An Adaptive Neural Network Approach to Hypertext Clustering

Natalija Vlajic and Howard C. Card
Internet Innovation Centre
Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Manitoba, Canada R3T 5V6
vlajic@ee.umanitoba.ca, hcard@ee.umanitoba.ca

Abstract

The WWW is an on-line hypertextual collection, and a more sophisticated algorithm for Web page clustering may have to be based on combined term-similarity and hyperlink-similarity measures. It has been observed that nearly all currently employed techniques for document classification on the Web make use of textual information only. In addition, most of these techniques are incapable of discovering the real nature of the collection to which they are applied due to rather inefficient clustering algorithms employed. This paper describes a novel technique for hypertext clustering, called an adaptive hypertext clustering (AHC) algorithm. This algorithm has been derived from a modified neural network algorithm, and adjusted to the problem of combined term-similarity and hyperlink-similarity measures. The results presented in the paper show that AHC can be easily adapted to enable the most appropriate Web page classification within collections of various thematic and functional profiles, suggesting its main benefits over the traditional techniques.

1. Introduction

In the general case of on-line document collection, document clustering is shown to be the main precondition for sophisticated and efficient access to the information of interest. Thus, when dealing with a clustered collection, a new search can be confined only within the groups that appear to be sufficiently close to a given query, instead of comparing the query against each individual item. In addition to more rapid document retrieval, this strategy makes the collection very convenient for browsing operations, since each recognized cluster may be conceived as a compact representation of a number of documents.

It has been proven that a clustering technique aimed at enabling both rapid and accurate document retrieval has to be entirely independent of document distribution density (for more see [1], and [2]). This requirement has been inherited from the traditional libraries, and can be justified by the following. Although a library, for example, may contain hundreds of books on one topic, and just one book on another topic, a separate category has to be established for each, in order to obtain a satisfactory registration (file) system. Furthermore, a number of newly arrived books on any of the two topics may be expected to cause some changes within the corresponding category only, but should not effect the general organization of the overall file system.

Most of the existing digital libraries employ either *simple* or *complete link* clustering algorithms [1]. These algorithms exhibit no dependency on input data distribution density, and accordingly seem to be an appropriate solution to document classification tasks. However, the main disadvantage of single and complete link clustering is lack of learning, which renders them incapable of discovering the real nature of the information space to which they are applied. In addition, for every new item added to a previously classified collection they require a complete reclassification. That may be very computationally expensive and inefficient, especially for large collections, or collections that are frequently changing.

From the perspective of artificial neural networks (ANNs), document classification is a standard unsupervised learning problem. On the other hand, from the perspective of document clustering and retrieval, ANNs based on self-organization are appealing for several reasons. First, providing cluster identification, ANNs can undoubtedly enable cluster search and reduce search time. Furthermore, cluster identification is the key to efficient browsing. Finally, the generalization in terms of a previously clustered collection means that new documents can be automatically added and classified, without a complete reclassification as required with some other techniques in statistical pattern recognition. All this

suggests that the utilization of unsupervised ANN learning for document clustering tasks can offer a number of advantages.

2. Unsupervised ANN Learning for Document Clustering

Among the most widespread unsupervised ANN learning methods used are standard (hard) competitive learning (HCL), the self-organizing map (SOM) algorithm, and adaptive resonance theory (ART). (In the remainder of the paper we will refer to ART2, a class of adaptive resonance architectures that is capable of learning analog input patterns.) Although all are based on the so-called *winner-take-all* concept, these algorithms provide quite different results. HCL and the SOM algorithm, for example, are proven to be strongly dependent on the input data density, such that small variations in the distribution of training data may result in clusters of altered sizes and positions (for more see [2]). Therefore, it would not be appropriate to employ HCL or the SOM for the purpose of improved information clustering and retrieval. However, they might be very convenient for information browsing tasks.

