Improvement of Vector Representations of Legal Documents with Legal Ontologies

Erich Schweighofer, Gottfried Haneder
Institute of Public International Law
University of Vienna
Research Center for Computers and Law
Universitätsstraße 2, A-1090 Wien, Austria
erich.schweighofer@univie.ac.at
gottfried.haneder@telering.at

Andreas Rauber, Michael Dittenbach
Institute of Software Technology, Vienna
University of Technology
Favoritenstraße 9-11, A-1040 Wien, Austria
andi@ifs.tuwien.ac.at
michael.dittenbach@ec3.at

Abstract

In previous research, a TFxIDF vector representation of legal documents was the basis for structuring and labelling a collection of legal documents. Extensive experiments with three text corpora of European law of about 580 documents in three languages have shown, that binary or weighted vector representation may not be sufficient. Even quite successful approaches of similarity computation have problems in identifying the best context of classification.

This paper deals with a new approach. The vector representation is modified using a simple ontology of the domain. The adaptation of the vector is done using the newly developed Data Enrichment feature of the SOM/KONTERM library.

This approach was evaluated with a small text corpus of 253 documents of cyberspace law. The improvement was significant in the quality of the clusters but also the labels.

1. Introduction

In these days the amount of legal information is growing faster and faster. Archiving and access to legal documents has much improved. Yet, it is becoming more and more difficult to search and select the appropriate information. The huge legal information systems require powerful instruments for classification and data analysis. Boolean search or other search engines are not sufficient for the needs of the legal community.

The IR community has invested much effort to improve the retrieval with AI methods (for an overview on related research see [1, 18, 19, 20, 22]). The best solution for the representation of documents and the similarity problem seems to be a vector space representation and susequent cluster analysis for classification and description of contents. One of the most distinguished unsupervised neural network in this domain certainly is the self-organising map [6]. It is a general unsupervised tool for ordering high-dimensional data in such a way that alike input items are mapped close to each other. In order to use the self-organising map to explore text documents, we represent the various texts as the histogram of its words. With this data, the artificial neural network performs the exploration task in a completely unsupervised fashion. The method *LabelSOM* [14] can properly describe the common similarities of the cluster. An extension to the SOM architecture, the GHSOM [4] can automatically represent the inherent hierarchical structure of the documents.

In this paper we deal with the improvement of the vector representation for legal documents. We have to be aware that legal documents are very diverse. A *TFxIDF* vector representation and cluster analysis may result in only a subset of many common similarities of a particular topical cluster being detected and used for cluster formation. In order to focus on the desired vector weights of the documents, we refine the automaticially computed weights with the data enrichment tools. Based on a very simple ontology, the weight values for important words are increased leading to a conceptual representation of the documents.

The material presented in the remainder of this paper is organised as follows. In Section 2 we give the details of the document vector representation and the neural network based system we use for the training and subsequent description of units (document clusters). Section 3 describes the ontology. In Section 4, we present details on the refinement of the vectors. In Section 5 we provide an analysis of our results. Finally, we provide some conclusions in Section 6.

2. Document vector, self-organising map and LabelSOM

Documents are represented as vectors in the self-organising map. Although no limitations exist for weighted vectors, we use *TFxIDF* weighting scheme vectors containing all words except very frequent ones.

The self-organising map [3, 6, 7, 8, 9, 10, 11, 12, 13, 17, 27] is a general unsupervised tool for ordering high-dimensional data in such a way that similar input items are grouped spatially close to one another. It consists of a layer of input units that receive the input patterns and propagate them as they are to a set of output units. 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 of the same dimensionality as the input space.

During each learning iteration, the unit c having the highest activity level with respect to a randomly selected input pattern $x = [\xi_1\ \xi_2, ..., \xi_n]^T$ 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 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.

In order to provide a convenient interface to larger document collections, a hierarchical representation is preferable. We thus use the Growing Hierarchical SOM (GHSOM) [4, 20], an extension of the SOM that grows a hierarchy of individual maps and adapts their size according to the input space. The GHSOM has a hierarchical structure of multiple layers where each layer consists of several independent growing self-organising maps.

