TOWARDS AUTOMATIC CONTENT-BASED ORGANIZATION OF MULTILINGUAL DIGITAL LIBRARIES: AN ENGLISH, FRENCH, AND GERMAN VIEW OF THE RUSSIAN INFORMATION AGENCY NOVOSTI NEWS

Andreas Rauber[1], Michael Dittenbach[2], Dieter Merkl[1]

[1] Department of Software Technology
Vienna University of Technology
Favoritenstr. 9-11 / 188, A-1040 Vienna, Austria
[2] E-Commerce Competence Center - EC3
Siebensterng. 21/3, A-1070 Vienna, Austria
e-mail: *{andi, mbach, dieter}@ifs.tuwien.ac.at*

Abstract

In this paper we present the application of the *SOMLib* digital library system to a multilingual document corpus from the Russian Information Agency Novosti. News articles in Russian, English, and German are automatically organized into separate topic hierarchies using a novel unsupervised neural network, namely the *Growing Hierarchical Self-Organizing Map*. Furthermore, machine translation is used to create a coherent corpus in a single target language. In spite of the "noise" introduced by the automatic translation a consistent topical structuring of the integrated document collection can be created by the neural network. This facilitates straightforward browsing and exploration of multilingual document collections in a given target language.

Keywords: Document Clustering, Neural Networks, Growing Hierarchical Self-Organizing Map, GHSOM, Machine Translation

1 Introduction

With the increasing amount of textual information being available electronically, methods for organizing these vast amounts of information to support browsing and exploration of topic spaces gained importance. These means of access complement more traditional search and retrieval approaches, as they were found to not fully satisfy all needs of the users. While full-text search provided tremendous benefits in terms of information selection, it still requires at least basic knowledge of query formulation techniques, as well as a rather precise idea of what kind of information one is looking for. While this may satisfy a significant proportion of users needs, explorative information search has also been found to merit closer consideration, allowing users to analyze and browse an unknown document repository, finding out which topics are covered by a collection, to what extent certain topics are dealt with, and in which way various concepts are related. This gave rise to considerable research efforts aiming at the provision of methods for exploring information spaces.

Amongst the wealth of techniques employed particularly successful we find the *Self-Organizing Map* (*SOM*) [Koh95], an unsupervised neural network providing a topology-preserving mapping from the high-dimensional document space onto a two-dimensional map space, where documents on similar topics are located in neighboring regions [CSO96,KKL$^+$00,MR00]. This spatial organization provides suitable interface to large document collections as it resembles the way information has been organized and used in conventional, i.e. printed, form, with topic-based organization being the dominant form of public library organization. (At least this applies to most public access libraries, as well as most private libraries and book stores, with exceptions to be found in large-scale libraries only accessible via card catalogs due to mere organizational necessities.) Out of these reasons this approach was also chosen as the core module of our *SOMLib* digital library system [RM99].[1] This system allows users to explore unknown document collections, providing an overview of available topics organized in an intuitively comprehensible way, both in terms of large library maps, as well as, with advanced neural network models such as the *Growing Hierarchical Self-Organizing Map* (*GHSOM*) [DMR00], by providing automatically detected topic hierarchies [RDM00].

Yet, with the Internet providing access to a vast range of information repositories world wide we would like to support users in using these resources even if they are available in foreign languages. Even if the user does not understand a given foreign language, being able to locate relevant information provides a considerable benefit. Cross-language information retrieval (CLIR) addresses this issue by allowing users to formulate queries in their native language and retrieving documents from foreign language document repositories, using a variety of machine translation (MT) approaches. After the retrieval step, the quality of state-of-the-art machine translation techniques is sufficient enough to allow a user to judge the relevance of a retrieved document and to understand the gist of it. While CLIR techniques allow the retrieval of foreign language documents, hardly any means of browsing and exploring multilingual document collections are available.

In this paper we present the application of the *SOMLib* digital library system to multilingual document repositories. More specifically, we demonstrate the language-independence of our digital library system by using it to separately organize collections of news articles from the Russian Information Agency Novosti (RIAN).[2] In spite of being a multilingual information repository, the archive does not constitute a parallel corpus, i.e. not all articles are available in all different languages, making it a challenging application scenario for multilingual information exploration. The documents are parsed to extract content-bearing features, with the resulting document representations being used to train a *GHSOM*. In the resulting map architecture, the articles are grouped into various topical branches, with topics in each branch being arranged on two-dimensional maps similar to conventional library organization, and keywords serving as labels for the respective topics.

