Journal of Universal Computer Science, vol. 4, no. 7 (1998), 630-651
submitted: 25/3/98, accepted: 22/5/98, appeared: 28/7/98

Link-based Shaping of Hypermedia Webs [1] Assisted by a Neural Agent

Alberto Faro
Universita' di Catania, Italy
afaro@k200.cdc.unict.it

Daniela Giordano
Concordia University, Montreal, Canada
giord@alcor.concordia.ca

Corrado Santoro
Universita' di Catania, Italy
csantoro@iit.unict.it

Abstract : The paper proposes a neural agent that performs self-organizing classification to assist in searching and contributing to webs of documents and in the process of document reuse. By applying the Kohonen self-organizing feature map (SOFM) algorithm to patterns of influence links among documents it is possible to originate clusters of documents that help infer the aspects that such documents implicitly share. The approach complements search techniques based on semantic indexes. The resulting classification is sensitive to the multiple aspects of a document that may belong to multiple classes with a varying degree and allows for treating effectively items that typically have a limited life span, either because they are means to the collaborative production of a more complex item, or because they belong to fast evolving domains. The method has been implemented by Lotus Notes Domino Web server for a case-based application in the domain of information systems design.

Key Words : information retrieval, neural networks, hypertext navigation

Categories : H.3.3, H.5.1, I.5.1, I.5.3

1 Introduction

Webs of hypermedia documents need support for interactive exploration, to orient the user and to facilitate effective document retrieval. Among the solutions that have been proposed recently are perspective walls [MacKinlay et al. 91], interactive dynamic maps [Zizi & Lafon 95], dynamic landscapes [Chalmers et al. 96]. Regardless of which specific front-end visualization technique is adopted, the critical issue for effective use of such webs is finding adequate forms of a document's organization to reflect the task domain and support different user typologies. The same collection of

[1] This is an extended version of a paper presented at the WebNet'97 conference in Toronto, Canada. The paper has received a "Best Full Paper Award".

documents may benefit from different, possibly co-existing, forms of organization that become more or less suitable according to what is the specific goal involved in retrieval, especially if each document is of a grain size such that its contents involve many facets. In particular, retrieval for the reuse of documents is a scenario that deserves attention because reuse is an activity integral to many tasks that can be supported by the web, such as case-based problem solving and those tasks that involve the collaborative production of documents (e.g., design specifications, building shared models, legal agreements).

Documents can be organized with a varying degree of semantic and structural constraints [Wang & Rada 95], nonetheless there are limitations inherent to retrieval based on semantic indexes. In fact, whether the documents are organized in a conventional database or in a hypertext, searches based on keywords are not robust because of the „vocabulary problem", i.e., the fact that spontaneous word choice for the same domain by different subjects coincides with less than 20% probability [Furnas et al. 87]. This can be ameliorated by techniques for generating particularly sophisticated thesauri such as, for example, the concept space proposed in [Chen et al. 96] or for performing automated semantic analysis of the text, such as the latent semantic analysis proposed in [Landauer & Dumais 97]. Without delving, for the moment, into a detailed analysis of the performance of each of the above techniques, the problem that remains open is that terms or indexes rarely support the psychological process of flexible framing of contents [Medin &Ross 89], and of perceiving their multiple facets. As a result, the set of documents retrieved after a search often share only a shallow semantics, in which the context that makes a particular document salient tends to be lost.

When the base of documents is fast evolving, because of content updates, or because the documents are temporary means to produce a deliverable in a cooperative setting, more flexible and evolving classification techniques are needed. Flexibility is required in order to track a classification process that is fundamentally emergent and to retain a discriminating power for the multiple aspects and issues coexisting in a document, or, with a small leap of abstraction, in a „case". For example, the same piece of information may become irrelevant with respect to a problem, but still retain some value with respect to an issue that was unforeseen at the time of document creation. Also, the same piece of information can become obsolete or become incorporated in the web in a more refined form, thus discarding the original source or precedent versions is warranted. Therefore the shortcomings of index based retrieval techniques with respect to capturing the temporal dimension of meaning (topicality, obsolescence, evolution with respect to an issue) are apparent.

On the other hand, it is practically impossible to classify documents according to all the facets that they may possess, since these facets do not have an ontological status and mainly emerge from the modes of discourse adopted by a specific community. Therefore we are faced with a twofold problem: i) what clustering of documents can reveal the aspects being adopted by a community for framing experiences, and ii) how can the users enter the setting resulting from this clustering in a point near to the documents conveying the aspects being sought? This latter issue is referred to, in the context of hypertexts, as the „entry point" problem [Carlson 89].

Bearing these problems in mind, the paper proposes an approach that is complementary to symbolic retrieval and is based on the influence links that trace the document evolution, and whose regularities may be used to discover aspects otherwise concealed. The underlying classification technique is based on a self-organizing mapping [Kohonen 89] of a web of documents linked by weighted reference relations into a set of neurons to highlight classes according to topological properties of the original data space and operates on the reference links that take into account the influence relations among the documents. Such links are generated by the documents' authors, who acknowledge influence relations by creating citation links to other documents when contributing to the web. Reference links are not typed, to avoid incurring in the indexing problems highlighted above (as the approach of treating a web as a semantic network would entail) and also because research shows that users resist creating and using typed links [Wang & Rada 95]. The goal is to let emerge from a geography of links a classification that:

- takes into account multiple aspects of a document, so that an item can be considered as belonging to more than one class, with a varying degree;
- allows for treating items in the web that typically have a limited life span, because they are means to the collaborative production of a more complex item or because they belong to fast evolving domains;
- facilitates searching the web and orients the process of contributing an item to the document base.

The remainder of the article is organized as follows. [Section 2] briefly characterizes some approaches based on lexical analysis to improve recall and precision in retrieval and points out how they stand with respect to the issues of support for flexible framing of contents and the overall organization of the base of documents. [Section 3] describes how the Kohonen self-organizing feature map (SOFM) algorithm can be applied to patterns of influence links, to originate documents' clusters that help infer aspects implicitly shared by the documents. [Section 4] discusses how the proposed self-organizing classification assists in consulting, reusing, and contributing to the web and how conventional retrieval methods can support the user in approximating the best entry point of the web. [Section 5] illustrates an implementation of the method by Lotus Notes Domino Web server and a neural agent performing a Kohonen-like classification applied to information systems design. Finally, [Section 6] deals with a case study to point out the strengths and weaknesses of the proposed methodology.