The fundamental difference between the ART2 learning model and the other two algorithms based on the winner take all concept is related to the adjustment of the winning node for each new training pattern. In contrast to the SOM algorithm and to HCL, the ART2 model initiates the adjustment of the winning node only if it deems this node to be an acceptable match. Put another way, a category modifies its previous learning only if the input vector is sufficiently similar to risk a further refinement of its profile. This principle protects earlier gained knowledge from being eliminated by new learning, while enabling new learning to be automatically incorporated into the total knowledge of the system in a self-consistent way.

Although, theoretically, ART2 appears to be an appropriate solution to document clustering tasks, primarily due to the stable nature of its learning rule, it exhibits some significant limitations when utilized in practice [2]. In order to preserve the conceptual advantages of ART2, while overcoming its main disadvantages, we have proposed a modified version of ART2. In contrast to the original algorithm, our modified ART2 is based on a recursive learning procedure, employing a dynamically changing vigilance parameter. These novelties have been proven to ensure absolutely stable and hierarchical clustering, and thereby provide for highly efficient multi-level document retrieval. (A detailed description of modified ART2 and its properties are given in [2] and [3].)

Most clustering algorithms have a common disadvantage: they do not provide any information on the spatial relationships among discovered clusters. When applied to ANN for clustering tasks, including modified ART2, this would mean that there is no clear indication how one reference vector is positioned with respect to others. However, the knowledge of spatial relationships among reference vectors, and the corresponding Voronoi regions, could undoubtedly help to obtain a better insight into the nature and complexity of input data.

Competitive Hebbian (CHL) learning is an unsupervised learning algorithm, usually not employed on its own but in conjunction with other methods (see [4]). This algorithm does not change reference vectors at all, but merely inserts a number of topological connections, or edges, among the units of the network. Thus, it has been shown that modified ART2 in conjunction with CHL can perform accurate and stable clustering while preserving the topology of input data. Accordingly, these two algorithms are capable of discovering related or relevant groups of documents when used for information clustering purposes.

2.1 Experimental Results

Figures 1 to 4 present the clustering results obtained applying the SOM algorithm, modified ART2, and modified ART2 with CHL respectively, to the collection of 25 Web documents presented in Table 1. In all the cases the documents were represented by 25-dimensional word vectors, where each vector dimension corresponded to one of the words with the highest discriminatory power according to the modified TF/IDF model. (More details on the experiment could be found in [2].)

topic (category)	number of documents
tennis	6
volleyball	4
accordion	4
jazz	5
philosophy	3
neural networks	2
java	1

Table 1

Based on Figure 1, the SOM algorithm evidently provided results that were dependent on the input data density. Thus, while four out of twelve reference vectors were positioned within or close to the group on tennis, which had the highest distribution density, none of them was placed close to the group on java, which contained a single document. In fact, the document on java was placed in the same group with the two documents on neural networks. This implies that the SOM algorithm was not completely efficient in discriminating among the

main topics of the collection. However, it reflected very accurately thematic proximity among the discovered groups: the nodes related to the documents on tennis and volleyball, jazz and accordion were placed next to each other in the map, forming larger clusters on sports and music.

Figure 2 shows that modified ART2 successfully coped with the problem of a considerably non-uniform distribution, and for seven output nodes it correctly classified all 25 documents into seven basic groups.

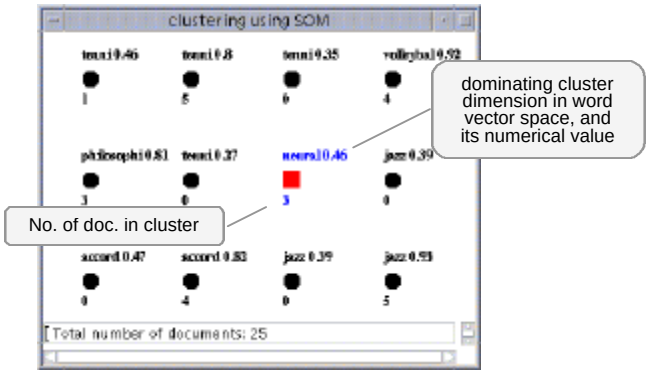

Figure 1 Clustering results obtained using SOM for the Web document collection from Table 1

