

# Tolerant and Adaptive Information Retrieval with Neural Networks

**Thomas Mandl**

*Information Science - University of Hildesheim  
Marienburger Platz 22 - 31141 Hildesheim - Germany  
e-mail: mandl@rz.uni-hildesheim.de*

**Abstract.** The COSIMIR (Cognitive Similarity Learning in Information Retrieval) model applies the powerful backpropagation algorithm to Information Retrieval, integrating human centered, soft and tolerant computing to the core of the retrieval processes. An overview of neural networks in Information Retrieval shows that current systems do not fully exploit the potential of neural networks. An empirical evaluation of COSIMIR has led to positive and promising results.

*Keywords: Backpropagation, Information Retrieval, Neural Networks*

## 1. Information Retrieval

The amount of knowledge available in the world is increasing at a fast pace. Most of it is still conveyed through text documents. Information Seekers need to be directed toward the relevant piece of information in an ocean of knowledge to be able to solve their problems. Therefore, Information Retrieval will be a key technology in the near future. Large scale experiments have shown that current retrieval engines only find a fraction of the relevant documents in a collection (Voorhees and Harman 1998). To deal with the ever growing amount of text, better suited models are necessary. The main weaknesses of today's Information Retrieval systems are:

- Cognitive processes are modeled mathematically:  
Query and document are matched by similarity functions which are not based on the human judgement of similarity. Therefore, the inherent vagueness of the Information Retrieval process is not appropriately modeled in current systems.
- Lack of Adaptivity:  
The mathematical models imperfectly adapt to the situation within a certain domain. Different importance of terms and their combinations are neglected.
- Treatment of heterogeneity:  
Traditional Information Retrieval systems assume a homogeneous and monolithic data source, although users want to retrieve documents of different types with one query. The semantic problems of linking multimedial, multilingual sources remain unsolved.

A sketch of the state of the art in chapter three shows that the current Information Retrieval models based on neural networks have considerable weaknesses. This analysis has led to the development of COSIMIR (Cognitive Similarity Learning in Information Re-

trieval), an innovative model integrating human knowledge into the core of the retrieval process.

## **2. Neural Networks**

Neural networks are an information processing technology within the framework of soft computing and computational intelligence. They are based on the parallel and distributed processing of information which leads to highly tolerant systems. Neural networks can learn from existing data even when humans find it difficult to identify rules. The back-propagation network specifically has been applied to a large number of problems. It learns the mapping between pattern spaces based on examples. Input and output are located in layers of neurons and the Backpropagation networks introduce a hidden layer, which increases the computing capabilities.

Neural networks are also appropriate from the perspective of cognitive science. Smolensky 1988 claims that the introduction of hidden neurons without symbolic equivalence leads toward an "intuitive processor" capable of implementing human expert knowledge.

## **3. Neural Networks in Information Retrieval**

The soft computing paradigm of neural networks seems to be well suited for Information Retrieval tasks. This particular field has attracted considerable research; however, the search for an appropriate architecture has proved to be difficult. Current systems can be grouped into four categories.

### **3.1 Kohonen Self-Organizing-Maps**

Several Researchers have implemented Information Retrieval systems based on the Kohonen Self-Organizing Map (SOM), a neural network model for unsupervised classification. Implementations for large collections can be tested on the internet (Chen et al. 1996, Kohonen 1998). The SOM consists of a usually two-dimensional grid of neurons, each associated with a weight vector. Input documents are classified into the most similar class and in the next step the algorithm adapts the weights of the winning class and its neighbor. As a result, the most similar classes are always the neighboring classes.

The Information Retrieval paradigm for the SOM is browsing. However, users of large text collections need search mechanisms and the SOM does not adapt to search.

### **3.2 Associative Memories**

Associative memories like the Hopfield-Network are powerful error-tolerant retrieval tools. Documents can be stored as energy minima. The query is considered as distorted pattern and serves as input. The network minimizes its energy and tends toward the closest minimum which represents the result document. As a consequence, only one pattern



The model of Mori et al. 1990 is an extension of the spreading activation systems and it includes several hidden layers. It learns to map from sets of query terms to sets of documents. The document layer of such a system seems to be too large to collect sufficient training data. In addition it is unclear how generalization can be guaranteed when the features of the objects are not integrated into the model.

