# DETECTING AND EXPLORING SEMANTIC CONCEPTS IN UNSTRUCTURED DATA

Andreas Rauber[1,2]

Michael Dittenbach[2], Helmut Berger[2], Andreas Pesenhofer[2]

Georg Pölzlbauer[1], Nataliya Sokolovska[1], Rudolf Mayer[1],
Thomas Lidy[1], Robert Neumayer[1]

[1] *Institute for Software Technology and Interactive Systems*
*Vienna University of Technology, Vienna, Austria*
www.ifs.tuwien.ac.at
[2] *Electronic Commerce Competence Center EC3*
*Vienna, Austria*
www.ec3.at

***Abstract.*** *With the abundance of data available in electronic form, sophisticated search techniques have become crucial to find useful pieces of information. Yet, this is now gradually being superseded by the need and desire to do more than merely locate information. The value does not reside primarily within single pieces of data to be located, but in the data collection as a whole. Obtaining an overview of the amount of data available, uncovering semantic concepts and relationships between groups of data allows us to make use of an amount of data as a whole.*

*In this paper we present an overview of recent works in the field of information extraction and visualization that allow us to analyze data, provide an overview, and unveil underlying concepts, focusing primarily on applications from the field of text mining.*

## 1. Introduction

Until some time ago, locating specific pieces of data has been the primary goal of information retrieval research. Especially with the ever increasing volumes of data, high precision retrieval has been gaining importance, and will continue to remain a challenging field for quite some time to come. Yet, with this increase, the focus is also moving from data to actual information retrieval, and continuing via information mining to, ultimately, knowledge retrieval. While we

most probably are – in spite of all promising claims and profound research results – still far from really achieving solid, automatic knowledge retrieval in general disciplines, advances in information mining are substantial, allowing us to slowly climb the ladder to higher levels of semantic concepts automatically extractable from data. Core to most of these techniques still is some kind of human involvement, adding, evaluating and/or confirming semantic knowledge to the information elicited from the data by sophisticated algorithms. While this is generally considered a weakness of current technology, the holy grail being a fully automatic knowledge acquisition and evaluation system, the strengths and merits of these systems should not be underestimated, and may actually prove to be more than a mere intermediary step. Similar to the somewhat surprising potential of – by now and by comparison – rather simple data mining algorithms, a range of other rather simple approaches and tools may be used to help with in interaction and analysis of the results of data mining algorithms. They provide flexible and powerful techniques for obtaining more explicit information, and eventually move us closer to the levels of knowledge we would like to see emerge from raw data.

In this paper we will give a brief overview of our work in this area, presenting tools and approaches that assist a knowledge worker in understanding the data available, in extracting semantic concepts and communicating these. Most of these approaches are based on the Self-Organizing Map (SOM) [6], a popular neural network model providing a topology-preserving mapping from a high-dimensional input space onto a usually two-dimensional output space. While this mapping already allows considerable insight into data at hand, it is only through sophisticated visualization techniques and in combination with other machine learning and natural language processing techniques that the real potential of this approach is unleashed. We will thus provide a brief overview of both existing as well as novel visualization schemes for SOMs supporting in the identification of structure in the data. These will be combined with tools to extract content information from the data to be able to explain the structures detected.

The remainder of this paper is structured as follows: Section 2 provides a brief introduction to the SOM, followed by an overview of visualization techniques in Section 3. Sections 4 and 5 present approaches to extract semantic information from the data, based on analyzing the structure on the map as well as the textual content on the map, respectively. Current case studies and pilot applications are presented in Section 6, followed by a brief summary of the benefits of these approaches in Section 7.

# 2. Self-Organizing Maps

The Self-Organizing Map (SOM) is an unsupervised neural network that provides a mapping from a high-dimensional input space to usually two-dimensional output space [5,6]. During that process topological relations are preserved as faithfully as possible. A SOM consists of a set of *i* units arranged in a two-dimensional grid, each attached to a weight vector $m_i \in \Re^n$. Elements from the high-dimensional input space, referred to as input vectors $x \in \Re^n$, are presented to the SOM and the activation of each unit for the presented input vector is calculated using an activation function such as e.g. the Euclidean Distance. In the next step, the weight vector of the winner is moved towards the presented input signal by a certain fraction of the Euclidean distance as indicated by a time-decreasing learning rate $\alpha$. Consequently, the next time the same input signal is presented, this unit's activation will be even higher. Furthermore, the weight vectors of units neighboring the winner, as described by a time-decreasing neighborhood function, are modified accordingly, yet to a smaller amount as compared to the winner. The result of this learning procedure is a topologically ordered mapping of the presented input signals in two-dimensional space.

