Bringing taxonomic structure to large digital libraries

David Sánchez* and Antonio Moreno

Department of Computer Science and Mathematics, University Rovira i Virgili, Avda. Països Catalans, 26, 43007, Tarragona, Spain
E-mail: david.sanchez@urv.cat
E-mail: antonio.moreno@urv.cat
*Corresponding author

Abstract: Digital libraries are invaluable repositories of information. However, in many situations, their size makes it difficult to access the desired resource. In this paper, we present an automatic, unsupervised, domain-independent and scalable approach for structuring the resources available in a certain electronic repository for a particular domain. The system automatically detects and extracts the main topics related to the desired domain, offering a taxonomical structure. This result is complemented by the library’s search engine, offering an integrated tool for accessing resources as an automatically composed directory service. The system has been tested for several digital libraries and domains of knowledge, providing good quality results in all cases.

Keywords: taxonomy learning; digital libraries; web mining; web search engines; resource indexing; knowledge acquisition; ontologies; semantic web.


Biographical notes: David Sánchez is a member of the Intelligent Technologies in Advanced Knowledge Acquisition (ITAKA) Research Group at the University Rovira i Virgili’s Department of Computer Science and Mathematics and a PhD student at the Technical University of Catalonia. His research interests are intelligent agents and ontology learning from the web. He received his MSc in Artificial Intelligence from the Technical University of Catalonia.

Antonio Moreno is a Lecturer at the University Rovira i Virgili’s Computer Science and Mathematics Department. His main research interests are the application of agent technology to healthcare and ontology learning from the web. He received a PhD in Artificial Intelligence from the Technical University of Catalonia.

1 Introduction

In recent years, the growth of the Information Society has been very significant, providing a way for fast data access and information exchange all around the world through the World Wide Web. In consequence, the web has become an indispensable knowledge tool where one is able to find resources for every possible domain. However, due to the lack of control over the content published on the web, in many situations, problems about dispersion and untrustworthiness of the information arise.

In this sense, web-based digital libraries (e.g., Citeseer, PubMed, etc.) provide an environment in which the scientific production for a particular domain is stored, configuring a trusted, updated and immediate repository of information. However, due to the success of these initiatives, the amount of available resources is beginning to be, in some cases, so huge that the difficulty of searching and obtaining the desired information has become a serious problem in a similar way as with the whole web itself but in a lower scale (Kobayashi and Takeda, 2000). That is why the need of tools for information retrieval that ease the way in which those resources are accessed and analysed has been growing in pair with the information itself.

Even though, in comparison to the web, these libraries incorporate a certain degree of structure (e.g., resources can be indexed by authors, title, references, etc.), many of these repositories lack a semantic structure to classify the resources according to the covered topics.

The most common way for accessing the resources is by means of keyword-based search engines that incorporate many of these libraries. They do a good job in retrieving relevant resources for the corresponding query. However, this type of search usually suffers from two problems derived from the nature of textual queries and the lack of structure in the documents:
• the difficulty to set the most appropriate and restrictive query
• the tedious evaluation of the huge amount of resources obtained.

So, in this paper, we present a general purpose tool for easing the accessing to electronic resources in digital repositories by automatically constructing structured representations – taxonomies– for those resources according to the main topics discovered for a particular domain. These results are used as concrete queries for retrieving resources from the library’s search engine, providing an access similar to a directory service but composed in a completely automatic and unsupervised way.

As a summary, the main features of our work are:

• Unsupervised operation. This is especially important due to the amount of available resources in many repositories, avoiding the need of a human expert on the searched domain.

• Completely automatic behaviour, allowing to perform easy executions at any time in order to maintain updated results. This characteristic fits very well with the dynamically changing nature of many domains of knowledge.

• Domain independent solution, because no domain related assumptions are formulated and no predefined knowledge is needed. This is especially interesting when dealing with highly specific domains where non widely-used concepts may appear. The only restriction here is that it can only be applied with resources written in English, due to the dependency of certain basic rules about word morphology and syntactic constructions.

