In Simoudis, E., Han, J., and Fayyad, U., editors, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 238-243. AAAI Press, Menlo Park, California.

Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration

Krista Lagus, Timo Honkela, Samuel Kaski, and Teuvo Kohonen
Helsinki University of Technology
Neural Networks Research Centre
Rakentajanaukio 2 C, FIN-02150 Espoo, Finland
e-mail: Krista.Lagus@hut.fi

Abstract

Powerful methods for interactive exploration and search from collections of free-form textual documents are needed to manage the ever-increasing flood of digital information. In this article we present a method, WEBSOM, for automatic organization of full-text document collections using the self-organizing map (SOM) algorithm. The document collection is ordered onto a map in an unsupervised manner utilizing statistical information of short word contexts. The resulting ordered map where similar documents lie near each other thus presents a general view of the document space. With the aid of a suitable (WWW-based) interface, documents in interesting areas of the map can be browsed. The browsing can also be interactively extended to related topics, which appear in nearby areas on the map. Along with the method we present a case study of its use.
Keywords: data visualization, document organization, full-text analysis, interactive exploration, self-organizing map.

Introduction

Finding relevant information from the vast material available, e.g., in the Internet is a difficult and time-consuming task. Efficient search tools such as search engines have quickly emerged to aid in this endeavor. However, the basic problem with traditional search methods such as searching by keywords or by indexed contents is the difficulty to devise suitable search expressions, which would neither leave out relevant documents, nor produce long listings of irrelevant hits. Even with a rather clear idea of the desired information it may be difficult to come up with all the suitable key terms and search expressions. Thus, a method of encoding the information based, e.g., on semantically homogeneous *word categories* rather than individual words would be helpful.

An even harder problem, for which search methods are usually not even expected to offer much support, is encountered when interests are vague and hard to describe verbally. Likewise, if the area of interest resides at the outer edges of one's current knowledge, devising search expressions is like finding a path in total darkness, and the results are poor. If there were something like a map of the document collection at hand, a map where documents were ordered meaningfully according to their content, then even partial knowledge of the connections of the desired information to something already familiar would be useful. Maps might help the exploration first by visualizing the information space, and then by guiding one to the desired information as well as to related subjects. A visualized map of the information landscape might even reveal surprising connections between different areas of knowledge.

The self-organizing map (SOM) (Kohonen 1982; 1995; Kohonen *et al.* 1996) is a means for automatically arranging high-dimensional statistical data so that alike inputs are in general mapped close to each other. The resulting map avails itself readily to visualization, and thus the distance relations between different data items can be illustrated in a familiar and intuitive manner.

When suitably encoded textual documents are organized with the SOM algorithm, the map of the document collection provides a general view to the information contained in the document landscape, a view where changes between topics are generally smooth and no strict borders exist. Easy exploration of the document landscape may then be provided, e.g., via the World Wide Web (WWW).

The WEBSOM Method

Before ordering the documents they must be *encoded*; this is a crucial step since the final ordering depends on the chosen encoding scheme. In principle, a document might be encoded as a histogram of its words. Often the computational burden required for treating the histograms would, however, be orders of magnitude too large with the vast vocabularies used for automatic full-text analysis. An additional problem with the word histograms is that the discrete, symbolic words as such retain no information of their relatedness. Synonyms appear to be as distant as any unrelated words, although in a useful full-text analysis method similar words should obviously be encoded similarly.

Since it is not currently feasible to incorporate references to real-life experience of word meanings to a text analysis method, the remaining alternative is to use the statistics of the *contexts* of words to provide information on their relatedness. It has turned out that the size of the word histograms can be reduced to a fraction with the so-called self-organizing semantic maps (Ritter & Kohonen 1989). At the same time, by virtue of the organizing power of the SOM, the semantic similarity of the words can be taken into account in encoding the documents.

Preprocessing Text

Before applying the SOM to a document collection (in this case 8800 articles from a Usenet newsgroup, with a total of 2 000 000 words) we automatically removed some non-textual information from the documents. In the remaining text, numerical expressions and several kinds of common code words were categorized with heuristic rules into a few classes of special symbols. To reduce the computational load the words that occurred only a few times (in this experiment less than 50 times) in the whole data base were neglected and treated as empty slots.

In order to emphasize the subject matters of the articles and to reduce variations caused by the different discussion styles, which were not of interest in this experiment, a group of common words that were not supposed to discriminate any discussion topics were discarded from the vocabulary. In the actual experiment we removed 1000 common words from the vocabulary of 3300 words that remained after discarding the rarest words.

