Generalizability of the WEBSOM Method to Document Collections of Various Types

Krista Lagus
Neural Networks Research Centre
Helsinki University of Technology
P.O.Box 2200
FIN-02015 HUT, Finland
E-mail: Krista.Lagus@hut.fi
http://websom.hut.fi/websom/

Abstract: WEBSOM is a method in which the self-organizing map algorithm is used to automatically organize collections of documents on a map to enable easy exploration of the collection. This article illustrates with case studies how collections of various types of text can be successfully organized using the WEBSOM. The emphasis is on describing the particular challenges that each type of material poses, as well as on identifying properties of a text collection that affect the choices made at each progessing stage. Properties such as the size of the document collection, the size of the vocabulary, the domain, the style of writing, and the language are considered.

1 INTRODUCTION

Large collections of text documents in electronic form have made it possible and necessary to develop methods for effective utilization of those collections. The WEBSOM (Honkela et al., 1996; Kaski et al., 1996; Kaski et al., 1998; Kohonen et al., 1996; Lagus et al., 1996; Lagus et al., 1999; Kohonen, 1997) can be used to organize collections of documents onto a two-dimensional display, *a map*, so that map regions that are close to each other contain similar documents. A document is represented statistically as the collection of words that appeared in the document, weighted suitably and in some cases clustered. The self-organizing map (SOM) algorithm (Kohonen, 1982; Kohonen, 1995) is used to organize such representations.

The main objectives in developing the WEBSOM method have been to devise *a method for exploring text collections* that differs from the query-result -approach, to be able to process *large document collections* efficiently and automatically, and not to set limits on the *type of text material* that the method can handle.

2 WEBSOM METHOD IN BRIEF

PREPROCESSING. Before encoding the documents, filters that are specific to the format of the documents are used to remove non-textual and structural information which is not considered relevant for organization of the map (e.g. images, signatures, message headers, numbers, URLs, and email addresses). Words occurring rarely in the material as well as a list of common words that are considered empty in content are also discarded. Language-specific tools might be utilized, e.g. for stemming words. This stage must usually be adjusted a little for each new material type, since texts come in varying formats.

DOCUMENT ENCODING. The most straightforward document encoding used in the WEBSOM follows Salton's vector space model (Salton et al., 1975), where each dimension in the document vector corresponds to a word in the vocabulary. The value of the dimension describes how many times the word occurs in the document, weighted suitably. E.g. inverse document frequency (IDF) can be used as word weight. When the documents fall into separate topic areas or classes, words that separate the topics well can be given a large weight and vice versa (an entropy-based measure, see e.g. Kohonen, 1997).

The dimensionality of the document vectors is of paramount importance, both due to time and space requirements during processing. We have used the following three main methods for dimensionality reduction: excluding rare words from the vocabulary, clustering words based on statistical similarity of their contents (e.g. by word category maps or WCMs, see Honkela, 1997), or performing a so-called *random mapping* to the

	Usenet 1	Usenet 2	WSOM	Patent abstr.	News items
Properties of collection					
type of material	colloquial	colloquial	scientific	scientific	editorial
variation in topics	medium	large	medium	small	large
variation in writing style	large	large	small	small	small
class information avail.	yes	yes	no	yes	no
language	English	English	English	English	Finnish
number of documents	8,800	1,124,134	58	10,074	18,677
words per document	227	218	106	126	64
number of words	1,973,555	245,592,634	6,148	1,266,094	1,198,254
number of unique words	899,358	1,127,184	1,723	17,234	38,276
Choices made					
frequency cutoff	50	50	1	10	10
final vocab. size	2,287	63,773	455	4,660	8,489
dimension reduction	WCM	WCM	none	rand.mapp.	rand.mapp.
document vector dim.	315	315	455	315	315
word weighting	none	class entropy	IDF	class entropy	IDF
document map size	768	104,040	60	1,008	1,620
processing time	1 day	1 month	1 hr	6 hrs	8.5 hrs

Table 1: An overview of various types of text collections organized using WEBSOM. *Frequency cutoff* tells how frequent a word had to be for inclusion in the vocabulary. *Final vocabulary size* is the number of unique words after removing the too general and too rare words. *Dimension reduction* refers to the method used for reducing the dimension of document vectors in the document encoding stage. The listed processing times should be interpreted with caution, as the speeds of the methods are constantly being improved. IDF stands for inverse document frequency.

document vectors (Kaski, 1998; Kaski et al., 1998). Most often a combination of excluding rare words and either WCM or random mapping has been used.

