Visualizing an Auto-Generated Topic Map

Nadine Amende [1]*, Stefan Groschupf* [2]

[1] *University Halle-Wittenberg, information manegement technology*
na@media-style.com
[2] *media style labs – Halle Germany*
sg@media-style.com

Abstract

The present work was developed within the scope of a practical training and in cooperation with media-style GmbH. The aim of this practical training was visualizing an auto-generated topic map to represent important information as topics. In times of information overload and huge amounts of documents, people have to receive the information they need. Topic maps are well suited to address this problem. We look at different approaches for visualizing a topic map: graphs, trees and maps. Graphs and trees are good for an efficient navigation and a quick access to information whereas trees provide the user with an understandable hierarchical structure. Maps represent a clear and simple overview about plenty of documents. In this paper, we discuss visualization approaches and select one being most suitable for our topic map. We elect an algorithm to realize this approach and explain its functionality. To test the quality of our map we wrote a test algorithm, which tests and if necessary improves our topic map.

1 Introduction

Everyday more information is produced. People are confronted with plenty of information, which are provided by the www as well as by companies. The result of the amount of information is information overload, which means that people can't handle all information they get and thus they can't decide which information is important or not. Additionally the user can't receive the information he needs, because he could not locate it or the information does not exist.
So, now, more than ever, is urging to take benefits from methods that can retrieve it [13]. With a topic map we can visualize essential information adapted to the users needs. Therefore it is important to create a visually comfortable topic map where users get an overview of all topics and their associations and find the information they need.
The goal of this practical training was to select a value able visualization technique to represent the topics, which were automatically extracted from a text-corpus with a text-mining tool named GATE [3,4].

The remainder of this paper is structured as follows. Section two introduces topic maps and discusses several techniques for the visualization of topic maps. A special form of map visualization is chosen for the work presented here. Section three outlines the algorithm for the generation of the map out of a given topic map. In section four some problems with the results of this special algorithm are outlined and a testing algorithm is presented, that can automatically decide, if the generated map is good enough or has to be recalculated. The paper ends with some conclusions and perspectives for future work.

2 Topic Map Visualization

To find a visualization approach for our topic map we first have to look at the basic concepts and review different visualization techniques. After that we can compare the different techniques and eventually select one, which meet our purpose to visualize the topic map.

2.1 Topic Map Basics

Topic maps are an ISO standard [10] which allows to describe knowledge and to link it to existing information resources. They are intended to enhance navigation in complex data sets and to help users identify interesting spots. Although topic maps allow organizing and representing very complex structures, the basic concepts – topics, occurrences, and associations – are simple [12].
Topics represent subjects, which can be names, organizations, concepts, locations or every other word. Information resources belonging to a topic are linked to it. These resources, called occurrences, are documents, web pages (URLs) etc. An association is a relationship between two or more topics. Topics can take a role within an association. [14]

2.2 Different Visualization Techniques for Topic Maps

We decided to choose graphs, trees and maps, because they are the main diffused techniques.