Starting from a top-level map, each map grows in size in order to represent a collection of data at a certain level of detail. In particular, starting with an initial 2x2 SOM, rows and columns of units are added to those areas of the map where input discrimination is rather poor. After a certain improvement of the granularity of data representation is reached, the units are analysed to see whether they represent the data at a specific minimum level of granularity. Those units that have too diverse input data mapped onto them are expanded to form a new small SOM at a subsequent layer, where the respective data shall be represented in more detail. Again, these newly created maps grow during their training process as described above. Units representing an already rather homogeneous set of data, on the other hand, will not require any further expansion at subsequent layers. The resulting GHSOM thus is fully adaptive to reflect, by its very architecture, the hierarchical structure inherent in the data, allocating more space for the representation of inhomogeneous areas in the input space.

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 [14, 15]), 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 relative importance of every feature and the mean quantisation error of that feature in the weight vector value of that unit, resembling the mean and the variance, respectively. 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 (resembling the variance) for that particular vector element 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 usage 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 (resembling the mean) such that the corresponding feature may be used for labelling.

3. Ontologies and vector representation

Using traditional binary or *TFxIDF* vector representation of documents the results may not be sufficiently accurate for a dynamic legal commentary. Length and variety of legal documents lead to this tentative conclusion. Some sort of an ontology may provide a solution for more efficient vector representation.

The concept of ontology is defined as an explicit specification of the conceptualisation of the legal domain. Valente [23] has proposed ontologies as the missing link between legal theory and AI & law. Valente's decomposition of legal functions leads to six categories of primitive legal knowledge. The frame based ontology of van Kralingen and Visser [24, 25, 26] distinguishes three classes of entities: norms, acts and concept descriptions. Frame structures list all attributes relevant for the entity. Both ontologies focus on knowledge engineering. The aspects of knowledge sharing and knowledge reuse as typical problems of knowledge engineering receive high priority.

Ontologies have not yet been widely used in the legal domain. Recent applications are the projects POWER [5] and CLIME/MILE [2] that have verified the potential but also the quite difficult knowledge engineering process.

These well developed ontologies are not appropriate for a first use of ontologies for vector representation. A quite useful first step may be a simple ontology like a thesarus representing the most important concepts of the domain.

Two approaches may be distinguished: feature vector creation based on ontologies and adaptation of *TFxIDF* weight vectors.

Feature vectors have quite often been used in legal document representation. In the FLEXICON project [21], a term extraction module recognising concepts, case citations, statute citations and fact phrases lead to a structured document profile. This profile is transformed into a weighted vector. A similarity computation between likewise-structured queries and documents is performed using the Cosine formula [16]. Another approach has been developed within the KONTERM project [11]. The various documents are represented as feature vectors of the form $x=\{t_1, ..., t_m, c_1, ..., c_n, m_1, ..., m_0\}$. The t_i represent terms extracted from the fulltext of the document, the c_i are the context-sensitive rules, and the m_i represent the meta rules associated with the document. Like in the FLEXICON project, concepts are recognised by matching a list but also by applying some heuristic rules. Linguistic templates are found by context-sensitive rules. The wording of rules is facilitated allowing also probabilistic expressions. Meta rules represent a concept that must be defined as a combination of rules occurring in the same document or section of a document. The result is a weighted vector giving more importance to linguistic templates and meta rules. In both approaches, evaluation results were very promising. The main problem remains the development of the knowledge base of lexica and rules. A major advantage may be the screening of huge document collections for particular problems showing connections between documents that are not detected by manual research.

The second approach on vector weights refines the *TFxIDF* vector representation. Previous research has shown that a good capture of all important concepts in a legal document based only on statistics seems impossible. The basic idea is that a small ontology will be integrated into the vector representation. The ontology describes properly the main concepts and facts of a legal domain. This information is used to give the vector components of important terms a significant higher value. A hierarchy can be also properly represented using different weighting values.

4. Data Enrichment: Refinement of Vector Weights

The input for the refinement of vector weights are the automatically generated vectors with *TFxIDF* weights. The vector components are represented in a template vector file and the *TFxIDF*-file which contains the values related to the words.

Our approach modifies the vector components in such a way that legal vocabulary is more properly represented. Two tools are offered to the user: replacement of words (Replace), and changing the weights of words (Weight).

Replace. This tool simply replaces one word with another and accumulates the weight vector values of these words. The most evident application are synonyms but also insufficient automatic stemming. The replacement is done using a replacement file provided by the user. The relevant values in the template vector file are changed by adding the term frequency of the source word to the term frequency of the destination word. In the same way, we change the vectors in the *TFxIDF*-file.