The SOM Algorithm

To provide a general understanding of what the self-organizing map is and why it is suitable for ordering large collections of text documents, the following expedition of thought may be helpful.

Consider an information processing system, such as a brain area, which must learn to carry out different tasks, each of them well. Let us assume that the system may assign different tasks to different sub-units that are able to learn from what they do. Each new task is given to the unit that can best complete the task. Since the units learn, and since they receive tasks that they can do well, they become even more competent in those tasks. This is a model of specialization by competitive learning. Furthermore, if the units are interconnected in such a way that also the (predefined) *neighbors* of the unit carrying out a task are allowed to learn some of the task, the system also slowly becomes ordered: nearby units have similar abilities, and the abilities change slowly and smoothly over the whole system. This is the general principle of the self-organizing map (SOM). The system is called a *map* and the task is to imitate, i.e., *represent* the input. The representations become ordered according to their similarity relations in an unsupervised learning process. This property makes the SOM useful for organizing large collections of data in general, including document collections.

In the simplest kinds of SOMs the map consists of a regular grid of units, and the task is to represent statistical data, described by vectors $x \in \Re^n$. Each map unit i contains a model vector $m_i \in \Re^n$, and the model vectors are used for representing the data. In the learning process the model vectors change gradually so that finally the map forms an ordered non-linear regression of the model vectors into the data space.

At each step t of the learning process, a data sample $x(t)$ is presented to the units. The node c that best represents the input is then searched for using, e.g., the Euclidean distance to define the quality of the representation: $||x - m_c|| = min_i\{||x - m_i||\}$. Next, the unit c as well as the neighboring units learn to represent the data sample more accurately. The model vector of unit i is updated according to the following learning rule:

$$m_i(t+1) = m_i(t) + h_{ci}(t)[x(t) - m_i(t)]. \qquad (1)$$

Here h_{ci} is a "smearing" or neighborhood function expressing how much the unit i is updated when unit c is the winner. The neighborhood function typically is a symmetric, monotonically decreasing function of the distance of units i and c *on the map grid*. During repeated application of (1) with different inputs, model vectors of neighboring map units become gradually similar because of the neighborhood function h_{ci}, eventually leading to global ordering of the model vectors.

The Word Category Map

The word category map that is used for document encoding is organized according to word similarities, measured by the similarity of the contexts of the words. Conceptually related words tend to fall into the same or neighboring map nodes. Nodes may thus be viewed as word categories. The ordering is formed by the SOM algorithm based on the average short contexts of the words (Ritter & Kohonen 1989).

In our experiment, the word contexts were of length three, and consisted of one preceding and one following word in addition to the word that was being encoded. The ith word in the sequence of words is represented by an n-dimensional (here 90) real vector x_i with random number components. The averaged context vector of a word, marked by w, may be expressed as follows:

$$X(w) = \begin{bmatrix} \mathrm{E}\{x_{i-1}|x_i = w\} \\ \varepsilon w \\ \mathrm{E}\{x_{i+1}|x_i = w\} \end{bmatrix} , \quad (2)$$

where E denotes the conditional average over the whole text corpus, and ε is a small constant, here 0.2. Now the real-valued vectors $X(w)$, in our experiment of dimension 270, constitute the input vectors given to the word category map. During the training of the map, the averaged context vectors $X(w)$ were used as input.

After the map has self-organized, each map node corresponds to a set of input vectors that are close to each other in the input space. A map node may thus be expected to approximate a set of similar inputs, here averaged word contexts. All the inputs for which a node is the best match are associated with the node. The result is a labeling of the map with words, where each node on the map is associated with all the inputs $X(w)$ for which the node is the closest representative. The word in the middle of the context, that is, the word corresponding to the w part of the context vector, was used as a label of the node. In this method a unit may become labeled by several symbols, often synonymic or describing alternative or opposing positions or characteristics. Examples of map nodes are illustrated in Fig. 1. Interrelated words that have similar contexts tend to appear near each other on the map.

