Chapter 8

Self-organizing map

Teuvo Kohonen, Samuel Kaski, Panu Somervuo, Krista Lagus, Merja Oja, Vesa Paatero

8.1 Self-organizing maps: introduction

Teuvo Kohonen

The name Self-Organizing Map (SOM) signifies a class of neural-network algorithms in the unsupervised-learning category. In its original form the SOM was invented by the founder of the Neural Networks Research Centre, Professor Teuvo Kohonen in 1981-82, and numerous versions, generalizations, accelerated learning schemes, and applications of the SOM have been developed since then.

The central property of the SOM is that it forms a nonlinear projection of a high-dimensional data manifold on a regular, low-dimensional (usually 2D) grid. In the display, the clustering of the data space as well as the metric-topological relations of the data items are clearly visible. If the data items are vectors, the components of which are variables with a definite meaning such as the descriptors of statistical data, or measurements that describe a process, the SOM grid can be used as a groundwork on which each of the variables can be displayed separately using grey-level or pseudocolor coding. This kind of combined display has been found very useful for the understanding of the mutual dependencies between the variables, as well as of the structures of the data set.

The SOM has spread into numerous fields of science and technology as an analysis method. We have compiled a list of over 5000 scientific articles that apply the SOM or otherwise benefit from it.

The most promising fields of application of the SOM seem to be

- data mining at large, in particular visualization of statistical data and document collections,
- process analysis, diagnostics, monitoring, and control,
- biomedical applications, including diagnostic methods and data analysis in bioinformatics, and
- data analysis in commerce, industry, macroeconomics, and finance.

References

[1] Teuvo Kohonen, *Self-Organizing Maps*, Springer Series in Information Sciences, Vol. 30, Springer, Berlin, Heidelberg, New York, 1995, 1997, 2001, 3rd edition.

8.2 5384 works on SOM

Merja Oja, Samuel Kaski, Teuvo Kohonen

The Self-Organizing Map (SOM) algorithm has attracted a great deal of interest among researches and practitioners in a wide variety of fields. The SOM has been analyzed extensively, a number of variants have been developed and, perhaps most notably, it has been applied extensively within fields ranging from engineering sciences to medicine, biology, and economics. We have collected a comprehensive list of 5384 scientific papers that use the algorithms, have benefited from them, or contain analyses of them. The list is intended to serve as a source for literature surveys.

The collection is available at the WWW address `http://www.cis.hut.fi/nnrc/refs/` (cf. [1, 2]).

A SOM of SOM references. The SOM references were organized onto a document map to study the relationships between the topic categories, and to provide an interface for browsing and searching the collection. A WEBSOM [3] was computed using the titles of the documents. For some documents also an abstract was available and it was used in the computation.

The map is available for browsing and search in the address `http://websom.hut.fi/websom/somref/search.cgi`.

References

[1] Merja Oja, Samuel Kaski, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1998–2001 Addendum *Neural Computing Surveys*, Volume 3, pages 1–156, 2003. Available in electronic form at http://www.cse.ucsc.edu/NCS/: Vol 3, pp. 1–156.

[2] Samuel Kaski, Jari Kangas, and Teuvo Kohonen. Bibliography of self-organizing map (SOM) papers: 1981–1997. *Neural Computing Surveys*, 1(3&4):1–176, 1998. Available in electronic form at http://www.cse.ucsc.edu/NCS/: Vol 1, pp. 102–350.

[3] Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Jukka Honkela, Vesa Paatero, and Antti Saarela. Self organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11:574–585, 2000.

8.3 Median self-organizing map of human endogenous retroviruses

Merja Oja, Panu Somervuo, Samuel Kaski, Teuvo Kohonen

Only about two percent of human DNA codes for proteins. The function of the rest is unknown, and it has been called “junk DNA.” It is, however, far from random, and numerous studies (for a review see [1]) have already shown that it may serve for meaningful functions.