• Efficient and scalable analysis of resources but keeping a good coverage for the specific domain. This is achieved thanks to the intensive use of search engines that provide an immediate way of obtaining resources.

The rest of the paper is organised as follows. Section 2 presents related works in structuring web resources automatically. Section 3 introduces the bases of our process based on linguistic patterns and statistical analyses. Section 4 presents the proposed novel automatic and unsupervised methodology for obtaining topic taxonomies from digital libraries. Section 5 introduces the developed prototype and its application to several particular electronic repositories. Section 6 discusses the evaluation of the results, comparing them to other approaches. The final section contains the conclusions and proposes lines of future work.

2 Related work

Nowadays, there exist several initiatives for structuring electronic repositories and, more generally, the whole web in an unsupervised way. However, most of the approaches are based on clustering techniques: they group similar objects into sets. In the web search context, this implies the organisation of web pages into groups, so that different groups correspond to different user needs. Some systems of this kind are: Scatter/Gather (Cutting et al., 1992), which provides online -fast- and offline -slow- clustering algorithms; Grouper (Zamir and Etzioni, 1999) which operates on query result snippets and clusters together documents with large common subphrases; Carrot2 (Stefanowski and Weiss, 2003); Vivisimo/Clusty (http://www.vivisimo.com; http://clusty.com), for commercial purposes; Mapuccino (Maarek et al., 2000), using similarities based on the vector-space model; and SHOC (Zhang and Dong, 2004) based on key phrase discovery.

The main problem of those approaches is the poor semantics of the structure presented to the user; clusters obtained for a searched domain tend to be unconnected, with very different degrees of granularity. This hampers the comprehension of the domain structure and the browsing of the available resources.

On the other hand, taxonomies are a typical way of structuring knowledge. In Artificial Intelligence systems, in fact, they are the first step for composing ontologies (Fensel, 2001) and a crucial component for many knowledge intensive tasks as the Semantic Web (Berners-Lee et al., 2001). Nowadays, there exist taxonomy generator packages such as Autonomy Verity Thematic Mappings (Chung et al., 2002). These allow the construction of topic hierarchies that can be used for browsing or classification. Taxonomical approaches have been applied for structuring web resources in directory services (like Yahoo), offering a very intuitive way of browsing. However, they are generally constructed offline and/or manually by subject matter experts (Freeman, 2006). This requires a considerable amount of human effort, and it is hardly scalable in such an enormous and dynamic repository as the web. In consequence, directory services have a limited scope, in comparison to a general web search engine, and incomplete categories, especially for specific technological domains. The same situation, albeit on a smaller scale (millions instead of billions of resources), applies to digital repositories.

Due to the described limitations presented by cluster based methods and manually composed taxonomical approaches, as mentioned in the introduction, we intend to design an automatic and domain independent tool for structuring taxonomically the resources of an electronic library, complementing the repository search engine.

Our approach is based on knowledge acquisition from text. The most well-known approaches in this area are: pattern-based extraction (Morin, 1999; Hearst, 1992) where a relation is recognised when a sequence of words in the text matches a pattern; association rules (Agrawal et al., 1993), which have been used (Maedche and Staab, 2001) to discover non-taxonomic relations between concepts, using a concept hierarchy as background knowledge; ontology pruning (Kietz et al., 2000) that is based on refining a general ontology using heterogeneous sources, and concept...
learning (Hahn and Schulz, 2000) where a given taxonomy is incrementally updated as new concepts are acquired from texts. However, most of these approaches have been applied for developing methodologies that deal with preselected domain corpus and/or use some degree of supervision and previous knowledge (Gómez-Pérez et al., 2004). Those aspects hamper their performance when dealing with a huge, noisy and unstructured repository, such as the web.

Recently, some authors have been using the web as a learning corpus for developing (Navigli and Velardi, 2004) or enriching knowledge structures (Agríre et al., 2000), proposing techniques adapted to this particular environment (Sánchez and Moreno, 2006). In particular, statistical analysis have been applied for ranking synonyms sets (Turney, 2001) or checking the relevance of pattern based extracted candidates for taxonomic relationships (Cimiano and Staab, 2004). These statistics, in conjunction with an exhaustive utilisation of web search engines, have been used for acquiring large lists of facts using a set of predefined linguistic patterns (Etzioni et al., 2005).