2.2.1 Graphs

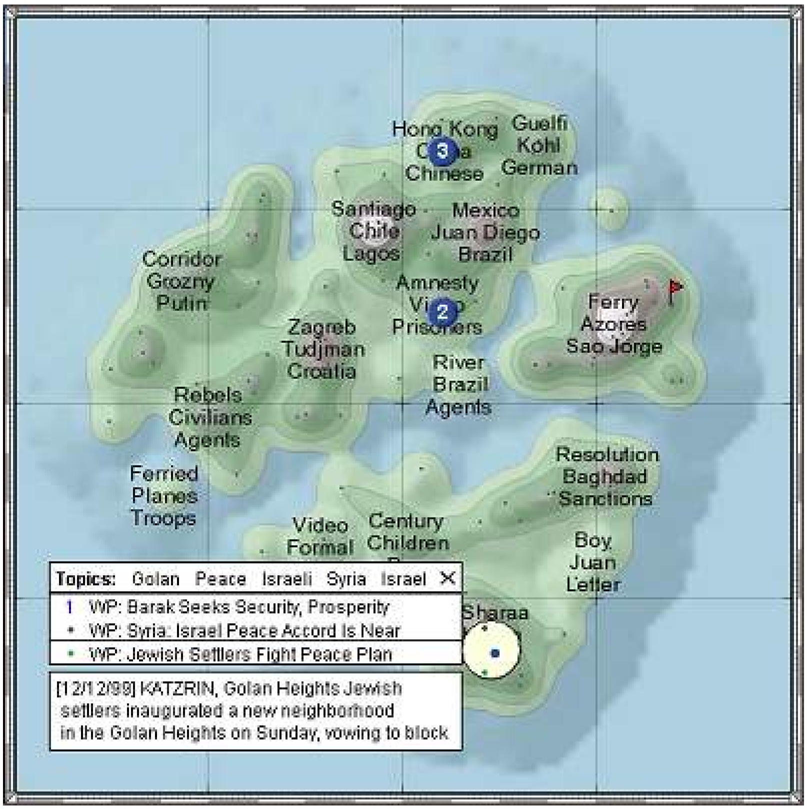
Graphs visualize topic maps as a network of nodes and edges. Whereas nodes represent topics and edges the associations between the topics. Static graph visualization shows all nodes with their associations. To avoid clutter and complexity dynamic graph visualization displays only a limited scope of nodes and associations starting from the topic of interest and its related topics. [1] The Brain [17] is an example for a dynamic graph. It creates a large browse able and searchable human-edited directory, providing a visual map of the web and documents stored on the users pc [13].

2.2.2 Trees

Trees arrange the topics and edges in a hierarchical structure, making it easier for users to interpret the topic map [12]. In this way information can be better structured. Trees are often used to visualize organization structures, computer file systems, interlinked Web hierarchies and communication hierarchies [16]. Hyper-linked trees (site maps) guide a visitor through a web site using hyper-links between nodes, which represent a structured form of the content list referent to the Web site [13]. Examples for tree visualizations used for navigation are the Microsoft Windows Explorer and Inxight’s hyperbolic tree [9].

2.2.3 Maps

Topics are arranged at a certain position on a 2- or 3-dimensional grid. The Self-Organizing Maps (SOM) Algorithm [11] can be used to calculate optimal coordinates for the topics. Cartia’s ThemeScape Map is an example for a 3-dimensional landscape map (figure 1). It is arranged like an original topographical map. Mountains display topics, whereas related mountains (topics) are placed close to each other. The mountain’s height is depending on the degree of closely related documents (occurrences) to one topic. The valleys between mountains can be interesting, because they contain fewer documents and more unique content. Avoiding complexity labels reflect only the biggest mountains on the map. The user reveals additional labels by zooming into the map. [12]



Topics: Golan Peace Israeli Syria Israel ✕

1 WP: Barak Seeks Security, Prosperity

• WP: Syria: Israel Peace Accord Is Near

• WP: Jewish Settlers Fight Peace Plan

[12/12/99] KATZRIN, Golan Heights Jewish settlers inaugurated a new neighborhood in the Golan Heights on Sunday, vowing to block

(Figure 1: ThemeScape Map, Cartia Inc., [5])

2.3 Comparison of the Techniques

Using graphs or trees for visualizing a topic map, users will benefit from a good navigation between the topics. Graphs and trees techniques concentrate on navigation through hyperlinks whereas maps or landscape maps concentrate on topics representation. Good representation helps users to find interesting topics. An efficient navigation is important for quick access to the topic of interest.

An advantage of graphs and trees is that they can display more details like roles, different associations as well as association types, topic types and occurrence types using different colors or shapes. [15] Whereas maps and landscape maps can display proximity and the amount of occurrences for one topic or a topic cluster. The amount of occurrences shows the importance of a topic, which can be significant information [5].