We then use standard machine translation techniques to create a monolingual text corpus in the target language of the user. In spite of the inacurracies, i.e. “noise”, introduced by MT, the neural network is capable of organizing both original as well as translated documents as faithfully as possible into topical hierarchies. This facilitates straight-forward browsing and exploration of multilingual document collections in a given target language.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work in the field of document clustering, cross-language retrieval, and machine translation. This is followed by a brief introduction into the various modules of the *SOMLib* digital library system in Section 3. Some considerations concerning the organization of multilingual document archives are addressed in Section 4. Section 5 reports on our experimental results using a collection of RIAN news articles, detailing the capabilities of the presented approach as well as identifying some pitfalls with respect to machine translation. The paper is rounded off by providing some conclusions as well as an outlook on future work in Section 6.

[1] The SOMLib project homepage is available at `http://www.ifs.tuwien.ac.at/~andi/somlib`

[2] http://www.rian.ru

2 Related Work

Document clustering has been identified as one of the key issues in digital library exploration and has thus been addressed in a number of projects like the BEADS system [CC92] using multidimensional scaling or the BiblioMapper [Son98] using a hierarchical clustering algorithm. A technique classifying documents in a hierarchical topic structure is presented in [KS97], the application of the multiple cause mixture model for text categorization using the Reuters document collection is reported in [SHS96]. One of the most prominent systems for exploring large document spaces is Scatter/Gather [HP96] employing clustering of retrieved documents in combination with a form of relevance feedback by allowing users to select relevant sub-clusters. The *Self-Organizing Map* and related models have been used successfully in a number of systems for the classification and representation of document collections [CSO96,KKL$^+$00,LSG91,MR00,RM99]. A variation of this approach using hierarchically organized *SOMs* is described in [MR98] using data form the CIA world factbook. Yet, most of these systems so far address only monolingual collections.

Cross-language information retrieval (CLIR) describes the task of finding documents written in one language with queries formulated in a different language. A good overview on the subject of CLIR and machine translation is provided in Oard's CLIR bibliography [Oar97]. Basically, there are two different approaches, namely translation of the documents and translation of the queries, with most approaches following the query-translation principle due to memory efficiency reasons as the source documents are stored only once. Query terms are translated using machine translation, to retrieve documents in a language other than the query itself. Yet, this usually has the disadvantage of word sense disambiguities caused by the lack of context a translation engine could use [HG96]. Especially short queries, which are most commonly issued by users of search engines, suffer from bad automatic translation. Dictionary-based methods combined with query expansion techniques [BC97] or structured translation [SO00] try to reduce the ambiguity of the translated query and therefore, to increase retrieval performance. Another technique is using parallel text corpora to select the most appropriate query terms from the set of possible translations.

A different, yet essential, approach to access a multilingual document collection is interactive exploration. However, only little research work has been reported in this field so far. Some initial experiments comparing *SOM* clustering performance on a small parallel Chinese - English corpus are reported in [LY00], supporting the general feasibility of multilingual corpora analysis by comparing results in parallel collections, with analogous experiments in the legal domain using English, German and French corpora being reported in [RSM00].

3 The SOMLib Way of Document Organization

3.1 Document content representation

The core metaphor of the *SOMLib* digital library system is a map of documents organized by topic using an unsupervised neural network, namely the *Self-Organizing Map* (*SOM*) [Koh82,Koh95] and variants thereof, specifically the *Growing Hierarchical Self-Organizing Map* (*GHSOM*) [DMR00]. In order to utilize the *SOM* for organizing documents by their topic, a vector-based description of the content of the documents needs to be created. While manually or semi-automatically extracted content descriptors may be used, research results have shown that a rather simple word frequency based description is sufficient to provide the necessary information in a very stable way. For this word frequency based representation a vector structure is created consisting of all words appearing in the document collection. This list of words is usually cleaned from so-called stop words, i.e. words that do not contribute to content representation and topic discrimination between documents. Again, while manually crafted stop word lists may be used, simple statistics allow the removal of most stop words in a very convenient language- and subject-independent way. On the one hand, words appearing in too many documents, e.g. in more than half of all documents, can be removed without the risk of loosing content information, as the content conveyed by these words is too general. On the other hand, words appearing in less than a minimum number of, say, 5 to 10 documents, can be omitted for content-based classification, as the resulting subtopic granularity would be too small to form a topic cluster of its own.

