Parameter für die Visualisierung von Dokumenten-Mengen

Parameters for the Visualization of Document Sets

Maximilian Eibl
IZ Sozialwissenschaften
GESIS Außenstelle
Schiffbauerdamm 19
10117 Berlin
Germany
Tel.: ++49-30-233611-323
Fax: ++49-30-233611-310
E-mail: eibl@berlin.iz-soz.de

Thomas Mandl
Universität Hildesheim
Angewandte Informationswissenschaft
Marienburger Platz 22
31141 Hildesheim
Germany
Tel.: ++49-5121/883-837
Fax: ++49-5121/883-802
E-mail: mandl@uni-hildesheim.de

Zusammenfassung:
Visualisierung wird im Allgemeinen als eine gute Möglichkeit angesehen Anwender von Informationssystemen in der Nutzung zu unterstützen. Semantische Information kann in Visualisierungen auf vielfältige Art transportiert werden: Farbe, Form, Position, etc. In den letzten Jahren fanden zweidimensionale Darstellungen von Dokumentverteilungen immer mehr Bedeutung in der Forschung. Hier wird semantische Ähnlichkeit zwischen Dokumenten üblicherweise durch räumliche Nähe ausgedrückt. In diesem Beitrag werden verschiedene Möglichkeiten diskutiert, solche Dokumentverteilungen graphisch umzusetzen.

Abstract:
Visualization is generally regarded as a good technique to support user interaction in complex information systems. Many design features may transmit semantic information in a visualization including distance, movement, color, size and orientation. Two-dimensional display where semantic similarity is expressed by means of distances between objects have gained considerable attention in the last years. The visual approach to information processing in these systems requires consideration of aesthetic aspects. We provide an overview on current research topics involving these two-dimensional maps and present an implementation where many design parameters can be modified in order to control the visual appearance.

Visualization plays a crucial role in knowledge management. The visual presentation of semantics is an important method for transferring knowledge explicitly stored in machine readable formats to humans. Semantics is mostly displayed in two ways depending on the certainty of the knowledge. Most often, concepts, ontologies and the objects represented are described by certain knowledge which is formalized by hard facts, rules or relationships. However, more and more often knowledge in computing systems is presented in a more natural way and is represented by vague relations or facts.

Especially, for systems dealing with vague knowledge, visualization presents a high challenge. In our research we focus on visualization of document sets where the relations between documents are not assessed by humans. The relations need to be calculated from the property values of the objects. For text objects, usually similarity measures are calculated on the basis of the indexing terms found [cf. Mandl 2001].

In order to allow the projection of the objects onto two dimensions these usually highly dimensional spaces need to be reduced. Methods of dimensionality reduction are for example Latent Semantic Indexing (LSI) or Kohonen Self Organizing Maps (SOM) [cf. Kohonen 1997]. That way, computer graphic algorithms also lead to new opportunities for user interfaces. Unlike common WIMP (windows, icons, menus, pointers) interfaces, advanced graphic elements consist of more than lines

and rectangles and partly rely on the aesthetic perception by the user. Vague human needs are addressed through advanced computer graphic routines.

Two-dimensional document maps have received considerable attention in recent years. Several experimental and commercial systems have been developed (e.g. http://www.cartia.com, http://websom.hut.fi/websom). Document maps try to exploit the visual capabilities of humans in order to create interfaces which are easy to use. Objects closely related are located next to each other, geometric distance becomes a metaphor for semantic similarity. This basic design strategy is cognitively plausible and is usually well understood by users. 2D maps have been applied to a variety of different object types including text documents [Honkela et al. 1997], software code, economic time series, authors of scientific literature and their position in a social network of discourse [Mutschke 2001] and multimedia internet content. Kohonen networks have been applied to software code in order to improve software reusability [Merkl 1995] as well as to music pieces [Rauber & Frühwirth 2001]. They seem to be especially suited for images. When visual perception of the objects is necessary, this form of presenting similarity is quite natural. As a consequence, there are applications for image retrieval [Ojala et al. 2001] as well as for pattern.

