Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 41 (2005) 587–598
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Exploration of textual document archives using a fuzzy hierarchical clustering algorithm in the GAMBAL system

Vicenç Torra [a,*], Sadaaki Miyamoto [b], Sergi Lanau [a]

[a] Institut d'Investigació en Intelligència Artificial—CSIC, Campus UAB s/n, 08193 Bellaterra, Catalonia, Spain
[b] Institute of Engineering Mechanics and Systems, University of Tsukuba, Ibaraki 305-8573, Japan

Received 25 August 2003; accepted 23 January 2004
Available online 5 March 2004

Abstract

The Internet, together with the large amount of textual information available in document archives, has increased the relevance of information retrieval related tools. In this work we present an extension of the Gambal system for clustering and visualization of documents based on fuzzy clustering techniques. The tool allows to structure the set of documents in a hierarchical way (using a fuzzy hierarchical structure) and represent this structure in a graphical interface (a 3D sphere) over which the user can navigate.

Gambal allows the analysis of the documents and the computation of their similarity not only on the basis of the syntactic similarity between words but also based on a dictionary (Wordnet 1.7) and latent semantics analysis.

Keywords: Information retrieval; Hierarchical clustering; Fuzzy clustering

1. Introduction

Searching on the Internet and in archives of documents is increasingly becoming a frequent activity by most people. Due to this, information retrieval tools and methods have gained importance, and research is pushed forward, to increase the performance and friendliness of existing systems. Nevertheless, information retrieval has a long tradition and has been studied for a long time. See Salton and McGill (1983); Van Rijsbergen (1979) for some classical references on information retrieval and Baeza-Yates and Ribeiro-Neto (1999), Crestani, Vegas, and de la Fuente (2003) and Miyamoto (1990) for a state-of-the-art descriptions of the field. Crestani and Pasi (2000) is a useful collection of recent studies of soft computing applications to the field.

At present research on information retrieval gathers a broad set of interests. For example, to name a few, it encompasses work on clustering and classification (Ko, Park, & Seo, 2004; Nieto Sánchez,

* Corresponding author. Tel.: +34-93580-9570; fax: +34-93580-9661.
E-mail addresses: vtorra@iiia.csic.es (V. Torra), miyamoto@esys.tsukuba.ac.jp (S. Miyamoto).

0306-4573/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2004.01.001

Triantaphyllou, & Kraft, 2002, Soonthornphisaj & Kijsirikul, 2004), filtering systems, ranking methods (and also methods to combine/aggregate rankings (Bosc, Liétard, & Pivert, 2003; Herrera-Viedma & Peis, 2003)) and on developing test collections (Bailey, Craswell, & Hawking, 2003). Finally, a relevant line of research is on tools for modeling the processes related with information retrieval (e.g., using multisets (Miyamoto, 2003)). In this work we focus on clustering and visualization.

Due to the fact that a simple search in any search engine usually retrieves a large collection of documents, clustering becomes an essential tool. In fact, some of the recognized search engines (e.g., Alltheweb, 2004) include nowadays categorizations with the results of a search. In this way, instead of visiting all pages, a user can select first which cluster matches best her interests and forward the corresponding links. By the way, navigation in such clusters is still a difficult task because similarities among clusters are not represented and interfaces are not user friendly.

In fact, clustering methods have been intensively studied in information retrieval for textual documents (Agosti, Crestani, & Pasi, 2001; Kohonen, 1997; Merkl & Rauber, 2000). Nowadays, the research literature describes several clustering approaches based on alternative assumptions about the clustering process (e.g., agglomerative vs. divisive methods), the way documents are represented and how similarity is computed. For example, Tombros, Villa, and Van Rijsbergen (2002) uses agglomerative methods and Known and Lee (2003) and Merkl and Rauber (2000) use divisive ones. Kraft, Chen, and Mikulcic (2000) compares methods of both approaches. Most advanced methods include a graphical visualization system (see Leide, Large, Beheshti, Brooks, & Cole, 2003 for a study on visualization and navigation). This is the case, for example, of the systems described in Crestani et al. (2003), Merkl and Rauber (2000), Rodden (2002) and Stan and Sethi (2003).

As the number of available documents nowadays is large, hierarchical approaches are better suited because they permit categories to be defined at different abstraction levels. Moreover, divisive methods with a smaller computational cost are also preferable. Due to this, the HSOM (Hierarchical Self-Organizing Maps or Hierarchical SOM) system (Merkl & Rauber, 2000) can be considered as the most relevant approach because it offers a hierarchical decomposition of the set of documents embedded in a graphical interface so that the user can navigate through the interface to select the most appropriate document. The eID (Stan & Sethi, 2003) is a similar exploration system but for image databases.