The documents are described within the resulting feature space of commonly between 2,000 and 15,000 dimensions, i.e. distinct terms, by the words they are made up of. While a basic binary indexing may be used to describe the content of a document by simply stating whether a word appears in the document or not, more sophisticated schemes, such as *tf x idf*, i.e. term frequency times inverse document frequency [Sal89], provide a better content representation. This weighting scheme assigns higher values to terms that appear frequently within a document, i.e. have a high term frequency, yet rarely within the complete collection, i.e. have a low document frequency. Usually, the document vectors are normalized to unit length to make up for length differences of the various documents.

3.2 A self-organizing map of document collections

The resulting vector representations are fed into a standard *Self-Organizing Map* for cluster analysis. The *Self-Organizing Map* (SOM) [Koh95] is an unsupervised neural network model that provides a topology-preserving mapping from a high-dimensional input space to a usually 2-dimensional output space. It consists of a grid of units with *n*-dimensional weight vectors. During the training process input data are presented to the map in random order. An activation function based on some metric, such as the commonly used Euclidean distance between the presented input vector and a unit's weight vector, is used to determine the winning unit. Next, the weight vectors of the winner and of neighboring units are modified to represent the presented input signal more closely, following a time-decreasing learning rate.

Text documents can be thought of topical clusters in the high-dimensional feature space spanned by the individual words in the documents. A trained *SOM* thus represents a topological ordering of the documents, meaning that documents on similar topics are located close to each other on the 2-dimensional map. This is comparable to what one can expect from a conventional library, where we also find the various books ordered by some contents-based criteria. In the simplest form, a document collection may then be represented as a rectangular table with similar documents being mapped onto the same cells. Using this model, users find a document collection to be automatically structured by content in a way similar to how documents are organized into shelves in conventional libraries.

While *SOM* based architectures found wide appreciation in the field of text classification, their application had been limited by the fact that the topics of the various cluster were not evident from the resulting mapping. In order to find out which topics are covered in certain areas of the map, the actual articles had to be read to find descriptive keywords for a cluster. To counter this problem, we developed the *LabelSOM* method, which analyses the trained *SOM* to automatically extract a set of attributes, i.e. keywords, that are most descriptive for a unit [Rau99]. Basically, the attributes showing a low quantization error value and a high weight vector value, comparable to a low variance and a high mean among all input vectors mapped onto a specific unit, are selected as labels. Thus, the various units are characterized by keywords describing the topics of the documents mapped onto them.

3.3 Detecting topic hierarchies using the GHSOM model

While the *SOM* has proven to be a very suitable tool for detecting structure in high-dimensional data and organizing it accordingly on a two-dimensional output space, some shortcomings have to be mentioned. These include its inability to capture the inherent hierarchical structure of data. Furthermore, the size of the map has to be determined in advance ignoring the characteristics of an (unknown) data distribution. These drawbacks have been addressed separately in several modified architectures of the *SOM* [BM93,Fri95,Mii90]. However, none of these approaches provides an architecture which fully adapts itself to the characteristics of the input data. To overcome the limitations of both fix-sized and non-hierarchically adaptive architectures we developed the *GHSOM* [DMR00], which dynamically fits its multi-layered architecture according to the structure of the data.

The *GHSOM* has a hierarchical structure of multiple layers where each layer consists of several independent growing self-organizing maps. Starting from a top-level map, each map, similar to the *Growing Grid* model [Fri95], grows in size in order to represent a collection of data at a certain level of detail. In particular, starting with an initial *2 x 2 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 analyzed 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. The growth process of these new maps continues again in a *Growing Grid* like fashion. 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 map space for the representation of inhomogeneous areas in the input space.

3.4 A metaphor-graphics based visualization of document archives

Although the spatial organization of documents on the 2-dimensional map in combination with the automatically extracted concept labels supports orientation in and understanding of an unknown document repository, much information on the documents cannot be told from the resulting representation. Information like the size of the underlying document, its type, the date it was created, when it was accessed for the last time and how often it has been accessed at all, its language etc. is not provided in an intuitively interpretable way. Rather, users are required to read and abstract from textual descriptions, inferring the amount or recentness of information provided by a given document by comparing size and date information.

