Discovering Web Document Clusters with Self-Organizing Maps

Emil Şt. CHIFU*, Călin CENAN*

Abstract
The self-organizing maps (SOM) represent a data mining and visualization method for complex high dimensional data sets. We have applied the SOM model in Web mining, by giving sets of documents as input data space for SOM. The result of applying SOM on a set of documents is a map of documents, which is organized in a meaningful manner so that documents with similar content appear at nearby locations on the two-dimensional map display. A document map clusters the data, resulting in an approximate model of the data distribution in the high dimensional document space. The paper describes some promising experimental results, where a couple of meaningful clusters have been discovered by our system in a subset of the “20 newsgroups” data set. The clustering capability of our system allows users to find out quickly what is new in a Web site of interest by comparing the clusters obtained from the site at different moments in time.

1 Introduction
The self-organizing map (SOM) is a very popular unsupervised neural network model for the analysis of high dimensional input data [7]. It is a clustering, visualization and abstraction method based on displaying the data set in another, more usable representation form. SOM allows mapping the high dimensional input data onto a lower dimensional (more concrete, two-dimensional) output space. The resulting map is a two-dimensional grid of arrays, which preserves the structure of the input data as faithfully as possible: data items – represented as vectors of numerical attribute values – which are close to each other (i.e. similar) in the high dimensional data space tend to be also close to each other on the map. The main advantage of the self-organizing maps is that large quantities

* Department of Computer Science, Technical University of Cluj-Napoca, Romania, {Emil.Chifu, Calin.Cenan}@cs.utcluj.ro
of data can be organized quickly into a compact form that reveals the structure within the data. As such, a SOM map displays an overview of the data.

A somehow non-classical approach in the mining of Web documents is the one based on the self-organizing maps [4, 5, 7]. Our implemented SOM-based system is able to manage large collections of (hyper) text documents by ordering them semantically on a SOM map. We show that the system is capable of finding semantically meaningful clusters on a map of documents. The cluster visualization capability is based on applying the unified-distance matrix (U-matrix) algorithm on a SOM map [3, 12, 13]. The flat clusters can be visually discovered with the help of different grey-levels on the map as induced by the U-matrix algorithm.

2 Self-organizing maps

Teuvo Kohonen has created the self-organizing maps as a particular kind of neural networks [7]. A SOM map visualizes important relationships among the data – which are latent in the input data set – in an easily understandable way. Even though there are no explicit clusters in the input data set, important relationships are nevertheless latent in the data. SOM can discover and illustrate these latent structures of an arbitrary data set. SOM can describe different aspects of a phenomenon in any domain, provided that vectors of numerical attributes can represent the data in the domain.

The map learns by a self-organization process. No a priori knowledge about the membership of any input data item (vector) in a particular class or about the number of such classes is available. Hence, the training proceeds with unlabeled input data like any unsupervised learning. The clusters (classes) are instead discovered and described with gradually detected characteristics during the training process.

The map consists of a regular two-dimensional (rectangular) grid of processing units – the neurons. Each unit has an associated model of some multidimensional observation, represented as a vector of attribute values in a domain. SOM learning is an unsupervised regression process that consumes at every iteration one available observation represented as a vector of values for the attributes in a given domain. The role of a learned map is to represent all the available observations with optimal accuracy by using a restricted set of model vectors associated to the map units.

2.1 The learning algorithm

The initial values for the model vectors – also referred to as reference vectors – of the map units can either be chosen depending on the problem domain or they can be taken randomly. Each iteration of the learning algorithm processes one input (training) vector as follows. Like usually for unsupervised neural
networks, some form of a competitive learning takes place: the winner unit index \( c \), which best matches the current input vector, is identified as the unit where the model vector is the most similar to the current input vector in some metric, e.g. Euclidean. Then all the model vectors or a subset of them that correspond to units centered around the winner unit \( c \) – i.e. units in the neighborhood area of \( c \) –, including the winner itself, are adjusted in the direction of the input vector. This updating forms a globally ordered map in the process of learning. A map unit has six immediate neighbors in a hexagonal map topology, which is usually the preferred topology. This is merely a hexagonal lattice type of the two-dimensional array (grid) of neurons, and the SOM map is kept as a planar rectangle.