Weight. The second tool increases the weight of important words. Here, the user must provide a file with a multiplicator for the *TFxIDF*-value of the chosen words (i.e. the thesaurus). Using this input, the programme changes automatically the *TFxIDF* file by multiplying the values of the vector components. In our tests, we have used a small ontology of European law, but also cyberspace law.

Create replace file: Assisting this procedure, some help is provided with the WordSOM tool. The output of a WordSOM is word clusters, rather than document clusters,

by grouping words that are used in similar sets of documents. This results in words with same meaning or similar function being grouped together. The WordSOM can be described as a different way to find synonyms. This output can be used for the replace function and can be quite easily automated if required.

5. Experiments

The test environment for our approach comprises a text collections of the most important documents of Cyberlaw in English (253 documents) in HTML format.

No preparation of the documents for further analysis is necessary. We represent each document by a vector with a length of 6669 words using a *TFxIDF* weighting scheme [16]. Four test cycles were performed: standard, with segmentation (4000 byte blocks), with segmentation (2000 byte blocks), with weighting (double value). The vectors were improved with a very small ontology of 204 words in test cycle 4. The relevant weight of the vector components was doubled to represent the much higher importance of these descriptors.

In general, a significant improvement of classification and labelling can be detected. As an example for the automatic approach, we present the map of test 1 on data protection (1,1) (2,4)[1] in more detail:

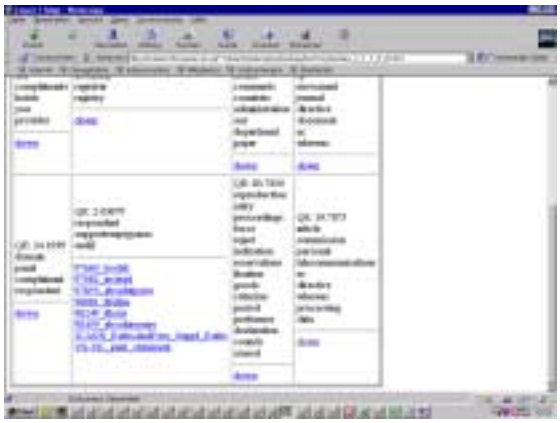

Figure 1. Cyberlaw, test 1, layer 1

The map of layer 1 (*Figure* 1) has put the documents on data protection together with those of copyright and telecommunications in cluster (2,4). The labels are not very precise:

Article, commission, personal, telecommunications, ec, directive, whereas, processing, data

Besides the stopswords *article* and *whereas* the remaining descriptors contain the broad terms of *commission, ec* and *directive* that are not very indicative of the documents.

The second layer of unit (2,4) (*Figure* 2) with 3 rows and 2 colums shows very clearly the various topics of the broad cluster:

(1,1) Mobil communications	(1,2) Copyright
(2,1) Telecommunications, data protection	(2,2) Telecommunications
(3,1) Data protection, e-commerce	(3,2) Data protection

Figure 2. Cyberlaw, test 1, unit (2,4), topics

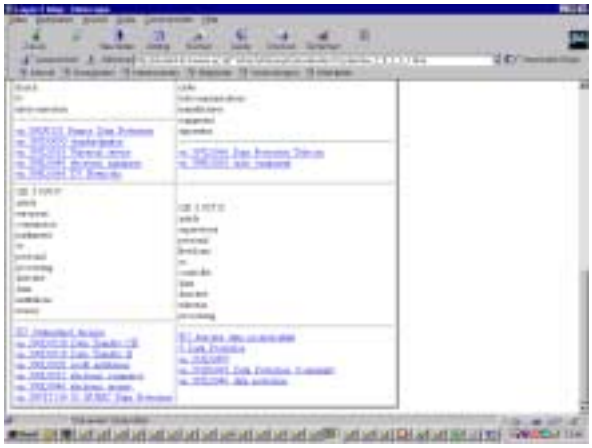

Figure 3. Cyberlaw, test 1, layer 2, unit (2,4)

As mentioned above, stopwords *(oj, article, whereas)* and broad descriptors *(commission, ec, directive, french, institutions, parliament)* form a significant part of the labels. Aprt from that, the descriptors are quite helpful. It should be mentioned that the clusters (2,1) und (2,2) contain documents on e-commerce law that is properly indicated by the labels.

Summing up, the helpful clusters and labels are blurred by some documents and labels not fitting very well into the main topic.