Our approach, as will be described in the following section, is also based on these techniques. Adapting them adequately to our goals and working environment, we configure a novel methodology for structuring digital libraries by resolving some of the limitations presented by the introduced approaches.

3 Taxonomy learning framework

The base of our proposal is the analysis of the resources available for a specific domain in an electronic repository to detect the main topics. In order to perform this process automatically and unsupervised, two main tasks are performed:

- extraction of suitable candidates that represent different topics for the domain
- evaluation of their relevance in order to select the most representative ones for composing a taxonomy.

In this section we present the main techniques used for tackling each task and justify why they are adequate for our purposes.

3.1 Linguistic patterns for taxonomy learning

Learning taxonomies without previous knowledge is a hard task when dealing with natural language texts like web resources. In this case, a common methodology for obtaining predefined relationships (in our case, taxonomical ones) between concepts is the use of linguistic patterns associated with the language in which documents are presented. In pattern-based approaches, the text is scanned for instances of distinguished lexico-syntactic patterns that indicate a relation of interest. This is especially useful for detecting specialisations of concepts that can represent is-a (taxonomical) relationships or even named entities (instances). Several authors such as Grefenstette (1997) and Hearst (1992) have studied, and have used exhaustively, those patterns for obtaining taxonomical relationships.

A particular case of linguistic pattern for obtaining specialisation relationships is the use of noun phrases (e.g., credit card) and adjective noun phrases (e.g., local tourist information office). Concretely, in the English language, the immediate anterior word for a keyword is frequently classifying it (expressing a semantic specialisation of the meaning) (Grefenstette, 1997). So, the previous word for a specific keyword can be used for obtaining the taxonomical hierarchy of terms (e.g., pressure sensor will be a subclass of sensor). If the process is repeated recursively we can create deeper-level subclasses (e.g., air pressure sensor is a subclass of pressure sensor).

This particular pattern will be applied in our proposal over the resources available in an electronic repository to detect candidate concepts that will compose the taxonomy of topics for the particular domain.

3.2 Statistical analysis for inferring information's relevance

Once a set of candidates for composing the taxonomic structure is retrieved, we need to select the most relevant ones in order to obtain a consistent structure.

The use of statistical measures (typically about co-occurrence of terms) for inferring the relevance of concepts and the degree of relationship between terms is a common technique when processing unstructured text (Lin, 1998). However, statistical techniques suffer from the sparse data problem, i.e., they perform poorly when the words are relatively rare, due to the scarcity of data. Some authors such as Brill et al. (2001) have demonstrated the convenience of using a large amount of texts to improve the quality of classical statistical methods. Concretely, in Turney (2001), methods are proposed to address the sparse data problem by using the biggest available data source: the web.

However, the analysis of such an enormous repository is, in most cases, computationally unviable. Here is where the use of lightweight processing algorithms, that can scale well enough with high amounts of information, in addition to the statistical information provided by web search engines, can suppose a good deal. In fact, on the one hand, some authors such as Paşca (2005) have justified the need of using simple processing analysis when dealing with such a huge and noisy repository like the web; on the other hand, Etzioni et al. (2005), Cimiano and Staab (2004) and Cilibrasi and Vitanyi (2004) have shown the convenience of using web search engines to obtain robust statistics.