2.2 Cluster visualization

A subset of data items that are close to each other in a high dimensional input data space – and thus defines a cluster in the input space – are arranged to a map area consisting of nearby neurons in the two-dimensional SOM display. As a consequence, the problem of discovering a cluster in a high dimensional data set with the help of the self-organizing maps reduces to the problem of discovering the map area whose neurons to contain all the data in the cluster. Actually, we have to find the boundaries of the map cluster. Finding the boundaries of a SOM map cluster is based on applying the unified-distance matrix (U-matrix) algorithm on a SOM map [12].

U-matrix visualizes the map in grey-levels, in order to express how similar or dissimilar adjacent neurons are [3, 13]. In a hexagonal self-organizing map topology, six extra neurons around each neuron separate geometrically the neuron from its six immediate neighbors and show by grey-levels its similarity with each neighbor. The lighter a separating hexagon, the bigger the similarity of the reference vectors of the two separated neurons, and the darker the hexagon, the bigger the dissimilarity of the reference vectors. This way, SOM map clusters can be discovered visually as “valleys” or “depressions” (light areas) separated by “hills” (dark areas or borders). Moreover, the higher (i.e. darker) a hill separating two clusters, the more dissimilar the two adjacent clusters in the multidimensional input data space. There is also a large dissimilarity between two non-adjacent clusters when they are far away from each other on the map.

3 Self-organizing maps in web mining

Applying SOM on natural language data means doing data mining on text data, for instance Web documents [8]. The main problem of SOM as applied to natural language is the need to handle essentially symbolic input such as words. If we want SOM to have words as input then SOM will arrange the words into
word categories. But what about the input (training) vector associated to each input word? What should be the vector components, i.e. the attributes of a word?

When classifying words by SOM, the result is a word category map. The attributes of the words in our experiments were the count of the word occurrences in each document in a collection of documents. As such, we have chosen to represent the meaning of each word as related to the meanings of text passages (documents) containing the word. Symmetrically, the semantic content of a document is represented by a bag-of-words style function of the meanings of the words in the document. The lexical-semantic explanation of this contextual usage meaning of words is that the set of all the word contexts in which a given word does and does not occur provides a set of mutual constraints that captures the similarity of meaning of words and passages to each other. The measures of word-word, word-passage and passage-passage relations are well correlated with several cognitive phenomena involving semantic similarity and association [9].

4 System architecture

The architecture of our system is based on two self-organizing maps. The first one creates a semantically ordered spread of all the word forms in a large collection of Web documents. This is also called the map of word categories or level 1 SOM. The second SOM (called document map or level 2 SOM) represents a semantically ordered spread of all the documents in the collection, where the documents are codified as vectors that are histograms of word categories. The word categories are the ones as already induced into the word category map units. This way we have reduced the dimensionality of the document vectors from thousands of components that would correspond to thousands of different word forms in a classical bag-of-words approach. The dimensionality is reduced to around 200 or 300 components, which correspond to 200 or 300 different word categories, enough to express the number of different concepts in a shallower or wider domain. Thus the reduced dimensionality removes the noise caused by the variability in word usage; since the number of dimensions is much smaller than the number of word forms, minor differences in terminology will be ignored.

4.1 System implementation

The system is written in C, and the LEX software package [11] is used for implementing the preprocessing module, which reads and counts the word occurrences in all the documents in a collection, by ignoring all the HTML tags. The preprocessing module also ignores 450 common words, i.e. English words having no semantic load. These words have been taken from the information retrieval software package GTP [2]. Finally, the preprocessing also means a
stemming phase that uses a morphological analyzer for English, which is part of the GATE system [1]. The stemming is done in order to reduce the number of word forms by keeping only their stem.

