In: Zimmermann, Harald H.; Schramm, Volker (Hg.): Knowledge Management und Kommunikationssysteme, Workflow Management, Multimedia, Knowledge Transfer. Proceedings des 6. Internationalen Symposiums für Informationswissenschaft (ISI 1998), Prag, 3. – 7. November 1998. Konstanz: UVK Verlagsgesellschaft mbH, 1998. S. 312 – 325

Vague Transformations in Information Retrieval

Thomas Mandl

Social Science Information Centre, Bonn
Lennéstrasse 30; 53113 Bonn; ma@bonn.iz-soz.de

Content

Abstract

Information Retrieval (IR) is often confronted with different and heterogeneous collections of multilingual and multimedial content. Transformations are necessary in order to map between different representation schemes which are the result of different content analysis methods and different terminologies. Vague methods which have been sucessfully applied to IR seem especially suited to tackle the vague nature of such transformations. Statistical methods are reviewed and two experimental systems based on neural networks are presented.

1 Introduction

The advent of multimedia and the ever growing number of database providers shape today's reality in Information Retrieval. Users face an overwhelming number of resources in the internet and other information markets. In many cases, users want to query a wide range of resources as it is not known in the beginning of an information process whether the information need can be satisfied by text documents, factual data or multimedia objects. However, information resources

and their interfaces are often organized by formal criteria, such as numerical data, text documents and collections of internet links. End users as well as information brokers prefer one integrated interface over many interfaces. Alta Vista and Yahoo are popular examples for systems integrating many internet sources.
The heterogeneous nature of the data is a great challange for IR which works on representations of the indexed objects. The representation of text documents consists most times of keywords or index terms which are chosen by a human indexer or derived by an automatic indexing algorithm. Multimedia objects like pictures or video clips may also be represented by keywords and new representation schemes based on their specific features such as color distribution or video scene analysis have been developed. Mappings between keywords and specific multimedia representations nevertheless remain difficult.
In addition, problems can occur when integrating different text collections which have been indexed by different methods or using different thesauri. For example, the meaning of the same word in two fields can differ considerably. Aligning words found by automatic indexing with corresponding terms assigned by human indexers will likely result in inconsistencies and poor retrieval performance. Currently, these differences are widely neglected. Although the Internet proves that searches over heterogeneous documents are possible, the quality of the results also shows that many of the conceptual problems of data integration remain yet to be solved. The next chapter further elaborates on the problems arising from various types of heterogeneity in IR. The layer model deals with text collections of varying data quality and different content analysis methods. The text-fact integration in ELVIRA points towards the integration of multimodal data types. Chapter three reviews vague methods which can be employed to transform between different representation schemes. The family of statistical methods has already been utilized in real world IR problems and seems therefore very appropriate. However, their computational abilities and their learning capabilites suggest neural networks as a candidate for transformations. Two backpropation networks allowing integrated IR are discussed. The transformation network already applied for an experimental system has a high plausibility. The adapted COSIMIR (COgnitive SIMilarity learning in Information Retrieval) model is interesting as it does not need an explicit transformation step, however, so far it has only succeed when used with small data sets.

2 Heterogeneous Datasources in IR

A classical example for retrieval from heterogeneous document collections is Multilingual Information Retrieval. It allows users to formulate their query in one language and to retrieve documents in various others (cf. Hull/Grefenstette 1996). The multilingual corpus can be seen as a meta-database consisting of several monolingual datapools. The most interesting case is the retrieval of a document in a language other than the query language. In this case, the query or the documents must be transformed into a representation in the other language. After that, techniques from monolingual IR can be applied.
The need for transformations in Multilingual Information Retrieval is evident as the vocabulary between languages is basically distinct. In areas where the vocabulary partly overlaps the heterogenity is often ignored. Two examples for such problems are introduced in the remainder of this chapter. Text-Fact-Integration provides an

example for multimodality whereas the layer model integrates datapools of different quality.

