An Architecture for Retrieving and Organizing Web Resources for Didactic Purposes

José Ignacio Mayorga, Javier Vélez Reyes and Miguel Rodríguez-Artacho

Juan Manuel Cigarrán
Natural Language Processing and Information Retrieval group, http://nlp.uned.es
Universidad Nacional de Educación a Distancia (UNED)
16, Juan del Rosal
28040 Madrid, Spain
{nmayorga, jvelez, juanci, miguel}@lsi.uned.es

This paper proposes a retrieval architecture based on clustering techniques to search the web for resources that are potentially useful for educational purposes. This architecture makes use of Formal Concept Analysis to search the web as a sort of learning object repository and returns a clustering of results based on pedagogical criteria that can help providing a rich environment for content authoring, enhancing the possibilities of traditional search based on metadata. The retrieval system is designed to be seamlessly embedded in virtual learning environments or authoring tools, and combines information retrieval technologies with pedagogical objectives to help organizing and sorting search results in a meaningful way for educational content authors as well as grounding an approach for semantic-aware learning content retrieval.

Keywords: Learning Objects, e-learning authoring tools, learning design, Formal Concept Analysis (FCA), Clustering, Information Retrieval (IR).

ACM Classifications: K.3.1 Computer uses in education, H.3.3 Information Search and Retrieval

1. Introduction

During the last decades, we have witnessed the evolution of educational material creation and instructional processes. In an early phase, teachers prepared their classes using pre-designed instructional material (digital or not). In a second phase, technology allowed them to use authoring tools and utilities, which caused teachers to acquire the new role of instructional designers. Thus, they started creating and designing their own contents to be interrelated but self-contained pedagogical units. In this manner, these instructional materials would become artefacts – named learning objects (Wiley, 2000) – which were thought to foster instructional content reusability in different learning settings (Polsani, 2003) as well as allowing the interoperability between different systems (Friesen, 2005; Hatala, 2004; Sicilia et al, 2003).

As a consequence of their popularity, learning objects have multiplied their numbers in such a way that storage and retrieval solutions ought to be applied to manage them. Learning object
repositories appeared to face the first problem (Schell and Burns, 2001; Sarasa et al, 2009; Duval et al, 2001; Klebl and Krämer, 2010) whilst metadata standards were devised to cope with the latter one (IMS-LOM, 2001; SCORM, 2004; Duval et al, 2002).

Nevertheless, along the aforementioned second phase, some drawbacks appeared related to searching and retrieval issues. On one hand there is a difficulty to tag learning objects due to the variety of classifications and the high amount of data of the metadata fields. Also, there must be a trade-off between granularity and cataloguing effort. Thus, the smallest the resource, the less instructional context, but at the same time a meaningful tagging of small components is costly and difficult to maintain (Artacho et al, 2008). Some other critics to manual tagging come from the pedagogical perspective based on the inability to model in a flexible way the learning processes and methods and their lack of a pedagogical context (Yahya and Yusoff, 2008; Sicilia and García, 2003). As it has already been said, in this authoring process, tools used by authors and creators of educational material rely heavily in mechanisms of retrieval based on metadata sets by matching search keywords with the appropriate metadata field (Wiley, 2000; Bodas et al, 2007).

In a more natural way nonetheless, teachers have always realized that instructional processes can be enriched by including resources accessible on the Internet, but authoring tools and virtual learning environments lack of such a module. Locating those resources on the Internet requires, however, an interacted and heavy search process articulated by means of common Web search engines such as Google or Bing, so finally the retrieved resources are not presented in a meaningful way for educational purposes, nor is there any navigational model based on instructional relationships. This could motivate a need for software components that improve the search process adding some didactic information such as the pursued instructional goal, the difficulty level or the current moment within the overall instructional process. A way for undertaking this enrichment would start by separating the space of retrieved resources according to their potential uses. This arrangement would permit to associate each resource in the space with a set of didactic goals according to a given instructional framework. Hence, retrieval actions can be didactically oriented to get resources suited to a specific instructional requirement. One way to achieve these capabilities is making use of external frameworks such as ontologies or semantic web technologies (Kinshuk, 2004) or provide inference mechanisms to overcome manual tagging work using rules and inference engines (Henze, 2004).

In this paper, however, we explore a different approach to classify unstructured web content for e-learning using Formal Concept Analysis (FCA) (Cole and Eklund, 1996; Ganter and Wille, 1999) instead of relying on semantic web techniques. Our architectural proposal combines FCA and a didactic organizer to support the structuring of retrieved content potentially useful for instructional purposes by means of concept lattices. These lattices can be used to extract, from a given dataset, a conceptual hierarchy, in some ways similar to an ontology, to be exploited for automatic classification or tagging mechanisms for instructional content. There are not many works on this research line making a heavy use of FCA, and those that make a significant use of this theory are more centred on personalisation rather than in authoring (Boon, 2006). Our intention is twofold, (a) to explore a different approach for gathering instructional information, making no use of harvesting techniques or metadata fields for authoring of educational material and (b) to build an interoperable architecture capable of serving as an embedded subsystem within any e-learning content authoring tool or virtual environment. Hence, we describe a software architecture that is able to transform a previously random set of resources within a domain into a structured space that could be navigated and explored in an easy and meaningful
way from an instructional perspective. Nevertheless, it is important to highlight that this framework is focused on helping instructional designers to create whole instructional processes on a given subject matter rather than to satisfy specific retrieval necessities within a narrower knowledge area. This is due to the underlying used techniques application scope and limitations. The rest of the paper is structured as follows. In Section 2, we describe the architecture of the conceptual framework used to organize the resource space in terms of didactic goals and uses. Section 3 explains the uses of formal concept analysis exemplifying in a realistic learning domain. Section 4 proposes a number of possible exploitation scenarios, and some conclusions and future works.

