Semantic indexing and searching using a Hopfield net

Hsinchun Chen, Yin Zhang and
Andrea L. Houston

University of Arizona, Tucson, USA

Received 16 June 1997

Revised 8 September 1997

Abstract.

This paper presents a neural network approach to document semantic indexing. A Hopfield net algorithm was used to simulate human associative memory for concept exploration in the domain of computer science and engineering. INSPEC, a collection of more than 320,000 document abstracts from leading journals, was used as the document testbed. Benchmark tests confirmed that three parameters (maximum number of activated nodes, ϵ – maximum allowable error, and maximum number of iterations) were useful in positively influencing network convergence behavior without negatively impacting central processing unit performance. Another series of benchmark tests was performed to determine the effectiveness of various filtering techniques in reducing the negative impact of noisy input terms. Preliminary user tests confirmed our expectation that the Hopfield net algorithm is potentially useful as an associative memory technique to improve document recall and precision by solving discrepancies between indexer vocabularies and end-user vocabularies.

1. Introduction

The effective and efficient use of information is no longer merely a strategic advantage for corporations

and individuals, it has become a necessity in the normal course of doing business. Quality information use begins with effective information storage, exploration and retrieval, which in turn depends on having an intelligent and highly efficient information indexing, searching and retrieval mechanism for the information source. Traditionally, the predominant information indexing and retrieval method has been keyword-based (using keyword indexes manually created by domain experts). Today, the vast amount of available information and the constant influx of new information have created a situation where the sheer volume of information overwhelms both the typical user and manual indexing methods. This phenomenon is known as 'information overload' [6].

To successfully index, store, locate and retrieve information, indexers and users need to know two things about the information space they are using. First, they need to have a working knowledge of the system where the information is stored; in particular, how to navigate through that information system. This requires an understanding of how the information is indexed, categorized or organized. Second, they must have subject or domain knowledge; in particular, the domain-specific vocabulary and domain-specific indexing terminology. Users with different levels of subject expertise and system familiarity [28] combine with the often imprecise nature of language to create what is known as the 'vocabulary problem' [11, 20, 27] also referred to as a semantic barrier [37].

Human indexing depends heavily on the domain knowledge that a given indexer possesses at a particular point in time, thus making it subject to human error and inconsistency. For example, previous research [4] has shown that well-trained individual indexers often assign different indexing terms to the same document (*synonymy*) and that the same indexer may use different terms for the same document at different times. Meanwhile, different users tend to use diverse terms to seek identical information (*polysemy*). Because of these discrepancies, an exact match

Correspondence to: A.L. Houston, Management Information Systems Department, 430 McClelland Hall, University of Arizona, Tucson, Arizona 85721, USA. Tel: +1 (520) 621 2748. Fax: +1 (520) 621 2433. E-mail: ahouston@bpaosf.bpa.arizona.edu

between a searcher's terms and an indexer's terms is unlikely, resulting in poor document recall and precision. The problems of polysemy (which reduces document precision) and synonymy (which reduces document recall) make it extremely difficult for novice users or for users searching in a field outside their domain knowledge to retrieve relevant information. Furthermore, manual indexing is too time-consuming for processing large volumes of information, or information that is volatile (i.e. the Internet).

To overcome the inefficiency of human indexing, a major effort has been put into developing automatic indexing techniques which can substantially accelerate information processing by increasing the volume of information indexed per unit of time [44]. However, most current automatic indexing programs use only keyword indexing – *fetching* keywords physically present in a document – and therefore do not solve the vocabulary problem. One way to address the vocabulary problem is to index documents semantically, permitting users to search by concept meanings as opposed to keywords.

In this research, a different approach was taken. An automatic indexing program [15] was enhanced by introducing an innovative parallel search algorithm (i.e. the Hopfield net algorithm [31]) as an associative memory for concept exploration in a neuron-like representation of the knowledge base. The Hopfield net is a parallel search algorithm which activates, directly as well as indirectly, terms associated with an initial set of concepts in the concept space. The net's spreading activation process can be used to identify a set of terms that are semantically relevant to the input terms.

The next section contains a literature review of semantic indexing. A framework for using a Hopfield net algorithm to semantically index the INSPEC testbed collection is presented in Section 3. Section 4 discusses the tests used to determine network convergence parameters and to refine network input terms. The results from a preliminary set of user evaluation experiments is presented in Section 5. Section 6 presents our conclusions and Section 7 discusses semantic interoperability, the ultimate goal of this research stream.

2. Semantic indexing

For decades, the 'vocabulary problem' has plagued information retrieval systems. In a 1987 study, Furnas *et al.* [28] showed that the probability of two people using the same term to classify an object was less than

20%. Individual terms are not adequate discriminators of semantic content, because the indexing relationship between words and document content (meaning) is many-to-many [3]. Various methods have been developed to address the vocabulary problem, most of which concentrate on solving synonymy by adding associative terms to keyword indexes. The major drawback of these methods is that the added terms often have multiple meanings that are different from the intended meanings, resulting in rapid precision degradation [24]. A common example of this technique is the domain expert indexing method that begins with a standard set of predetermined subject terms in an existing domain-specific thesaurus. Domain experts typically assign two to four of these terms to each document that they review. Accuracy is therefore heavily dependent on an indexer's expertise, both in the subject domain and with the domain-specific standard thesaurus. Domain expert indexing typically only works well in small domains with a limited number of documents and is too cumbersome and time-consuming to be used for processing large and varied collections. More effective and efficient techniques are needed to supplement or replace domain expert methods.

Other approaches to the vocabulary problem use either a thesaurus or a vector space representation (based on Salton *et al.*'s pioneering work [44]). Thesauri, which can be generated either manually or automatically, are mainly used to expand users' queries by translating query terms into alternative domain-specific standard indexing terms. Most vector space methods are based on syntax analysis and incorporate statistical analysis techniques (e.g. cluster analysis, co-occurrence efficiency analysis and factor analysis). These techniques attempt to represent relationships between documents, between terms and documents, and between terms and terms in mathematical matrices. Although vector space models have been quite successful in small domains with limited numbers of documents, computer resource constraints have made scaling up to large information spaces a challenge.

2.1. Thesaurus approach to semantic indexing

A variety of thesaurus structures apply different representation schemes to model terms (document descriptors) and their relationships. Three of those, and their major applications, are as follows.

