X. Self Organizing Maps,
a Visual Exploration Tool in Datamining

drs. Martijn Schuemie[1], dr. ir. Jan van den Berg[2], drs. Roger Tan[3]

x.1 Introduction

We humans are good at detecting visual patterns. Driving a car, scanning the headlines of newspapers or noticing that a single tile is out of alignment in your bathroom, during most of our daily activities we rely heavily upon the information we can extract from what our eyes can see. Our affinity for visual imagery can be used to effectively and efficiently convey information. This has been done for ages. Graphs and charts have been used to give better insight into data than is possible with numbers alone.

This paper is about Self-Organizing Maps (SOMs), an algorithm that can be used to turn data into something we can see and understand: a two-dimensional map. Tools like SOMs are invaluable in datamining. They provide the means to make high-dimensional data understandable and, perhaps more importantly, explorable for the user. These visual exploration tools give the opportunity to first get a look and feel of the complex data; they do not require the user to ask the right questions right away.

First of all, this paper will elaborate on what SOMs are and how they work, followed by some examples of SOMs applied in datamining and closing off with a look at the future of visual exploration tools.

x.2 Self-Organizing Maps

The Self-Organizing Map algorithm was introduced around 1982 [Kohonen, 1982]. Ever since, several thousands of scientific papers have been written both on applications of SOMs and on mathematical analyses of the algorithm (for an overview, we refer to [Kaski et al., 1998). In addition, several textbooks were published containing the theory and/or applications developed using a SOM. Last but not least, various software packages have been developed facilitating the application of the SOM algorithm (for an overview of these, see [Deboeck et al.,1998], ch. 13).

Basically, a Self-Organizing Map consists of a grid of units. This usually two-dimensional grid is put in a high-dimensional data space, also called the input space. Our goal is to infer 'knowledge' from the input space, i.e., we want to understand the structure of this space. By putting the grid in

[1] Delft University of Technology, Faculty of Information Technology and Systems, http://is.twi.tudelft.nl/~schuemie/

[2] Erasmus University Rotterdam, Faculty of Economics, Department of Computer Science,
http://www.few.eur.nl/few/people/jvandenberg/

[3] Robeco Group N.V., Quantitative Research Dept., tan@mediaport.org.

the input space, every two-dimensional unit of the grid corresponds - at the same time - to a high-dimensional 'reference vector' in the given data space.

To illustrate, the height, weight and age of a person can be used respectively as *x*, *y*, and *z* coordinate of a point in a three-dimensional input data space. Figure 1a shows a number of samples of another population plotted in some three-dimensional input space. In an attempt to map this data onto a two-dimensional map, the reference points of the SOM-units are placed in this space. In the beginning these points can be placed at random or, to speed up the algorithm, using some form of linear initialization as depicted in figure 1b. SOMs use an iterative algorithm to move the reference points from the initial state to a state where each reference point is close to a cluster of data points in a way as has been shown in figure 1c.

Figure 1a: Points in a data space

Figure 1b: Linear initializations of the reference vectors of a SOM

Figure 1c: Reference vectors after applying the SOM algorithm

To a certain extent, the structure of the three-dimensional space is mapped on the two-dimensional SOM: the distribution of the reference points of the SOM mimics that of the original data. In addition, these reference points are also part of the two-dimensional grid and linked together in a way as shown in figures 1b and 1c. By doing so, any data point in the input space, even a new one, can now be represented on a unit in the grid, simply by finding the reference point closest to that data point. The result is a two-dimensional map of the three-dimensional space. The most important feature of this map is that it is *topology-preserving*: points that are near each other in the input space are also near each other on the map.

If we look at figure 1c, we see that there are few units with reference vectors in between the clusters. The result on the map will be that units on the edge of one cluster are only a few units away from units on the edge of another cluster, even though in the data space the distance between them is relatively big. In order to visualize the distances between the reference vectors in the original input space, a color-coding is introduced in the map as shown in figure 2. Here the darker colors indicate a large distance between reference vectors and bright colors a short distance

Figure 2: Two-dimensional map of the three-dimensional space with color-coding.