The SOM is traditionally being used intensively in the field of text mining [4,16] due to its capabilities of handling the very high-dimensional data spaces common in this discipline. In order to cluster text documents by content one needs to obtain a representation of their content. One of the most common representations uses word frequency counts based on full text indexing. A list of all words present in a document collection is created to span the feature space within which the documents are represented. This may be reduced by various techniques such as thresholds, stemming, and stop-word lists to obtain a representation of the documents in a feature space conventionally spanning several thousand dimensions. The words are further weighted according to the standard *tfxidf,* i.e. term frequency times inverse document frequency, weighting scheme [17]. The resulting feature vectors may further be used for SOM training.

As a result we obtain clusters of text documents on a map display, where the clusters in turn are placed according to their mutual similarity. This follows very closely the concept of traditional organization of information. These document clusters may serve as a basis for further processing such as automatic summarization or manual exploration.

# 3. Visualization Techniques for SOMs

While the SOM offers itself for direct exploration, identifying the cluster structure and the similarities on the maps still is a non-trivial task. To this end, over the years, numerous visualization techniques have been developed for the SOM. Most commonly, component planes and the U-Matrix, which both take only the prototype vectors and not the data vectors into account, are applied to visualize the map. Component planes show projections of individual dimensions of the weight vectors. If performed for each vector component, they are the most precise and complete representation available. However, cluster borders cannot be easily perceived, and high feature space dimensions result in lots of plots, a problem that many visualization methods in multivariate statistics, like scatterplots, suffer from. The U-Matrix technique [18] is a single plot that shows cluster borders according to dissimilarities between neighboring units. The distance between each map unit and its neighbors is computed and visualized on the map lattice, usually through color coding.

Recently, an extension to the U-Matrix has been proposed, the U*-Matrix [20], that relies on yet another visualization method, the P-Matrix [19]. Other than the U-Matrix, it is computed by taking both the prototype vectors and the data vectors into account and is based on a concept of data density around the model vectors. It is designed for use with Emergent SOMs [18], which are SOMs trained with a high number of map units compared to the number of data samples. Interestingly, both the U*-Matrix and our novel method, among other goals, aim at smoothing the fine-structured clusters that make the U-Matrix visualization for these large SOMs less comprehensible, although the techniques are conceptually totally different.

Methods that rely more heavily on the distribution of the data on the map are hit histograms and Smoothed Data Histograms [9], and recently proposed methods that directly show the density as graphs on top of the map [14]. Other visualization techniques include projections of the SOM codebook with concepts like PCA or Sammon's Mapping. For an in-depth discussion, see [21].

# 4. Vector field visualizations of SOMs

While all of the above visualizations assist in interpreting the thematic clusters on the SOM, sometimes a different perspective is required. This may be motivated by a different application domain where different visualization metaphors prevail. An example would be many engineering disciplines, where flow representations are a common visualization concept. Another reason may be the need to combine and superimpose different visualizations to analyze various aspects of the map structure. We thus developed a vector field visualization of SOMs which assists in finding cluster boundaries and fine-grained structures in maps at different levels of detail.

The U-Matrix is the most common tool to highlight cluster boundaries, i.e. thematic groups, on Self-Organizing Maps. However, it fails to discover these boundaries in certain circumstances, like large or sparse maps, due to the fact that only differences between direct neighbors are taken into account. We thus developed a vector field visualization of SOMs which assists in finding cluster boundaries and fine-grained structures in maps at different levels of detail. The Gradient Field method [11,13] visualizes vector fields on top of the SOM lattice such that each arrow points in the direction of the most likely cluster center and away from dissimilar regions. Also, a smoothing parameter can be adjusted to fit the desired level of detail. The more smoothing is applied, the more the resulting visualization will highlight the stronger and more global cluster boundaries. Figure 1 (left) shows examples of different levels of smooting. As an extension to the Gradient Field method, the Borderline method emphasizes on displaying the cluster boundaries instead of the direction of the most likely cluster center, as shown in Figure 1 (right). The resulting graphs may furthermore be superimposed on other visualizations, such as e.g. the aforementioned U-Matrix.

