(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

Thematic clustering of geographic resource metadata collections

J. Lacasta, J. Nogueras-Iso, P.R. Muro-Medrano, and F.J. Zarazaga-Soria

Computer Science and Systems Engineering Dept., University of Zaragoza, Spain
{jlacasta, jnog, prmuro, javy}@unizar.es

Abstract. Spatial Data Infrastructures at national or higher levels usually comprise the access to multiple geographic data catalogs. However, with classical search systems, it is difficult to have a clear idea of the information contained in the metadata holdings of these catalogs. This paper describes a set of clustering techniques to create a thematic classification of geographic resource collections based on their associated metadata.

1 Introduction

Spatial Data Infrastructures (SDI) provide the framework for the optimization of the creation, maintenance and distribution of geographical information at different organization levels (e.g., regional, national, or global level) and involving both public and private institutions [1]. Geographic data catalogs are one of the main components of an SDI as they provide services for searching geographic information by means of their metadata and according to particular search criteria. But from the perspective of a European or Global SDI, the problem is that usually there is not a unique geographic data catalog. Quite the opposite, hundreds of geographic data catalogs contribute from the different nodes of an SDI. In such situations, where searches are distributed to different metadata catalogs, it is difficult to have a clear idea of the information described in each metadata collection. Therefore, it would be interesting to provide the mechanisms that give a general overview of the contents of each geographical catalog. Furthermore, this is even useful for individual geographic data catalogs in order to facilitate a summary of the contents for those users accessing for the first time. As mentioned in [2, p. 20], the user task in a discovery scenario might be one of browsing instead of retrieval (i.e., browsing instead of performing blind searches).

Focused on the context of geographic metadata, this paper studies the thematic classification and structuring of resources by means of pattern analysis techniques, known as clustering. These techniques are useful to find groups of metadata records (clusters) that share similar values in one (or more) metadata elements according to a mathematical correlation measure (e.g. the Euclidean distance, Bernoulli, Gaussian or polynomial functions among others). It can be said that elements in a cluster share common features according to a similarity

(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

criteria [3]. The clusters obtained as output of these techniques may help in two important issues of information retrieval systems operating on a network of distributed catalogs. On the one hand, the information of the clusters (e.g., names) may serve as metadata for describing the whole collection of metadata records. On the other hand, if the user is interested in browsing (instead of searching), the output clusters provide the means for structure guided browsing.

The novelty of the paper is to improve clustering techniques with the hierarchical relations that may be derived from keywords found in metadata records, whenever these keywords have been selected from well-established lexical ontologies. As mentioned in [4], an ontology is defined as an explicit formal specification of a shared conceptualization; i.e. the objects, concepts, and other entities that are assumed to exist in an area of interest and the relationships that hold among them. And lexical ontologies could be defined as ontologies with weak semantics where the main aim is to establish the terminology used in a domain together with a reduced set of general relations (hypernymy, hyponymy, synonymy, meaning associativity). In the geospatial community it is widely accepted to use thesauri, which can be considered as lexical ontologies, to facilitate the creation of metadata describing geospatial resources. The techniques presented in this paper make profit of this frequent use of thesauri.

The rest of this paper is organized as follows. Section 2 revises the state of the art in classification techniques. Section 3 proposes several methods of clustering taking into account the use of lexical ontologies. Section 4 shows some results obtained from the application of these techniques. Finally, the last section draws some conclusions and further work is discussed.

2 State of the art in classification techniques

This section reviews the main scientific contributions about classification of metadata collections making a special emphasis in those focused in the graphical visualization of the obtained groups.

For instance, [5,6] show a system to access metadata collections through a network of intelligent thumbnails, being those thumbnails either concepts or locations. For instance, in the case of locations, the network of thumbnails corresponds to the hierarchical structure of administrative toponyms.