On the map we can now clearly see the distinction between the lower cluster in the data space (on the right of the map) and the two top clusters (on the left). The distinction between these last two is somewhat vaguer since these clusters are not too far apart in the input data space.
Additionally, other color coding schemes can be applied. For instance, a coding based on the value of one of the components in the input space can provide the user insight in the distribution of this one characteristic over the data space. A good example of the use of such 'component planes' can be found in chapter Y of this book.
The example used here is a relatively simple one; a three-dimensional space is still comprehendible and the clusters already lie almost in the same linear plane, making it also possible to solve the problem using a simple linear mapping.

Figure 3a: Linear regression with a non-linear data set

Figure 3b: A one-dimensional SOM with the same non-linear data set

Figure 3a and 3b show a - in a certain sense - more complicated example where linear methods such as linear regression fall short. The input-space is two-dimensional here while the grid (or map) is simply one-dimensional this time. The three clusters do not lie on the same line, and the regression model depicted in figure 3a retains very little of the information in the original data. In figure 3b, we see a one-dimensional SOM plotted in the two-dimensional space, providing a better fit for the data. In the SOM, it is still clear that medium *x* values are related to higher *y* values and that both lower and higher values of *x* correspond to lower values of *y*.

As described here, SOMs can make a non-linear projection of a higher dimensional space onto a space with lower dimensionality while preserving topology. Sometimes this output space is used as input for other algorithms, for instance to classify input data. Most often however, SOMs are used as an interface to the user. The often two-dimensional map, when provided with appropriate labels, can provide the user with unique insights into the data. The next paragraph will show some applications where the SOM is used specifically to communicate complex information to the user.

x.3 Examples of SOM datamining applications

x.3.1 Information Retrieval using ACS-WEBSOM

The field of Information Retrieval (IR) is concerned with enabling users to find specific information in large collections of documents. Good examples of IR-systems are search-engines on the web such as Altavista and Yahoo, and catalogue-systems in libraries. Most such systems require the user to formulate a query, and words in this query are matched against words in the document title, authorname, abstract and/or keyword list. This requires the user to already have a fairly good idea of what to look for. The user must already be somewhat familiar with the domain and the terminology used in it.

A more visually oriented alternative is offered by algorithms such as the WEBSOM algorithm [Honkela et al., 1998] and our own extension, the ACS-WEBSOM algorithm [van den Berg et al., 1999] The notion underlying these systems is that documents can be characterized by the concepts addressed in them. Concepts are represented by clusters of words. The clusters themselves are found using a SOM and an additional learning and forgetting algorithm. For a detailed description of this process, we refer to [Schuemie, 1998].

Having done this, documents can be represented by points in another high-dimensional space where each dimension represents a concept as found in the previous step. For example: if a document contains many words found in the word cluster of concept *x*, then it’s *x*-coordinate will be relatively high. This high-dimensional document space contains all the documents, arranged in such a way that documents that address similar topics are found close to each other. Again a SOM is used to make this space, that usually has several hundred dimensions, available to the user.

Figure 4: User interface of an IR-system using the WEBSOM algorithm. From left to right: 1. Starting screen with an overview of the collection. 2. Close-up of a specific region of the map. 3. A list of the documents related to a specific unit on the map. 4. Text of a document

The resulting document-map is labeled either manually or using a simple algorithm. If the map is very large it can be made zoomable, as shown in figure 4. The user can now explore the document collection by investigating the map. Clicking with the mouse on a unit on the map results in a list of the documents related to that unit. Documents in the same unit or nearby units are semantically related, allowing the user to quickly find related information.

(An online demonstration of the WEBSOM is available at http://websom.hut.fi)

x.3.2 Other Applications of the SOM

In chapter Y of this book another application of the use of SOMs is presented, namely in the area of assessing the creditworthiness of a company based on a large set of financial data concerning that company.