In this work we present a novel tool for navigation in textual document archives. The system offers a method for document categorization that is tightly embedded into a visualization system. On the one hand, the system includes a clustering method for textual documents. On the other hand, it includes a visualization system that offers a 3D representation of the clusters (on the surface of a sphere) in terms of their centroids. Such a representation allows the user to navigate over the clusters and select the one that suits him the most. Then, the description of the cluster is supplied and, if required, a refinement of such a cluster is computed. Such refinement consists in the computation of a partition of the documents in the cluster. Then, the refinement will be represented in another graphical interface.

Our system uses a fuzzy clustering technique so that elements can belong to multiple clusters with different membership degrees. Such a clustering technique fits well with our visualization system because it uses a *continuous space*. This is, we consider the surface of the sphere as a continuous region because there are no buckets or cells as in the case of HSOM (HSOM considers a finite number of cells) or other neural network based systems.

The fuzzy clustering system and its visualization tool is currently a working prototype for the Gambal system interface. The Gambal system is an interactive environment for information retrieval that clusters and visualizes sets of HTML documents (either directly retrieved from the web or supplied by Google (Google, 2004) after a particular search). Nevertheless, the system can also be used to represent other kinds of data, such as images or numerical data.

The structure of this paper is as follows. In Section 2 we give a short overview of the Gambal system. Then, in Section 3 we introduce our new approach for clustering and visualization. Section 4 describes some

experiments and analyzes the results obtained. This paper is completed in Section 5 with some conclusions and a description of future work.

2. The Gambal system: an overview

Gambal (*GenerAdor d'estrats per a recuperacio d'inforMacio BAsat en metodes de cLassificacio*), or, in English, (*clustering based stratum generator for information retrieval*) is a system for accessing, clustering and visualizing HTML documents. The system encompasses several tools. Some of them are described below as separated entities (see Lanau, 2003 for details).

Crawling the web: Two alternative mechanisms can be used. The user selects the appropriate option. One mechanism consists in gathering web HTML documents from an initial file of URL and then extending the search to pointers in the corresponding documents and so on. The second mechanism consists in accessing web pages retrieved by Google for a query. In this case, the user gives a query to the system. Then, this query is forwarded to Google and the results from Google are retrieved and visualized by Gambal.

For each retrieved HTML document, the language used is determined and words are preprocessed accordingly (stop words are removed and a stemming algorithm is applied). At present, three languages are considered: Catalan, Spanish (official languages in Catalonia) and English. Stemming algorithms are currently applied only to texts detected as English.

Document representation: All documents of interest are transformed into an unified vector representation. At this point, non-relevant words (frequency under a given threshold) are removed. Two representations can be selected by the user. The normalized frequency (TF) and the inverse document frequency (IDF). See e.g., Baeza-Yates and Ribeiro-Neto (1999) for details. Additionally, inverse indices are built. These indices are used so that a user can access our database and retrieve documents from the words appearing in these documents.

Similarity computation: Three alternative ways are considered for computing similarity between documents. The first one consists in the comparison of the lexical elements in the two documents (due to the internal vectorial representation, this is just an elementwise comparison). In this case, Euclidean distance is used. The second one compares words using a dictionary (Wordnet 1.7 (Fellbaum, 1998; Wordnet, 2004)). The third one compares words taking into account latent semantics analysis (LSA, 2004). It has to be said that the LSA decomposition is restricted to the available web pages instead of considering a large corpus of documents as in LSA (2004) and related works. Wordnet is only applied to English texts (although there exists a EuroWordNet (2004) that includes lexical information for both Catalan and Spanish).

Similarity between words is based on the shortest path in WordNet hierarchy of concepts. This approach was previously used in Mandala, Tokunaga, and Tanaka (2000).

Clustering and/or visualization: Three clustering/visualization methods were implemented: (i) Hierarchical Spherical Clustering (HSC); (ii) Self-Organizing Maps (SOM); (iii) Hierarchical Self-Organizing Maps (HSOM). HSC (see Torra & Miyamoto, 2002a, 2002b) is based on the hard (crisp) c-means algorithm on a Sammon's map (a method for multidimensional scaling). SOM and HSOM are described, respectively, in Kohonen (1997) and Merkl and Rauber (2000). SOM defines a crisp partition on a grid structure in such a way that neighbors contain similar documents and HSOM is a hierarchical variation of the former (a cell in the structure can be split into a new grid).

Visualization tools permits us to navigate on the hierarchy (zoom and rotation is available for the HSC interface), change the level of detail (available in both HSC and HSOM), and click and display particular documents (for all three interfaces). When a document is selected, its HTML page is loaded and displayed to the user. Additionally, documents similar to the one clicked are also listed in a sensitive list of URL pages. In this way, the user can select any of them for display.

