Interactive Two-Level WEBSOM for Organizational Exploration

Timo Honkela[1] and Michael Knapek[1,2]

[1] School of Science, Department of Information and Computer Science
[2] School of Arts, Design and Architecture, Department of Media
Aalto University, P.O. Box 15400, FI-00076 Aalto, Finland

Abstract. Among the large number of applications of the self-organizing map (SOM) algorithm, creating maps of document collections have become commonplace since the introduction of the WEBSOM system. This article presents a novel development in WEBSOM research. The Interactive Two-Level WEBSOM, I2WEBSOM, includes two main components, a map of terms, and a dynamic map of documents. The map of terms is used to enable interactive feature selection and weighting. The map of documents is calculated using terminology-based feature vectors where their weights can be changed using the first-level map. In the experimental part, we focus on the application of creating maps of people based on their interest or competence profiles.

1 Introduction

We will describe in the following the classical WEBSOM method for information visualization, and how maps of people can be created using WEBSOM, for instance, to facilitate competence management.

1.1 WEBSOM in Exploration of Text Collections

The WEBSOM method was developed to facilitate interactive visual exploration of large document collections [1]. A central component of the WEBSOM is the self-organizing map (SOM) [2] that has proven to be an efficient and reliable means for projecting high-dimensional data into a low-dimensional space. The map consists of a number of model vectors $m_i(t)$ that are organized typically into a regular two-dimensional grid. For input data item $x(t)$ updated value $m_i(t+1)$ is computed iteratively using the well known update rule $m_i(t+1) = m_i(t) + (t)h_{ci}(t)[x(t)m_i(t)]$, where (t) is a scalar factor that defines the size of the update, index i is the model vector under processing, c is the index of the model that has the smallest distance from $x(t)$, and the factor $h_{ci}(t)$ is a smoothing kernel, also called the neighborhood function [2]. It has been shown that the SOM is a viable alternative to more recently developed methods that are based on information-theoretical or probability-theory principles especially when the trustworthiness of the visualization is used as the quality criterion [3].

The number of applications in analyzing and visualizing numerical data was already substantial by the beginning of 1990s even though the popularity of the SOM has since

V. Mladenov et al. (Eds.): ICANN 2013, LNCS 8131, pp. 579–585, 2013.

then grown to cover most areas of science and technology. The second important enabling step that led into the development of the WEBSOM method was the advent of the self-organizing semantic maps or, in other words, maps of words [4]. It was shown that the SOM can be used to create meaningful analysis of individual words based on the contextual statistics of each word. The relative syntactic or semantic similarity of two words can be detected by comparing the patterns of sentential contexts in which the words appear in text. The more similar the context patterns, the closer the relationship between the words is. This basic idea has been known already for a long time [5] and has been applied extensively during the recent years [6]. In the seminal work, artificially generated sentences were used [4]. The first map of words in written texts was based on analyzing the English translations of the fairy tales by Grimm brothers [7].

The idea of maps of documents had been brought up in early 1990s [8] and the first published experiments were based on the titles of documents [9]. The development of the WEBSOM method started in 1995 with the idea of creating "maps on information highways". The underlying motivation was based on the experiences in building a natural language database interface using traditional natural language processing and knowledge representation methods [10] and in applying developed modules in information retrieval [11]. A straightforward approach for creating maps of documents is to take the words appearing in the documents (or usually a subset of them) and to create a vector space in which each word corresponds to a dimension.

The first WEBSOM architecture had two levels: map of words was used to model similarities between words so that each word would not give rise to a single dimension but semantically related words could be grouped together [1]. The same kind of motivation is widely used in information retrieval applications that apply latent semantic analysis (LSA) [12].

In the two-level WEBSOM architecture, each document is encoded based on a map of words. A histogram of of the words in the document on the map is formed, and the histogram is normalized. This histogram resembles the vectors used in the vector space model but in this case the components of the vectors correspond to groups of words instead of single words [13].

In the later developments of the WEBSOM method, the two-level architecture was abandoned partly due to computational complexity issues. With the simplified architecture it was possible to create a map of millions of patents abstracts in which the number of connections between input and output layers was in the order of 10^{10} [14]. Since then, the WEBSOM concept has been applied in a large body of research (see, e.g., [15,16,17]).