It furthermore allows to simultaneously analyze groups of component planes (e.g. according to specific concept words), revealing the semantics of various subsets of attributes in feature space [12]. This approach allows to break down the overall clustering structure into contributing factors. Thus, the correlation and potential semantic relationships between groups of variables can be investigated. An example can be seen in Figure 2, where two sets of variables are contrasted, exhibiting orthogonal behaviour.

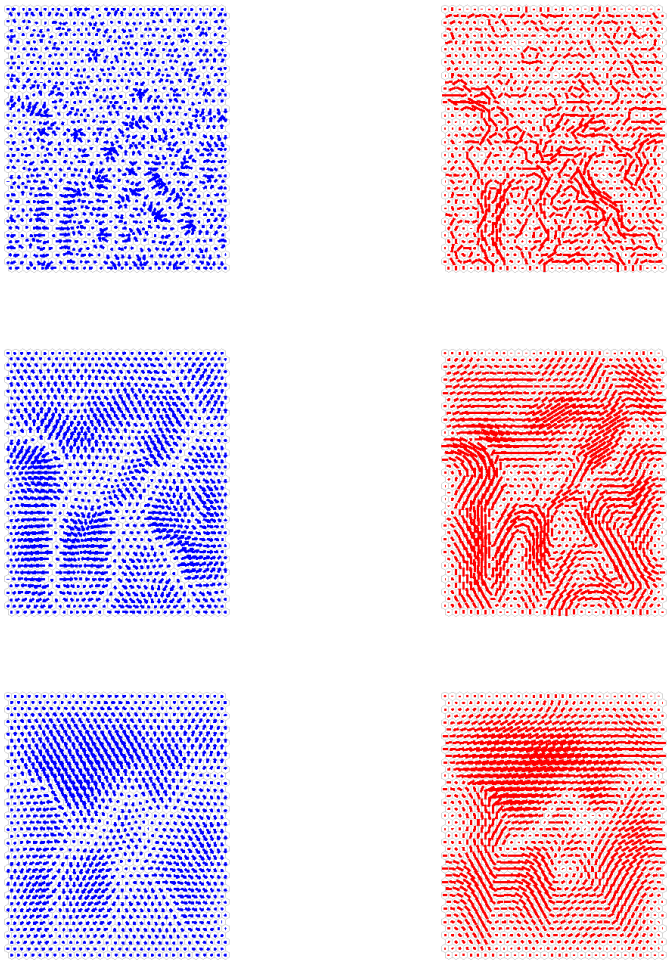


**Figure 1: Gradient Field (left) and Borderline (right) visualizations; smoothing parameter (from top to bottom): 1 (local), 5, 15 (global)**

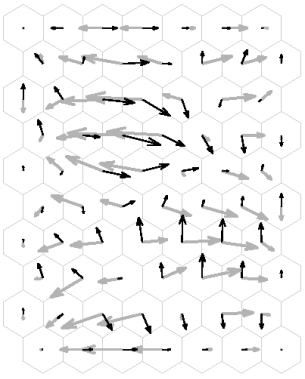


**Figure 2: Dual Gradient Field with two groups of components**

# 5. Natural Language Descriptions of Text-based SOMs

While the visualizations mentioned above provide information on the thematic structure of the documents, there still is a need for more natural descriptions of the clusters and their common characteristics. Several techniques may now be used to label the resulting maps, i.e. to extract semantic information on the clusters. Two of the most prominent techniques in this field are the keyword selection proposed by Lagus et al. [7] as well as the LabelSOM method [15]. In a nutshell, both techniques try to measure the descriptiveness and specificity of each dimension (i.e. term) for a unit or cluster. As a result, each unit or region is labelled with a set of keywords.