2. The Architecture of the Didactic Organizer

According to the ideas stated on the introduction, our purpose is to organize unstructured resources – web documents – within an easy and straightforward navigable structure aligned to pedagogical needs. For undertaking this task, firstly, we must characterize instructional uses – i.e. goals – by means of a set of terms accurately selected by instructional designers. Secondly, we must perform a number of search engine queries related to a domain keyword and enriched with the previous terms. As a working hypothesis, we assume that the resources retrieved by each query will be bound to a goal if there is a sufficient overlapping (see subsection 2.2) between the enriching terms on the queries and those characterising the goal. This global retrieval process can be managed by two complementary subsystems that work together to create the pursued resource space separation. Next, a description of both responsibilities and their collaboration with each other are presented:

• The didactic knowledge manager (DKM). This subsystem is in charge of supplying the collection of terms that is expected, as the construction hypothesis, to refer to the set of resources best suited for a given didactic goal. The set of so returned terms will be aligned with a certain goal in order to enrich a retrieval process. This subsystem is discussed in depth in the subsection 2.1.

• The domain organizer (DO). This subsystem is responsible for creating a navigable and explor- able representation of a certain cognitive domain based on the results found by the collection of queries ran on the space of Internet resources. These queries are previously enriched with the terms provided by the didactic knowledge manager as it has been previously stated. The result, as a final outcome, is an artefact containing information about certain interesting resources properly clustered according to didactic knowledge and expectedly aligned with each didactic goal prescribed by the didactic knowledge manager. This subsystem is discussed in depth in the subsection 2.2.

Figure 1 shows the collaboration taking place between these two subsystems when a new request is launched. The process starts by accepting a high level query – i.e., a typical external search engine keyword-based query. This query refers, essentially, to a cognitive domain. As an example, let us suppose that the user wants to perform a query with the domain keyword “algebra” because it is intended to get a collection of interesting resources aligned with their potential didactic uses in this cognitive area. This high level query is taken by the domain organizer subsystem that firstly requests the target goals from the didactic knowledge subsystem. Then, for each goal, the didactic knowledge subsystem is asked to return a set of terms to enrich a new query along with the aforementioned didactic terms. Once a set of queries has been obtained, the domain organizer starts an exploration process by executing them against external
search engines, such as Google or Bing. As a result, a collection of didactically enriched resources is obtained, which is used to create the final outcome representing the overall domain that will be delivered to the user as a clustered result.

2.1 The Didactic Knowledge Manager

As it was stated in the previous section, the didactic knowledge manager is a system that provides the didactic knowledge necessary for the aforementioned clustering purposes. Specifically, the system returns a collection of terms used to assist to cluster the resource space around a set of instructional uses.

As we aim to build a software solution able to cover every possible instructional framework, it is necessary to establish previously the general requirements those frameworks should share. First, the framework must be centred on didactic goals. That is, it is necessary to provide a clearly established vocabulary of goals. Second, these goals shall be characterized by a number of terms closely related to and indicative of resources thought to be useful for didactic purposes. Finally,
goals must be bound to each other by different relationships depending on the distinct frameworks.

Instructional frameworks considering goals can be organized along two dimensions. On the one hand, the structural axis determines the kind of relationships among goals. Thus, in a matrix-structured framework, goals are aligned to several dimensions, each one attending to a specific classification criterion while in a relational framework; these goals are bound together along relationships, whose semantics keep didactic purposes. On the other hand, the evolutive axis deals with the nature, either static or dynamic, of the goal characterisation. Thus, a framework is said to be static whenever the allocation of resource terms within the goals does not change over time whilst in a dynamic framework, these terms can vary depending on environmental conditions. Table 1 classifies some of the most popular frameworks according to these organizational criteria.

To support the heterogeneous spectrum of features within instructional frameworks, we propose a general model that revolves around the concept of goal. As can be seen in the UML class diagram depicted in Figure 2, this entity is responsible for maintaining the bag of terms that are to be used to recall the proper resources from the Internet according to the instructional purpose that this goal bears. Terms internally contain the string that must be used to enrich the query during the retrieval process. Furthermore, each goal includes a set of relations that binds it with other ones according to certain relational semantic. To do this, each relation is characterized by a set of the related goals belonging to the relation. We omit for this sketch the specific mechanism by which terms associated to goals can change over time in order to capture the system evolution according to the instructional state.