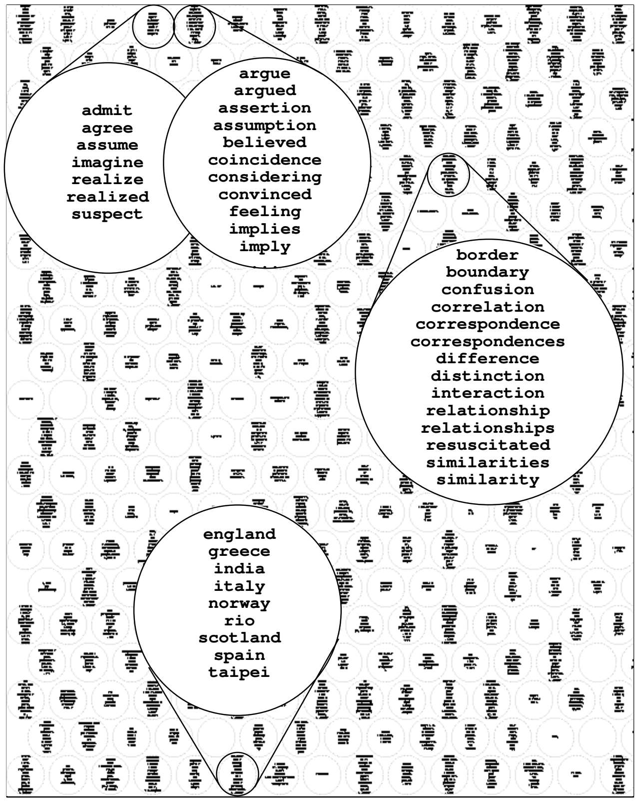

Figure 1: *Sample categories illustrated on a word category map. Similar words tend to occur in the same or nearby map nodes, forming "word categories". The map was computed using a massively parallel neurocomputer CNAPS, and fine-tuned with the SOM_PAK software (Kohonen* et al. *1996).*

The Document Map

With the aid of the word category map the documents can be encoded as *word category histograms.* Very closely related words (that are in the same category on the map) then contribute identically to the code formed for a document.

It would also be advantageous if words in *similar* categories contributed similarly to the codes of the documents. This is indeed possible to achieve since the word category map is ordered. The relations of the categories are reflected in their distances on the map. Therefore, the contributions of nearby categories can be made similar by *smoothing* the histogram on the word category map, whereby the encoding becomes less influenced by the choices of words made by the authors of the documents. A moderately narrow (e.g., diminishing to the half of the maximum value between two neighboring map units) Gaussian function was found to be a suitable smoothing kernel in our recent experiments.

A representative sample of the encoded documents is presented as input to the SOM, which organizes them by unsupervised learning. After the map is formed, also documents that were *not* used in the learning may be automatically positioned onto the map.

When the learning process is complete the density of the documents in the document space can be visualized as shades of gray on the document map, to illustrate the relations of different map locations (see Fig. 3).

Previous Work on Organizing Documents with the SOM

Several studies have been published on SOMs that map words into grammatical and semantic categories (e.g., by Honkela, Pulkki, & Kohonen 1995; Miikkulainen 1993; Ritter & Kohonen 1989; 1990; and Scholtes 1993). The SOM has also been utilized previously to form a small map based on the titles of scientific documents (Lin, Soergel, & Marchionini 1991). Scholtes has applied the SOM extensively in natural language processing, e.g., developed a neural filter along with a neural interest map for information retrieval (1992a; 1992b). Merkl (1993; Merkl, Tjoa, & Kappel 1994) has used the SOM to cluster textual descriptions of software library components. Quite recently we have also been informed of an approach studied in the University of Arizona AI Lab for organizing WWW-pages.

Exploring Document Maps

We have designed a browsing environment which utilizes the order of the document map. The document collection may be explored with any graphical WWW browser.

Example: Documents from the Usenet Newsgroup sci.lang

To ensure that the WEBSOM method works in realistic applications, we performed a case study with material that is difficult enough from the textual analysis point of view. We organized a map with a collection of 8800 full-text articles that appeared in the Usenet newsgroup "sci.lang" during the latter half of 1995, containing approximately a total of 2 000 000 words. The articles are colloquial, mostly rather carelessly written short documents that contain little topical information to organize them properly. Furthermore, spelling errors are not uncommon. Currently (1 May) over 13 000 articles have been organized by the map.

Viewing the Document Collection

The document maps may be explored via a point-and-click interface: the user may zoom in on any map area by clicking the map image to view the underlying document space in more detail. Fig. 2 presents the four different view levels that one encounters when exploring the material. The view of the whole map (part 1 of Fig. 2) offers a general overview on the whole document collection. The display may be focused to a zoomed map view, deeper to a specific node, and finally to a single document.

In a typical session, the user might start from the overall map view, and proceed to examine further a specific area, perhaps later gradually wandering to close-by areas containing related information. Clickable arrow images are provided for moving around on the map. After finding a particularly interesting node, one may use it as a “trap” or “document bin” which can be checked regularly to see if new interesting articles have arrived. An exploration example of the "sci.lang" map is shown in Fig. 3.

