



PERGAMON

Information Processing and Management 38 (2002) 79–89

**INFORMATION
PROCESSING
&
MANAGEMENT**

www.elsevier.com/locate/infoproman

Document organization using Kohonen's algorithm

Vicente P. Guerrero Bote^{a,*}, Félix de Moya Anegón^b, Victor Herrero Solana^b

^a *Library and Information Science Faculty, University of Extremadura, C/José María Alcaráz y, Alenda 1 (pasaje), 06011 Badajoz, Spain*

^b *Library and Information Science Faculty, University of Granada, Campus Cartuja, Granada, Spain*

Received 4 August 2000; accepted 29 November 2000

Abstract

The classification of documents from a bibliographic database is a task that is linked to processes of information retrieval based on partial matching. A method is described of vectorizing reference documents from LISA which permits their topological organization using Kohonen's algorithm. As an example a map is generated of 202 documents from LISA, and an analysis is made of the possibilities of this type of neural network with respect to the development of information retrieval systems based on graphical browsing. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Neural networks; Self-organizing maps; Document clustering; Vectorization; Browsing

1. Introduction

As is now well-known, the first information retrieval systems were based on uniterms and exact matches. One of the problems that this involved was that the retrieval depended on whether one or various exact terms existed in the query. When users require information, however, they are looking for concepts rather than terms. With the techniques of partial matching there has been some progress in the direction of conceptual searches, although the basis has been metric rather than cognitive. When one considers the amount of electronic information available world-wide, it is clear that information retrieval needs to overcome these problems to improve the users' access to that information.

Current technology with its sequential algorithms has far surpassed humans in tasks such as mathematical operations. There are other tasks, however, which are easy for humans but very

* Corresponding author. Tel.: +34-2-425-9910; fax: +34-2-428-6401.

E-mail addresses: vicente@alcazaba.unex.es (V.P. Guerrero Bote), felix@goliat.ugr.es (F. Moya Anegón).

difficult to solve using these classical methods. Examples are optical character recognition, image processing, speech, etc.

The current progress in the treatment of images has been achieved mainly by emulating the characteristics of the human process. For instance, it is known that the human eye responds instantaneously to the characteristics of the image (Julesz, 1991). This emulation has involved the development of high-order algorithms which have sometimes been based on neural networks (García del Valle Alfageme, 1996). In this sense, pattern recognition is presently reasonably developed, and from pattern recognition the recognition of characters and images which appear to be crucial for the development of the basic recognition routines to facilitate image retrieval (Myler & Weeks, 1993).

Likewise one also foresees the development of similar routines for the recognition of a concept in a text (Kantor, 1994). This development will at least have to involve studies of language and of psychology. In this sense, it seems suggestive that the form in which a document is left after pre-processing (removing the stop words, and unifying the different morphological variants) is much like the language used by children who are beginning to speak. It is thus a matter of imitating human reasoning capacities with the corresponding advantages and disadvantages.

Neural networks are designed to simply emulate the internal functioning of the brain, and they have been at the base of most successes achieved in image processing. We therefore believe that these methods may lead to progress in information retrieval, and allow some of the problems currently being tackled to be solved.

These networks perform thousands of processing steps of local information at each element. Their sum gives rise to the overall intelligent behaviour of the network (Doszkocs, Reggia, & Lin, 1990). The said forms of processing are found to be best suited to tasks with a great degree of complexity, but not to traditional mathematical operations or similar tasks. In light of the problems that they are able to solve, some researchers have seen them as a new paradigm of artificial intelligence. Recent studies, however, show that their current citation environment is quite distinct from that of artificial intelligence (Van der Besselaar & Leydesdorff, 1996).