Concerning this last point, one of the most important precedents can be found in Turney (2001), where several heuristics for employing the statistics provided by web search engines are presented (aka ‘web scale statistics’ (Etzioni et al., 2005)). The conclusion of that work is that the degree of relationship between a pair of concepts can be measured through the analysis of the results of combinations of queries made to a web search engine (involving those
Bringing taxonomic structure to large digital libraries

As an example, a typical score measure of co-occurrence between an initial word (problem) and a related candidate concept (choice) is equation (1).

\[
Score(choice) = \frac{\text{hits(\text{problem AND choice})}}{\text{hits(choice)}}.
\] (1)

This score is derived from the probability theory. Here, \( p(\text{problem AND choice}) \) is the probability that problem and choice co-occur. If problem and choice are statistically independent, then the probability that they co-occur is given by the product \( p(\text{problem})p(\text{choice}) \). If they are not independent, and they have a tendency to co-occur, then \( p(\text{problem AND choice}) \) will be greater than \( p(\text{problem})p(\text{choice}) \). Therefore the ratio between \( p(\text{problem AND choice}) \) and \( p(\text{problem})p(\text{choice}) \) is a measure of the degree of statistical dependence between problem and choice. We can drop \( p(\text{problem}) \) because it has the same value for all choices, obtaining the final expression by associating probabilities with the number of hits returned by a web search engine.

Applying this idea to our case, we use this statistical approach to infer the relevance of candidate topics against the web, through appropriate queries performed into the search engine. With this approach we are able to obtain relevant measures immediately that involve, in most cases, millions of resources (an order of magnitude much higher than the particular explored library), providing a very robust measure. Those results will be then considered in order to select which candidates will be included in the taxonomy.

4 Taxonomy construction methodology

In this section, we present a detailed description of the proposed methodology for constructing topic taxonomies automatically, and without supervision, from large digital libraries by means of web search engines.

The process starts, as shown in Figure 1, by specifying a domain (e.g., cancer) to be explored in a particular digital library. The system uses the keyword or set of keywords that specify the domain for querying the internal library’s search engine. The set of returned resources is used as the source from which to extract the appropriate subtopics for the domain. In order to perform this process, the text content from those resources is processed to detect instances of the mentioned linguistic patterns (e.g., breast cancer is a subtopic of cancer). However, as the list of resources to be analysed can be potentially large (e.g., near 2 millions documents for the ‘cancer’ domain in the PubMed repository), this approach may not be scalable enough for dealing with those general domains.

For that reason, we introduce two aspects that certainly improve the performance of this process:

- On the one hand, due to the nature of the resources contained in those digital repositories (typically scientific publications), in many cases, it is possible to obtain previews, abstracts or summaries of the particular resources. As we only want to extract the main subtopics for the specified domain, those pieces of text (in conjunction to the title) are typically enough for detecting them. It is commonly possible to specify to the repository’s search engine to show that information for each item in the result page. Considering only that page (containing dozens of search engine results), we can extract valuable knowledge without having to analyse large amounts of redundant information and to perform additional access to the web to download each resource.

- On the other hand, even though considering the mentioned optimisation, there can be thousands of potential result pages that should be accessed and processed adequately. However, for many domains, the main interesting topics are a reduced set that can be mostly detected at the beginning of the analysis. For that reason, in our approach, only a reduced set of resource summaries is analysed. As will be described below, the system automatically decides the number of analysed resources according to the domain’s generality and the potential amount of available subtopics.

4.1 Subtopic extraction

Taking those aspects into consideration, the system starts by analysing the first set (e.g., 20) of resource summaries presented by the library’s search engine for the queried domain. The content is processed in order to extract clear
text and then it is parsed for finding all matchings of the domain’s keyword(s). For each matching, the immediate anterior word is extracted and selected if it is a noun or an adjective (distinguished by means of a syntactic analyser) and not a stop word. Those new candidates for subtopics (e.g., breast cancer) are processed by an English stemming tool in order to detect and avoid repetitions of the same concept expressed by different morphological forms.