The transformation network (Crestani 1994) maps between two different representation schemes of documents. It does not implement the central process in Information Retrieval, although it can be used for pre-processing in an environment of heterogeneous documents.

#### 4. The COSIMIR-Model

The COSIMIR model (COgnitive SIMilarity Learning in Information Retrieval) implements the central process in Information Retrieval in a backpropagation network and avoids the weaknesses of the discussed models. An Information Retrieval system calculates the similarity between a query and a document representation. COSIMIR learns to calculate this similarity by making use of many examples given by humans.

The input for COSIMIR consists of one query representation and one document representation. Both are fed in parallel into the input layer. The activation spreads through one hidden layer into the output layer consisting of one unit, representing the similarity between both objects. The similarity calculated can be interpreted as relevance of the document for the query. This step needs to be repeated for each document in the collection. By using the backpropagation algorithm, COSIMIR can form sub-symbolic representations in the hidden layer and can implement a complex function.

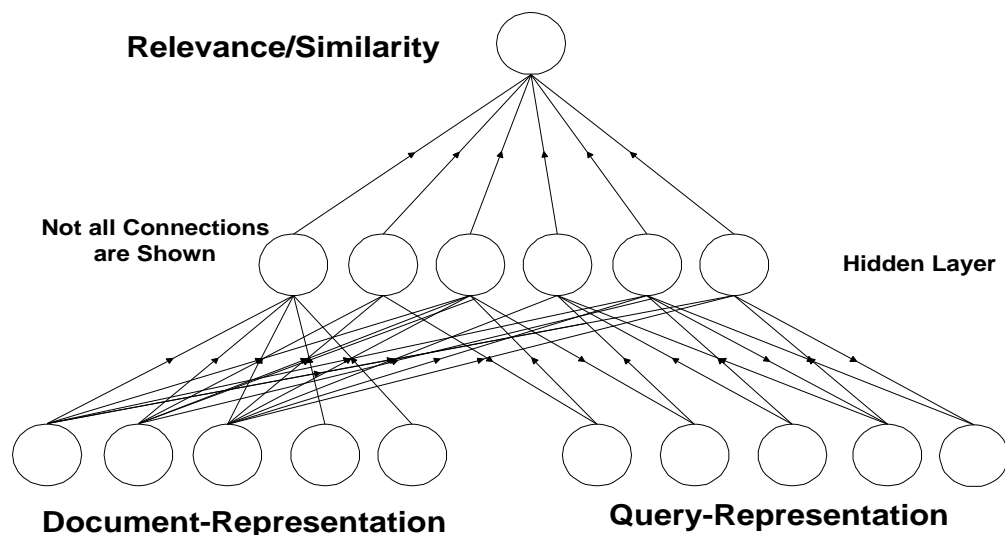


Fig. 3: The COSIMIR Model

## **4.1 The suitability of COSIMIR for Information Retrieval**

The COSIMIR model uses the traditional knowledge source in Information Retrieval, which means that the weights of the terms for the documents are derived by an indexing method. In addition, it integrates a large number of relevance judgements. That way, it makes use of more knowledge than a traditional Information Retrieval system.

Neural networks are a very tolerant processing method. As one result of this tolerance, different evaluations of different users will not dramatically decrease the performance of COSIMIR.

Traditional Information Retrieval systems use a mathematical similarity function like the cosine to calculate the relevance of a document for a query; however, these formulas do not account for the complexity of human similarity judgements. Tversky 1977 showed that similarity is often perceived neither as transitive nor symmetrical; however, most mathematical functions have these properties. COSIMIR does not need to make these assumptions and does not need to model them explicitly. According to the set of relevance judgements, the resulting similarity function implemented in the neural network may be transitive or not. And the choice of one similarity function based on some heuristics can be avoided, as well.

Common Information Retrieval models assume that terms are independent and that all have the same importance for the similarity. Neither of these assumptions is true. COSIMIR does not rely on such assumptions and rather learns the complex relationships between terms. As a result, a cognitive similarity function is implemented rather than a mathematical one.