We realize that the foundational principles of any framework are arguable. That is the reason why we do not defend any particular framework over the rest. For this paper and just as a concept proof, we have used a static matrix framework inspired in Bloom (1965), Anderson et al (2001) and Churches (2009) works that characterizes the goals in terms of two orthogonal dimensions:

- **The cognitive processes.** This dimension comprises all the identified types of knowledge acquisition processes that can be undertaken by a student in a certain moment within a learning experience. Each cognitive process captures a variety of capabilities and skills that can be organized according to an increasing level of abstraction in such a manner that to a certain degree each one can be said to rely on the previous one. Specifically, there are six cognitive processes:

  1. **Remember.** At this level, students are trained to be able to recall from memory certain knowledge that is considered necessary for the acquisition processes at higher levels.
An Architecture for Retrieving and Organizing Web Resources for Didactic Purposes

2. Understand. The understanding level aims to provide students with means and tools that can be useful in the cognitive task of capture the essentials of certain domain area from basic concepts to semantic relationships among them.

3. Apply. The apply cognitive processes teach students how to use the knowledge at the two previous levels in both real and hypothetical scenarios of experimentation.

4. Analyse. In the analysis processes, students learn how to differentiate and attribute different concepts, ideas and models according to their characteristic properties or comparatively with respect to another one.

5. Evaluate. In the evaluation cognitive processes students make use of the previously acquired skills to get a clear comprehension of the differential characteristics of various concepts, ideas or conceptual models and learn how to judge or evaluate them according to certain goodness metrics.

6. Create. In the creation processes, students are formed to get skills related with the production of solutions to fulfil certain requirements obeying certain design principles.

- The knowledge types. The knowledge types dimension characterizes the different varieties of knowledge to be acquired, in terms of its intrinsic nature. Each type of knowledge along this axis covers the overall spectrum of possible epistemological knowledge that can be found to be bound to domain areas (Anderson et al., 2001). Next, we go in depth in the description of these different types of knowledge:

1. Factual. The factual knowledge refers to basic cognitive elements that are considered necessary to starting the cognitive processes described in the aforementioned dimension. Facts, dates, places, events, and basic vocabularies are typical factual knowledge that can be found within each domain area.

2. Conceptual. The conceptual knowledge allows establishing the semantic relationships that are naturally found among basic concepts within a domain area in order to constitute a closed interrelated cognitive corpus. Taxonomies, classifications, models, theories, semantic relations and concept maps belong to this type of knowledge.

3. Procedural. The procedural knowledge refers to those cognitive elements that describe a manner in which tasks can be carried out in order to achieve certain outcomes. Algorithms, sequences of steps or recipes among others are considered procedural knowledge.

4. Strategic. The strategic knowledge, which is a part of the so-called meta-cognitive knowledge (Anderson et al., 2001), helps deciding what kind of activities must be undertaken in order to obtain certain objectives. Plans, heuristics, tenets or principles are typical examples of strategic knowledge.

The previously stated dimensions can be seen as two complementary axes that when crosscut make emerge a matrix of goals, which can be understood as “explicit formulations of the ways in which students are expected to be changed by the educative process” (Bloom, 1956). In this manner, each cell within this matrix represents a learning goal. For instance, one such goal would be REMEMBER x FACTUAL, meaning that the student would be expected, after achieving this goal, to be able to REMEMBER a series of FACTs from the specific domain to be studied.

According to the searching requirements imposed by our architecture, we characterize each of these goals by a collection of terms. Table 2 summarizes the terms that we have used to define each goal. It is important to notice that these values respond to our specific chosen settings.
indicating how to select resources that can help students to acquire a given goal. In fact, the matrix of goals it is expected to be filled by an instructional designer with expertise both in didactic knowledge and the domain area. In any case, the concrete set of terms for each goal can be selected at design time to tune up the framework to suit given pedagogical settings.

We must highlight that this framework does not aim either to provide a roadmap of how students do acquire goals along time or to track the actual learning cognitive state the student is in at any given moment in time. This framework must be understood as a conceptual artefact devised to arrange a set of resources retrieved from the Internet according to a set of potential didactic uses referred as goals.

A final aspect that must be determined deals with the way in which goals can be related within our instructional framework. Specifically, as the concept proof explained in this paper is based on a matrix arrangement of goals, it can be said that each goal will belong to two relations, one for each value in each dimension. For instance, the goal at row 2 and column 3 will appear within both the UNDERSTAND and the PROCEDURAL relations. The former it is made up of all goals in the second row whilst the latter shares relation with all those in column 3. As it was depicted in Figure 2, this knowledge will be mapped onto the architecture by means of the relation entity. Section 3 will explain how Formal Concept Analysis will take profit of this knowledge in order to aggregate goals into either more general or specific concepts for distributing and organizing resources by means of subsumption associations.