1.2 Maps of People

One early application of the WEBSOM was creation of maps of people. In WSOM'97 conference, the participants were mapped based on the contents of their abstracts [18]. The idea has been extended in competence analysis [19] which is a central task in the area of human resource management (HRM). In HRM, data mining methods are becoming increasingly popular [20].

In the following, a novel development in WEBSOM research is described. The two-level architecture is re-introduced but based on a different approach in which the map of

words is used for feature selection and weighting. Moreover, the user interface is built to visualize also the organization process, not only the end result. This visualization is based on another two-level process in which the map units are adapted according to the SOM learning rule and the dynamics is shown as a movement of the data points on the map.

2 Interactive Two-Level WEBSOM

In the original WEBSOM two-level architecture, the map of words was used to find synonyms and otherwise closely related words [1,13]. This procedure helps in lowering the dimensionality of the input vector for the document map and in finding relationships between semantically related documents even when different words are used to describe the same thing. The random projection method also proved to be feasible in dimensionality reduction [13].

Since the early developments, much more powerful computational resources and automatic term selection methods have become available. Moreover, there is an increasing number of Semantic Web based and other terminological resources. Therefore, in the Interactive Two-Level WEBSOM (I2WEBSOM), the terminology is extracted or defined beforehand. The map of words becomes effectively a map of terms in which each term is related to the domain(s) of the application at hand. The I2WEBSOM has a two-level architecture:

- The map of terms is calculated to enable interactive feature selection and weighting.
- The map of documents is calculated using terminology-based feature vectors where their weights can be changed using the first-level map. In this paper, we focus on the application of creating maps of people.

The I2WEBSOM prototype has been fully implemented in Javascript with a json interface to the person database (people.aalto.fi). The profiles of the people in the system consist currently of a subset of the staff in the six schools of Aalto University. The descriptive documents are a concatenation of each person's description, associated terms, and titles of their publications. The I2WEBSOM architecture and a snapshot of the functional system is shown in Fig. 1.

2.1 Maps of Words as a Feature Selection Tool

The WEBSOM method has also been used to support qualitative research [21]. It was concluded that the utility of the SOM in improving inference quality follows from the fact that the method can easily be used to generate multiple well-grounded perspectives on the data. These perspectives are not a collection of random views but form an organized whole [21]. In a map of documents, different points of view can be obtained through different weightings of the terms. A seemingly neutral starting point is to weight all features equally or to use a weighting scheme that is based only on statistical criteria such as different variants of tf-idf.

The user often has preferences which makes some conceptual domains covered by the text collection more relevant than the others. As one of the strengths of the maps of

Fig. 1. An Interactive WEBSOM interface with three main elements: (1) A term map for choosing term weightings, (2) A dynamic people map, the order of which depends on the chosen term weights, and (3) Person information that can be viewed by clicking a person on the people map

documents is the possibility to obtain an overall view, this question of relevance is not related to individual queries but to the overall mapping function. Moreover, when the number of terms is in hundreds or thousands, one may wish to change the weights of groups of terms at once. A map of terms (Fig. 2) can be used to accomplish such task. Contextual information gives rise to a map in which related terms are close to other and groups of them form conceptual clusters. In an interactive interface, one can choose an area on the map and apply a weight changing operation on the selected terms. For instance, in relation to a map of university staff, one may choose to give increased weight to terms, e.g., related to business and software engineering. The effect of the change is such that the areas that contain persons working in these areas will be magnified.

2.2 Visualizing Map Organization

The vast majority of SOM-based data analysis and visualization tools are based on the idea that only the end result, an organized map is given to the user. The dynamic process of self-organization is typically shown only when the algorithm is demonstrated in educational settings. In I2WEBSOM, the organization and re-organization of the document map is shown to the user. The main motivation is to help the user in detecting the effects of the weight changes and to provide an understanding of the nature of dimensionality reduction.

The dynamic visualization of the organization process takes place by showing the position of each input vector x_j. The positions are changed based on the iterative updates of the model vectors $m_i(t)$. If the change of the best-matching units would be visualized straightforwardly, it would be impossible in practice to follow the process in sufficient detail. Therefore, the changes in the locations of the data points, loc_i, are updated based on the update rule $loc_i(t+1) = loc_i(t) + \beta[BMU_i(t)loc_i(t)]$. For the parameter β, the step size used in the visualization, the value 0.1 has been found to be appropriate. Due to the limitations of a printed document, the organization process is not shown here but examples can be found at $http://research.ics.aalto.fi/cog/websom/$.