2 Semantic retrieval based on lexical analysis

The two key measures of performance in retrieval are recall and precision. Recall is defined as the ratio between the number of retrieved relevant documents and all of the existing relevant documents. Precision is defined as the ratio between the number of

retrieved relevant documents and the total number of retrieved documents. In the classical retrieval techniques, any set of keywords is able to recall the documents in which such keywords appear most frequently, and precision is obtained by restricting the retrieved documents to the ones that share specific keywords, i.e., keywords that are present in a few documents of the database [Salton 92]. In the modern retrieval techniques, such as the „concept space“ for automated thesaurus generation [Chen et al. 96], any set of keywords that belong to the thesaurus are able to recall not only the documents in which the set of keywords appear most frequently but also the documents in which the keywords that co-occur with the original ones appear most frequently. This increases the recall power of the keywords but, generally, it does not increase precision since the ratio of relevant documents to non-relevant ones is approximately the same both before and after consultation of the thesaurus [Chen et al. 96]. In fact, keyword co-occurrence or term co-occurrence with respect to the entire document has been proved to be unsatisfactory to grasp the „similarity“ between the keywords or the terms of a thesaurus [Salton et al. 96]. For this reason, techniques in which word similarity is computed based on their co-occurrence within the paragraphs of the documents rather than within the entire document (such as the Latent Semantic Indexing (LSI) [Landauer & Dumais 97]) are becoming increasingly adopted to improve precision for information retrieval.

Modern information retrieval techniques, including the aforementioned concept space and latent semantics indexing approaches, are not necessarily bound to generate a flat list of documents in response to a query. On the contrary, they are able to map both the query and the retrieved documents onto a n-dimensional space in such a way that the closer the documents are to the query, the more likely they correspond to the user's needs. In particular, the concept space organizes the documents in a two-dimensional setting by using a Kohonen neural net [Kohonen 89]. The N input neurons of the neural net are exposed to Nt vectors, each input vector Vi being associated to one of the Nt documents of the base and each vector's component Vij measuring the occurrence of the j-th term of the thesaurus in the whole document i. The Kohonen net is trained in such a way that each vector (i.e., each document) activates only one neuron of the output array and similar vectors (i.e., similar documents) activate close output neurons. After the neural network has been trained, any term of the thesaurus (i.e., a vector having only one non null element corresponding to this term) or any query issued by the users (i.e., a vector whose non null elements are only the ones corresponding to the features of the query) may be associated with only one output neuron. Therefore the two dimensional array of output neurons originates a concept space in the sense that it may be subdivided into areas representing similar terms conveying a concept [Chen et al. 96], whereas any query activates one neuron that is in an intermediate positions with respect to the ones activated by the elementary terms contained in the query. In this condition, the documents that correspond to a given query are the ones that activate the output neurons belonging to a small area around the neuron activated by that query. Let us note that the documents retrieved in this way are not necessarily labelled by terms contained in the query, but also (according to the concept space approach) by terms that frequently co-occur with the ones issued by the user. A simplification of this

approach may be found in [Kohonen 96] where the thesaurus is restricted to the set of keywords labeling the documents under classification.

So far, approaches that apply the Kohonen neural net have been proposed to reduce the complexity of n-dimensional classification by mapping it into a bi-dimensional plane. This allows the discovery of more or less obvious clusterings but has limitations in pointing out the different kinds of analogies that may co-exist among the documents. This is the reason why latent semantics analysis and latent semantic indexing (LSI) use a 300-dimensional space identified by a specific factor analysis of the word/paragraph co-occurrence matrix. However, the n-dimensional clustering of the latent semantics approach, as the mentioned concept space, has essentially a lexical basis that might fail in pointing out other important ways of clustering (e.g., behavioral, ontological, emotional) that often are more consonant with the user's query. In fact, in textual communication important elements (such as the activities, the situations and the emotions that are distinctive of the practices of the community addressed by the document) are often left implicit. Even if these elements are not written in the documents, they are active in the background and play an important role in framing both narration - on the author's part, and comprehension - on the reader's part. Some weaknesses of lexical clustering are also outlined in [Foltz 90], where it is pointed out that documents close to each other in the n-dimensional LSI setting cannot always be considered satisfactory responses to the query issued by the user. Additionally, it must be noted that the lexical approaches above do not take into account notions of relevance that may originate from the documents being organized in a hypertext structure, and do not support the process of contributing to such web.

3 Shaping the Web by Neural Classification

In this section we illustrate a method for inferring a similarity degree among documents from the information embedded in the references links and for creating clusters of related documents. The method starts by asking the author to link every new document to those documents dealing with relevant ontologies, by using a quantifier I (Influence weight) defined as follows: $I = 0.5$ if the new document takes into account some marginal aspects of the referenced item, $I = 1$ if the new document inherits several important aspects of the referenced item, and $0.5 < I < 1$ for the intermediate situation. The influence weights in the range between 0 and 0.5 are not used since they do not produce any practical effect, as explained below. [Fig. 1a] shows the influence weights placed on the reference links between documents.

There are several ways to classify elements in classes not known *a priori*, for example, methods based on the information exchanged between the new element and the other ones [Alexander 64]. Here we adopt a neural approach based on Kohonen self-organizing networks (or maps) [Kohonen 89] which aggregates the documents in classes, not known *a priori,* that preserve a meaningful topological distribution, i.e., in our case, the more aspects are shared, the closer the classes. By this approach, self-organization of the input documents is obtained by fixing the synaptic weights among the input and output neurons of the neural network as resulting from an unsupervised learning process that depends on the difference between the synaptic weights and the

values of the input neurons, rather than resulting from a supervised learning process that depends on the difference between the actual output and the desired one (since the desired output is not known *a priori*). Thus the learning process terminates when the synaptic weights of the output neurons are not significantly changed by the input vectors that, as shown in [Fig. 1b], we propose to be associated to the documents in such a way that vector i represents document i and the j-th component of the input vector i (i.e., X_{ij}) represents the influence degree between document i and document j. At the end of the learning process a document i is characterized by a vector Di whose general element Dik measures the distance between document i and neuron k as follows:

$$Dik = \text{Const.} [(X_i - W_k)^T (X_i - W_k)]^{1/2}$$

being W_k the vector representing the set of the synaptic weight between the output neuron k and the input neurons. The class of the document i is the class associated to the output neuron k to which document i is closest, i.e., the output neuron k that satisfies the following expression:

$$\min Dik, \quad k=1 \text{ to total number of output neurons.}$$

The implementation of this algorithm is as follows:

- the reference graph consisting of documents interconnected by the above influence weights [Fig. 1a] is extracted from the web;
- this graph is transformed in matrix form [Fig. 1b];
- this matrix is given, as the input space, to the neural network [Fig. 1c], whose number of output neurons is set equal to the number of classes in which we want to classify these documents.

Under these conditions, the neural algorithm not only classifies documents according to the aspects they share, but also operates a classification in which the spatial distance among the neurons that represent the classes mirrors the one of the input space, i.e., close output neurons represent ontologically close classes. This latter feature depends on the fact that the learning algorithm reinforces not only the synaptic weights of the output neuron to which the present document is closest (i.e., the winner neuron) but also it reinforces the synaptic weights of the neurons in the neighborhood of the winner (even if by an intensity lower than the one adopted for the winner neuron).

Let us note that the activation function adopted for the output neurons is a sigmoid, thus the input neurons whose value is less than 0.5 barely affect the output neurons; consequently, influence weights between documents range from 0.5 to 1, unless two documents are unrelated, in which case their influence weight is zero.