CREATING DOCUMENT MAPS. Document maps are constructed automatically using the SOM algorithm. The appropriate size of the document map is mostly a usability factor: a map unit should contain few enough articles for convenient browsing. An average of 10-15 articles per node appears suitable. The map size affects also processing speed, but speedups exist for teaching large SOM:s (Kohonen et al., 1996).

CREATING THE INTERFACE. An interface consisting of HTML pages enables easy exploration of the map (see Fig. 1). The labels written on the map display are automatically selected words that appear often in the articles in that map region and rarely elsewhere. Decisions on how many zoom levels to use or how densely to write labels on the map can mostly be based on the size of the document map.

EVALUATING DOCUMENT MAPS. The optimal measure for evaluating document maps would be user satisfaction, but often it can be properly measured only after full-fledged product development, not at an early stage in scientific research. Furthermore, with specialized material such as the patent abstracts or scientific documents only a domain expert is able to evaluate the maps. However, if the documents naturally fall into topic areas or classes, the goodness of the document maps may be measured by class separability, i.e. the percentage of articles that fall into a node with their own class as the majority. Since a good categorization into topic areas is often not available, it is important to develop also other evaluation methods.

3 EXPERIMENTS ON DIFFERENT TYPES OF TEXT COLLECTIONS

In the following several case studies of applying WEBSOM to various kinds of document collections will be described. An overview of the properties of the experiments discussed here is presented in Table 1.

3.1 USENET DISCUSSION GROUP ARTICLES

Usenet discussion group articles were the first document collections organized by WEBSOM, and the collections outlined in table 1 have been documented in detail elsewhere: Usenet 1 in (Lagus et al., 1996) and Usenet 2 in (Kohonen, 1997). The material that appears in discussion groups can be literally anything, from short remarks, jokes, or questions to elaborate discussions or long-winded "wars" between individuals. In addition, the material may contain FAQ's (frequently asked questions and their answers), digests, mathematical formulas, program code, ASCII images, signatures of authors etc. The actual text is often carelessly written, full of spelling errors and of poor style. For these reasons the vocabularies are generally huge. The variation and poor quality of this type of material pose a challenge to an automatic text exploration system, especially the preprocessing: the system cannot assume practically anything about the structure, the quality, or the vocabulary. Furthermore, the preprocessing stage, however well designed, cannot be expected to weed out all the "noise", and thus the system itself must tolerate quite noisy data.

In the larger experiment with over million articles it became necessary to develop and utilize computational speedups in teaching the maps, and emphasized the need to keep the method very streamlined. For example, application of language-specific technologies was not deemed worth it—in addition to the time expenditure, the poor quality of the texts would have made them difficult to analyze for a system that expects correct language.

3.2 ABSTRACTS OF SCIENTIFIC ARTICLES

An experiment on organizing abstracts of scientific articles was reported in (Lagus, 1997). The collection was very small, only 58 abstracts, and on a narrowly defined topic, i.e. the SOM algorithm. Scientific abstracts differ from Usenet news articles in that a single abstract may have connections to many other abstracts, based on both the methodology as well as the application area. Thus it may not be possible to display all the connections at the same time. Furthermore, in some map region the articles may appear together because of a common application area, whereas in another region the commonality may be methodological.

With a large text collection two documents that are related but do not utilize similar language may nevertheless appear close to each other on the map because of another "bridging" article that has similarities with both documents. However, with a small collection these connections should be obtained explicitly, e.g. by using a thesaurus to identify synonyms (with a large collection the word category map, WCM, may be used since there is enough information on word statistics for clustering the words). In organizing a small collection it would also be possible to process each article more thoroughly in order to identify the topics of an article, possibly by syntactic analysis or some other advanced but more time-consuming method.