The quality of such labels can be drastically improved by using e.g. phrases instead of single keywords. A further step is the inclusion of part-of-speech analysis. By integrating NLP tools such as GATE [1] we can identify specific word categories such as persons, organizations, countries, dates, which are furthermore used to characterize units and regions on the map. This serves as a basis for comparative descriptions of regions on the map, highlighting common as well as distinctive semantic concepts. They may be presented both in tabular form as well as in the form of automatically generated natural language descriptions based on templates.

Figure 3 shows the different kind of annotations that are created for individual map units. By interpreting these, high-level concepts can be deducted either manually or through the use of domain ontologies, resulting e.g. in the map depicted in Figure 4.

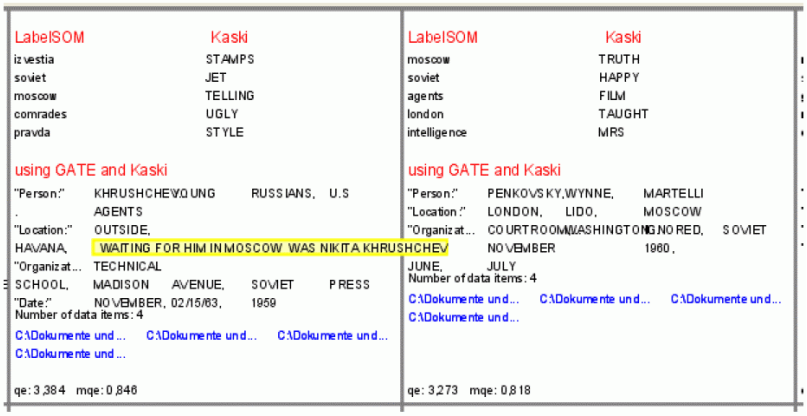



**Figure 3: Units automatically annotated with semantic concepts**

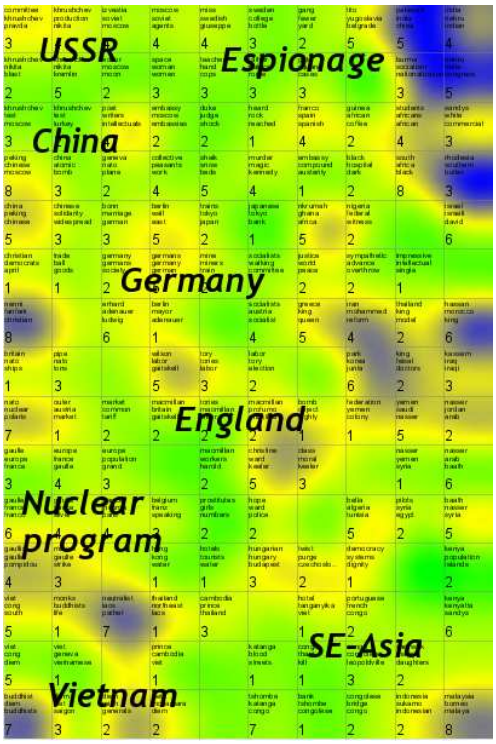



**Figure 4: Manual descriptions based on labels**

## 6. Case Studies

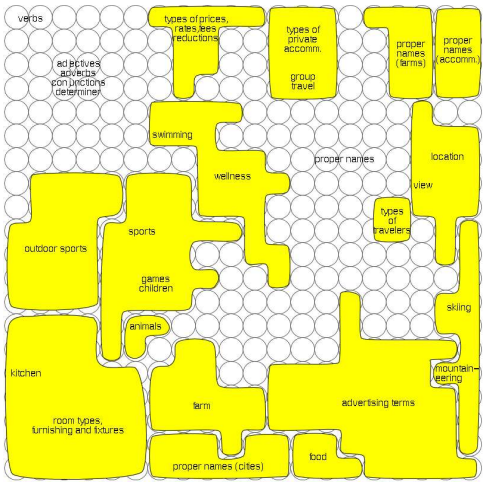
The tools and methods described above are being utilized in a number of case studies and prototype applications in research as well as industry settings, both at the Electronic Commerce Competence Center (EC3) and at the Vienna University of Technology. Apart from serving as a general tool for data mining applications, such as e.g. sensor data analysis, they are particularly powerful when it comes to the analysis of large document collections.