In [Deboeck et al.,1998], several other applications in mainly the area of finance are presented ranging from a SOM clustering of more than 100 Scotch whiskies based on 72 different whisky features, to a SOM of about 50 mutual funds based on 15 fund features and intended to create a better basis for portfolio selection, and to a SOM intended to get a better understanding of the trends and patterns among today's emerging markets.

In [Kaski et al, 1998], an extended overview of SOM applications of all kind can be found. The articles cited concern applications in fields like machine vision and analysis, optical character recognition, speech analysis, signal processing and telecommunications, process control, robotics, mathematical problems, neurobiology, and more. In addition, many papers are devoted to mathematical analyses of the SOM algorithm and its extensions including convergence proofs, its relation to other mathematical fields like Markov-processes, energy-function formalisms, and Baysian learning approaches.

In general, one may conclude that all SOM applications exploit one or more of its three 'basic properties' [Haykin, 1994]:

1. The Self-Organizing Map (represented by the grid points in the output space, each of which corresponds to a reference vector in the input space) provides a good approximation of the input space.
2. A Self-Organizing Map is topology preserving which means that nearby grid points on the map correspond to nearby patterns in the input space.
3. Regions in the input space having a high density of data points are mapped onto larger domains of the output map. In this sense, a Self-Organizing Map reflects the statistics of the input data distribution.

x.4 Future

Around 20 years ago, the SOM-algorithm was invented, but it took quite a long time for getting this algorithm available in an 'easy-to-use' way. Fortunately, nowadays user-friendly SOM–type software packages are available like Viscovery SOMine [Eudaptics, 1999]. This type of packages has a visual interface simplifying the work of the data miner: (S)he is now able to concentrate on what (s)he is interested in, namely, on the discovery of knowledge. Besides creating the map (after automatically having executed several preprocessing tasks), Viscovery helps the data miner by converting the trained SOM into *visual information.* Several advanced data analysis tools are available within this package like data cluster search, numerical information retrieval including cluster statistics, data dependency evaluation, and more.

However, interpreting all emerging views on the data remains almost completely a task for the data miner. Like in all areas of data mining, *model validation* is of great importance here. One could imagine that future software packages will become more intelligent by, for example, warning the user automatically when some structure has become visible which – after a more thorough statistical or other type of analysis – appears to be accidental instead of structural. In addition, techniques like cross-validation [Haykin, 1994] in order to assess the generalization capabilities of a SOM found, might be performed in an automatic way to further support the future data miner.

Other improvements of the SOM can be achieved by further facilitating the human-computer interaction. Generating a SOM requires interaction with the user. Parameters for the SOM algorithm have to be set, and for example the user has to determine the basis for the color-coding, and this basis is often changed when trying to interpret the map. This interaction currently requires the user to have a great deal of knowledge of the algorithm. Perhaps in the future the computer can aid the user in a dialogue fashion in refining and adapting the map in an interactive way to fit the users need.

Also, the SOMs displayed to the user currently are all two-dimensional maps, while our perceptual system is based primarily on our environment that is three-dimensional. A 3D SOM could, in some applications, not only fit our way of viewing better, but could also be able to show intricacies of the input data not visible on a 2D map. How best to display a 3D map is still unclear however.

Summarizing, we think that SOMs will become a standard tool for dataminers probably both in 3D and in 2D, available within standard office packages. In addition, we foresee that other machine learning algorithms will be invented based on the idea of visualizing information since visual information is an efficient and effective ways for rapid and correct data interpretation.

