

Text Classification and Labelling of Document Clusters with Self-Organising Maps

Andreas Rauber⁽¹⁾, Erich Schweighofer⁽²⁾, Dieter Merkl⁽¹⁾

(1) *Institute of Software Technology, Vienna University of Technology*
Favoritenstraße 9-11, A-1040 Wien, Austria
www.ifs.tuwien.ac.at/~andi *www.ifs.tuwien.ac.at/~dieter*

(2) *Institute of Public International Law, University of Vienna*
Research Center for Computers and Law
Universitätsstraße 2, A-1090 Wien, Austria
www.univie.ac.at/RI/erich.html

Abstract. The freely available law on the Internet could be one of the best application areas of text classification and labelling. This paper explores the high potential of the self-organising map for information reconnaissance by classifying and describing unknown legal text collections. The maps can be seen as topic-oriented libraries that are automatically created without intellectual input. The clustered topics - units of the self-organising map - are labelled with the most appropriate keywords. Extensive tests have shown the potential of this approach.

Introduction

More and more legal information is available in electronic form. This welcome situation offers a high potential of more assistance for lawyers and citizens in getting access to law. However, while the size and availability of electronic information has changed a lot, ways for representing and interacting with those collections could not keep pace. Therefore, more attention has to be devoted to challenges relating to the interaction with these digital legal libraries. The link between the complex legal order and information needs is quite difficult to handle. Searching these collections requires users to define their queries in some Boolean logic based expressions, specifying large sets of keywords and synonyms, requiring both knowledge of the problem domain as well as basic query formulation experience. Results of queries are usually presented as long lists of (both relevant and irrelevant) retrieved documents sorted according to some ranking criteria, with the large overall number of documents retrieved usually inhibiting efficient search. The IR community has invested much effort to improve the retrieval with AI methods (for an overview on related research see [9, 10, 12]).

With conventional libraries and the long time they had to evolve according to the needs of their users, we find a rather different situation. Documents are mostly sorted by topic, allowing users to easily orient themselves in large, unknown collections, finding out which topics are covered and where specific documents are to be found, additionally to other search indices like author and title catalogues. Furthermore, searching these libraries by asking a librarian provides the user with a pointer to relevant sections in the library rather than a (1-dimensionally relevance-sorted) pile of books. Adopting these characteristics of conventional libraries for electronic document archives to combine the benefits of the evolved structures of conventional systems with the benefits of digital library systems has proven to be difficult. Obviously, a way to automate this process is needed, automatically classifying documents according to their content, and arranging those clusters of documents on a two-dimensional map in such a way, that similar topics are located next to each other. This representation may then be considered as a map of a document archive [4].

The content of documents is conveyed by the words that the documents are made up of. Documents on similar topics will thus contain, by and large, a highly similar set of words. As a consequence of this we can think of text documents as forming topical clusters in a high-dimensional feature space spanned by the individual words in the documents based on the so-called vector space model of information retrieval [7]. Detecting the patterns, i.e. clusters within a collections of documents allows us to organise these according to their topical similarity. Among the large number of algorithms available for cluster analysis the self-organising map (SOM) [2], an unsupervised neural network, has shown to perform excellent with respect to the organisation of document archives. Being an unsupervised neural network it requires no manually pre-classified or other supervised training data. It rather learns the structure of a high-dimensional feature space, which in our case is the feature space spanned by the text documents' feature vector representations. This high-dimensional feature space is mapped onto a two-dimensional map-space while preserving the overall topology as faithfully as possible. This leaves us with a two-dimensional organisation of documents similar to conventional libraries, facilitating interactive exploration. With the *LabelSOM* technique, we are furthermore able to automatically extract descriptions of the clusters found by the SOM.

Text representation

What we need is a way to (preferably automatically) obtain a representation of the document's contents in this vector space. This is achieved by using full-term indexing, that is, the documents are described by all words that appear in the document collection. Since the dimensionality of the resulting feature spaces will be rather high (i.e. the number of distinct words appearing in a document collection is rather high), some statistical methods may be employed to reduce the size of the feature space. Furthermore, weighting schemes may be used to automatically assign different weights to words contributing stronger to topic description or rather, topic discrimination for a given collection. The most prominent family of weighting schemes is the so-called term frequency times inverse document frequency weighting, or *TFx-IDF* scheme, where words are considered highly important if they appear frequently within one single document, yet rarely in the collection as such. Unimportant words, such as pronouns or articles, are assigned a lower weight, as they appear in many documents. The same applies for misspelled words, as they usually occur only very rarely within a document.