About 45 per cent of the DNA [2] is derived from *transposons*, parts of genome capable of moving or copying themselves in the genome. About eight per cent consists of specific kinds of transposons, called *human endogenous retroviruses (HERV)*. Human retroviruses such as HIV in general are viruses capable of copying their genetic code to the DNA of humans, and they become endogenous once they have been copied to the germ-line. Human endogenous retroviruses, in contrast to some other human transposons, are not capable of moving any longer but it has been suggested that they may have functions in regulating the activity of human genes, and may produce proteins under some conditions [3]. It is important to learn more about the HERVs and their effect on our genome.

We have started studies on human endogenous retroviruses (HERVs) by exploring their mutual relationships and their similarities to other DNA elements [4]. We demonstrated that a completely data-driven grouping is able to reflect same kinds of relationships as more traditional biological classifications and phylogenetic taxonomies. The clusters and their visualization were computed with the Median Self-Organizing Map algorithm [5] of pairwise FASTA-based distances [6]. The whole-sequence distances were able to distinguish between the different known types of endogenous elements, and exogenous retroviruses. The HERVs became grouped meaningfully (see Figure 8.1).

References

[1] Roswitha Löwer. The pathogenic potential of endogenous retroviruses: facts and fantasies. *Trends in Microbiology*, 7(9):350–56, September 1999.

[2] E.S. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[3] David J. Griffiths. Endogenous retroviruses in the human genome sequence. *Genome Biology*, 2:1017.1–1017.5, 2001.

[4] Merja Oja, Panu Somervuo, Samuel Kaski, and Teuvo Kohonen. Clustering of human endogenous retrovirus sequences with median self-organizing map. In *WSOM’03 Workshop on Self-Organizing Maps, 9-14 Sep 2003, Hibikino, Japan*, 2003.

[5] Teuvo Kohonen and Panu Somervuo. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–52, 2002.

[6] W. Pearson and D. Lipman. Improved tools for biological sequence comparision. *Proc. Natl. Acad. Sci. USA*, 85:2444–8, 1988.

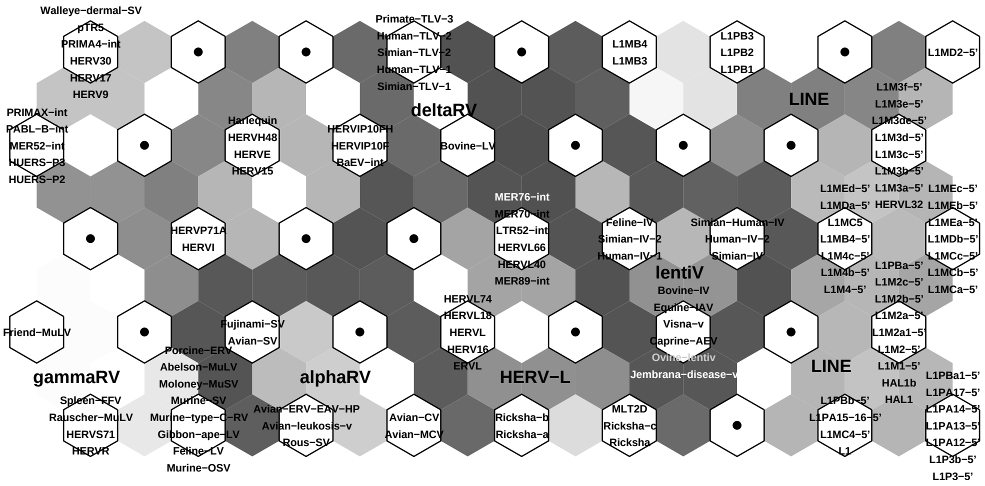


Figure 8.1: Part of the Median SOM of HERV, LINE, and exogenous retrovirus sequences. Every second (bordered, and dotted if not being a best match for any sequence) hexagon denotes a SOM unit, and the rest are U-matrix entries indicating distance between the units. The resulting light areas are clusters and black stripes borders between them. Symbols of the sequences have been inserted to the locations where the sequences have been mapped. Manually assigned names for the clusters are presented on the map. (V=virus, RV=retroV, SV=sarcomaV, OSV=osteoSV, LV=leukemiaV, Mu=murine, TLV=T-lymhocytic V, CV=carcinomaV, AEV=arthitis-encephalitis V, IAV=infectious anemia V, MCV=myelocytomatosis V, FFV=focus forming V.)