Neural networks can learn to assign multidimensional outputs to multidimensional inputs, and they do so while maintaining a great capacity for generalization. This is because they first perform a phase of training or learning, which in most networks is done offline. There is thus a differentiation between the production phase and the training phase in almost all neural network architectures. This division is due to the difficulty in keeping a network sufficiently plastic for it to be able to continue learning, and sufficiently stable for it not to forget what it has learnt. Indeed, this is known as the stability–plasticity dilemma. There also exist other types of neural network such as those corresponding to Adaptive Resonance Theory which are able to overcome this dilemma and continue learning while they are in production.

As mentioned above, in the first stage the networks learn to assign outputs to the inputs. This training may be supervised or unsupervised. In supervised training, one uses a set of input–output pairs so that the network progressively adapts towards the desired outputs on the basis of the errors that it makes. In unsupervised training, only the inputs are presented to the network (without their corresponding outputs) – hence the name. The network, in most of these training methods, forms clusters with the said inputs. One might therefore say that the first potential application of this type of algorithm with document input would be simply clustering. What the training phase really does that gives rise to these clusters is, however, to enable the network to find

the closest cluster to a document vector, whether it belongs to the training set or not. This means that it can find the cluster that corresponds to a document or to a request for information (which can be transformed into a document vector).

The clusters that are produced by an application of neural networks have certain highly desirable traits from the perspective of document clustering. One is that they are adaptive, i.e., they do not have to be provided with parameters of size, overlap, etc., but to a great degree adapt to the database in question. One could say that they learn from the documents they have to deal with. One consequence is that the clusters are closer together where there are more documents. Thus, the zone a cluster covers may shrink or grow according to this density of documents. To this end, the vectors are subjected to an overall comparison which is able to include in a cluster documents that do not have exactly the same terms. This can allow a query to be answered with documents that do not include the term in question but are, nevertheless, closely related, i.e., while the response is the cluster that is closest to the query, this cluster contains documents that are included because of their overall similarity with the rest although they do not contain the terms of the said query.

We focused on one of these algorithms – Kohonen’s model. This not only clusters the inputs, but also organizes the resulting clusters topologically. In the present work, we study specifically the latter process. We shall first describe the model, then the generation of the database for the experiment, and finally comment on the results that we obtained.

2. Materials and methods

Despite the enormous complexity of the cerebral cortex microscopically, macroscopically its structure is uniform, even from one brain to another. The centres corresponding to specific activities such as thinking, vision, hearing, motor functions, etc., are located in specific regions of the cortex with a determined spatial relationship between them. One example is the so-called tonotopic map of the auditive regions, in which neurons which are close to each other respond to similar sound frequencies. Another is the somatotopic map whose artistic representation is the well-known homunculus.

The cortex is an extensive (approximately 1 m² surface area) thin (between 2 and 4 mm thickness) layer consisting in turn of six layers of different types of neurons. Its folds maximize the area that fits inside the cranium. For our purposes, however, we shall treat it as simply a surface.

It may well be that this map is to a great degree foreordained by genetics. Nevertheless, Kohonen’s interest in discovering how an organization of this type might arise led him to investigate the subject (Kohonen, 1982, 1989, 1990, 1995). The product of those researches was the network model that bears his name, and which is capable of performing a topological organization of the inputs presented to it.

This type of network has recently been used in documentation for the analysis of domains (White, Lin, & McCain, 1998), for textual data mining (Lagus, Honkela, Kaski, & Kohonen, 1999), to extract semantic relationships between words from their contexts (Honkela, Pulkki, & Kohonen, 1995; Ritter & Kohonen, 1989), and in particular to generate topological maps of sets of documents, even labeling the zones of influence of each word or term (Kohonen et al., 1999a; Kaski, 1999; Lagus & Kaski, 1999; Moya, Herrero, & Guerrero, 1998; Moya Anegón et al., 1999;

Chen, Houston, Sewell, & Schatz, 1998; Lin, 1997; Guerrero Bote, 1997; Orwig, Chen, & Nunamaker, 1997; Lin, Soergei, & Marchionini, 1991).