4.2 **Subtopic selection**

Once the list of candidates from this first set of resources has been compiled, the next step is to figure out which the correct and the most suitable ones for the specific domain are. In order to present an unsupervised method, we use a statistics based approach for performing the selection of the appropriate candidates. As mentioned in Section 3.2, we use web scale statistics obtained directly from a web search engine to obtain very robust measures (considering the whole web and not only the scope of the digital library) in a very efficient way. As this approach requires composing adequate queries for a web search engine involving the extracted candidates and the initial keyword, the quality of the statistical values about candidate relevance for the domain will depend on the way in which they have been obtained. Due to the nature of the linguistic patterns involved (based on noun and adjective phrases) we have defined the following score as the most suitable one:

\[
\text{Score}(\text{candidate}) = \frac{\text{hits}(\text{"candidate\ keyword"})}{\text{hits}(\text{"candidate"})}.
\]  

(2)

This score involves the formulation of two queries for a web search engine giving us a value that measures the degree of relatedness between the candidate concept and the domain. Those candidates that exceed an empirically set threshold are finally selected. The threshold configures the behaviour of the system in terms of potential precision and recall of the results in relation to the list of extracted candidates. A minimum number of hits for the constructed queries are also required in order to avoid misspelled terms.

4.3 **Incremental learning**

At this point, we have to decide whether the set of selected results is enough for the domain or a deeper analysis should be performed to consider more resources. However, how big should this set of web resources be in order to obtain a set or results with good recall? It will depend on many factors, like the domain’s generality or the web’s coverage. Due to the automatic and domain independent nature of our proposal this parameter can not be set a priori, so we need a mechanism that regulates the process, to provide feedback about how the learning is evolving in order to decide whether to continue evaluating resources or not.

In order to tackle this problem, we propose an incremental analytic methodology in which the amount of web resources analysed in each learning step is increased until most of the knowledge for the domain has been acquired. More concretely, after the selection procedure is performed, the percentage of selected terms from the list of extracted candidates is computed. If it is high, this indicates that the domain is particularly productive and a deeper analysis will potentially return more results. In this case, we query the digital library search engine again to obtain the next set of resources and repeat the full process. The process is iteratively executed until the global percentage of selected terms, computed from the accumulation of results of each iteration, falls below a certain threshold, indicating that most of the knowledge for the domain has already been acquired. The process also finishes if during ten iterations we are not able to retrieve any new candidate. Again, the learning threshold can be configured adequately to tune up the system’s behaviour in relation to the performance of the process runtime and the coverage of the results (the amount of potential topics retrieved for the domain). Using this feedback mechanism we ensure the correct finalisation of the automatic algorithm.

4.4 **Putting all together**

At the end, we obtain a first level taxonomy that includes the main subtopics available for the particular electronic repository for the specific domain. Each subtopic represents a specialisation of the initial term. Querying those terms into the repository’s search engine, we are able to retrieve resources corresponding to that specialisation. In this manner, considering each discovered topic as a new query for the search engine, the user is able to browse the available resources in the same way as a directory service. The important point here is that taxonomies have been obtained automatically and unsupervised with independence to the particular domain. In this manner, we complement the functionality of the keyword based search engine but overcome its main limitations (mentioned in the introduction), which derive from its lack of semantics.

In addition, for each new subtopic of the hierarchy (that, at the same time, represents a new more specific domain of knowledge), the same process can be repeated recursively, obtaining a more detailed multi level taxonomy. In this manner the user can request further details on the particular topics in which he is especially interested.

As a final note, the characteristics that a particular electronic repository should fulfil in order to be able to apply our methodology appropriately are:

- It must have an internal search engine that allows standard query formulations. This is mandatory as it is a crucial part of the proposed methodology.
- It should be possible to present the result in a summarised form, in terms of abstract, previews, etc.
- It must allow external access to perform queries and retrieve result sets via a computer program.
5 The prototype

The proposed methodology has been implemented as a web interface that is placed on top of a particular digital library and provides a way to access its resources in a taxonomical directory service fashion. The system controls the access to the library’s search engine to retrieve resources according to the extracted topics transparently.

The interface (as shown in Figure 2) provides the main functionalities for managing searches, allowing to refine a particular subtopic or to specify different predefined settings for the mentioned selection and learning thresholds, controlling the behaviour of the system. Concretely, ‘Search width’ controls several predefined learning thresholds (from 80% to 50% learning rates), resulting in simple, medium and complex searches (with better domain coverage at the cost of increasing the processing time). On the other hand, ‘Search precision’ controls the selection threshold (between 0.001 and 0.00001), allowing high, medium and low precision (with increasing potential recall). Results are presented as a multilevel hierarchy (on the left) in which each item represents a hyperlink to the results of the search associated with the automatically extracted subtopic into the electronic repository.