A disadvantage of static graphs and trees is that they present all topics, which could be very complex and confusing. Dynamic graph and tree visualization reduces this problem, because it limits the scope of visualized topics and thus avoids cluttering and shows the user only the topic of interest with its directly related topics [1]. Furthermore users have only access to a limited amount of occurrences. Maps and landscape maps nearly completely avoid the problem of complexity. They provide the user with a clear and easy overview. Despite reduced complexity users can access millions of documents (any amount of documents can be linked to the visualized topic).

Another disadvantage of graphs and tress is that the user can feel kind of lost because he has to navigate through plenty of topics. However, maps and landscape maps can use a zoom function (cf. Cartia's ThemeScape Map [5]), which means that the overview of the map presents only topic clusters by label. Clusters are combined topics having similar content or occurrences. When we use the zoom function the cluster will be split in other sub-clusters and several topics.
Using zoom the user will keep an overview of his position in the map and can easily go back to the starting point. [12]

2.4 Selection of a Visualization Technique

We chose overview, representation and direct access to all occurrences as criteria for selecting a visualization technique.
Visualizing a topic map means to focus on representation or navigation purpose. We decided to generate a landscape map based on representation purpose according to the ThemeScape Map from Cartia Inc. [5].

Landscape maps are not too complex and provide the user with a clear and easy to understand overview. Graphs and trees are not very well suited for representing a topic map, which contains millions of topics and occurrences [12]. An advantage of landscape maps is that they have an unlimited scope of visualization, so users get access to all topics and occurrences from every position. Another point for electing landscape maps is that they can visualize the proximity between topics and they can present the amount of occurrences belonging to a topic. Thus users will receive essential information about the importance of one topic and the context of all topics [5].

3 Approach for Visualizing an Auto-Generated Topic Map

The chosen visualization technique and additional concepts are the basis for the visualization approach of our topic map. To realize this approach we need an algorithm.

3.1 Algorithm

At first we elect an algorithm, which meets our visualization purpose. After that we describe the functionality of the algorithm.

3.1.1 Election of an algorithm

For visualizing our topic map as a landscape we used the force directed placement algorithm [2,6] based on numerical scaling, which computes optimal coordinates for the position of every topic depending on its associations. Another common algorithm for the landscape metaphor is called Self-Organizing Map (SOM) [11]. The use of taking force directed placement instead of SOM depends on our purpose for visualizing the Topic map. SOM is a neural network model, which analyses every document (using high-dimensional input vectors) and puts it on a 2-dimensional grid depending on the other documents already present. Similar documents are placed near each other whereas dissimilar documents get a long distance between them. [2,8,11]

The purpose of this work was to find a possibility to visualize topics (e.g. Germany, Berlin) and associations (e.g. Berlin is the capital of Germany.) extracted from documents by the text-mining tool GATE. GATE was developed further by media-style [3,4]. Therefore our topic map doesn't focus on visualizing document collections (Comp. SOM), but on visualizing topics and associations extracted from the documents.

Force directed placement was well suited for our purpose, because it arranges objects in a low-dimensional space using weighted graphs or object relationships described by similarity. The object similarities are interpreted as forces and the algorithm tries to find a force-balanced state [2]. In our case forces are the associations (relationships) between two topics (objects). For calculating the proximity the number of associations between two topics are handed over as the input distance value. All types of associations were equally weighted.

Because of using associations as a factor for calculating our map we can express the coherence of the topics so that users easily get to know how and why topics are connected. While SOM displays only similar documents in the same neighborhood [16], our topic map shows different documents (occurrences) belonging to one topic, because a topic could appear in different documents. This will provide the user with a better result of what he is searching for. He will find every document belonging to the topic of interest.

Another shortcoming of SOM approaches is that they don't allow the user to influence the document comparison [2]. Our force directed placement algorithm operates with topics stored in a database. Thus users can select the topics, which should be visualized.

