A&QT-R 2002 (THETA 13)
2002 IEEE-TTTC International Conference on Automation, Quality and Testing, Robotics
May 23 – 25, 2002, Cluj-Napoca, Romania

A SYSTEM TO RETRIEVE WEB DOCUMENTS FROM A LARGE COLLECTION

Emil Şt. Chifu, Ioan Alfred Letia, Viorica R. Chifu

Technical University of Cluj-Napoca, Department of Computer Science
Baritiu 28, RO-3400 Cluj-Napoca, Romania
Fax: +40-64-194491, Email: Emil.Chifu@cs.utcluj.ro, Ioan.Alfred.Letia@cs.utcluj.ro,
viorica@observ.obs.utcluj.ro

Abstract: We present a new approach in information retrieval of Web documents, together with an implemented system. The method is applicable to any collection of (hyper) text documents and is especially suitable when the user has rather limited knowledge about the domain or the contents of the text collection. The approach is based on the self-organizing maps (SOM) [2, 6, 10]. Our system manages a large collection of HTML documents by spreading them on a SOM map. Semantically similar documents occupy the same position or neighbor positions on the map, depending on the degree of semantic content similarity. The system allows the user to navigate on the document map, in order to retrieve relevant documents.

Keywords: information retrieval, self-organizing maps (SOM), unsupervised statistical machine learning, data mining on (hyper) text documents, semantic similarity.

1. INTRODUCTION

One of the most time consuming task for Web users is to find relevant information from a vast quantity of available information. Creating efficient search engines is the aim of the information retrieval area, a recently emerged application of natural language processing.

One of the problems of traditional keyword search methods in information retrieval is the difficulty to pose the most suitable keywords in the query, so that the answer doesn't miss relevant documents, and also doesn't produce a long list of irrelevant ones. This problem can be resolved by coding semantic content of text documents based on word categories rather than on individual word forms. By a word category we mean a set of semantically homogeneous words, i.e. words either having similar meanings or having the same meaning. In our approach, producing the word categories is the result of a machine learning process applied on collections of text documents. Using word categories rather than actual words in coding the semantic information space for texts means removing the noise and reducing the dimensionality in this information space. It is more reasonable to think of text documents as talking about two or three hundreds of different concepts or topics rather than as talking about thousands of concepts expressed by thousands of individual word forms in a document of reasonable size.

Another difficulty in keyword search information retrieval methods occurs when the Web user has only vague knowledge about the field to search in. In such cases the search expression tries to find documents in a vaguely delimited field and the answers will obviously be of poor quality. This problem is resolved by allowing the Web user to browse a map where documents are ordered by their semantic content.

We will present a method and a system to retrieve Web documents from a large collection. The approach starts from the two above-mentioned problems. The method is based on the self-organizing maps (SOM) [2, 6, 10]. Our implemented system manages a large collection of HTML documents by spreading them on a SOM map. Semantically similar documents occupy the same position or neighbor positions on the map, depending on the degree of semantic content similarity. The system allows the user to navigate on the document map, in order to retrieve relevant documents. This "navigation" is done by using any Web browser and consists in the facility to see the list of document names contained at any location on the map and to display the content of any document in that list.

2. SELF-ORGANIZING MAPS

The self-organizing maps have been created by Teuvo Kohonen as a particular kind of neural networks [6, 10]. There are multiple views on SOM; the different definitions are the following. SOM is a model of specific aspects of biological neural nets (the ordered "maps" in the cortex). SOM is a model of unsupervised machine learning (and an adaptive knowledge representation scheme). SOM is a tool for statistical analysis and visualization: it is both a projection method which maps a high dimensional data space into a lower dimensional one and a clustering method so that similar data samples tend to be mapped on nearby neurons. SOM is a data mining and visualization method for complex high dimensional data sets.

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

2.1. The Learning Algorithm

The initial values for the models – also referred to as reference vectors - of the map units can either be chosen depending on the problem domain or can be taken randomly. Each iteration of the learning algorithm processes one training vector (one sample) $x(t)$ as follows. First, the winner unit index c, which best matches the sample, is identified as the unit where the model vector is most similar to the current training vector in some metric, e.g. Euclidean.

$$\| x(t) - m_c(t) \| \leq \| x(t) - m_i(t) \| , \text{ for any unit index } i \qquad (1)$$

Then all model vectors or a subset of them that belong to units centered around the winner unit c – i.e. units in the neighborhood area of c - are updated as follows.

$$m_i(t + 1) = m_i(t) + h_{ci}(t) * [x(t) - m_i(t)] \qquad (2)$$

h_{ci} is the neighborhood function, which is a decreasing function on the distance between the *i*-th and *c*-th units on the map grid. In practice, the neighborhood area is chosen to be wide in the beginning of the learning process, and both its width and height decrease during learning. A map unit has six immediate neighbors in a hexagonal map topology, which is usually the preferred topology. In a rectangular topology, a map unit has only four immediate neighbors and consequently the number of neighbor units affected during the learning is smaller.

