

Efficient Preprocessing for Information Retrieval with Neural Networks

Thomas Mandl

Information Science – University of Hildesheim

Marienburgplatz 22 - 31141 Hildesheim - Germany

Tel.: ++49-5121-883-837, e-mail: mandl@rz.uni-hildesheim.de

Abstract: Neural networks are well suited for Information Retrieval (IR) from large text or multimedia databases. Their capacity for tolerant and intuitive processing offers new perspectives in IR where the vague nature of human relevance judgements has confronted theory and systems with considerable problems. Most models use the keyword representation vector as input or output. However, fulltext indexing brings forth large vectors which are difficult to handle for neural networks. This article discusses methods for dimensionality reduction used in IR and applies one of them, Latent Semantic Indexing (LSI) to information retrieval using a neural backpropagation network. The transformation between two representation schemes is enabled through preprocessing by LSI which is based on Singular Value Decomposition (SVD).

1 Introduction

IR is concerned with the analysis, representation and retrieval of texts. The increasing amount of machine readable text documents available is a great challenge for computer and information scientists as improved retrieval quality may have a considerable impact on the typical computer user. Hence, many methods from Artificial Intelligence have been tested in IR. Neural networks in particular seem to have the properties needed for more intelligent IR.

Neural networks with large dimension spaces as they commonly occur in IR are difficult to handle and require either expensive hardware or a considerable amount of time. This article introduces an approach to overcome the issue of high dimensional spaces by preprocessing the IR data using Latent Semantic Indexing (LSI). LSI was originally developed as a method for the entire IR process. In the approach here presented, it is combined with a backpropagation neural network which implements an IR model.

2 Neural Networks in Information Retrieval

Neural networks have been applied to Information Retrieval in different manners. Overviews can be found in Doszkocs et al. 1990, Chen 1995 and Mandl 1998a. The three principal approaches are:

- Spreading activation models are basically Hopfield networks customarily used with nodes for query terms and document nodes to retrieve the most activated documents. The links are weighted according to the document-term-matrix determined by a common indexing algorithm. Kwok/Grunfeld 1996 and Boughanem/Soulé-Dupuy 1997 applied spreading activation models for large real world data and achieved results comparable to those of statistical models dominating in IR research and development.
- Transformation network has been suggested by Crestani/Rijsbergen 1997 for query optimization. It consists of a backpropagation network with one or more hidden layers where input and output are representation schemes.
- The COSIMIR model (COgnitive SIMilarity learning in IR; Mandl 1998a,b,c) utilizes backpropagation for the match between query and document representation. Both query and document serve as input to the network which calculates their similarity in the output layer. The similarity serves as measure for the relevance of the document to the query. The training data needs to be collected from a large amount of relevance judgements from users. As a result, COSIMIR implements a cognitive similarity function.

All models use the term vector in which a document is represented by few terms within a large vector of all terms occurring in the document collection. The models encounter difficulties deriving from the size of these sparsely coded vectors. Especially fulltext retrieval results in large vectors potentially containing each word of a natural language. Even manual indexing using a controlled thesaurus often leads to large vectors. For example, the thesaurus of the

Social Science Information Center in Bonn contains some 22.000 entries. A neural network with such a large number of nodes requires significant resources. Therefore, dimension reduction is a promising approach for a successful utilization of neural networks in IR.

3 Dimension Reduction

Regardless of the requirements of neural networks, the reduction of dimensions has been a successful approach to IR. The following dimension reduction methods have been applied in IR:

- Context Vector
- Compression by Backpropagation
- Latent Semantic Indexing (LSI)

The context vector (cf. e.g. Gallant et al. 1993) approach roots in symbolic AI. The index terms are intellectually mapped onto a vector of important basic categories which can be used to describe the features of each term. Each term can be assigned to various categories with a certain weight. Thus, the local representation of the index term, in which the index term is one element in a vector is transformed into a distributed representation where each term activates several different categories and resultantly different elements in the representation vector.