- (1) **Hierarchical thesauri – term disambiguation into term senses:** term disambiguation addresses the

problems of polysemy and synonymy by using the context of words to determine meaning. The representation scheme of this type of thesaurus is tree-like, using IS-A notation [8] to represent relationships. Voorhees investigated replacing document terms and query terms with meanings or 'term senses' after automatically indexing a document collection [50]. WordNet, a thesaurus created at Princeton University by Miller [35], illustrates Voorhees' procedure. WordNet is a *manually* created hierarchy of sets of terms, with the primary relation between terms represented as an IS-A relationship.

Experience has demonstrated that it is difficult to disambiguate word senses in short query statements because they do not contain enough context. The IS-A relationship that typically represents a generalization/specialization hierarchy is not strong enough to consistently disambiguate word senses. Typically, tools based on this type of thesaurus reduce precision but fail to increase recall.

- (2) **Relational thesauri:** Wang, Vandendorpe and Evens created a number of relational thesauri to supplement semantically-based information retrieval queries [51]. Relational thesauri group words according to their relations with one another (i.e. parts-to-whole relations, collocation relations, paradigmatic relations, taxonomy and synonymy relations, antonymy relations, etc). The relationships are typically manually created term-to-term matrices of relations using lists of index terms from a collection of abstracts. Wang *et al.* used two non-parametric tests (the Sign Test and the Wilcoxon Signed Rank Test) to analyze results [51]. They determined that the best results with relational thesauri occurred when all but the antonymy relation were used together. Even though improved recall and precision was demonstrated, the scale of the experiment was limited (222 documents). The advantage of this technique is that it is truly based on semantic analysis. A disadvantage is that the thesauri must be generated *manually*.
- (3) **Automatic thesaurus creation:** as readily available computer processing and storage power has increased, research into automatically-created thesauri has also increased. Neural-like thesauri or concept spaces have generated the most interest. In a neural knowledge base, concepts or terms are represented as nodes and their relationships are represented as weighted links. This associative

memory feature has created a new paradigm for knowledge discovery using spreading activation algorithms such as the Hopfield net [16].

Virtually all techniques for automatic thesaurus generation are based on syntactic analysis using statistical word co-occurrence [12, 22, 41]. Coefficients between pairs of distinct terms are usually obtained based on coincidences in term assignments to the documents of a collection. Despite two decades of research, the usefulness of terms generated by co-occurrence analysis has not been clearly demonstrated. Some research has shown that co-occurrence terms produce poor document retrieval results [36, 38]. Other research has demonstrated significant improvements in document retrieval effectiveness [21]. Document recall improvements of the order of 10–20% have been demonstrated when a thesaurus is used in an environment similar to one in which the original thesaurus was constructed [22, 40, 42].

Recently, Chen *et al.* have experimented with parallel supercomputers to address the scalability issues related to large-scale information retrieval and have successfully created 575 fine-grained concept spaces for engineering communities [17]. The significance of using this automatic thesaurus creation method is its ability to create concept spaces that represent knowledge in a neuron-like network [2, 25, 29], allowing the potential use of neural network algorithms for knowledge discovery. In this research, a concept space (or thesaurus) was automatically created from a portion of INSPEC (a collection of citations and abstracts). Instead of using the concept space for interactively suggesting terms, a Hopfield net algorithm was used to traverse it automatically in search of concepts associated with document keyword indexing terms.

2.2. Multi-dimensional semantic space approach to semantic indexing

Multi-dimensional semantic space techniques go beyond Salton and McGill's 1983 term-vector representation, where two documents *must* share at least one term in order to be considered relevant [43]. The multi-dimensional semantic space methods relate documents (or document vectors) via their term meanings, not simply via document keywords. Two examples of multi-dimensional semantic space techniques are metric similarity modeling (MSM) and latent semantic indexing (LSI).

- (1) **Metric similarity modeling** uses a multi-dimensional semantic space where vectors represent both queries and documents, ensuring that common keywords are not the only measurement of document similarity. 'MSM allows semantic associations to directly determine the interpretation of terms and the representations of text in the multi-dimensional space' [3].

MSM uses techniques from multi-dimensional scaling, where 'objects are represented as points in a multi-dimensional space; points are chosen so that the inter-point similarities meet a set of externally imposed constraints on the similarities' [3]. MSM document vectors are computed using standard statistical techniques by Luhn [34] and Salton and McGill [43]. These vectors are then placed into the multi-dimensional semantic space, with their positions determined by similarity constraints.

One disadvantage of MSM is that it can only be applied when it is possible to use external sources to determine similarity constraints. Co-citation analysis, document classification information such as Library of Congress Subject Headings or Compindex Classification Codes and relevance feedback can be used to determine similarity constraints.

- (2) **Latent semantic indexing** is an optimal method of MSM. It represents documents, queries and terms as vectors in a matrix determined by multi-dimensional singular value decomposition. LSI assumes that a latent semantic structure of relationships exists within a document and uses this assumed structure to place documents which are similar to one another in close proximity in some type of 'space'. Most of the research on the application of latent semantic indexing has been done by Deerwester, Dumais, Furnas, Landauer and Harshman, who considered a number of alternative models and settled on a two-mode factor analysis based on singular value decomposition (SVD): 'SVD represents both terms and documents as vectors in a space of choosable dimensionality, and the dot product or cosine between points in the space gives their similarity' [24]. In previous information retrieval research, Borko and Bernick [7] used one-mode factor analysis to reduce a matrix of document terms to orthogonal factors which were then used to classify a document collection. The intuition behind this matrix transformation method is to generate a relatively densely distributed matrix from an original sparse matrix so that documents which share no

keywords may, nonetheless, be considered similar.

Deerwester *et al.* [24] experimented with LSI, using a rectangular matrix where the number of documents was between 1,000 and 2,000 and the number of terms was 5,000–7,000. They have tested LSI on two standard document collections, MED and CISI, with promising results in both document recall and precision. Their LSI results were at least as good as simple term matching, at least as good as SMART [9], and better than Voorhees' term disambiguation process [50].

3. A framework for semantic indexing

An automatically generated concept space has already been shown to be an effective tool for term suggestions in the CSQuest project (<http://ai.bpa.arizona.edu/CSQuest>). This research expands that technique by adding a Hopfield net to explore the INSPEC Concept Space automatically in search of a set of concepts most relevant to the document's keyword indexing terms.

Human indexing involves the use of human-accumulated knowledge for inferences and associations which exist as a set of interconnected neurons linked by synapses. Human brains function in an associative or content-addressable fashion (i.e. are capable of generating associative terms from any initial concept). In this research, a knowledge base was created that represented knowledge as a single layer of interconnected neurons (nodes) and weighted synapses (links) resembling neuron patterns in human brains. A Hopfield net algorithm was applied to activate neighboring nodes in the network [5, 48]. The basic steps for the semantic indexing program are as follows.