3.1.2 Functionality

The force directed placement algorithm supports our idea of a landscape map best. We wrote the algorithm in java, because of java's compatibility with other tools, its simple ease of learning and understanding and its object-orientation.

The algorithm assigns random start coordinates for every topic being extracted and saved in a database. For computing the distance between two topics the algorithm needs to have an input distance. This input is depending on the number of associations between a topic and its direct related topic. The more associations the less distance between two topics, which means that they occur in similar documents and thus they are related closer to each other on the grid. Based on the input values the algorithm computes the optimal coordinates.

We have to consider that the algorithm will not compute a perfect topic map, but a value able and useful one with some restrictions, which means that the coordinates of the topics have to adapt to their environment. Two Topics, which do not have any associations between them, should have a long distance between each other, but because of their related topics, which can be linked to each other, they can have a smaller distance than expected.

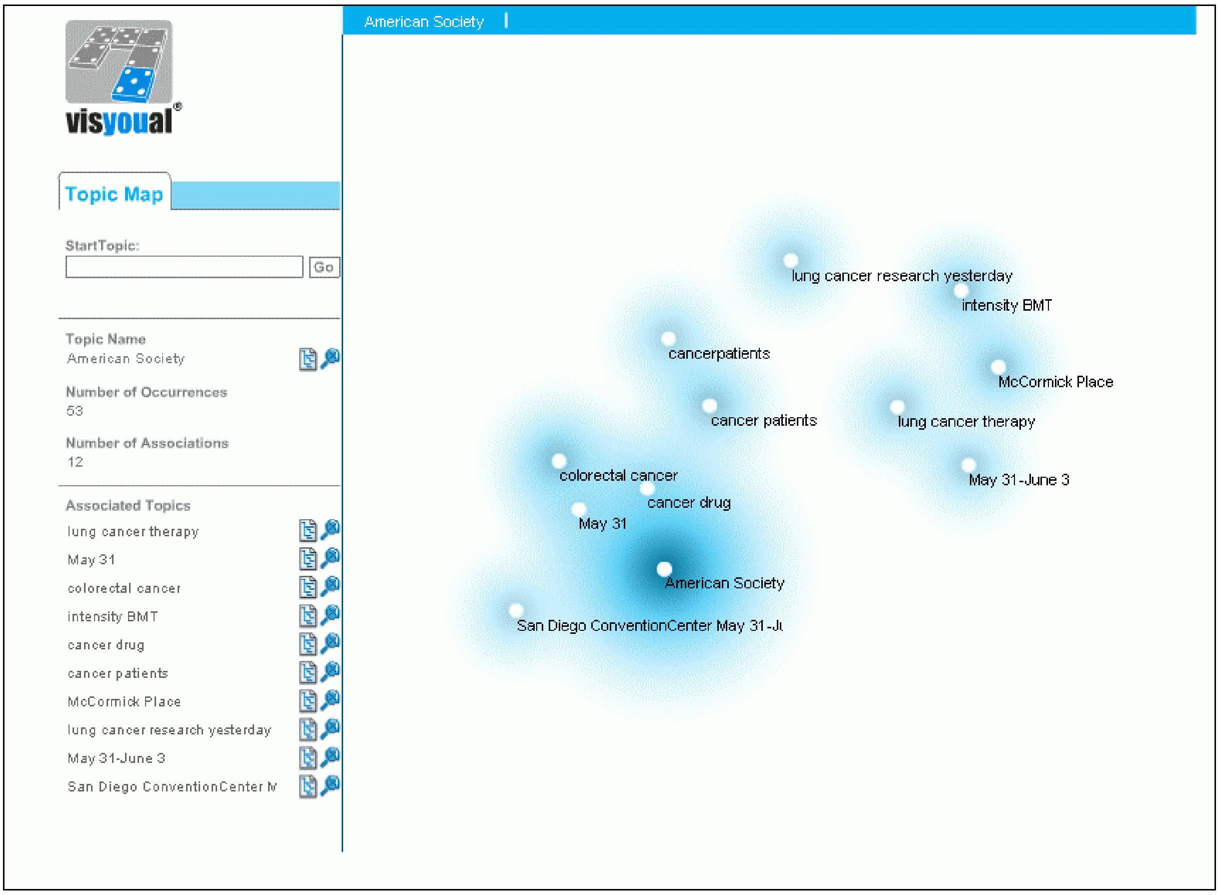
3.2 Visualization

The optimal coordinates being computed by the force directed placement algorithm are arranged on a 2-dimensional grid. Only the 30 biggest topics are visualized in every zoom level, namely those that have the most direct associations. We decided to use a zoom function and not to visualize every topic because there could be thousands of topics being extracted from the text corpus and thus the map will not provide the user with a clear and understandable presentation.

For creating a zoom function we were inspired by WEBSOM. [18] There, users may view any area of the map in more detail by simply clicking on the map image with the mouse [8]. By doing that smaller topics (labels), which were not visualized before, can appear on the map. The new generated map is a part of the initial map and thus displays only topics belonging to the chosen section of the map.

We decided to provide users with an advanced function for zooming in our map. The basic function was developed according to the WEBSOM zoom function. It allows us to reveal every topic especially the smaller ones with fewer associations. The second possibility to zoom in is to click on a node (white dot in the our map cf. figure 2) and reveal a new map showing only nodes, which are directly related to the parent node. This saves time for searching the associated topics, because the basic zoom function will only display the 30 biggest nodes (with the most associations) in every zoom level.
If the user wants to avoid zooming, he can also just search for one topic by using the search function.