The present work was designed to test the capacities of this algorithm under conditions as close as possible to reality, as well as to study a reliable process for the generation of document vectors from the database. Strictly for experimental purposes, we created a test-bed database from records of the LISA bibliographic database, so that we could later compare the organization that we obtained with that resulting from the descriptors assigned to the said records. We took each of the abstracts contained in the records to be independent documents. Of the remaining fields, we only took the descriptors into account for purposes of the subsequent comparison. To achieve as much generality as possible, instead of retrieving documents by topic, we chose the last 954 records of LISA (Summer 96 version), these being the records with Accession Number greater than or equal to 9605000. This led to the database that we created containing the aforementioned 954 records, with a total of 7758 different words.

2.1. Document vectorization

The next step following the creation of the document database was to transform these documents into vectors to use as inputs to our algorithms or networks. To this end, we applied the vector space model which transforms each document into a vector. First, each term in each document was assigned a corresponding weight. Then a set of terms was determined which we could use to describe all the documents of the database.

We chose a weighting scheme very close to the classical IDF which was found to be one of those that gave the best results in the study by Noreault, McGill, and Koll (1981). Each component's weight is given by

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

where a_{ij} is the weight assigned to the term t_j in document D_i , t_{ij} the frequency of appearance of the term t_j in document D_i , f_j the frequency of appearance of the term t_j in the whole database and F is the total number of words (repeated or not, tokens) in the whole database.

We decided to use this scheme instead of the classical IDF because the latter assigns the greatest weights to the terms that appear in a single document since it uses n_j , the number of documents that appear in the term, instead of f_j , *the frequency of appearance of the term t_j in the whole database*. Since we wish to perform a clustering procedure, however, these terms which appear in only one document are translated into differences between their corresponding document and the rest, and are of no use in forming clusters.

The number of terms in the database was 7758. This was computationally unfeasible for us, so that it became necessary to determine a set of terms that we could use to describe all the documents of the database, and then assign them the corresponding weights in each document.

We think that the ideal way to reduce the said number of terms is to use the discrimination value (Salton & McGill, 1983) calculated with the cosine function as the measure of similarity. This leads to an angle comparison which is desirable in similarity functions according to the study of Noreault et al. (1981). Ranking the calculated discrimination values for all the terms in decreasing order, we found that in the top ranked there was a drastic reduction in the number of

terms of frequency one (we found none in the first 100). There was also a marked reduction in those that appeared in only one document (but not as large as for the frequency-one case). These findings confirm the results of studies on the value of discrimination (Salton & McGill, 1983; Moya Anegón, 1994) and are encouraging pointers that this indeed must be the ideal way to reduce the number of terms. In line with those studies, one may expect that the results will improve as the database increases in size.

The entire process carried out on the terms can be summed up in three phases:

- *Elimination of stop words.* This first phase deals with eliminating functional words of the language which have no meaning. Since the language was English, we used the frequency dictionary of Kucera and Francis (1967), with which we generated a list of 200 stop words corresponding to the most frequent words of that language. This reduced the number of different terms in the database from 7758 to 7577.
- *Stemming.* We used the *Porter Stemmer* which is the most commonly used stemmer in English (Porter, 1980; Frakes & Baeza-Yates, 1992). This reduced the number of terms to 5052.
- *Extraction of the terms with the greatest discrimination value.* We calculated the discrimination values, and extracted the 1200 words with the greatest value.

After this process, there were only 10% left of the words that appear in only one document. Given the generality of the procedure, there will be a corresponding reduction in the computational burden.

In the study of Noreault et al. (1981), the best similarity measures were those that made angular comparisons. We therefore next normalized the resulting vectors by dividing them by their Euclidean norm.

As we noted above, the intention of the present study was to test the topological organizations that can be obtained with this algorithm. For this purpose, we have to simulate a network with more hidden neurons than input patterns. In view of the results of experiments carried out previously (Guerrero Bote, 1997), these 954 documents would be suitably classified using a network with 1500 neurons in the hidden layer. Given the dimension (1200) and number (954) of the vectors, however, such a high number of clusters would make the process unfeasible in terms of computation time. There is the additional difficulty that one would have in subsequently contrasting the results with the complete trial database. We therefore decided to reduce the number of vectors to be processed. This lessens both the computation times and the difficulty in examining the results. It does not lead to any loss in generality, however, since in obtaining the said vectors we used all the documents and hence all the terms that they contributed.