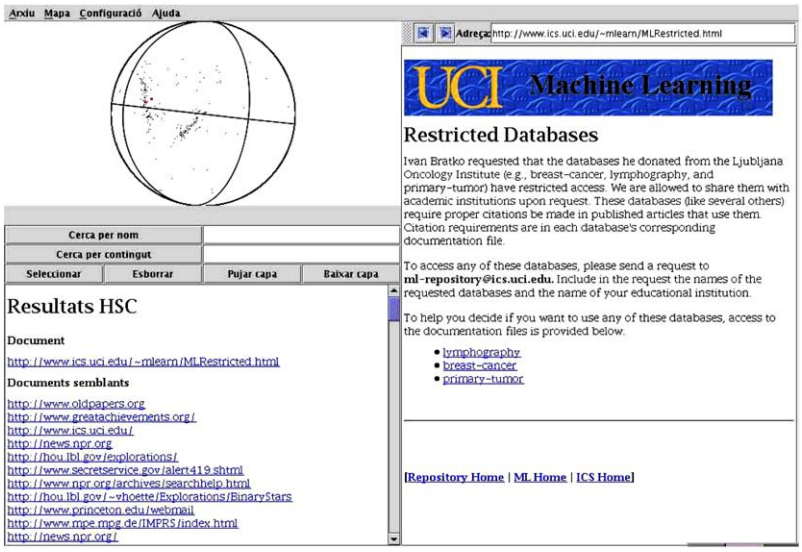


Fig. 1. The Gambal system for information retrieval and web clustering. A document has been selected in the map, corresponding to the Machine Learning UCI repository.

Fig. 1 gives a snapshot of the Gambal system. The user interface of the system is in Catalan. On the left-hand side, the figure shows a particular layer of the HSC (with a representation of the documents). One document has been selected (the one corresponding to the Machine Learning UCI repository) and, then, a list with the selected document and similar ones is given (in the left-hand side, bottom). On the right-hand side, the figure displays the selected web page.

3. Fuzzy clustering for indexing

This section describes our approach of document clustering using fuzzy *c*-means. We start giving a motivation of such approach and then describing our proposal for building hierarchical fuzzy clustering. The section finishes with a comparison with our previous approach in HSC (Torra & Miyamoto, 2002a, 2002b).

3.1. Motivation

Fuzzy clustering (Kraft, Bordogna, & Pasi, 1999) is a relevant tool for information retrieval. As a document might be relevant to multiple queries, this document should be given in the corresponding response sets. Otherwise, the users would not be aware of it. Due to this, fuzzy membership, and therefore, fuzzy clustering, seems a natural technique for document categorization.

Moreover, the fuzzy approach fits well with the continuous representation we use for visualization. This is, we use the surface of a sphere to represent the objects. In particular, we represent on it the centroids of the clusters. Then, in this continuous space, documents are located on the surface according to their membership to clusters. In particular, a document is located on the centroid (or very near to it) if the membership to the cluster is one and farther away when membership decreases. According to this, fuzzy clustering and our continuous representation support each other.

3.2. Fuzzy clustering

We have based our approach in standard fuzzy clustering techniques. In fact, we have used the fuzzy c-means algorithm, a well-known method that generalizes the crisp clustering algorithm k-means so that partial membership is allowed. In this way, elements can belong to several clusters. Multiple membership is modeled considering fuzzy partitions. Thus, objects have membership degrees to all clusters in the [0,1] interval (0 stands for no membership and 1 for total membership) in such a way that the summation of all the memberships is 1. Fuzzy c-means is a method to find such fuzzy partition from a set of data. We give below a short overview of the method, see Klir and Yuan (1995) or Miyamoto and Umayahara (2000) for a more detailed description.

To formalize the method, we need to consider p-dimensional objects in X (in our case, documents represented in a p-dimensional space), and fuzzy partitions on X. Let $A = \{A_1, \ldots, A_c\}$ denote a fuzzy partition of X into c clusters. This is, for all $x_k \in X$ it holds:

$$\sum_{i=1}^{c} A_i(x_k) = 1 \quad \text{for all } x_k \in X$$

where $A_i(x_k) \in [0,1]$ is interpreted as the membership of the kth element to the ith set.

Now, fuzzy c-means is to find the fuzzy partition A that minimizes the following expression:

$$J(A,V) = \sum_{k=1}^{n} \sum_{i=1}^{c} (A_i(x_k))^m \cdot \|x_k - v_i\|^2$$

subject to the following constraints:

$$M_{\mathrm{f}} = \left\{ (A_i(x_k)) \mid A_i(x_k) \in [0,1], \sum_{i=1}^{c} A_i(x_k) = 1 \text{ for all } k \right\}$$

Here, $\|\cdot\|$ corresponds to the Euclidean distance, $V = \{v_i\}_{i=1,\ldots,c}$ are the centers of the clusters and m is a real number ($m \geqslant 1$) that influences the membership values. With $m = 1$, the solution is a crisp partition and, then, the larger is m, the more fuzzy the clusters we obtain.