After having presented our classification method it might be evident that one main difference between our method and other ones that apply the Kohonen neural net [Chen 96]; [Kohonen 96] is the nature of the input space, since we use an inter-document influence matrix, whereas the other methods use a term (or keyword) occurrence matrix. This is an important difference because the influence matrix

allows us to discover evolving classifications, whereas thesauri or keywords are fixed *a priori* and therefore they are not able to capture rapidly evolving clustering. However, this is not the only difference since we believe that linear or two dimensional settings are not able to represent effectively the neighborhood of a class: in fact, the complexity of the class relations requires that they be represented in more than two dimensional spaces, some times even in non-Euclidean ones. For this reason when we need to navigate from a class to its neighboring classes we do not move in a linear or planar space, as requested by the Kohonen approach, rather we use inter-class links, whose weights are computed by a n-dimensional formula that measures the link weights between class i and class r as follows:

$$L_{ir} = [\Sigma_j \, Erj] / n_i \quad \text{where } j = 1 \text{ to } n_i$$

$$Erj = 1/Drj = 1/[(X_{ij} - W_r)^T (X_{ij} - W_r)]^{1/2} \quad \text{if } Drj > \theta \text{ otherwise } Erj = 1$$

where X_{ij} is the vector characterizing item j of class i (e.g., vector (1,0.8,0.5,1) characterizes item 1 in [Fig. 1b]), W_r is the vector of the synaptic weights of the neuron representing class r, whereas n_i is the number of the items belonging to class i. Constant θ may be fixed in several ways, e.g., as the minimum Drj or, in the case min(Drj)=0, as the average of the two or three lowest Drj.

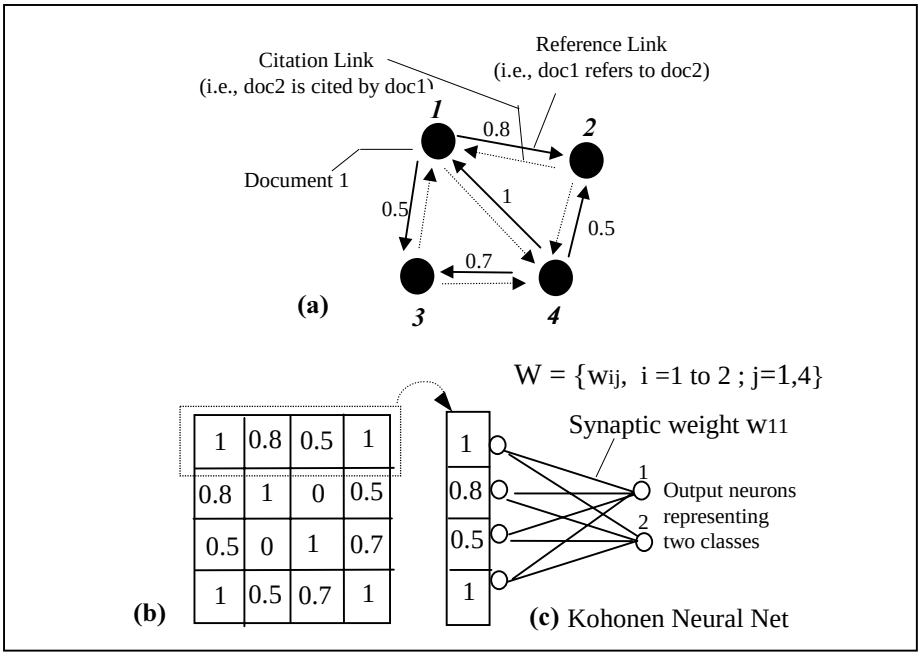

Figure 1: a) Space of references and citations in the web; b) influence matrix between documents; and c) self-organizing feature map W to classify documents in classes (input space = influence matrix, output space = neurons representing the classes).

[Fig. 2] shows how the method works assuming a binary decomposition scheme. Starting from the initial class containing all the cases (Class1), the algorithm

subdivides it into two classes and then re-subdivides Class1.1 and Class1.2 in other two classes and so on. Thus, to classify a new document it is sufficient to start from the class that contains all the items referenced by the new one. For example, if the new document refers to items in Class1.2.1 and Class1.2.2 the method restarts classification from Class1.2.

When a new document is inserted in the web with its reference links, the set of the already existing classes to which it belongs is computed. Thus the neural classification is repeated every time a new document enters the web; the classes are created and dynamically refined with the web evolution. If the new document does not belong to any existing class, the author is invited to introduce a general description (pattern) to provide some clues concerning the meaning of the newly created class. If the document is placed on an existing class, but the author does not agree with the proposed patterns, s/he can add a new version of the patterns that presently denote the class. If the document belongs to a class not denoted by a pattern yet, the author is „challenged" to identify a general pattern, which is likely to emerge if all the documents referenced by the new one with I > 0.8 belong to the same class.

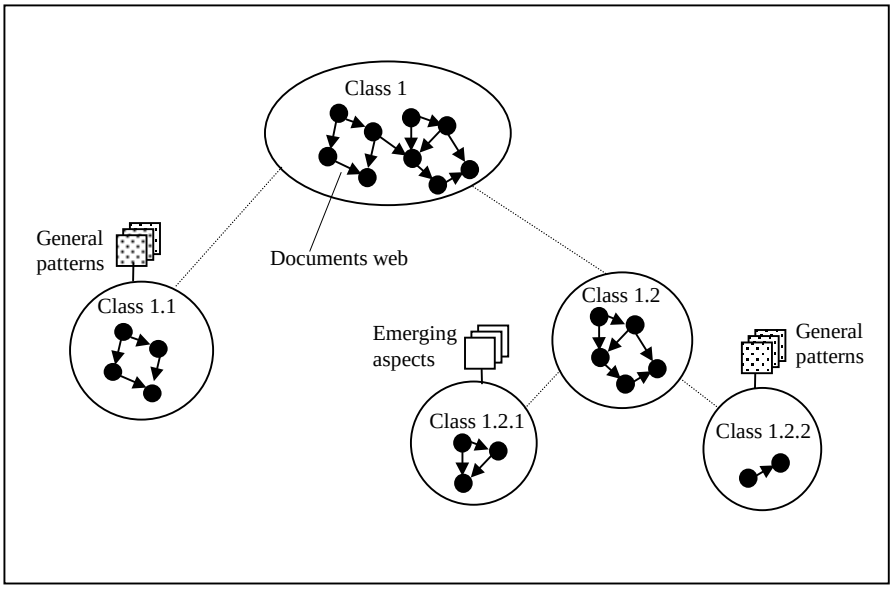

Figure 2: Classification of the documents in the web by the neural algorithm.

Other outputs of the classification are: a) for each class, a measure of the interconnectedness Ai of the elements in the class i (aggregation factor) and, b) for each element j, the degree E_{ij} with which it belongs to all the existing classes j. A discussion on the possible ways of measuring these factors is outside the scope of the paper, however, as an example, we show the ones adopted in our classification algorithm, i.e.,:

$$A_i = [\Sigma_j Eij] / n_i \quad \text{where } j= 1 \text{ to } n_i$$

$$Eij = 1/Dij = 1/[(X_{ij} - W_i)^T (X_{ij} - W_i)]^{1/2} \text{ if } Dij > \theta \text{ otherwise } Eij = 1.$$

where X_{ij} is the vector characterizing document j of class i, W_i is the vector of the synaptic weights of the neuron representing class i, whereas n_i is the number of the items belonging to class i. Constant θ is fixed as described above. By this convention, fully aggregated classes are characterized by Ai = 1. Analogously, item j completely fits to a class i if its Eij=1. Let us note that even if Eij=1, it is possible that $Erj \neq 0$ since item j could convey all the aspects of class i but also some aspects of class r.