x.5 References

- van den Berg, J., Schuemie, M., *Information Retrieval Systems using an Associative Conceptual Space,* Proceedings of the 7th European Symposium on Artificial Neural Networks (ESANN'99*),* ed. M. Verleysen, pp. 351-356, Bruges, Belgium (1999).
- Deboeck, G., Kohonen, T., *Visual Explorations in Finance with Self Organizing Maps*, Springer-Verlag (1998).
- Eudaptics, *Viscovery SOMine*, http://www.eudaptics.com/ (1999).
- Honkela, T., Lagus, K., Kaski, S., *Self-Organizing Maps of Large Document Collections*, in: [DeBoeck et al., 1998], ch. 13, pp. 168-178.
- Haykin, S., *Neural Networks, A Comprehensive Foundation*, MacMillan (1994).
- Kaski, S., Kangas, J., Kohonen, T, *Bibliography of Self-Organizing Map (SOM) Papers*: 1981-1997, available at http://www.cis.hut.fi/research/refs/ (1998).
- Kohonen, T., *Self-Organized Formation of Topologically Correct Feature Maps*, Biological Cybernetics, **43**, 59-69 (1982).
- Schuemie, M, *Associatieve Conceptuele Ruimte, een vorm van kennisrepresentatie ten behoeve van informatie-zoeksystemen*, Master Thesis, Erasmus University Rotterdam (1998).
  (available at
  http://www.few.eur.nl/few/people/jvandenberg/masters.htm (#23))