We excluded therefore those documents that contained none of the descriptors listed in Table 1. This table also gives the number of documents in which each descriptor appears.

Table 1
Descriptors used to reduce the document set and the number of documents in which each descriptor appears

Acquisitions	37
Artificial Intelligence	28
Business Management	20
Computerized Information Storage and Retrieval	23
Conferences	20
Periodicals	22
World Wide Web	58

We were finally left with 202 documents with which to generate our database. This number is not the simple sum of the above figures because there were documents containing more than one of the descriptors.

3. Results

We made two trials using networks of different dimensions. Firstly we used a 20×20 network, i.e., one with 400 neurons. The result is shown in Fig. 1. We have marked the nodes where the documents were classified with (redundantly) both a colour and a letter associated to the descriptor assigned to the documents, and with the number of documents classified at the said node. At node (19,19) for instance, there were two documents classified which had the descriptor *Business Management* assigned them. One sees in the figure how nodes with documents containing the same descriptors cluster together.

We should first clarify that there are a number of neurons which are winners for more than one descriptor. Some of these cases are due to the existence of documents that have two or more descriptors. When we eliminate these, in which the confusion was inevitable, the following cases of confusion remain:

- (8, 0) Periodicals/World Wide Web;
- (8, 15) Periodicals/Adquisitions;
- (4, 15), (7, 8), (7, 10), (15, 15) Conferences/Periodicals;

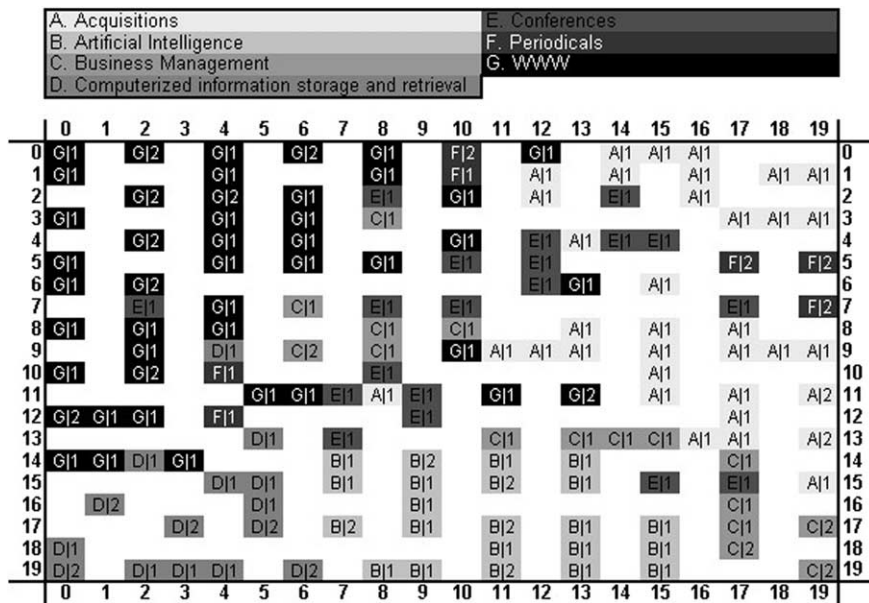


Fig. 1. Document Organization of a 20×20 network. The nodes where the documents were classified have been marked with (redundantly) both shading and a letter associated to the descriptor assigned to the documents, and with the number of documents classified at the said node.

- (5, 10) Conferences/Computerized Information Storage and Retrieval;
- (6, 15) Acquisitions/World Wide Web;
- (7, 6), (13, 11) Business Management/World Wide Web;
- (11, 11), (13, 5) World Wide Web/Computerized Information Storage and Retrieval;
- (14, 2) Artificial Intelligence/Computerized Information Storage and Retrieval.