3.3 PATENT ABSTRACTS

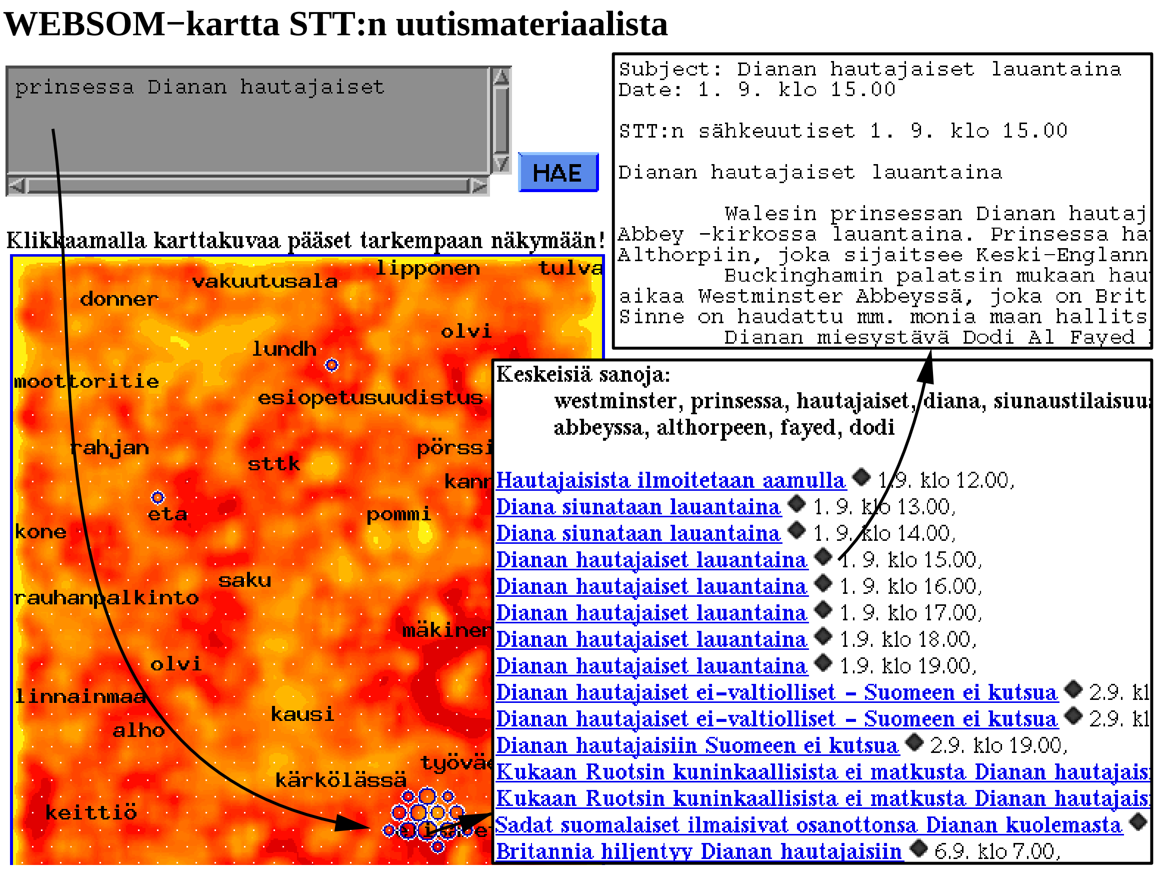
Managing patents costs large amounts of time and money for those applying for a patent because of the great number of existing patents that must in principle be examined to identify and avoid collisions. As part of the process, each incoming patent application is assigned by a patent engineer (a domain expert) to a single primary category and sometimes to secondary categories according to different aspects of the invention. For preliminary experiments ten thousand patent abstracts were drawn from four such related but non-overlapping categories.

Patent abstracts are typically very technically written. Many abstracts seem to have commonalities with many other abstracts, sometimes also accross category borders. Knowing which of these similarities are relevant on a deeper technical level is very difficult for a non-expert. Although there are methods for measuring the goodnes of SOMs, here the results of the entire process need to be evaluated. So far the separability of classes on the map has been used as the goodness criterium. When relevance information regarding the documents is available, one might also consider utilizing it to measure the effectiveness of searches on the document maps.