- (1) Identify a collection of documents in a specific domain to be indexed.
- (2) Analyze the co-occurrence probabilities of key indexing terms in this domain's documents and automatically create a concept space for them. Key indexing terms in this instance were generated from the INSPEC record fields of 'title' and 'abstract'. A key indexing term can consist of multiple word phrases (up to three words). Term thresholds (term frequency in a given document and term frequency in the overall collection) were used to determine if a key indexing term should be included in the concept space.
- (3) Apply a Hopfield net parallel search algorithm for concept exploration in the knowledge base [13].

Fig. 1 illustrates this semantic indexing process.

Fig. 1. From keyword indexing to semantic indexing.

3.1. INSPEC collection or database

In any automatic concept space building effort, the first task is to identify a collection of documents in a specific subject domain that can serve as the source for the vocabulary. It is extremely difficult to approach the semantic indexing problem using a large and complex information space because of heterogeneous subjects and domains. Therefore, this prototype focused on a very specific domain of the information space – computer science and computer engineering. The testbed was the INSPEC database, which is approximately 1.3 gigabytes and contains a collection of 328,181 journal research paper abstracts, most of which were published between 1992 and 1994.

3.2. INSPEC Concept Space

The INSPEC Concept Space is a computer-generated neural net knowledge base built from the INSPEC database using co-occurrence analysis. More than 328,181 documents in this database were used as vocabulary sources to produce over 270,000 unique concept space term nodes and over 10,000,000 relationship links. The result is a computer engineering concept space which consists of nodes connected by weighted links that represent weighted relationships between computer science concept terms. The total size of this concept space is more than 240 megabytes.

3.3. Hopfield net – a concept exploration method

The richness of semantically associative terms presented in a neural-like concept space enables users to get into the concept space easily and to explore and navigate it interactively. However, since humans typically use a *serial* search strategy, users can get

lost in a large information space. Serial searching behavior also leaves many promising paths unexplored. Identifying relevant concepts effectively and efficiently in large information spaces requires an intelligent method that can navigate *multiple* links in *parallel*. The Hopfield net algorithm is an excellent candidate for this kind of parallel searching [16].

Spreading activation, a memory association mechanism originating in human memory research, has been used successfully in neural net applications (e.g. image classification, character recognition, robotics and concept-based information retrieval [16]). The Hopfield net [31, 49] is a neural net that can be used as a content-addressable memory. Since knowledge can be stored in single-layered interconnected neurons (nodes) and weighted synapses (links) [48], information can be retrieved based on the network's parallel relaxation process. Nodes are activated in parallel and activation values from different sources are combined for each individual node (see Fig. 2) until the activation levels of nodes on the network reach a stable state (convergence). To accommodate the unique characteristics of semantic indexing, modifications (see [13] for details) were made to the original Hopfield net algorithm.

This research used the Hopfield net activation algorithm in the following manner.

- (1) **Initialization of the network with automatic indexing terms:** the automatic indexing program generates a set of key indexing terms for each document in the collection. In this case, the terms were automatically generated from the INSPEC record fields 'title' and 'abstract'. Co-occurrence analysis and term frequency thresholds were used to determine if the terms would be included in the concept space. Each key indexing term generated from the automatic indexing program was treated as a neuron in the network and the links were assigned a random weight value.
- (2) **Iterative activation and weight computation:** the formulas in Fig. 3 show the parallel relaxation property of the Hopfield net. At each iteration, nodes in the concept space are activated in parallel and activated values from different sources are combined for each individual node. Neighboring nodes are traversed in order until the activation levels of nodes on the network gradually 'die out' and the network reaches a stable state (convergence). The weight computation scheme ($net_i = \sum_{j=0}^n t_{ij} u_j(t)$) is unique to the Hopfield net algorithm. Each newly activated node computes its new weight based on the summation of the

Fig. 2. Hopfield net parallel relaxation process.

products of its neighboring nodes' weights and the similarity of its predecessor node to itself.

- (3) **Convergence condition:** the above process is repeated until there is no significant change in terms of output between two iterations, which is accomplished by checking the following formula:

$$\sum_{j=0}^{n-1} |u_j(t+1) - u_j(t)| \leq \epsilon$$

where ϵ is the maximum allowable error (used to indicate whether there is a significant difference between two iterations). Once the network con-

verges, the output represents the set of terms most relevant to the starting input terms.

3.4. Human indexing vs. automatic indexing

To study the effectiveness of semantic indexing using a Hopfield net, the results of human indexing and other methodologies were used as benchmarks for comparison and evaluation. This research compared five categories of indexes: keywords chosen manually from each document abstract; thesaurus terms picked from the manually created INSPEC thesaurus; keyword

$$u_i(t) = x_i, \quad 0 \leq i \leq n-1$$

$u_i(t)$ is the output of node i at time t . x_i (which has a value between 0 and 1) indicates the input pattern for node i .

$$u_j(t+1) = f_s \left[\sum_{i=0}^{n-1} t_{ij} u_i(t) \right], \quad 0 \leq j \leq n-1$$

where f_s is the continuous SIGMOID transformation function as shown below ([23, 33]):

$$f_s(net_j) = \frac{1}{1 + \exp \left[\frac{-(net_j - \theta_j)}{\theta_0} \right]}$$

where $net_j = \sum_{i=0}^{n-1} t_{ij} u_i(t)$, θ_j serves as a threshold or bias, and θ_0 is used to modify the shape of the SIGMOID function.

