WSOM 2011
This is a self-archived pre-print version of this article.
The final publication is available at Springer via
https://doi.org/10.1007/978-3-642-21566-7_24.

On Wires and Cables: Content Analysis of WikiLeaks Using Self-Organising Maps

Rudolf Mayer and Andreas Rauber

Institute of Software Technology and Interactive Systems
Vienna University of Technology, Austria

Abstract. The Self-Organising Map has been frequently employed to organise collections of digital documents, especially textual documents. SOMs can be employed to analyse the content and relations between the documents in a collection, providing an intuitive access to large collections.
In this paper, we apply this approach to analysing documents from the Internet platform WikiLeaks. This document collection is interesting for such an analysis for several aspects. For one, the documents contained cover a rather large time-span, thus there should also be an quite divergence in the topics discussed. Further, the documents stem from a magnitude of different sources, thus different styles should be expected. Moreover, the documents have very interesting, previously unpublished content. Finally, while the WikiLeaks website provides a way to browse all documents published by certain meta-data categories such as creation year and origin of the cable, there is no way to access the documents by their content. Thus, the SOM offers a valuable alternative mean to provide access to the content of the collection by their content.
For analysing the document collection, we employ the Java SOMToolbox framework, which provides the user with a wealth of analysis and interaction methods, such as different visualisations, zooming and panning, and automatic labelling on different levels of granularity, to help the user in quickly getting an overview of and navigating in the collection.

1 Introduction

Self-Organising Maps (SOMs) enjoy high popularity in various data analysis applications. Due to its interesting properties, the SOM has been used in several applications to automatically organise documents in a digital library by their content. Examples of text document organisation with the SOM include the WEBSOM project [3], where the contents of a newsgroup collection containing a million of articles were clustered on the map, or the SOMLib digital library system, which has been applied to various collections, such as news texts [1].

Organising the content of textual documents, aided by fitting mechanisms to visualise the structures and to label the map, allows for an intuitive approach to explore the contents of a previously unknown collection. The SOM provides a two-dimensional knowledge map interface to the collection, where users can

quickly identify sets of related or distinct documents. Labelling methods function as an aid in identifying the topics of these groups of documents.

We apply this technique to the collection of diplomatic cables published by the Internet Platform WikiLeaks. These documents are comparable to longer newspaper articles, in both length, writing style, and information content. The collection is rather diverse in several aspects. It is diverse in temporal origin, with the single documents being authored in a span of several years to decades. Secondly, the documents were authored by many different authors, albeit from the same profession and country of origin; still, they exhibit different styles of writing. Moreover, the documents cover topics from different regions all over the world.

There are several ways to browse the collection on the WikiLeaks website – by creation year, by origin, and also by tag. While the latter seems most promising to access documents of a certain topic, they are in fact rather unusable for that purpose: there is a total of 590 tags, most of them are non-obvious 3-4 letter acronyms. Only some of the tags contain e.g. names of people. Thus, other means of organising the collection are necessary. The Self-Organising Map provides a valuable means to organise the documents by their content, and present these to the user in a way that they can on the one hand quickly get an overview of the topics in the collection, and one the other hand efficiently navigate through the single documents.

The remainder of the paper is structured as follows. In Section 2 we briefly describe the SOM framework employed. Section 3 then provides details on the WikiLeaks diplomatic cable collection, and the feature extraction method employed. In Section 4 we then present an analysis of the collection's content.

2 SOM Framework

We employ the Java SOMToolbox framework[1], developed at the Vienna University of Technology. Besides the standard SOM learning algorithm, the framework includes several implementations and modifications to the basic algorithm, such as the Growing Grid or the Growing Hierarchical Self-Organising Map (GHSOM). The core of the framework is an application that supports the user in an interactive, exploratory analysis of the data organised by the map training process. This application allows for zooming, panning and selection of single nodes and regions among the map.

To facilitate the visual discovery of structures in the data, such as clusters, a wealth of approximatively 15 visualisations are provided. The visualisations utilised in the experiments later in this paper are now described briefly.