The use of topic maps [7] is another way to structure the classification of a resource collection. Topic maps are a representation of knowledge, with an emphasis on the find-ability of information. They can be directly used for exploratory search of a resource collection, providing an overview of the collection content. For instance, [8] presents a system that creates a graphic topic map to enhance the access to a medical database using a graphical representation to locate the different topics over a graph of the human body. Within the context of the geospatial community, [9] describes a method to generate a topic map from a metadata collection based on the keywords section of metadata. It makes profit of the Knowledge Organization System (KOS) (e.g., classification schemes, taxonomies, thesaurus or ontology) that has been selected to pick up those terms

(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

in the keywords section. The assumption of using a selected vocabulary or KOS enables the creation of hierarchical topic maps thanks to the semantic relations existent in the selected KOS. Moreover, [?] proposes the use of XTM [10] as the topic map exchange format, which can be easily visualized by a wide range of tools compliant with this format. However, in distributed systems like an SDI, where the different catalogs store thousands of metadata records, the topic map summarizing the thematic contents of a collection may still be fairly complex. In this sense, [9] already identifies the necessity of obtaining a reduced set of representative terms from the generated topic map.

Clustering techniques [11] have proved to produce good results in the classification of big collections of resources for different purposes (data mining, signal analysis, image processing. . .). Two main different types of classification can be highlighted:

Hard clustering: It associates each element of the collection to a unique cluster. An example of this category is *K-means* and its variants [12][13].

Fuzzy clustering: It associates each element to each cluster with a different probability. It includes fuzzy and probabilistic techniques between others. Some examples are fuzzy *C-means* [14] and finite mixture models (as Bernoulli or polynomial) [15].

The MetaCombine project [16] is a good example of building services over heterogeneous metadata. In particular, it focuses on providing a browsing service, i.e. the exploration or retrieval of resources (OAI and Web resources) through a navigable ontology. In addition, it shows the effectiveness of different clustering techniques for heterogeneous collections of metadata records according to different factors, as the time used to locate a resource, the number of clicks needed to reach it or the number of failures in locating the resource.

Other works to be highlighted in this area are the following. [3] provides a visual data mining approach to explore in a easier way the complex structure/syntax of geographical metadata records. [17] also describes a system to cluster a collection of metadata into clusters of similar metadata elements using the cluster algorithm of [18]. [19] describes a system to show graphically traditional representations of statistical information about a collection to facilitate the identification of patterns.

3 Techniques of classification

The different classification techniques shown in the state of the art section group the resources by the similarity of some properties but do not have into account the relations that can exist between the analyzed properties. This section describes how to adapt some of those techniques to make profit of the hierarchical structure of the thesaurus concepts used to fill the keyword section of metadata records. The two following techniques make use of this information to improve the results obtained by traditional hard and fuzzy clustering approaches:

- **Clustering keywords selected from thesauri:** The keywords of the metadata records are transformed into numerical codes that maintain the hierarchical relations between thesaurus terms. These numerical codes are then grouped by means of clustering techniques.
- **Clustering keywords as free text:** Hard and fuzzy clustering techniques are directly applied over the text found in the keywords section. But previous to the clustering, some keyword expansion techniques are used to improve the results.

3.1 Selection of algorithms for hard and fuzzy clustering

As mentioned in [11], there are several techniques of hard clustering. From them, the *K-means* family of algorithms [12,13] has been selected for the test with hard clustering algorithms. It has been selected by its simplicity and by its general use in other areas of knowledge to find patterns in collections of data and by the availability of numerous tools to perform it. The *K-means* algorithm is based on the partitions of N records into K disjoint subsets S_j (being $j \in 1..K$) that minimize the sum of the distances of each record to the center of its cluster. The mathematical function used to measure this distance varies depending on the *K-means* implementation (Correlation functions, Spearman Rank, Kendall's Tau...). The function that has been used here for the experiments is the Euclidean distance, which is one of the most frequently applied.

The implementation of *K-means* used in the experiment section is the proposed in equation 1. The objective of this equation is to minimize the value of J, where x_n is a vector representing the *n-th* data record and v_j is the S_j centroid, being v_j calculated with formula 2, where N_j is the number of elements contained in S_j. The algorithm starts assigning randomly the records to the K clusters. Then two steps are alternated until a stop criterion is met (number of iterations or stability of J). Firstly, the centroid is computed for each record. Secondly, every record is assigned to the cluster with the closest centroid according to the Euclidean distance.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - v_j|^2 \tag{1}$$

$$v_j = \frac{\sum\limits_{n \in S_j} x_n}{\mathrm{N_j}} \ . \tag{2}$$

The main problem of this implementation is the need to select the number of clusters to create, because it is not able to automatically adjust the number of cluster returned. Other more advanced implementations use adaptive techniques. For instance, ISODATA algorithm [20] creates or joins clusters when needed, returning a number of clusters adjusted to the collection distribution.