The second example consists of test cycle 4 with the modification of the vector files. The very small ontology of 204 words has significantly improved the clustering and

[1] We will use the notation *[x,y]* to refer to the unit in row *x* and column *y*. For the GHSOM we list the path through the hierarchy, i.e.(a,b)(c,d) describing unit (c,d) on the map originating from unit (a,b) in the previous layer of the hierarchy.

labels. The comparison remains difficult because clusters and labels have changed. For reasons of convienience, we represent again the example of data protection (*Figure* 4):

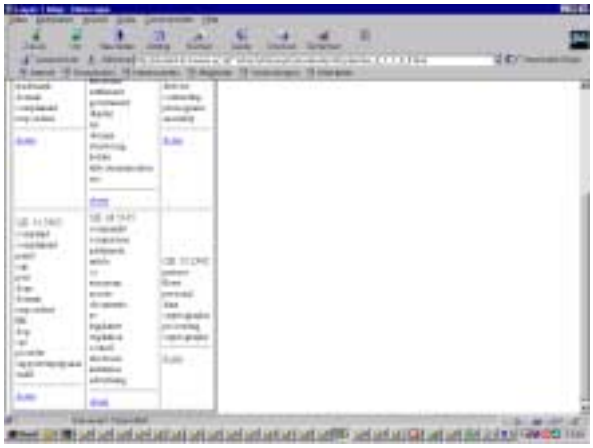

Figure 4. Cyberlaw, test 4, layer 1

All documents on data protection are now found in cluster (2,3). The labels are more precise:

Privacy, flows, personal, data, cryptographic, processing, cryptography

Quite evident seems the missing of stop words. The labels itsself are very indicative of the topic on data protection and cryptography. Most labels are descriptors of the small ontology.

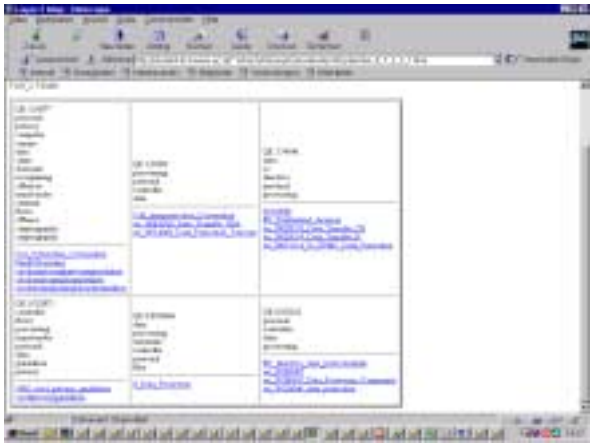

Figure 5. Cyberlaw, test 4, layer 2, unit (2,3)

Figure 5 shows the layer 2 of unit (2,3). The various clusters describe the main topics of international data protection (*Figure* 6):

(1,1) Cybercrime, crypthography, transborder data flow	(1,2) Transborder data flow, telecom data protection	(1,3) EC data protection, transborder data flow
(2,1) OECD guide-lines on trans-border data flow	(2,2) Swedish data protection law	(2,1) EC data protection law

Figure 6. Cyberlaw, test 4, layer 2, unit (2,3)

The labels clearly show the main content. The labels of the clusters (2,3) (1,1) und (2,3) (2,1) give more information on the subtopics:

Cluster (2,3) (1,1)
Labels: personal, privacy, computer, europe, data, crime, domestic, recognising, offences, transborder, criminal, flows, offence, cryptographic, cryptography
Documents: Cybercrime Convention (2 versions), OECD Guidelines for Cryptography Policy, OCED Recommendation for Crypography Policy, OECD Declaration on Transborder Data Flows

Cluster (2,3) (2,1)
Labels: controller, flows, processing, transborder, personal, data, guidelines, privacy
Document: OECD Privacy Guidelines (2 versions)

Summing up, the Data Enrichment tool provides a useful framework for improvement of vector representations. However, it should be noted that the description becomes more subjective. Clusters and labels are more focused on the small thesaurus but less on the text of the documents.

6. Conclusions and future work

Our new Data Enrichment tool provides a solution for the difficult vector representation of legal documents. With quite small intellectual input the vectors can be significantly improved leading to better clustes and labels. Our experiments are only indicative and will be deepened in the near future. The main advantage of such semiautomatic analysis may be the creation of appropriate vector values that may be used for similar collections.

Acknowledgements

This research was supported by the Jubiliäumsfonds der Oesterreichischen Nationalbank, Vienna, research project no. 6888.