The unified distance matrix, or U-Matrix [9], is one of the earliest, and most popular visualisations of the SOM. It aims at visualising cluster boundaries in the SOM grid. It is calculated as the input-space distance between the model vectors

[1] `http://www.ifs.tuwien.ac.at/dm/somtoolbox/`

vectors of adjacent map units. These distances are subsequently displayed on the map by colour-coding. Figure 1 depicts an U-Matrix with a grayscale colour map, where lighter shades of grey indicate high distances, and thus cluster boundaries.

The objective of Smoothed Data Histograms[5] (c.f. Figure 2(a)) is to uncover the clusters in the data through an estimation of the probability density of the data on the map. They build on the basic principle of an hit histogram, but they rather rely on a smoothed data histogram which is computed by not only increasing the histogram of the best-matching unit of an input vector, but up to s additional units.

Gradient Fields [6] aims at visualising cluster structures of a SOM, while using a special analogy for the markers utilised in the method. The rationale is that many persons with engineering background are not familiar with colour-coded maps as used e.g. in the U-Matrix, while they are generally familiar with the concept of vector fields. The gradient is displayed on the map with one arrow per unit, with the length and direction of each arrow indicate the location of cluster centres. The arrows at large form a smooth vector field. The arrows are computed based on the model vectors, the map topology, and a neighborhood kernel. Each arrow a_i for a map unit is obtained by computing weighted distances between the units model vector and all other model vectors, which are then aggregated and normalised.

The Thematic Classmap visualisation [4] shows the distribution of meta-data labels or categories attached to the data vectors mapped on the SOM. It colours the map in continuous regions in such a way that the regions reflect the distribution and location of the categories over the map, similar as e.g. a political map does for countries. The method is based on a segmentation of the SOM grid using Voronoi diagrams. A Voronoi diagram of a set of Points $P = p_1, \ldots, p_n$ partitions a plane in exactly n Voronoi regions, each being assigned to one point $p \in P$, so that all the points in a region are closest to p_i. Applied to SOMs, the plane is the visual representation of the map, and the number of regions is equal to the number of units with at least one data item mapped onto. Units with no data items will be split by the algorithm to become parts of other, adjacent regions. The voronoi cells are then coloured according to the categories attached to the data mapped onto the units in each voronoi cells. Details on the colouring can be found in [4].

To assist the user in interpreting the content of the SOM, we automatically generate labels for the map. We employ the LabelSOM method [7], which assigns labels to the units of the SOM describing the features of the data items mapped onto them. The method utilises the *quantisation error* of the vector elements. The quantisation error(q_{i_k}) is the sum of the distances for a feature k between the node's weight vector m_i and the input vectors x_j mapped onto the node. A low quantisation error thus characterises a feature that is similar in all input vectors to the weight vector, which is assumed to be a descriptive feature. To eliminate features that have low quantisation error due to not being present on that node, i.e. having zero values, we require the mean value of the feature to

be above a certain threshold.

As the SOM does not generate a partition of the map into separate clusters, a clustering of the units is applied to identify the regions in the map computationally. Of advantage are hierarchical algorithms, which result in a hierarchy of clusters the user can browse through, allowing different levels of granularity; the framework provides the Ward's linkage [2] and several other linkage methods. Having clusters or regions identified, the framework also provides labelling of these entire regions. Making use of the properties of the hierarchical clustering, we can also display two or more different levels of labels, some being more global, some being more local.

Even though labelling the map regions assists the user in quickly getting a coarse overview of the topics, labels can still be ambiguous or not conveying enough information. Thus, the framework also employs Automatic Text Summarisation methods to provide a summary of the contents of the documents of interest, allowing the user to get a deeper insight into the content. The summarisation can either be on single documents, documents from a certain node, a cluster, or from a user-selected set of nodes or documents. Different summarisation algorithms are provided; the user can also specify the desired length of the summary.

3 Collection

The WikiLeaks diplomatic cable collection[2] is composed of United States embassy cables, allegedly "the largest set of confidential documents ever to be released into the public domain". The cables date from the 1960s up until February 2010, and contain confidential communications between 274 embassies in countries throughout the world and the State Department in Washington DC.