To find A and V that minimize this objective function constrained in M_{f}, the following algorithm is used:

Step 1: Generate an initial A and go to step 3 or an initial V and go to step 2
Step 2: Solve $\min_{A \in M_{\mathrm{f}}} J(A,V)$ computing (A_{ik} stands for $A_i(x_k)$):

$$A_{ik} = \left(\sum_{j=1}^{c} \left(\frac{\|x_k - v_i\|^2}{\|x_k - v_j\|^2} \right)^{\frac{1}{m-1}} \right)^{-1}$$

Step 3: Solve $\min_V J(A,V)$ computing:

$$v_i = \frac{\sum_{k=1}^{n} (A_{ik})^m x_k}{\sum_{k=1}^{n} (A_{ik})^m}$$

Step 4: If the solution does not converge, go to step 2; otherwise, stop

This algorithm does not give an optimal solution but only a suboptimal one. See e.g. Miyamoto and Umayahara (2000) for details.

3.3. Hierarchical fuzzy clustering in Gambal

The Gambal approach for clustering and visualization consists on building a dendrogram (a hierarchy) of the original set of multidimensional data on the surface of a set of concentric spheres. In this formulation, the inner sphere corresponds to the root of the dendrogram (this is the node where all data is gathered together). Then, in the first sphere, we have a first partition of the elements, and in subsequent spheres we have more refinements of such partitions. In this way, the most external sphere, if generated, corresponds to the leaves in the dendrogram. This is so because leaves correspond to clusters with single documents (or, in general, a single information datum).

The process for building the hierarchical fuzzy clustering and the corresponding representation starts with the construction of a fuzzy partition and the location of the corresponding centroids on the surface of a sphere. Then, when a cluster is selected, a fuzzy partition of such cluster is computed and a new sphere is constructed where all previous centroids are included together with the new ones (the ones corresponding to the new clusters). The location of the initial centroids is based on the Sammon's map algorithm (see e.g., Kohonen, 1997). The location of subsequent points is based on the location of the previously located ones and on the memberships of the new points to the previous clusters.

We give a more detailed description of this process below:

Step 1: Apply fuzzy clustering to the complete set of documents and obtain for each cluster a centroid. An heuristic approach is used for selecting an appropriate number of clusters.

Step 2: Represent the centroids on the sphere using an adaptation of Sammon's mapping for a spherical surface. We use the Sammon's mapping described in Torra and Miyamoto (2002a, 2002b). In short, the centroids in the original high dimensional space are located on the surface of the sphere in such a way that distances on the surface are roughly proportional to the distances on the high dimensional space. Formally speaking, let x_i denote data in the original space and z_i their representation on the surface of the sphere, then if $d^{\mathrm{o}}(x_i, x_j)$ is the distance between x_i and x_j in the original space and if $d^{\mathrm{s}}(z_i, z_j)$ is the distance of the same data on the surface of the sphere (i.e., the angle that define z_i and z_j), then the goal is to locate all data x_i on the surface (or, equivalently, to find z_i) so that the following expression is minimized:

$$\sum_{i>j} \frac{\left(d^{\mathrm{o}}(x_i, x_j)/d^{\mathrm{o}}_{\max} - d^{\mathrm{s}}(z_i, z_j)/d^{\mathrm{s}}_{\max}\right)^2}{d^{\mathrm{o}}(x_i, x_j)/d^{\mathrm{o}}_{\max}}$$

Here, $d^{\mathrm{o}}_{\max}$ is defined as $d^{\mathrm{o}}_{\max} = \max_{i,j} d^{\mathrm{o}}(x_i, x_j)$ (i.e., the maximum distance on the original space) and, similarly, $d^{\mathrm{s}}_{\max}$ is defined as π (i.e., the maximum distance on the surface).

Step 3: When a cluster is selected, the following steps are applied:

- *Step 3.1:* Fuzzy clustering is applied to associated documents. Although standard fuzzy c-means is used at this point, clustering methods considering element importance (Keller & Klawonn, 2000) are also appropriate. In such case, elements with a small membership would have less influence on the final partition than elements with a large membership.
- *Step 3.2:* Compute for all new centroids their membership to already located clusters.
- *Step 3.3:* Use these values to locate the new centroids on the surface of the sphere. Note that for doing so, we need the fuzzy clustering algorithm to not only compute the fuzzy partition but also permits to compute membership values for any arbitrary data point (in the original high dimensional space). Fuzzy c-means (among others, e.g., entropy based fuzzy c-means) permits such computation. See Bezdek (1981) or Miyamoto and Umayahara (2000) for details.