Self-organising map

The self-organising map [1, 2] is one of the most distinguished unsupervised artificial neural network models. It basically provides a form of cluster analysis by producing a mapping of high-dimensional input data onto a usually 2-dimensional output space while preserving the topological relationships between the input data items as faithfully as possible. These output units are arranged according to some topology, the most common choice of which is a two-dimensional grid. Each of the output units i is assigned a weight vector m_i .

During each learning iteration, the unit c having the highest activity level with respect to a randomly selected input pattern $x = [\xi_1, \xi_2, \dots, \xi_n]^T$ is selected and adapted in such a way as to decrease the difference between that unit's weight vector m_c and the input pattern x . Unit c is further referred to as the winning unit, the winner in short. A common choice to compute the activity level of a unit is marked by the Euclidean distance between the input pattern and that unit's weight vector. Adaptation takes place during each training iteration and is realised as a gradual reduction of the difference between the respective components of input and weight vector. The amount of adaptation is guided by means of a learning-rate α that gradually decreases in the course of training. In addition to adapting the winner, a number of units in a time-varying and gradually decreasing neighbourhood of the winner is adapted too. Thus, during the training steps, a set of units around the winner is tuned towards the currently presented input pattern. This leads to a spatial arrangement of the input patterns such that alike inputs are mapped onto regions close to each other in the grid of output units. As a consequence, the training process results in a topological ordering of the input signals. The spatial range of units around the winner that are subject to adaptation may be described by means of a time-decreasing neighbourhood function h_{ci} taking into account the distance (in terms of the output space) between unit i currently under consideration and unit c , the winner of the current learning iteration.

It still remains, however, a challenging task to label the map, i.e. to determine those keywords of input patterns mapped onto a particular unit that are characteristic for the cluster. With our *LabelSOM* approach (for a more detailed description see [5, 6]), every unit of the map is labelled with the keywords that best characterise all documents that are mapped onto that particular unit. This is achieved by using a combination of the mean quantisation error of every feature and the relative importance of that feature in the weight vector of the unit. Vector elements having about the same value within the set of input vectors mapped onto a certain unit describe the unit in so far as they denominate a common feature of all input patterns of this unit. The mean quantisation error for that particular vector value will be small. The corresponding feature, i.e. index term in our application, may be used as a label for the unit. Thus, index terms that have a deviation δ below a certain threshold τ_1 are candidates for labelling.

A specific problem with a keyword-based document classification is that a large number of features in the document vectors will have a weight of zero (keywords not appearing in the respective documents). In order to avoid the use of these features with the minimal quantisation error as labels, a threshold parameter τ_2 is introduced describing the minimum value for a weight vector element so that only features exhibiting a certain importance with respect to their *TFx-IDF* are used for labelling.

Evaluation

At present, experiments have been performed using text collections from 43 to about 6000 documents. The first test circle with a small text corpus has proven the feasibility of the approach and will be describe in detail below. The test environment for the second test comprises of three text collections of the most important documents of European law in English (583 documents), German (572 documents) and French (626 documents) in HTML and TXT-format. In the third run, we have enlarged these document collection but taken only the ECJ judgements. The text collections consist of 317 (English), 383 (German) and 342 (French) documents. These documents have been segmented into sections according to their logical structure as well as into paragraphs of comparable length allowing multiple assignments of one single document according to several topics covered. The numbers of documents thus increased to 5087 (English), 5961 (German) and 5978 (French) documents. Due to space considerations, only the first test environment will be presented here.

The text corpus consists of 43 European law documents concerning public enterprises from the databases EUR-Lex and CELEX. We represent each document by a vector with a length of 2160 words using a *TFxIDF* weighting scheme [7]. The word list was selected automatically. Words are stemmed to their approximate root form. Firstly, these vectors were used for a classification of the document space. The self-organising map gives a useful overview about the document space showing good concentrations of documents. We provide the user with some instruments for refining the exploratory analysis: size of the self-organising map and threshold value for labelling. In our experiments, we have worked with 3x3 to 6x6 maps. The results of a 5x5 self-organising map are presented hereafter, the other results are quite similar. A bigger size of the map leads to a higher granularity of the topic-oriented library with more book shelves, e.g. a 5x5 map is equivalent to 25 book shelves.