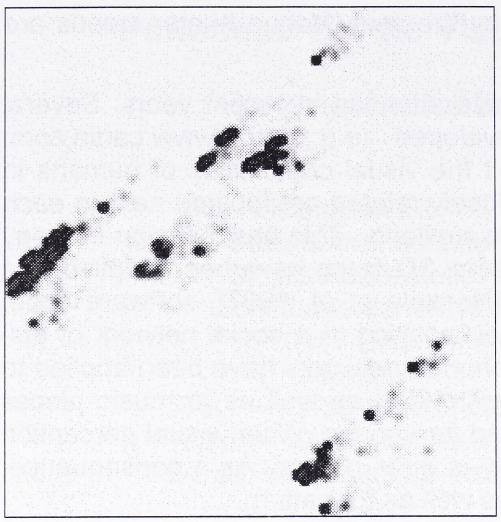
The evaluation of visualizations is a crucial aspect in the development. In order to ensure users satisfaction and effective use, graphic interfaces need to be thoroughly evaluated. Our evaluation strategy focuses on the selection of the proper algorithms for dimensionality reduction. This evaluation joins the users perspective with the mathematical foundation of 2D displays. Different strategies for the reduction of dimensions lead to very different results [Mandl & Eibl 2001]. Thus, in an LSI-map the documents are located totally differently to a SOM-map. We also need to consider the users way of interacting with a map. Most users will focus on one document at one point. From there they may browser to a close document and then again to a close document. In that process they may pursue a local or a global browsing strategy. A locally oriented user will lose sight of the initially evaluated documents and step to next closest document form the document momentarily under focus. In contrast, a globally oriented user will return to the first document and visit the next closest document to that one [Eibl & Mandl 2002].

In this article we want to demonstrate different methods of visualizing these data. We investigate several two and three dimensional displays. All of the six discussed visualizations are based on a set of some 10.000 documents from the social sciences. The dimensionality of this document set was reduced to two (fig.1-5) resp. three (fig.6) dimensions by an LSI algorithm. In the following the six resulting visualizations are displayed. Each visualization as accompanied by a short description a short summary of the main advantages and disadvantages. Our implementation of 2D maps in JAVA allows the modification of many design aspects which are often neglected. It can therefore serve as a investigation tool for further experiments especially on the users interaction with such systems as well as on the aesthetic dimension of these tools.

Description: The documents are aligned according to a Latent Semantic Indexing (LSI) algorithm for reducing dimensions. Each dot represents a document:

Advantages: The scatterplot gives a good impression of the overall distribution of the documents.

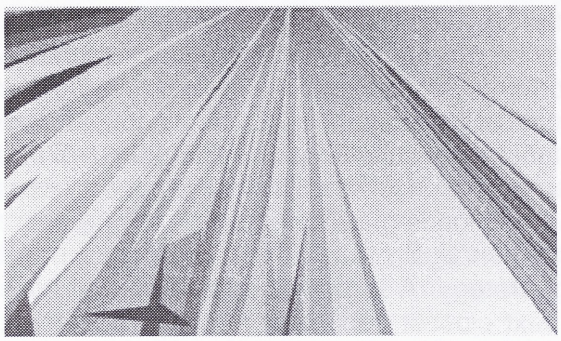
Disadvantages: The single documents are hard to recognize due to the small appearance of the dots. Bigger dots would lead to overlapping documents. In order to get to a detailed view of a special area some kind of zoom-mechanism would have to be introduced. This would harden the navigation.



Description: Same as above. In order to emphasize the documents they are displayed in red with a kind of corona.

Advantages: The documents and document clusters are easier to recognize.

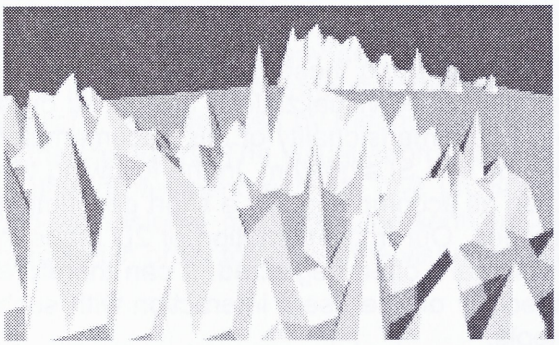
Disadvantages: Same as above.



Description: Now the document clusters are displayed as a 3D landscape. High mountains show dense clusters of documents.

Advantages: none

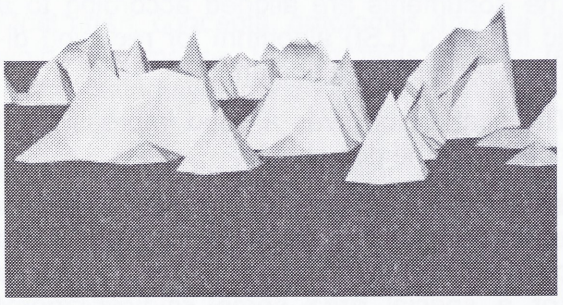
Disadvantages: This visualization is hard to navigate. The extremely high mountains are fairly confusing.



Description: The height of the mountains is reduced by introducing an algorithm based on a logarithm.