We thus developed the *libViewer*, a metaphor-graphics based interface to a digital library [RB00]. Documents are no longer represented as textual listings, but as graphical objects of different *representation types* such as binders, papers, hardcover books, paperbacks etc, with further metadata information being conveyed by additional metaphors such as *spine width, logos, well-thumbed spines*, different degrees of *dustiness,* *highlighting glares, position in the shelf* and others. Based on these metaphors we can define a set of mappings of metadata attributes to be visualized, allowing the easy understanding of documents.

4 Supporting Exploration of Multilingual Document Collections

As the basic principles of *SOM* based document clustering are language independent, the *SOMLib* digital library system can be applied to collections in any language, provided that words as primary tokens can be identified. (This may require special preprocessing steps for languages such as Chinese, where word boundaries are not eminent from the texts.) Thus multilingual document collections can be clustered as split collections in different languages. This results in separate maps for e.g. English and German documents, thus actually forming separate monolingual collections. On the other hand, a multilingual collection may be organized on one single map. Still, on such a map documents will be primarily organized by language, as the words in different languages will form the primary separators in the feature space. This will only be followed by topical sub-clusters within each language set. Yet, topical organization of different language documents within one single map in a users target language would allow straight-forward exploration of a multilingual collection.

While machine translation may still be far from perfect regarding the high standards expected and provided by human translation, it usually suffices for human readers to allow topic detection of foreign language documents. Where humans can abstract and make up for the imperfections provided by machine translation programs, it still remains a fascinating challenge to analyze the capabilities of MT with respect to topic preservation and automatic topic recognition. Although the translations provided by MT programs will be imprecise and noisy, and sometimes even utterly wrong, we may expect the neural network to make up for these imprecisenesses. This is particularly due to the high-dimensional feature spaces encountered in text classification.

Generally, several weaknesses of current MT systems may be considered. The most obvious to human readers is their weakness in providing grammatically nice or at least correct sentences, with the most noticeable weak point being word order. As the order of words does not influence the performance of the bag-of-words approach used in the presented document clustering procedure, there is no effect on the performance of the system. Some problems are caused by the way proper nouns are treated. On the one hand, they may have a special, yet unintended meaning in one of the target languages, such as for example, the name of the Russian space station *“Mir”* meaning *“World”* or *“Peace”*. However, *“Mir”* also happens to be the German word equaling the English term *“me”*, with the space station being translated into *space station “Me”* from German sources, whereas it ends up as *space station “World”* following Russian-English translations. A similar situation is encountered with proper names being spelled differently. As there are usually not available in dictionaries, they will usually end up untranslated, providing further mismatch. For example, we find the country of *Vietnam* also to be spelled as two separate words *Viet Nam* after translation into English. Yet the correct translation of these key terms may be crucial for correct topic identification. Other difficulties result from synonyms, such as the *space station* also being an *orbital station*. While synonymy causes only little disturbances in monolingual collections, as the content is usually carried by a large number of words, this problem becomes more prominent in multilingual collections with subsequent translations. This is because consistent usage of specific synonyms might set translated texts apart from original versions, where broader vocabulary is used. The same applies for utterly “wrong” translations, where words are mistaken for specific meanings and result in completely different translations. While none of these should cause problems on their own, as they also appear in monolingual collections in the form of proper nouns, synonyms and misspellings, the consistency with which they might appear in translations poses a challenge to the clustering approach. However, we may expect the *SOM* to make up for the noise introduced, specifically due to the high-dimensional feature vectors involved.

5 Experiments

5.1 Data Sources and Setup

For the following experiments we use a collection of news articles provided by the Russian Information Agency Novosti (RIAN) covering the period of March 1 - 14, 2001. Articles are made available in a variety of different languages, namely Russian, English, French, German, and Arabian. For the experiments provided hereafter we selected the Russian, English, French, and German versions. These collections do not form parallel corpora, i.e. not all documents are available in all languages, with the most extensive set being available in Russian (1387 articles), followed by the English collection (973 documents), French (712 documents), and the German article set (485 documents) for the two-week period.