8.4 Self-organization of very large document collections

Teuvo Kohonen, Samuel Kaski, Krista Lagus, Vesa Paatero

Text mining systems are developed to aid the users in satisfying their information needs, which may vary from searching answers to well-specified questions to learning more of a scientific discipline. The major tasks of web mining are *searching*, *browsing*, and *visualization*. Searching is best suited for answering specific questions of a well-informed user. Browsing and visualization, on the other hand, are beneficial especially when the information need is more general, or the topic area is new to the user. The SOM, applied to organizing very large document collections, can aid in all the three tasks.

The WEBSOM method

In a method that we have called the WEBSOM [1], a massive collection of documents can be organized efficiently on a large self-organized map.

The computation of document maps. In short, the method is as follows: Encode each document using the vector space model [2] with word weighting. Rare words and a stoplist of common words are excluded. The document vectors are condensed for computational reasons by applying the random projection [3] method. Finally, the document vectors are automatically ordered on a self-organizing map. Various shortcut methods are applied in the construction of large SOMs, including application of a pointer representation in the random projection for fast generation of the document vectors, utilization of the Batch Map algorithm for SOM learning, accelerated winner search, speeded distance computation by neglecting zero-valued elements in the vectors, rapid estimation of a larger map based on a smaller one, and saving memory by using reduced accuracy in storing the maps. It has been shown [1] that while these shortcut methods reduce computation time by an order of $O(d)$ where d is the vocabulary size (nearly 50,000 in our largest experiment), the quality of the maps is practically the same as with a computation where no shortcut methods have been applied.

User interface. The final document map is presented as a series of HTML pages and clickable images that enable exploration of the grid points: a mouse click on a grid point brings to view the links to documents residing in that grid point. The documents, stored in a database, can then be read by following the links. A large map can be first zoomed to view subsets of it more closely. For the largest maps we have used several zooming levels. To provide guidance in the exploration, an automatic method, described in [4], has been utilized for selecting keywords to characterize map regions. The selected words have been marked on the map display.

Content-addressable search. The interface to the map has been provided with a form field into which the user can type a query in the form of a short “document.” This query is preprocessed and a document vector is formed in the exactly same manner as for the stored documents. The resulting vector is then compared with the “models” of all SOM grid points, and the best-matching points are marked with circles on the map display: the better the match, the larger the circle. These locations provide good starting points for browsing.

Keyword search. If the user wants to find documents containing a single keyword or very few keywords, one can search the map using a more conventional *keyword search mode* which is provided as an alternative to the content addressable search. The keyword search is performed by accessing an index from each word to the map units where that word occurs.

Experiments

The largest published map

The largest WEBSOM map made so far is a map of 6,840,568 patent abstracts that were available in electronic form and written in English. The size of the SOM was 1,002,240 models (neurons), and the dimensionality of each model was 500. The representation for each document has been made by projecting the 43,222-dimensional word histogram randomly onto the final 500-dimensional space. Formation of the document map and the interface took altogether about 6 weeks with the newest speedup methods; searching occurs in a few seconds.

The Britannica map

For this map, published in [5], the collection consisted of about 68,000 articles from the Encyclopaedia Britannica, and additionally summaries, updates, and other miscellaneous material of about 43,000 items. Very long articles were split into several sections, resulting in a total of about 115,000 documents.

The documents were preprocessed to remove HTML markup, links and images. Inflected word forms were converted to their base forms using a morphological analyzer. The average length of the documents was 490 words. The size of the finally accepted vocabulary was 39,058 words. The words were weighted by the inverse document frequency (IDF). The representation for each document was made by projecting the 39,058-dimensional weighted word histogram randomly onto the final 1000-dimensional space.

The size of the SOM was 12,096 units. Speedups were employed in the creation of the map, namely SOM magnification, and the batch map algorithm where the speeded winner search was employed for fast convergence. The model vectors were represented using reduced accuracy to decrease the memory requirements.