Multiple Applications of the WEBSOM Method

The WEBSOM method is readily applicable to any kind of a collection of textual documents. It is especially suitable for exploration tasks in which the users either do not know the domain very well, or they have only a limited or vague idea of the contents of the full-text database being examined. With the WEBSOM, the documents are ordered meaningfully according to their contents, and thus related documents are located near each other. Maps also help the exploration by giving an overall view of what the document space looks like.

In the World Wide Web, one target of application for the WEBSOM is the organization of home pages instead of newsgroup articles. Also electronic mail messages can be automatically positioned on a suitable map organized according to personal interests. Relevant areas and single nodes on the map could be used as "mailboxes" into which specified information is automatically gathered.

The method can also be used to organize official letters, personal files, library collections, and corporate full-text databases. Especially administrative or legal documents may be difficult to locate by traditional information retrieval methods, because of the specialized terminologies used. The category-based and redundantly encoded approach of the WEBSOM is expected to alleviate the terminology problem.

Conclusions

In this work we have presented a novel method for organizing collections of documents into maps, and a browsing interface for exploring the maps. The method, called the WEBSOM, performs a completely automatic and unsupervised full-text analysis of the document set using self-organizing maps. The result of the analysis, an ordered map of the document space, visualizes the similarity relations of the subject matters of the documents; they are reflected as distance relations on the document map.

The present version of the WEBSOM interface has the basic functionality needed for exploring the document collection: Moving on the document map, zooming in on the map and viewing the contents of the nodes. Many different directions for enhancing the interface are possible. For example, ideal starting points for exploration could be provided by finding the position where a user-specified document would fall on the map. The WEBSOM (Honkela *et al.* 1996) tool is already available for exploring collections of Usenet news articles.

References

Honkela, T.; Kaski, S.; Lagus, K.; and Kohonen, T. 1996. Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo. WEBSOM home page (1996) available at http://websom.hut.fi/websom/.

Honkela, T.; Pulkki, V.; and Kohonen, T. 1995. Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulié, F., and Gallinari, P., eds., *Proceedings of the International Conference on Artificial Neural Networks, ICANN-95*, volume 2, 3–7. Paris: EC2 et Cie.

Kohonen, T.; Hynninen, J.; Kangas, J.; and Laaksonen, J. 1996. SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo.

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

Kohonen, T. 1995. *Self-Organizing Maps.* Berlin: Springer.

Lin, X.; Soergel, D.; and Marchionini, G. 1991. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research & Development in Information Retrieval.* 262–269.

Merkl, D.; Tjoa, A. M.; and Kappel, G. 1994. A self-organizing map that learns the semantic similarity of reusable software components. In *Proceedings of the 5th Australian Conference on Neural Networks, ACNN'94.* 13–16.

Merkl, D. 1993. Structuring software for reuse - the case of self-organizing maps. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN-93-Nagoya*, volume III, 2468–2471. Piscataway, NJ: IEEE Service Center.

Miikkulainen, R. 1993. *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory.* Cambridge, MA: MIT Press.

Ritter, H., and Kohonen, T. 1989. Self-organizing semantic maps. *Biological Cybernetics* 61:241–254.

Ritter, H., and Kohonen, T. 1990. Learning 'semantotopic maps' from context. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN-90-Washington-DC*, volume I, 23–26. Hillsdale, NJ: Lawrence Erlbaum.

Scholtes, J. C. 1991a. Kohonen feature maps in full-text data bases: A case study of the 1987 Pravda. In *Proc. Informatiewetenschap 1991, Nijmegen*, 203–220. Nijmegen, Netherlands: STINFON.

Scholtes, J. C. 1991b. Unsupervised learning and the information retrieval problem. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN'91*, 18–21. Piscataway, NJ: IEEE Service Center.

Scholtes, J. C. 1993. *Neural Networks in Natural Language Processing and Information Retrieval.* Ph.D. Dissertation, Universiteit van Amsterdam, Amsterdam, Netherlands.