For visualizing the topics they are displayed as blue circles presenting the mountains of our landscape map (figure 2). These circles can have different sizes depending on the number of occurrences (e.g. documents, web pages) belonging to one topic. More occurrences create bigger mountains (circles) colored with a darker blue. Topics or mountains having many associations between them are closely related to each other. [5]
Additionally users will get important information to a topic by a tool tip, e.g. the names of the ten biggest direct associated topics, the number of occurrences and the total number of associated topics.



Topic Name
American Society

Number of Occurrences
53

Number of Associations
12

Associated Topics
- lung cancer therapy
- May 31
- colorectal cancer
- intensity BMT
- cancer drug
- cancer patients
- McCormick Place
- lung cancer research yesterday
- May 31-June 3
- San Diego ConventionCenter M

(Figure 2: Visualized Topic Map)

4 Testing the Results

The elected algorithm causes some visualization problems. So it is important to test the results of the algorithm and to recalculate the topic map in order to get improved results.

4.1 Visualization Problems

A major problem of the algorithm is that topics can be displayed at a wrong position in the map. This means that two topics, which should be near one another can be far away from each other, or a topic with fewer associations is nearby its associated topic instead of being far away. [7]
These problems can arise, because the algorithm calculates adapted and not perfect coordinates. During the calculation the coordinates of the topics adapt to and organize each other. For visualizing the topics on a 2-dimensional finite map, the adaptation process can't consider each constellation of topics and associations.

For improving the quality of our landscape map, we developed a test algorithm (cf. section 4.2), which can detect wrong positions (coordinates) and calculate them again in order to get better results.

Another problem of the force directed placement algorithm is that every new calculation can cause a different visualization of the same topics. This depends on the arbitrary initial values of coordinates [2] at the beginning of the calculation. So if the user adds some topics in the database and calculates the map again, the map's appearance wouldn't be the same. This won't meet our goal for creating a simple, understandable and visually valuable topic map. The old topics should be displayed at the same position in the map and the new added topics have to fit in.

We solved this problem by giving the existing topics their old coordinates stored in the database instead of arbitrary ones. The newly added topics will receive arbitrary coordinates. The result is a map with the minimum of changes necessary for placing the new topics.