2.1 The Layer Model

The layer model (Krause 1996) has been developed for information service institutions. It suggests deregulation to overcome centralistic structures and its political dimension calls for a new information-science model of information provision. Its key idea is the replacement of a monolithic database by a layer structure. The layer model allows various levels of relevance and content analysis where norms and quality are not imposed, but rather coordinated and administered. That means, an information provider does not need to fulfill all the requirements (use of a certain thesaurus, data quality) to be included in the monolithic database of an information service center. Rather, any information provider can choose to be included in an outer layer where requirements are relaxed. The innermost layer contains the nucleus of the most relevant documents where content analysis is deep and of high quality. Towards the outer layers, relevance and quality of content analysis continually decrease. To allow flexible querying over several layers, transformations between different content analysis schemes (e.g. thesuari) need to be established.
The layer model provides a framework for the common situation of many information centers which have performed intellectual indexing for a corpus of controlled quality. In many cases, users are willing to relax their requirements on quality of the data when they can query a larger document base. Therefore, information centers include new data pools, which have been either indexed by other institutions using different thesauri or which have not been indexed yet. Often, these new data sources cannot be subjected to the same deep content analysis due to economical or political reasons. As a result, automatic indexing is used. Nonetheless, the intellectual work for thesaurus development and indexing should not be lost, but rather included in the retrieval process.
This leads to non trivial problems as the same term used in different thesauri can have different meanings. In addition, a term found in many documents by an automatic indexing algorithm has a different value for retrieving documents than the same term given to few documents by a human indexer.

2.2 Text-Fact-Integration

The integration of textual and factual data (statistical time series data and other numerical data in tables) is the focus of the project ELVIRA[1] (Scheinost et al. 1998; Krause et al. 1998). ELVIRA is an online information system for time series distributed by three German industrial associations to their member companies. Empirical studies with users have shown that the market researchers in companies also want to retrieve text documents to satisfy their information needs. A prototype of ELVIRA including text retrieval as well as time series has been implemented. To achieve a user friendly system, the integration needs to be further developed and should allow the retrieval of both data types using one query.
First empirical studies have shown that the indexing vocabularies of texts and facts differ greatly due to the different ways of content analysis. The text documents are indexed automatically resulting in a weighted document-term-matrix, whereas the time series are indexed intellectually using official hierarchical nomenclatures. For example, the controlled nomenclatures for the time series contain only one term for a product whereas the text collection typically contains many synonyms for the same product. Often, the nomenclatures contain an official term not commonly used in everyday language. Therefore, transformations from the indexing terms of one data type to another are necessary.

3 Transformation Methods

The IR process in an integrated system which takes into account the different representation schemes must include a transformation step proceeding the standard IR process. The transformation step needs to assure that both query and document are represented within the same scheme to enable standard IR processes. The transformation of the query into the representation scheme of the according document collection seems to be the most efficient way. However, this method raises the question of whether a user really uses a specific representation scheme when formulating his query. This is indeed the case when he is either an expert in the use of one thesaurus and its terminology or when the interface solely allows thesaurus based access, like the time-series tool of ELVIRA. It is often the case however, that the user will use everyday language and the system transforms his terms into specific thesaurus terms used in the collections in question.

Which methods can be used to implement transformations? One approach is either to construct a single thesaurus for the field or develop concordances between existing thesauri. Multilingual IR often utilizes these approaches. The construction of a single thesaurus is expensive and is not advisable for each field as each existing thesaurus represents a viewpoint which can be appropriate for certain user intentions. Concordances are basically rule-based systems where a rule transforms one term into its synonym in the other thesauri. Although this

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

seems a straightforward method, it bears several problems besides the cost factor. For instance, the mapping may depend on the context and rules may be more complex than simple synomym to synonym transformations. Many thesauri have been developed for distinct purposes and are differ so much that simple mappings are rarely successful and that is especially the case for the large vocabulary derived from automatic free text indexing in which intellectually constructed concordances are not feasible.
In the context of ELVIRA, a concordance between the old nomenclature for statistics of industrial production which was valid until 1994 and the new nomenclature introduced in the European Union in 1995 was developed. Groups of products were broken up and the single products were distributed over distant positions in the new nomenclature where they formed new groups with different products.
Although concordances may work well in some areas, empirical investigations suggest that transformations are a fundamentally vague problem. Therefore, vague information processing methods already successfully applied to IR should be tested in order to implement transformations.
The following paragraphs discuss three vague methods for transformations. The first is based on statistics and has been used for real world multilingual IR (Sheridan/Ballerini 1996). The second and the third approaches are based on neural networks and have so far been only tested with small experimental data sets.

3.1 Statistical Methods

In cases where manual construction of concordances is difficult e.g. between a thesaurus and the indexing vocabulary of automatic indexing, the automatic construction of concordances becomes desirable. Several statistical apporaches designed to achieve this goal have been proposed.
Biebricher et al. 1988 developed the system AIR/PHYS to increase the quality of manual indexing by automatically suggesting appropriate indexing terms. Some already manually indexed documents were additionally analyzed by an automatic indexing algorithm to calculate association factors between free text terms t and thesaurus descriptors d:

$$z(t,d) = \frac{h(t,d)}{f(t)}$$ (Biebricher et al. 1999:334)

f (t) number of documents containing t
h (t,d) number of documents from f (t) indexed with d