With respect to the fuzzy clustering family, it includes techniques as fuzzy *C-means* (a fuzzy variant of *K-means*) or finite mixture models (Bernoulli, Gaus-

(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

sian or Polynomial). They assign a metadata record to each cluster with a probability value. Usually, the cluster that has the higher probability is considered as the cluster to which the record belongs, but in doubtful situations more than a cluster can be selected. Fuzzy clustering allows to measure to what extent a metadata record belongs to a cluster, distinguishing between the records that are clearly contained in a subgroup from those that may belong to several clusters. This distinction makes possible to sort the records of each subgroup by their degree of membership and to include a record in more than a cluster with different levels of relevance.

The fuzzy algorithm selected for the experiments has been the *C-means* algorithm proposed in [14]. It minimizes the distance of each record to every cluster centroid, adding the probability of the record to belong to each cluster. It minimizes the function in equation 3, where $A_j(x_n)$ stands for the probability of record n to be in the cluster j. Additionally, this algorithm takes into account the constraints in equation 4. The exponent m is a weighting parameter used to adjust the fuzziness of the clustering algorithm. As it can be seen, the function to minimize is similar to the one used in the *K-means* algorithm. The difference in this case is that the Euclidean distance of each record to a cluster center is multiplied by the probability of being such element in that cluster.

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} (A_j(x_n))^m \left| x_n - v_j \right|^2 . \qquad (3)$$

$$(A_j(x_i)) \in [0,1] \ , \ \sum_{j=1}^{K} A_j(x_i) = 1 \ \mathit{for\ all}\ j \ . \qquad (4)$$

3.2 Thematic clustering using the thesaurus structure

To detect the main themes of the collection, this first approach encodes the keywords of the metadata records into numerical values that preserve the hierarchical relations among the thesaurus concepts. Then, these encodings are clustered. This process let us group metadata records that share similar thematic characteristics without losing the benefits provided by the hierarchical structure of the thesaurus used to fill the metadata records.

The goal here is to generate a numerical identifier for each concept of the thesaurus that describes the hierarchical relations of the concepts, that is, its position in the thesaurus). This identifier is similar to the Dewey Decimal Classification System [21]. It consists of a set of numbers where each number indicates the position of the term in the thesaurus branch to which it belongs (a branch is a tree whose root is a top concept with no broader concepts and contains all the descendants of this concept in the "broader/narrower" hierarchy). For example, as shown in figure 1, the identifier *"02 03 00"* indicates that the term *Geotechnology* is the third child of the second top term (*Earth Science*) of the thesaurus.

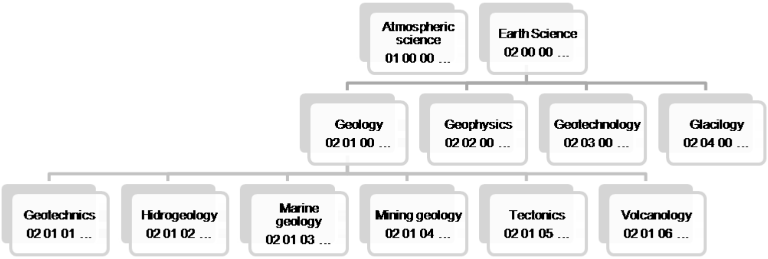

Fig. 1. Numerical encoding of concepts

The result of clustering using these identifiers is the detection of groups of records containing concepts that are close in the thesaurus structure. These clusters are then processed to obtain a reduced set of branches that concentrate most of the keywords used in the metadata records. Since these selected branches group most of the keywords in the metadata collection, they can be considered as its main themes. The approach consists of four sequential steps:

- As an initial step of this approach, the metadata collection has been processed to extract a list of pairs $< keywordInRecord\_URI, keyword\_id >$. In these pairs, $keywordInRecord\_URI$ represents a string that consists of a record identifier and a keyword term found in a record with this identifier. The second element in the pair ($keyword\_id$) is the numerical identifier of the term found in the record. This list of pairs has been used as input for the hard and fuzzy clustering.
- The second step of the approach has been the application of the hard and fuzzy clustering algorithms. The result of hard clustering is a list of clusters indicating the set of $keywordInRecord_URI$s contained in each cluster. Fuzzy clustering returns a matrix where the rows represent the different $keywordInRecord_URI$s and the columns represent the identified clusters. Thus, each row in the matrix contains the membership degree of $keywordInRecord\_URI$ to each cluster. The cluster with the highest probability is selected as the right cluster for each $keywordInRecord\_URI$.
- Obviously, the results obtained in the second step are not the final output, because each record may have been assigned to several clusters, as many clusters as the number of times it appears in a $keywordInRecord\_URI$. Therefore, in order to detect the most proper cluster for each record, a third step has consisted in applying the following heuristic: the records have been associated to the clusters where they appear more times.
- Finally, the fourth step determines the names of the clusters. The numerical codes representing the keywords in each cluster have been processed back to obtain the terms behind them. And the common ancestors of the keywords have been analyzed to select the cluster names. The name given to each cluster is the most specific common ancestor of the keywords in the records grouped in the same cluster. To make the name more representative, there

is an exception in this rule when all the records in a cluster can be grouped under a maximum number of three sub-branches. In that case the name of the cluster is the concatenation of the sub-branches names (e.g., "Product, Materials, Resource" cluster in section 4.2).

3.3 Thematic clustering using keywords as free text

The main problem in the previous approach is that it is very much dependent on taking as input metadata collections that contain terms from selected vocabularies, and that the choice for those vocabularies is relatively small. But, in practice, one can find metadata containing terms from a wide range of thesauri, or even keywords typed randomly by the users. Therefore, it seems relevant to explore other alternatives with not so restrictive prerequisites that facilitate the thematic characterization of collections described with heterogeneous metadata. This section studies the application of clustering techniques considering the keywords section as a free set of terms that do not belong necessarily to a selected vocabulary or thesaurus structure.

The keywords of the metadata records are clustered using the vector space model. In this model, documents are encoded as N-dimensional vectors where N is the number of terms in the dictionary, and each vector component reflects the relevance of the corresponding term with respect to the semantics of each document in the collection [22]. This relevance is directly proportional to the number of occurrences of a keyword in a metadata record. Additionally, not only the terms that exist in the keywords section of the metadata records are included but also all the ancestors (*broader* terms) of those terms extracted from a thesaurus. That is to say, making profit of thesaurus structure, we expand the text found in keywords sections when the terms belong to selected thesauri. Additionally, the use of ancestors help to increase the relevance of some keywords.

Since metadata collections can be very heterogeneous, the output clusters can be seen as a set of different homogeneous sub-collections centered in very different themes. Thus, the naming of clusters may be quite problematic. The names of clusters should represent the theme of each cluster.

The generation of the main theme of each sub-collection (its name) is created differently when hard or fuzzy clustering is applied. When hard clustering is used, the most frequent keywords in the output cluster are the ones selected to form part of the name. With fuzzy clustering the selected name contains the terms with the highest degree of cluster membership.

4 Experiments

The techniques described in the previous section have been validated using hard *K-means* and fuzzy *C-means* clustering techniques. The obtained results have been analyzed to find the situations in which the use of each technique is most adequate.

(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

4.1 Description of the metadata corpus

In order to validate the presented techniques, the content of the Geoscience Data Catalog at the U.S. Geological Survey (USGS) has been used. The USGS is the science agency of the U.S. Department of the Interior that provides information about Earth, natural and living resources, natural hazards, and the environment. Despite being a national agency, it is also sought out by thousands of partners and customers around the world for its natural science expertise and its vast earth and biological data holdings. The USGS metadata collection has been processed as indicated in [23] until a collection of 753 metadata compliant with CSDGM [24] standard were obtained. From them, the 623 keywords that have values from the GEMET thesaurus [25] have been selected for the experiments.

4.2 Thematic clustering using the thesaurus structure

The collection selected for the experiment contains keywords with terms picked up from the GEMET thesaurus. Therefore, this thesaurus has been processed to generate identifiers in the form already described in section 3.2. Because none of the branches of this thesaurus has more than 99 terms as *narrower* of a concept, two digits were enough to encode each level of the thesaurus.

In the case of hard *K-means* algorithm, five clusters (K=5) were asked, with the objective of obtaining the five most used branches of GEMET in the collection. However, the results obtained were disappointing because the algorithm did not converge, producing clusters with heterogeneous keywords.

