

Mandl, Thomas (1999): 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.

*6th European Congress
on Intelligent Techniques
& Soft Computing
Aachen, Germany,
September 7 - 10, 1998
Proceedings
Volume 2*

**EUFIT
1998**

Learning Similarity Functions in Information Retrieval

Thomas Mandl
Social Science Information Centre
Lennéstraße 30 - 53113 Bonn - Germany
Tel.: ++49-228-2281-179, e-mail: ma@bonn.iz-soz.de

Abstract: Most models for Information Retrieval (IR) using neural networks are simple spreading activation models. Some of them were successfully applied to real world document collections. Nevertheless, they do not exploit the subsymbolic paradigm of neural processing. In this paper a model using a simple backpropagation network for IR is proposed. The COSIMIR model implements the central process in IR. It is a backpropagation network which calculates the similarity between a document and a query representation. The similarity function is learned through examples. Hence, it implements a cognitive similarity function. The first evaluation demonstrates that COSIMIR works well for short vectors.

1 Introduction

The COSIMIR (Cognitive SIMilarity Learning in Information Retrieval) model is intended for similarity calculation in IR. It needs to learn from a large number of relevance judgements on query/document combinations. Thus, COSIMIR learns the complex dependencies between terms which are ignored by most IR models.

Evaluation with real world data is essential for every IR model. More elaborated system design does not always result in higher retrieval quality. For decades evaluation in document retrieval suffered from incomparable results as each researcher used his own data set for evaluations. Since the advent of TREC (Text REtrieval Conference), the situation has improved significantly. TREC provides a general testbed for IR by maintaining a real world document collection, queries and relevance judgements. TREC is organized by the National Institute of Standardization (NIST) in Gaithersburg, USA (for an overview cf. Womser-Hacker 1996; newest results cf. Harman 1997, 1998). The TREC studies have shown, that the average recall (percentage of relevant documents retrieved from the corpus) is about 30%. Unfortunately, many IR experiments have shown that results from one corpus cannot be transferred to other corpora. Thus, TREC can be only one factor when choosing a retrieval model.

The evaluation of COSIMIR has so far been restricted to factual data. Experiments with e.g. TREC data require far more hardware and software resources.

2 Neural Networks in Information Retrieval

In recent years many models for IR based on neural networks have been proposed. Overviews can be found in Doszkocs et al. 1990, Chen 1995 and Mandl 1998. The following discussion does not include semantic networks which use labelled links.

2.1 Spreading Activation Models

Most models are based on the spreading activation. This term is commonly used although every neural network is based on spreading activation. From the neural network point of view, these networks are Hopfield-networks with layer structure. The basic architecture is shown in figure 1. The objects from an IR-System are represented in several layers and the term layer is present in all models. Most models also include a document layer, unless they are used solely for term expansion. Other models have an additional layer for queries. The connection strengths are initialized with knowledge derived by typical indexing methods, such as inverse document frequency (Womser-Hacker 1996).

The basic functionality of these models is straightforward. The user or his query initiate activation in the network. The activation spreads through the net and according to some rule, the spreading is halted. The most activated document units are the result. The advantage of the spreading activation model is the natural integration of term expansion and

relevance feedback in a framework which provides a good metaphor for IR. Term expansion occurs, when a query term activates documents which again activates the terms associated with it. After only two steps, the original query terms have already activated other terms, which occur in the same documents. Some models allow connections within layers. Such, associative relationships between terms or documents can be realised directly.