For the thesaurus of 22 000 terms, 800 000 association factors were calculated, of which the 350 000 most important were included in the system. The transfer function was enriched with 50 000 USE and 170 000 BROADER TERM relations from the thesaurus. Biebricher et al. 1988 report that the indexers could use many of the suggestions of AIR/PHYS.
Statistical methods with their inherent vagueness have also been utilized for multilingual IR. Sheridan/Ballerini 1996 developed a system similar to the first one where they changed the role of documents and terms, such that each term is

represented by a feature-vector in the document space. A similarity matrix for all terms can be calculated. The knowledge for the transformation is derived using a corpus with comparable documents. Sheridan/Ballerini 1996 used news stories which were available in two languages. Thus, the document space is the same for each language and each term can consequently be transformed into its document vector. Since the same documents exist in the other corpus, the document vector can be used to calculate the corresponding term vector for the second language. As a result, query terms in one language are not translated into a single word using a dictionary, but rather transformed into a weighted vector of terms in the second language representing the use of the query term in the document collection (see figure 1).
Sheridan/Ballerini 1996 tested this method with a collection of 93 000 news stories available in German and Italian. As expected the retrieval results were outperformed by the best mono-lingual experiment, but the overall quality was nevertheless satisfying.

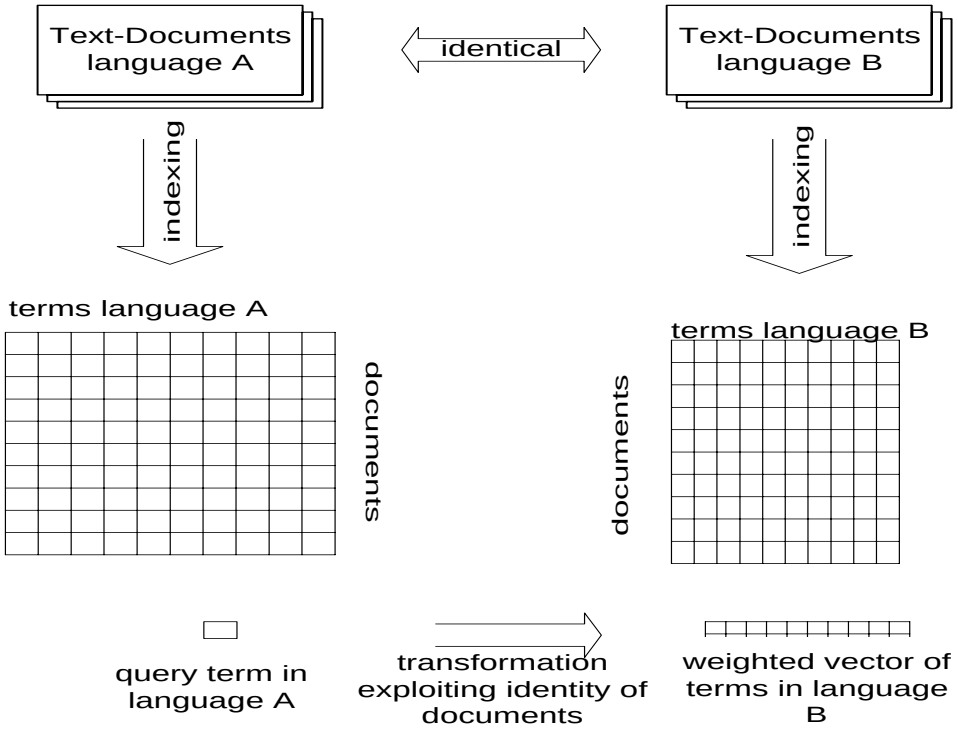

Figure 1: Transformation for multilingual IR (cf. Sheridan/Ballerini 1996)

3.2 Transformation Network

Although statistical methods are well established for vague transformations it is not clear whether they can implement every transformation. The unknown transformation function can be rather complex.
Statistical transformations or concordances implement basically a large rule based system. Therefore, only functions which can be expressed in such a system can be implemented. As the formal nature of the transformation function is unknown, it is advisable to choose a powerful mechanism for its implementation.
Neural networks and in particular the multilayer backpropagation network can learn complex functions such as non linear separable problems (cf. Zell 1994:99).