The results obtained with fuzzy *C-means* algorithm were quite better. The clusters produced contained groups of records with similar keywords. The clusters obtained as output are the following:

Hydrosphere, Land: 19 records about the following topics (including hierarchical path):
- Natural Environment, Anthropic Environment. Hydrosphere
- Natural Environment, Anthropic Environment. Land

Biosphere: 10 records about the following topic (including hierarchical path):
- Natural Environment, Atrophic Environment. Biosphere

Product, Materials, Resource: 211 records about the following topics (including hierarchical path):
- Human activities and products, effects on the environment. Products, Materials. Materials
- Human activities and products, effects on the environment. Products, Materials. Product
- Human activities and products, effects on the environment. Resource

Chemistry, Substances and processes: 39 records about the following topics (including hierarchical path):
- Human activities and products, effects on the environment. Chemistry, Substances and processes

Research Science: 291 records about the following topic (including hierarchical path):

- Social Aspects, Environmental, politics measures. Research Science

These results exhibit how the metadata records of the collection are related, showing a broad thematic view of the collection. The use of upper level branches (the most generic) indicates disperse keywords in the metadata records, the use of lower level branches (more specific) expose a high relation with a specific theme and the lack of a branch of the thesaurus indicates that the metadata in the collection are not related to that theme.

4.3 Thematic clustering using keywords as free text

In this second approach, the use of hard and fuzzy clustering algorithms have produced quite different results. The details of each experiment are detailed next.

Keywords as free text and hard clustering The keywords of the USGS metadata records were transformed into a $C(NxM)$ matrix where N is the number of metadata records to cluster, and M the set of keywords contained in the collection plus the terms added from the GEMET thesaurus. An element $C(i, j)$ of the matrix takes value 1 when the term j is contained in the metadata record i or j is an ancestor in GEMET hierarchy of a term in record i. Otherwise, $C(i, j)$ takes value 0. The *K-means* algorithm were applied to the matrix asking for five clusters ($K = 5$). Then, the clusters obtained were processed to generate a representative name (combination of the most frequent keywords, or the name of the thesaurus branch that contains them), producing the following results:

Natural gas, Geology, Earth Science: This cluster consists of 151 metadata records all of them containing the keywords *Natural gas*, *Geology* and *Earth Science.*

Coal: 116 of its 117 metadata records contain the *Coal* keyword, and the other has *Lignite*, a keyword related hierarchically with *Coal* (its father).

Geology: This third cluster contains 128 metadata records from which 92 contain *Geology* and 34 *Marine geology* (narrower term of *Geology*). The remaining two records contain keywords with terms related to *Geology.* One of them contains *Earth Science* (broader term of *Geology*) and the other one contains *Mineralogy* (sibling term of *Geology*, i.e. having *Earth Science* as broader term).

Chemistry, substances and processes: This cluster contains 53 metadata records with keywords from the *Chemistry, substances, and processes* branch of GEMET.

Parameter & Others: This cluster of 32 metadata records is heterogeneous. It contains 14 metadata records from the *Parameter* branch of GEMET, however, the other 18 metadata records have no relation with them or between them.

Using the result obtained, the generated classification from the whole collection of metadata would be *Geology* (contained in 151+92 records), *Natural gas*, *Earth Science*, *Coal* and *Chemistry, substances and processes*. Each one being less relevant than the previous one (it appears fewer times).

The obtained results also show that inside the set of metadata records about *Geology* an important part of them is more specialized (they are also about *Natural gas* and *Earth Science*)

The expansion through the GEMET hierarchy in conjunction with clustering techniques allows detecting semantic aggregations not directly visible. For example, the cluster *Chemistry, substances, processes* is not detected when no hierarchy relations are considered. In addition, this technique also separates in different clusters records not sharing enough keywords. An example of this occurs in the *Geology* cluster. There are records that contain the terms *Geology* or *Earth Science*, also present in the *Natural gas* cluster, but do not contain the *Natural gas* term. In order to distinguish them from the set of records that always include *Natural gas* two different clusters have been created.

This separation causes that the two clusters share part of name, given that both contain the *Geology* term in most of their metadata. This can cause confusion when accessing the information, given that two subsets are said to be about the same matter. When this happens, the solution adopted is to include all the elements of the more specific cluster inside the most general considering the *Natural gas, Geology, Earth Science* as a subset of the *Geology* cluster.