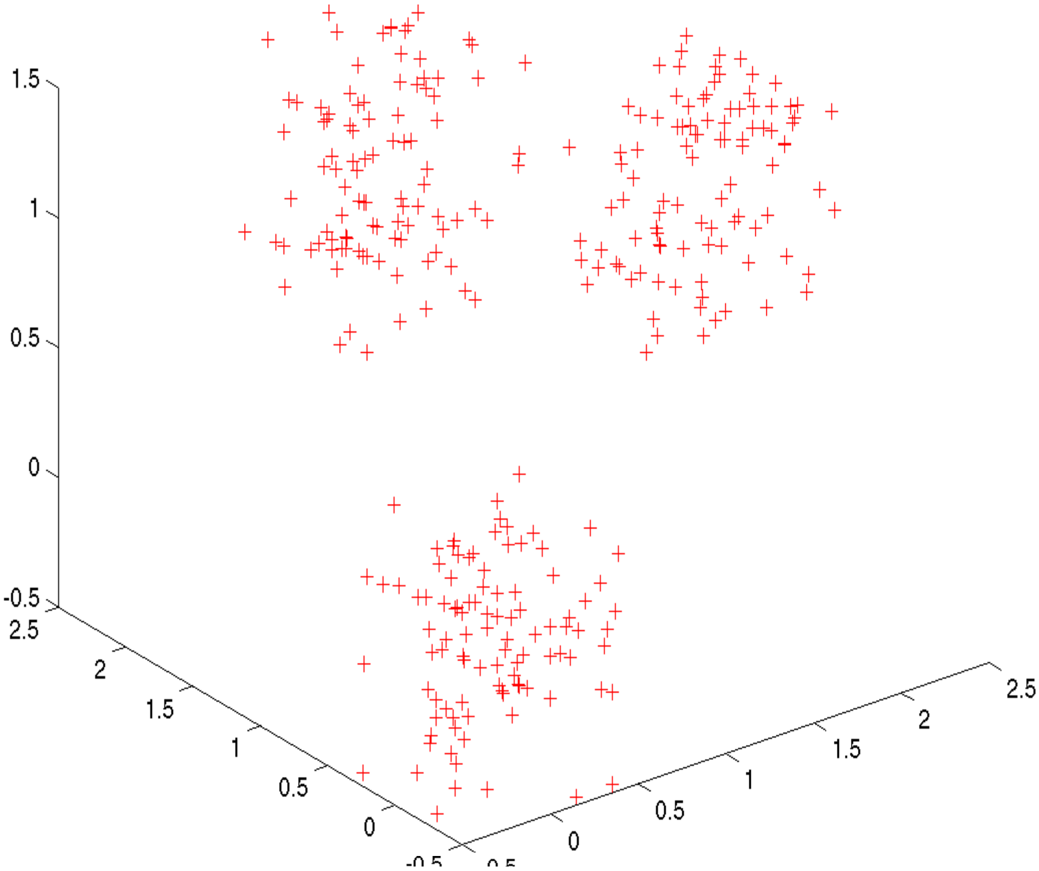


Figure 1a

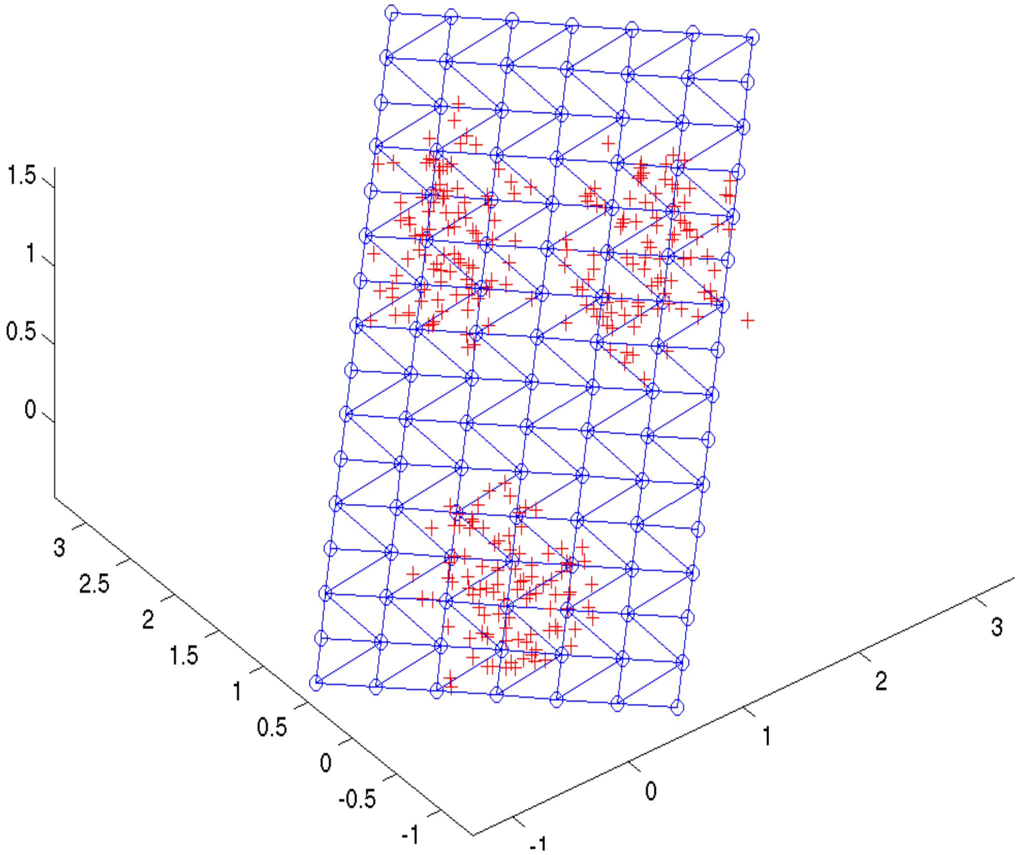


figure 1b

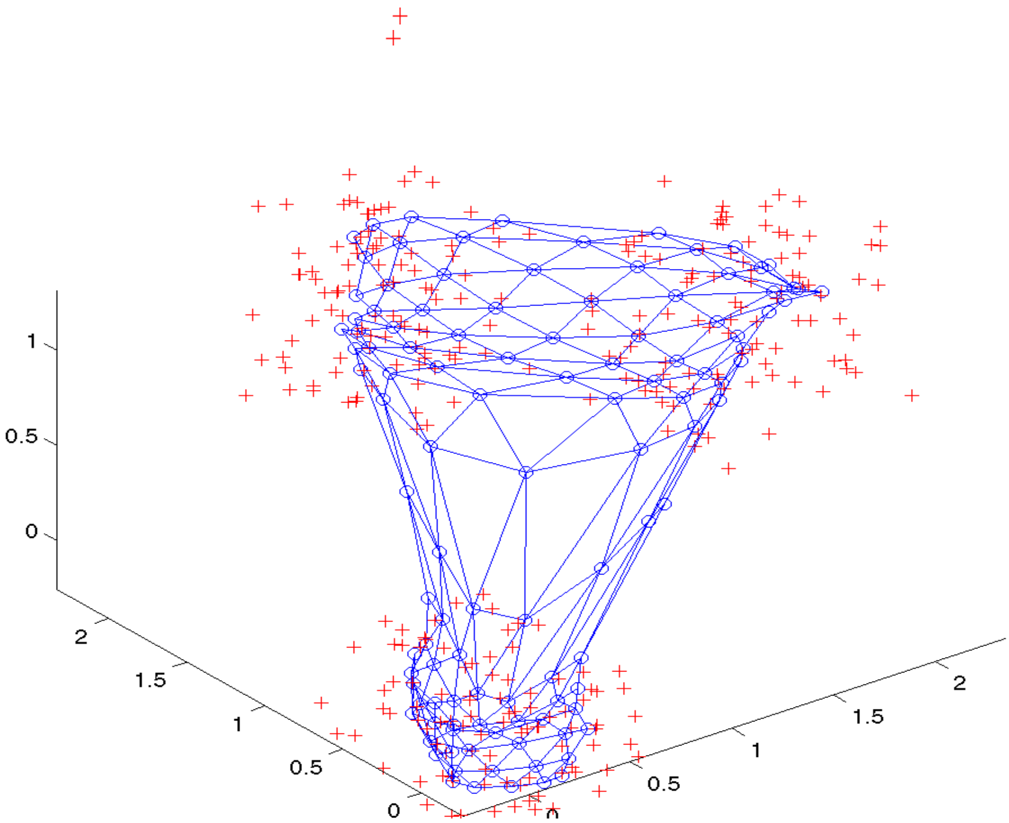


Figure 1c

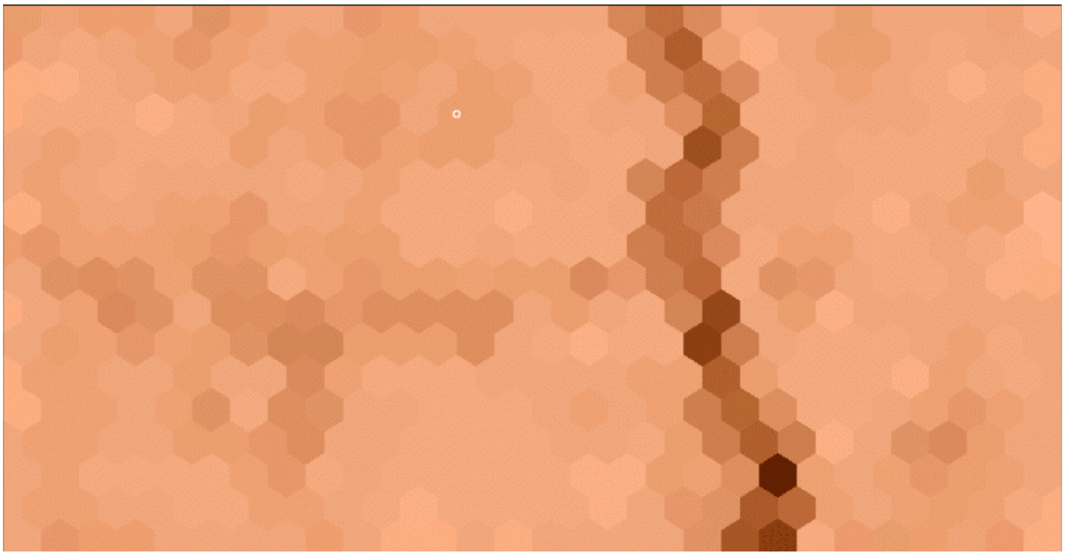


Figure 2

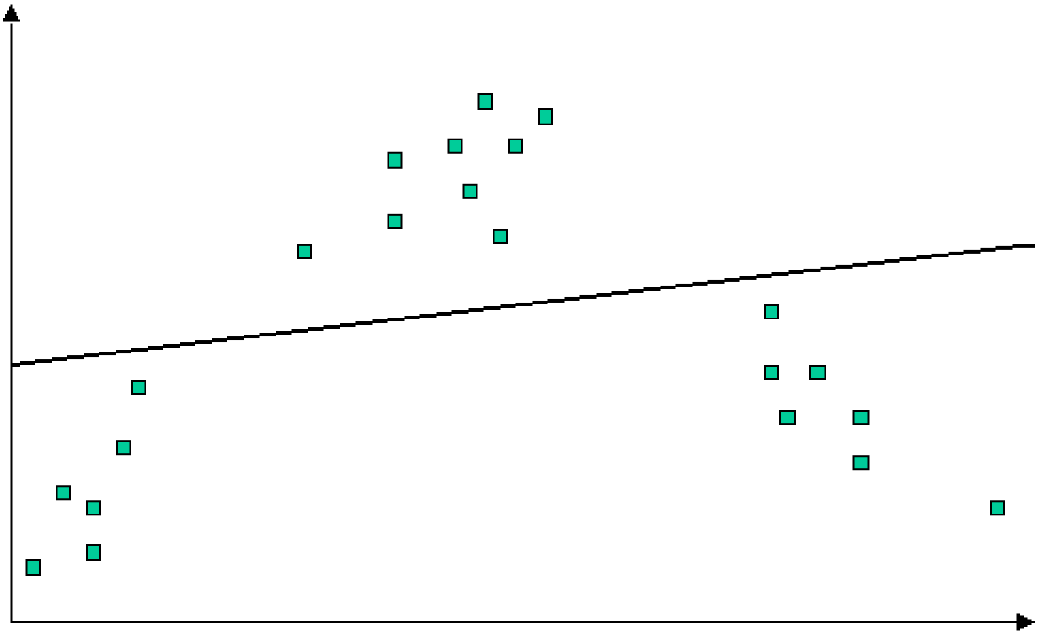


Figure 3a

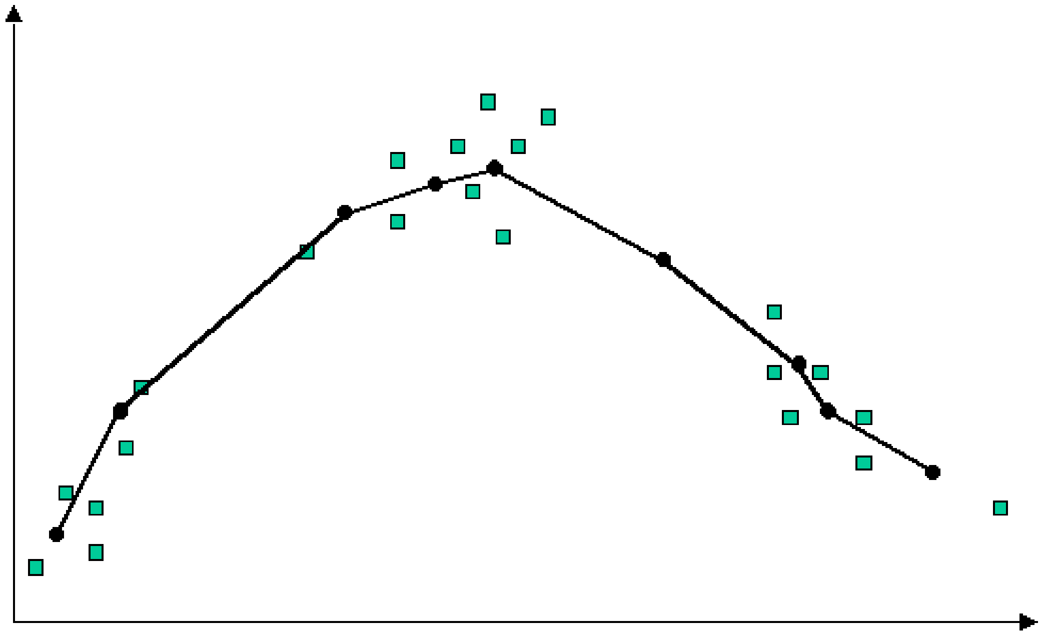
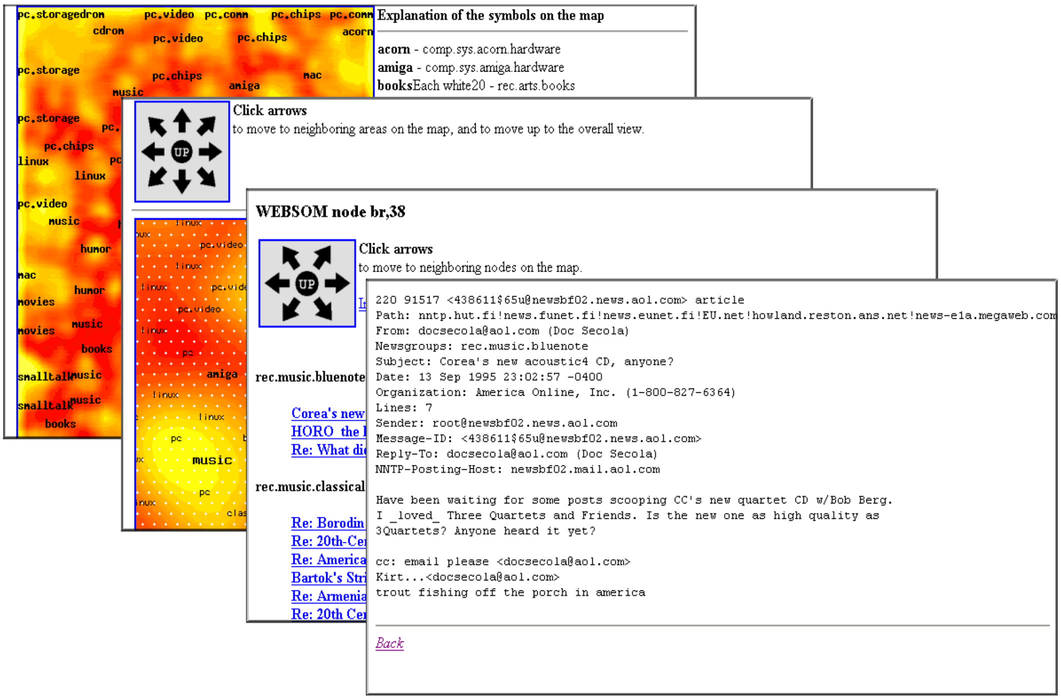


Figure 3b



Explanation of the symbols on the map