Moreover, Smolensky 1988 argues that connectionist systems such as neural networks including hidden units realize the subsymbolic paradigma of computing and can implement an intuitive processor with capablities beyond those of symbolic processing. Their knowledge is not derived from rules but from examples which are used as training data. Thus, neural networks can approximate intuitive expert knowledge which cannot be formalized in a rule based system.
All calculations within neural networks are local. Each unit gathers its input from incoming connections and calculates its own activation and output. Signals travel along connections and are modified according to the connection strengths. A typical backpropagation network consists of an input and an output layer and of one or more hidden layers. The input vector is propagated into the network which calculates the output vector. In a training phase, examples for the desired mapping are presented enabling the network to compare the actual output and the desired output. The error is used to tune the connection strengths of the network such that it is closer to the teaching output and that it can better approximate the desired function. After training the network, its generalization capabilities need to be tested with another data set (for details cf. Zell 1994).
Crestani/Rijsbergen 1997 present a backpropagation network for the transformation between different queries in which the representation schemes are equivalent. The same architecture can be modified for general transformations between different representation schemes. The goal of Crestani/Rijsbergen 1997 was to transform actual user queries into queries which achieved better results for the users' information needs. A set of query pairs is necessary in order to train the network, each consisting of the original user query and of an improved one. The improved query was constructed using relevance feedback information. The query tranformation reported in Crestani/Rijsbergen 1997 led to queries with similar performance as the original query, although the adapted queries were considerably different from the original ones. This shows that the network is not merely an implementation of query expansion where additional terms are addded to a query. The Cranfield I collection served as a testbed. Because of its limited size, more testing is necessary in order to show the feasibility of this approach.
The following figure shows an adaption of the network for general transformations between two representations. This architecture is currently used to test mappings between the thesausrus and the classification for the documents of the Social Science Information Centre.

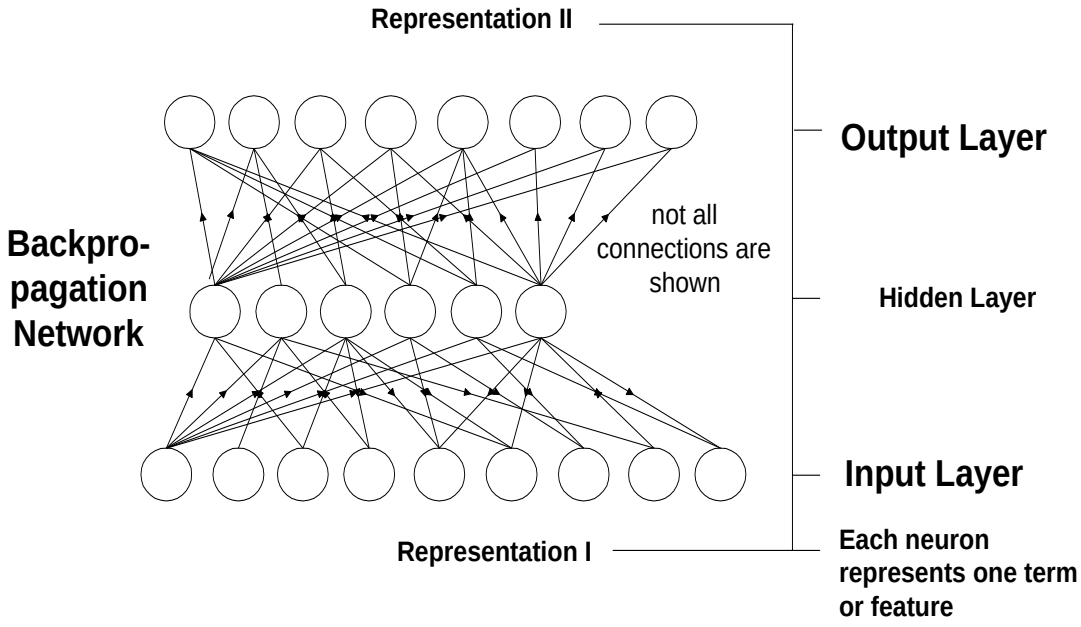

Figure 2: Transformation-Network

The transformation network depends on the same training data as the statistical method for multilingual IR, a corpus of comparable documents in the same feature space.

3.3 The COSIMIR-Model for Transformations

In the same way as the transformation network, the COSIMIR model is based on backpropagation. Currently, most IR-systems consisting of neural networks use Hopfield networks or Self Organizing Maps. These networks do not include hidden units and, therefore, they cannot implement an intuitive processor. COSIMIR is a simple backpropagation network which takes document and query representation as input vector and the relevance between them as the only output unit. So far COSIMIR has only been tested with small problems; however, it can be easily modified to allow retrieval over different representation schemes without an explicit transformation step.

