



Pergamon

Library & Information Science Research  
24 (2002) 235–250

**Library &  
Information  
Science  
Research**

## Automatic extraction of relationships between terms by means of Kohonen's algorithm

Vicente P. Guerrero<sup>a,\*</sup>, Félix Moya-Anegón<sup>b</sup>, Victor Herrero-Solana<sup>b</sup>

<sup>a</sup>*Facultad de Biblioteconomía y Documentación (Alcazaba de Badajoz),  
Universidad de Extremadura, 06071 Badajoz, Spain*

*E-mail address: vicente@alcazaba.unex.es (V.P. Guerrero)*

<sup>b</sup>*Facultad de Biblioteconomía y Documentación, Universidad de Granada, Colegio Máximo de la Cartuja,  
18071 Granada, Spain*

---

### Abstract

This article describes a method of finding the contextual relationships among different terms in a database. First, the vector model is used to represent the terms as vectors according to which documents they appear in. Second, these vectors are used as the input to a Kohonen network, which organizes them topologically. This organization, in turn, generates term clusters arranged on a grid, so that each term is not only related to the others in its own cluster but also to those of neighboring clusters. © 2002 Elsevier Science Inc. All rights reserved.

---

The exponential growth of information has been a topic of concern and interest for more than 25 years (Price, 1973). The availability of electronic information and the process of digitalization have contributed in large part to this growth, with the transformation of documents “based on atoms to ones based on bits” (Negroponte, 1995). Computer use is, of course, not restricted to editorial production; it is present in all aspects of life—from the workplace, where there is often one computer per person, each of whom is generating new documents, to the home, where, increasingly, people not only have a computer but multimedia equipment as well. There is also the distribution of information via the so-called information highway and the effect of the increasingly lower cost of storage media. One is thus in the midst of a developing environment of electronic information that can be accessed automatically. Another consideration is the diversification of media, which has the collateral effect of producing a greater amount of nonnormalized information (e.g., images, sound, and text).

---

\* Corresponding author.

There is a need for automatic procedures that allow such an immense quantity of information to be processed. One of these procedures is the automatic classification of the terms that are present in a database. This type of automatically generated thesaurus or “concept space approach” (Chen, Ng, Martinez, & Schatz, 1997) is in great demand for databases with very specific scientific content, as an aid to searching, browsing, or simply representing the contents of a document database, or of results. There are two ways to put this procedure into practice:

*Independently of the corpus:* These relationships depend on the language in which the documents are written. There have been studies ranging from the construction of a thesaurus to the processing of natural language. Vocabulary networks (WordNet) have also been used to form a database (Miller et al., 1990; Voorhees, 1993). With respect to performance, however, the results have not been as expected. As with artificial intelligence, specific relationships are required for the domain of the database.

*With the aid of the corpus:* In this case, the relationships are developed from the information contained in the database. The result will be specific relationships in the domain of the corpus. This can be carried out by evaluating the following:

The *lexical similarity* of the words, employing techniques of: stemming (words with the same stem are considered to be related), suffix elimination (words are related after eliminating suffixes or prefixes), or n-grams (relating words with similar pieces). This type of technique gives good results if the terms in the corpus have a strong lexical relationship (e.g., those relating to chemistry and pharmaceutical products; Oakes & Taylor, 1999).

The *semantic similarity*, which can only be done using co-occurrence studies. The bases for the technique are described in Van Rijsbergen (1977) and Peat and Willett (1991). Natural-language processing techniques have been used in some cases (Strzalkowski, 1995), and neural networks have begun to be used (Chen, Houston, Sewell, & Schatz, 1998; Honkela, Pulkki, & Kohonen, 1995; Kaski, 1999; Kohonen et al., 1999; Lagus, Honkela, Kaski, & Kohonen, 1996; Lagus & Kaski, 1999; Lagus, Kaski, Honkela, & Kohonen, 1999; Lin, 1997; Moya-Anegón et al., 1999; Moya-Anegón, Herrero-Solana, & Guerrero, 1998; Muñoz Garcia, 1994; Orwig, Chen, & Nunamaker, 1997; White, Lin, & McCain, 1998; Wong, Cai, & Yao, 1993).