Another problem found in the obtained results is that the *K-means* algorithm assigns very heterogeneous records to the smallest cluster (in the example, the *Parameter* cluster), because they can not be classified in the rest of clusters. The result is that the smallest cluster is quite useless. Therefore, the name obtained for the *Parameter* cluster is not adequate as it contains many heterogeneous elements. In order to remark this heterogeneity, the generic term *Others* has been added to the cluster name.

Keywords as free text and fuzzy clustering The same $C(NxM)$ matrix generated for the previous example was used with the fuzzy *C-means* algorithm asking for five clusters (*K=5*) with *m=2*. But due to the low degree of cluster membership of the records in two of the clusters, the experiment was redone with $K = 4$. In the results obtained, these two clusters were joined producing a new cluster whose main keyword was *parameter* with a probability of 0.5. There were no changes in the other clusters. The following results were obtained; they include the cluster name and the number of records that have been assigned to each cluster with the highest probability:

Natural gas, Earth Science (151 records) : The occurrences of *Natural gas* and *Earth science* keywords are contained in metadata records near the cluster center, with a mean probability of being in the cluster of 0.99, *Geology* is the following with a probability of 0.64.

Coal (117 records) : The occurrences of *coal* keyword are contained in metadata records near the cluster center, with a mean probability of being in the cluster of 0.8.

Marine geology (128 records) : The occurrences of *marine geology* are contained in metadata records near the cluster center, with a mean probability of being in the cluster of 0.75.

Others (85 records) : The other two clusters are not adequate given that they contain more or less the same keywords but with low probability of belonging to the clusters.

In order to generate the name of each cluster (its main themes), the concept names present in the records with the highest cluster membership were selected. Depending on the requirements of the systems, concepts in records with a lower degree could be also included (e.g. *Geology* in the first cluster). In this situation, the membership degree could be displayed to indicate that not all the categories represent the collection in the same way.

4.4 Comparison of results

The results obtained in each situation are quite different. Figure 2 displays an skeleton of the GEMET hierarchy and how the clusters of each approach are located in this thesaurus hierarchy. Regarding the first approach, it has been shown that hard clustering does not work, and that fuzzy clustering produces very general results. In the second approach, fuzzy clustering produces more specific results than hard clustering. However, all the results are valid. Each one provides a vision from a different perspective of the collection: ranging from the more general vision of the first approach to the more specific vision of the last experiment.

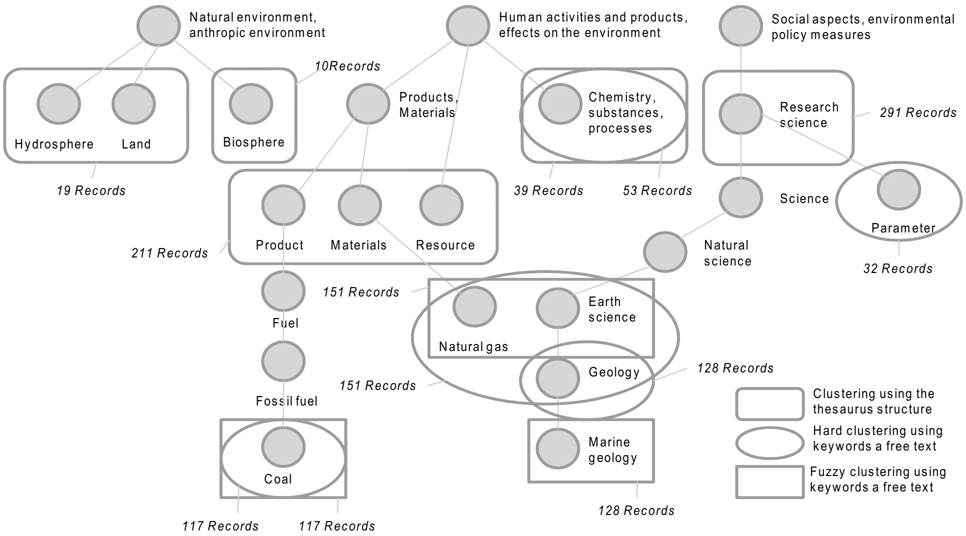

Fig. 2. Clusters generated with the different approaches proposed

The first approach only works correctly if the keyword section has been created with terms of a single thesaurus. But, given that in practice metadata contain terms from a wide range of thesauri, or even keywords randomly typed by the users, the second approach is more flexible. This second approach expands the keywords contained in the metadata (to improve the results) using the thesauri used as source. Nevertheless, it also works properly with several thesauri or even with keywords not contained in a controlled vocabulary.