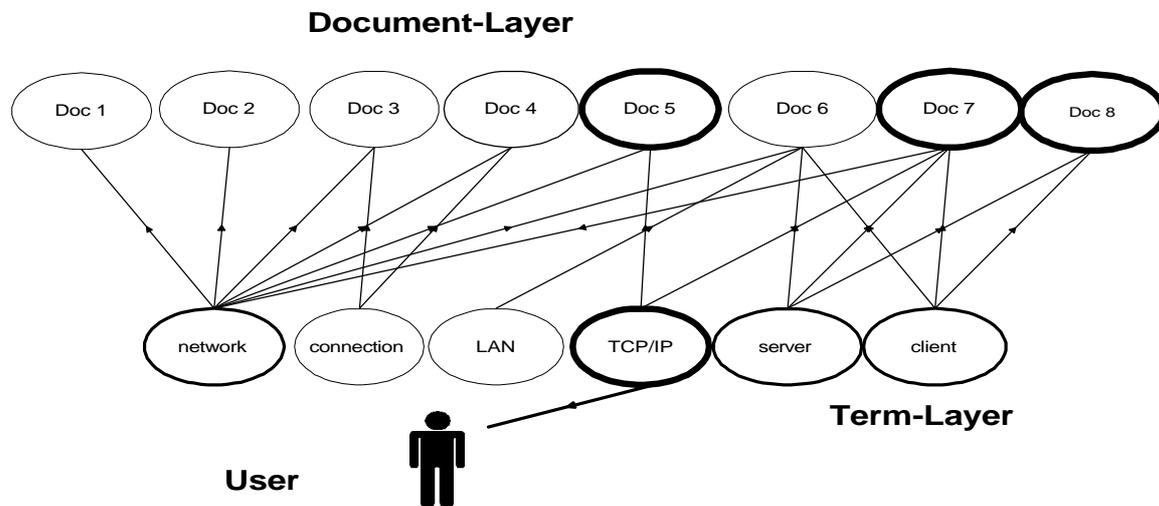


figure 1: Typical Spreading Activation Network for IR: The original query term has led to the activation of several documents and some further terms.

Relevance feedback is usually implemented as follows: after some steps, the result is presented to the user. He marks the most relevant documents. These receive a high activation. Such, they can activate the terms associated with them which in turn activate more documents. Some models learn from the feedback information (e.g. Kwok/Grunfeld 1996; Layaida/Caron 1994), others incorporate no learning (e.g. Salton/Buckley 1988; Wilkinson/Hingston 1992). Further differences between the models arise from the use of different activation functions and indexing schemes.

Several spreading activation models have been successfully applied to the TREC tasks. Especially Kwok/Grunfeld 1996 and Boughanem/Soulé-Dupuy 1997 have achieved good results.

However, the spreading activation approach is not a new paradigm in IR. Many researchers have pointed out, how closely it resembles the vectorspace model. Mothe 1994 proves that spreading activation and vectorspace are two formalisms to describe the same.

2.2 Other IR Models Based on Neural Networks

The spreading activation has been extended hybrid systems, which include semantic knowledge. An advanced system is SCALIR, which was developed for the legal domain. It includes knowledge on court decisions and their relations, such as *overturned* or *followed*. This data is also used during spreading activation.

A preprocessing system for IR based on Backpropagation has been suggested by Crestani/Rijsbergen 1997. Their network transforms a query into an optimized query which is then used in a conventional IR system. For the query expansion process, the model can learn subsymbolic relationships between terms from relevance feedback information. A similar model based on a Hopfield network has been proposed by Bordogna et al. 1996.

3 The COSIMIR-Model

COSIMIR intends to exploit the subsymbolic processing capabilities of neural networks to calculate the similarity calculation between query and document. It consists of a backpropagation network with only one output node for relevance. In the input layer, both the query and document representation vector are fed into the network. Training data can be collected from relevance feedback as well as from jurors judgements. Typically, a query will be used along with many documents in the training set. In the case of shortage of relevance judgements, even similarity calculated by

other approaches can be used to supplement the training set. Note that the COSIMIR-model needs also examples for zero or small relevance. During recall, users can formulate their queries. They are then compared to all documents in the collection and COSIMIR computes the similarity value, which represents the relevance. The collection is then ordered according to these values. Thus COSIMIR needs to make no heuristic choice of a mathematical similarity function as other IR systems. Jones/Furnas 1987 have shown that there is no ideal similarity function for IR. They claim that different measures have different sensitivity to e.g. „within- and between-object term weight relationships“ (Jones/Furnas 1987:423) and stress the importance of empirical validation to find an appropriate measure for the specific task. However, this selection is based solely on meta analysis of the collection and the indexing method. Furthermore, one similarity measure is unlikely to be optimal throughout an entire collection. Different terms might require different mathematical models.