2.2 The Domain Organization Subsystem

The main goal of the domain organization subsystem is to query commercial Internet search engines to retrieve relevant resources according to a specific topic and didactic goals. Retrieved resources will be enriched according to their specific dimensions (i.e. cognitive process and knowledge type) and their representative terms or descriptors and, finally, they will be organized and clustered into a navigable concept lattice by means of Formal Concept Analysis Theory (FCA). In this subsection we will describe each of these sub-processes and we will also give a brief introduction to FCA in the context of this e-learning software application.
2.2.1 Lookup of Didactic Resources on the Internet

In order to retrieve a set of didactic resources from the Internet, the first goal is to produce accurate queries on the domain to teach. We are interested in maximising the retrieval precision, that is, a small quantity of didactic resources but well aligned with the didactic goals, instead of the retrieval recall, more focused on getting a lot of topic related resources but maybe not useful for a learning task. Given the computational cost of building a lattice, reducing the number of resources will help making it affordable. A similar approach has already been used in such systems as CREDO (Carpineto and Romano, 2004).

The query building approach is a critical step on our system to achieve the objectives of a high precision and a limited number of resources. This process will be supported by the Didactic Knowledge Manager. Supposing that the user is able to express her initial needs about a topic accurately and without ambiguities, the queries are built combining the searched domain keyword provided by the user with the set of terms associated to each specific didactic goal extracted from the DKM. More specifically, the system produces a query for each didactic goal by conjoining the searched domain keyword with the disjunction of the set of terms associated with the didactic goal. Examples of such queries would be "Algebra AND (mindmap, video, example) to obtain resources associated to the didactic goal UNDERSTAND x CONCEPTUAL or "Algebra AND (Classification, scheme, theory) if the didactic goal were APPLY x CONCEPTUAL. The set of queries produced will be used to retrieve the first n ranked resources using an external search engine. These resources will be enriched and clustered on the next steps of the process.

2.2.2 Enrichment of Retrieved Resources

Once the system has retrieved all the resources for each didactic goal, it is necessary to enrich those resources in order to describe each one in terms of their terms, their associated cognitive processes and their associated knowledge types (i.e. which conform the resource as suitable for some didactic goals). To do this we do not use semantics but we apply a shallow approach based on the terms occurrence into the whole resource. Basically, the system downloads and analyse each of the retrieved resources to find matches between their textual information and the set of terms associated to each didactic goal. As we introduced in Section 2, a resource is assigned a particular goal if there is a sufficient overlapping between the resource contents and the set of terms associated to a didactic goal. As we introduced in Section 2, a resource is assigned a particular goal if there is a sufficient overlapping between the resource contents and the set of terms associated to a goal in Table 2. Currently, if a single match is found, then the system will enrich the resource with the terms and also with its associated cognitive process and knowledge type (i.e. column and row headers of Table 2). This approach not only allows to characterize resources at a very high level with information related to the dimensions described above (e.g. CONCEPTUAL and UNDERSTAND) but also with specific information about the contents of the resource (e.g. mind map or video or example). Notice that the enriched queries are never persisted as the proposed technique starts the process from scratch by searching the resources available on the web at the time of executing the query. Hence, we delegate on the external search engine the actual retrieval of the most promising resources according to the query.

2.2.3 Organization and Clustering Process

As a result of the enrichment process it is possible to find resources suitable for more than one didactic goal. In our particular case, it is interesting to find a way to organize all those resources in order to ease the user the browsing and selection tasks. The third sub-process described here is focused on this task. To do this we apply the FCA Theory (Wille, 1982; Wille, 1992; Ganter, 1999).
It is a mathematical theory of concept formation derived from lattice and ordered set theories that provide a theoretical model to build complex data structures. From our e-learning and, more specifically, didactic goals and content discovering point of view, FCA can be seen as a powerful tool to automatically structure and classify all the resources retrieved and enriched from the Internet. This theory fits on a lattice-based clustering approach improving information access and exploratory tasks on pure Information Retrieval (IR) scenarios (Cigarran, 2004; Cigarran, 2005; Carpineto, 2004).

The main actor of FCA is the formal concept. A formal concept is defined by its extent and its intent (or comprehension). The extent of a concept covers all the formal objects (resources in our case) that belong to the concept, while the intent comprises all the formal attributes (in this case, terms, knowledge types and cognitive processes) shared by all the objects under consideration. Formal concepts can automatically be derived using a mathematical entity called formal context \( K \) that can be understood in the following way. Given a set of objects \( G \), and a chosen set of features \( M \), there can be defined an attribution relationship which connects elements from both sets. If a formal object \( g \) is linked to a formal attribute \( m \), that relation can be read as \( g \) has the attribute \( m \) or \( m \) is a feature of \( g \).

As an example to understand the theory basics, let’s suppose our system has retrieved five resources from the Internet and it has enriched them as it can be seen in Table 3. Then, the formal context \( K \) would be represented as it is shown in Table 4.