In general, hard clustering has the advantage of being more simple but fuzzy clustering enables the identification of records with doubtful allocation (they have similar probability for two or more clusters). This is a clear advantage of fuzzy with respect to hard clustering. Fuzzy clustering detects a cluster not only on the basis of the number of times a keyword appears in a metadata record, but also taking into account that a keyword can not be found in other clusters. An example of this is the treatment of the *Geology* concept in the second approach. When hard clustering is applied, *Geology* was as important for the cluster theme as the rest of the main keywords. However, with fuzzy clustering it is shown that the records containing *Geology* are less centered in the cluster than the other ones.

5 Conclusions

This paper has shown some techniques to automatically generate a classification of a collection of metadata records using the elements of their keywords section. These techniques are based on hard (*K-means*) and fuzzy (*C-means*) clustering and have into account the hierarchical structure of the concepts contained in the Knowledge Organization Systems (KOS) used to select these keywords.

Two different approaches have been analyzed. In the first approach, the keywords in the metadata records picked up from a single thesaurus have been encoded as a numerical value that maintains the inner structure of the thesaurus. Then, this encoding has been used as the property for the definition of the thematic clusters. In the second approach, all the keywords in the metadata collection have been considered as free text. Additionally, in order to improve the results, when it has been recognized that a term belongs to a KOS, the terms in the hierarchy of ancestors have been added to the set of keywords. Then, these sets of expanded keywords have been clustered using hard and fuzzy clustering techniques. This last approach aimed at avoiding the deficiencies of the first approach that assumed the use of a single selected vocabulary.

For the experiments, the USGS metadata collection corpus has been used. This collection has been processed to identify terms from the GEMET thesaurus. Then, each approach has been tested with hard and fuzzy clustering processes to compare the obtained results, and to show the advantages and disadvantages of each technique. The results obtained have been quite different, ranging from general to specific classifications. Depending on the specific needs of the final application where the classification is integrated, the use of one or another will be more appropriate.

Further work will be focused on two different aspects: the improvement of clustering techniques, and the integration of these techniques within the services offered by an SDI.

With respect to the improvement of clustering techniques, we expect to adjust automatically the number of clusters. For instance, clustering algorithms requiring no previous knowledge on the number of clusters (e.g., Extended Star [26]) could be executed to estimate the initial number of clusters before applying the approaches presented in this paper. Additionally, we will study the effect of introducing modifications in the way to compute the relevance of an expanded keyword in the second approach. The modifications will take into account the distance from the ancestor to the original term in the "broader/narrower" hierarchy.

Last, concerning the integration of clustering within SDI services, future work will tackle, among other issues, the use of clustering for the improvement of the human computer interaction. For users more interested in browsing than in performing blind searches, thematic clustering will help them in this browsing task. Besides, in a distributed network of geographic catalogs, thematic clustering will help to filter those catalogs that are clearly related to user queries, and discard those not related at all. In fact, the names of the clusters obtained as a result of the techniques proposed in this paper may serve as the metadata that describes the whole collection of records in a repository.

6 Acknowledgements

This work has been partially supported by the University of Zaragoza through the project UZ 2006-TEC-04.

References

1. Nebert, D., ed.: Developing Spatial Data Infrastructures: The SDI Cookbook v.2.0. Global Spatial Data Infrastructure (GSDI), http://www.gsdi.org (2004)
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. New York. ACM Press, Addison Wesley (1999)
3. Demšar, U.: A visualization of a Hierarchical Structure in Geographical metadata. In: Proceedings of the 7th AGILE Conference on Geographic Information Science, Heraklion, Greece (2004) 213–221
4. Gruber, T.: A translation approach to portable ontology specifications. ACM Knowledge Acquisition, Special issue: Current issues in knowledge modeling **5, Issue 2**(KSL 92-71) (1993) 199–220
5. Schlieder, C., Vögele, T., Visser, U.: Qualitative spatial representation for information retrieval by gazetteers. In: Proceedings of Conference of Spatial Information Theory COSIT. Volume 2205., Morrow Bay, CA (2001) 336–351
6. Schlieder, C., Vögele, T.: Indexing and Browsing Digital Maps with Intelligent Thumbnails. In: Proceedings of Spatial Data Handling 2002 (SDH'02), Ottawa, Canada (2002) 69–80