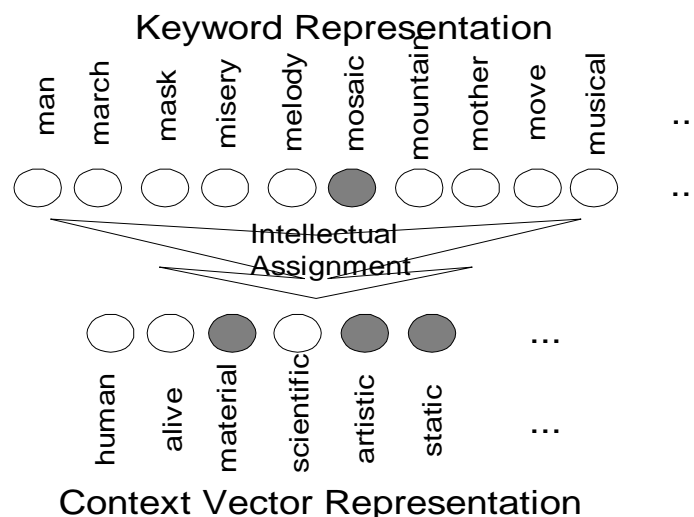


Figure 1: One keyword is mapped onto several context vector categories

The advantage of the context vector is the creation of symbolically interpretable representations. However, the intellectual mapping is rather costly. Both other methods work automatically and also provide distributed representation in a space with less dimensions. Yet, they create dimensions which cannot be interpreted symbolically. Merkl 1995 reports experiments with the Backpropagation compression method. The domain is software reuse in a large software library from which 489 keywords were extracted. He implemented a Kohonen SOM to allow retrieval of software classes. The process of self organization is very time consuming. Therefore, Merkl 1995 reduced the dimensionality of the data by compressing it with a backpropagation network with equal input and output. The compressed version can be read from the hidden units after successful training. Merkl 1995 used 75 and 30 hidden units for the original 489 dimensions and reduced training time significantly. However, he does not report on the retrieval quality.

Latent Semantic Indexing (LSI) performs Singular Value Decomposition (SVD) of the document-term matrix resulting in a reduced space with lower dimension (cf. Syu et al. 1996; Gordon/Dumais 1998). The technique is related to factor analysis and produces orthogonal factors. The SVD of the document-term matrix D is defined as follows:

$$D = U W V$$

W is a diagonal matrix containing the singular values of D ordered by size. By considering only the k largest and therefore most important dimensions, a k -dimensional approximation D' of the original matrix D can be calculated. In IR, between 100 and 300 dimensions are typically used instead of several thousands.

4 LSI as Preprocessing for a Neural Information Retrieval Systems

Syu et al. 1996 have incorporated LSI as preprocessing for a competition based Hopfield-style network for IR. They experimented with some of the classical and often used IR collections and reported higher retrieval quality and substantial improvements in efficiency measured through time. The approach presented here extends the idea of using LSI as preprocessing for a neural network IR system to IR systems based on backpropagation. The transformation network is a simple backpropagation network which maps between two different representation schemes. One representation of an object can be transferred into a different representation of the same object.

This transfer is often necessary in IR and mostly implemented with statistical methods based on cooccurrences. Transformations are needed when documents are manually or automatically indexed according to different thesauri. Users accustomed to working with one thesaurus often have difficulties finding the appropriate terms in another thesaurus where the items of their interest may be at different positions or even named differently. Examples and methods are provided e.g. by Yang 1995, Ferber 1997, Krause et al. 1998 and Buckland et al. 1999.

A solution based on a backpropagation network has been proposed by Mandl 1998a who adapted a model suggested by Crestani/Rijsbergen 1997. This transformation network learns to map between two representations; one situated in the input and the other one in the output.

4.1 Description of the data

The data used for the experiments is part of the literature and research project databases of the Social Science Information Center, Bonn (IZ). The IZ manually indexes incoming documents following two representation schemes (cf. Kluck 1998):

- the thesaurus: a collection of some 22.000 keywords and synonym relations
- the classification: a hierarchy of scientific disciplines from the perspective of the social sciences containing 157 entries

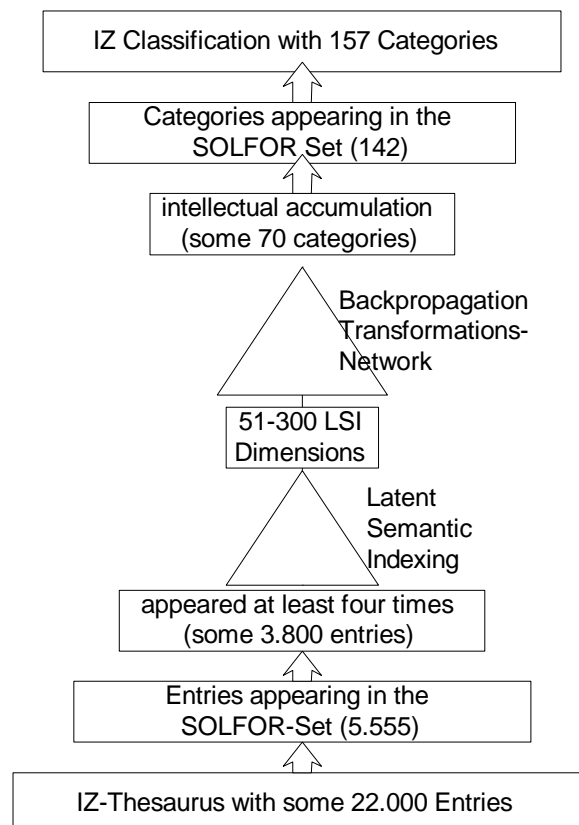


Figure 2: Schema of the Transformation from Thesaurus to Classification

According to the human indexers, the assignment of terms within both schemes are independent. The experiment reported tries to find a transfer function automatically. Such a system can be useful, when the IZ or another information provider e.g. indexes some documents only conforming to the thesaurus. For users, who are familiarized with the classification, an automatic transformation would be a value added service.