3.3.1 Neural Networks in IR

Neural networks have been applied to IR in many experimental systems (for overviews cf. Doszkocs et al. 1990, Chen 1995, Mandl 1998a). Most systems use either Self Organizing Maps (Kohonen networks) or the so called spreading activation networks (Hopfield networks with layer structure). The latter have been successfully introduced at TREC (Text Retrieval Conference) where their performance is comparable to that of standard algorithms (Mercure, cf. Boughamen/Soule-Dupuy 1997; PIRCS, cf. Kwok/Grunfeld 1996). TREC is an annual meeting in which a large real world test collection as a common testbed for IR systems is provided.
Neither Kohonen nor Hopfield networks include hidden units. Consequently, they cannot exploit the advantages of subsymbolic processing like the backpropagation algorithm. Backpropagation is seldomly part of an IR system and it is usually not

used for the central task in IR which is the match between query and document, but rather for other tasks as in Crestani/Rijsbergen 1997. Although these systems are based on neural networks, they very much resemble the standard vector space model. In Mothe 1994 partial formal equivalence of both was proved.

3.3.2 The COSIMIR Model

The COSIMIR (COgnitive SIMilarity learning in Information Retrieval) model (Mandl 1998a) intends to use backpropagation for the central problem in IR, to calculate the similarity between query and document as a measure for relevance. In most cases, mathematical functions like the cosine or the inner product are used to calculate the similarity in IR. The decision for the use of one or the other can only be made based on empirical data. Jones/Furnas 1987 discuss the sensitivity of various similarity measures to between-object and within-object differences of term weights. However, their analysis provides little guidance for the developer of an IR system. Additionally, Tversky 1977 reports experiments suggesting that the human similarity judgement lacks formal properties such as symmetry and transitivity which mathematical similarity functions embody.

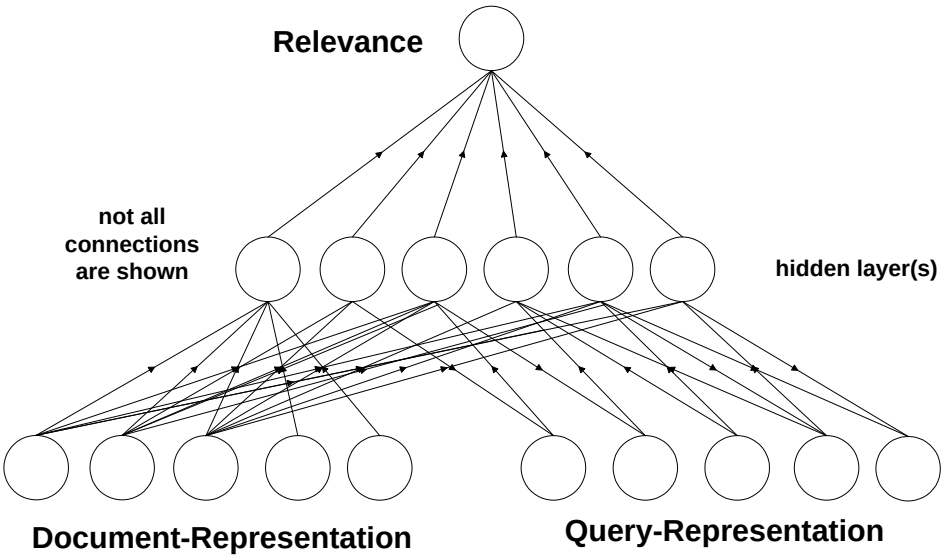

Figure 3: COSIMIR Model

COSIMIR consists of a simple backpropagation neural network which learn the similarity function from users judgements. At the input layer, the query as well as one document representation are presented to the network. The similarity is then calculated in the output unit. Training data for COSIMIR are pairs of documents and queries with relevance (similarity) judgements. The training set also needs to contain cases of little and zero relevance. During the training process, COSIMIR can learn, how the combination of terms in document and query influences relevance. No assumptions on binary independence are necessary, as it is the case in many other IR models.
First experiments with very short vectors from a materials database have shown positive results (see Table). However, experiments with the Cranfield II collection have not produced satisfying results. The number of relevance judgements for the Cranfield II collection is not sufficient for COSIMIR. Further experiments focus on

the use of compressed representations thus reducing the number of input nodes and links in the network.
An equivalent model was proposed by de Jong et al. 1996 as cognitve similarity function within a Case Based Reasoning system for industrial processes.