Naturally, this procedure reduces the number of displayed information, and, therefore, it is possible to display large data sets. However, this is not a restriction because selecting a relevant set, the user can list its

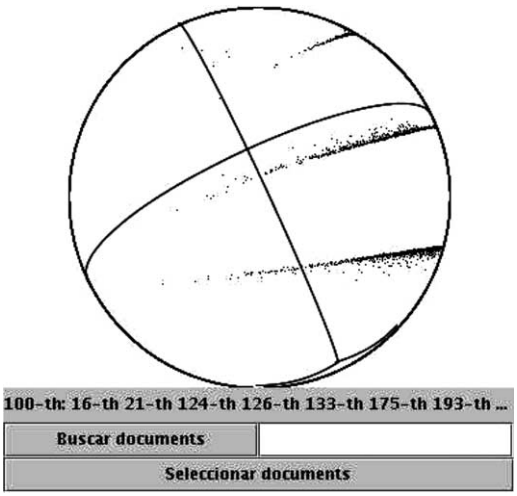


Fig. 2. Original Sammon's map (Abalone data file).

documents or force its refinement and move to smaller sets (sets with less documents). Additionally, as said before, the proposed procedure integrates well with the continuous representation on the sphere because at any time (on the first layer or in subsequent ones) documents might be located at any position of the surface.

3.4. Comparison with HSC

The fuzzy clustering approach suggested in Torra, Lanau, and Miyamoto (2003) is based on the Hierarchical Spherical Clustering (HSC) described in Torra and Miyamoto (2002a, 2002b). Nevertheless, substantial differences can be found in both methods. The main one corresponds to the clustering process itself. In the system described here, we apply a fuzzy clustering technique while HSC uses a crisp one. Additionally, the fuzzy clustering algorithm only uses the data in the original high dimensional space (the space where documents are represented) while the crisp technique uses information about the region where the cluster and subclusters are located.

Besides of the particular clustering algorithm, the procedure for computing the whole hierarchy also diverges from the HSC approach. While in HSC, all documents are first located on the surface of the sphere using a Sammon's map and then clustered; at present, we proceed first with the clustering. Then, only the corresponding centroids are located on the sphere.

This change on the strategy is motivated by the results we obtained with several data files. The analysis of our clustering results showed that the HSC approach is appropriate when the number of elements is of reasonable size and when the documents and the information levels satisfy some properties about their distribution in the space. Otherwise, the visualization of the clusters is not easy and the Sammon's map does not give a good performance. A typical problem with files of large dimension is that data tends to be accumulated in a subregion of the surface. In this case, documents are difficult to be distinguished and similarities between documents can not be properly exploited. Fig. 2 shows this case for the Abalone data file (from Murphy & Aha, 1994). Each dot in the figure corresponds to a record in the file. From our point of view, some of the inconveniences of the original HSC approach are due to a large dimensionality reduction and to the presence of too many elements. Therefore, HSC was not suitable when large sets of documents were considered. Our new approach tries to solve this drawback.

4. Experiments and analysis

Our system has been tested using three different kind of data. We have applied it to a set of web pages downloaded from Internet. We also used data sets available from public repositories: (i) data files from the UCI repository (Murphy & Aha, 1994) (we used the *iris*, *abalone* and *ionosphere* data files) and (ii) data corresponding to the 1990 Edition of the CIA World Factbook. These latter data were previously used in Merkl and Rauber (2000) to analyze the Hierarchical SOM and they are available at CIA WorldFact (2004).

Here we describe the results obtained using the CIA World factbook data because they are publicly available and correspond to textual information. The original CIA World factbook data corresponds to textual records for 245 *countries* and *regions* (e.g., oceans). Each record includes information about one country on several different categories (e.g., Geography, People, Government, ...). To apply our approach, data has to be preprocessed. Processed files can be downloaded from CIA WorldFact (2004) and they consist on frequency vectors for each record. Four files are available, where differences correspond to different lengths (1056, 690, 977 and 889) of the vectors. Here we have used the file with the largest dimension, 1056.

In our experiments, we have tuned the system (heuristically) starting with 10 clusters and then, in subsequent steps, using the number of clusters equal to 4.

In Table 2 we show two snapshots of the first sphere (two projections). This is, the result of the Sammon's map applied to the centroids of the fuzzy c-means. Table 1 gives a description of the two clusters (clusters 0-1Lclass and 3-1Lclass) located in the center of the figure at the bottom of Table 2. Identifiers i-1Lclass indicate the ith cluster in the first layer (first sphere). It can be observed that these clusters correspond, respectively, to the oceans and to some (small) island territories (except for the last record in the list). These two clusters are the ones that appear more near in the surface of the sphere. Therefore, they are clusters that are similar.