Figure 8.2 exemplifies a case of keyword search. The map and the collection can be explored using a WWW-browser. Further examples of document maps can be found at `http://websom.hut.fi/websom/`.

Once an interesting region has been located e.g. by the search facility, it can be explored by zooming on the map. Figure 8.3 shows an example of how the local ordering of the map may be useful for examining a topic.

Using the document maps for improving search results

Previously, it has been shown how the maps can be utilized for exploration of a large document collection with the help of a browsing interface and the visualized map display. However, as described in [6], the document maps can also be applied for searching without the benefit of the visual interface.

When using the small CISI test collection intended for information retrieval tests, a statistically significant improvement was found when comparing to the standard vector space model. The favourable effect is considered to be due to the fact that the document map brings into the result set similar, relevant documents that do not contain otherwise sufficient amount of the particular words utilized in the search expression.

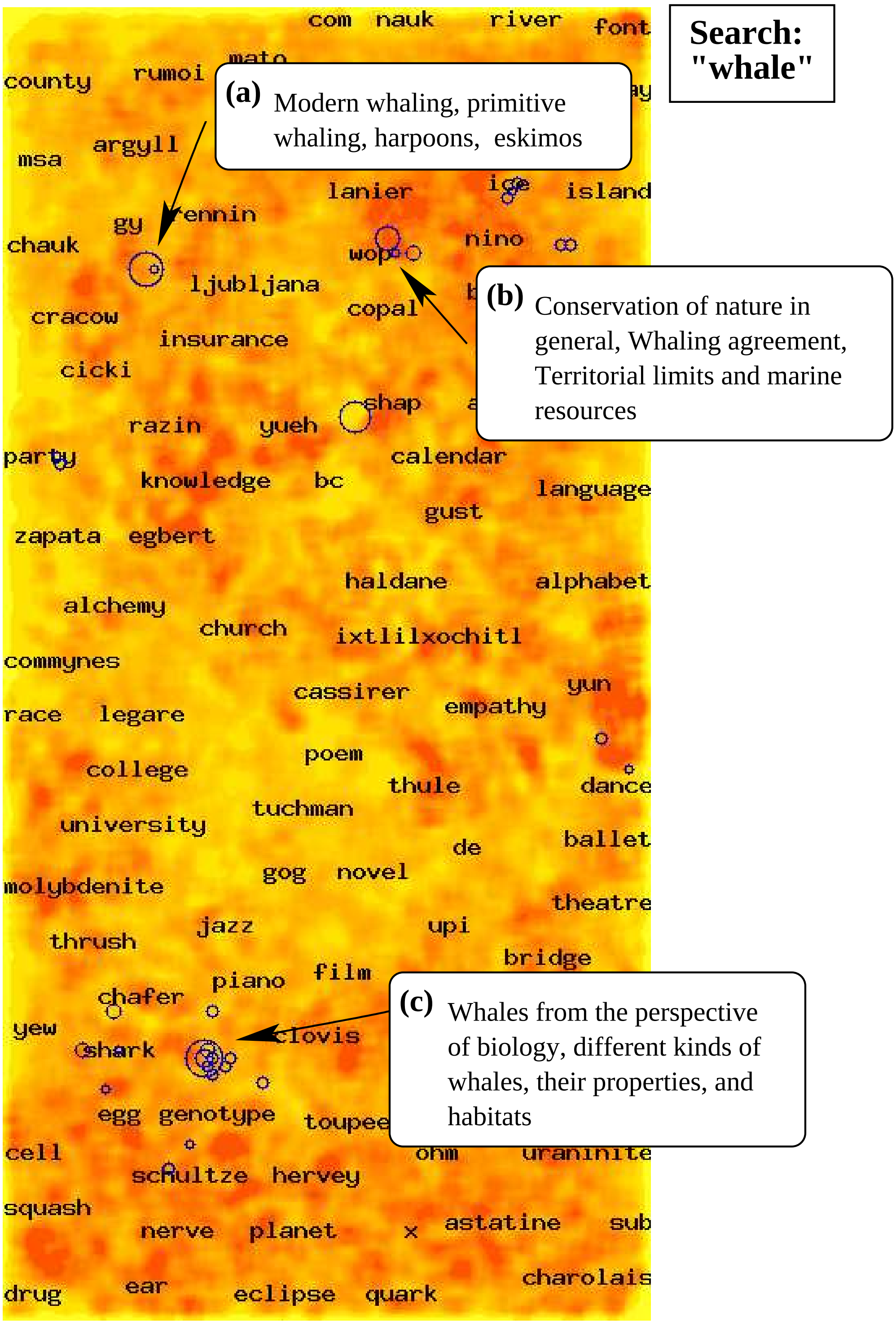
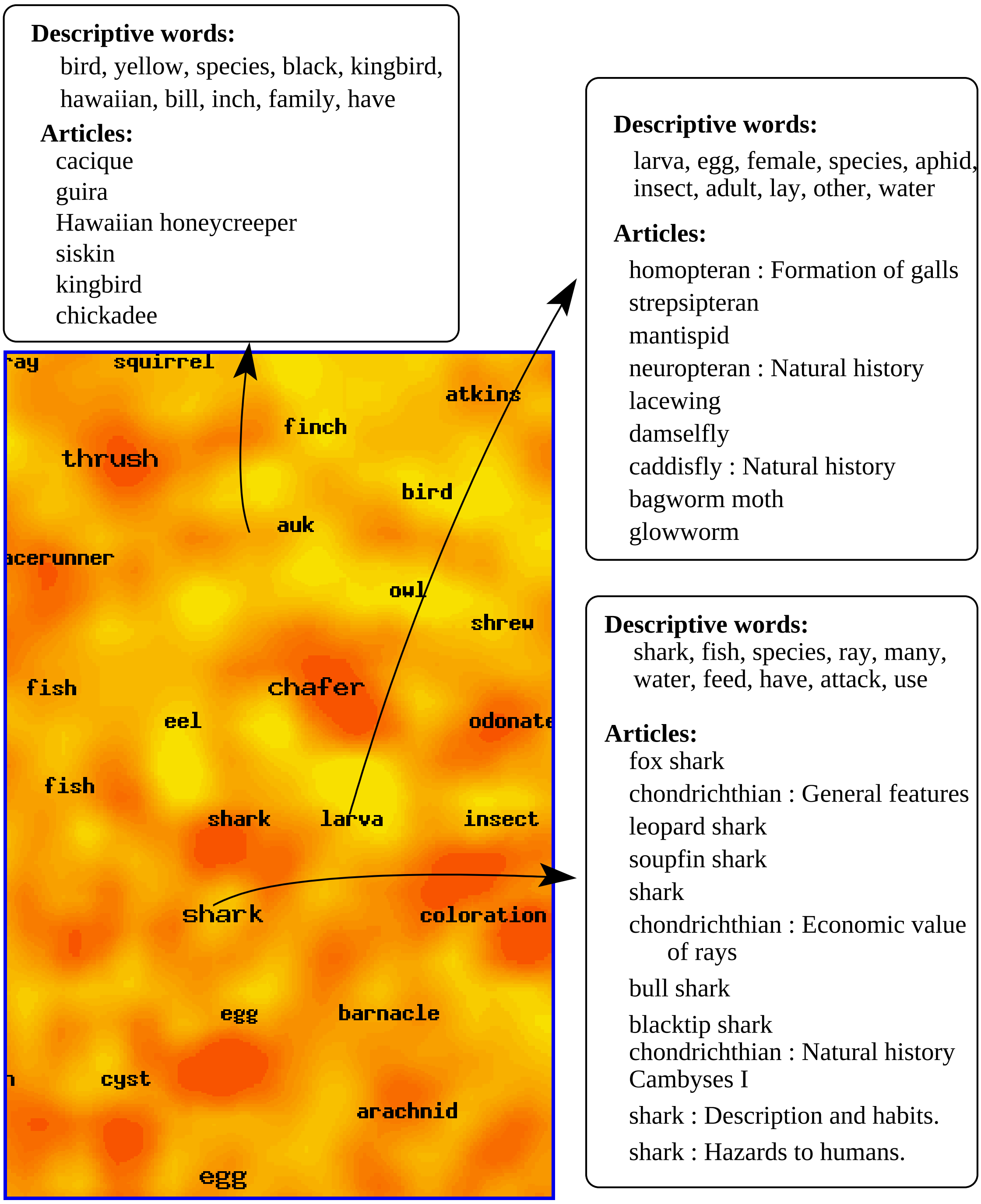