The documents were parsed to extract feature vectors. During the parsing process we removed words that appeared in more than 10% or in less than 0.7% of all documents in the respective collection. This resulted in feature vector dimensionalities of 2869, 2163, 2271, and 1979 for the Russian, English, French, and German set, respectively. These document vectors were further used to train separate *GHSOM* architectures to analyze the topical hierarchies present in each document set. The topical hierarchies detected in the separate document collections are discussed in Section 5.2.

To provide a single coherent interface for browsing and exploring the multilingual collection, the various documents were translated into a single target language, i.e. English. We refrained from using domain specific machine translation techniques. Rather, all documents were translated by the publicly available *Babelfish* translator provided by AltaVista.[3] Since *Babelfish* is not capable of translating latin encoded Russian documents into English, and since we did not have access to another Russian-to-English translator accessible via scripting interfaces, we had to omit the original Russian version articles from the automatic translation process. The combined set of original English documents together with the French and German articles translated into English comprised 2170 files. These were again parsed using the same parameter settings as for the individual collections, resulting in a feature space of 1959 dimensions. This integrated document collection was again used to train a *GHSOM* hierarchy of newspapers, which is described in more detail in Section 5.3.

5.2 Russian, English, French, and German Novosti News

Due to space considerations we omitted figures of the German, Russian, and French hierarchies but they are provided online along with the other hierarchies described in this paper.[4]

In Figure 1(a), the first-layer map of the *GHSOM* trained with the English set of articles is depicted. This map provides a rough overview of the main topics present in the document collection. In the top-left corner of the

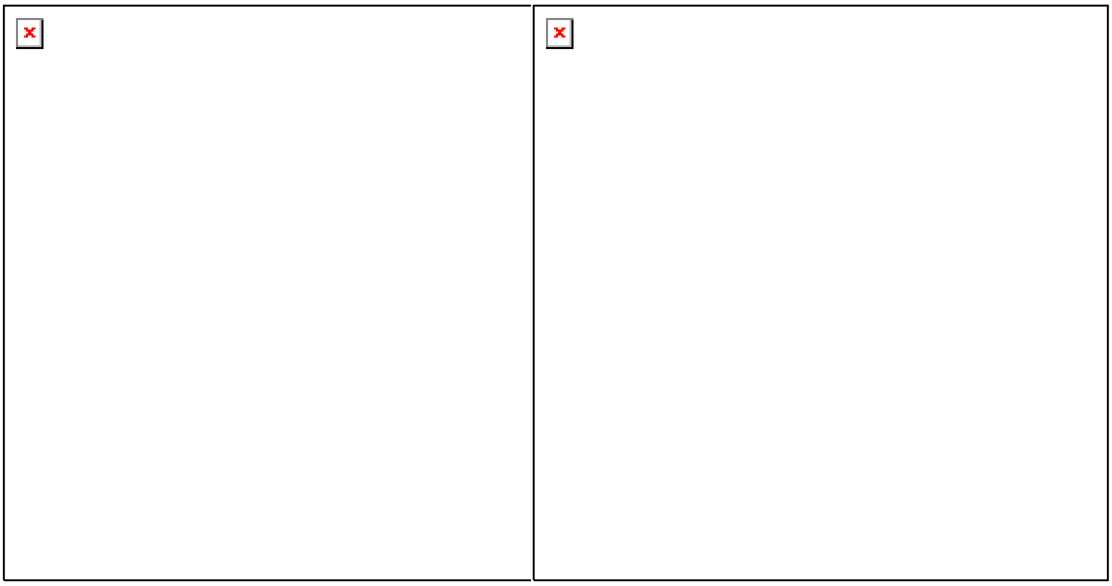
[3] http://babelfish.altavista.com

[4] http://www.ifs.tuwien.ac.at/~andi/somlib/experiments_rian.html

map on unit (1/1)[5] we find documents covering various aspects of Russian - Vietnamese relations. This unit has been automatically labeled with, for example, *vietnam*, *trade*, *hanoi*, and *export*.