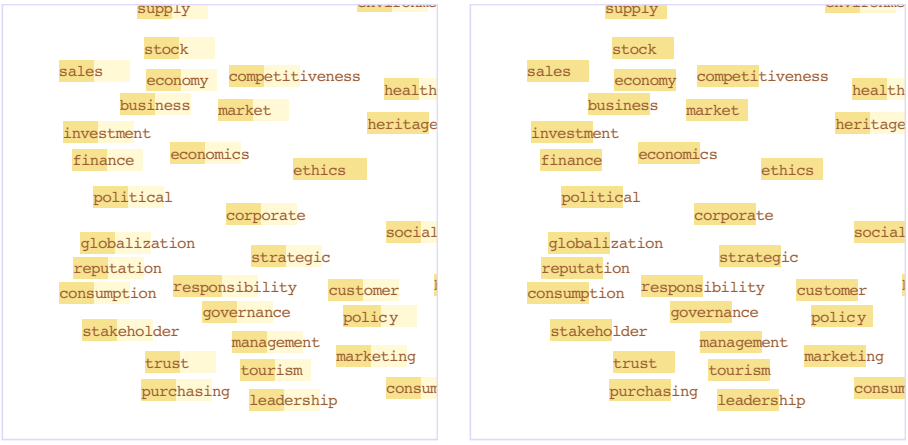

Fig. 2. An example of a map of terms where a part of the map is shown. Each term is associated with a slide bar that can be used to change the weight of the term. Also a group of terms can be manipulated simultaneously. On the left hand side, a neutral weighting is in use, and on the right hand side, the weights of a selected set of terms have been increased.

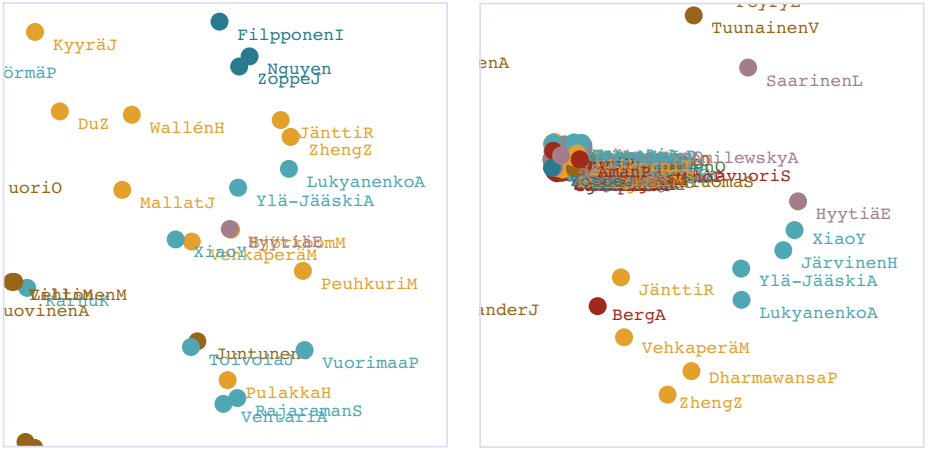

Fig. 3. Two versions of a zoomed map of people. The coloring of each node indicates the school of the person. On the left hand side, the map has been created with a neutral weighting in which each term has a similar weight. On the right hand side, the map has become restructured based on reweighting of the terms. One notable effect is that those people whose interests are defined by terms with a low weight are grouped tightly together. This helps in exploring the relevant parts of the map.

The end result of the organization process is in this case a map of people where two persons are close to each other if they have relatively similar interests or competences. Fig. 3 shows two examples of an area of a map of people.

3 Conclusions and Discussion

In this article, a novel development related to using the self-organizing map in the visualization of document collections has been presented. Different points of view into a document collection can be formed by the I2WEBSOM method. An interactive term map can be used in changing the weights of terms and terms groups.

The usefulness of the SOM in data visualization has been shown in practice through numerous applications as well as quantitatively [3]. There is also a large number of applications of the WEBSOM and closely related methods, and their usefulness has been shown in relation to information retrieval tasks [16]. The I2WEBSOM method increases the users' control over the organization of the document map and helps in creating transparency between data and the end result.