acorn - comp.sys.acorn.hardware
amiga - comp.sys.amiga.hardware
booksEach white20 - rec.arts.books

Click arrows
to move to neighboring areas on the map, and to move up to the overall view.

WEBSOM node br,38

Click arrows
to move to neighboring nodes on the map.

rec.music.bluenote

Corea's new
HORO the I
Re: What di

rec.music.classical

Re: Borodin
Re: 20th-Cer
Re: America
Bartok's Stri
Re: Armenia
Re: 20th Cer

```
220 91517 <438611$65u@newsbf02.news.aol.com> article
Path: nntp.hut.fi!news.funet.fi!news.eunet.fi!EU.net!howland.reston.ans.net!news-e1a.megaweb.com
From: docsecola@aol.com (Doc Secola)
Newsgroups: rec.music.bluenote
Subject: Corea's new acoustic4 CD, anyone?
Date: 13 Sep 1995 23:02:57 -0400
Organization: America Online, Inc. (1-800-827-6364)
Lines: 7
Sender: root@newsbf02.news.aol.com
Message-ID: <438611$65u@newsbf02.news.aol.com>
Reply-To: docsecola@aol.com (Doc Secola)
NNTP-Posting-Host: newsbf02.mail.aol.com

Have been waiting for some posts scooping CC's new quartet CD w/Bob Berg.
I _loved_ Three Quartets and Friends. Is the new one as high quality as
3Quartets? Anyone heard it yet?

cc: email please <docsecola@aol.com>
Kirt...<docsecola@aol.com>
trout fishing off the porch in america
```

Back

Figure 4