The cables are released subsequently, thus currently a subset of 3,319 documents is available. The subset contains cables originating from 165 different sources (embassies, consulates and other representations), and covers mostly the last few years. Details on release year and origin of the dataset are given in Table 1.

It can be noted that a rather large portion of approximatively 12% of the cables were issued by the embassy in Tripoli. A large numbers of documents also originates from Brazil (10.4%, including the cables from the consulates in Sao Paolo and Rio de Janeiro), and Iceland (8.6%). Countries where the USA are involved in military actions, such as Afghanistan or Iraq, have not been published yet in large quantities, thus distinguishing this collection from the Afghan and Iraq war diaries published earlier by WikiLeaks.

[2] `http://wikileaks.ch/cablegate.html`

(a) Documents per year

period	documents
1960s & 1970s	6
1980s	6
2000-2002	11
2003	24
2004	100
2005	167
2006	292
2007	378
2008	684
2009	1270
2010	434

(b) Document origin

Origin	Documents
Libya	406
Brazil	351
Iceland	290
Spain	202
Secretary of State	158
The Netherlands	146
France	122
Russia	93
Egypt	81
Afghanistan	77
UK	75
Pakistan	59
China	58

Table 1. Cablegate document collection as of February 2011

To obtain a numeric representation of the document collection for our experiments, we used a bag-of-words indexing approach [8]. From the resulting list of 65,000 tokens, the features for the input vectors were selected according to their document frequency, skipping stop-words, as well as too frequent (in more than 50% of the documents) and too infrequent (in less than 1% of the documents) terms. This resulted in a feature vector of approximately 5,500 dimensions for each document, which formed the basis of the maps subsequently trained. The values of the of the vector are computed using a standard $tf \times idf$ weighting scheme [8], which assigns high weights to terms which appear often in a certain document (high tf value), and infrequent in the rest of the document collection (high idf value), i.e. words that are specific for that document.

4 Experimental Analysis

We trained a map of the size of 35 $\times$ 26 nodes, i.e. a total of 910 nodes for the 3,319 text documents. Due to the uncertain legal situation of the Wikileaks documents, we have to refrain from publishing any quotes from the cables, or other details, in this paper.

After inspection of the initial map, it became obvious that the map was dominantly organised along the origin of documents. The reason is that most cables describe events in the country the embassies are located in, thus the names of such countries are too predominantly represented. Thus, for having a more topic-oriented map, we decided to remove the most frequent country names from the feature vector. While this step influences the content of cables that might talk about foreign countries, this side-effect seems acceptable.