Fig. 3. Hopfield net parallel relaxation formulas.

indexing terms generated by the automatic indexing program, as well as directly related terms in the automatically generated INSPEC Concept Space, and the indirectly related terms suggested by the Hopfield net (refer to Fig. 4). The five categories are illustrated with the following example of an INSPEC document:

Document 4687935

Author : Chen, H, Nunamaker, J.F., Jr, Sprague, R.H., Jr

Title : A machine learning approach to document retrieval: an overview and an experiment

Journal : Los Alamitos, CA, USA : IEEE Comput. Soc. Press, Proceedings of the Twenty-Seventh Hawaii International Conference on, System Sciences. Vol.III: Information Systems: Decision Support and, Knowledge-Based Systems (Cat. No.94TH0607-2) | d(1994) | pp631-40

Abstract: We provide an overview of artificial intelligence techniques and then present a machine learning based document retrieval system we developed. GANNET (Genetic Algorithms and Neural Nets System) performed concept (keyword) optimization for user-selected documents during document retrieval using genetic algorithms. It then used the optimized concepts to perform concept exploration in a large network of related concepts through the Hopfield net parallel relaxation procedure. Our preliminary experiment showed that GANNET helped improve search recall by identifying the underlying concepts (keywords) which best describe the user-selected documents

- (1) **Human indexer terms (categories 650 and 651):** these terms are generated by professional indexers. There are two categories: category 650, which contains thesaurus terms manually chosen from

the manually created INSPEC thesaurus, and category 651, which consists of keywords manually selected from the documents. The terms for each category from the sample INSPEC document above are as follows:

CATEGORY 650 TERMS:

1. DATABASE MANAGEMENT SYSTEMS
2. GENETIC ALGORITHMS
3. HOPFIELD NEURAL NETS
4. LEARNING (ARTIFICIAL INTELLIGENCE)
5. QUERY PROCESSING

CATEGORY 651 TERMS:

1. MACHINE LEARNING APPROACH
2. DOCUMENT RETRIEVAL
3. ARTIFICIAL INTELLIGENCE TECHNIQUES
4. MACHINE LEARNING BASED DOCUMENT RETRIEVAL SYSTEM
5. GANNET
6. NEURAL NETS
7. CONCEPT KEYWORD OPTIMIZATION
8. KEYWORD OPTIMIZATION
9. USER-SELECTED DOCUMENTS
10. GENETIC ALGORITHM
11. CONCEPT EXPLORATION
12. HOPFIELD NET PARALLEL RELAXATION PROCEDURE
13. SEARCH RECALL

- (2) **Automatic keyword indexing program terms:** these terms are key indexing terms extracted from documents by the automatic indexing program. Single word terms are excluded because there are too many of them and user feedback indicated that, in general, their meanings are too generic (i.e. not specific enough) to be considered useful as document indexing terms. The automatic keyword indexing program terms for the sample INSPEC document above are:

MACHINE-GENERATED KEYWORD INDEXING TERMS:

1. PARALLEL RELAXATION PROCEDURE
2. NEURAL NETS SYSTEM
3. NET PARALLEL RELAXATION
4. HOPFIELD NET PARALLEL
5. DOCUMENT RETRIEVAL SYSTEM
6. ARTIFICIAL INTELLIGENCE TECHNIQUES
7. USER-SELECTED DOCUMENTS
8. SEARCH RECALL
9. RELAXATION PROCEDURE
10. PRELIMINARY EXPERIMENT
11. NEURAL NETS
12. MACHINE LEARNING

Fig. 4. Human indexing vs. automatic indexing.

13. KEYWORD OPTIMIZATION
14. HOPFIELD NET
15. GENETIC ALGORITHMS

(3) **Concept exploration – directly linked terms:** these terms come from the INSPEC Concept Space and represent concepts directly associated with the keyword indexing terms generated by the automatic indexing program. They are similar to related terms in a thesaurus and, as such, represent a set of terms that are semantically related to the input keyword indexing terms according to the INSPEC Concept Space. The set of terms in this category for the sample INSPEC document above are:

DIRECTLY ASSOCIATIVE TERMS FROM INSPEC
CONCEPT SPACE (SORTED BY WEIGHT):

1. |0.29313| INFORMATION RETRIEVAL SYSTEMS
2. |0.27411| HOPFIELD NEURAL NETS
3. |0.23803| RELAXATION THEORY
4. |0.22979| INFORMATION RETRIEVAL
5. |0.22700| FUZZY LOGIC
6. |0.20259| DOCUMENT RETRIEVAL
7. |0.20206| GENETIC ALGORITHM
8. |0.17221| ARTIFICIAL INTELLIGENCE
9. |0.16194| LEARNING (ARTIFICIAL INTELLIGENCE)
10. |0.15935| LEARNING SYSTEMS
11. |0.14322| INDEX TERMS
12. |0.13779| KIROUSIS, L.
13. |0.13524| INHERENTLY SEQUENTIAL
14. |0.12227| NEURAL NETS
15. |0.12158| NEURAL NETWORKS
16. |0.12153| CONSTRAINT SATISFACTION PROBLEM
17. |0.12073| RELAXATION
18. |0.12060| HINGSTON, P.
19. |0.12050| FULL-TEXT DATABASES
20. |0.11564| KNOWLEDGE ACQUISITION

(4) **Concept exploration – indirectly linked terms:** these terms are generated by using automatic indexing program terms as input to the Hopfield net which then in parallel activates neighboring nodes in the INSPEC Concept Space until the network converged. The final set of terms consisted of the network terms most semantically related to the input terms. This set of terms that was suggested by the Hopfield net algorithm after using the top seven automatic indexing terms from the sample document and allowing the net to iterate eight times is:

INPUT TERMS FOR HOPFIELD NET (Source: Automatic Indexing Program):

1. DOCUMENT RETRIEVAL SYSTEM
2. ARTIFICIAL INTELLIGENCE TECHNIQUES

3. RELAXATION PROCEDURE
4. NEURAL NETS
5. MACHINE LEARNING
6. HOPFIELD NET
7. GENETIC ALGORITHMS

TOP 20 TERMS OF ITERATION 8 ($\epsilon = 0.64$ # OF
ACTIVATED TERMS = 25):

1. |0.87690| INFORMATION RETRIEVAL
2. |0.81544| DOCUMENT RETRIEVAL
3. |0.79935| NATURAL LANGUAGE DOCUMENTS
4. |0.77354| DOCUMENT RETRIEVAL SYSTEM
5. |0.77503| NATURAL LANGUAGES
6. |0.77223| PATTERN MATCHER
7. |0.74429| INFORMATION RETRIEVAL SYSTEMS
8. |0.73760| NEURAL NETS
9. |0.73656| NEURAL NETWORK MODEL
10. |0.72664| STANDARD STRATEGIES
11. |0.71219| NEURAL NETWORK
12. |0.70533| NATURAL LANGUAGE QUERY
13. |0.67523| NEURAL NETWORKS
14. |0.64539| NATURAL LANGUAGE
15. |0.61330| COSINE MEASURE
16. |0.60220| VECTOR SPACE MODEL
17. |0.60082| NEURAL NETWORK MODEL
18. |0.59710| NATURAL LANGUAGE TEXT
19. |0.59292| COMPUTATIONAL LINGUISTICS
20. |0.58769| QUERY PROCESSING

4. System implementation

Like most neural network algorithms, the Hopfield net is computationally expensive and its performance is highly dependent on hardware configuration. The prototype system was developed in C on a SGI workstation running UNIX system V 4.0 (150 MHz with one gigabyte of memory). The critical issue in implementation was the trade-off between performance (speed) and memory (space usage). The relatively 'small' size of the prototype concept space allowed it to be left in the main memory during processing. The size of the INSPEC database testbed (over 1.3 gigabytes) made it necessary to pre-process the data files to reduce disk access time (sequential searching would have been impossibly slow in real time). Prototype performance was improved by indexing the testbed file, loading the index into memory, and performing a binary search on the index file.