**Table 3: A set of enriched resources**

<table>
<thead>
<tr>
<th>RESOURCE1</th>
<th>RESOURCE2</th>
<th>RESOURCE3</th>
<th>RESOURCE4</th>
<th>RESOURCE5</th>
</tr>
</thead>
<tbody>
<tr>
<td>{FACTUAL, REMEMBER, list, CONCEPTUAL, ANALYZE, survey}</td>
<td>{FACTUAL, ANALYZE, timeline, outline, CONCEPTUAL, REMEMBER, outline}</td>
<td>{CONCEPTUAL, UNDERSTAND, mindmap, video, CONCEPTUAL, REMEMBER, mindmap}</td>
<td>{FACTUAL, ANALYZE, timeline}</td>
<td>{PROCEDURAL, CREATE, algorithm}</td>
</tr>
</tbody>
</table>

**Table 4: Formal context representation example**

<table>
<thead>
<tr>
<th>FORMAL OBJECTS</th>
<th>KNOWLEDGE TYPES</th>
<th>COGNITIVE PROCESSES</th>
<th>DESCRIPTORS</th>
</tr>
</thead>
<tbody>
<tr>
<td>RESOURCE1</td>
<td>FACTUAL</td>
<td>CONCEPTUAL</td>
<td>CREATE</td>
</tr>
<tr>
<td>RESOURCE2</td>
<td>EVENT</td>
<td>REMEMBER</td>
<td>SURVEY</td>
</tr>
<tr>
<td>RESOURCE3</td>
<td>PROCEDURAL</td>
<td>UNDERSTAND</td>
<td>OUTLINE</td>
</tr>
<tr>
<td>RESOURCE4</td>
<td>CONCEPTUAL</td>
<td>ANALYZE</td>
<td>MINDMAP</td>
</tr>
<tr>
<td>RESOURCE5</td>
<td>PROCEDURAL</td>
<td>CREATE</td>
<td>ALGORITHM</td>
</tr>
</tbody>
</table>
An Architecture for Retrieving and Organizing Web Resources for Didactic Purposes

Being $K$ a formal context and $A$ a subset of objects, we define $A'$ (i.e. read as prime operator) as the set of all the formal attributes of $M$ applied to all objects of $A$ (i.e. dually, we can define $B'$ using a subset $B$ of the formal attributes). For instance, if $A=\{\text{RESOURCE2}, \text{RESOURCE4}\}$, then $A' = \{\text{FACTUAL}, \text{ANALYZE}, \text{timeline}\}$ or if $B=\{\text{FACTUAL}, \text{timeline}\}$, then $B' = \{\text{RESOURCE2}, \text{RESOURCE4}\}$.

A formal concept is defined by a pair extension-intension, although not all pairs obtained from a formal context are allowed. Being $A$ a subset of formal objects and $B$ a subset of formal attributes, the pair $(A, B)$ defines a formal concept $c(A, B)$ if and only if:

$$c(A, B) \iff (A' = B) \land (B' = A)$$

If the pair $(A, B)$ is a formal concept, then the sets $A$ and $B$ define the extent and the intent of the formal concept and they hold $A'' = A$ and $B'' = B$. For instance, $\{(\text{RESOURCE2}, \text{RESOURCE4}), (\text{FACTUAL}, \text{ANALYZE}, \text{timeline})\}$ is a formal concept in our particular example.

Especially useful types of formal concepts are the object and attribute concepts. An object concept $(g', g)$, where $g$ belongs to the set of formal objects, is the most specific concept that includes $g$ in its extension. On the other hand, an attribute concept $(m', m)$, where $m$ belongs to the set of formal attributes, is the most generic concept that contains the formal attribute $m$ on its intension. For instance, the formal concept $\{(\text{RESOURCE2})(\text{FACTUAL}, \text{CONCEPTUAL}, \text{REMEMBER}, \text{ANALYZE}, \text{outline}, \text{timeline})\}$ is an object concept for object RESOURCE2 as it cannot be more specialized to another formal concept containing RESOURCE2 on its extension. An example of an attribute concept for the formal attribute timeline is $\{(\text{RESOURCE2}, \text{RESOURCE4}), (\text{FACTUAL}, \text{ANALYZE}, \text{timeline})\}$ as it is not possible to find a more general formal concept with timeline on its intension.

It is possible to define a binary relationship on the set of all the formal concepts for a given context $K$. This binary relationship allows implementing the definition of what a superconcept or a subconcept is. In our example $\{(\text{RESOURCE1}, \text{RESOURCE2}), (\text{FACTUAL, ANALYZE, timeline})\}$ is a superconcept of $\{(\text{RESOURCE2}, \text{FACTUAL, ANALYZE, timeline, outline})\}$, which means that the first formal concept is more generic than the second one. Finally, the Fundamental Theorem of FCA (Ganter, 1999) defines the correspondence between ordered sets theory and the general lattice theory. It asserts that the set of formal concepts for a given context is a complete lattice where meets and joins (i.e. infimum and supremum) are always guaranteed for any set of formal concepts. The meet of a set of concepts can be understood as a specialization. In other words, it is the larger concept that is a subconcept of the original concept set. On the other hand, the join of a set of concepts can be understood as a generalization, which means that it is the smaller concept that is a superconcept of the original concept set. We will use meets and joins as powerful operators to navigate and explore the lattice finding upper and lower concepts sharing features with an initial set of concepts.