This article describes an automatic procedure that uses a document corpus and allows one to find the contextual semantic relationships between the terms. These relationships will only be valid for that corpus.

## 1. Method

The study method is divided into two parts: First, the classification is made by clustering technique, whether hierarchical or heuristic (Moya-Anegón, 1994; Salton & McGill, 1983). Second, the objects to classify need to have a vector representation, or, similarly, a representation as lists of characteristics.

Artificial neural networks are currently attracting particular interest as mechanisms for performing certain basic intellectual tasks. These networks were designed to emulate how the brain functions, learning to assign multidimensional outputs to multidimensional inputs, which they do with a great capacity for generalization.

The training of the neural networks may be supervised or unsupervised (according to the type of network). For the supervised training, one uses a series of pairs formed by an input and its corresponding output, so that the network gradually adapts to the desired outputs on the basis of the mistakes that it makes. In unsupervised training, however, only the inputs are presented (without their corresponding outputs) for the network to cluster. These clusters have certain advantages over their classic counterparts. For example, they are adaptive—they do not have to be provided with parameters of size or overlap, but, to a great degree, they adapt to the database in question. The neural networks therefore learn from the objects they have to process. One consequence of this is that the clusters are closer together where there is a greater density of terms. In addition, the zone that a cluster covers may expand or contract according to this density. To this end, the vectors are compared globally to allow terms to be included in a cluster, even though they do not exactly appear in the same documents.

One of the models that has been used for the classification of terms is the Hopfield (1982) network. This model is designed to assist the user by memorizing or generating thesauri or heading lists, so as to subsequently suggest alternative terms and expand or rephrase the query (Chen, 1995; Chen & Lynch, 1992; Chen, Lynch, Basu, & Ng, 1993; Chen & Ng, 1995; Chen, Schatz, Yim, & Fye, 1995; Chen et al., 1997).

Teuvo Kohonen introduced the Kohonen model for the classification of terms according to their syntactic function in texts (Honkela et al., 1995; Kaski, 1999; Kohonen et al., 1999; Lagus et al., 1996, 1999; Lagus & Kaski, 1999; Ritter & Kohonen, 1989) by using the information derived from the previous and the following terms. This model has also been used for the classification of a database's keywords (Moya-Anegón et al., 1998, 1999). The Kohonen model allows both clustering and topological classification (Guerrero, Moya-Anegón, & Herrero-Solana, 2002; White et al., 1998). It can be applied to the semantic classification of terms, for which purpose it is necessary to study a procedure to select representative terms of the database and their vectorial representation.

Recently, Kohonen networks have been used to generate topological maps of a set of documents, even labeling the zones of influence of each word or term (Chen et al., 1998; Honkela et al., 1995; Kaski, 1999; Kohonen et al., 1999; Lagus et al., 1996, 1999; Lagus & Kaski, 1999; Lin, 1997; Lin, Soergei, & Marchionini, 1991), and have also been used in automated domain analysis (White et al., 1998).

### *1.1. Selection and representation of terms*

This study's objective is not to test the Kohonen algorithm, which has been used for similar purposes on other occasions, but to study a general process with bases that are closer to those that have been used in information retrieval and which therefore results in semantic classifications.

To test the capacities of this algorithm under conditions as general and close to reality as possible, the authors generated a test-bed database; they chose an English-language database because there is no broadly accepted stemming algorithm in Spanish. The authors therefore decided to use records from the bibliographic database *Library and Information Science Abstracts* (LISA) to form their test bed. Each of the summaries in the records were treated as independent documents; the remaining fields were ignored. To maintain as great a generality as possible, instead of performing a retrieval by topic, the authors selected the last 954 records of LISA. The retrieval was then performed by selecting the references that had an accession number greater than or equal to 9,605,000. The database generated then consisted of the previous 954 records, which contained 7,758 different words. This number of terms in the database was unmanageable for the later stage of representation and study, so the authors had to reduce the number until they could be satisfied that they had the most representative terms of the database. For this purpose, they used the Salton discrimination value (Salton & McGill, 1983).