Experimental evaluation of the classification method that we have implemented has shown that binary decomposition of the initial class into 2^k classes (after k successive refinements) is more accurate than the one step classification obtained by using a Kohonen network with 2^k output neurons, i.e., the $N=2^k$ classes obtained by applying k times the binary classification are more aggregated than the ones obtained by subdividing the initial set of documents into N classes in only one step. Depth of classification, e.g., the number of levels, can be fixed by the user. In any case, classification is stopped when all the subclasses cannot be further subdivided due to their high aggregation factor (lowest level classes).

Adding a new document could modify the structure of the existing classes, i.e., some old document could pass from a class to a different one. However, this phenomenon involves only few documents of the existing classes, and modifies only marginally the structure of the classes. This happens because as long as classes become consolidated, the links introduced by the new item are significantly less in number with respect to the existing ones. The documents that migrate to new or different classes are important to originate new ontologies or to reinforce the existing ones. At the end of the decomposition, we have these types of classes:

- classes that are denoted by general patterns, i.e., ontological descriptions that shape the web (e.g., class 1.1 or class 1.2.2 in [Fig. 2]); such classes are characterized by a high aggregation factor;

- classes that cannot be denoted by a single description, either because there is no underlying ontology or because their ontology is so ill-structured that it cannot be expressed explicitly (e.g., class 1.2 in [Fig. 2]);

- classes that are denoted by partial descriptions pointing out particular aspects that can be taken into account when authoring documents that will be aggregated in the same or the neighboring classes (e.g., class 1.2.1 in [Fig. 2]); such classes are characterized by an intermediate aggregation factor.

4 Use and Reuse in Self-Organizing Webs

Our scenario, which emphasizes web documents retrieval for reuse, is inspired by the case-based reasoning (CBR) paradigm [Kolodner 93], i.e., an approach to problem solving based on finding the best similar „case“ matching the current problem and then adapting it to solve the problem. The new generated case and the „lessons“ it

conveys can be contributed to the base of cases, which thus learns the new experience and makes it available for future use.

CBR can be considered an effort in the direction of querying the system in cognitively plausible ways, by resorting to sophisticated indexing schemes and to a carefully chosen vocabulary to ensure a proper level of abstraction. In fact, too abstract indexes may collapse the difference among cases and over-generalize them, thus providing little heuristic power in finding few best matching cases; on the other hand, highly specific indexes may fail to capture relevant similarities. Although indexing has been criticized as not being a psychologically plausible model of analog retrieval [Thagard & Holyoak 91], still it proves useful whenever the adopted classification scheme is stable and sufficiently descriptive of the problem and of the domain. For example, a fixed classification scheme, e.g., indexing, can be adequate for the retrieval of documents based on stable categories such as authors, title or date.

The way in which the proposed neural approach is complementary to symbolic retrieval is explained in the following. By creating a web of documents linked by references that do not have an explicit semantics but that only capture strength of influence, it is possible to originate, as discussed in [Sect. 3], a space that can be dynamically classified by extending the self-organizing Kohonen map. Following the metaphor of conventional „folders“, one might think of a folder as representing, more or less explicitly, the aspects shared by the documents contained in it. The assumption of the paper is that „folders“ do not have an *a priori* ontological status, and the method attempts to support the processes underlying the folders origination and evolution and the placing of documents in multiple folders. This is helpful especially in two situations : 1) when there is a huge quantity of documents to scan (consultation mode) and 2) when an author, or a team, wants to place a document in context (contributing mode). Conventional retrieval methods allow the user to find some relevant documents, whereas neural classification helps the user in: a) identifying the folders to which the relevant documents belong, i.e., the folders that highlight aspects co-existing in the documents relevant to the users' needs, and b) passing to neighboring folders where the user may discover other important facets of the problem under study.

In particular, the proposed approach supports a search and retrieval mechanism based on four main steps:

1. first the documents are classified by the neural net described in [Sect. 3] and the interclass distances are computed as proposed in the same section;

2. a document, or a set of documents, is identified based on semantic/lexical conventional criteria (e.g., by full text search or conventional indexes);

3. each document is proposed with the context (i.e., the class) to which it more strongly belongs;

4. finally, the closest classes are highlighted also, to suggest other relevant items or contexts.

This approach allows us to implement a sort of spreading activation mechanism that makes it likely to find the information, suggestion, or item being sought in the surroundings of the initially retrieved documents. One peculiar advantage afforded by the proposed technique is to provide this neighborhood.

Step 1 of the approach has been widely described in [Sect. 3]. How to enter and to navigate the web by taking advantage of the global forms emerging from the self-organizing classification is discussed in [Sect. 4.1]. The conditions under which a class can be considered a stable form that facilitates the search process and the conditions under which considering only one or few forms may be too restraining for solving the problem under study are pointed out in [Sect. 4.2]. A concrete example of entering, navigating, contributing and innovating the web is discussed in the case study, illustrated in [Sect. 6], which is concerned with a community of learners who use a web of linked design artifacts to get inspiration for solving problems of data and interactions modeling in information systems design. [Sect. 6] also discusses in detail why the net's topological self-organization in classes provides an implicit representation of the aspects shared by the documents classified as belonging to that class.

4.1 Entering and Navigating the Web

Generally, the more structured the documents, the more precisely they can be retrieved from the set of words (or keywords) that co-occur together with the query words (or keywords) with respect to the structural units in which the documents are subdivided [Salton et al. 96]; [Faro & Giordano 97a]; [Landauer & Dumais 97]. The internal structure of the documents in a web generally depends on the modes of discourse adopted by the community to whom the documents belong and on the type of document. For example, articles are subdivided in sections, the notes of meetings are organized according to agendas, playwrights progress through scenes, narration consists of intertwining stories that evolve though episodes. However, many documents have only a lexical structure, i.e., they are subdivided in paragraphs.

In [Foltz 90] it is shown that retrieval methods that take advantage of the internal structure of the documents, such as Latent Semantic Indexing (LSI) [Landauer & Dumais 97], increase both the recall and, especially, the precision performances of the keywords matching, respectively, by 13% and by 26%. Given that the figures for keywords matching are about 65% for recall and 54% for precision [Chen et al. 96] we can expect about 70% of average recall and 65% of overall precision by using LSI.