4.2 Test Algorithm

Our test algorithm is based on a comparison between input and output values. Whereas the input value is the initial distance (edge) between two topics depending on the number of their associations and the output value is the calculated distance (edge). Let d be the input value and h the output value. For comparing the input values with the output values it is not sufficient to consider only the absolute values, as the calculated output distances can have smaller or larger values than the input distances. We have to consider the ratio ($R_{i,k}$) between the distances (edges *i* and *k*) before and after calculation, because output values can differ from input values, but the ratio between two edges has to remain constant. A ratio is the calculated difference between the input and output values quotient of two edges (*i,k*). If the ratios remain constant, the topics (nodes) will be displayed at the right position in the map.

$$R_{i,k} = \frac{d_i}{d_k} - \frac{h_i}{h_k} \qquad i > 0; i = 1(1)n$$

$$k > 0; k \neq i; k = 1(1)m$$

For getting a newly calculated, high quality map, every distance (edge) has to be compared with every other distance.
Because of the increasing calculation time caused by the amount of topics and their distances, we decided to run a smaller test. Each distance is only compared to 1000 randomly selected distances of all existing distances (edges). We chose 1000 because it leads to the lowest calculating time in relation to significant test results, so that the cost-utility ratio is optimal. We can say that the results are significant, because they don't vary too much from the test results, which compares every distance with all other distances.

After calculating all ratios we sum them up and divide this value by the number of ratios (*n*). The resulting value is the average of all ratios (Avg) and thus says something about the average deviation of a ratio between two edges before and after calculation.

$$Avg = (\sum_{j=0}^{n} R_{i,k}) \Big/ n \qquad j = 1(1)n$$

If there is no deviation the value is 0. The more average deviation the worse is the quality of the map, which means that many topics (nodes) are displayed at a wrong position.

By looking at each ratio deviation and knowing how bad every single ratio is we can recalculate the topics being involved in this ratio. Having not enough calculation capacity, because this increases proportional to the number of topics, we decided to specify a scope for the ratio deviation. We chose 5% of the largest deviation value as the scope, whereas the largest deviation arises from the screen width. All ratios being out of this scope are potentially bad. Based on these potential ratios we determine the 20% worst topics whereby 4 topics are involved in one ratio. Here we are again restricted by the capacity of the computer and the computing time.
The next step is to recalculate the coordinates of these topics and thus improve the quality of the map.

4.3 Recalculation

Within the recalculation of the topic map the coordinates of the worst topics are recalculated based on random start values, whereas the remaining topics are recalculated based on their coordinates in the database. In this way the remaining topics do not change their position or change it minimal and the worst topics will adapt to them. After the recalculation the new coordinates of all topics are recorded into the database.

After the recalculation the topics are tested again using their new coordinates. If the new average deviation is smaller than the previous the map quality has improved.

5 Conlusions and Perspectives

We developed a landscape map providing the user with a simple and clear overview about thousands of topics automatically extracted from a huge text-corpus. The user can identify the coherence between the topics and the importance of single topics. Besides this we enable a comfortable navigation to all documents being linked to the topics.

During our work we found out that some topics are displayed at a wrong position on the map. So we wrote an algorithm testing the coordinates of the topics. After doing this test multiple times with the old and the new recalculated coordinates, it is clear that the quality of the map has risen. However, using only restricted values and algorithms to test and recalculate the map, we will not receive a map reaching the quality we aim for.
For calculating a high-quality topic map we didn’t have enough time and calculating capacities. So we decided to set some restrictions in our test algorithm, which produces not a perfect but an acceptable map. The aim of this work was to develop a Topic map prototype, not a final solution.