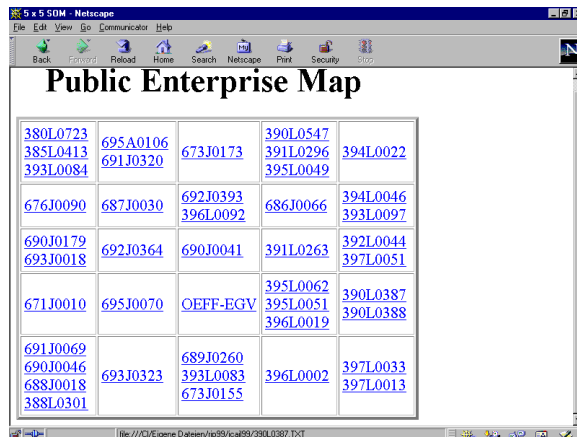


Figure 1: Self-organising Map, 5x5 units

Figure 1 shows the quite satisfying results of the 5x5 self-organising map. An indication of the good classification may be the units of transparency of financial relations of public enterprises and member states [1,1]¹, energy single market (transit of electricity and gas through grids) [1,4], electricity single market [2,3]), satellites [2,5], telecommunication liberalisation (units at the lower right-hand site of the map, e.g. voice telephony [4,4]) and judgements concerning art. 90 ECT (arranged around the related provisions of the Treaty on the European Community, here called OEFF-EGV). The labelling and description of the various units is a very ambitious goal. We may note that the *LabelSOM* method works quite well in the case of a cluster with only one or more but very similar documents (e.g. unit [5,4] concerning mobile and personal communications). A much more demanding situation is present, however, when more documents are represented by the same unit.

We have performed a series of experiments with different threshold values regarding the selection of particular keywords as unit labels. As an example, we show the labels² of the 5x5 self-organising map shown in figure 1 with threshold values of 0.1 (figure 2).

Good examples of the potential of this approach are the following:

In the map shown in figure 1, unit [1,4] describes documents concerning the liberalisation of the energy market (electricity and gas). The corresponding labels (as shown in figure 2) *energy*, *transit*, *grid*, *pressure* and *electricity* give an indication of the concerned energy and the most important problem of the regulation of transit. *Entity*, *list* and *update* are related to implementation techniques. The obligation of transit applies only to a list of entities (in the annex) that has been updated in the latest directive.

As another example, unit [4,4] describes documents concerning the liberalisation of voice telephony. The labels (figure 2) are *organization*, *interconnection*, *voice*, *liberalisation*, *telephone*, *regulatory*, and *cable*. *Organization* refers to the main addressee of these directives – telecommunication organisations. The term *regulatory* focuses the necessary establishment of regulatory authorities and a regulatory framework. The other terms could be quite easily recognised as the main descriptive terms of telephone liberalisation.

Lowering the threshold value a bit lists substantially more labels and provides more information on the particularities of the regulation.

To summarise our observations, the automatically produced labels give a fine list of important keywords describing the contents of the document. This result is very promising but some insufficiencies remain that are the focus of our present and future research. A more detailed description of the experiments may be found in [11].

¹ We will use the notation [x,y] to refer to the unit in row x and column y.

² We provide the (quasi)root forms of the index terms as determined by our document parser.