Figure 2  Web interface provided by our system for the PubMed electronic library

As a web search engine for obtaining web scale statistics, we use Google through Google API, even though other search engines are also supported (more details later).

We have adapted the system to the following digital libraries (that fulfil the requisites exposed in Section 4):

- The Association for Computing Machinery (ACM) (http://www.acm.org/): ACM provides the computing field’s premier Digital Library and serves its members and the computing profession with leading-edge publications, conferences, and career resources.
- IEEE Computer Society (http://www.computer.org/portal/site.ieeeecs/index.jsp): With nearly 100,000 members, the IEEE Computer Society is the world’s leading organisation of computer professionals. Founded in 1946, it is the largest of the 39 societies of the IEEE.
- NASA Astrophysics Data System (http://adswww.harvard.edu/): is a NASA-funded project which maintains three bibliographic databases containing more than 4.7 million records: Astronomy and Astrophysics, Physics, and ArXiv e-prints.
From the user’s point of view, the process starts by specifying through the web interface a particular digital library from the list of supported ones. Then, a particular query and the search parameters can be specified in the top frame. Once the search is confirmed, the system executes the described taxonomy learning methodology. When the process is finished the resulting one level taxonomy is presented. By clicking over each topic the system automatically retrieves (by querying the library’s search engine) the associated available resources, which are presented in the main frame. At this point the user also has the opportunity of refining a specific subtopic by selecting it and defining a new search (with the desired parameters), in order to obtain a multilevel hierarchy as shown in Figure 2. It is also possible to save and store, in HTML format, the taxonomies obtained through several recursive searches.

This results in a system that is able to return automatically, depending on the specific library and searching parameters, a hierarchy of topics for every possible domain from less than one minute for small general searches useful for casual users to half an hour for enormously detailed searches useful for researchers or web managers. However, most of this time is dedicated to query Google (or other web search engines), which heavily slows down the process. Certainly, the time required to perform a query is several orders of magnitude (a few seconds) higher than for performing the text analysis. This is because web search engines introduce courtesy waits between consecutive queries in order to ensure the performance of the server. Without this introduced overhead, we can expect a speed up that certainly approaches an immediate response, even for the largest searches.

In relation to the computational complexity of the implemented algorithm, the variable that defines the algorithm’s execution is the number of correct or incorrect candidates that are available in a certain repository for a domain. From the initial set – fixed number of resources, depending on the domain, a number of candidates are retrieved. In order to evaluate them, queries to a web search engine are performed to compute statistical measures. As described previously, due to the web search engine behaviour, this is the most important parameter when considering the response time. As a result of the queries, items are selected and rejected items are rejected in function of the specified selection threshold. Then, depending on the learning threshold, the algorithm may decide to evaluate an additional set of resources (resulting in selected and rejected items) or stop the analysis. At the end, a total of terms are returned to the user and a total of rejected items have been performed. In summary, the response time of an individual execution of the algorithm grows linearly on the number of performed web queries, which depends on the number of extracted candidates; at the same time, this last number depends on the domain’s diversity of candidates on the particular repository and the specified learning and selection thresholds.

6 Evaluation

Concerning the evaluation of the results returned by our system, the suitability of the extracted concepts to compose the domain taxonomy is checked manually at this moment. In general, whenever it is possible, a representative human made classification for the domain is taken as the ideal model of taxonomy to achieve (Gold Standard).

As the domain’s coverage of available standard classifications is typically limited to the most general subclasses and the learning process is recursively repeated for each new discovered concept (that itself characterises a new, more concrete, subdomain of knowledge), only the first level of the obtained taxonomy is evaluated against the Gold Standard or manually evaluated by a domain expert. Concretely, the list of subtopic candidates of the initial concept, which are finally selected or rejected, is evaluated. Checking the presence and absence of the extracted concepts in the domain’s standard classification and comparing it with the decision of the selection procedure, we can compute the amount of correctly and incorrectly classified terms and measure the performance of the proposed algorithm. Concretely, we apply standard measures of recall and precision.