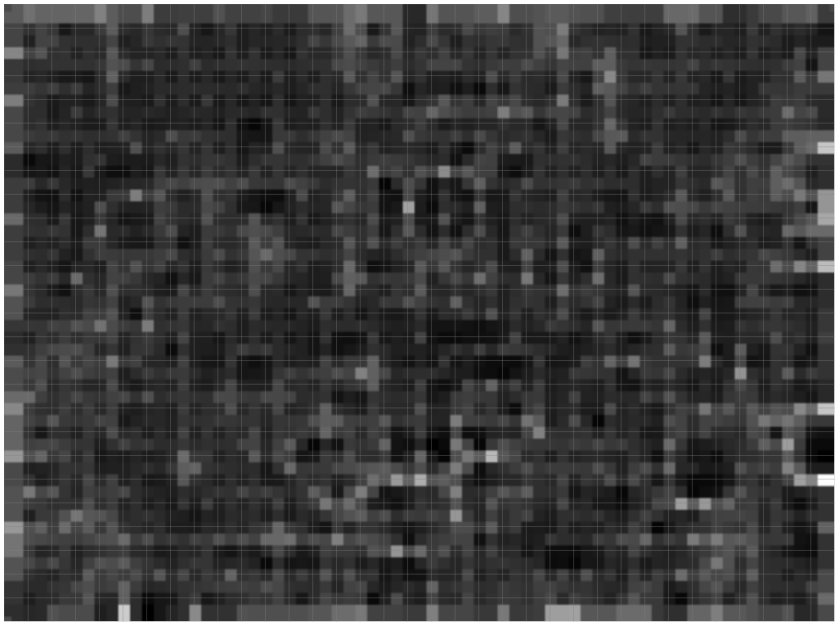
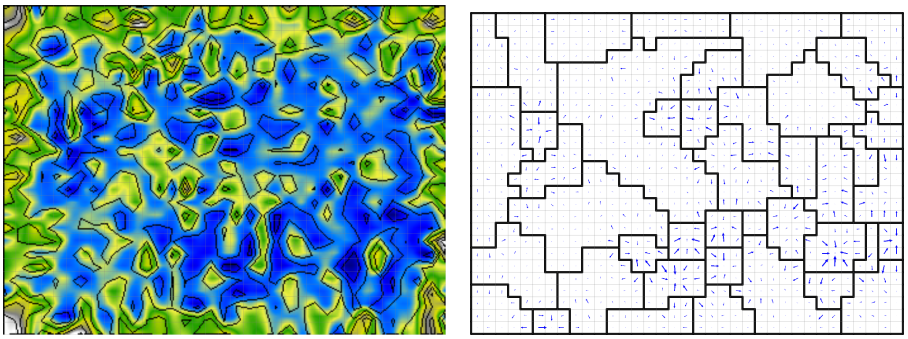


Fig. 1. U-Matrix of the Cablegate SOM

The U-Matrix visualisation of this map is depicted in Figure 1. However, on this data set, only a few local boundaries become apparent. The existence of smaller, interconnected clusters is also confirmed by the Smoothed Data Histograms, which visualises density in the map, in Figure 2(a). Figure 2(b) shows the Vector Fields visualisation, where the arrows point towards local cluster centres. These clusters overlap very well with a clustering of the weight (model) vectors of the map, with the clustering for 40 target clusters being superimposed in the same illustration. It can be observed that especially the centre area does not seem to have a clear cluster centre.

The Thematic Classmap visualisation depicted in Figure 3 shows the distribution of the origin of the cables. It can be observed that the SOM manages to separate the classes very well, especially on the edges of the map. Overlapping areas are mainly found in the centre of the map, which has previously been identified as an area without a clear cluster centre, and on the upper-left corner. It is often those areas, where the external classification scheme contradicts the topical similarity, which are the most interesting to uncover unexpected relations.

Figure 4 finally shows the Cablegate map with 40 clusters, each of which has two labels assigned, using the LabelSOM method described in Section 2. The display of labels on regions helps to quickly get an overview on the contents of



(a) (b)

Fig. 2. Smoothed Data Histograms (a) and Vector Fields (b) visualisations

the map, and where to find them. We will describe some of the regions in detail now.

The upper-left corner of the map prominently features diplomatic cables discussing nuclear programs, both of Iran and North Korea, and related issues, such as sanctions and the role of the International Atomic Energy Agency (IAEA). As this is a topic which involves international diplomacy on a large scale, also the sources of origin mentioning the topic are manifold – from the secretary of state and embassies of countries involved into the UN proceedings to cables from the UN representation in Vienna, seat of the IAEA. Topics also dealt with in this area of the map are weapons and the military in general.

The cluster on the central upper edge of the map features reports on the Russian-Georgian war in 2008, and other topics related to Russia. The neighbouring cluster, holding messages mostly about energy such as oil and gas, also features Russian politics, and Russian companies, as well as cables from other countries, such as Venezuela, Nigeria, and Libya. The cluster right next to it, in the top-right corner, then deals with further topics concerning the North-African country ('gol' stands for government of Libya). One topic is for example the diplomatic crisis between the country and Switzerland, which resulted in Switzerland refusing Schengen-Visa.

To the left of this, towards the centre of the map, are two clusters with reports on Iceland, one of them identified by the names of the former prime minister Geir Haarde and the former minister for foreign affairs, Ingibjörg Gísladóttir, who had to step down from office due to the financial crisis hitting Iceland in 2008. The other cluster deals with reports on the Icelandic banks, which suffered intensely from the crisis.