On careful observation, one finds that a great many of these cases are caused by the descriptors Conferences and Periodicals, which are in turn the descriptors that are the most spread out over the map and most coincide with each other (they participate in seven coincidences, of which they occur together in four). The explanation might be that they are both descriptors of form, so that it would be natural to expect a topical dispersion of the documents that contain them.

Looking at the above theoretically confused documents in more detail, such as for instance those corresponding to the node (8, 0) (Accession Numbers 9605017 and 9605899), one finds that they share the descriptor “Hypertext”. The network, however, classified them according to the terms contained in the abstracts which are apparently not closely related to the descriptor.

One can thus see that when there is apparent confusion, and two descriptors win at the same node, the documents are usually related. It is not limited to this, however, since, as was to be expected, the documents of bordering clusters are also related.

Simple observation shows that the zone assigned to Acquisitions is split into two. When one examines the documents of the two zones (such as, for example, those whose Accession Numbers are 9605223 and 9605247), one finds that the upper zone (that of document 9605247) is mainly based on topics that are economic, budgetary, technological, etc., while the lower zone (corresponding to 9605223) is based on aspects relating to the development of a library’s collection.

Another noteworthy aspect is the placement of the different descriptors winning zones on the map. For example, the zones corresponding to the descriptors World Wide Web, Computerized Information Storage and Retrieval, and Artificial Intelligence, which are the most technology related, border each other. Also that order seems to be the most logical, since one may say that the WWW is related to Information Retrieval given the importance that information seeking systems are gaining on the Web. Also, Information Retrieval is related to Artificial Intelligence, and after all the present work is an artificial intelligence technique applied to information retrieval. In the zone bordering Artificial Intelligence there happens to appear a zone dedicated to Business Management which expands inwards to the centre, touching all the other zones.

There exist many other interesting questions in the organization, some of which also seem inexplicable. The intention of the present study, however, was not to make an exhaustive study (which will be the subject of later work), but to perform a trial and calculate the possibilities offered by this type of network.

The second trial was to classify the same database using a 30×30 neuron network. The result is shown in Fig. 2.

One sees that the general behaviour is very similar to the previous case. As the dimension is greater, however, the documents are spread out over the whole map, and no white zones are generated that might indicate the distances between the different topics. This is caused by the “conscience” with which the units were endowed (a mechanism which reduces the probability that a neuron will win a competition as the number of competitions that it has already won rises, which is sometimes used to avoid the problem of the stuck vector (Freeman & Skapura, 1991; Muñoz García, 1994)), which ensures that the competition, and therefore the effort, is shared out over the