7. International Organization for Standardization (ISO): Information technology – SGML applications – Topic Maps. ISO/IEC 13250, International Organization for Standardization (1999)
8. Kamel Boulos, M.N., Roudsari, A.V., Carson, E.R.: Towards a Semantic Medical Web: HealthCyberMaps Dublin Core Ontology in Protégé-2000. In: Fifth International Protégé Workshop, SCHIN, Newcastle, UK (2001)
9. Lacasta, J., Nogueras-Iso, J., Tolosana, R., Lopez, F., Zarazaga-Soria, F.: Automating the Thematic Characterization of Geographic Resource Collections by Means of Topic Maps. In: Proceedings of 9th AGILE International Conference on Geographic Information Science: Shaping the future of Geographic Information Science in Europe, Visegrád, Hungary (2006) 81–127
10. Pepper, S., Moore, G.: XML Topic Maps (XTM) 1.0. Technical report, http://www.topicmaps.org (2001)
11. Steinbach, M., Karypis, G., Kumar, V.: A comparison of document clustering techniques. In: Proceedings of the KDD Workshop on Text Mining, Boston, USA (2000) http://www.cs.cmu.edu/ dunja/KDDpapers/Steinbach_IR.pdf.
12. Dubes, R.C., Jain, A.K.: Algorithms for Clustering Data. Prentice Hall (1988)
13. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons (1990)
14. Torra, V., Miyamoto, S., Lanau, S.: Exploration of textual document archives using a fuzzy hierarchical clustering algorithm in the GAMBAL system. Information Processing and Management **41** (2005) 587–598
15. Juan, A.: Reconeixement de formes. Technical report, Universidad Politecnica de Valencia (2005)
16. Krowne, A., Halbert, M.: An Evaluation of Clustering and Automatic Classification For Digital Library Browse Ontologies. Metacombine project report, htttp://metacombine.org (2004)
17. Podolak, I., Demšar, U.: Discovering structure in geographical metadata. In: Proceedings of the 12th International Conference on Geoinformatics, Gävle, Sweden (2004) 805–811
18. Fisher, D.H.: Knowledge acquisition via incremental conceptual clustering. Machine Learning **2** (1987) 139–172
19. Albertoni, R., Bertone, A., Demšar, U., Martino, M.D., Hauska, H.: Knowledge extraction by visual data mining of metadata in site planning. In: Knowledge Extraction by Visual Data Mining of Metadata in Site Planning, Espoo, Finland (2003) 119–130
20. Ball, G., Hall, D.: ISODATA, A novel method of data analysis and pattern classification. NTIS AD699616, Standford Research Institute, Standford, California (1965)
21. OCLC: Dewey Decimal Classification System, 22nd edition. Online Computer Library Center (2003)
22. Berry, M., Drmac, Z., Jessup, E.: Matrices, Vector Spaces, and Information Retrieval. SIAM Review **41** (1999) 335–362
23. Nogueras-Iso, J., Zarazaga-Soria, F.J., Muro-Medrano, P.R.: Geographic Information Metadata for Spatial Data Infrastructures - Resources, Interoperability and Information Retrieval. Springer Verlag (2005)
24. Federal Geographic Data Committee (FGDC): Content Standard for Digital Geospatial Metadata, version 2.0. Document FGDC-STD-001-1998, Metadata Ad Hoc Working Group (1998)

(Draft) Lecture Notes in Computer Science (LNCS). 2007, vol. 4857, p. 30-43. ISSN 0302-9743.

25. European Environment Agency: GEneral Multilingual Environmental Thesaurus (GEMET). Version 2.0. European Topic Centre on Catalogue of Data Sources (ETC/CDS), http://www.eionet.europa.eu/gemet (2004)
26. Gil-García, R.J., Badía-Contelles, J.M., Pons-Porrata, A.: Extended Star Clustering Algorithm. Lecture Notes in Computer Science (Progress in Pattern Recognition, Speech and Image Analysis) **2905** (2003) 480–487