Two major challenges are involved in use of a Hopfield net search algorithm for semantic indexing: selecting parameters that can effectively control the network search behavior (in particular, convergence behavior) and refining initial input to overcome the

problem of noisy input. Benchmark testing was performed to determine the parameter configurations which best yield consistent system performance and to determine the best technique for refining network input to filter out possible noisy terms.

Partial input terms tend to have more generic meaning than do the complete terms, and appear to be more likely to survive and dominate the network search behavior, probably because generic terms tend to have more neighbors with higher weighted links than more specific terms. In some extreme cases, too generic an input term can lead the Hopfield net to search outside the mainstream of the initial set of concepts. This can be illustrated with the following sample document:

Document 4190999

Author : George, J.F, Nunamaker, J.F., Jr, Valacich, J.S

Title : Electronic meeting systems as innovation: a study of the innovation

Journal : Information and Management | vvol.22, no.3 | d(March 1992) | pp187-95

Abstract: Electronic meeting systems (EMS) are slowly moving out of university environments into work organizations. They constitute an innovative method of supporting group meetings. The authors report on the innovation process in one organization that recently adopted and implemented an EMS. They trace the innovation process through four stages: conception of an idea; proposal; decision to adopt; and implementation. Important factors from the innovation literature are considered as an explanation of the innovation process involving EMS in this particular organization (Indian Health Service in US)

From the above document, the automatic indexing program generated the following list of key indexing terms:

1. INDIAN HEALTH SERVICE
2. ELECTRONIC MEETING SYSTEMS
3. UNIVERSITY ENVIRONMENTS
4. SLOWLY MOVING
5. PARTICULAR ORGANIZATION
6. INVOLVING EMS
7. INNOVATION LITERATURE
8. HEALTH SERVICE

The highlighted terms existed in the automatically generated INSPEC Concept Space and served as input to the Hopfield net network for concept exploration. *Health service* proved to be a noisy term, with many highly weighted links to its neighbors. This term soon

dominated the network search behavior, misleading the network into producing a final set of terms in the *health care* domain. After seven iterations of spreading activation process, the system suggested the following associative concepts along with their weights which subjects rates as irrelevant to the initial sample document:

1. |0.92367| HEALTH CARE
2. |0.87074| HEALTH SERVICE
3. |0.85545| CLINICAL ACTIVITIES
4. |0.83459| CLINICAL CARE
5. |0.83454| MEDICAL ADMINISTRATIVE DATA PROCESSING
6. |0.77699| CLINICAL STAFF
7. |0.76236| COMPLEX KNOWLEDGE BASE
8. |0.74955| CLASSIFICATION STRUCTURE
9. |0.74582| OO APPROACH
10. |0.74422| RECURSIVE MANNER
11. |0.74022| MEDICAL COMPUTING
12. |0.72347| GENERIC MODEL
13. |0.70577| OBJECT-ORIENTED PROGRAMMING
14. |0.67990| HEALTH SERVICES
15. |0.67908| HEALTH CARE PROFESSIONALS
16. |0.66852| DISCRETE EVENT SIMULATION MODELS
17. |0.66446| PATIENT FLOW
18. |0.66226| EQUALLY SPACED INTERVALS
19. |0.65703| COMMUNICATION GAP
20. |0.65703| CLINICAL SITUATIONS

4.1. Parameter configuration tests

The major research objective for this set of experiments was to create a system capable of generating an adequate number of associative concepts while at the same time having the network converge within a limited number of iterations in order to conserve central processing unit (CPU) cycles. Controlling network search behavior is essential for two reasons. First, a network that does not converge in a limited number of iterations consumes too many CPU cycles. Second, long distance traversal in the concept space often produces a set of final terms that are semantically far away from the intended meanings of the original set of concepts.

A series of benchmark tests was run on candidate parameters. Both convergence behavior and CPU utilization were recorded. After analyzing the benchmark test results, three parameters were chosen for controlling network convergence behavior: maximum number of activated nodes, ϵ (used to indicate when the network stabilized) and maximum number of iterations. Test results indicated that the maximum number of terms allowed for each iteration was the most effective

control parameter, but it could not guarantee network convergence in a limited number of iterations and had to be discarded. A limit for maximum number of iterations was set, which forced the network to stop processing if necessary. Benchmarking results indicated that the appropriate value for maximum number of nodes should be between 25 and 50, and ϵ should have a value between one and three. These parameter values produced network convergence within ten iterations in most cases. When 20 was chosen as the default value for maximum number of iterations, benchmark results further indicated that, for these parameter settings, the network converged in less than one second of CPU time in most cases. This level of performance is tolerable for real-time situations.

4.2. Input refinement

Two major factors can have an effect on the quality of Hopfield net semantic indexing results: the concept space that represents knowledge in a neural network and the initial input to the network for concept exploration. Like human memory, the machine-generated knowledge base does not always represent knowledge accurately. It is inevitable that some mistakes in knowledge representation will cause the network to traverse inappropriate paths in the concept space. Noisy initial input can also substantially affect the result. In most cases, noisy terms gradually die out after several iterations. Unfortunately, in other cases, noisy terms survive and end up influencing network search behavior.

In an attempt to improve the indexing result, another set of benchmark tests aimed at refining and filtering input was designed. More than 30 INSPEC documents (covering neural network, electronic meeting systems, human computer interactions and database management systems) were used as the testbed. For each case, keyword indexing terms generated by the automatic indexing program were used as input to the Hopfield network (single terms were excluded). Many of the input terms used in the testing were related to each other as concept and sub-concept (e.g. *business process re-engineering* and *process re-engineering*). Three filters were tested: using all terms generated by the automatic indexing (essentially no filter), using only complete terms (eliminating all partial terms), and using a limited filter (keeping some of the partial terms). Benchmark testing indicated that a limited filter produced the best results. Complete concepts were still used as often as possible, but if a complete term did not exist in the concept space then

one of its existing partial terms was used instead. A wider exploration space was therefore available for searching, while starting terms were kept as specific as possible.