3.4 SHORT NEWS ITEMS IN FINNISH

As all the other experiments have been on English material, it was interesting to see whether Finnish texts could be organized with as good results. Finnish poses a challenge due to its high degree of inflections: according to Lindén (1997) "a verb root in Finnish may have 18000 inflected forms, and one noun some 2000 forms".

The material in this experiment consisted of 18,677 short news items written by journalists at a Finnish news agency in the first 9 months of year 1997 (see Fig. 1). The style of writing was clear, to the point, and consistent,


WEBSOM−kartta STT:n uutismateriaalista

prinsessa Dianan hautajaiset

Klikkaamalla karttakuvaa pääset tarkempaan näkymään!

```
Subject: Dianan hautajaiset lauantaina
Date: 1. 9. klo 15.00

STT:n sähkeuutiset 1. 9. klo 15.00

Dianan hautajaiset lauantaina

     Walesin prinsessan Dianan hautaj
Abbey -kirkossa lauantaina. Prinsessa ha
Althorpiin, joka sijaitsee Keski-Englann
     Buckinghamin palatsin mukaan hau
aikaa Westminster Abbeyssä, joka on Brit
Sinne on haudattu mm. monia maan hallits
     Dianan miesystävä Dodi Al Fayed
```

Keskeisiä sanoja:
westminster, prinsessa, hautajaiset, diana, siunaustilaisuu
abbeyssa, althorpeen, fayed, dodi

Hautajaisista ilmoitetaan aamulla ◆ 1.9. klo 12.00,
Diana siunataan lauantaina ◆ 1. 9. klo 13.00,
Diana siunataan lauantaina ◆ 1. 9. klo 14.00,
Dianan hautajaiset lauantaina ◆ 1. 9. klo 15.00,
Dianan hautajaiset lauantaina ◆ 1. 9. klo 16.00,
Dianan hautajaiset lauantaina ◆ 1. 9. klo 17.00,
Dianan hautajaiset lauantaina ◆ 1.9. klo 18.00,
Dianan hautajaiset lauantaina ◆ 1.9. klo 19.00,
Dianan hautajaiset ei-valtiolliset – Suomeen ei kutsua ◆ 2.9. kl
Dianan hautajaiset ei-valtiolliset – Suomeen ei kutsua ◆ 2.9. kl
Dianan hautajaisiin Suomeen ei kutsua ◆ 2.9. klo 19.00,
Kukaan Ruotsin kuninkaallisista ei matkusta Dianan hautajais
Kukaan Ruotsin kuninkaallisista ei matkusta Dianan hautajais
Sadat suomalaiset ilmaisivat osanottonsa Dianan kuolemasta ◆
Britannia hiljentyy Dianan hautajaisiin ◆ 6.9. klo 7.00,

Figure 1: Over 18000 Finnish news bulletins were organized on a document map of 30 by 54 units. A click on the document map first reveals a zoomed view of the area (not shown here), then a view of the contents of a single unit, and last the contents of a document. Interesting starting-points for exploration can be found using the search facility. The user was interested in Lady Diana's accident and especially the funeral, and thus entered "prinsessa Dianan hautajaiset" as a search expression. The map area with similar documents is marked with circles, and the titles of articles (about the funeral) in the best matching unit are shown in the box. In the surrounding area there are other articles related to the accident, i.e., actions of paparazzies, Queen Elisabeth etc.

which is hardly surprising given that all of the texts were written by only a handful of interacting journalists. Starting to process Finnish texts called for two changes in the preprocessing stage. The list of common words to be excluded was created again for Finnish. Furthermore, each inflected word form was replaced with the corresponding base form obtained using the TWOL morphological analyzer by Lingsoft (Koskenniemi, 1983). A simple rule which utilizes the marginal densities of various inflection types was used for deciding cases where the same inflected form had several alternative interpretations.

The lack of a predefined classification of the documents made the evaluation of the map difficult. On the other hand, subjective evaluation was easy in this case because the news bulletins were generally understandable and interesting. Thus it was possible to obtain feedback from actual users, and for the first time also the search facility of WEBSOM was put to a more general test. User responses regarding the initial map identified a problem: many of the words that people used as search keys were not found on the map due to the fact that words occurring less than 50 times had been excluded for computational reasons. This problem was alleviated by starting to use the random mapping method of the document vectors for controlling the dimensionality of map input, which allowed the occurrence limit to be dropped to 10. As a result the user satisfaction improved considerably since searches had a much higher probability of success. If obtaining well organized document maps is sufficient, a large portion of the rare words may well be discarded; the documents will still be well organized due to redundancy of expression in the documents. However, many good search keys appear rarely in a document collection. Therefore, when the search facility is of particular importance, rare words should be included in the document representations or utilized in the search in some other way.