After creating the document database, the next step was to transform the documents into vectors to calculate the cited discrimination value of the terms. The authors applied the Vector Space Model, which transforms each document into a vector.

The authors chose a weighting scheme that is very close to the classic Inverse Document Frequency (IDF) method (denoted  $tf \cdot idf$ ), and which was among those weighting schemes that gave the best results in Noreault, McGill, and Koll's (1981) study. The weight of each component is as follows:

$$a_{ij} = t_{ij} \log \frac{F}{f_j}$$

where

$a_{ij}$  = weight assigned to the term  $t_j$  in document  $D_i$

$t_{ij}$  = number of times that the term  $t_j$  appears in document  $D_i$

$f_j$  = number of times that the term  $t_j$  appears in the entire database

$F$  = total number of words (whether repeated or not, tokens) of the whole database.

The authors decided to use this scheme instead of the IDF method because the latter assigns the greatest weights to those terms that appear in a single document. Because the aim is to perform a classification and a topological organization that will show the semantic relationships existing between all the terms, those which appear in only one document are related to those of the same document, so that neither a classification nor a topological organization seems necessary. Hence, if the said discrimination values of all the terms are calculated and ranked in decreasing order, among the top values classified there will be a sharp reduction in terms of frequency 1 (among the first 100, there were none). There will also be a great reduction in those terms that appear in a single document (though not as great as that for those of frequency 1).

The entire procedure of creating the representation of the terms to be used as input to Kohonen's neural network can be summarized in four phases: elimination of stopwords, stemming, selection of words, and vectorization. These are described as follows:

*Elimination of stopwords:* This first phase is aimed at eliminating the functional words of the language that have no meaning. Because the language was English, the authors used the frequency dictionary of Kucera and Francis (1967), with which they generated a list of 200 stopwords corresponding to the words of greatest frequency in that language. The authors thereby reduced the number of different terms in the database from 7,758 to 7,577.

*Stemming:* The authors used the Porter Stemmer, which is the most-used stemming algorithm in English (Frakes, 1992; Porter, 1980). The number of terms is reduced to 5,052.

*Extraction of those terms of greatest discrimination value:* The authors calculated the discrimination value of all the terms, and extracted those 1,200 with the greatest value.

In the resulting database, the authors were left with 10% of the words appearing in a single document only. They took this as an acceptable extra computational load, given the generality of the procedure that had been carried out.

*Creation of the vector representation vectorial of the terms:* If one writes the vectors as horizontal arrays, and stacks them vertically, the rows of the resulting matrix each correspond to the same document, and the columns to the same term (i.e., each element indicates the weight of the term corresponding to its column within the document corresponding to the row). Thus, although the rows represent the document vectors by construction, the columns can also represent term vectors.

If the initial document vectors are binary, the term vectors thus generated would also be binary. In the  $tf \cdot idf$  weighting, each component consists of two parts: one local, which indicates the importance of the term in each document (and which would vary with the frequency in each document) and the other global, which indicates the importance of the term in the entire database (and which would be the same for all the coordinates of a term). The study of Noreault et al. (1981) showed that the best similarity functions are those that use angular measures, so that the global component of  $tf \cdot idf$  is of no use (it only modifies the modulus of the vector), and one would therefore be in a situation equivalent to using the frequency in the document for each component.

For this reason, similarly to what was done in the document vectors for the terms, one can generate different document weights, based for instance on the number of terms that they contain (Kantor, 1994). Logically, those vectors containing many terms have less weight than those that contain few terms. One can therefore use a weighting scheme similar to the  $tf \cdot idf$  for the terms

$$a_{ij} = t_{ij} \log \frac{D}{ND_i}$$

where

$a_{ij}$  = weight assigned to document  $D_i$  in term  $t_j$

$t_{ij}$  = number of times that term  $t_j$  appears in document  $D_i$

$ND_i$  = number of terms that appear in document  $D_i$

$D$  = total number of terms in the database

Once the terms have been represented as vectors, analogously to how similarities between documents were calculated using functions defined for that purpose, the similarities between terms using the same functions can be calculated.