However, meaningful results can be attained also by using a full text searching engine and a contextualized filter [Bourigault 92] consisting of a set of words (and their synonyms) each referring to a unit of the document. This is especially true if these units are connected in such a way as to form a unique context, such as, for example: the first word refers to the title/abstract of a paper, the second one to the title/introduction of a chapter and the third one to the title/body of a section. In this case it is reasonable to expect some increase in the above performances with respect to the uncontextualized filters powered with some synonyms, whose recall and precision typically are, respectively, in the range 15%-30% and 30%-50% [Salton 92]; [Chen et

al. 96]. Thus we can assume that a reasonable order of magnitude of the performance of contextualized filter-based searching is 30% in recall and 50% in precision, i.e., that performance is at least shifted towards the highest levels of the ranges above.

By entering the web by a contextualized filter, firstly we have to discover the relevant documents (about 50% of retrieved items) and then we should move around these documents to increase recall without greatly decreasing precision. Local links certainly assist in understanding better the meaning and context of the retrieved documents, but the mechanism of local exploration is costly, thus it is recommended to resort to it especially when the organizing principle underlying the current class (returned with each retrieved document) is not evident *yet*. This is likely to occur when the documents are not stable, or when the user has some difficulty in framing the search problem. In this sense, after entering the base of documents it may be helpful to navigate in the neighborhood of the retrieved documents by using the hypertextual influence links proposed in the paper, whereas using links automatically generated by the lexical similarity of documents, such as the ones proposed in [Goffinet & Noirhome-Fraiture 96], would allow the user to visit only lexically relevant documents, thus loosing many other documents important for the query. However, as long as the documents' configuration, following the self-organizing classification, evolves towards more stable forms, local links become less useful in the search process, even if they still play a role in letting the „global forms" of the web emerge. As discussed in the previous section, the evolution towards clearer forms of the documents' organization will be determined by the insertion of new elements that will update the pre-existing configuration of links.

Identifying stable or emergent global forms may improve recall and precision in information retrieval. In fact, knowledge of what are the other items of the classes, say entry classes, to which the entry set documents belong facilitates discovering the most relevant documents among the ones retrieved by the full text searching mechanism mentioned above. As explained above, to identify the relevant documents in the entry set amounts to having more or less 1/3 of all the relevant items. The other 2/3 of the relevant documents may belong to the entry classes or to the neighboring classes. Thus a good heuristic may be moving from the set of the entry classes to the neighboring classes only if the total number N of documents of the entry classes is significantly less than three times the number M of documents in the entry set, i.e., $N << 3M$. In this case the other relevant items have to be discovered in the neighborhood of the entry classes. If $N >> 3M$, it may be useful to decompose the global forms into sub-forms that make more specific aspects emerge, as discussed in [Sect. 3]. Of course, $N \cong 3M$ is the termination condition for the searching process. This way of proceeding aims at recalling more or less all the relevant documents and at discovering the most significant facets for the user query.

If the query is not expressed by significant words and synonyms or if the documents are either not structured or structured by a superficial framing, precision and recall can dramatically decrease. Low precision makes it difficult to discover important items among the ones of the entry set, whereas low recall may increase navigation in the neighborhood of the entry classes to find other important facets for the query. In these cases using more sophisticated thesauri (e.g., the concept space or

LSI) could be, even if it is costly, the only way for obtaining an entry set that facilitates the discovery of the most relevant items for the query.

4.2 Emergent vs. Stable Global Web Forms

It must be noted that meaningful global forms tend to emerge and become stable when a huge quantity of interrelated documents is available. Under these circumstances large scale dynamic classification cannot be the sole responsibility of a human processor. In real life, a small scale approximation of the dynamic classification process occurs when a problem is framed and solved by incorporating incrementally the suggestions coming from peer reviewing and expert consultations, each suggestion highlighting some particular aspect of the problem. The process' validity increases when the number of consultations increases, and when everybody is aware of each other's suggestions, as in a meeting or brainstorming session. This is quite rare and quite costly, but, fortunately, CSCW technologies and models now make it possible to collect contributions in a shared electronic environment, in which the role of the above neural agent is justified, also to support the asynchronous sharing of experiences for reuse.

However, reasoning on a knowledge base for reuse cannot take place successfully if one is not able to manage the contradictions that are inevitably present in every dynamic collection of documents. This reverberates as a potential weakness in the use of global forms for search and retrieval. Solving this problem is not easy. In fact, contradictions can arise because of item „misplacement", or because the item contains errors or misconceptions. The first problem can be solved by a finer classification of the space of the documents, or by „migration" of the item to a more appropriate partition of the documents' space. The second one can be solved either by document elimination or by amendment, to inhibit the creation of a new class that would be based on faulty hypothesis. Related to both these points is the observation that neural classification assists in managing the documents' space growth by allowing the elimination of obsolete documents only when they belong to classes that are consolidated in stable ontologies, thus keeping the overall web organization stable in spite of the deleted links. Thus, on the one hand, neural classification of a documents' web assists the user in discovering contradictions by comparing an item with the other items of a class and, on the other hand, it avoids eliminating „productive" contradictions, i.e., ones that may let emerge other forms of documents' organization in which they are possibly resolved (web progress).

Another important consideration in reusing experience is whether one has to restrict the analysis to the items (and related classes) directly linked to the user's problem or whether consulting other items may be fruitful even if they are not linked very precisely to the query issued by the user. Precision and recall are certainly important performances when the web is used within the context of a bibliographic search, for example to know the state of the art of the problem under study (as in the scientific practice) or to identify relevant precedents (as in the legal practice) or to highlight the origins of a question (as in the historiographic practice). On the other hand, in the context of CBR-like searches, especially for design purposes, re-using

experience embedded in documents retrieved under „quasi-perfect“ precision and recall may promote uniformity or idiosyncratic ways of approaching design problem-solving. Thus a certain degree of fuzziness in information retrieval may be desirable in some cases so that opportunities for incidental learning are not cut off [Levitt & March 88] and the user can be exposed to items that are possibly suggestive of new aspects or solutions even if they do not appear relevant immediately. This implies that when the web is used as a shared design memory it is important to encourage the users to enter the web by assuming an exploratory disposition towards *a priori* „not relevant“ documents that may be conducive to new solutions. For this reason if all of the entry classes in response to a query are stable global forms it may be fruitful to explore the neighboring classes even if the condition N<< 3M does not hold. This suggestion supplements the heuristics discussed in [Sect. 4.1]. If the innovative links generated in this way among the documents are gradually reinforced by the consensus of the other users, such links may underscore the development of new practices within an organization.

5 Self-organizing Documents' Webs in a Lotus Notes Based Environment

A Lotus Notes based environment, called StoryNet [Faro & Giordano 97b] for the collaborative production of documents structured in stories and episodes, has been enhanced by a neural agent performing classification according to the SOFM algorithm previously outlined. The StoryNet's architecture has been conceived to manage evolving systems, and is proving useful for information systems (IS) collaborative design. The rationale for the story based organization is that in such a format experiences can be cognitively represented and recollected [Bruner 90]. In the application of StoryNet to IS design, a project consists of a set of use stories and episodes of the information system. Each episode is linked to the ones it refers to, and may be reused for specifying analogous episodes. The episode's categories (title, assumptions, what, who, why, when, where, rituals, how, what can go wrong, exception handling) are used as a probe to extract the episodes that best fit the specific design needs [Faro & Giordano 97b]. After adapting these episodes, the designer inserts the new episodes in StoryNet. To support reuse, any new document should be inserted as a justified evolution of the previous ones, i.e., as an enhancement of the experience already captured in the knowledge base. This can be pointed out in comments mediating the references links.