5. System experiment and evaluation

5.1. Experimental design

The goal of this experiment was to let users judge the *quality* of terms suggested by the Hopfield net algorithm in order to assess whether the Hopfield net can substantially help to identify more relevant terms by exploring associative concepts in the knowledge base. Term (as opposed to document) recall and precision were used as performance measures.

Eight subjects, all graduate MIS students considered knowledgeable in the subject domain, were chosen to evaluate the output from a Hopfield net application. The testbed consisted of 20 documents pertaining to machine learning algorithms, human-computer interaction, electronic meeting system or database management systems for a total of 160 comparisons. These documents were abstracts from the INSPEC collection of journal research papers published between 1991 and 1994. The experiment consisted of two phases. Phase one determined subject recall. Subjects were presented with the title and abstract from each of the 20 documents. They were asked to read each document carefully and then to generate as many related concepts or terms as possible through free association.

Phase two of the experiment involved the subjects' judgment of a list of suggested terms and their relevance to a given document for each of the 20 documents. Terms were generated in five different ways:

- (1) chosen by professional indexers from the manually created thesaurus (650);
- (2) manually picked from each INSPEC abstract (651);
- (3) produced by the automatic indexing program (AUT);
- (4) first-level associative terms selected from an automatically created computer science concept space (or thesaurus) (THE); and
- (5) terms suggested by the Hopfield net (HOP).

The five types of terms were integrated by removing duplicates and were presented as an alphabetized list. Subjects were asked to evaluate the relevance of each term to the content of a given document using a Likert-like scale: 'Irrelevant', 'Somewhat Relevant' and 'Very Relevant'. Terms considered too general were to be classified as 'Irrelevant'.

Term Recall: ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
RECALL	6	28.3270	4.7212	346.44	0.000
ERROR	1099	14.9766	0.0136		
TOTAL	1105	43.3036

Fig. 5. ANOVA analysis of recall.

Term Precision: ANALYSIS OF VARIANCE

SOURCE	DF	SS	MS	F	P
Precision	6	7.5032	1.2505	17.44	0.000
ERROR	1099	78.7950	0.0717		
TOTAL	1105	86.2982

Fig. 6. ANOVA analysis of precision.

Two broader term categories were created to compare manually generated terms (i.e. both sets of INPSEC terms – category 650 and category 651) with automatically generated terms (automatic keyword indexing

terms (*AUT*), first layer concept space terms (*THE*) and Hopfield net terms (*HOP*)). In Figs. 5 to 8, these two categories are denoted by 'INS' (manually generated terms) and 'US' (automatically generated terms).

Fig. 7. Comparison of recall.

5.2. Experimental results

Term recall was defined as the portion of the target list that was generated by each of the five term categories. Term precision was defined as the portion of each list that appeared in the final target list. Overall, the manually generated term category (*INS*) had greater precision than the automatically generated term category (*US*), but the automatically generated term category had greater recall. Recall for automatically generated terms was about 66%, while recall for manually generated terms was approximately 34% (see Figs. 5 and 7). In contrast, the precision for the automatically generated category was 80%, whereas precision for manually generated terms was about 64% (see Figs. 6 and 8). This difference was statistically significant.

The manually generated category (*INS*) had recall values of 10% for category 650 and 25% for category 651 (see Figs. 5 and 7). The low recall values may indicate that human beings' ability to independently generate associative terms is limited. However, the precision values for both categories were similar: 80% and 81% (see Figs. 6 and 8).

Within the automatically generated category, recall scores for automatic keyword indexing (*AUT*), the INSPEC Concept Space (*THE*) and Hopfield net category (*HOP*) were 23%, 30% and 31% respectively. Higher recall values for both the Hopfield net and the concept space categories indicate that both of those methods generated more meaningful and relevant concepts than the automatic keyword indexing technique.

Fig. 8. Comparison of precision.

The automatic keyword indexing (*AUT*) category had the highest precision (72%) among the three sub-categories (60% for the concept space (*THE*) and 65% for the Hopfield net (*HOP*)). This could have been due to the fact that this category contained keywords extracted directly from the document.

Overall, the human beings appeared to generate more precise indexing terms than those produced automatically. However, automatic term generation outperformed manual term generation by producing more meaningful concept associations without adversely affecting precision. In fact, the precision of automatic term generation was very high in terms of its absolute value (66%) although lower than that of manual term generation (80%). This is reasonable, as higher recall inevitably leads to the introduction of more irrelevant terms.

The term overlap among the three automatic methods (*AUT*, *THE* and *HOP*) was small (see Fig. 5). This implies that the Hopfield net did help to generate more diversified terms and significantly increased the variety of expressions for the same concept. As both the recall and precision of the concept space category (*THE*) and the Hopfield net category (*HOP*) were similar, it probably would be necessary to use both categories' terms for improved concept exploration. Figs. 7 and 8 illustrate the precision and recall values for each method.

Overall, we believe this research has provided *insights* concerning developing a robust and *intelligent* network knowledge management and inferencing

system. An algorithmic approach to concept exploration in a large network (Hopfield net) can be performed under real-time computation constraints and could *potentially* be useful for developing a semantic indexing system. More extensive research is required to confirm this belief.

6. Conclusions

This research presents a framework for semantic indexing in a large-scale knowledge network using a Hopfield net and an example of an implementation on the INSPEC collection. Using initial keywords generated by the automatic indexing program for a given document, the Hopfield net is designed to simulate human associative memory functions when exploring a large knowledge base in search of related concepts. We believe that using a Hopfield net for concept exploration can improve the quality of machine information processing and substantially enrich an indexing vocabulary, thereby representing the knowledge incorporated in a database more accurately.

The Hopfield net algorithm was tested in an application using a single large knowledge base for concept exploration. The results indicate that using a Hopfield net to automatically traverse multiple promising paths in parallel is an improvement over direct thesaurus checking for semantic indexing. A future application of automatic semantic indexing might involve embedding *intelligence* into the system in order to allow it to choose correctly the most promising paths for further traversing.

Another future application might involve assisting indexers in improving efficiency by interactively indexing information. It is difficult to dynamically envision the parallel paths taken during the generation of a final set of terms. A user interface that can help users to visualize different paths leading toward a final term and allow their dynamic intervention in the spreading and activating processes should be very useful. The essential issue in visualizing the Hopfield net activating and spreading processes is to present *all* the different paths in an intuitive and concrete manner: a complicated issue when there are hundreds of different paths in the network. Applying visualization techniques to user interface designs is an interesting current research issue [30, 39].