Ideally, COSIMIR learns the complex interdependencies between the terms within its simple network. Using the backpropagation learning algorithm and hidden units with no symbolic representation it can implement a large number of functions. It makes no assumptions about the identity of the two representations. Query and document vector could have different lengths representing different document types, indexing methods or languages.

When applied to text retrieval a COSIMIR network needs to be rather big and thus need a large amount of training data. However, in modern text collections this seems to be available. TREC consists of 350 queries (topics) by now. For each, a set of 1000 relevant documents is available. Documents not occurring in the relevant set can be considered not relevant. Another option to cope with this difficulty is the use of reduced or compressed representations which are already popular in IR research (e.g. Latent semantic indexing, Dumais 1994; context vector, Caid et al. 1994). Reduced vectors have commonly about 100 to 300 dimensions. The input layer of a COSIMIR network would therefore shrink to about 600 units.

Obviously, COSIMIR can be used as a similarity tool in various other domains where similarity is calculated based on vector representations, where choice of a mathematical model is problem and where enough training data is available.

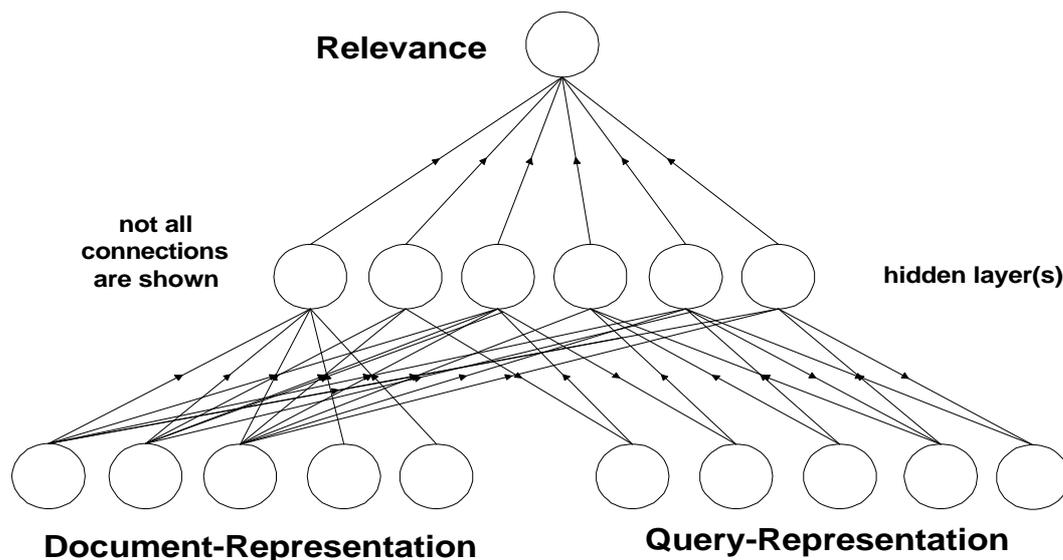


figure 2: The COSIMIR Model

3 Evaluation of COSIMIR

The first evaluation of COSIMIR used short vectors as the available hardware resources were not sufficient for text retrieval. The data set represents materials used for the construction of engines and was used in the project WING-IIR¹ (Materials Information System with Natural Language/Graphical User Interface and Intelligent Information Retrieval, cf. Krause/Womser-Hacker 1997). The component NEURO-WING was developed to detect similar materials (Ludwig/Mandl 1997). When users were looking for the data of a material, they were sometimes confronted with Null answers because of gaps in the database. In this situation NEURO-WING offers a similar material. The internal structure of NEURO-WING is shown in figure 3. The primary definition of similarity in the domain is based on the usage of the material. A backpropagation network learned to map a feature vector onto a usage profile, a task which

¹ The project WING was funded by the German Ministry of Economy, grant no. WI 712.50.

requires expert knowledge. The similarity was then calculated based on the usage profiles using a mathematical measure. Experts were satisfied with the quality of the tool.

This data set was used to evaluate COSIMIR in two experiments. In both cases, the input consisted of two material vectors. The first task was to reimplement a mathematical similarity measure. Although COSIMIR is intended to implement a cognitive similarity function, it must also serve as a general similarity tool and therefore be able to approximate e.g. the cosine measure. The second experiment tested, whether COSIMIR could implement a similarity function where the teacher and input values did not correlate. The feature vectors of the materials were the input and the similarity calculated using the usage vectors were the output (see figure 4). Thus, COSIMIR did not learn directly based on user judgements of similarity. However, the basis of the similarity calculation were the usage profiles, which are result of the features and an expert judgement.