Figure 3 shows the concept lattice of our example depicted using a Hasse diagram (Wille, 1982; Wille, 1992). A Hasse diagram is a compact way to represent concept lattices where the extent and intent sets for a specific concept can be easily read. Each node of the graph represents a formal concept where the existence of labels indicates if it is an object concept, an attribute concept or both. In addition, if a formal concept is an attribute concept then the node will be also depicted colouring its upper half in blue colour. On the other hand, an object concept will be represented colouring the lower half of the node in black colour. A node coloured with both blue and black colours represents a concept that is an attribute and an object concept at the same time. Small
nodes without colouring represent formal concepts that are not attribute concepts nor object concepts. A formal concept can be reconstructed from the diagram in the following way. The extension is created reading all the object labels (i.e., resources) from the target node to the bottom of the lattice, whereas the intension is generated reading all the attribute labels (i.e., terms, knowledge types and cognitive processes) from the target node to the top of the lattice. For instance, the selected concept (i.e. rounded with a dotted circle) on Figure 3 could be defined as \(((\text{RESOURCE1, RESOURCE2}), \{\text{FACTUAL, ANALYZE, CONCEPTUAL, REMEMBER}\})\).

Hence the Domain Organization Subsystem will produce the formal context using the output provided by the enrichment sub-process. Then the formal context will be used to generate the concept lattice corresponding to the set of resources retrieved (Figure 4). Our system implements the Next Neighbours algorithm that has a $O(|C| |G| |M|^2)$ complexity (Carpineto, 2004) and
returns not only the formal concepts set \( C \), but also the relations between formal concepts. The lattice, seen from the software artefact perspective, also implements a set of useful capabilities for manipulating and exploring the result. Moreover, any third party developer could use the results provided by the proposed framework to implement graphical views of the lattice and navigational or exploring paradigms to easily integrate the lattice into any e-learning instructional authoring tool, enabling the interoperability with Learning Management Systems.

3. Asking the Architecture for Instructional Content

In this section, we are going to present a realistic example of the proposed framework to illustrate its functionalities and how it can help an expert user to create contents meant for an instructional process. First, we will present a basic use story and then, we will show how the system is going to produce a concept lattice that organizes all the retrieved resources allowing the user to select those didactic goals necessary to build a specific learning experience.

3.1 A Typical User Story

“Amy is a teacher who wants to create a learning experience dealing with Algebra but she has not enough contents to build it from scratch. She also wants to illustrate that experience with resources according to didactic goals from the Revised Bloom’s taxonomy (Churches, 2009; Anderson et al., 2001). She has tried to look for these contents on the Internet but she has found that this task is time consuming; she has to manually filter a lot of results and, in most cases, the retrieved resources do not fit into any of the desired didactic goals. Hence, she decides to test the proposed framework in order to have a set of resources automatically retrieved and organized from the Internet. She will inspect the outcome by browsing the resulting lattice, which is labelled using the well-known Revised Bloom’s taxonomy dimensions enriched with a set of terms associated to the selected goals. That will suggest to her the potential usage of the retrieved resources”.

3.2 The Framework in Action

The first step is to run a query focused on a domain keyword to get didactic resources. In this example, it will be Algebra. Then, the system begins its work carrying out the following tasks on the Domain Organizer Subsystem:

1. The original query is enriched using the Didactic Knowledge Manager that supplies the subsystem with the set of terms for each didactic goal. As a result, a set of queries is built upon each predefined didactic goal (see subsection 2.2.1).

2. The retrieval process is performed. The framework queries an external search engine using the set of queries built. As a result, a set of resources suggested by the user-selected goals (terms associated to these goals on Table 2) is available to be enriched and organized. In this real example, the number of Internet resources retrieved was 42.

3. The set of retrieved resources is enriched according to the set of terms on the dimensions described in Section 2 and the methodology described in subsection 2.2.2. This will be the input to create the formal context.

4. The set of enriched resources is used to populate the formal context by mapping the enrichments into attribution relationships between the formal objects – i.e. resources – and the formal attributes – i.e. terms, and the goals dimension labels, that is, knowledge types and cognitive processes. The obtained formal context had 42 formal objects and 55 formal attributes.
5. The formal concepts and the concept lattice are generated. In our example 1674 formal concepts and their relationships were created.

3.3. Exploring the Results

Although our approach is applied to a limited number of resources as we have explained in subsection 2.2.1, it is noticeable that the number of relationships to be discovered between formal concepts could be potentially high. Hence, it could prove complex for the user to explore the resulting concept lattice directly. Our approach to tackle this inherent difficulty consists in filtering out those relations and sub-lattices falling out of the user’s intended focus as expressed in their query. As we are describing a framework and not an end-user application, we are going to focus on specific didactic interests instead of dealing with the user’s actions on a graphical interface. Let’s suppose that the user is interested in the following didactic goals FACTUAL x UNDERSTAND (i.e., the goal having FACTUAL as knowledge type and UNDERSTAND as cognitive process) and CONCEPTUAL x REMEMBER and let’s explore the lattice. In other words, we are going to navigate and browse the concept lattice that has been already calculated by the Knowledge Manager Subsystem using a subset of its formal attributes (i.e., terms, knowledge types and cognitive processes). This query method will produce views by filtering the organized resources, which are projections of the original lattice obtained in step 5 of the previous section. These are focused on specific features and they will add no computational cost as the concept lattice will not need to be re-constructed but redrawn.