CONCLUSIONS

The WEBSOM method has been successfully applied to text materials that are quite different both in terms of the size of the collection as well as the type, domain, and language of the text material. Based on the case studies it seems that the following properties of the document collection suggest that extra attention should be paid on choosing the most suitable processing strategy at each stage: 1) the topics of the documents are very close to each other, 2) the writing style of the documents varies considerably, 3) each document is about many different topics, and 4) there are very few documents and so the sample of documents does not give a statistically reliable picture of the topic area. The case studies illuminate how to choose an appropriate strategy to achieve best qualitative results and processing speed for a particular type of collection. Language-specific technology may be helpful in the preprocessing of some types of material, but the core method appears to work well independent of language. In the future fast language-specific preprocessing methods could be studied further. It is also important to develop additional methods for evaluating the document maps.

ACKNOWLEDGEMENTS

I would like to thank Dr. Timo Honkela, Dr. Samuel Kaski, and Academy Professor Teuvo Kohonen for insightful conversations regarding many of the issues discussed in this article. I would also like to thank the National Board of Patents and Registration of Finland for the collection of patent abstracts.

REFERENCES

Honkela, T. (1997). *Self-Organizing Maps in Natural Language Processing.* PhD thesis, Helsinki Univ. of Technology, Espoo, Finland.

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996). Newsgroup exploration with WEBSOM method and browsing interface. TR A32, Helsinki Univ. of Technology, Lab. of Computer and Information Science, Espoo, Finland.

Kaski, S. (1998). Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of IJCNN'98, International Joint Conference on Neural Networks*, vol. 1, pp. 413–418. IEEE Service Center, Piscataway, NJ.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1996). Creating an order in digital libraries with self-organizing maps. In *Proc. of WCNN'96, World Congress on Neural Networks, Sept. 15-18, San Diego, CA*, pp. 814–817.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). WEBSOM—self-organizing maps of document collections. *Neurocomputing.* Accepted for publication.

Kohonen, T. (1982). Self-organizing formation of topologically correct feature maps. *Biol. Cyb.*, 43(1):59–69.

Kohonen, T. (1995). *Self-Organizing Maps.* Springer, Berlin, Heidelberg.

Kohonen, T. (1997). Exploration of very large databases by self-organizing maps. In *Proc. of ICNN'97, Int. Conf. on Neural Networks*, pp. PL1–PL6. Piscataway, NJ.

Kohonen, T., Kaski, S., Lagus, K., and Honkela, T. (1996). Very large two-level SOM for the browsing of newsgroups. In von der Malsburg, C., von Seelen, W., Vorbrüggen, J. C., and Sendhoff, B., eds., *Proc. of ICANN96, Int. Conf. on Artificial Neural Networks, Bochum, Germany, July 16-19, 1996*, Lecture Notes in Computer Science, vol. 1112, pp. 269–274. Springer, Berlin.

Koskenniemi, K. (1983). *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production.* PhD thesis, Univ. of Helsinki, Dept. of General Linguistics.

Lagus, K. (1997). Map of WSOM'97 abstracts—alternative index. In *Proc. of WSOM'97, Workshop on Self-Organizing Maps, Espoo, Finland, June 4-6*, pp. 368–372. Helsinki Univ. of Technology, Espoo, Finland.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In Simoudis, E., Han, J., and Fayyad, U., eds., *Proc. of the Second Int. Conf. on Knowledge Discovery and Data Mining*, pp. 238–243. Menlo Park, CA.

Lagus, K., Kaski, S., Honkela, T., and Kohonen, T. (1999). WEBSOM for textual data mining. *AI Review—special issue on data mining on the Internet.* Accepted for publication.

Lindén, K. (1997). Language applications with finite-state technology. *Int. Journal of Corpus Linguistics*, 2:1–7.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620.