StoryNet has been implemented by Lotus Notes Domino Web server, to afford easy access to the designers without requiring a Lotus Notes client. The story-episode organization is easily supported by the Domino Web server, as it is capable of full text search on all the documents. To link episodes belonging to different stories it is necessary to extend the Domino Web server by a suitable software library of C modules that supports the referencing process as follows :

1. the designer first creates special documents to comment the episodes that have been proposed as the result of the search; typically only the subset considered

potentially relevant to the current purposes is marked by a comment ([Fig. 3], step 1);

2. while detailing the new episodes, the designer may scan the comments for possible suggestions ([Fig. 3], step 2);
3. after having specified the episode, the designer creates references to the comments that were taken into account ([Fig. 3], step 3).

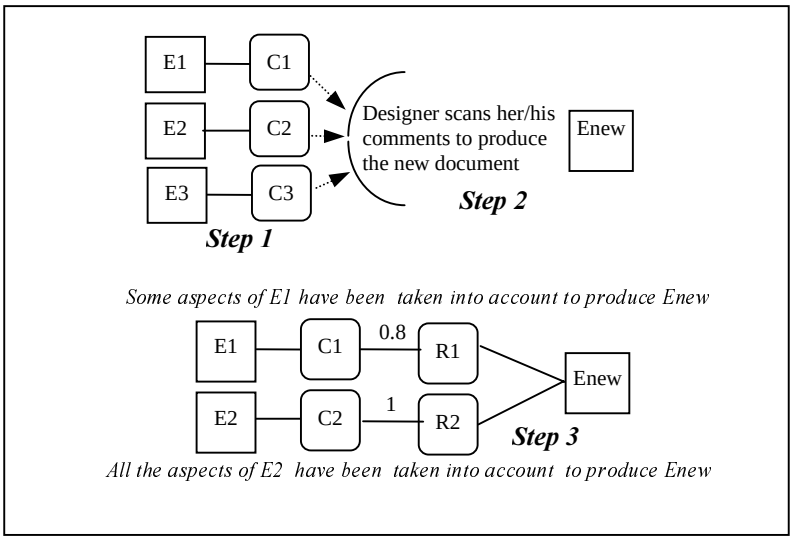

Figure 3: Supporting the referencing process in StoryNet (Ei = episode, Ci = comment, Ri = reference).

Passing from an episode to its comments and to its references is supported by the Domino Web server facilities; passing from an episode to the referred ones is supported by the above extension. For example, to pass from a Enew to its referenced items one can obtain the list of all the references, i.e., R1and R2 [3], then pass from Ri to Ci by simply clicking a special field inside the reference Ri. After reaching Ci it is easy to pass to episode Ei by the Lotus Notes facilities.

[Fig. 4a] shows how the user can navigate from a document, e.g., „driving lesson reservation“, to its source, e.g., „flight lesson reservation“, via a reference link. Episodes are organized in a graph whose oriented arcs are labeled by a number measuring how much an existing episode has influenced the new one. The graph is put in a inter-episodes influence matrix stored into a file external to StoryNet, to be elaborated at regular intervals by the neural agent. The agent stores the hierarchical classification of the episode into another file, so that StoryNet can superimpose this classification scheme on the existing episodes. The current version of StoryNet labels each episode by the lowest level class it mainly belongs to, and provides all the classes to which the episodes belong. [Fig. 4b] shows the StoryNet's user interface for the

classification results. Note that „driving lesson reservation“ and „flight lesson reservation“ belong to the same class, due to the reference link. If an episodes belongs to a class with a degree greater than 0.8, the two relevant lowest level classes are shown too.

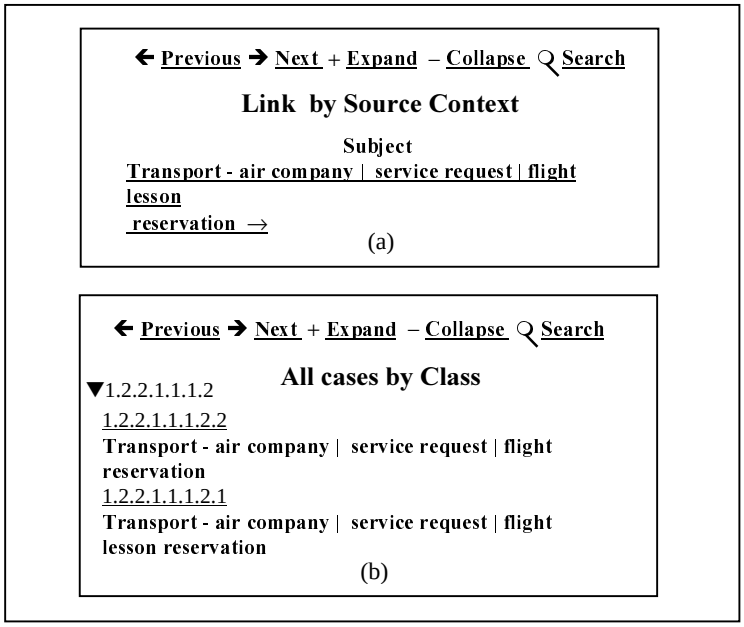

Figure 4: (a) StoryNet reference links; (b) StoryNet classification performed by the neural agent.

6 Case Study

In this section we illustrate how the proposed method works for a realistic Web consisting of 150 design documents whose links simulate the consultation process of the students of a course in information systems design. In particular, each document describes either a „use story“, i.e., a story specifying the scenario of a use case of the information system as a sequence of episodes, or a „use episode“ detailing the actions of the users to achieve goals productive for the progress of the higher-level scenarios. According to the method exposed in [Sect. 3], two documents are linked if the old document was deemed relevant by an author for producing a new design.

This example refers to a simulation performed during the preliminary phase of StoryNet, during which it was important to prove the feasibility of the method. Presently StoryNet is populated by about 1500 documents and an accurate evaluation is in progress to highlight the subtle process of concept formation and knowledge transfer within a community of learners. However, the simulation is illustrative

enough to clarify the strengths and weakness of the neural classification and the related heuristics to support the reuse of experiences. The documents of the simulation belong to the following information systems' application areas:

1. Community development in the city
2. Hospital first aid department
3. Walk-in clinic
4. Music (CDs and recordings) store
5. Musical instruments store
6. Electronic equipment store
7. Aquarium
8. Video rental service
9. Photographic studio
10. Car rental service
11. Automobile equipment and supplies
12. Ski school
13. Vocational training center
14. Conventions and exhibits center

Each of the above projects consists of about 3 use stories, each consisting of about 4 use episodes. For example, one story of the „Ski School" is „Equipment rental", which entails the episodes of „Reserving the equipment", „Equipment preparation", „Pick-up and Payment", and „Restitution". In the simulation each use episode is linked to the story to which it belongs (with weight 0.5) and to the relevant (with weight 0.8) and highly relevant (with weight 1) episodes or stories of other projects. The overall number of links between episodes belonging to different projects is approximately 150. This number was purposely so small to reflect the average number of links that the students tend to deploy, as it was observed in StoryNet. A discussion on how to facilitate the insertion of links is outside the scope of the paper, however, because links affect both the recall and precision of the method it is useful to report that particular attention was devoted to this problem by interviewing the students about the reasons of the apparent reluctance in placing the links. From the analysis of their responses we have concluded that the majority of the students are interested in consulting the links, and don't consider difficult the identification of the relevant sources of inspiration, if any, of their projects [Giordano 98]. Rather, it is the process of linking documents that is considered to be somewhat laborious. For this reason we are improving the interface for linking the items and we expect that the present linking hindrance will be overcome with the next version of StoryNet.