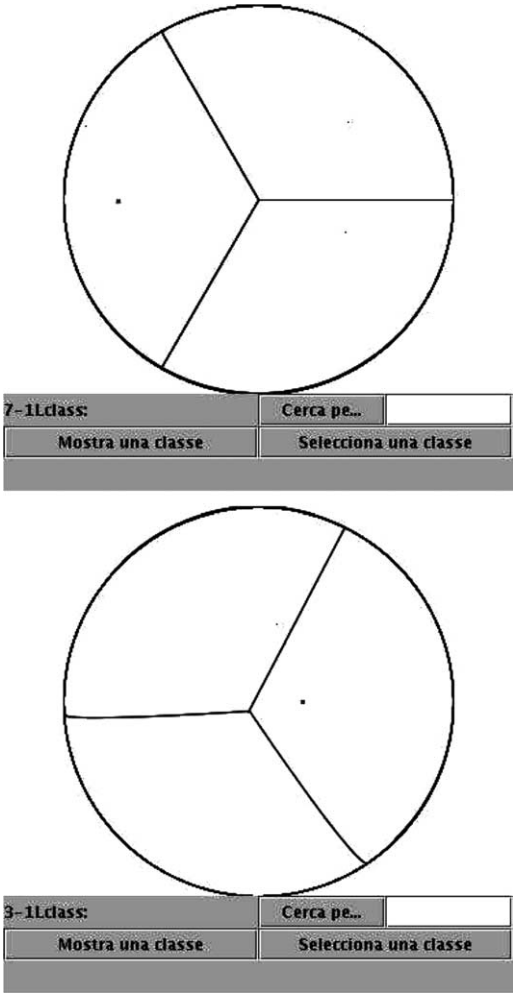
Figure in the top of Table 3 gives a snapshot of the same sphere when the cluster 4-1Lclass is located in the center of the figure. This cluster, described in Table 4 top corresponds to most developed countries. Then, figure in Table 3 bottom gives a snapshot of the same region of the sphere once this cluster has been split into four new clusters. Now, the identifiers i-2L have been used. They indicate that the cluster belongs to the second layer (second sphere). The figure only includes three of these clusters (the ones with identifiers 1-2L, 2-2L and 3-2L) because the fourth one (4-2L) has been located in the other side of the figure near the cluster 5-1Lclass. This location is natural. Note that cluster 4-2L is also a subset of 5-1Lclass (due to the use of fuzzy clustering countries can belong to two clusters at the same time). The description of cluster 4-1Lclass and of their four subclusters 1-2L, 2-2L, 3-2L and 4-2L is given in Table 4. It can be seen that the countries in the same subcluster are somehow similar to each other being 4-2L the most dissimilar of the four clusters.

The example illustrates how the approach is applied and proves its applicability to the field of information retrieval. The results obtained with our clustering algorithm are similar to the ones obtained with

Table 1
Description of clusters 0-1Lclass and 3-1Lclass

Cluster	Countries and regions
0-1Lclass	Ashmore and Cartier Islands, Baker Island, Bassas da India, Bouvet Island, Clipperton Island, Coral Sea Islands, Europa Island, French Southern and Antarctic Lands, Glorioso Islands, Heard Island and McDonald Islands, Howland Island, Jan Mayen, Jarvis Island, Johnston Atoll, Juan de Nova Island, Kingman Reef, Midway Islands, Navassa Island, Palmyra Atoll, Paracel Islands, South Georgia and the South Sandwich Islands, Spratly Islands, Tromelin Island, Wake Island, Iraq—Saudi Arabia Neutral Zone (land)
3-1Lclass	Antarctic ocean, Arctic ocean, Atlantic ocean, Indian ocean, Pacific ocean

Table 2
Sammon's representation of the centroides of each cluster: the two sides of the sphere

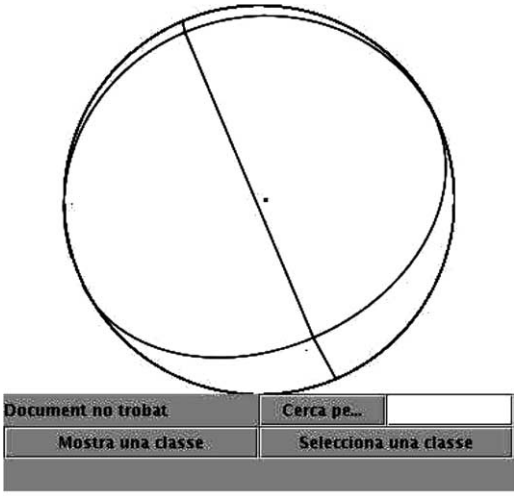


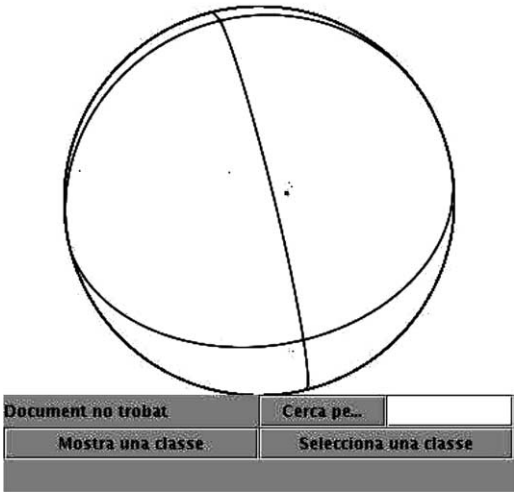
the HSOM, SOM (in such case, the space is divided into buckets—cells—and there is no multiple membership) and by similar clustering algorithms. However, our approach integrates well clustering (and, particularly, hierarchical clustering) and visualization. In particular, it exploits the continuity of the surface of the sphere locating new centroids so that their membership is represented in terms of the distance to already located centroids. In this way, the similarity between clusters (and the categories they represent) can be seen graphically. Moreover, our system is user friendly for accessing the documents.