Case studies include the detection of topics and semantic concepts in abstracts of scientific conferences, such as the European Congress of Radiology, which features about 1.000 scientific presentations every year. A multi-annual corpus is used both for automatic topic assignment via text categorization tools [10] and for cluster analysis. With the latter we are able to obtain an idea of the topics covered as well as trends present in the event over the years.

Another case study in this direction is currently in progress, creating an information map of Austrian research activities based on the national research documentation (FODOK). Here, the various research activities are analyzed based on publication data and confronted with the classification scheme provided

by Statistics Austria for reporting purposes. Related techniques have also been used in the analysis of corpora of news articles form the Austrian Daily Newspaper "Der Standard" [2].

Similarly, the tools are being used to detect semantic concepts in free-form textual descriptions of accommodations hosted by the Tourism Information System TISCOVER. The descriptions were indexed according to text windows and clustered using the SOM. The resulting cluster structure, depicted in Figure 5 reveals groupings of terms according to concepts such as room types and furniture; sports and leisure activities; wellness offers; food; etc. These, in turn, are used to improve an ontology for natural language based searching [3].




| | |
|---|---|
| kochnische | bad |
| wanne | stockbett |
| sofa | doppelzimmern |
| badewanne | dusche |
| waschraum | schlafraeume |
| doppelbett | zimmerausstattung |
| schlafmoeglichkeiten | dreibettzimmer |
| hotelzimmer | wohnschlafraum |
| essraum | schlafzimmer |
| kochecke | zimmer |
| duschen | fliesswasser |
| kinderzimmer | einbettzimmer |
| schlafraum | komfortzimmer |
| wohnschlafzimmer | doppelschlafzimmer |
| badezimmer | schlafraeumen |

**Figure 5: Concept clusters in accommodation descriptions**

While most of the semantic concepts analyzed are on a textual level, the system is also employed on different types of data, such as e.g. audio collections. Music files are analyzed with respect to their frequency spectra characteristics and clustered by the SOM according to their sound similarity. The resulting maps can be labelled with music genre information as well as acoustic characteristics, and serve directly as an intuitive interface to large music collections [8].

# 7. Conclusions

Making use of the information buried in large volumes of data requires new techniques to assist in handling and structuring it. The Self-Organizing Map offers itself as a basic tool to provide a representation of the data on a 2-dimensional map display. Similar data items are grouped close to each other and thus relationships between data items are revealed.

By utilizing combinations of different visualization techniques, these structures can be made visually explicit and analyzed from different perspectives. This allows us to get a comprehensive understanding of the structures present in an information repository. By analyzing the content of textual clusters we are able to extract characteristic phrases as well as specific concepts. These provide a descriptive explanation of the text clusters and their specific relationship to each other via common as well as disjunct concepts. They may be used as a basis for constructing or extending formal representations of a given information repository.

While in most cases the fully automatic extraction of semantic concepts, and especially the automatic detection of sound relationships between these still remains a challenging task, these tools are available to ease many of the laborious tasks. They can assist in the comprehension of semantic structures present in large-volume document collections, as well as in the semi-automatic creation of formal representations of it.

# 8. References

[1] Bontcheva, K., Tablan, V., Maynard, D., Cunningham, H. 2004: Evolving GATE to Meet New Challenges in Language Engineering, Natural Language Engineering 10(3-4):349-373, 2004.

[2] Dittenbach, M., Rauber, A., Merkl, D. 2001: Business, Culture, Politics, and Sports -- How to Find Your Way Through a Bulk of News? On Content-Based Hierarchical Structuring and Organization of Large Document Archives In: Proceedings of the 12th International Conference on Database and Expert Systems Applications (DEXA01), Sept. 3-7 2001, Munich, Germany, Springer Lecture Notes in Computer Science, Springer, 2001.

[3] Dittenbach, M., Berger, H., and Merkl, D. 2004: Improving domain ontologies by mining semantics from text. In Hartmann, S. and Roddick, J., editors, Proceedings of the 1st Asia-Pacific Conference on Conceptual Modelling (APCCM 2004), volume 31 of Conferences in Research and Practice in Information Technology, pages 91-100, Dunedin, New Zealand, January 18-22. Australian Computer Society Inc.