Figure 8.2: The map of Encyclopaedia Britannica articles where the results of a search for 'whale' are depicted. The document map is visualized in the background, and the lighter shades of colour correspond to document clusters. The words written on the document map have been selected automatically using the method described in [4]. The search hits are indicated with blue circles, the size of which describes the goodness of the hit. Three different aspects regarding whales are described in the insets.



Descriptive words:
bird, yellow, species, black, kingbird, hawaiian, bill, inch, family, have

Articles:
cacique
guira
Hawaiian honeycreeper
siskin
kingbird
chickadee

Descriptive words:
larva, egg, female, species, aphid, insect, adult, lay, other, water

Articles:
homopteran : Formation of galls
strepsipteran
mantispid
neuropteran : Natural history
lacewing
damselfly
caddisfly : Natural history
bagworm moth
glowworm

Descriptive words:
shark, fish, species, ray, many, water, feed, have, attack, use

Articles:
fox shark
chondrichthian : General features
leopard shark
soupfin shark
shark
chondrichthian : Economic value of rays
bull shark
blacktip shark
chondrichthian : Natural history
Cambyses I
shark : Description and habits.
shark : Hazards to humans.

Figure 8.3: A close-up of the map of Encyclopaedia Britannica articles. The user has clicked a map region with the label 'shark', obtaining a view of a section of the map with articles on sharks, various species of fish and eel (in the middle and left); insects and larvae (lower right corner); various species of birds (upper right corner); etc. Searches performed on the map confirm that also whales and dolphins can be found nearby (not shown). A topic of interest is thus displayed in a context of related topics. The three insets depict the contents of three map units, i.e., titles of articles found in the unit. By clicking the title, one may read the article. The 'descriptive words' list was obtained with the labeling method [4] and contains a concise description of the contents of the map unit.

Conclusions

We have demonstrated that it is possible to scale up the SOMs in order to tackle very large-scale problems. The strength of the large map displays is in "finding" rather than "searching for" relevant information. Nevertheless, experiments on a small reference collection indicate that the obtained clusters may serve as meaningful sub-topics that can be used to improve accuracy also in a more focused search task.

Although initially designed for text mining, WEBSOM document maps have additional applications in other fields of natural language processing. Examples of such applications are described in Section 12 where the document maps have been aplied to improving speech recognition and for word sense disambiguation.

References

[1] Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., and Saarela, A. Self Organization of a Massive Document Collection. *IEEE Transactions on Neural Networks*, vol. 11, number 3, pp. 574–585. May 2000.

[2] Salton, G., and McGill, MJ. *Introduction to modern information retrieval.* McGraw-Hill, New York, 1983

[3] Kaski, S. Dimensionality reduction by random mapping. *In Proc of IJCNN'98, Int Joint Conf on Neural Networks.* IEEE Press, Piscataway, NJ, 1998, pp. 413–418

[4] Lagus, K., and Kaski, S. Keyword selection method for characterizing text document maps. In *Proc. of ICANN99, Ninth Int. Conf. on Artificial Neural Networks* vol. 1, pp. 371–376. IEE, London, 1999.

[5] Lagus, K., Kaski, S., and Kohonen, T. Mining massive document collections by the WEBSOM method *Information Sciences.* In press.

[6] Lagus, K. Text retrieval using self-organized document maps. *Neural Processing Letters*, vol. 15, no. 1, pp. 21-29. February 2002.