7. Semantic interoperability: the ultimate goal

While this research was limited to a single subject domain, the ultimate research goal is to achieve semantic indexing on multiple domains with multiple knowledge bases. Many recent information retrieval research projects have adopted multiple existing thesauri for term suggestion. Chamis discussed the issues of thesaurus compatibility and system strategies that might be developed to overcome difficulties in searching multiple incompatible databases [10]. In particular, she described the effectiveness of the vocabulary switching system (VSS) [10], an integrated vocabulary consisting of twelve existing thesauri in four diverse subjects areas (business, social science, life science, physical sciences). Knapp's BRS/TERM vocabulary database [32] maps natural language synonyms and controlled vocabulary descriptors from seven bibliographic databases in the social and behavioral sciences.

Several research projects have attempted to incorporate existing thesauri in the design of knowledge-based information retrieval systems. Fox *et al.* [26] focused on creation of so-called 'relational thesauri'. Ahlswede and Evens parsed *Webster's Seventh New Collegiate Dictionary* to obtain a 'lexical database' containing lexical or lexical-semantic relationships from dictionary definitions [1]. Chen *et al.* [18, 19] conducted a series of experiments which included several large-scale, domain-specific thesauri. In [16], a portion of the Library of Congress Subject Headings in the computing area was incorporated into a system that used branch-and-bound and spreading activation algorithms to assist users in query formulation. More recently, Chen *et al.* [14] developed a concept-based document retrieval system using multiple thesauri.

Several research groups recently have experimented with an algorithmic approach to cross-domain term switching. Chen *et al.* are investigating the generation, integration, and activation of multiple thesauri (including both existing thesauri and automatically generated thesauri in computing-related areas) [13]. Both Kim and Chen have proposed treating thesauri as neural networks and applying spreading activation algorithms for term-switching. We believe that automatically creating robust and useful *domain-specific* thesauri is feasible and can potentially pave the way for cross-domain scientific information retrieval. The crucial issue for semantic indexing across multiple domains is to find a path that will connect different domains.

The compatibility of multiple domains and the establishment of paths that cross over different domains for concept exploration based on some overlap among multiple knowledge bases are two critical issues that need to be addressed. In [14], Chen and his team experimented with using two automatic thesauri created for worm and fly biology for cross-domain concept-based retrieval. The first experiment was aimed at understanding fly-worm biologists' cross-domain term association patterns and their similarity to the terms and associations represented by the fly-worm thesaurus. The second experiment involved implementing the team's thesauri on the operational Worm Community System and investigating the retrieval performances of fly biologists when using fly terms to retrieve worm documents. Their research demonstrated promise for accessing the same quality of information by using different queries specific to different domains.

The NSF/DARPA/NASA-funded Digital Library Initiative Project [46, 47] includes a multiple-year plan to attack semantic interoperability that concentrates on scalable semantics across objects, domains and types in order to enable users to search information on unfamiliar subjects by inputting terms from their (known) domain and have the system determine semantically similar terms in the target (unknown) domain [45]. For document text, there is a particular need to develop new algorithms for vocabulary switching across information spaces, using deeper mathematics for space intersection. Techniques may include neural nets, genetic algorithms and differential manifolds. Using the Hopfield net algorithm as an associative memory is a good candidate for traversing a single knowledge base and has potential to be used in a multiple cross-over domain environment where overlap between domains can serve as a bridge for term switching.

Acknowledgements

This project was supported mainly by the following grants:

- (1) NSF/ARPA/NASA Digital Library Initiative, IRI-9411318, 1994-1998 (B. Schatz, H. Chen *et al.*, *Building the Interspace: Digital Library Infrastructure for a University Engineering Community*);
- (2) NSF CISE, IRI-9525790, 1995-1998 (H. Chen, *Concept-based Categorization and Search on Internet: A Machine Learning, Parallel Computing Approach*);

- (3) NSF CISE Research Initiation Award, IRI-9211418, 1992-1994 (H. Chen, *Building a Concept Space for an Electronic Community System*).

References

- [1] T. Ahlswede and M. Evens, Generating a relational lexicon from a machine-readable dictionary, *International Journal of Lexicography* 1(3) (1988) 214-237.
- [2] J.R. Anderson, *The Architecture of Cognition* (Harvard University Press, Cambridge, MA, 1983).
- [3] B.T. Bartell, G.W. Cottrell and R.K. Belew, Representing documents using an explicit model of their similarities, *Journal of the American Society for Information Science* 46(4) (1995) 254-271.
- [4] M.J. Bates, Subject access in online catalogs: a design model, *Journal of the American Society for Information Science* 37(6) (1986) 357-376.
- [5] W. Bechtel and A. Abrahamsen, *Connectionism and the Mind: An Introduction to Parallel Processing in Networks* (Basil Blackwell, Cambridge, MA, 1991).
- [6] D.C. Blair and M.E. Maron, An evaluation of retrieval effectiveness for a full-text document-retrieval system, *Communications of the ACM* 28(3) (1985) 289-299.
- [7] H. Borko and M. Bernick, Automatic document classification, *Journal of the ACM* 10 (1963) 151-162.
- [8] R.J. Brachman, What IS-A is and isn't: an analysis of taxonomic links in semantic networks, *IEEE COMPUTER* (October 1983) 30-36.
- [9] C. Buckley, G. Salton, J. Allan and A. Singhal, Automatic query expansion using SMART: TREC 3. In: *Text Retrieval Conference (TREC-3)*, Bethesda, MD, November 2-4 1995, pp. 69-80.
- [10] A.Y. Chamis, *Vocabulary Control and Search Strategies in Online Searching* (Greenwood Press, New York, 1991).
- [11] H. Chen, Collaborative systems: solving the vocabulary problem, *IEEE COMPUTER* 27(5) (1994) 58-66. [Special issue on computer-supported cooperative work (CSCW).]
- [12] H. Chen and K.J. Lynch, Automatic construction of networks of concepts characterizing document databases, *IEEE Transactions on Systems, Man and Cybernetics* 22(5) (1992) 885-902.
- [13] H. Chen, K.J. Lynch, K. Basu and D.T. Ng, Generating, integrating, and activating thesauri for concept-based document retrieval, *IEEE EXPERT, Special Series on Artificial Intelligence in Text-based Information Systems* 8(2) (1993) 25-34.
- [14] H. Chen, J. Martinez, D.T. Ng and B.R. Schatz, A concept space approach to addressing the vocabulary problem in scientific information retrieval: an experiment on the Worm Community System, *Journal of the American Society for Information Science* 48(1) (1997) 157-170.