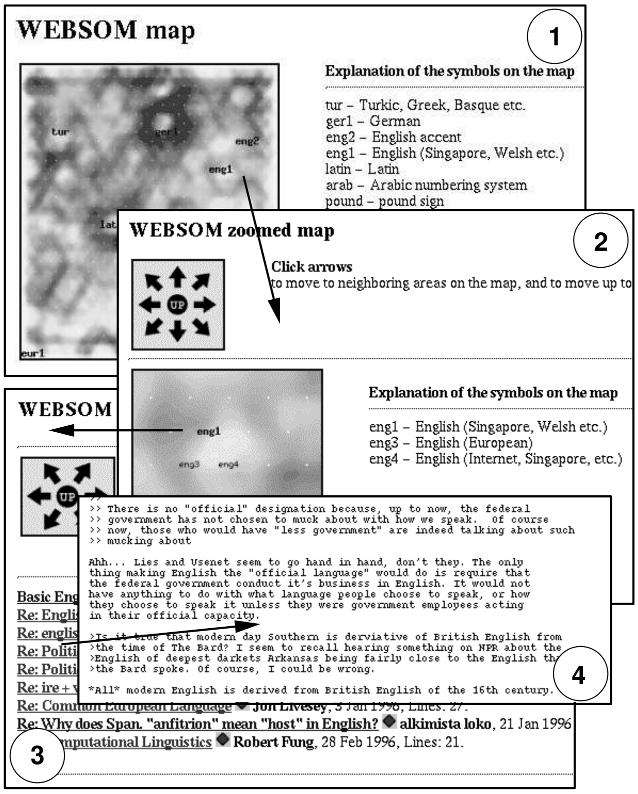

Figure 2: *The different view levels, (1) whole map, (2) zoomed map, (3) map node, and (4) single document, in the order of increasing detail. Moving between the levels or to neighboring areas on the same level is done by mouse clicks on the images or on the document links. Once an interesting area on the map has been found, one may use the arrow images to proceed to explore neighboring areas containing related documents. This can be contrasted with the traditional information retrieval techniques, where the users cannot know whether there are considerable numbers of relevant documents just "outside" the search results.*

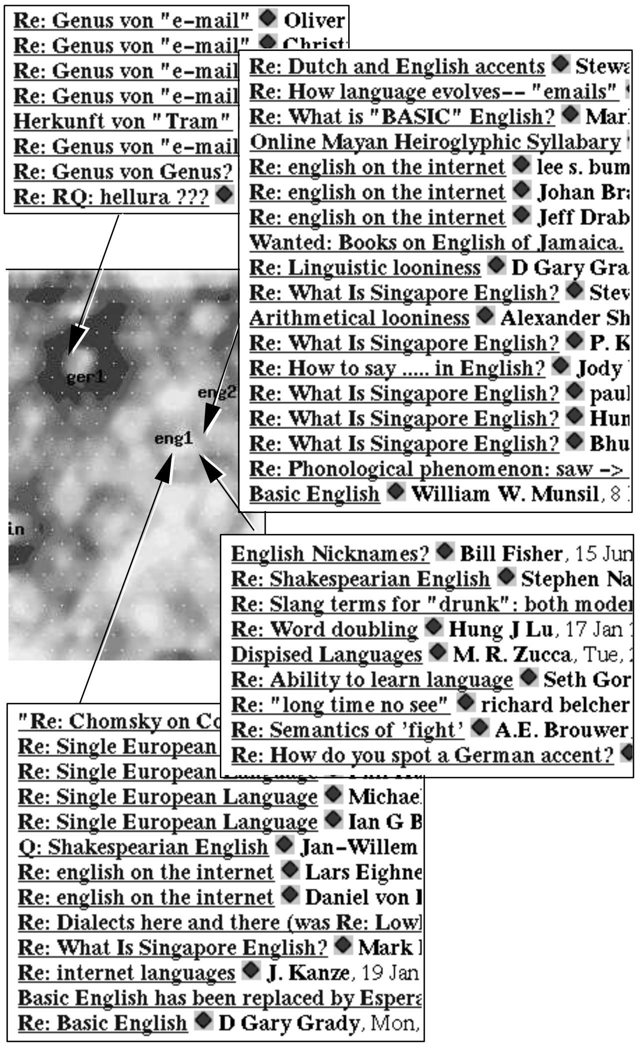

Figure 3: *An example of map exploration. The contents of four map nodes in the upper right corner of the sci.lang map are shown. The shades of gray reflect distances between the map locations in the document space, a dark shade corresponding to a long distance. In a distinct area of its own (the upmost arrow), one can find articles on German or written in German. The remaining three neigboring nodes contain articles related to English, for instance, its usage around the world, historical English, and English as a common language for European Union, or for the Internet community. It is to be noted that some of the titles may be quite disassociated from the contents. In long discussions the original title is often used even after the conversation has diverged into several threads.*