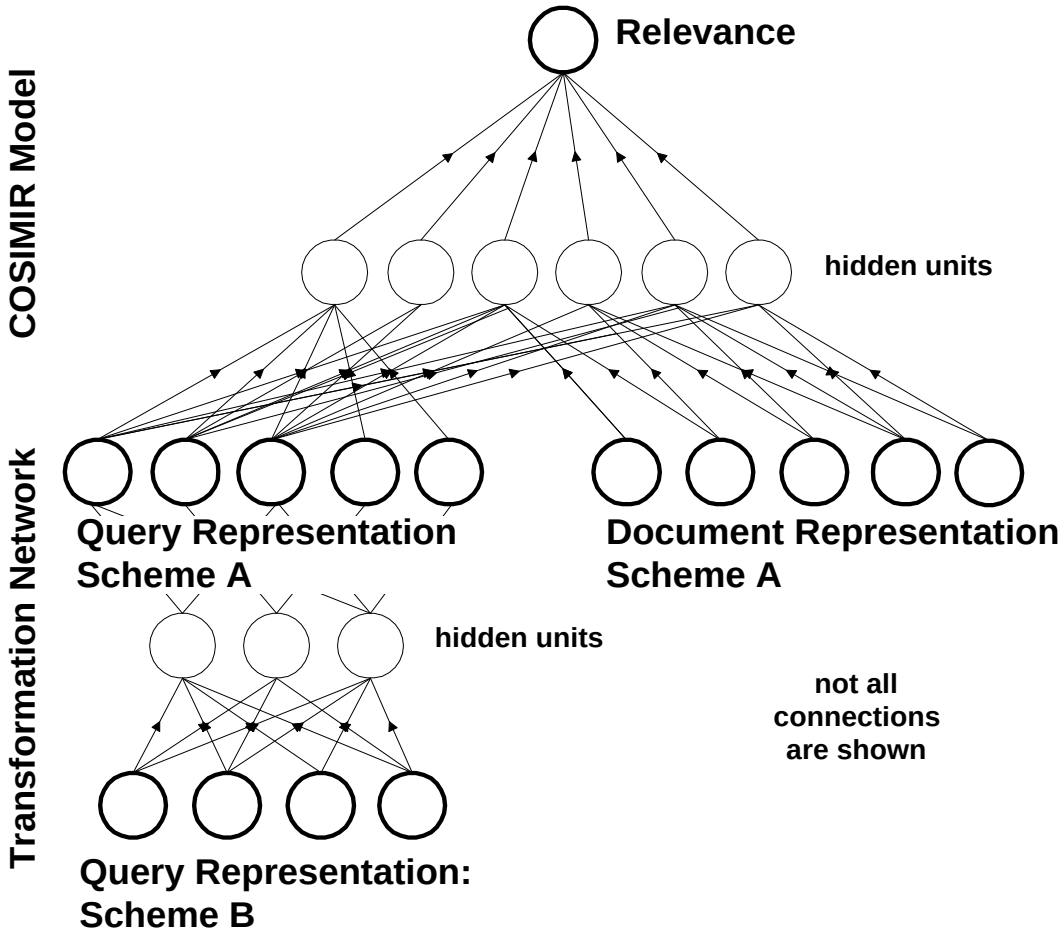

Figure 4: Combination of COSIMIR and a Transformation Network

3.3.3 Adaption of COSIMIR for Heterogeneous IR

COSIMIR can be combined with the transformation network as well as with any other transformation method. Either query or document is transformed and afterwards query and document are compared. Both networks can be combined such that the output of the transformation network will provide part of the input for COSIMIR (fig. 4).
COSIMIR can be adapted to allow queries in heterogeneous datapools, thus requiring only one network. The query and document do not need to be represented within the same scheme as in classical IR systems using mathematical similarity functions. Such, a query representation in one scheme can be directly compared to documents indexed differently once sufficient training data has been collected (fig. 5). No explicit transformation step is necessary. Like the other transformation methods, COSIMIR also needs a corpus of comparable

documents. However, the relevance judgements can be gradual whereas the other methods require binary decisions on comparability.