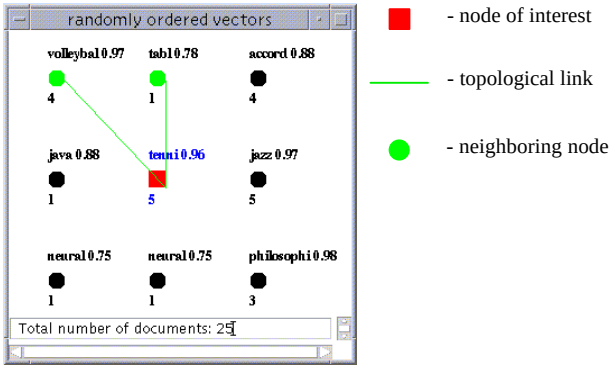

Figure 3 Clustering results obtained using modified ART2 with CHL for the Web collection from Table 1; Reference to tennis

3. Hypertext Clustering

The WWW is a global hypertext. In other words, it is hypertext enriched with the ability to connect with sites around the world. Accordingly, the content and importance of each individual Web document (page) is a compound function of its textual and hyper portions, in which the hyper part is the information determined by the position and function of the Web page with respect to the surrounding Web space.

Modified ART2 in conjunction with CHL, as indicated in Figures 3 and 4, was able to recognize certain relatedness among the discovered groups, when the number of output nodes mismatched the number of actual clusters. Thus, for the case when the group on tennis was the group of interest, the network pointed to the groups on volleyball and table tennis as being the topological neighbors, i.e. thematically closest nodes. Similarly, when the group on accordion was the group of interest, the groups on jazz, philosophy, and java were found to be in its spatial (thematic) neighborhood.

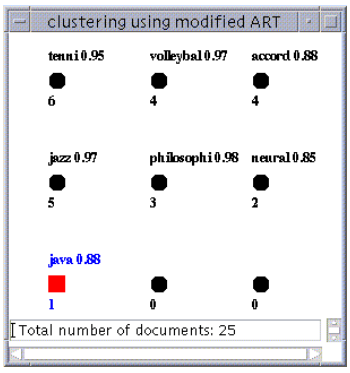

Figure 2 Clustering results obtained using modified ART2 for the Web document collection from Table 1

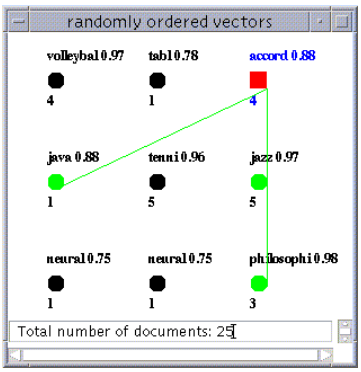

Figure 4 Clustering results obtained using modified ART2 with CHL for the Web collection from Table 1; Reference to accordion

The functionality of a Web page is exclusively based on the features it exhibits, and the same functional categories (such as *organizational home pages* (OHP), *departmental home pages* (DHP), *link* or *index pages* (IP), *content pages* (CP), *pages of special content* (PSC)) can be found in different thematic domains. Experimenting with Web documents of various profiles and sizes we proved that characteristics such as page length, number of images and links and the percentage of upward, downward and outward links, etc. gave valuable

information on Web page functionality, and thereby provided for more accurate clustering.

In particular, we have introduced 12 *hyper dimensions* for the purpose of hyper-dimension based clustering. Each dimension is defined with a function that takes one of the features presented in Table 2 as its argument. Numerical (quantitative) values related to these features can be calculated using information retrievable directly from the corresponding html source codes, which presents the main advantage of our technique over some other techniques for hypertext classification.

node type	OHP	DHP	LP	CP	PSC
depth	Sm	Me	Me	Me/Lg	Me/Lg
size	Sm/Me	Sm/Me	Me	Me/Lg	Me/Lg
links	M/S	M/S	M	S/N	S/N
% outward links	S/N	S/N	M	N	N
% upward links	N	S	S	S/N	S/N
% crosswise & downward links	M	S	S	S/N	S/N
anchors	S/N	S/N	N	S/N	S/N
images	M/S	S	N	S/N	S/N
e-mail or news link	M/S	M/S	S	S/N	S/N
other protocols	N	N	S/N	N	S
.java, .exe, … files	N	N	N	N	S