In the lower-centre of the map, a large area is dedicated to topics regarding Brazil. These are dealing with ethanol and other biofuels, which Brazil is a major producer of. Other topics include the defense sector (Nelson Jobim serves as the Minister of Defense). On the left, two clusters deal with other South American issues, namely Bolivian politics, or the crisis between Colombia and Venezuela,

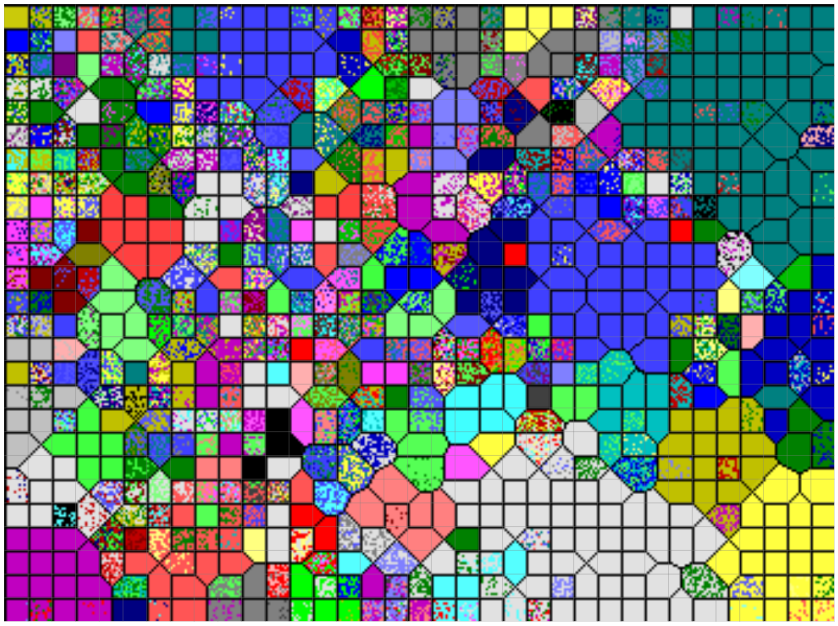


Fig. 3. Thematic Class Map showing the origin of the cables

reported by cables from both countries. Another topic in that region is the Venezuelan president Hugo Chavez.

Towards the left, certain documents talking about Afghanistan are located. Several of them deal with drugs, while others talk about the involvement of the UK and Spain in the war. Just above that, cables report on Taliban activity, and the situation in Pakistan, as well as cables from India about the attacks in Mumbai, which are linked to terrorists in Pakistan. The region right of that, towards the centre of the map, generally gathers cables from many different sources, all talking about terrorism and criminal activities, without a major topic dominating.

Another interesting arrangement of documents can be found on the left-centre area, which features the previously mentioned documents from Iran and closely to it also Sweden. Many of the documents in the cluster about Sweden deal with the Swedish stance towards the sanctions against Iran due the nuclear programme of the latter.

5 Conclusions

In this paper, we presented a case study for analysing text documents with Self-organising Maps. We employed a framework that provides extensive support in visualisations that uncover structures in the data, and other methods which help to quickly communicate the contents of the collection as a whole, and certain