Recall (3) shows how much of the existing knowledge is extracted. To calculate recall we count the number of correctly selected concepts and divide it by the overall number of taxonomical terms of a Gold Standard (whenever it is available). Sometimes, especially if there does not exist a standard classification that can tell us the full set of expected terms for the domain, we can compute the Local Recall (4) for the specific execution. In this case, we compute this measure as the ratio between the number of correctly selected concepts against the full set of correct entities included in the candidate list. This score can also give us a measure of how good the selection procedure is in accepting or rejecting candidates. This metric is consistent with the recall metric used in TREC conferences (Voorhees, 2001) and has been used by several authors, as shown in Etzioni et al. (2005), for evaluating automatically obtained knowledge.

$$Recall = \frac{\#correctly\ selected\ entities}{\#domain\ entities}$$

(3)

$$Local\ Recall = \frac{\#correctly\ selected\ entities}{\#extracted\ correct\ entities}.$$  

(4)

Precision equation (5) specifies the extent to which the knowledge is extracted correctly. In this case we compute the ratio between the correctly extracted concepts and the whole number of extracted ones.

$$Precision = \frac{\#correctly\ selected\ entities}{\#total\ extracted\ entities}.$$  

(5)

As an example of evaluation, we present the results obtained in two well differentiated domains over their more suitable repositories: a technological one (Sensor) for the NASA
Bringing taxonomic structure to large digital libraries

As shown in Tables 1 and 2, respectively, we include measures of precision and local recall to evaluate the quality of the proposed selection process in addition to other statistics such as the number of extracted topics or the searching time in function of the size of the search.

Observing the results, we can conclude that the correctness of the candidate selection procedure is high as the number of mistakes (incorrectly selected and rejected concepts from the candidate list) is maintained at around a 10–20%. This is an important conclusion as it indicates that the base of our learning method and, at the same time, one of our contributions (a web statistically based selection procedure) behaves well with independence of the domain.

Concerning the number of correct extracted topics, as expected, it grows in relation to the number of explored resources that, at the same time, require more searching time. Here we can see how the system adapts its behaviour to the domain generality, analysing, more or less, resources according to the search parameters and the feedback provided by learning rates.

Concerning the number of correct extracted topics, as expected, it grows in relation to the number of explored resources that, at the same time, require more searching time. Here we can see how the system adapts its behaviour to the domain generality, analysing, more or less, resources according to the search parameters and the feedback provided by learning rates.

Figure 3  One level taxonomy of sensor subtopics discovered in the NASA library with medium precision and medium search

Table 1  Evaluation results for several search sizes for the Bacteria domain in the PubMed digital library (High search precision)

<table>
<thead>
<tr>
<th>Sensor search size</th>
<th>Precision (%)</th>
<th>Local recall (%)</th>
<th>#Correct topics</th>
<th>#Analysed resources</th>
<th>Search time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>83</td>
<td>100</td>
<td>10</td>
<td>20</td>
<td>12 s</td>
</tr>
<tr>
<td>Medium</td>
<td>87.5</td>
<td>87.5</td>
<td>14</td>
<td>60</td>
<td>45 s</td>
</tr>
<tr>
<td>Complex</td>
<td>91.4</td>
<td>82.3</td>
<td>107</td>
<td>1260</td>
<td>6 min</td>
</tr>
</tbody>
</table>

Comparing our results with existing approaches, we have queried the same domains of knowledge for a widely recognised cluster-based search engine: Vivisimo (considered as one of the most capable available systems (Freeman, 2006)). The results obtained are presented in Figure 5. Note that categories are obtained from the analysis of the full web. When restricting the search scope to the mentioned digital libraries, the results present a very limited amount of categories (between 2 and 8). One can see that

Table 2  Evaluation results for several search sizes for the Sensor domain in the NASA Astrophysics digital library (Medium search precision)