Other topics on this map can be identified as discussions about military cooperations between Russia and Iran on unit (2/1), the Kosovo Crisis (3/1), Ukrainian issues (1/2), articles on Afghanistan and Tajikstan (2/2), or articles on the situation in Israel and Palestine on unit (2/3). Russian politics is located on unit (3/4). On unit (2/4) we find articles on the International Space Station (ISS) and the Russian Space Station MIR. Following the link on this unit to the according second-layer map (cf. Figure 1(b)), a more detailed representation of these articles can be found. This map evolved to a *3 x 3* map during training, showing articles about the IIS on the units in the top row, whereas articles on the deorbiting of the MIR are located in the bottom row. Showing the topology-preserving feature of the *SOM*, on unit (1/2) in the middle row, articles about international cooperations between Russia and other countries concerning space exploration can be found.

Analogous, in the German hierarchy, the main topics can be found again, such as the Vietnam topic on unit (1/1), articles about the situation in Kosovo (1/3), or the Iran subject on unit (3/4). The articles concerning the MIR space station are located in the top row of a second-layer map below unit (1/2) on the top-layer map. The French and Russian hierarchies are quite similar regarding their topical structure. Due to the different numbers of articles present in the respective languages, the cluster structure of topics found on the maps may vary between the different hierarchies.



(a) First-Layer Map: *3 x 4* units; Main topics of the English Novosti news.

(b) Second-Layer Map: *3 x 3* units; International Space Station (ISS) and MIR.

Figure 1: English Novosti news: The top-layer map and a second-layer map covering articles about space flight.

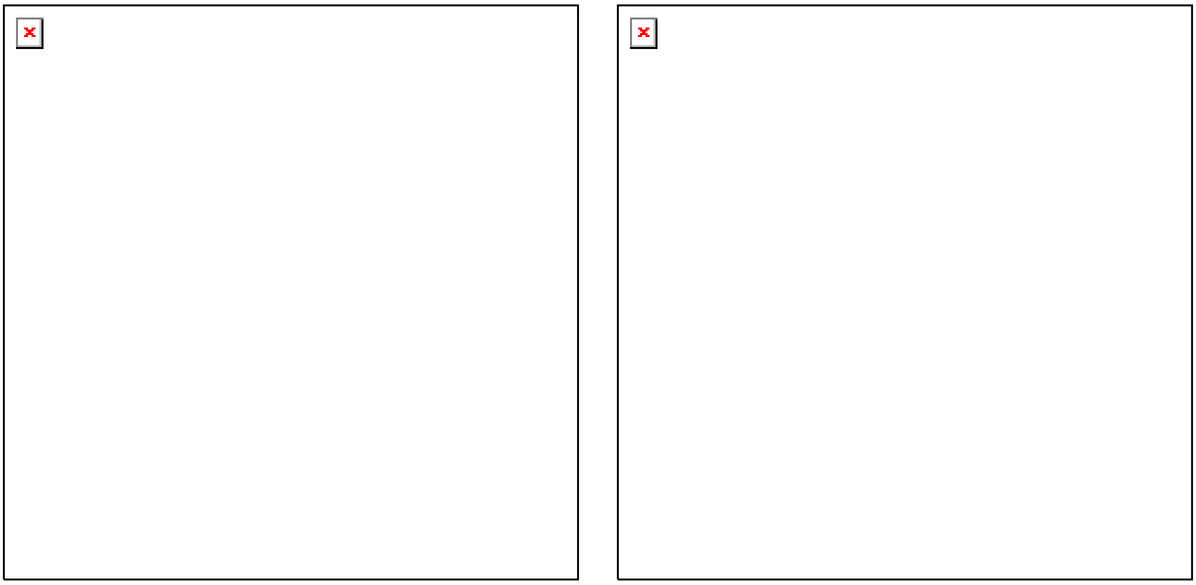
5.3 Integrating Multilingual Document Collections

Figure 2(a) depicts the top-layer map of the *GHSOM* trained with the collection consisting of the English articles as well as the French and German articles translated to English. This map has grown to a size of *3 x 4* units during training. Here, we find the Vietnam topic located on unit (3/1), articles about the Ukraine on unit (2/1), Chechnia (1/2), the Caspian Sea and Kazhakstan (3/2). The Kosovo subject can be found on unit (1/3) and

[5] We refer to a unit located in column *x* and row *y* as *(x/y)*, starting with (1/1) in the upper-left corner.

articles about Iran on unit (1/4). Interestingly, unit (3/3) represents articles about the Foot and Mouth Disease, currently being a big issue in Europe, and the hard and frosty winter in Mongolia where a lot of animals, important to the Mongolian people, died because of the cold temperatures.