3.1 Comparing Similarity Matrices

Result of both experiments were different similarity matrices for the same objects. As COSIMIR serves as a similarity retrieval tool in this case, the users perspective was chosen as basis for the evaluation. Not the matrices as a whole were compared more the absolute quality of the approximation, but how well the ranked lists based on each material matched. Thus, the matrices was compared row by row. The correlation between two ranked lists was measured using the following formulas:

$$\text{Spearman: } r = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)} \quad \text{Kendall: } t = 1 - \frac{4 \sum_{i=1}^n q_i}{n(n - 1)} \quad (\text{Hartung 1984})$$

Both correlation measure were calculated for each row. The average was of all rows was then calculated as the correlation between the matrices.

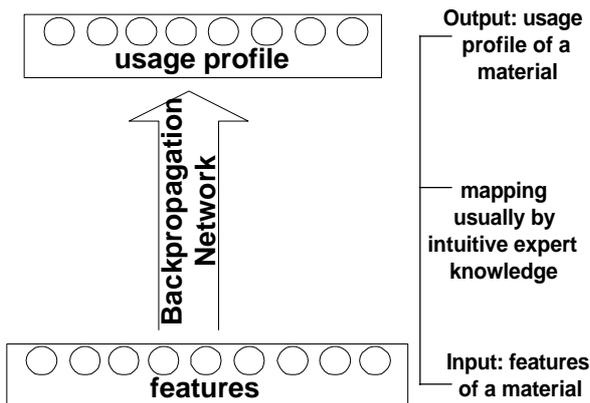


figure 3: Funcionality of NEURO-WING

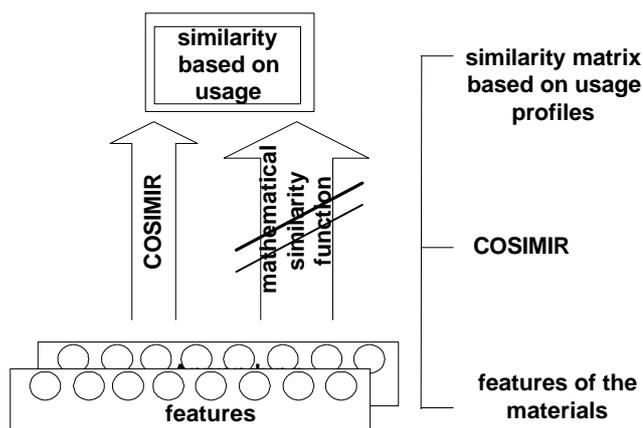


figure 4: Evaluation of COSIMIR

3.2 Evaluation Results

The evaluations show that the COSIMIR model works well for short vectors. The results reported are based on der Spearman Koefficient and were measured in a test set of 23 materials after training the net with all combinations between 46 materials. In the first experiment COSIMIR and Dice correlated with 82% and COSIMIR and Cosine with 79%. This is satisfying considering the small data set and that the correlation among different similarity measure is sometimes lower (cosine-dice: 95%; cosine-pearson 89%, cosine-euklid 65%). The results for the second taks are lower. Both dice and cosine were approximated with 70%. This are also good values as matrices calculated for feature vectors and for usage profiles correlate much less (cosine 37% and dice 6%).

4 Conclusion

This article introduced the COSIMIR model that was developed considering the state of the art of neural networks in IR outlied in the second chapter. First experiments hint, that COSIMIR is a successful approach. Further experiments

with larger data sets and longer vectors are necessary. A detailed description of the state of the art of neural networks in IR, the COSIMIR model and its evaluation can be found in Mandl 1998.

COSIMIR also offers a framework to match documents of different lengths. Therefore it can be applied for multimodal information systems, like ELVIRA¹, in which combined retrieval of text and factual documents is envisioned. ELVIRA is an information system used by three German industrial associations (cf. Scheinost et al. 1998). It offers statistical time series and is currently extended to handle text documents (Krause et al. 1997).