Sm – small, Me – medium, Lg – large
M – many, S – several, N - none

Table 2

3.1 Adaptive Hypertext Clustering (AHC) Algorithm

The main problem of combined hyper-text clustering is regarding vector representation of Web pages. Namely, Web documents may be of various lengths (sizes), and in order to enable their comparison the text-related dimensions (word vectors) are required to be normalized according to (1).

$$x_{normalized\_i} = \frac{x_i}{\sqrt{x_1 + x_2 + \ldots + x_n}} \qquad (1)$$

Word vector normalization actually means that only the relative information on the content of a Web page is preserved. Accordingly, the numerical value of each word vector dimension describes the percentage to which the corresponding concept is present in the document. If the presence of one concept is significant it may imply that the presence of all the other concepts in negligible. From a mathematical point of view, normalization based on (1) means that all word vectors lie on the unit hypersphere in n-dimensional word vector space, as depicted in Figure 5 a).

In contrast to text-related dimensions, hyper dimensions cannot be normalized in the same manner, since most of these are not mutually dependent of related. (For example, if a Web page has significant depth weight it does not imply that its outward link density weight is insignificant, etc.) However, hyper dimensions require a different type of normalization, in order to maintain their numerical values within the limits 0 to 1, and make the corresponding vectors comparable. (For more see [2].) Consequently, hyper vectors may take any position within the unit hypercube in the 12-dimensional hyper space, as illustrated in Figure 5.b).

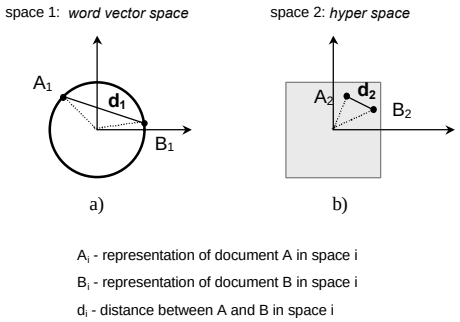

Figure 5 Web page representation in word vector space and in hyper space

It is apparent that due to the different nature of word vector and hyper dimensions it would be inappropriate to simply combine them in a unique hyper-text space. Therefore, a Web page clustering algorithm intended to utilize all available information has to incorporate simultaneous performance in separate word vector and hyperspace. In addition, the overall distance between two Web pages has to be defined as a compound function of the distances in the two spaces.

The modified ART2 algorithm is proven to provide perfectly stable learning (clustering) for a sufficiently small value of the so-called dynamic parameter (for more see [2]). This property makes modified ART2 very convenient for the problem of hyper-text classification, since an algorithm can ensure stable multi-space clustering only if it guarantees stable clustering within each individual space.

In our work we have introduced an adaptive hypertext clustering algorithm (AHC), which may be defined as modified ART2 adjusted to the problem of multi-space vector representation. In particular, the main difference between our earlier modified ART2 and AHC is in their respective measures for the distance between training vectors. While the earlier modified ART2 uses a Euclidean metric, the AHC algorithm employs the formula given in Figure 6 as its distance measure.

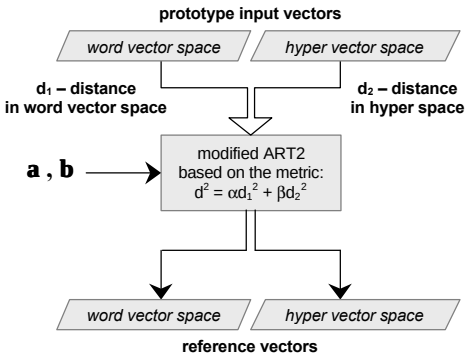

Figure 6 System for adaptive hypertext clustering

Parameters $\alpha, \beta \in [0,1]$ (Figure 6) are adjustable, and they determine the nature of clustering. For example, if $\alpha=1$ and $\beta=0$ then AHC produces pure text-based clustering, and Web pages are categorized according to their content. However, if α and $\beta \neq 0$, both word vector and hyper dimensions influence the grouping of Web pages. In that case the ratio α/β determines if textual or structural information is more decisive. Accordingly, the AHC algorithm has the ability to switch between text-based and hyper dimension-based clustering, and thereby provide the most appropriate classification within collections of various thematic and functional profiles.