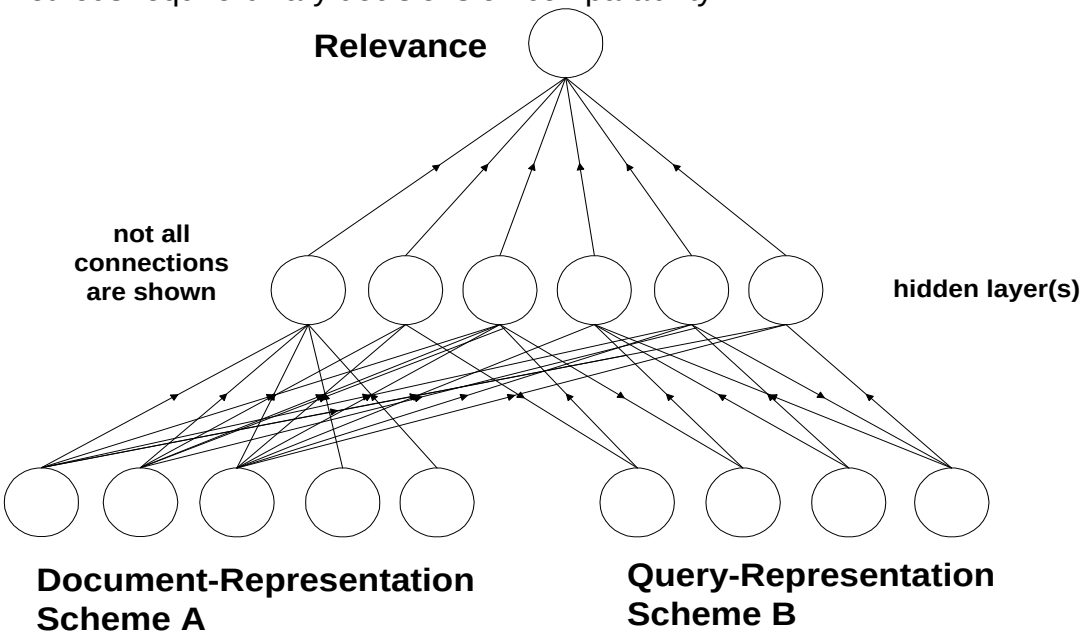

Figure 5: Adapted COSIMIR-Model

This approach was first validated with a small dataset described in Ludwig/Mandl 1997, which represents materials. For each material, two representations were available, namely the physical features and the usage profile. For the users, the usage profile was usually the basis for a similarity judgement. In Ludwig/Mandl 1997, a neural network is discussed which transforms the feature vector into the usage profile. The network reached an overall recognition rate of 85%. As not enough human similarity judgements were availbale, the similarity was calculated based on the usage profile using a standard similarity function. This heuristic approach seems valid as human judgement is usually based on the usage profile. Three experiments with the same target data could be carried out. First, COSIMIR learned to implement a mathematical similarity function, receiving the same input as the cosine. Second, a standard COSIMIR network learned to calculate the usage-similarity based on the feature profiles, thus going beyond the pure implementation of a mathematical function. Third, COSIMIR learned to calculate the usage-similarity receiving one feature vector and one usage profile, thus relying on heterogeneous representations. The data set consists of 72 materials of which one third was used for the test set. Input and output vector each have 22 elements.
The evaluation ot the results did not consider the absolute value of the similarity as it has little importance in IR. It seemed rather appropriate to compare only the ranked lists, where all materials are ordered according to their similarity to one query material. The average correlation between the in target and result lists were calculated and are shown in the following table.

Table: Results with the Materials Dataset

Target	First Input	Second Input	Output	Correlation
implement mathematical similarity function	usage profile	usage profile	cosine based on usage profiles	79%
standard COSIMIR-model	features	features	cosine based on usage profiles	70%
heterogeneous representations	usage profile	features	cosine based on usage profiles	50%

Although the difficulty of the task reduces the quality of the mapping, the results are satisfying, considering that the correlation between the calculated usage-similarity and the calculated feature-similarity is below 40%. Details on these experiments can be found in Mandl 1998a and Mandl 1998b. Further experiments are nevertheless needed.

4 Conclusions

This article discusses three methods for transformations between heterogeneous representation schemes. Whereas the statistical methods are well tested, the neural networks are still in an experimental stage. However, their capabilites to approximate complex functions and first results with small data sets encourage further investigation.
The construction of a proper knowledge base with double representations for aligned or identical documents is crucial for the success of any vague transformation. In the context of ELVIRA, this is currently a priority. There are not enough user or expert judgements available, therefore aligned corpora need to be constructed heuristically. In one text collection of ELVIRA, many texts contain tables which can be identified. As the tables deal with same topic as the text and may contain terms which are more typical for time series descriptions, text and corresponding table may be used as training pair. In addition, this text collection uses its own thesaurus, so that relations between its terms and free text can be determined. For a first approach, statistical methods are applied to the ELVIRA data.