The SOM_PAK [6] system is used for the training of all our SOM maps. The result of training the document SOM is a text file containing for every document category a list of document names that belong to that category, i.e. the list of documents managed into the corresponding map unit. The format of this text file is exemplified by seven document categories in Fig. 1, where each row corresponds to a different map unit. The first two integer numbers in each row represent the rectangular coordinates $(x$ and $y$) of the current unit. The document category name follows the coordinates of the unit and becomes the identification label of the unit. The document category name is given by the name of the first document in the (training) data set that “hit” the unit during the training process.

The seven document categories in Fig. 1 are semantically related as they all contain as documents emails from one and the same newsgroup (talk.politics.mideast) in the “20 newsgroups data set” [10]. The seven corresponding map units are neighbors on the document map and they constitute together an area or cluster. The aggregation of the neurons in this cluster is noticeable from their coordinates and from the hexagonal topology adopted.

4.2 Graphical user interface

The graphical interface has been implemented by using the PHP language. The interface reads the text file of document categories (exemplified in Fig. 1) and translates this document classification into a dynamical HTML file, which is the interactive graphical display of the document map itself. Every map unit is labelled with the associated document category name. A second label on each map unit represents the number of documents in the corresponding category.

```
8  3 75381  :  75381
8  4 75382  :  75369  75382
9  4 75394  :  75371  75389  75394
10  4 75393  :  75393
11  4 75400  :  75370  75392  75400
8  5 75395  :  75395
9  5 75399  :  176854  75366  75387  75388  75390  75399
```

Fig. 1. Example document categories
The interface allows the navigation on the document map from any Web browser. A click on a map unit gives access to an index of documents in that unit, which is also a dynamically generated HTML file. The index contains a list of links that point to the documents themselves.

5 Experimental evaluation

The experiments reported here take as test data the “20 newsgroups data set” [10]. This data set contains 20,000 UseNet news postings having the form of email messages. The 20,000 messages were collected randomly from 20 different Netnews newsgroups, 1000 messages from each newsgroup [10]. The data set is “labelled”, by being already partitioned into twenty categories. This labelling helped us to evaluate the clustering of email documents as discovered by our document SOM. In one of our most successful experiments, we have selected randomly 40 documents from each newsgroup, summing up a total set of 800 message documents. This balanced subset of the original “20 newsgroups” data set has been taken as input data space for our SOM-based system in order to arrive at an email document SOM map.

An important question in this experiment was to choose a size for the SOM map, in order to arrive at a map with the highest degree of visual expressiveness for clustering [13]. The map size means the total number of neurons of the rectangular grid. For a given data set, different map sizes mean different granularity levels, in terms of the average number of data items to belong to a neuron. If the map is too small, it is too rough and consequently it might hide some important differences that should be detected in order to separate the clusters. This is because too many non-similar data items could belong to the same neuron. When the map is too big, then it is too detailed and, besides the important differences, the map displays also too small differences, which are often unimportant for clustering. This is because data items that are very similar could belong to different neurons, when normally we expect them to belong to one single neuron.

We have chosen a map size of 16 (columns) times 12 (rows) considered as suitable for the input data space of 800 data items (800 email documents). This also conforms to the suggestions in [13], where a suite of experiments with input data sets of different cardinality and different SOM map sizes is described. Fig. 2 shows the result email document SOM map image, where grey levels occur as an effect of applying the U-matrix algorithm for cluster visualization. The U-matrix algorithm used here is included in the SOM_PAK program package [6], which is part of our system. The algorithm conforms to the description in Section 2.2.
5.1 Clustering results

The document map in Fig. 2 clearly illustrates five clusters, as discovered by the document map. All the 17 documents in Cluster 2 belong to the newsgroup `talk.politics.mideast`, but there are 41 messages in the input data that belong to this newsgroup. Cluster 2 contains seven neurons whose explicit description in terms of email messages grouped in document categories in each neuron is given in Fig. 1. Actually only 40 input messages are “officially” labelled by the authors of the “20 newsgroups” data set to belong to the newsgroup `talk.politics.mideast`. One more message, named 176854, and found out by our map to belong to Cluster 2, has been “abusively” put by the authors into another newsgroup, namely `talk.politics.misc`. The header of this email indicates explicitly `Newsgroups: talk.politics.mideast, misc.headlines, talk.politics.misc`.