During the learning of a map, the reference vectors of the map units become ordered. Similar reference vectors become close to each other and dissimilar ones become far from each other on the map.

3. APPLYING SOM ON NATURAL LANGUAGE DATA

Applying SOM on natural language data means doing data mining on text data, for instance Web documents. The role of SOM is to cluster numerical vectors given at input and to produce a topologically ordered result. 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 cluster the words in 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? Similarity in word appearance is not related to the word meaning, e.g. “window”, “glass”, “widow”.

We have chosen to classify (cluster) words by SOM, creating thus word category maps. The attributes of the words in our experiments were the count of the word occurrences in each document in a collection of documents. Consequently, we have chosen to represent the meaning of each word as related to the meanings of text passages (documents) containing the word and, symmetrically, the meaning of a document as a 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 (documents) 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 [8]. The meaning of semantically similar words is expressed by similar vectors.

After training a SOM on all the words in a collection of documents – where the vectorial coding of words represents the contextual usage -, the result self-organizing map clusters words in semantic categories. There are also other possibilities to code words, which lead to grammatical or semantic word categories [2, 3, 6, 11].

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 will be referred to as level 1 SOM. The second SOM represents a semantically ordered spread of the documents themselves – level 2 SOM. The final aim of this architecture is to classify the document collection by using the criterion of semantic similarity. Hence the graphical browsing interface of our system is in essence the level 2 SOM, i.e. a document map. In the experiment illustrated here we started from a collection of 139 Web news documents in the politics and sport domains, the majority of the documents being taken from CNN.

First, a SOM map of word categories (level 1 SOM) is created by training on words coded as vectors as explained in section 3. We had 7820 vectors of 139 components each, corresponding to 7820 distinct word forms in 139 documents. We have created a map with 16x12 units with hexagonal topology. Out of the 192 units (neurons) on the map, 190 represented word categories. The remaining two units have been left empty, unused by this learning process, i.e. no word clustered into these two units. Two examples of word categories as discovered by the level 1 SOM follow.

```
 15  0 game  :  stanley  lemieux  penguin  playoff  central  win  var
cup  hit  nhl  jersey  season  hockey  adsync  si  event  score  devil
final  illustrate  textarray  function  snow  linkarray  http  team
goal  game                                                        (3)
 1  5 macedonium  :  country  reuter  south  month  human  fight
plane  rebel  military  international  move  macedonium           (4)
```

This is a text representation of two word clusters corresponding to two units on the word category map. The first two integer numbers represent the coordinates of the current map unit. The most representative word in the cluster (category) follows after the coordinates. The enumeration of words in the cluster follows the colon character.

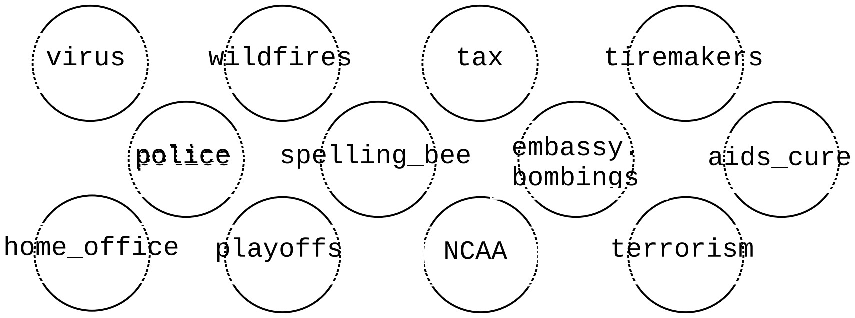

Figure 1. A document map that organizes 139 documents.

Next, the system codifies the documents in the collection as vectors that are histograms of word categories. Such a histogram is a vector having as many components as are word categories on the word category map (level 1 SOM). For every word category, the histogram contains the number of word form occurrences in the current document which belong to that word category on the level 1 SOM. This way, we have reduced the dimensionality of the information space in which the semantic content of our documents is represented: the document information space uses 190 learned word categories rather than 7820 word forms. The 139 document vectors (word category histograms) are used for training the SOM map of documents (level 2 SOM). This time we had 139 vectors with 190 components each, the result being a map with hexagonal topology having 4x3 units, illustrated in figure 1. The 12 units correspond to 12 document categories. Three examples of document categories as discovered by the level 2 SOM follow.

```
 2  1 embassy.bombings  :  Hanssen  accused_spy  embassy.bombings (5)
 1  2 playoffs  :  Berlin  Bucks_Sixers  Devils  Diamondbacks_Giants
Hornets  TripleCrown  XFL  playoffs                               (6)
```

```
 3  2 terrorism  :  CeBit  Indonesia  Macedonia  Microsoft  Mid_East
Milosevic  Mir  Mori  NSA  Napster  Oklahoma_City  Oscar  UK_vote
US_China  aids.research  energy  energy_crunch  global_warming
immigration  power_crisis  stadium  submarine  terrorism          (7)
```

Here the text representation of the document categories is the same as for the word categories (3), (4), except that document names – without their `.html` or `.htm` extension - occur here in place of word forms. Notice that the semantically related document clusters `embassy_bombings` and `terrorism` correspond to neighbor units on the map.

