) Bottom-up clustering: Data elements are clustered automatically, and no label for the identified clusters is available.
c) Top-town categorization: Categories are defined a priori, and learned using standard machine learning algorithms using manually labeled training data.
d) Differential placing of elements: data elements are placed according to their similarity with a predefined bipolar concept.
e) Non-exclusive grouping: Data elements are cluster automatically into groups, however in contrast to b) the grouping is non-exclusive i.e. the categories do overlap.
f) Time-series visualization: The development of data structures across time is visualized

We hope to identify further promising approaches as part of the discussion in the doctoral consortium.

Methods used. The research approach follows a standard process consisting of the design of new approaches, a prototypical implementation of these approaches, optimization of these prototype implementations through fine-tuning of parameters based on informal user tests and a formal empirical evaluation with real end users in a realistic application context.

Empirical evaluation will target three main areas: performance, subjective satisfaction and analysis of the perception process. Performance will be operationalized as task completion time and error rate, subjective satisfaction with the approach will be explicitly asked in a questionnaire and analysis of perception process will be done using eye tracking technology. The main evaluation methodology will be similar to the approaches applied in the already performed evaluations as reported in [5] and [7].

A sketch of the proposed solution. In detail the following test prototypes have been implemented respectively are planned as of now:

a) Semantic placing of elements [5]: A thirst approach was to use semantic placing of data elements in a tag cloud. We used the getrelated-function of flickrs API to retrieve a list of the tags most related to each word within the tag cloud. Then based on the number of co-occurring related tags a measure for the relatedness of two tags was calculated. An alternating least-squares algorithm to perform multidimensional scaling (ALSCAL) was used to compute a two-dimensional arrangement of the tags. In the third step we used the value on the y-axis to form 7 groups of 11 resp. 10 tags each. Next tags within each group were sorted according to their value on the x-axis. The result provided an 11 times 7 arrangement that was used to generate the tag cloud.

b) Clustered tag clouds (related submission #232): Here we calculate tag similarity using a well proven method known as Jaccard coefficient. Similarity between tags is measured by the intersection divided by the union of the sample set. Based on this similarity measures clusters of tags where calculated using the bisecting k-means approach. For a discussion of different clustering approaches and their pros and cons see Steinbach et al. [8]. The clusters were calculated using the CLUTO-Toolkit provided and described by [4]. Basically the N-dimensional similarity matrix of tags was used as an input for the clustering algorithm. The target number of clusters to calculate was specified as 20. This number was chosen to form clusters of about 5 tags, which informal pre-test showed to be a good size for clusters.

c) Categorized tag clouds (under construction): The clustering process used in b) doesn't identify the topic of a cluster but only groups similar items together. To be able to also study the effects of labeled groups we plan to apply machine learning algorithms for the classification process, in which a subset of the data is categorized and labeled manually and used as training data.

d) Differential tag clouds (planned): Within this approach the idea is to use the calculated similarity to predefined bipolar concepts (e.g. work versus leisure) to arrange the items on the screen. Similarity is expected to be calculated by the same

method as in c), however more extensive training data on the bipolar concepts are needed. We also plan to use interactivity in this approach, i.e. to allow the user to select different concepts and to directly adapt the display.

e) Non-exclusive grouping (planned): In this approach we plan to apply non-exclusive clustering algorithms to the data set and to develop different visualizations for the users. A possible approach is to implement a slider, which dynamically changes the threshold above which the groups are displayed, or to show the related items only on mouseover to minimize visual clutter. We also plan to explore the possibilities of automatic drawing approaches for Euler graphs [5].

f) Time-series visualization (planned): Here we plan to introduce a dynamic element. The basic idea is to use the developed automated classification and placement algorithms to visualize the development across time within a field. For example, when using a differential tag clouds approach (compare item d) with the concept work – leisure and items are arranged horizontally regarding their distance to the concept (e.g. work-related displayed left, leisure related displayed right) and to also indicate the allocation to one of these concepts using color. One then could analyze chronological data sets (e.g. articles in a newspaper per day) and show a dynamic display of the development of the topics over time.

The expected contributions of the PhD research

a) New methods for displaying data that combine data mining and information visualization approaches that allow developing a faster understanding of patterns within the data and support users in complex analysis tasks.
b) Improved understanding of perception processes in information visualization especially with regard to the perception of modified tag clouds

References

1. Fujimura, K., Fujimura, S., Matsubayashi, T., Yamada, T., Okuda, H.: Topigraphy: visualization for large-scale tag clouds. In: Proc. WWW 2008 (2008)
2. Hassan-Montero, Y., Herrero-Solana, V.: Improving tagclouds as visual information retrieval interfaces. In: Proc. InfoSciT 2006 (2006)
3. Lohmann, S., Ziegler, J., Tetzlaff, L.: Comparison of Tag Cloud Layouts: Task-Related Performance and Visual Exploration. In: Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P., Prates, R.O., Winckler, M. (eds.) INTERACT 2009. LNCS, vol. 5726, pp. 392–404. Springer, Heidelberg (2009)
4. Karypis, G.: CLUTO - a clustering toolkit. Technical Report #02-017 (November 2003)
5. Rodgers, P., Mutton, P., Flower, J.: Dynamic Euler Diagram Drawing. In: Visual Languages and Human Centric Computing (2004)
6. Schrammel, J., Leitner, M., Tscheligi, M.: Semantically structured tag clouds: an empirical evaluation of clustered presentation approaches. In: Proc. CHI 2009 (2009)
7. Schrammel, J., Deutsch, S., Tscheligi, M.: The Visual Perception of Tag Clouds - Results from an Eye Tracking Study. In: Proc. INTERACT 2009 (2009)
8. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: Grobelnik, M., Mladenic, D., Milic-Frayling, N. (eds.) KDD 2000 Workshop on Text Mining, Boston, MA (2000)