5. Conclusions and future work

At present we have considered for experimentation data from the UCI repository, data from the CIA World Fact (CIA WorldFact, 2004) and small sets of web pages. The largest data set (with respect the number of records) was the Abalone data file from the UCI repository. It contains about 4000 records.

Table 3
Representation when the cluster 4-1Lclass is in the center of the figure: first layer and second layer when 4-1Lclass is split





Instead, in relation to the number of variables, the file with more variables was the CIA World Fact file with 1056 variables but only 245 documents. Future work includes working with larger data sets. Recent work on clustering (Alsabti, Ranka, & Singh, 1998; Ganti, Ramakrishnan, & Gehrke, 1999) includes the development of efficient algorithms that would be appropriate in this context.

Additionally, at present we only highlight those documents that are near the centroids. Thus, implicitly, these documents are the only ones considered as *relevant* by our system. Although this can be appropriate in some cases to avoid information overload, in some applications, for example when surfing on the web, it is also important to detect elements with *potentially* relevant characteristics even though these documents are not in a central position with respect to any cluster. At this point there exists, chances (in the sense of Ohsawa & Fukuda (2000)) for improving the system. To highlight such elements, we need to develop methods that analyse the data in more detail and compare isolated data with neighboring one and nearby cluster centroids. This would permit the detection of documents with novel or complementary approaches.

Table 4
Description of clusters 4-1Lclass and 1-2L, 2-2L, 3-2L and 4-2L

Cluster	Countries and regions
4-1Lclass	Austria, Argentina, Australia, Belgium, Brazil, Canada, Chile, China, Colombia, Denmark, Ecuador, Finland, France, Fed. Rep. Germany, Greece, Iceland, India, Indonesia, Ireland, Israel, Italy, Japan, Liechtenstein, Luxembourg, Mexico, Monaco, Morocco, New Zealand, Norway, Pakistan, Paraguay, Peru, Portugal, Saudi Arabia, South Korea, Spain, Sweden, Switzerland, Turkey, United Kingdom, United States, Uruguay, Venezuela
2-2L	Liechtenstein, Monaco
1-2L	Australia, Belgium, Canada, China, France, Fed. Rep. Germany, Greece, Iceland, India, Indonesia, Ireland, Israel, Italy, Japan, Luxembourg, Morocco, New Zealand, Norway, Pakistan, Portugal, Saudi Arabia, South Korea, Spain, Turkey, United Kingdom, United States
3-2L	Austria, Belgium, Denmark, Finland, Sweden, Switzerland
4-2L	Argentina, Brazil, Chile, Colombia, Ecuador, Indonesia, Mexico, Pakistan, Paraguay, Peru, Uruguay, Venezuela

Acknowledgements

Partial support by Generalitat de Catalunya (AGAUR, 2002XT 00111 and 2002BEAI400017) and by Grant-in-Aid for Scientific Research (c), Japan Society for the Promotion of Science no. 13680475 are acknowledged.

References

Agosti, M., Crestani, F., & Pasi, G. (2001). Lectures on information retrieval. *Lecture notes in computer science 1980*.

Alltheweb. (2004). Available: http://www.alltheweb.com.

Alsabti, K., Ranka, S., & Singh, V. (1998). An efficient *K*-means clustering algorithm. In *Proceedings of the 11th international parallel processing symposium (IPPS)*.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Addison-Wesley.

Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for Web retrieval experiments. *Information Processing and Management, 39*, 853–871.

Bezdek, J. C. (1981). *Pattern recognition with fuzzy objective function algorithms*. New York: Plenum Press.

Bosc, P., Liétard, L., & Pivert, O. (2003). Sugeno fuzzy integral as a basis for the interpretation of flexible queries involving monotonic aggregates. *Information Processing and Management, 39*, 287–306.

CIA WorldFact. (2004). Available: http://www.ifs.tuwien.ac.at/~andi/somlib/experiments_wfb90.html.

Crestani, F., & Pasi, G. (Eds.). (2000). *Soft computing in information retrieval* (pp. 102–121). Germany: Physica Verlag and Co, ISBN:3790812994.

Crestani, F., Vegas, J., & de la Fuente, P. (2003). A graphical user interface for the retrieval of hierarchically structured documents. *Information Processing and Management, 40*(2), 269–289.

EuroWordNet. (2004). Available: http://www.illc.uva.nl/EuroWordNet/.

Fellbaum, C. (1998). *WordNet: an electronic lexical database*. The MIT Press.

Ganti, V., Ramakrishnan, R., & Gehrke, J. (1999). Clustering large datasets in arbitrary metric spaces. In *The proceedings of international conference on data engineering*.

Google. (2004). Available: http://www.google.com.