Figure 5 shows the filtered lattices for each didactic goal. The bottom of each lattice represents a formal concept whose extent contains resources that are suitable for both dimensions of the goal (i.e. knowledge type and cognitive process). In this figure we have also represented the number of owned object for each formal concept (i.e. the number of formal objects for which it is an object concept) and its percentage over the total number of retrieved resources. It can be seen that the bottom concepts own a high percentage of the retrieved resources, which clearly indicates that those resources are suitable for more than one didactic goal at the same time. On the other hand, only a few resources are owned by upper concepts. This situation indicates that these resources are not appropriate for the targeted didactic goals as they can be specialized by selecting new elements of the complementary dimension. For instance, the second goal CONCEPTUAL x REMEMBER has the concept with intent Conceptual Knowledge, which owns a 12% of the resources. These resources cannot be more specialized with the cognitive process Remember so it is expectable that they will fit well for any other cognitive process such as Understand, Apply, Analyse, Evaluate or Create.

![Figure 5: Lattice views for each of the desired didactic goals](image-url)
Figure 6a shows the fusion of the two didactic goals FACTUAL x UNDERSTAND and CONCEPTUAL x REMEMBER. Although the bottom concept is still the most populated, a careful exploration of the resulting lattice can give us extra information. For instance, Figure 6b shows in bolder lines the closure set for the formal concept having [FACTUAL KNOWLEDGE, UNDERSTAND] as intent, which includes all its subconcepts and superconcepts. It can be seen that a 5% of the resources can be used exclusively for this didactic purpose (i.e. for these resources the system has identified only those descriptors belonging to such didactic goal). In addition, a 10% of the resources can also be used for this didactic goal but taking into account that they are also suitable for a Remember cognitive process (Figure 6c). On the other hand, the concept lattice shows that it is not possible to isolate the didactic goal CONCEPTUAL x REMEMBER. Hence, all the retrieved resources suitable for such goal (that is, the bottom concept of the lattice that owns a 67% of the resources) could also be useful for a FACTUAL x UNDERSTAND didactic goal.

Figure 7 shows how the framework can help the user to find resources not only well suited for a specific didactic goal but also to find relevant resources according to a specific type (e.g. images, definitions, lists, theorems, etc.). In this figure we are depicting the didactic goal FACTUAL x UNDERSTAND with its associated terms. For instance, if the user would like to use those resources with Definitions he could easily identify that the system has retrieved two such
resources (a 5% of the retrieved resources) for her. Moreover, she could also find that there is not too much material including Timelines (only on the 2% of the resources) and that they always occur with Definitions and Images. Finally, he could find that Images are the most frequent types of content not only for this goal but also for other goals (i.e. comparing the owned objects between the Images attribute concept and its lower neighbour).

4. Conclusions and Future Work

In this paper, an approach has been presented for an e-learning material retriever and organizer that makes use of formal concept analysis to arrange a collection of web resources as a set of interrelated clusters linked by subsumption relationships. In our opinion, this is a fresh promising solution because it makes up a proper tool to assist teachers and students to leverage learning experiences with richer and more significant assets. It is also a step forward in the attempt to use the web as a learning object warehouse.

We can highlight several advantages of using FCA as a mediator in this solution. Firstly, with this approach it is possible to get a grasp on one entire domain by means of a single retrieval and organization task, instead of trying to filter and arrange a large number of results from blind queries to external search engines, as it has been a common practice. Secondly, this process is also oriented towards learning goals, that is, the internal queries, that is, the ones not defined by the user, ran within the aforementioned single task are enriched by proper terms aligned to learning goals, which shifts the outcome so that the results are biased to the current didactic purposes. Thirdly and finally, FCA supplies, by constructing a lattice, a visual representation, which is likely to be helpful for both instructors and students.

In addition, the outcome provided by FCA, that is, a latticed clustering, can be computationally processed for any other purposes not explicitly covered here. As typical examples of such processing, we can propose the following scenarios that could be considered as future work:

- Domain indexing: the most direct and intuitive way of exploiting this tool coming to mind is that of indexing a knowledge domain for the sake of getting resources rankings, taxonomies,
An Architecture for Retrieving and Organizing Web Resources for Didactic Purposes

classifications and other possible arrangements useful for instructional designers.

- Domain exploring and browsing: the lattice visual nature makes it easy to navigate the resource collection retrieved by the domain organization process. Documents are classified under a number of didactically relevant concepts, such as-in our didactic framework-cognitive processes, types of knowledge and terms describing resources, which adds meaning to this navigation. As this solution neither claims nor uses any formal domain representation, a navigational model based on lattices rather than a plain list is definitely a step forward in searching for useful resources. This organization allows choosing a given kind of resource appropriate for a didactic goal, such as examples or definitions by navigating to the formal concept that is labelled with tags referring the aforementioned goal. The downset of this concept will contain all the retrieved resources referring to that kind of entity.

- Repository population: an unsolved problem in Learning Object and Digital Repositories is achieving an automatic way of populating the metadata associated with new learning objects where those metadata are somehow extracted from the context. Particularly, in our approach the FCA clustering information could be used to provide a context whence metadata would be extracted.