### 1.2. *The Kohonen algorithm*

Despite the enormous complexity of the cerebral cortex under the microscope, macroscopically it has a uniform structure, even when comparing one brain with another. The centers corresponding to specific activities, such as thought, vision, hearing, or motor functions, are located in specific zones of the cortex, and each of these zones has a particular placement with respect to the others. An example is the tonotopic map of the auditory regions, in which neurons that are close to each other respond to similar sound frequencies.

This map may, to a great degree, be predestined by genetics. Nonetheless, it was his interest in looking at how organization of this type might come about that led Kohonen (1982, 1989, 1995) to investigate the subject. The result of these investigations was the neural network model that bears his name. Such networks are capable of topologically organizing their inputs.

The Kohonen model is a competitive model in which the influence of each neuron on the other neurons of its layer will be a function of the distance between them (i.e., each neuron will have a positive influence on itself and on neurons which are topologically close). This influence will decrease as the distance between the neurons increases, reaching negative values, but then will increase again so that finally there is a positive influence on the most distant neurons.

In addition, the Kohonen model has a biological foundation; it has been found that in certain primates there occur lateral interactions between neurons that are excitatory within a radius of 50 to 100  $\mu\text{m}$ , inhibitory within a circular corona of a 150- to 400- $\mu\text{m}$  thickness around the previous circle, and very weakly excitatory, or practically null, from that point out to a distance of several cm (Hilera & Martínez, 1995). There is therefore a bubble of activity in the layer, formed by all the units that are close to the winner and that participate in the reinforcement corresponding to learning.

The simulation hardware involves creating a competitive layer of a certain complexity. Nevertheless, the process that it performs during learning each time that it is presented with a vector can be summarized in the following steps:

Select as the winning node (each node of the network represented by a different weight vector of the same dimension as the input vectors) that which is closest (typically using the Euclidean distance) to the presented input vector.



facilit, optim, tool, nlm, evid	physician, patient, clinic	18
medic, medicin, biomed, partli, ebm	health, care, informat	19
18	19	

Fig. 2. Enlargement of the lower right corner of the topological organization from Figure 1.

Adjust the weight vectors of the winning node and the nodes corresponding to its neighborhood by adjusting them toward the input vector values (in some cases the reinforcement is the same for the whole neighborhood and in others it falls off with distance from the winner).

During this phase, the training vectors are presented repeatedly to the network at random, simultaneously with a gradual reduction of the neighborhood and of the learning rate to force the network’s stability. After this phase, the configuration is one in which the neurons that are topologically close in the network (arranged as a two-dimensional lattice) are winners with respect to clusters of vectors that are close to each other in the input space. Occasionally, as in the present case, the neurons are endowed with a “conscience.” This is a mechanism that reduces the probability that a neuron will win a competition as the number of competitions that it has already won rises, and which is sometimes used to avoid the problem of the stuck vector (Freeman &

	7	8	9	10	11	12	13	14
0	sun, java, hotjava	devic, pda, browser	client, distribut	chicago, illinoi, mosaic, second		html, macro, word, markup	hypertext	schema, brows, exploit, conceptu
1	realiti, assist	germani, carri, uniform	usabl, week, converg	server, collabor	webmap, immedi, window, frog, file	graphic, automat, sgml	interac, expres, manipul, abil	map, hypermedia, semant, navig, illustr

Fig. 3. Enlargement of the lower left corner of the topological organization from Figure 1.

Skapura, 1991; Guerrero, Moya-Anegón, & Herrero-Solana, 2002; Guerrero & Moya-Anegón, 2001; Muñoz Garcia, 1994), allowing the victories to be shared out over the whole network.

Because of this topological organization, at times the only item of interest is the clustering carried out by the hidden layer, and one selects the whole set of vectors for training for the sole purpose of seeing the resulting topological organization. In this last application, there exist two possibilities: if there are more hidden units than training vectors, one obtains an optimal projection on the chosen topology, or, if the number of units is less than the number of vectors, one obtains a layer that performs the clustering and orders each cluster topologically. The resulting number of clusters will be equal to the number of neurons making up the hidden layer.