Now we turn to explain the method in practice. Let us assume that we are interested in designing the registration procedure to a training course. In this case it is quite natural to issue the query „course AND registration" to the Lotus Domino full

text search engine. This query behaves as a contextualized filter since „course“ refers to a story and „registration“ deals with an episode of this story. After processing the query, the engine finds the four items that seem to address some aspects relevant for query, that is:

I1) registration to a course organized in the convention center (episode)

I2) registration to vocational courses (story)

I3) application for registering to the first year of a vocational course (episode)

I4) application for registering to the second year of a vocational course (episode).

By subdividing the 150 items in 5 classes applying the proposed neural classification we find that I1 belongs to a class containing 18 items, whereas items I2, I3 and I4 belong to another class containing 25 items. Thus we are faced with 43 items, whereas we expect [see Sect. 4] that the relevant items be about three times the number of the items of the entry set, i.e., about 12 documents. Therefore, following the heuristic proposed in [Sect. 4], we further subdivide the above 5 classes into 5 inner sub-classes thus partitioning the 150 documents into 25 classes. In this case we obtain that I1 belongs to a class, say „A“, containing 5 items, I2 and I4 to a class „B“ containing 3 items, and I3 to a class „C“ containing 7 items. By inspecting these classes we find that:

- class A is especially dedicated to the problem of allocating space and staffing to support a course in a convention center, or ski runs and instructors for the ski school;
- class B deals with the problem of students' enrollment in vocational courses;
- class C deal with the problem of students' enrollment in courses held in the convention center.

Class A is of partial interest for the query issued. On the contrary, both classes B and C convey relevant aspects. Since the partial interest towards class A may be interpreted either statistically (i.e., about 3 of the 7 items of class A could be of interest) or as a fuzzy expression (i.e., each item of class A could have some aspect of minor interest) we decide to weight class A by 0.5. Classes B and C are weighted by 1, thus obtaining about 11 relevant items for the query. Since this number is of the same magnitude of the overall number of the expected relevant items (i.e., 12), we could decide to terminate the searching process.

Let us note that in this case we have taken advantage of the high precision of the entry set. Generally, we may have less precision. This causes a higher effort in identifying the relevant items among the ones of the entry set and increases the need for navigating in the neighborhood of the entry classes.

However, as pointed out in [Sect. 4], entering the web by a crisp query, as is the previous one, can obscure some other relevant aspects. Thus students with a broadening attitude could decide to repeat the searching process by issuing a more general query consisting of only one word, i.e., „registration“. After processing this

new query, the Lotus Domino engine finds 7 items: the four mentioned items I1, I2, I3 and I4 plus three new items. Two of these new items, say I5 and I6, do not produce any interesting consequence since they belong to the class A mentioned above. On the contrary, the third new item, say I7, deals with the registration at a distance to the activities organized in the convention center, and thus is useful to point out another aspect relevant to the query, i.e., the one of allowing people to register to a training course either by post or electronic mail. This aspect emerges because after partitioning the 150 documents into 25 classes we would find item I7 belonging to a class „D" whose seven items aim at providing services via data nets.

By applying the heuristics mentioned above and weighting the items of class D by 0.5, we would find that the number of the retrieved relevant items is about 15, which is not so close to the expected number of the overall relevant items (i.e., 21). Thus we could decide to continue the search by adopting the exploratory disposition typical of Web's navigators rather than the point of view derived from the traditional information retrieval methods. In this case the neural classification has to be considered as a way suggestive of interesting routes to discover relevant aspects unforeseen *a priori*.

Accordingly, we have to focus our attention not only on the classes to which the documents initially retrieved by the full text-searching engine belong but also on the neighboring classes, and in particular to the ones that are more tightly linked to the entry classes. [Fig. 5] shows the classes' geography around the entry classes A, B, C and D mentioned in the discussion above. These classes are surrounded by six other classes whose links significantly differ from zero. Let us note that the neural agent computes the links' weights Lir between classes i and r by using the formula introduced in [Sect. 4].

To increase precision, it may be useful to apply the following formula:

$$L_{ir} = [\Sigma_j \, Erj] / n_I$$

by imposing that j ranges over the set of the n_I relevant documents belonging to class i; this avoids to pass to classes that are neighbors of the initial ones but are not close to the relevant items. By reflecting on this geography and on the related descriptions it seems quite natural that one may perceive the suggestions emerging from the shared design memory as follows: „The main aspect of the problem of registration to a course deals with the procedures for managing the applications coming from users (see entry classes B and C). However, some special consideration should be given to the procedures that facilitate the registration of people who have taken part to previous courses (see classes E and F surrounding class C). Registration at distance could be supported (see entry class D and its neighboring classes G an H), whereas allocating necessary resources and people has not to be overlooked (entry class A, and its neighboring classes I and L)". Concrete suggestions for implementing such a scenario can be found in the items belonging to the classes already mentioned .

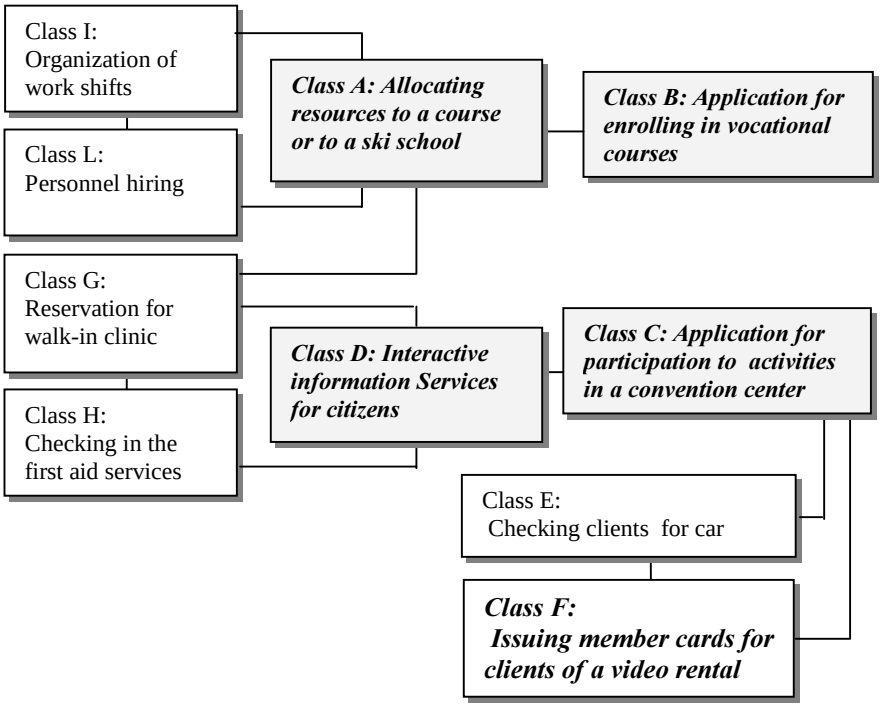

Figure 5: Class geography around the entry classes A, B, C and D. These classes are retrieved by the neural agent in response to the query „registration AND course".