Again, we find the articles on ISS and MIR represented by unit (2/2), depicted in Figure 2(b). Despite the size, the according second-layer map of size *2 x 4*, has a similar structure compared to the English-only map (see 1(b)) covering the articles about space flight. MIR-related documents are located in the lower half, whereas articles about the IIS can be found in the upper half of the map.



(a) First-Layer Map: *3 x 4* units; Main topics of the English, French, and German Novosti news.

(b) Second-Layer Map: *2 x 4* units; International Space Station (ISS) and MIR.

Figure 2: English, German, and French Novosti news: The top-layer map and a second-layer map covering articles about space flight.

6 Conclusions

We showed that content-based clustering using the *Growing Hierarchical Self-Organizing Map* works with documents in different languages due to a language-independent method for vector-space creation. Furthermore, translating documents, available in different languages to a common target language using automatic machine translation systems, is sufficient enough to get a satisfying content-based organization employing this unsupervised neural network model. However, it has to be noted that the quality of the document organization can be improved, because some restrictions of most of nowadays MT systems remain.

First, the translation of proper names is sometimes difficult, either due to different notations in different languages (e.g. *Vietnam* - *Viet Nam*), or because a proper name in one language has a certain meaning in another (e.g. *Mir* - *Peace, World*). Secondly, although the problem of wrong translations is mostly alleviated by the high-dimensional feature spaces providing sufficient overlap, the quality of data representation could be improved if the translation engine would provide one or more synonyms of uncertain words.

Yet, our approach facilitates straight-forward browsing and exploration of multilingual document collections in a given target language.

Bibliography

[BC97] L. Ballesteros and B. Croft.
Phrasal translation and query expansion techniques for cross-language information retrieval.
In *Proc ACM Special Interest Group on Information Retrieval (SIGIR 1997)*, pages 84 - 91, Philadelphia, PA, 1997.

[BM93] J. Blackmore and R. Miikkulainen.
Incremental grid growing: Encoding high-dimensional structure into a two-dimensional feature map.
In *Proceedings of the IEEE International Conference on Neural Networks (ICNN'93)*, volume 1, pages 450-455, San Francisco, CA, USA, 1993.

[CC92] M. Chalmers and P. Chitson.
Bead: Exploration in information visualization.
In *Proc. of the 15th Annual Int'l. ACM SIGIR Conf.*, pages 330 - 337, Copenhagen, Denmark, 1992.

[CSO96] H. Chen, C. Schuffels, and R. Orwig.
Internet categorization and search: A self-organizing approach.
Journal of Visual Communication and Image Representation, 7(1):88-102, 1996.

[DMR00] M. Dittenbach, D. Merkl, and A. Rauber.
The growing hierarchical self-organizing map.
In S. Amari, C. L. Giles, M. Gori, and V. Puri, editors, *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2000)*, volume VI, pages 15 - 19, Como, Italy, July 24. - 27. 2000. IEEE Computer Society.

[Fri95] B. Fritzke.
Growing Grid - A self-organizing network with constant neighborhood range and adaption strength.
Neural Processing Letters, 2(5):1 - 5, 1995.

[HG96] D.A. Hull and G. Grafenstette.
Querying across languages: A dictionary-based approach to multilingual information retrieval.
In *Proc ACM Special Interest Group on Information Retrieval (SIGIR 1996)*, pages 49 - 57, Zьrich, Switzerland, 1996.

[HP96] M.A. Hearst and J.O. Pedersen.
Reexamining the cluster hypothesis: Scatter/Gather on retrieval results.
In *Proceedings of the 19. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 76-84, Zьrich, Switzerland, August 18. - 22. 1996. ACM.

[KKL$^+$00] T. Kohonen, S. Kaski, K. Lagus, J. Salojдrvi, J. Honkela, V. Paatero, and A. Saarela.
Self-organization of a massive document collection.
IEEE Transactions on Neural Networks, 11(3):574-585, May 2000.

[Koh82] T. Kohonen.
Self-organized formation of topologically correct feature maps.
Biological Cybernetics, 43, 1982.

[Koh95] T. Kohonen.
Self-organizing maps.
Springer-Verlag, Berlin, 1995.