Thus, one achieves, in an iterative but straightforward manner, not only a good cluster analysis but also a good topological organization. As with other algorithms, however, there are certain characteristics that are somewhat unsatisfactory mathematically: the termination is forced by the number of iterations, there is no guarantee of convergence, there is dependency on the order in which the data are input, stability is attained at the cost of slower learning, or a classic instead of a fuzzy partition is generated. There have been attempts at improvement and at "fuzzification" (see, e.g., Bezdek, Chen-Kuo Tsao, & Pal, 1992) where the aim was to incorporate some of the characteristics of the c-mean method, which yields a fuzzy output. Although all of these proposals lead to certain improvements, they all tend to have the common denominator of the loss of the topological organization that characterizes this algorithm.

The authors chose to link in an output fuzzification module that allows each input to have a certain degree of membership in each cluster. The algorithm then functions during learning according to Kohonen's original picture. It varies only in the production phase, where, instead of providing the winning cluster (the closest), it generates a fuzzy output, which indicates the input's degree of membership to each cluster. This fuzzy output is calculated in the same way

	2	3
0		hong, kong, proceed, asian
1	moslem, ac, com	southeast, republ, asia, singapor

Fig. 4. Enlargement of a zone of the topological organization from Figure 1 containing political-religious terms.

	0	1	2
3	western, economi, low	china, internetwork, chines, pose, market	financi, countri, connect, compani
4	profit	cee, eastern, overview, cash	
5	payment, europ	foundat, opportun, aslib, secur	copyright, ronchei
6	retrospect, conver	tackl, gatewai, theme	
7	ensur	societi, democrat, easi, deliveri	workplac, highlight
8	danish, realis, ministri	australian, nordic, uncov, readi, erin	polit, administr, feder
9	self, housekeep, fast	initi, competit	hour
10	darmstadt, hess, allegro	easili, expand, member	superhighwa i

Fig. 5. Enlargement of a zone of the topological organization from Figure 1 related to the capitalist economy and to democratic countries.

as with the c-mean algorithm (Bezdek, 1981) in the iterative process for  $m = 1$ . In a summarized form,

$$u_{ij} = \frac{1}{\sum_{k=1}^c (D_{ij}/D_{kj})^2}$$

where

$u_{ij}$ , is the degree of membership of the term  $t_j$  to the cluster  $i$

$D_{ij}$  is the distance between the vector corresponding to the term  $j$  and the weight

vector of the neuron  $i$  (which is simply the centroid of the cluster that it gives rise to)

As in the previous algorithm, in case a term vector coincides with the centroid (the neuron’s weight vector) of a cluster, the degree of membership to this cluster is set to unity and the rest to zero.

In the present study, the vectors described previously were applied as the input to this type of network and grouped into clusters according to the winning neuron. This type of cluster has the advantage compared with those arising from other algorithms of being topologically organized on the grid produced by the network (Guerrero et al., 2002). In addition, the fuzzy

	16	17	18	19
13	referenc, mark, rapid, analys		edition, revision, japanes	cite
14	rule, inqueri, iso, failur, congress	canadian, descript, archivist, entri	format, appear, classif, danbib	aacr, marc, quarter, festschrift, centuri
15	minim, mlc, less, whole, risk		index, bibliograph, descrip	

Fig. 6. Enlargement of a zone of the topological organization from Figure 1 related to information science.

		0	1
0		health	patient, care
1		partli, informat	clinic, ebm
2		medic	facilit
3		medicin, nlm, biomed	advanc, optim

Fig. 7. Enlargement of a zone of a topological organization performed by a 30 × 30 node network related to medicine and health care.

partition that results from the fuzzification module allows the terms belonging to each cluster to be ranked according to their degree of membership.