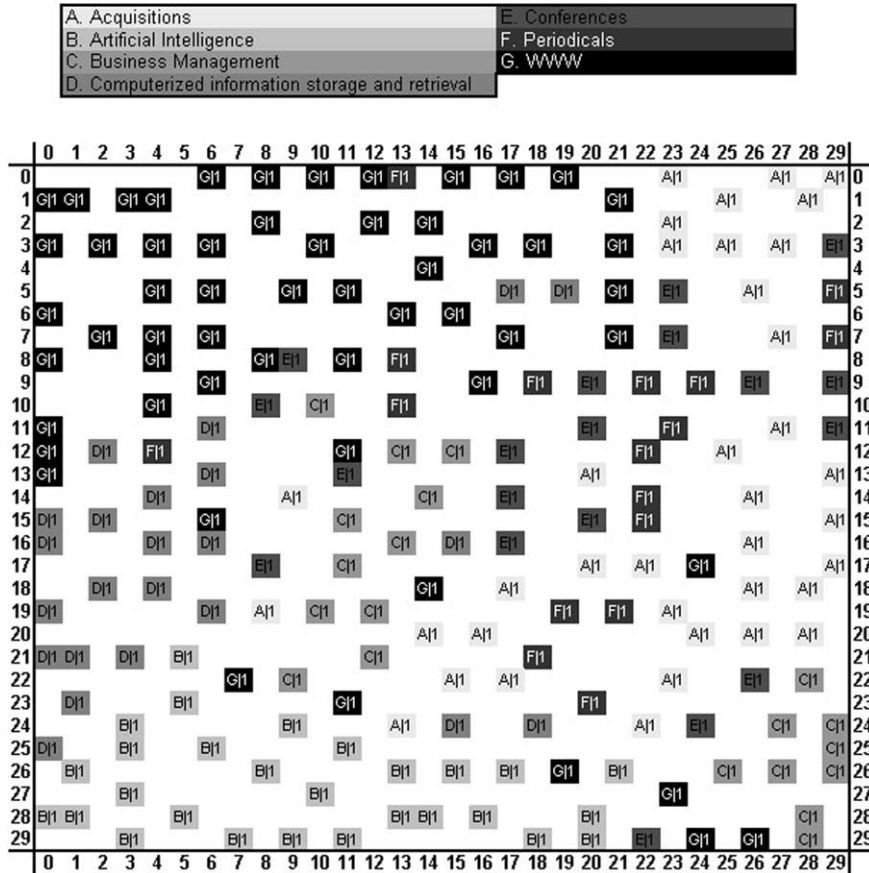


Fig. 2. Document Organization of a 30 × 30 network. The nodes where the documents were classified have been marked with (redundantly) both shading and a letter associated to the descriptor assigned to the documents, and with the number of documents classified at the said node.

whole network. Unlike the previous case, now no node wins more than one competition, so that the only descriptor confusions are those due to documents that have more than one descriptor.

One notable difference is the tendency for the occurrence of microzones where various documents of one descriptor are clustered in the midst of zones corresponding to other descriptors. This is the case, for instance, in the zone of (29, 24) where there is a cluster of a set of documents relating to the WWW in the middle of a zone corresponding to Business Management. In the previous case, the documents corresponding to this latter descriptor clustered in an elongated but more or less continuous zone that rose from the bottom right-hand corner towards the centre. Now they have split into two groups, one in the centre and the other in the aforementioned corner. Also the distances between the two groups of Acquisitions has widened, and so forth. This is so because of the existence of more gaps, which allows a different form of organization to occur.

A glance through the groups of documents corresponding to the Business Management descriptor shows that they all deal with computer-related applications. Those of the lower left corner deal with outsourcing, however, while the others do not.

As in the previous trial, when one looks in detail at documents situated in zones corresponding to other descriptors, one sees that there indeed exist relationships between the two. For instance, that corresponding to node (0, 13) of Periodicals which appears in the WWW zone deals with electronic publications and hypertext, that corresponding to node (8, 9) of Conferences which appears in this same zone corresponds to a congress on the HTTP and HTML standards, etc.

4. Conclusion

A neural network is able to learn to respond to a certain type of input or question on the basis of a set of training examples. This response capacity is not constrained to the training set examples but may be generalized for other similar cases or inputs. In the unsupervised learning case, it achieves a suitable clustering of the inputs. If one applies this process to document vectors, the result is a fairly good document organization, better than that provided by a descriptor of the document (with which we contrasted the present results). It even manages to find relationships between documents without their sharing many terms.

In the learning process, as well as clustering the inputs, the Kohonen network generates a topological organization of those clusters. When we apply this to documentation the result will be the creation and organization of clusters in a manner that those which are topically close will also be close in the network. We may use this to *expand the query*, or rather the *results*: once one has found the cluster that best fits the query, one may extend the activation to those which are topologically close. On occasions, this has come to be used as a sort of *system of navigating* through the document database, indicating which zones of our topological organization we are visiting with the consequent aid to browsing, and even to choose the location and size of the zone of the database that we wish to visit. Recently, networks similar to these have been used to generate *topological maps* of a database (Kohonen et al., 1999a,b; Kaski, 1999; Lagus & Kaski, 1999; Moya et al., 1998; Moya Anegón et al., 1999; Chen et al., 1998; Honkela, Kaski, Lagus, & Kohonen, 1996; Kaski, Honkela, Lagus, & Kohonen, 1996; Lagus, Kaski, Honkela, & Kohonen, 1996), of the results of a search (Lin, 1997), or even of the comments contributed during a brainstorming session (Orwig et al., 1997). In some of these cases, as well as performing a document classification, one determines for each node which unitary term vector produces the greatest activation. One may thereby generate each term's zone of influence, providing a graphical view of the database on which one could even select the zone that one wants to visit.