3.2 Experimental Results

This experiment concerned the data collection presented in Table 3. The collection consisted of documents on four different topics (tennis, jazz, volleyball, and neural networks), and the group on neural networks consisted of two functional subcategories.

topic (category)	number of documents
tennis	5
jazz	5
volleyball	5
NN theory	5
NN companies	5

Table 3

Figures 7 and 8 present the dendograms (search trees) learned with AHC, when operating in two different modes. (The documents of the collection from Table 3 are represented by D_i (i=1,..,25), and the corresponding groups are indicated with the shadowed vertical strips.)

It can be observed that pure text-based clustering (Figure 7) provided perfect separation among the main thematic categories, but it failed to recognize the subgroups within the group on NN. (The results regarding pure hyper dimension based clustering are presented in [2]. They show that hyper dimension based clustering was more successful in discriminating among the functional subgroups of the group on NN, but it failed to recognize the main thematic categories.) However, combining the information on the content and functionality of Web pages (Figure 8), the AHC algorithm properly identified both the main groups and the subgroups.

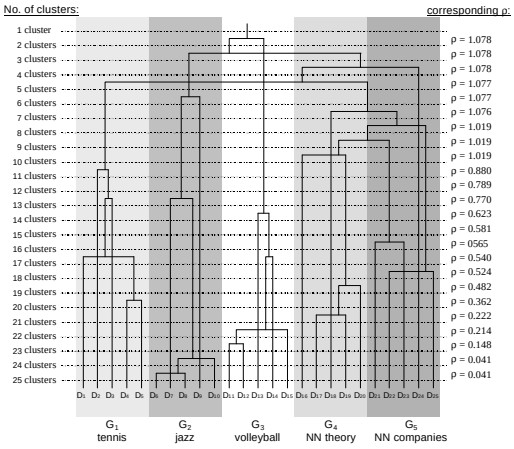

Figure 7 Dendogram obtained with AHC (**a=1, b=0**) on the collection from Table 3

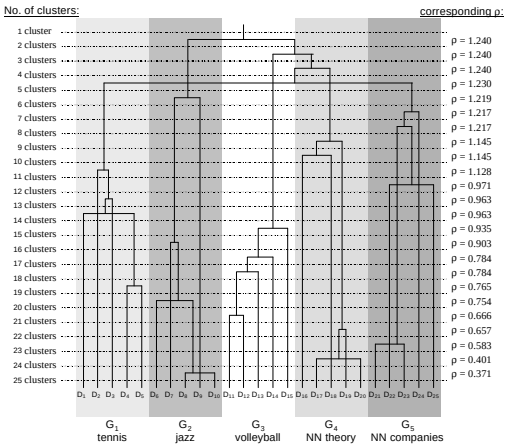

Figure 8 Dendogram obtained with AHC (**a=1.0, b=0.6**) on the collection from Table 3

References

[1] C. J. van Rijsbergen 1975, *Information Retrieval*, Boston: Butterworths.

[2] N. Vlajic 1998, *Adaptive Algorithms for Hypertext Clustering*, M.Sc. Thesis, University of Manitoba, Canada.

[3] N. Vlajic and H. C. Card, *Adaptive Algorithms for Hypertext Clustering*, International Conference on Computational Intelligence for Modeling Control and Automation CIMCA'99, February 1999, Vienna, Austria.

[4] B. Fritzke 1997, *Unsupervised Ontogenic Networks*, IOP Publishing Ltd. and Oxford University Press: Handbook of Neural computation, Release 97/1.