A quantitative evaluation of the I2WEBSOM method in a traditional manner would be extremely difficult as the motivation is to enable different points of view into the same data, based on interactions with the user. Moreover, the method has not been developed in one particular information retrieval task in mind but it can serve several purposes simultaneously. One may wish to obtain an overall view on the people in an organization or look for people with a particular interest or competence profile. Therefore, it remains a future task to determine what kind of user studies are needed to evaluate in a detailed manner the method presented in this paper. These are naturally beyond the scope of this paper. Instead, with this method we hope to indicate novel kinds of future application possibilities for unsupervised neural network and machine learning techniques.

Acknowledgments. We are grateful to Oliver Manner and Jan Fabritius who have developed the Aalto People web service and provided access to the data used in the experiments. The support from Philip Dean and Juhani Tenhunen is gratefully acknowledged as well as the collaboration with Jorma Laaksonen and Hannele Törrö in the earlier stages of the project.

References

1. Honkela, T., Kaski, S., Lagus, K., Kohonen, T.: Newsgroup exploration with WEBSOM method and browsing interface. Technical Report A32, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland (1996)
2. Kohonen, T.: Self-Organizing Maps. Springer, Heidelberg (2001)
3. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Networks 19(6), 889–899 (2006)
4. Ritter, H., Kohonen, T.: Self-organizing semantic maps. Biological Cybernetics (1989)
5. Harris, Z.: Distributional structure. Word 10(23), 146–162 (1954)
6. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. J. of Artificial Intelligence Research 37, 141–188 (2010)
7. Honkela, T., Pulkki, V., Kohonen, T.: Contextual relations of words in Grimm tales analyzed by self-organizing map. In: Proc. of ICANN 1995, Paris, EC2 et Cie, vol. 2, pp. 3–7 (1995)
8. Honkela, T., Vepsäläinen, A.M.: Interpreting imprecise expressions: Experiments with Kohonen's self-organizing maps and associative memory. In: Proc. of ICANN 1991, vol. 1, pp. 897–902 (1991)

9. Lin, X., Soergel, D., Marchionini, G.: A self-organizing semantic map for information retrieval. In: Proc. of the 14th ACM SIGIR, pp. 262–269 (1991)
10. Jäppinen, H., Honkela, T., Hyötyniemi, H., Lehtola, A.: A multilevel natural language processing model. Nordic Journal of Linguistics 11, 69–82 (1988)
11. Alkula, R., Honkela, T.: Development of text storage and information retrieval methods with natural language processing components. Final report of the FULLTEXT project (in Finnish). VTT, Espoo, Finland (1992)
12. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. J. of the American Society of Information Science 41, 391–407 (1990)
13. Kaski, S., Honkela, T., Lagus, K., Kohonen, T.: WEBSOM–self-organizing maps of document collections. Neurocomputing 21(1), 101–117 (1998)
14. Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A.: Self organization of a massive document collection. IEEE Transactions on Neural Networks 11(3), 574–585 (2000)
15. Ong, T.H., Chen, H., Sung, W.K., Zhu, B.: Newsmap: a knowledge map for online news. Decision Support Systems 39(4), 583–597 (2005)
16. Saarikoski, J., Laurikkala, J., Järvelin, K., Juhola, M.: A study of the use of self-organising maps in information retrieval. Journal of Documentation 65(2), 304–322 (2009)
17. Ding, Y., Fu, X.: The research of text mining based on self-organizing maps. Procedia Engineering 29, 537–541 (2012)
18. Lagus, K.: Map of WSOM 1997 abstracts–alternative index. In: Proc. of WSOM 1997, vol. 97, pp. 4–6 (1997)
19. Honkela, T., Nordfors, R., Tuuli, R.: Document maps for competence management. In: Proc. of the Symposium on Professional Practice in AI, pp. 31–39 (2004)
20. Piazza, F., Strohmeier, S.: Domain-driven data mining in human resource management: A review. In: Proc. of ICDMW 2011, pp. 458–465 (2011)
21. Janasik, N., Honkela, T., Bruun, H.: Text mining in qualitative research application of an unsupervised learning method. Organizational Research Methods 12(3), 436–460 (2009)