![Document SOM map for 800 email messages taken from the “20 newsgroups” data set](image)

Fig. 2. Document SOM map for 800 email messages taken from the “20 newsgroups” data set
Cluster 3 contains 12 messages, 11 of them from the newsgroup rec.sport.hockey. This cluster is less clearly bordered on the map, because of the semantic overlap with other messages, some of them form the related newsgroup rec.sport.baseball. In fact, the only message in Cluster 3, which is outside of the expected newsgroup rec.sport.hockey, is from the related newsgroup rec.sport.baseball. Table 1 shows the classification quality parameters accuracy and coverage associated with the five clusters discovered by the map in Fig. 2.

5.2 Discussion of results

There are some more results found out from our document map induced from 800 news messages and illustrated in Fig. 2. For instance, there is one more cluster, Cluster 5, also mentioned in Table 1, which contains 26 email messages, 21 of them being a mixture of messages from three different newsgroups: talk.religion.misc, soc.religion.christian, and alt.atheism. The first two newsgroups are obviously related to each other, and they are also semantically related with the third, even if this relation sounds more like an antonymy. Similar topics are nevertheless discussed in messages about religion and atheism.

About 85% of the 800 email messages are contained in about the left half of the map, which is almost completely white, and constitutes a huge cluster. Such a cluster has no clear semantic content, because it contains messages from all the 20 newsgroups, including the messages left out from the five clusters already mentioned. The technical explanation for this phenomenon is that the document SOM map was unable to display semantic differences in this big cluster. The differences in the semantic content of the messages could be too small when the authors of the messages use too few words specific to the domain of the newsgroup or sometimes when they communicate announcements with no bearing with the domain of the newsgroup.

Table 1

Classification accuracy and coverage associated with document clusters on Fig. 2

<table>
<thead>
<tr>
<th>Cluster No.</th>
<th>Newsgroup</th>
<th>Accuracy (Correct/Actual)</th>
<th>Coverage (Correct/Predicted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>science.space</td>
<td>100% (15 / 15)</td>
<td>37.5% (15 / 40)</td>
</tr>
<tr>
<td>2</td>
<td>talk.politics.mideast</td>
<td>100% (17 / 17)</td>
<td>41.5% (17 / 41)</td>
</tr>
<tr>
<td>3</td>
<td>rec.sport.hockey</td>
<td>91.5% (11 / 12)</td>
<td>27.5% (11 / 40)</td>
</tr>
<tr>
<td>4</td>
<td>comp.windows.x</td>
<td>90.9% (10 / 11)</td>
<td>25% (10 / 40)</td>
</tr>
<tr>
<td>5</td>
<td>Combination of talk.religion.misc, soc.religion.christian, alt.atheism</td>
<td>80.8% (21 / 26)</td>
<td>17.5% (21 / 120) where 120 = 3*40</td>
</tr>
</tbody>
</table>

8
6 Conclusions

The self-organizing maps constitute a powerful model for Web mining by defining a visual overview of a collection of Web documents. By using the unified distance matrix (U-matrix) algorithm, our Web mining system is able to find semantically meaningful clusters on a map of documents. We have reported here some promising experimental results from document clustering.

As a further work, we can apply our Web mining system in order to arrive at a visual overview of all the documents in a given Web site. This overview is a snapshot image of the given site for a given moment in time. The clustering capability of our system allows users to find out quickly what is new in a Web site of interest by comparing the clusters obtained from the site at different moments in time.

In order to improve the ability to display semantic differences for clustering, we will introduce some weighting in our bag-of-words approach, for instance the inverse document frequency. Words or word categories occurring in too many documents will receive a low weight, because of their low discriminating power (i.e. low information gain). We can also ignore all the words that only occur once or only occur a few times in the whole collection of documents.

References