[4] Kaski, S., Honkela, T., Lagus, K., Kohonen, T. 1998: WEBSOM-self-organizing maps of document collections. Neurocomputing, 21:101-117, 1998.

[5] Kohonen, T. 1982: Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43:59-69, 1982.

[6] Kohonen, T. 2001: Self-Organizing Maps, 3rd edition. Springer, 2001.

[7] Lagus, K., Kaski, S. 1999: Keyword selection method for characterizing text document maps. In Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN), volume 1, pages 371-376, London, UK, 1999. IEEE.

[8] Neumayer, R., Dittenbach, M., Rauber, A. 2005: PlaySOM and PocketSOMPlayer: Alternative Interfaces to Large Music Collections.Proceedings of the Sixth International Conference on Music Information Retrieval (ISMIR 2005), pages 618-623, London, UK, September 11-15, 2005.

[9] Pampalk, E., Rauber, A., Merkl, D. 2002: Using smoothed data histograms for cluster visualization in self-organizing maps. In Intl. Conf. on Artifical Neural Networks (ICANN'02), 2002.

[10] Pesenhofer, A., Berger, H., Dittenbach, M., Rauber, A. 2005: Applying Text Classification in Conference Management: Some Lessons Learned Proceedings of the 7th Russian Conference on Digital Libraries (RCDL 2005), pages 126-133, October 4-6 2005, Yaroslavl, Russia.

[11] Pölzlbauer, G., Dittenbach, M., Rauber, A. 2005: A visualization technique for self-organizing maps with vector fields to obtain the cluster structure at desired levels of detail. In Proceedings of the International Joint Conference on Neural Networks (IJCNN'05), pages 1558-1563, Montreal, Canada, July 31-August 5 2005. IEEE Computer Society.

[12] Pölzlbauer, G., Dittenbach, M., Rauber, A. 2005: Gradient visualization of grouped component planes on the SOM lattice. In Marie Cottrell, editor, Proceedings of the Fifth Workshop on Self-Organizing Maps (WSOM'05), pages 331-338, Paris, France, September 5-8 2005.

[13] Pölzlbauer, G., Rauber, A., Dittenbach, M. 2005: A vector field visualization technique for self-organizing maps. In Huan Li Tu Bao Ho, David Cheung, editor, Proceedings of the Ninth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'05), pages 399-409, Hanoi, Vietnam, May 18-20 2005. Springer-Verlag.

[14] Pölzlbauer, G., Rauber, A., Dittenbach, M. 2005: Advanced visualization techniques for self-organizing maps with graph-based methods. In Zhang Yi Jun Wang, Xiaofeng Liao, editor, Proceedings of the Second International Symposium on Neural Networks (ISNN'05), pages 75-80, Chongqing, China, May 30 - June 1 2005. Springer-Verlag.

[15] Rauber, A., Merkl, D. 2001: Automatic Labeling of Self-Organizing Maps for Information Retrieval In: Journal of Systems Research and Information Systems (JSRIS), Vol. 10, Nr. 10, pp 23-45, OPA, Gordon and Breach Science Publishers, December 2001.

[16] Rauber, A., Merkl, D. 2003: Text mining in the SOMLib digital library system: The representation of topics and genres. Applied Intelligence, 18(3):271-293, May-June 2003.

[17] Salton, G. 1989: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison-Wesley, Reading, MA, 1989.

[18] Ultsch, A. 1999: Data mining and knowledge discovery with emergent self-organizing feature maps for multivariate time series. In Kohonen Maps. Elsevier, 1999.

[19] Ultsch, A. 2003: Maps for the visualization of high-dimensional data spaces. In Workshop on Self Organizing Maps (WSOM'03), 2003.

[20] Ultsch, A. 2003: U*-matrix: a tool to visualize clusters in high dimensional data. Technical report, Dept. of Mathematics and Computer Science, Philipps-University Marburg, 2003.

[21] Vesanto, J. 2002: Data Exploration Process Based on the Self-Organizing Map. PhD thesis, Helsinki University of Technology, 2002.