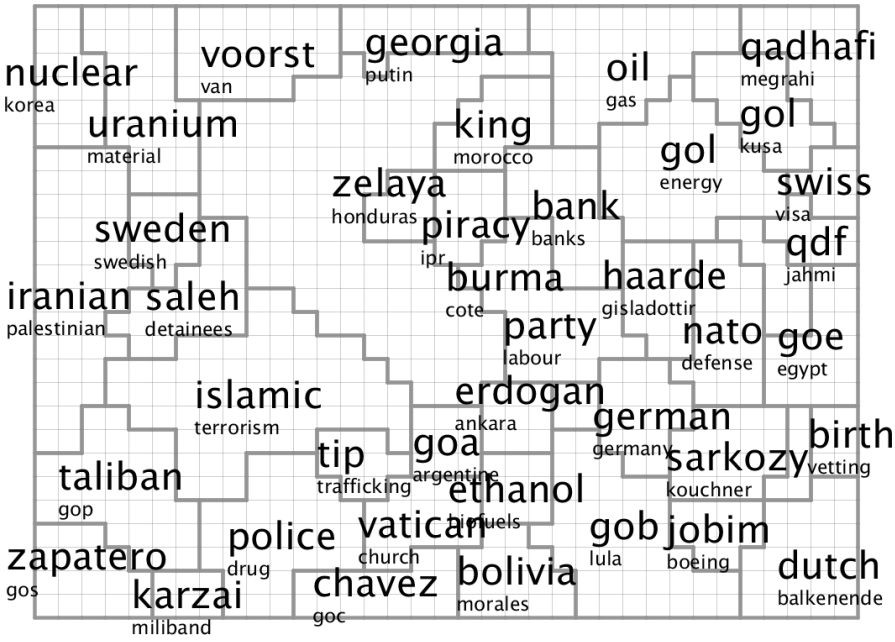


Fig. 4. Clustering of the map, with 40 clusters and two topic labels each

parts in particular, such as labelling on cluster level. With such an analysis tool, the user is able to rapidly get on overview on the interesting areas on the mapping, and gets access to the collection. This approach clearly exceeds the means available on the WikiLeaks website, which comprise of category-based browsing, but lack means to communicate the topics of the collection and exploring the collection by its content.

As the collection of cables is growing on a daily basis, an online version of the map is available at `http://www.ifs.tuwien.ac.at/dm/`, being regularily updated.

References

1. M. Dittenbach, A. Rauber, and D Merkl. Business, culture, politics, and sports – how to find your way through a bulk of news? On content-based hierarchical structuring and organization of large document archives. In *Proceedings of the 12th International Conference on Database and Expert Systems Applications*, Lecture Notes in Computer Science, Munich, Germany, Sept. 3-7 2001. Springer.
2. Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, March 1963.
3. Teuvo Kohonen, Samuel Kaski, Krista Lagus, Jarkko Salojärvi, Vesa Paatero, and Antti Saarela. Organization of a massive document collection. *IEEE Transactions on Neural Networks, Special Issue on Neural Networks for Data Mining and Knowledge Discovery*, 11(3):574–585, May 2000.

4. Rudolf Mayer, Taha Abdel Aziz, and Andreas Rauber. Visualising class distribution on self-organising maps. In Joaquim Marques de Sá, Luís A. Alexandre, Włodzisław Duch, and Danilo Mandic, editors, *Proceedings of the International Conference on Artificial Neural Networks (ICANN'07)*, volume 4669 of *LNCS*, pages 359–368, Porto, Portugal, September 9 - 13 2007. Springer.
5. Elias Pampalk, Andreas Rauber, and Dieter Merkl. Using Smoothed Data Histograms for Cluster Visualization in Self-Organizing Maps. In *Proceedings of the International Conference on Artifical Neural Networks (ICANN'02)*, pages 871–876, Madrid, Spain, August 27-30 2002. Springer.
6. Georg Pölzlbauer, Michael Dittenbach, and Andreas Rauber. Advanced visualization of self-organizing maps with vector fields. *Neural Networks*, 19(6–7):911–922, July–August 2006.
7. A. Rauber and D. Merkl. Automatic labeling of Self-Organizing Maps for Information Retrieval. *Journal of Systems Research and Inf. Systems (JSRIS)*, 10(10):23–45, December 2001.
8. Gerald Salton. *Automatic text processing – The Transformation, Analysis, and Retrieval of Information by Computer.* Addison-Wesley Longman Publishing Co., Inc., 1989.
9. Alfred Ultsch and H. Peter Siemon. Kohonen's Self-Organizing Feature Maps for Exploratory Data Analysis. In *Proceedings of the International Neural Network Conference (INNC'90)*, pages 305–308, Dordrecht, The Netherlands, 1990. Kluwer Academic Press.