5. IMPLEMENTATION

The system is written in C and bash script. We have used the LEX package [9] 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 package GTP [4]. Finally, the preprocessing also means a stemming phase that uses a morphological analyzer for English, which is part of the GATE system [1]. This is done in order to reduce the number of word forms by keeping only the stem of each.

The SOM_PAK [5] system is used for training both levels SOM maps. The result of training the level 2 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 on the corresponding map unit. The format of this text file is exemplified in (5)-(7).

5.1. Graphical User Interface

The graphical interface has been implemented by using the PHP language. The system creates the graphical interface as an interactive graphical display that is implemented as a dynamically created HTML file. This PHP module reads the document categories from the corresponding file (created by the level 2 SOM and exemplified in (5)-(7)) and translates this document classification into a dynamic HTML file which is the graphical display of the document map itself, i.e. the level 2 SOM. Every map unit is labeled with the most relevant and representative document in the corresponding category, i.e. the name of the document whose vector is closest to the model vector of that unit [7]. A second label on each map unit represents the number of documents in the corresponding category. For instance, the map unit for the document category `terrorism` (see (7)), having coordinates 3, 2 on the map, is also labeled `23`.

The interface allows the navigation on the document map from any browser. The aim of this navigation is the retrieval of relevant documents in two steps. Click on a map unit gives access to the list of documents in that unit, which is also a dynamically generated HTML file, containing a list of links, each of which having as text the document name and pointing to the document itself. Then click on a document name in this list allows viewing that document.

6. CONCLUSIONS

The system presented in this paper implements a new information retrieval method based on organizing and exploring large document collections. The system allows browsing a document map, where documents are grouped on semantic similarity criteria. Semantically similar documents occupy the same position or neighbor positions (units) on the map, depending on the degree of similarity between them. This way, groups of documents “talking” about different topics occupy different zones on the map.

As a future work, when having to deal with very large document collections, allowing the navigation on an interesting map area in more detail by zooming the map display becomes necessary. Such a map has a large number of units and is too vast to be looked at and browsed as a whole. This observation also leads to the idea of introducing searching by keyword queries (content addressable search) in order to identify from the beginning an area of a very large map a user is interested in.

Another extension necessary in the context of a very large document collection is the facility to browse in more detail a zoom area on the map. Hierarchically zooming the document map will lead to a hierarchical classification of documents. Hierarchically zooming the word category map will similarly lead to a hierarchical classification of words. This last observation will allow an experiment in which a hierarchical classification of words in a domain will be compared with the corresponding area in WordNet [10].

REFERENCES

1. R. Gaizauskas, P. Rodgers, H. Cunningham, K. Humphreys, S. Robertson, [1998], GATE User Guide, Department of Computer Science, University of Sheffield, UK, http://www.dcs.shef.ac.uk/~hammish/gate-rc1, http://gate.ac.uk.
2. T. Honkela, [1997], Self-organizing maps in natural language processing, *PhD thesis*, Neural Networks Research Center, Helsinki University of Technology, Finland.
3. T. Honkela, S. Kaski, K. Lagus, T. Kohonen, [1996], Exploration of full-text databases with self-organizing maps, In *Proceedings of the International Conference on Neural Networks (ICNN’96)*, vol. I, pp. 56-61.
4. S. Howard, H. Tang, M. Berry, D. Martin, [2001], General text parser, University of Tennessee, Department of Computer Science, http://www.cs.utk.edu/~lsi/gtp-request.html.
5. T. Kohonen, J. Hynninen, J. Kangas, J. Laaksonen, [1996], SOM_PAK, The self-organizing map program package, *Report A31*, Helsinki University of Technology, Faculty of Information Technology, Laboratory of Computer and Information Science.
6. T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Peetero, A. Saaerela, [2000], Self-organization of a massive document collection, *IEEE Transactions on Neural Networks*, **11**, 3.
7. K. Lagus, S. Kaski, [1999], Keyword selection method for characterizing text document maps, In *Proceedings of ICANN’99*.
8. T.K. Landauer, P.W. Foltz, D. Laham, [1998], Introduction to Latent Semantic Analysis, Discourse Processes, 25, 259-284.
9. M.E. Lesk, E. Schmidt, Lex – a lexical analyzer generator.
10. G.A. Miller, R.. Beckwith, Chr. Fellbaum, D. Gross, K. Miller, [1993], Introduction to WordNet: an on-line lexical database, *International Journal of Lexicography*, **3**, 4, pp. 235-244.
11. H.F. Pop, [2000], Self-organizing map in text mining, *Research report*, Natural Language Division, Department of Computer Science, University of Hamburg, Germany.