7 Concluding remarks

The use of neural networks for assisting people in finding information is becoming increasingly diffused. Kohonen networks allow the classification of web documents in a two-dimensional setting that has proved useful for information retrieval. Recall of relevant documents may be increased if one navigates in the two-dimensional setting. Also Hopfield neural networks have been used with the aim of increasing recall [Chen 96]. However, all the existing neural methods suffer from the lexical nature of the input space and from the reduced dimensions of the classification setting. The method illustrated in this paper aims at overcoming these limitations by proposing an input space that consists of documents interrelated by influence links and by referring the classes identified by the Kohonen net to a n-dimensional setting. This allows fine-grained navigation to discover relevant documents and allows capturing some

meanings implied in the evolution of the issues and aspects dealt with in the documents.

The neural agent has been tested on a small scale set of documents produced for information systems specification, and has generated classifications deemed plausible and useful for guiding searches. We are currently working at large scale testing in the context of collaborative design assisted by webs of design cases and at testing heuristics for deploying the links and policies for document elimination. If performance of the neural agent scales up, the next step is to find more effective visualization techniques to reflect the classes' topology and for highlighting items belonging to multiple classes.

Moreover, some studies are being conducted for identifying if and when automatic descriptions of the classes can be made available to the users without introducing unnecessary biases in their perception and exploration process. Finally, the efficiency of an algorithm that avoids repeating neural classification from scratch is now under evaluation. In particular, what conditions of the documents' distribution in the space generated by the neural classification warrant a new global classification rather than a local re-arrangement when a document is inserted or deleted from the Web is another related issue now being investigated.

References

[Alexander 64] Alexander, C.: „Notes on the Synthesis of Form"; Harvard University Press, Cambridge (1964).

[Bourigault 92] Bourigault, D.: „Lexter, un Logiciel d'Extraction de Terminologie"; Colloque International de TermNet, Avignon (1992), France.

[Bruner 90] Bruner, J.: „Acts of Meaning"; Harvard University Press, Cambridge (1990).

[Carlson 89] Carlson, A. P.: „Hypertext and Intelligent Interfaces for Text Retrieval"; in E. Barrett (Ed.) The Society of Text (pp.59-76); The MIT Press, Cambridge (1989).

[Chalmers et al. 96] Chalmers, M., Ingram R., & Pfranger, C.: „Adding Imageability Features to Information Displays"; Proc. UIST'96, ACM Press, Seattle (1996), 33-39.

[Chen et al. 96] Chen, H. et al.: „A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval"; IEEE Trans. on Pattern Analysis and Machine Intelligence, 18, 8 (1996), 771-782.

[Faro & Giordano 97a] Faro, A., & Giordano, D.: „From Documenting Design to Design by Documenting"; Proc. SIGDOC'97, ACM Press, Salt Lake City (1997), 45-54.

[Faro & Giordano 97b] Faro, A., & Giordano, D.: „StoryNet: an Evolving Network of Cases to Learn Information Systems Design"; Proc. 5th BCS Conference on Information Systems Methodologies; Springer-Verlag, Preston, UK (1997).

[Foltz, 90] Foltz, P.W.: „Using Latent Semantic Indexing for Information Filtering"; in R.B. Allen (Ed.), Proceedings of the Conference on Office Information Systems, Cambridge, MA (1990), 40-47.

[Furnas et al. 87]. Furnas, G. et al.: „The Vocabulary Problem in Human-system

Communication“; Communications of the ACM, 30, 11 (1987), 964-971.

[Giordano 98] Giordano, D.: „Bridging Qualitative and Quantitative Approaches in Evaluating the Instructional Effectiveness of a Shared Design Memory“; Journal for Universal Computer Science, 4, 4 (1998), 349-381.

[Goffinet & Noirhomme-Fraiture] Goffinet, L. & Noirhomme-Fraiture, M.: „Automatic Hypertext Link Generation based on Similarity Measures between Documents“; Research Paper RP-96-034, Technical Report, FUNDP, University of Namur (Belgium), December 96.

[Kolodner 93]. Kolodner, J.: „Case Based Reasoning“; Morgan Kaufmann, San Mateo, CA (1993).

[Kohonen 89]. Kohonen, T.: „Self Organization and Associative Memory“; Springer-Verlag, Berlin (1989).

[Kohonen 96]. Kohonen T., Lagus K., Honkela T., Kaski S.: „Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration“; Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon , AAAI Press (1996), 238-243.

[Landauer & Dumais 97] Landauer, T. K., & Dumais, S. T.: „A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge“; Psychological Review, 104 (1997), 211-240.

[Levitt & March 88] Levitt, B. & March, J.: „Organizational Learning“; Annual Review of Sociology, 14 (1988), 319-340.

[MacKinlay et al. 91]. MacKinlay, J.D., Robertson, G.G., & Card, S.K.: „The Perspective Wall : Detail and Context Smoothly Integrated“; Proc. CHI’91, ACM Press, (1991), 173-180.

[Medin & Ross 89]. Medin, D.L., & Ross, B.H.: „The Specific Character of Abstract Thought: Categorization, Problem Solving and Induction“; in Advances in the Psychology of Human Intelligence, Vol. 5, (pp.189-223), Lawrence Erlbaum, Hillsdale, (1989).

[Salton 92] Salton, G.: „The State of Retrieval System Evaluation“; Information Processing and Management, 28, 4 (1992), 441-450.

[Salton et al. 96] Salton, G., Singhal, A., Buckley, C. & Mitra, M.: „Automatic Text Decomposition Using Text Segments and Text Themes“; Hypertext (1996), 53-65.

[Thagard & Holyoak 91] Thagard, P. & Holyoak, K.: „Why Indexing is the Wrong Way of Thinking About Analog Retrieval“; DARPA Proceedings on Case-Based Reasoning, 1991, 36-40.

[Wang & Rada 95] Wang, W. & Rada, R.: „Experiences with Semantic Net Based Hypermedia“; International Journal of Human-Computer Studies, 43 (1995), 419-439.

[Zizi & Lafon 95] Zizi, M. & Beaudouin-Lafon, M.: „Hypermedia Exploration with Interactive Dynamic Maps“; International Journal of Human-Computer Studies, 43 (1995), 441-464.