- [15] H. Chen, J.P. Martinez, A.J. Kirchhoff, D.T. Ng and B.R. Schatz, Alleviating search uncertainty through concept associations: automatic indexing, co-occurrence analysis, and parallel computing, *Journal of the American Society for Information Science* (forthcoming 1998).
- [16] H. Chen and D.T. Ng, An algorithmic approach to concept exploration in a large knowledge network (automatic thesaurus consultation): symbolic branch-and-bound vs. connectionist Hopfield net activation, *Journal of the American Society for Information Science* 46(5) (1995) 348–369.
- [17] H. Chen, B.R. Schatz, D.T. Ng and M-S. Yang, *Breaking the Semantic Barrier: A Concept Space Experiment on the Convex Exemplar Parallel Supercomputers: Technical Report* (Department of MIS, College of Business and Public Administration, University of Arizona, Working Paper, March 1996).
- [18] H. Chen, B.R. Schatz, D.T. Ng, J.P. Martinez, A.J. Kirchhoff and C. Lin, A parallel computing approach to creating engineering concept spaces for semantic retrieval: the Illinois Digital Library Initiative Project, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(8) (1996) 771–782.
- [19] H. Chen, B.R. Schatz, T. Yim and D. Fye, Automatic thesaurus generation for an electronic community system, *Journal of the American Society for Information Science* 46(3) (1995) 175–193.
- [20] J. Courteau, Genome databases, *Science* 254 (1991) 201–207.
- [21] W.B. Croft and R. Das, Experiments with query acquisition and use in document retrieval systems. In: *Proceedings of the 13th Conference on Research and Development in Information Retrieval, Brussels, Belgium, 5–7 September 1990*, pp. 349–365.
- [22] C.J. Crouch, An approach to the automatic construction of global thesauri, *Information Processing and Management* 26(5) (1990) 629–640.
- [23] J. Dalton and A. Deshmane, Artificial neural networks, *IEEE Potentials* 10(2) (1991) 33–36.
- [24] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer and R. Harshman, Indexing by latent semantic analysis, *Journal of the American Society for Information Science* 41(6) (1990) 391–407.
- [25] S.E. Fahlman, *NETL: A System for Representing and Using Real-World Knowledge* (MIT Press, Cambridge, MA, 1979).
- [26] E.A. Fox, J.T. Nutter, T. Ahlswede, M. Evens and J. Markowitz, Building a large thesaurus for information retrieval. In: *Second Conference on Applied Natural Language Processing, Association for Computational Linguistics*, edited by B. Ballard (Bell Communications Research, Morristown, NJ, 1988), pp. 101–108.
- [27] K.A. Frenkel, The human genome project and informatics, *Communications of the ACM* 34(11) (1991) 41–51.
- [28] G.W. Furnas, T.K. Landauer, L.M. Gomez and S.T. Dumais, The vocabulary problem in human-system communication, *Communications of the ACM* 30(11) (1987) 964–971.
- [29] S.I. Gallant, Connectionist expert system, *Communications of the ACM* 31(2) (1988) 152–169.
- [30] N. Gershon, From perception to visualization. In: *Scientific Visualization: Advances and Challenges* (Academic Press and Harcourt Brace, 1994).
- [31] J.J. Hopfield, Neural network and physical systems with collective computational abilities, *Proceedings of the National Academy of Science, USA* 79(4) (1982) 2554–2558.
- [32] S.D. Knapp, Creating BRS/TERM, a vocabulary database for searchers, *DATABASE* 7(4) (1984) 70–75.
- [33] K. Knight, Connectionist ideas and algorithms, *Communications of the ACM* 33(11) (1990) 59–74.
- [34] H.P. Luhn, The automatic creation of literature abstracts, *IBM Journal of Research and Development* 2(2) (1958) 159–165.
- [35] G. Miller, WordNet: an on-line lexical database, *International Journal of Lexicography* 3(4) (1990). [Special issue.]
- [36] J. Minker, G.A. Wilson and B.H. Zimmerman, An evaluation of query expansion by the addition of clustered terms for a document retrieval system, *Information Storage and Retrieval* 8 (1972) 329–348.
- [37] S. Nadis, Computation cracks semantic barrier between databases, *Science* 272 (1996) 1419.
- [38] H.J. Peat and P. Willett, The limitations of term co-occurrence data for query expansion in document retrieval systems, *Journal of the American Society for Information Science* 42(5) (1991) 378–383.
- [39] R. Rau, J. Pedersen, M. Hearst, J. Mackinlay, S.K. Card, L. Masinter, P. Halvorsen and G. Robertson, Rich interaction in the digital library, *Communications of the ACM* (April 1995) 29–39.
- [40] G. Salton, Automatic thesaurus construction for information retrieval, *Information Processing* 71 (1972) 115–123.
- [41] G. Salton, *Automatic Text Processing* (Addison-Wesley, Reading, MA, 1989).
- [42] G. Salton and M.E. Lesk, Information analysis and dictionary construction. In: G. Salton (ed.), *The Smart Retrieval System – Experiments in Automatic Document Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1971), pp. 115–142.
- [43] G. Salton and M. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).
- [44] G. Salton, A. Wong and C.S. Yang, A vector space model for automatic indexing, *Communications of the ACM* 18(11) (1975) 613–620.
- [45] B.R. Schatz and H. Chen, Building large-scale digital libraries, *IEEE COMPUTER* 29(5) (1996) 22–26.
- [46] B.R. Schatz, B. Mischo, T. Cole, J. Hardin, A. Bishop and H. Chen, Federating repositories of scientific literature, *IEEE COMPUTER* 29(5) (1996) 28–36.

- [47] T.R. Smith, A digital library for geographically referenced materials, *IEEE COMPUTER* 29(5) (1996) 54–60.
- [48] J.F. Sowa, Issues in knowledge representation. In: J.F. Sowa (ed.), *Principles of Semantic Networks: Explorations in the Representation of Knowledge* (Morgan Kaufman, San Mateo, CA, 1991), pp. 1–12.
- [49] D.W. Tank and J.J. Hopfield, Collective computation in neuronlike circuits, *Scientific American* 257(6) (1987) 104–114.
- [50] E.M. Voorhees, Using WordNet to disambiguate word senses for text retrieval. In: *Proceedings of the ACM SIGIR, Pittsburgh, PA, 1993*, pp. 171–180.
- [51] Y-C. Wang, J. Vandendorpe and M. Evens, Relational thesauri in information retrieval, *Journal of the American Society for Information Science* 36(1) (1985) 15–27.