<table>
<thead>
<tr>
<th>Sensor search size</th>
<th>Precision (%)</th>
<th>Local recall (%)</th>
<th>#Correct topics</th>
<th>#Analysed resources</th>
<th>Search time (min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>90</td>
<td>96.6</td>
<td>29</td>
<td>40</td>
<td>1</td>
</tr>
<tr>
<td>Medium</td>
<td>87.7</td>
<td>93</td>
<td>93</td>
<td>240</td>
<td>4.5</td>
</tr>
<tr>
<td>Complex</td>
<td>77.4</td>
<td>88.8</td>
<td>429</td>
<td>3700</td>
<td>33</td>
</tr>
</tbody>
</table>
the classification presented by Vivisimo suffers from a poor structure, resulting in an arbitrary number of categories with a different degree of generality. From the semantic point of view and compared with the results by our approach, categories lack consistency, as different relationship types (specialisations, domains of application, institution names, barely related concepts, etc.) are presented completely merged. This hampers the comprehension of the domain and makes difficult the exploration of available resources.

Figure 5  Clusters returned by Vivisimo search engine when querying bacteria (on the left) and sensor (on the right) domains

Finally, one can wonder about the influence of the particular web search engine used during the learning process (for retrieving web resources and computing web scale statistics) in the final results. In our case, we have used Google because it provides a querying API and it is considered the best search engine available. However, we have also performed tests with other search engines like Altavista, Alltheweb and MSNSearch. Each one provides its own statistical measures about information distribution. We have drawn the following conclusion: although the absolute statistical values for a specific query may be quite different for each web searcher (due to the particular estimation algorithm employed), the final measures obtained during the selection about relatedness between concepts tend to be the same as they are always relative measures. However, an important difference between Google against other search engines is that, due to its high web coverage, for constrained queries (i.e., the deeper levels of the taxonomy), it is able to return a significantly larger amount of resources than other search engines, giving us better statistical measures in those cases.

7 Conclusion and future work

Nowadays, the amount of information available in many digital libraries can certainly overwhelm a user. This is why tools for easing the search and access to the appropriate resources have assumed paramount importance.

The proposed methodology provides a general purpose way for bringing structure to electronic repositories in a taxonomical way. Concretely, we have adapted classical knowledge acquisition techniques (linguistic patterns and statistical analysis) to the particular casuistry of electronic repositories (large amounts of unstructured results and available search engines). This configures a tool that can scale well in wide repositories.

As presented in Section 6, compared to cluster based approaches, it provides better and more comprehensive structures. Comparing it to classical taxonomy learning methodologies (Gómez-Pérez et al., 2004), it does not start from any kind of predefined knowledge and it works in a completely unsupervised way, resulting in an automatic and domain independent solution.

The resulting hierarchies of topics can certainly bring benefits for the users of a particular electronic repository. On the one hand, it allows to normal users to browse and access to the library’s electronic resources in a directory fashion in a very immediate way (performing short searches). On the other hand, it can also represent a valuable tool for web masters or domain experts that can automatically generate indexes for structuring large digital libraries (executing exhaustive searches).

In addition to those basic applications, if a particular repository has a large coverage for a specific domain it can also be possible to generate general taxonomies for that domain with a good recall and, potentially, a good precision. This is possible thanks to the high quality of information sources contained in digital repositories. This hierarchy can be applied to other knowledge tasks such as ontology learning and the Semantic Web (Berners-Lee et al., 2001).
As future lines of research, some topics can be proposed:

- The taxonomical analysis can be improved by adding other linguistic patterns for detecting hyponymy like Hearst’s ones (1992). In this manner, we will potentially detect more taxonomic relationships currently undetectable (e.g., a dolphin is a mammal).
- The presented methodology can be extended for performing different knowledge related tasks over the digital library. More concretely, automatic annotation of the evaluated resources according to the discovered topics can be performed. This may allow a faster a more immediate way of classifying and indexing electronic resources automatically.

Acknowledgement

The work has been supported by Departament d’Innovació, Universitats i Empresa de la Generalitat de Catalunya i del Fons Social Europeu.

References