References

[1] D. Austin, A. Mobray and Ph. Chung, "Scalability of Web Resources for Law: AustLII's Technical Roadmap. Past, Present and Future", *Journal of Information, Law and Technology*, 2000.

[2] A. Boer, “MILE Asessment: Turning Legal Information into Legal Advice”, *Proc. Int. Workshop on Database and Expert Systems Applications*, Munich, Germany, 2001.

[3] H. Chen, A. L. Houston, R. R. Sewell, and B. R. Schatz, ”Semantic search and semantic categorization”, *Proc. of the Int. ACM SIGIR Conf. on R&D in Information Retrieval (SIGIR’97)*, Philadelphia, PA, 1997.

[4] M. Dittenbach, D. Merkl and A. Rauber, "The growing hierarchical self-organizing map", *Proc. IEEE Int'l Joint Conference on Neural Networks (IJCNN 2000)*, Como, Italy, 2000.

[5] Tom M. van Engers et al.: “POWER : Using UML/OCL for Modeling Legislation – an application report”, *Proc. 8th Int. Conf. on Artificial Intelligence and Law*, St. Louis, MO, 2001.

[6] T. Kohonen, ”Self-organized formation of topologically correct feature maps”, *Biological Cybernetics*, Vol. 43, 1982.

[7] T. Kohonen, *Self-organizing maps*, Springer-Verlag, Berlin, 1995.

[8] T. Kohonen, ”Self-organization of very large document collections: State of the art”, *Proc. of the Int. Conf. on Artificial Neural Networks (ICANN’98)*, Skövde, Sweden, 1998.

[9] K. Lagus, T. Honkela, S. Kaski and T. Kohonen, ”Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration”, *Proc 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, Portland, OR, 1996.

[10] D. Merkl, ”Exploration of Text Collections with Hierarchical Feature Maps”, *Proc. Int. ACM Conf. on R&D in Information Retrieval (SIGIR’97)*, Philadelphia, PA, 1997.

[11] 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.

[12] D. Merkl, ”Text classification with self-organizing maps: Some lessons learned”, *Neurocomputing*, Vol. 21, No. 1-3, 1998.

[13] A. Rauber and D. Merkl, ”Creating an Order in Distributed Digital Libraries by Integrating Independent Self-Organizing Maps”, *Proc. Int. Conf. on Artificial Neural Networks (ICANN'98)*, Skövde, Sweden, 1998.

[14] A. Rauber and D. Merkl, ”Automatic Labeling of Self-Organizing Maps: Making a Treasure-Map Reveal its Secrets”, *Proc. Pacific Asia Conf. on Knowledge Discovery and Data Mining*, Beijing, China, 1999.

[15] 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.

[16] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA, 1989.

[17] 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.

[18] E. Schweighofer, ”The Revolution in Legal Information Retrieval, or: The Empire Strikes Back”, *Proc. Conf. The Law in the Information Society*, Florence, Italy, 1998.

[19] E. Schweighofer, *Legal Knowledge Representation*, Kluwer Law International, The Hague, The Netherlands, 1999.

[20] E. Schweighofer, A. Rauber and M. Dittenbach, "Automatic Text Representation, Classification and Labeling in European Law”, *Proc. 8th Int. Conf. on Artificial Intelligence and Law*, St. Louis, MO, 2001.

[21] J. Smith et al., "Artificial Intelligence and Legal Discourse: The Flexlaw Legal Text Management System, *AI & Law*, Vol. 3, 1995.

[22] H. Turtle, ”Text Retrieval in the Legal World”, *AI & Law*, Vol. 3, 1995.

[23] A. Valente and J. Breuker, “An Architecture for Modelling Legal Information”, *Proc. Fifth International Conference on Artificial Intelligence and Law*, Washington, DC, 1995.

[24] R. W. Van Kralingen, *Frame-based Conceptual Models of Statute Law*, 1995.

[25] P. Visser and T. Bench-Capon, “On the Reusability of Ontologies in Knowledge-System Design.” *Proceedings of the Seventh Int. Workshop on Database and Expert Systems Applications*, Zurich, Switzerland, 1996.

[26] P.R.S. Visser, *Knowledge Specification for Multiple Legal Tasks, A Case Study of the Interaction Problem in the Legal Domain*, 1995.

[27] P. Willet, ”Recend trends in hierarchic document clustering: A critical review”, *Information Processing & Management*, Vol. 34, 1988.