Turnover, financi, manufact, capit, consolid, loan, fund, enterpri, annu, credit,	mail, complain, corbeau, offic, french, concess, post, contest, applican, rigi, lihg	italian, submissi, industr, fami, reduct, becaus, devolv, upon, employer, parti, alter, applican, cannot, contest, aid, pertain,	energ, transit, grid, list, entit, pressur, updat, electric,	entit, prospect, ent, applicat, particip, procurem, energ, opt,
bureau, vehicl, accident, insu, handl, settleme, card, liabil, motor, damag,	extern, concessi, concess, group, monopol, giniral, parent, bodson, cass, belong, french, pric, cour	electric, producer, customer, generat, buyer, distribu, eligibl	Flight, schedul, airlin, exempt, tariff, concert, regul, rout, void, bilater, tarif, multilat,	terrestr, spac, station, mark, satellit, earth,
porto, port, genova, italian, genoa, corsica, ferr, fly, siderurg, worker, work, vessel, maritim	navig, convent, collect, jurisdic, spac, rout, cass, aircraft, admissib, cour, ruling, greek, contract, republic, belgian, internat, kingdom, control, proceedi	employme, agenc, procurem, consulta, german, macrotro, statutor	Manufact, test, notifi, qual, mark, certific, inspect, typ, termin, standard, symbol, conform, laborato, affix,	termin, organiza, acces, standard, regulato, lin, replac, leas,
port, luxembou, upon, tribun, apprais, privileg, conferr, divis, question, formulat, infer, waterwa, duch, crimin, cannot, particul, however, befor,	car, health, profit, hom, runn, reimburs, peopl, particip, regional, region, contract, arangem,	tax, compensa, sect, elimin, aid, disadvan, liv, caracte, promot, progress, approxim, quantita, distur, procur, good, time, monopol, qualifi, apply, unfair,	organiza, intercon, voic, liberali, telephon, regulato, cable	essenti, termin, telex, packet, acces, licens, telephon, data, resal, leas, organiza, messag, switch, interfac,
typ, approv, connect, belgian, radiocom, model, minister, hir, telephon, termin,	centr, artifici, health, cass, french, intra, stock, import, approv, coopirat, agricol	cable, satellit, retransm, televis, broadcas, advertis,	mobil, communic, licenc, frequenc, intercon, infrastr, band, licens,	committe, proporti, license, attach, stop, univers, regulato, organiza, shopp, licenc, intercon

Figure 2: Labels and Description of the Self-organising Map, threshold 0.1

The major remaining problem from a legal point of view concerns the diversity of legal documents. *LabelSOM* presents a very useful list of common elements. If documents cover different issues, *LabelSOM* selects the most important common features. This analysis may not be the same as those of a legal expert. These problem became apparent in our extensive second test circle. At present, we are improving the vector representation using segmentation techniques and mark-up with SGML or XML but are also experimenting with a combination of exploratory text analysis (previous research of KONTERM I and II [3, 8, 10]). First experiments with the segmented documents of our third test run have shown promising results.

Conclusions

The high potential of the self-organising map offers information reconnaissance by classifying and describing unknown text collections. Topic-oriented libraries can be automatically created. The units of the self-organising map are labelled with the most appropriate keywords. Extensive tests have shown the potential of this approach. In the future, we will improve the vector representation of the documents in order to improve the context of classification.

Acknowledgements

This research was supported by the Jubiläumsfonds der Oesterreichischen Nationalbank, Vienna, research project no. 6888. We have to thank the Office for Publications of the European Communities for providing us with a document collection.

References

- [1] T. Kohonen, "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, Vol. 43, 1982.
- [2] T. Kohonen, *Self-organizing maps*, Springer-Verlag, Berlin, 1995.
- [3] D. Merkl and E. Schweighofer, "The Exploration of Legal Text Corpora with Hierarchical Neural Networks: A Guided Tour in Public International Law", *Proc. Int. Conf. on Artificial Intelligence and Law*, Melbourne, Australia, 1997.
- [4] D. Merkl, "Text classification with self-organizing maps: Some lessons learned", *Neurocomputing*, Vol. 21, No. 1-3, 1998.
- [5] A. Rauber and D. Merkl, "Automatic Labelling of Self-Organizing Maps: Making a Treasure-Map Reveal its Secrets", *Proc. Pacific Asia Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 1999.
- [6] A. Rauber and D. Merkl, "Using self-organizing maps to organize document archives and to characterize subject matters: How to make a map tell the news of the world". *Proc. Int. Conf. on Database and Expert Systems Applications*, Florence, Italy, 1999.
- [7] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, 1989.
- [8] E. Schweighofer, D. Merkl and W. Winiwarter, "Information Filtering: The Computation of Similarities in Large Corpora of Legal Texts", *Proc. Fifth Int. Conf. on Artificial Intelligence and Law*, Washington, DC, 1995.
- [9] E. Schweighofer, "The Revolution in Legal Information Retrieval, or: The Empire Strikes Back", *Proc. Conf. The Law in the Information Society*, Florence, Italy, 1998.
- [10] E. Schweighofer, *Legal Knowledge Representation*, Kluwer Law International, The Hague, The Netherlands, 1999.
- [11] E. Schweighofer and D. Merkl, "A Learning Technique for Legal Document Analysis", *Proc. Int. Conf. on Artificial Intelligence and Law*, Oslo, Norway, 1999.
- [12] H. Turtle, "Text Retrieval in the Legal World", *AI & Law*, Vol. 3, 1995.