References

[Biebricher et al. 1988]
Biebricher, B.; Fuhr, N.; Lustig, G.; Schwantner, M.; Knorz, G.: *The automatic indexing system AIR/PHYS - from research to application*. In: Chiaramella, Yves (ed.) (1988): Proc. of the 11th Int. SIGIR Conf. ACM. New York. p. 333-342.
[Boughanem/Soulé-Dupuy 1998]
Boughanem, M.; Soulé-Dupuy, C.: *Mercure at trec6*. In: Harman, Donna (ed.) (1998): The Sixth Text Retrieval Conference (TREC-6).
[Chen 1995]
Chen, Hsinchun: *Machine Learning for Information Rerieval: Neural Networks, Symbolic Learning, and Genetic Algorithms*. In: Journal of the ASIS. vol. 46(3). p. 194-216.
[Crestani/Rijsbergen 1997]

Crestani, Fabio; Rijsbergen, Cornelis J. van: *A Model for Adaptive Information Retrieval*. In: Journal of Intelligent Information Systems.
[Doszkocs et al. 1990]
Doszkocs, T.E.; Reggia, J.; Lin, X.: *Connectionist Models and Information Retrieval*. In: Annual Review of Information Science and Technology (ARIST), vol. 25. p. 209-260.
[Frei et al. 1996]
Frei, Hans-Peter; Harman, Donna; Schäuble, Peter; Wilkinson, Ross (eds.): *Proc. of the 19th Annual Int. ACM SIGIR Conf. on Information Retrieval*. New York.
[Hull/Grefenstette 1996]
Hull, David; Grefenstette, Gregory: Querying Across Languages: *A Dictionary-Based Approach to Multilingual Information Retrieval*. In: Frei et al. 1996. p. 49-57.
[Jones/Furnas 1987]
Jones, William; Furnas, George: Pictures of Relevance: *A Geometric Analys-is of Similarity Measures*. In: Journal of the ASIS. vol. 38(6). p. 420-442.
[Jong et al. 1996)]
Jong, E. de; Keuken, H.; Pol, E. van der; Dekker, E. den; Kerckhoffs, E. J.: *Exergy Analysis of Industrial Processes Using AI Techniques*. In: Computers and Chemical Engineering. vol. 20. p. S1631-S1636.
[Krause 1996]
Krause, Jürgen: *Informationserschließung und -bereitstellung zwischen Deregulation, Kommerzialisierung und weltweiter Vernetzung. („Schalenmodell")*. IZ-Arbeitsbericht 6, IZ Sozialwissenschaften, Bonn.
[Krause et al. 1998]
Krause, Jürgen; Mandl, Thomas; Schaefer, André; Stempfhuber, Maximilian: *Text-Fakten-Integration in Informationssystemen*. In this volume.
[Krause/Womser-Hacker 1997]
Krause, Jürgen; Womser-Hacker, Christa (eds.): *Vages Information Retrieval und graphische Benutzeroberflächen - Beispiel Werkstoffinformation*. Konstanz.
[Kwok/Grunfeld 1996]
Kwok, K. ; Grunfeld, L.: *TREC-4 Ad-Hoc, Routing Retrieval and Filtering Experiments using PIRCS*. In: Harman, Donna (ed.) (1996): The Fourth Text Retrieval Conference.
[Ludwig/Mandl 1997]
Ludwig, Michaela; Mandl, Thomas: *Ähnlichkeit von Werkstoffen: Die Anwendung unterschiedlicher Wissensmodellierungstechniken für eine intelligente Komponente von WING*. In: Krause/Womser-Hacker 1997. p. 169-184.
[Mandl 1998a]
Mandl, Thomas: *Das COSIMIR-Modell: Information Retrieval mit Neuronalen Net-zen*. Informationszentrum Sozialwissenschaften Bonn, ELVIRA Arbeitsbericht 10. 1998.
[Mandl 1998b]
Mandl, Thomas: *Der Einsatz vager Verfahren für Transformationen*. Informations-zentrum Sozialwissenschaften Bonn, ELVIRA Arbeitsbericht 13. 1998.
[Mothe 1994]
Mothe, Josiane: *Search Mechanisms Using a Neural Network Model*. In: Intelligent Multimedia Information Retrieval Systems and Management. Proceedings of the RIAO ‘94. New York. p. 275-294.
[Scheinost et al. 1998]

Scheinost, Ulrich; Haas, Hansjörg; Krause, Jürgen; Lindlbauer, Jürg (eds.): *Marktanalyse und Marktprognose. Das ZVEI Verbandsinformationssystem ELVIRA*. Bonn: Informationszentrum Sozialwissenschaften.
[Sheridan/Ballerini 1996]
Sheridan, Páraic, Ballerini, Jean Paul: *Experiments in Multilingual Information Retrieval using the SPIDER System*. In: Frei et al. 1996. p. 58-65.
[Smolensky 1988]
Smolensky, Paul: *On the Proper Treatment of Connectionism*. In: Behavioral and Brain Sciences. vol. 11. 1988. p. 1-74.
[Tversky 1977]
Tversky, Amor: *Features of Similarity*. In: Psychological Review. vol. 84(4) July. p. 327.
[Zell 1994]
Zell, Andreas: *Simulation Neuronaler Netze*. Bonn et al.