For the experiment a subset of the IZ databases was used. The SOLFOR subset contains 12.000 documents, text documents as well as research project descriptions. The documents are manually indexed with both the thesaurus as well as with the classification. In average each document is assigned 13 descriptors from the thesaurus and 2.3 classification entries. Especially for the thesaurus descriptors, the variance is high though.

Because the subset did not contain all keywords and classification entries, the experiment was not set up as a direct transformation between thesaurus and classification. Of the 22.000 thesaurus entries only 5.555 occurred in the 12.000 documents. Those appearing at least four time were chosen and such, a vector of 3800 elements was created for each document. The data set contained 142 of all 157 classification categories. They were further accumulated to form 70 categories. Such, the task of the combined system (LSI and transformation network) was to transform a 3.800 element vector into a 70 element vector (see. Figure 2).

4.2 Experiments and Results

The LSI analysis was carried out with the Bellcore experimental software on a SUN workstation. Data sets with 50, 100 and 300 LSI dimensions were created which required little time. The resulting vectors were normalized before processed by the neural network. Eleven thousand documents were used for training the network and 1.000 for testing. The data was fed into a backpropagation algorithm. The experiments with 50 LSI dimensions gave the best results, although in most applications between 100 and 300 dimensions are used. However, the SOLFOR data set had only 3.800 dimensions - compared to many thousands in other applications - before entering preprocessing with LSI. As a result, a reduction to 50 was a sufficient approximation in this experiment. The experiments with the Backpropagation network and the analysis of the results were carried out on a Standard PC running under NT. Training one network required less than one hour. The best MLP had 20 hidden units. The results of the network were assessed by analyzing the classification quality. Table 1 shows the results.

Table 1: Results in the Test Set

Correct recognition of Mappings	63%
Correct recognition of non-Mappings	97%
Correct recognition overall	96%
Errors per document	1.6

Overall, the results are satisfying considering the difficulty of the task. Human experts thought there was little or no relationship between the classification and the thesaurus. Yet, the MLP reached a high recognition rate. Improvements may be achieved with larger data sets with a more balanced distribution of the documents over the categories. Figure 3 demonstrates how much the quality of the net depends on the number of examples presented for that class in the training set. The results certainly encourage further experiments.

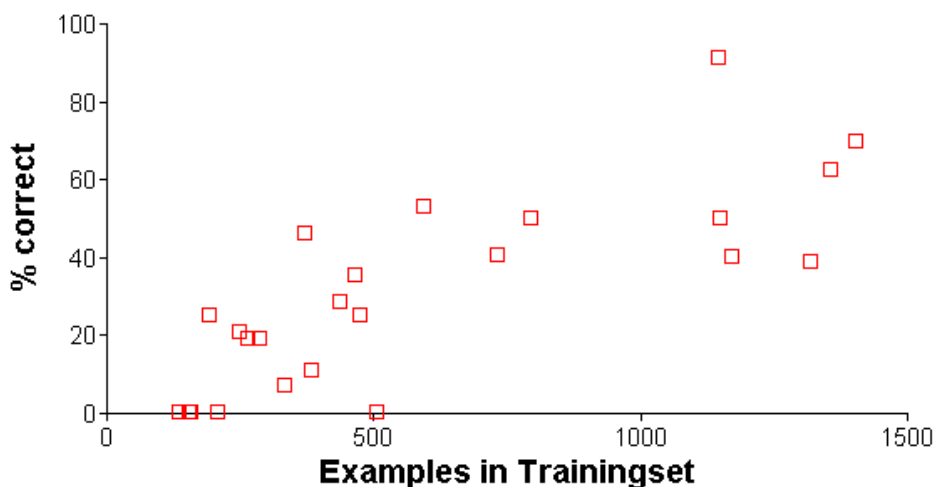


Figure 3: Relationship Between Number of Examples in the Training Set and Recognition in the Test Set

4 Conclusion

The presented approach may be used in daily indexing work at the IZ to suggest classification categories to a human indexer after being presented the thesaurus descriptors. The idea of using LSI for preprocessing needs to be tested for other neural network information retrieval systems using backpropagation. SVD may also be useful in other domains with large sparse vectors.