Advantages: The mountains and therefore the document clusters are very good to identify.

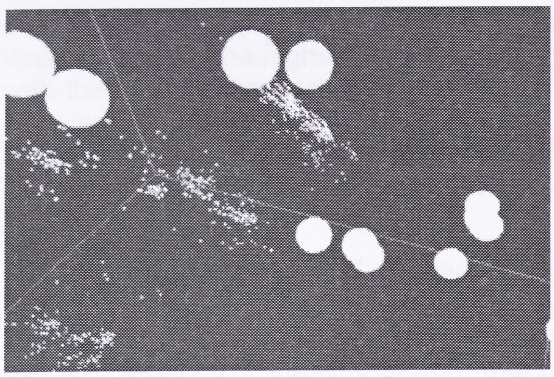
Disadvantages: Navigation in a 3D landscape is quite hard. Due to the 3D view (foreground – background) the height of the mountains is hard to compare.



Description: The height of the mountains is colour-coded according to topographic maps.

Advantages: The heights of mountains is easily comparable even if the mountains stand in different distances to the viewer.

Disadvantages: Navigation remains hard.



Description: The same document set reduced by LSI to three dimensions. The display employs the galaxy-metaphor: The documents in the front are closely related to each other and not related to the distant documents in the background.

Advantages: There is one more spatial dimension. This the dimensionality of the original set is less reduced.

Disadvantages: Very hard to navigate and very hard to interpret.

This overview on 2-dimensional maps shows the large variety of design decisions that the developer is confronted with. Depending on the requirements of the specific application, a tradeoff between the advantages and disadvantages associated with each presentation form needs to be found.

References:

Eibl, Maximilian; Mandl, Thomas (2002). Including User Strategies in the Evaluation of Interfaces for Browsing Documents. In: Journal of WSCG. Special issue: Proc of the 10th Intl Conf in Central Europe on Computer Graphics, Visualization and Computer Vision WSCG 2002, February 4-8, 2002, Plzen (Czech Republik) , vol. 10, no.1, pp. 163-169. (http://wscg.zcu.cz/wscg2002/Papers_2002/B89.pdf)

Kohonen, Teuvo (1997): Self-Organizing Maps. Springer: Berlin et al.

Honkela, Timo; Kaski, Samuel; Lagus, Krista; Kohonen, Teuvo (1997): WEBSOM-Self-Organizing Maps of Document Collections. In Proc of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6, Helsinki University of Technology, Neural Research Centre, Espoo, Finland. pp. 310-315.

Mandl, Thomas (2001): Tolerantes Information Retrieval: Neuronale Netze zur Erhöhung der Adaptivität und Flexibilität bei der Informationssuche. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft Bd. 39].

Mandl, Thomas; Eibl, Maximilian (2001). Evaluating Visualizations: A Method for Comparing 2D Maps. In: Smith, Michael J.; Salvendy, Gavriel; Harris, Don; Koubek, Richard J. (Ed.) Usability Evaluation and Interface Design. Proceedings of the 9th HCI International, New Orleans, August 5-10, 2001, Vol.1, p. 1145-1149.

Merkl, D. 1995. Content-Based Document Classification with Highly Compressed Input Data. In Proceedings of the International Conference on Artificial Neural Networks ICANN '95. Paris.. vol. 2, pp. 239-244.

Mutschke, P. (2001): Enhancing Information Retrieval in Federated Bibliographic Data Sources Using Author Network Based Stratagems. In: Constantopoulos, P.; Solvberg, I. (eds.): 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001) Darmstadt Sept. 4.-8. Berlin et al.: Springer. pp. 287-299.

Ojala, T.; Kauniskangas, H.; Keränen, H.; Matinmikko, E.; Aittola, M.; Hagelberg, K.; Rautiainen, M.; Häkkinen, M. (2001): CMRS : Architecture for content-based multimedia retrieval: In: Ojala, T. (ed.): Infotech Oulo International Workshop on Information Retrieval (IR 2001). Oulo, Finnland. Sept 19.-21. 2001. pp. 179-190.

Rauber, A.; Frühwirth, M., 2001. Automatically Analyzing and Organizing Music Archives. In 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001). pp. 402-414

Ojala, T.; Kauniskangas, H.; Keränen, H.; Matinmikko, E.; Aittola, M.; Hagelberg, K.; Rautiainen, M.; Häkkinen, M. (2001): CMRS : Architecture for content-based multimedia retrieval: In: Ojala, T. (ed.): Infotech Oulo International Workshop on Information Retrieval (IR 2001). Oulo, Finnland. Sept 19.-21. 2001. pp. 179-190.