For the future work it is important to use powerful computers or to minimize some iterations and functions in the algorithm. Another step could be to use the SOM algorithm instead of force directed placement, because SOM can handle thousand of documents [2] in a reduced calculation time. Therefore we have to transfer the visualization of documents [8,11] into a visualization of extracted terms and associations. To understand SOM’s functionality is more difficult than the force directed placement algorithm. We weren’t able to realize it in this short period of the practical training.

Another future feature will be to visualize the verbs between two topics by mouse over effect. This will provide the user with additional information about the coherence of two topics.

6 References

[1] Ahmed, Kal (2000): Topic Maps for Repositories, Proceedings of XML Europe 2000, URL: http://www.gca.org/papers/xmleurope2000/papers/s29-04.html, last access at: 10/08/2003, 17:27

[2] Becks, Andreas (2001): Visual Knowledge Management with Adaptable Document Maps, URL: http://sylvester.bth.rwth-aachen.de/dissertationen/2001/105/01_105.pdf, last access at: 10/31/2003

[3] Cunningham, H. et al. (2003a): GATE – A General Architecture for Text Engineering, Version 2.2, 2003, URL: http://gate.ac.uk, last access at: 10/30/2003

[4] Cunningham, H. et al. (2003b): Developing Language Processing Components with GATE (a User Guide), For GATE Version 2.1, 2003, pp 86-98, URL: http://gate.ac.uk/sale/tao/tao.pdf

[5] Dodge, Martin (2000): NewsMaps: Topographic Mapping of information, 2000, URL: http://mappa.mundi.net/maps/maps_015/#ref_2, last modified at: 2000

[6] Force-directed Layout, aiSee User Manual, 2003, URL: http://www.aisee.com/manual/unix/56.htm, last access at: 10/31/2003

[7] Germano, Tom (1999): Self Organizing Maps, URL: http://davis.wpi.edu/~matt/courses/soms/#Scale%20Neighbors, last access at: 10/09/2003, 12:51

[8] Honkela, T.; Kaski, S.; Lagus, K.; Kohonen, T. (1997): WEBSOM – Selforganizing Maps of Document Collections, in: Proceedings of WSOM'97, Finland, pp. 310-315, 1997, URL: http://www.websom.hut.fi/websom/doc/publications.html, last modified at: 10/26/2000

[9] Hyperbolic Tree Map, 2002, URL: http://www.inxight.com/map, last access at: 10/30/2003

[10] ISO/IEC 13250: Topic Maps – Information Technology, Document Description and Processing Languages, URL: http://www.y12.doe.gov/sgml/sc34/document/0129.pdf, last access at: 1999

[11] Kohonen, T. (2001): Self-Organizing Maps, Berlin et al., 2001

[12] Le Grand, B.; Soto, M. (2003): Topic Maps Visualization in: Geroimenko, V.; Chen, C. (Eds.): Visualizing the Semantic Web. XML-based Internet and Information Visualization, Great Britain 2003

[13] Oliveira, Rocha Alexandre (2000): Deployment of Topic Maps for Navigation and Searching in Huge Information Spaces as Component of Learning Environment, 2000, URL: http://www.ccg.pt/Publications/_PDFs/Theses/2000/Diploma_Thesis.pdf, last access at:

[14] Pepper, Steve; Moore, Graham (2001): XML Topic Maps (XTM) 1.0, 2001

[15] Pepper, Steve: The TAO of Topic Maps. Finding the Way in the Age of Infoglut, URL: http://www.ontopia.net/topicmaps/materials/tao.html, last access at: 10/17/2003, 12:51

[16] Rohrer, R. M.; Swing, E. (1997): Web-Based Information Visualization, in: IEEE Computer Graphics and Application URL: http://www.cs.duke.edu/courses/spring03/cps296.8/papers/web_infoVis.pdf, last access at: 10/30/2003

[17] The Brain Homepage, URL: http://www.thebrain.com, 2000, last access at:

[18] The WEBSOM research group (1999): http://websom.hut.fi/websom/comp.ai.neural-nets-new/html/root.html, last modified at: 10/17/1999