Acknowledgements:

This research was supported in part by grants from the German Ministry of Economy (no. II C7-003060/10 and IV C2-003060/22). I would like to thank the Social Science Information Center in Bonn for providing the document collection and Bellcore for providing their experimental LSI software.

References:

- Buckland, Michael et al. (1999): Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies. In: D-Lib Magazine vol. 5(1) Jan. URL: <http://mirrored.ukoln.ac.uk/lis-journal...lib/january99/buckland/01buckland.html>
- Boughanem, M.; Soulé-Dupuy, C. (1997): MercureO2: adhoc and routing tasks. In: Harman, Donna (ed.): The Fifth Text Retrieval Conference (TREC-5). URL: <http://trec.nist.gov/pubs/trec5/papers/irit.ps>
- Chen, Hsinchun (1995): Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. In: Journal of the American Society for Information Science vol. 46 (3). pp. 194-216.
- Crestani, Fabio; Rijsbergen, Cornelis van (1997): A Model for Adaptive Information Retrieval. In: Journal of Intelligent Information Systems.
- Doszkocs, T.E.; Reggia, J.; Lin, X. (1990): Connectionist Models and Information Retrieval. In: Annual Review of Information Science and Technology (ARIST), vol. 25. pp. 209-260.
- Ferber, Reginald (1997): Automated Indexing with Thesaurus Descriptors: A Cooccurrence Base Approach to Multilingual Retrieval. In: Peters, Carol; Thanos, Constantino (eds.): Research and Advanced Technology for Digital Libraries. 1st European Conf. ECDL'97. pp. 233-252.
- Gallant, Stephen; Caid, William; Carleton, Joel; Hecht-Nielsen, Robert; Qing, Kent; Sudback, David (1993): HNC's MatchPlus System. In: Harman, Donna (ed.): The First Text Retrieval Conference (TREC-1). NIST Special Publication 500-207. pp. 107-111.
- Gordon, Michael; Dumais, Susan (1998): Using Latent Semantic Indexing for Literature Based Discovery. In: Journal of the American Society for Information Science. vol. 49(8). pp. 674-685.
- Kluck, Michael (1998): German indexing and retrieval test database: some results of the pre-test. In: Discovering new worlds of IR. IRSG98. Grenoble, France. 25-27.3.1998.
- Krause, Jürgen; Mandl, Thomas; Schaefer, André; Stempfhuber, Maximilian (1998): Text-Faktenintegration in Informationssystemen. In: Zimmermann, Harald; Schramm, Volker (Hrsg.) (1998): Knowledge Management und Kommunikationssysteme. Proc. 6. Int. Symposium für Informationswissenschaft. 3.-7.11.98, Prag. S. 413-426.
- Kwok, K.L.; Grunfeld, L. (1996): TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In: Harman, Donna (ed.): The Fourth Text Retrieval Conference (TREC-4). URL: <http://trec.nist.gov/pubs/trec4/papers/queenst4.ps>
- Mandl, Thomas (1998a): Das COSIMIR-Modell: Information Retrieval mit Neuronalen Netzen. Informationszentrum Sozialwissenschaften Bonn, Arbeitsbericht, Feb. 1998. URL: <http://www.uni-hildesheim.de/~mandl/cosimir/>
- Mandl, Thomas (1998b): Learning Similarity Functions in Information Retrieval. In: Zimmermann, Hans-Jürgen (ed.): EUFIT '98. 6th European Congress on Intelligent Techniques and Soft Computing. Aachen, Germany, 8.-10.9.1998. pp. 771-775.
- Mandl, Thomas (1998c): Vague Transformations in Information Retrieval. In: Zimmermann, Harald; Schramm, Volker (Hrsg.) (1998): Knowledge Management und Kommunikationssysteme: Workflow Management, Multimedia, Knowledge Transfer. Proc. 6. Int. Symposium für Informationswissenschaft. (ISI '98). 3.-7.11.98, Prag. S. 312-325.
- Merkel, Dieter (1995): Content-Based Document Classification with highly Compressed Input Data. In: ICANN '95. Paris. October 9-13 1995. vol. 2. pp. 239-244.
- Syu, Inien; Land, S. D.; Deo, Narsingh (1996): Incorporating Latent Semantic Indexing into a Neural Network Model for Information Retrieval. In: ACM CIKM 96. Rockville MD. pp. 145-153.
- Yang, Yiming (1995): Noise Reduction in a Statistical Approach to Text Categorization. In: Fox, Edward, Ingwersen, Peter; Fidel, Raya (eds.): 18th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval. New York 1995. (SIGIR'95. Jul. 9-13, 1995. Seattle, USA). pp. 256-263.