In this present work, we used a reduced set of selected documents so as to be able to evaluate the result more easily, not because more documents cannot be processed. If more documents are processed with a network of the same dimension, there will be a greater density of documents. Furthermore, if one does not start with a set of selected documents, the evaluation will become more difficult, since the network will find relationships which were not determined a priori and which therefore would have to be studied case-by-case. The network will generally be able to classify much larger sets than that used here (Kohonen et al., 1999a,b; Kaski, 1999; Lagus & Kaski, 1999), even using classifications of different levels (Moya et al., 1998; Moya Anegón et al., 1999; Chen et al., 1998) to map or access a given document collection.

We have here used networks that yield a two-cal organization. The generalization to three-dimensional organizations is immediate, so that the clusters could be organized three-dimensionally.

To display that structure, however, 3D or virtual reality systems would be needed. The human difficulty in comprehending the structure means that there is no sense in opting for a higher dimension.

Neither should one forget that a network is trained to give responses. In the present case, the response may be the calculation of the nearest cluster to a document or query.

Lastly, these topological organization processes involve massive calculation. They have the advantage, however, that these calculations can be performed in parallel. This needs purpose-designed machines, which, despite their production still not being en masse, are now reasonably priced. There exist cards for PCs (the machines on which we carried out the trials) which are moderately priced and achieve increases in speed of the processes of factors of several thousands, even tens of thousands.

Acknowledgements

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

References

- Chen, H., Houston, A., Sewell, R., & Schatz, B. (1998). Internet browsing and searching: user evaluations of category map and concept space techniques. *Journal of the American Society for Information Science*, 49(7), 582–603.
- Doszkocs, T. E., Reggia, J., & Lin, X. (1990). Connectionist models and information retrieval. In M. E. Williams (Ed.), *Annual review of information science and technology* (Vol. 25, pp. 209–260). Amsterdam: Elsevier.
- Frakes, W. B., & Baeza-Yates, R. (1992). In W. B. Frakes, & R. Baeza-Yates (Eds.), *Information retrieval: data structures & algorithms*. Englewood Cliffs, NJ: Prentice-Hall.
- Freeman, J. A., & Skapura, D. M. (1991). *Neural networks algorithms applications and programming techniques*. Reading, MA: Addison-Wesley.
- García del Valle Alfigeme, M. (1996). Diseño de un modelo genérico de red neuronal y desarrollo de un sistema para su simulación. Ph.D. thesis, Universidad de Extremadura.
- Guerrero Bote, V. P. (1997). Redes neuronales aplicadas a las técnicas de recuperación documental. Ph.D. thesis, Universidad de Granada.
- Honkela, T., Kaski, S., Lagus, K., & Kohonen, T. (1996). Self-organizing maps of document collections. *Alma*, 1(2); URL <http://www.diemme.it/luigi/alma.html>.
- 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 international conference on artificial neural networks, (ICANN95)* (pp. 3–7). Paris: EC2 et Cie.
- Julesz, B. (1991). Early vision and focal attention. *Reviews of Modern Physics*, 63(3), 735–772.
- 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, Inc.
- Kaski, S. (1999). Fast winner search for SOM-based monitoring and retrieval of high-dimensional data. In *Proceedings of the ninth international conference on artificial neural networks (ICANN99)* (pp. 940–945). London: Institution of Electrical Engineers.
- Kaski, S., Honkela, T., Lagus, K., & Kohonen, T. (1996). Creating an order in digital libraries with self-organizing maps. In *Proceedings of world congress on neural networks (WCNN96)* (pp. 814–817). San Diego, CA: Lawrence Erlbaum.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1), 59–69.
- Kohonen, T. (1989). *Self-organization and associative memory* (3rd ed.). Berlin: Springer.

- Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9), 1464–1480.
- Kohonen, T. (1995). *Self-organization maps*. Berlin: Springer.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (1999a). Self organization of a massive text document collection. In E. Oja, & S. Kaski (Eds.), *Kohonen maps* (pp. 171–182). Amsterdam: Elsevier.
- Kohonen, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., & Saarela, A. (1999b). WEBSOM – A novel SOM-based approach to free-text mining. Neural Networks Research Centre (NNRC) at Helsinki University of Technology (HUT). <http://websom.hut.fi/websom>.
- Kucera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RD: Brown University Press.
- Lagus, K., Kaski, S., Honkela, T., & Kohonen, T. (1996). Self-organizing maps of document collections: a new approach to interactive exploration. In E. Simoudis, H. Jiawei, & U. Fayyad (Eds.), *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 238–243). Menlo Park, CA: AAAI Press.
- Lagus, K., Honkela, T., Kaski, S., & Kohonen, T. (1999). WEBSOM for textual data mining. *Artificial Intelligence Review*, 13(5&6), 345–364.
- Lagus, K., & Kaski, S. (1999). Keyword selection method for characterizing text document maps. In *Proceedings of the ninth international conference on artificial neural networks (ICANN99)* (pp. 371–376). London: Institution of Electrical Engineers.
- Lin, X., Soergei, D., & Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In Paper presented at the *Proceedings of the 14th annual international ACM/SIGIR conference on research and development in information retrieval*, Chicago, IL.
- Lin, X. (1997). Maps displays for information retrieval. *Journal of the American Society for Information Science*, 48(1), 40–54.
- Moya Anegón, F. (1994). *Sistemas integrados de gestión bibliotecaria*. Madrid: Anabad.
- Moya, F., Herrero, V., & Guerrero, V. (1998). Virtual reality interface for accessing electronic information. *Library and Information Research News*, 22(71), 34–39.
- Moya Anegón, F., Moscoso, P., Olmeda, C., Ortiz-Repiso, V., Herrero Solana, V., & Guerrero Bote, V. (1999). NeuroISOC: un modelo de red neuronal para la representación del conocimiento. In M. J. López Huertas, & J. C. Fenández 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 (EOCONSID'99)* (pp. 151–156). Granada: ISKO-España.
- Muñoz García, A. (1994). Redes neuronales para la organización automática de información en bases documentales. Ph.D. thesis, Universidad de Salamanca.
- Myler, H. R., & Weeks, A. R. (1993). *The pocket handbook of image processing algorithms in C*. Englewood Cliffs, NJ: Prentice-Hall.
- 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 first joint British Computer Society (BCS) and Association for Computing Machinery (ACM) symposium: research and development in information retrieval*, 1980 June, St. John's College, Cambridge, England (pp. 57–76). London: Butterworths.
- Orwig, R., Chen, H., & Nunamaker, J. F., Jr. (1997). A graphical, self-organizing approach to classifying electronic meeting output. *Journal of the American Society for Information Science*, 48(2), 157–170.
- Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3), 130–137.
- Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61(4), 241–254.
- Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.
- Van der Besselaar, P., & Leydesdorff, L. (1996). Mapping change in scientific specialties: a scientometric reconstruction of the development of artificial intelligence. *Journal of the American Society for Information Science*, 47(6), 415–436.
- 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, & A. S. Pollit (Eds.), *Structures and relations in knowledge organization: Proceedings of the fifth International ISKO Conference*, 25–29 August 1998, Lille, France; *Proceedings of the fifth international ISKO conference*. Würzburg: Ergon Verlag.