References:

- Bordogna, Gloria; Pasi, Gabriella; Petrosino, Alfredo (1996): Relevance Feedback Based on a Neural Network. In: Zimmermann (ed.) (1996). pp. 846-849.
- Boughanem, M.; Soulé-Dupuy, C. (1997): MercureO2: adhoc and routing tasks. In: Harman 1997.
- Caid, William R.; Dumais, Susan T.; Gallant, Stephen I. (1995): Learned Vector-Space Models for Document Retrieval. In: *Information Processing & Management*. vol. 31 (3).1995. pp. 419-429.
- Chen, Hsinchun (1995): Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. In: *Journal of the American Society for Information Science*. JASIS 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.
- Dumais, Susan (1994): Latent Semantic Indexing (LSI) and TREC-2. In: Harman 1994. pp. 105-115.
- Escobedo, Richard; Smith, Scott; Caudell, Thomas (1993): A Neural Information Retrieval System. In: *International Journal of Advanced Manufacturing Technology* vol. 8 (4). pp. 269-274.
- Harman, Donna (ed.) (1996): The Fourth Text Retrieval Conference (TREC-4).
- Harman, Donna (ed.) (1997): The Fifth Text Retrieval Conference (TREC-5).
- Harman, Donna (ed.) (1998): The Sixth Text Retrieval Conference (TREC-6).
- Hartung, Joachim (1984): *Lehr- und Handbuch der angewandten Statistik*. München, Wien.
- Jones, William; Furnas, George (1987): Pictures of Relevance: A geometric Analysis of Similarity Measures. In: *Journal of the American Society for Information Science*. JASIS vol. 38(6). pp. 420-442.
- Krause, Jürgen; Christa Womser-Hacker (eds.) (1997): *Vages Information Retrieval und graphische Benutzungsoberflächen - Beispiel Werkstoffinformation*. Konstanz.
- Kwok, K.L.; Grunfeld, L. (1996): TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS. In: Harman 1996.
- Layaida, Redouane; Caron, Armand (1994): Applications of the Backpropagation Algorithm to an Information Retrieval System. In: *Proceedings of the RIAO '94 (Recherche d'Information assistée par Ordinateur)*. Rockefeller University. New York. pp. 161-171.
- Ludwig, Michaela; Mandl, Thomas (1997): Ähnlichkeit von Werkstoffen: Die Anwendung unterschiedlicher Wissensmodellierungstechniken für eine intelligente Komponente. In: Krause/Womser-Hacker (1997). pp. 169-184.
- Mandl, Thomas (1998): *Das COSIMIR-Modell: Information Retrieval mit Neuronalen Netzen*. Informationszentrum Sozialwissenschaften Bonn, Arbeitsbericht, Feb. 1998.
- Mothe, Josiane (1994): Search Mechanisms Using a Neural Network Model. In: *Proceedings of the RIAO 94 (Recherche d'Information assistée par Ordinateur)*. Rockefeller University. New York. pp. 275-294.
- Salton, Gerard; Buckley, Chris (1988): On the Use of Spreading Activation Methods in Automatic Information Retrieval. In: Chiaramella, Yves (ed.): *11th Int. Conf. on Information Retrieval*. New York 1988. pp. 147-160.
- Scheinost, Ulrich; Haas, Hansjörg; Krause, Jürgen; Lindlbauer, Jürg (eds.) (1998): *Marktanalyse und Marktprognose: Das ZVEI Verbandsinformationssystem ELVIRA*. Bonn.
- Wilkinson, Ross; Hingston, P. (1992): Incorporating the vector space model in a neural network used for document retrieval. In: *Library HiTech News* vol. 10 (1-2). pp. 69-75.
- Womser-Hacker, Christa (1996): *Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval*. Habilitationsschrift. Universität Regensburg, Informationswissenschaft.
- Zimmermann, Hans-Jürgen (ed.): *EUFIT '96. 4th European Congress on Intelligent Techniques and Soft Computing*. Aachen, Germany, 02.-05. Sept. 1996.

¹The project ELVIRA is funded by the German Ministry of Economy, grant no. II C7-003060/10 and IV C2-003060/22.