[KS97] D. Koller and M. Sahami.
Hierarchically classifying documents using very few words.
In *Proceedings of the International Conference on Machine Learning (ML97)*, 1997.

[LSG91] X. Lin, D. Soergel, and Marchioni G.
A self-organizing semantic map for information retrieval.
In *Proceedings of the 14. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR91)*, pages 262-269, Chicago, IL, October 13 - 16 1991. ACM.

[LY00] C.H. Lee and H.C. Yang.
Towards multilingual information discovery through a SOM based text mining approach.
In T. Ah-Hwee and P. Yu, editors, *Proceedings of the International Workshop on Text and Web Mining (PRICAI 2000)*, pages 80-87, Melbourne, Australia, August28 - September 1 2000. Deakin University, Australia.

[Mii90] R. Miikkulainen.
Script recognition with hierarchical feature maps.
Connection Science, 2:83 - 101, 1990.

[MR98] D. Merkl and A. Rauber.
CIA's view of the world and what neural networks learn from it: A comparison of geographical document space representation metaphors.
In G. Quirchmayr, E. Schweighofer, and T.J.M. Bench-Capon, editors, *Proceedings of the 9. International Conference on Database and Expert Systems Applications (DEXA98)*, number LNCS 1460 in Lecture Notes in Computer Science, Vienna, Austria, August 24. - 28. 1998. Springer.

[MR00] D. Merkl and A. Rauber.
Document classification with unsupervised neural networks.
In F. Crestani and G. Pasi, editors, *Soft Computing in Information Retrieval*, pages 102-121. Physica Verlag, 2000.

[Oar97] D. W. Oard.
Cross-language information retrieval bibliography.
`http://www.clis.umd.edu/dlrg/filter/papers/clirbib.ps`, 1997.

[Rau99] A. Rauber.
LabelSOM: On the labeling of self-organizing maps.
In *Proceedings of the International Joint Conference on Neural Networks (IJCNN'99)*, Washington, DC, July 10. - 16. 1999.

[RB00] A. Rauber and H. Bina.
Visualizing electronic document repositories: Drawing books and papers in a digital library.
In H. Arisawa and T. Catarci, editors, *Advances in Visual Database Systems: Proceedings of the IFIP TC2 WG2.6 5. Working Conference on Visual Database Systems*, pages 95 - 114, Fukuoka, Japan, May, 10.- 12. 2000. Kluwer Academic Publishers.

[RDM00] A. Rauber, M. Dittenbach, and D. Merkl.
Automatically detecting and organizing documents into topic hierarchies: A neural-network based approach to bookshelf creation and arrangement.
In J. Borbinha and T. Baker, editors, *Proceedings of the 4. European Conference on Research and Advanced Technologies for Digital Libraries (ECDL2000)*, number 1923 in Lecture Notes in Computer Science, pages 348-351, Lisboa, Portugal, September 18. - 20. 2000. Springer.

[RM99] A. Rauber and D. Merkl.
The SOMLib Digital Library System.
In S. Abiteboul and A.M. Vercoustre, editors, *Proceedings of the 3. European Conference on Research and Advanced Technology for Digital Libraries (ECDL99)*, number LNCS 1696 in Lecture Notes in Computer Science, pages 323-342, Paris, France, September 22. - 24. 1999. Springer.

[RSM00] A. Rauber, E. Schweighofer, and D. Merkl.
Text classification and labelling of document clusters with self-organising maps.
Journal of the Austrian Society for Artificial Intelligence (ЦGAI), 19(3):17-23, October 2000.

[Sal89] G. Salton.
Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.
Addison-Wesley, Reading, MA, 1989.

[SHS96] M. Sahami, M. Hearst, and E. Saund.
Applying the multiple cause mixture model to text categorization.
In *AAAI Spring Symposium on Machine Learning in Information Access*, Stanford, USA, 1996.

[SO00] R. Sperer and D.W. Oard.
Structured translation for cross-language information retrieval.
In *Proc ACM Special Interest Group on Information Retrieval (SIGIR 2000)*, Athens, Greece, 2000.

[Son98] M. Song.
Bibliomapper: A cluster-based information visualization technique.
In *IEEE Symposium on Information Visualization (INFOVIS'98)*, North Carolina, 1998.