## 2. Results

The term vectors were applied to two networks. The first test used a network of dimension 20 × 20 nodes of hidden-layer neurons and, hence, clusters. Figure 1 shows the results showing each cluster as a cell in which the corresponding terms have been included (these are the terms that are closer to the corresponding centroid than to the others). Within a given

	6	7	8	9	10	11	12	13	14	15
0	devic, browser, pda		mosaic, second, illinoi, chicago	webmap	immedi	collabor, emploi, workgroup		sun, hotjava, java	netscap	b
1	scholarli, distribut	client		window		power	week	releas	soft	
2	sever	scholar	point, server	file, visit, coolist	code, version	converg, uniform	basisplu, due, page		c, com	msn, compuserv, prodigi

Fig. 8. Enlargement of a zone of the topological organization performed by a 30 × 30 node network related to the World Wide Web.

cluster, the terms have been ordered according to their degree of membership to that cluster as returned by the fuzzification module. It would actually be more correct to speak of word-stems rather than terms, since stemming was carried out before applying the network.

Because there are many cells and terms, just those parts of the classification that seem most representative will be commented on. First, in the lower right-hand corner of Figure 2, in the cluster of position (19,19), and in some of those nearby (19,18), (18,19), there appear terms (or word-stems) related to medicine: physician, patient, clinic, medic, biomed, care, and health. There also appear other words, such as informat, which is relatively logical since the instrumental nature of informatics may appear here and anywhere else that it is used.

Another part of the grid, which is quite different, is shown in Figure 3, where many terms appear that are related to the World Wide Web, such as java, hotjava, sun, browser, client, mosaic, html, markup, hypertext, map, hypermedia, navig, interact, sgml, webmap, or server; Here again there appear other words with a secondary and quite circumstantial relationship.

Close to this zone (see Figure 4), there exist some cells that contain political-religious terms related to Asia. This is not necessarily the zone exclusively containing terms of this type. There exists a nearby zone, which is apparently related to the capitalist economy and to democratic countries, in which there appear terms related to China, as shown in Figure 5. Last, Figure 6 shows an area in which there appear terms related to library and information science.

The second test is exactly the same but uses a network with a  $30 \times 30$  hidden-layer neurons. The result is very similar to that obtained in the former test, although because there are more cells (900 vs. 400), these contain fewer terms and more gaps. Nonetheless, one can locate zones that are similar to the previous cases. Figure 7, for instance, shows the zone relating to medicine, and which has a greater number of cells compared with Figure 2. Figure 8 shows the zone corresponding to the Web. The rest of the zones commented on before behave similarly.

### 3. Conclusion

The extraction of relationships between terms has been a long-sought goal. When the attempts were general in scope, covering an entire language, the desired results were not obtained. The demand, therefore, is always for application to particular specialized databases.

This study's procedure proved useful, because it uncovered quite a few semantic relationships between terms. It did this on the basis of the information present in the database itself, so that the results are especially suited to the corpus corresponding to that database. At the same time, given that the network performs a kind of co-occurrence analysis (although different in detail), the authors came across a number of anomalous relationships mainly due to the small size of their database. This problem will diminish as the database grows (as with other methods that use co-occurrence).

This procedure was aided by the Kohonen algorithm. The Kohonen algorithm is not only able to cluster nearby terms but also provides a topological organization of the clusters that it creates, and therefore has many potential uses in information science: for example, a search query can be expanded by including the terms of the corresponding cluster, or of a whole zone;

and, in browsing, the algorithm gives a concrete representation of the content of the database, so that training will not have to be carried out online but could be done periodically offline.

Also important is the result of the fuzzification module. Although it was used in the present study only to rank the terms of each cluster, it gives the same number of values of degrees of membership for each term as there are clusters. These values can thus be used to represent the terms, thereby reducing the dimension for subsequent calculations (Guerrero & Moya-Anegón, 2001).

Although the response times that were obtained were fairly long, with the rapid increase in calculation speed and memory capacity of computers, these times will be considerably reduced. Also, one of the characteristics of these algorithms is that they can be run in parallel on some of the already not-too-expensive neurocomputers that are currently available on the market. The present study used a small database. An application to a larger database would be influenced by an increase in the number of terms, which would only be reflected proportionally in the term selection phase, and by an increase in the number of documents whose consequence would be a corresponding increase in the dimension of the vectors representing the terms affecting proportionally both the selection and the network training phases.