Herrera-Viedma, E., & Peis, E. (2003). Evaluating the informative quality of documents in SGML format from judgements by means of fuzzy linguistic techniques based on computing with words. *Information Processing and Management, 39*, 233–249.

Keller, A., & Klawonn, F. (2000). Fuzzy clustering with weighting of data variables. *International Journal of Uncertain Fuzziness and Knowledge-Based Systems, 8*(6).

Klir, G., & Yuan, B. (1995). *Fuzzy sets and fuzzy logic: theory and applications*. UK: Prentice-Hall.

Known, O.-W., & Lee, J.-H. (2003). Text categorization based on *k*-nearest neighbor approach for Web site classification. *Information Processing and Management, 39*, 25–44.

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management, 40*, 65–79.

Kohonen, T. (1997). *Self-organizing maps* (2nd ed.). Germany: Springer-Verlag.

Kraft, D. H., Bordogna, G., & Pasi, G. (1999). Fuzzy set techniques in information retrieval. In J. C. Bezdek, D. Didier, & H. Prade (Eds.), *The Handbook of Fuzzy Sets Series* (vol. 3). *Fuzzy Sets in Approximate Reasoning and Information Systems*. Norwell, MA: Kluwer Academic Publishers.

Kraft, D. H., Chen, J., & Mikulcic, A. (2000). Combining fuzzy clustering and fuzzy inferencing in information retrieval. In *Proceedings of FUZZ-IEEE2000*, San Antonio, Texas, CD-ROM.

Lanau, S. (2003). *Clasificación y visualización de datos complejos*. Ms. Thesis, Universitat Autònoma de Barcelona, Catalonia, Spain.

Leide, J. E., Large, A., Beheshti, J., Brooks, M., & Cole, C. (2003). Visualization schemes for domain novices exploring a topic space: the navigation classification scheme. *Information Processing and Management, 39*, 923–940.

LSA. (2004). Available: http://lsa.colorado.edu/.

Mandala, R., Tokunaga, T., & Tanaka, H. (2000). Query expansion using heterogeneous thesaury. *Information Processing and Management, 36*, 361–378.

Merkl, D., & Rauber, A. (2000). Document classification with unsupervised neural networks. In F. Crestani & G. Pasi (Eds.), *Soft computing in information retrieval* (pp. 102–121). Germany: Physica Verlag and Co, ISBN:3790812994.

Miyamoto, S. (1990). *Fuzzy sets in information retrieval and clustering analysis*. Kluwer Academic Press.

Miyamoto, S. (2003). Information clustering based on fuzzy multisets. *Information Processing and Management, 39*, 195–213.

Miyamoto, S., & Umayahara, K. (2000). Methods in hard and fuzzy clustering. In Z.-Q. Liu & S. Miyamoto (Eds.), *Soft computing and human-centered machines* (pp. 85–129). Tokyo: Springer.

Murphy, P. M., & Aha, D. W. (1994). *UCI repository machine learning databases*. Irvine, CA: University of California, Department of Information and Computer Science. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html.

Nieto Sánchez, S., Triantaphyllou, E., & Kraft, D. (2002). A feature mining based approach for the classification of text documents into disjoint classes. *Information Processing and Management, 38*, 583–604.

Ohsawa, Y., & Fukuda, H. (2000). Potential motivations as fountains of chances. In *Proceedings of the IEEE international conference on industrial electronics, control and instrumentation (IECON 2000)* (pp. 1626–1631).

Rodden, K. (2002). *Evaluating similarity-based visualizations as interfaces for image browsing*. University of Cambridge, Computer Laboratory, Technical report 543, UCAM-CL-TR-543, ISSN 1476-2986.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Soonthornphisaj, N., & Kijsirikul, B. (2004). Iterative cross-training: an algorithm for learning from unlabeled web pages. *International Journal of Intelligent Systems, 19*, 131–147.

Stan, D., & Sethi, I. K. (2003). eID: a system for exploration of image databases. *Information Processing and Management, 39*, 335–361.

Tombros, A., Villa, R., & Van Rijsbergen, C. J. (2002). The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing and Management, 38*, 559–582.

Torra, V., Lanau, S., & Miyamoto, S. (2003). Fuzzy clustering for indexing in the GAMBAL information retrieval system. In *Proceedings of the EUSFLAT 2003* (pp. 54–58). Zittau, Germany, ISBN 3-9808089-4-7.

Torra, V., & Miyamoto, S. (2002a). Hierarchical Spherical Clustering. *International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 10*(2), 157–172.

Torra, V., & Miyamoto, S. (2002b). On increasing the performance of spherical Sammon's mapping. In *Proceedings of the 7th meeting of the EURO working group on fuzzy sets, workshop on information systems* (EUROFUSE, ISBN 88-87237-02-06) (pp. 17–22). Varenna, Italy, 2002.

Van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

Wordnet. (2004). Available: http://www.cogsci.princeton.edu/~wn/.