COSIMIR is also very flexible and can even process heterogeneous representations as long as enough human similarity judgements are available. Hence, it can be applied to a heterogeneous system where the user can form the query in a representation scheme or thesaurus different from the document representation scheme. This scenario will become more and more common when information sources are connected and users are able to query a number of them with one action.

## **4.2 Empirical Evaluation**

COSIMIR was first evaluated with a data set on materials used in the construction of airplane engines. These materials were characterized by two vectors, one representing their features and the other representing the parts for which the materials can be used. According to experts, the similarity of materials in this area is primarily based on the usage of a particular material. Thus, a COSIMIR model which took two feature vectors as input and was trained to calculate the similarity based on the usage vectors was implemented. The performance was measured by comparing the original similarity ranking with the one obtained by COSIMIR on a test set. The correlation reached a percentage of 79%, a result which can be considered very satisfying.

COSIMIR networks for text retrieval tend to become very large, as the number of terms is usually higher than 5000 even for controlled vocabulary. This results in a large number of connections which need to be trained using a sufficient number of training examples.

Therefore, a statistical dimensionality reduction based on Singular Value Decomposition is used for the experiments (Deerwester et al. 1990). Using this method, the term space can be reduced to some 300 dimensions, something which can be handled by COSIMIR. Details of COSIMIR and the experiments carried out can be found in Mandl 1998.

## 5. Conclusion

The COSIMIR model for information retrieval has the potential to improve current systems in order to lead to better results for information seekers. It integrates human knowledge and experience in the form of relevance judgements on documents and queries into the core of the system, and thus it eliminates some heuristic choices in the implementation phase of an information retrieval system. A cognitive similarity function is implemented by learning the regularities of human similarity judgement with a neural network.

## References

- Boughanem M; Soulé-Dupuy C (1998): Mercure at trec6. In: Voorhees and Harman 1998.
- Belew R (1989): Adaptive Information Retrieval: Using a Connectionist Representation to Retrieve and Learn about Documents. In: Belkin and Rijsbergen 1989. pp. 11-20.
- Chen H; Schuffels C; Orwig R (1996): Internet Categorization and Search: A Self-Organizing Approach. In: J of Visual Communication and Image Representation. 7(1). pp. 88-101.
- Crestani F (1994): Domain Knowledge Acquisition for Information Retrieval Using Neural Networks. In: Int J of Applied Expert Systems 2(2). pp. 100-115.
- Deerwester S; Dumais ST; Harshman R (1990): Indexing by Latent Semantic Analysis. In: Journal of the American Society For Information Science 1990. vol. 41 (6). pp. 391-407.
- Kohonen T (1998): Self-organization of Very Large Document Collections: State of the art. In: Niklasson L; Bodén M; Ziemke T (eds.): Proc ICANN98, 8th Int Conf on Artificial Neural Networks, Springer, London. vol. 1, pp. 65-74.
- Kwok K. L. (1989): A Neural Network for Probabilistic Information Retrieval. In: Belkin and Rijsbergen 1989. pp. 21-30.
- Mandl T (1998): Das COSIMIR Modell: Information Retrieval mit dem Backpropagation Algorithmus. ELVIRA-Arbeitsbericht 10, IZ Sozialwissenschaften, Bonn.
- Mori H; Chung CL; Kinoe Y; Hayashi Y (1990): An Adaptive Document Retrieval System Using a Neural Network. In: International Journal of Human-Computer Interaction 2 (3). pp. 267-280.
- Mothe J (1994): Search Mechanisms Using a Neural Network Model. In: Intelligent Multimedia Information Retrieval Systems and Management. Proc. of RIAO '94. New York. pp. 275-294.
- Smolensky P (1988): On the Proper Treatment of Connectionism. In: Behavioral and Brain Sciences vol. 11. pp. 1-74.
- Tversky A (1977): Features of Similarity. In: Psychological Review vol. 84 (4). pp. 327
- Voorhees E; Harman D (eds.) (1998): The Sixth Text Retrieval Conference (TREC-6). NIST Special Publication 500-240. National Institute of Standards and Technology. Gaithersburg. Nov. 19-21 1996.