In this study, the terms of a database were represented using Salton's vector space model. These terms were then applied to a Kohonen's feature map, which performed a topological organization yielding clusters and groups of neighboring clusters with related terms. Also, the fuzzification module gave a fuzzy partition, which allows one to obtain the same number of degrees of membership for each term as the number of clusters that have been generated, without losing the topological organization.

## Acknowledgments

This study was funded by the Junta de Extremadura-Consejería de Educación Ciencia y Tecnología and the Fondo Social Europeo, as part of research project IPR99A047.

## References

- Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum.
- Bezdek, J. C., Chen-Kuo Tsao, E., & Pal, N. R. (1992). Fuzzy Kohonen clustering networks. *Proceedings of the First IEEE Conference on Fuzzy Systems: March 8–12, 1992, Town & Country Hotel, San Diego, California* (pp. 1035–1043). New York, NY: IEEE.
- Chen, H. (1995). Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms. *Journal of the American Society for Information Science*, 46, 194–216.
- Chen, H., Houston, A. L., Sewell, R. R., & Schatz, B. R. (1998). Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49, 582–603.
- Chen, H., & Lynch, K. J. (1992). Automatic construction of networks of concepts characterizing document databases. *IEEE Transactions on Systems, Man and Cybernetics*, 22, 885–902.

- Chen, H., Lynch, K. J., Basu, K., & Ng, T. D. (1993). Generating, integrating, and activating thesauri for concept-based document retrieval. *IEEE Expert. Special Series on Artificial Intelligence in Text-Based Information Systems*, 8, 25–34.
- Chen, H., & Ng, T. D. (1995). 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, 348–369.
- Chen, H., Ng, T. D., Martinez, J., & Schatz, B. R. (1997). 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, 17–31.
- Chen, H., Schatz, B. R., Yim, T., & Fye, D. (1995). Automatic thesaurus generation for an electronic community system. *Journal of the American Society for Information Science*, 46, 175–193.
- Frakes, W. B. (1992). Stemming algorithms. In: W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms* (pp. 131–160). Englewood Cliffs, NJ: Prentice Hall.
- Freeman, J. A., & Skapura, D. M. (1991). *Neural networks: Algorithms, applications, and programming techniques*. Reading, MA: Addison-Wesley.
- Guerrero, V. P., & Moya-Anegón, F. (2001). Reduction of the dimension of a document space using the fuzzified output of a Kohonen network. *Journal of the American Society for Information Science*, 52, 1234–1241.
- Guerrero, V. P., Moya-Anegón, F., & Herrero-Solana, V. (2002). Document organization using Kohonen's algorithm. *Information Processing & Management*, 38, 79–89.
- Hilera, J. R., & Martínez, V. J. (1995). *Redes neuronales artificiales, fundamentos, modelos y aplicaciones*. Madrid, Spain: Rama.
- Honkela, T., Pulkki, V., & Kohonen, T. (1995). Contextual relations of words in Grimm tales, analysed by self-organizing map. In F. Fogelman-Soulié & P. Gallinari (Eds.), *Proceedings of the International Conference on Artificial Neural Networks* (pp. 3–7). Paris, France: EC2 et Cie.
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79, 2554–2558.
- Kantor, P. B. (1994). Information retrieval techniques. In M. E. Williams (Ed.), *Annual review of information science and technology: Vol. 29* (pp. 53–90). Medford, NJ: Learned Information.
- Kaski, S. (1999). Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. *Proceedings of the Ninth International Conference on Artificial Neural Networks* (pp. 940–945). London, UK: Institution of Electrical Engineers.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59–69 (Reprinted from *Neurocomputing*, by J. Anderson, & E. Rosenfeld, Eds., 1988, Cambridge, MA: MIT Press).
- Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). Berlin, Germany: Springer Verlag.
- Kohonen, T. (1995). *Self-organization maps*. Berlin, Heidelberg, Germany: Springer Verlag.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (1999). Self-organization of a massive text document collection. In E. Oja, & S. Kaski (Eds.), *Kohonen maps* (pp. 171–182). Amsterdam, Holland: Elsevier.
- Kucera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*, 13, 345–364.
- Lagus, K., & Kaski, S. (1999). Keyword selection method for characterizing text document maps. *Proceedings of the Ninth International Conference on Artificial Neural Networks* (pp. 371–376). London, UK: Institution of Electrical Engineers.
- Lagus, K., Kaski, S., Honkela, T., & Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In E. Simoudis, J. Han, & U. Fayyad (Eds.), *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (pp. 238–243). Menlo Park, CA: AAAI Press.

- Lin, X. (1997). Maps displays for information retrieval. *Journal of the American Society for Information Science*, 48, 40–54.
- Lin, X., Soergei D., & Marchionini, G. (1991). *A self-organizing semantic map for information retrieval*. Paper presented at the Proceedings of the Fourteenth Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (pp. 262–269). Chicago, IL.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3, 235–244.
- Moya-Anegón, F. (1994). *Sistemas Integrados de Gestión Bibliotecaria*. Madrid, Spain: Anabad.
- Moya-Anegón, F., Herrero-Solana, V., & Guerrero, V. P. (1998). Virtual reality interface for accessing electronic information. *Library and Information Research News*, 22, 34–39.
- Moya-Anegón, F., Moscoso, P., Olmeda, C., Ortiz-Repiso, V., Herrero-Solana, V., & Guerrero, V. P. (1999). NeuroISOC: Un modelo de red neuronal para la representación del conocimiento. In M. J. López Huertas, & J. C. Fernández Molina (Eds.), *La representación y la organización del conocimiento en sus distintas perspectivas: Su influencia en la recuperación de la información. Actas del IV Congreso ISKO-España* (pp. 151–156). Granada, Spain: ISKO-España.
- Muñoz García, A. (1994). *Redes Neuronales para la Organización Automática de Información en Bases Documentales*. Unpublished doctoral dissertation, Universidad de Salamanca, Spain.
- Negroponte, N. (1995). *El Mundo Digital*. Barcelona, Spain: Ediciones B (Grupo Z).
- Noreault, T., McGill, M., & Koll, M. B. (1981). A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In R. N. Oddy, S. E. Robertson, C. J. Van Rijsbergen, & P. W. Williams (Eds.), *Information retrieval research: Papers given at the 1st Joint British Computer Society (BCS) and Association for Computing Machinery (ACM) symposium: Research and development in information retrieval* (pp. 57–76). London: Butterworths.
- Oakes, M. P., & Taylor, M. J. (1999). *Clustering of thesaurus terms using adaptive resonance theory, fuzzy cognitive maps and approximate string-matching techniques*. Liverpool, UK: University of Liverpool.
- Orwig, R. E., Chen, H., & Nunamaker, J. F. (1997). A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science*, 48, 157–170.
- Peat, H. J., & Willett, P. (1991). Occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42, 378–383.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14, 130–137.
- Price, D. J. S. (1973). *Hacia una ciencia de la ciencia*. Barcelona, Spain: Ariel.
- Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61, 241–254.
- Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Strzalkowski, T. (1995). Natural language information retrieval. *Information Processing & Management*, 31, 397–417.
- Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33, 106–119.
- Voorhees, E. M. (1993). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 27 June–1 July 1993* (pp. 171–180). New York: ACM.
- White, H., Lin, X., & McCain, K. (1998). Two modes of automated domain analysis: Multidimensional scaling vs. Kohonen feature mapping of information science authors. In W. Mustafa el Hadi, J. Maniez, & S. Pollit (Eds.), *Structures and relations in knowledge organization: Proceedings of the Fifth International ISKO Conference, Lille, France* (pp. 57–61). Würzburg, Germany: Ergon Verlag.
- Wong, S. K. M., Cai, Y. J., & Yao, Y. Y. (1993). Computation of term associations by neural network. In R. Korfhage, E. Rasmussen, & P. Willett (Eds.), *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 27 June–1 July 1993* (pp. 107–114). New York: ACM.