- Knowledge synthesis and elicitation: formal concept analysis can be used to potentially “restructure” an existing repository as well as to correlate concepts into machine-understandable statements, which can be exploited to help to perform knowledge synthesis or inference. A possible scenario in this case could be a combination of firstly implicit instructional knowledge classification in lattices and secondly, knowledge elicitation by means of a set of inference rules.

- Instructional workflows and strategies: these kind of workflows can be seen as an ordered set of pedagogical activities arranged after a given sequencing logic where students must visit subject concepts to fulfil some didactic goals. These workflows would go along with exploration strategies over the lattices that drive the retrieval of proper resources. In fact, when undertaking an activity, the lattice could be browsed to recall those resources properly fitting the goal associated with that activity to enrich the learning experience. For instance, when presenting a new concept in a behaviourist learning set up, the framework could suggest the resources closest and possibly best fitted for a given cognitive process and type of knowledge focused on a given domain. Nevertheless, if this were a collaborative learning setting, the system could surely suggest resources loosely fitted to the didactic goal so that some discussion could be fostered.

As we have tried to show throughout this paper, the described work is an approach to overcome the problems of creating instructional content. Clustering capabilities provided by FCA can create navigable lattices and sub-lattices useful enough for instructional designers. For example, within our realistic concept proof, they are supplied with a meaningful classification based on cognitive processes and knowledge types.

Nevertheless, we must highlight that the lattice is built once at the beginning of a single system request. This construction process cannot be carried out otherwise, as this is a heavyweight task, whose computational cost depends on the number of formal objects and attributes. This fact yields two consequences. On the one hand, the usage methodology for instructional designers changes with respect to typical solutions. If in those solutions, designers ran a number of blind web queries whenever they needed some content for their purposes, now they must perform a
single task, whose outcome covers the entire cognitive domain and organizes it around a number of conceptual clusters, bound to didactic goals. This mode of usage must be taken into account by the users in order to get the most appropriate results for their purposes.

On the other hand, this methodological shift imposes some constraints on the users, as the resources are not recovered on demand as it is usual in typical e-learning applications but they are retrieved just once, at the beginning of the aforementioned single task, and they are later explored upon request. This procedure can be considered static because it limits the exploration with respect to the initial capturing mechanism, as the resource collection will be closed at the moment when the lattice is to be built up instead of being open and dynamic as it happens in on demand querying approaches.

Furthermore, users must also be aware that domain queries should be open enough to refer to a sufficiently wide subject matter to get the most significant results from this technique. For instance, in our examples, we used “Algebra” instead of, for instance, “Linear equations solving”, as the latter would be too narrow a query to be an effective input for this software application.

It is important to remark that the collection of terms selected to fill the matrix of goals is not a prescriptive element of this approach. It is rather a parametric dimension allowing to configure the didactic framework to fit a particular instructional setting. With this, we want to say that the value of this paper does not rely on those terms as they are only a proof of concept of the validity of our solution. It will be the instructional designer’s responsibility to tune them up according to the requirements of the specific instructional experience where this framework is to be used. Our preliminary tests carried out with the configuration shown in this paper have been promising. Nevertheless, more realistic tests should be undertaken in specific didactic settings to determine its pedagogic value and performance. There is still a huge field to explore based on the combination of instructional content and clustering with FCA and useful results are still not always achieved.

Acknowledgements

We thank Madrid Regional Government for the support of E-Madrid Network S2009-TIC-1650 (http://www.emadridnet.org) and MAVIR Network S2009/TIC-1542 (http://www.mavir.net/). We also thank Ministerio de Ciencia e Innovación (MICINN) for the support of CREASE project TIN-2009-14317-C03-03 (http://creaselsi.uned.es | http://crease.googlecode.com).

References


An Architecture for Retrieving and Organizing Web Resources for Didactic Purposes


SCORM (2004): ADL guidelines for creating reusable content with SCORM. On the web at legacy.adlnet.gov/Technologies/scorm/SCORMDocuments/SCORM Resources/Resources.aspx


Biographical Notes

José Ignacio Mayorga is a lecturer at the Universidad Nacional de Educación a Distancia (UNED, the Spanish Open University), where he has been teaching and researching since 1993. He is currently working within UNED’s Learning Technologies and Collaborative Systems Group to obtain his PhD in computer science. His research interests include the use of technology to assist teaching and learning (e-learning, technology-enhanced learning, learning design). He participates in the network of excellence NoE – Kaleidoscope wherein he has been involved in several projects for technology enhances learning.

Javier Vélez Reyes is a computer science engineer at the Universidad Politécnica de Madrid since 2001 and gained his PhD in Computer Science from the Open University of Spain (UNED) in 2009. Currently, he is working as a researcher in the latter university within the Learning Technologies and Collaborative Systems Researching Group. His specialization area is related to learning design technological solutions and the construction of systems for collaborative learning support. As a part of his researcher development, he has work as an architect in the design and construction of learning software solutions based on the use of learning objects framed in several researching projects with both national and international scope. Nowadays he belongs to the net of excellence NoE – Kaleidoscope wherein he has been involved in several projects for technology enhanced learning.

Juan Manuel Cigarrán is a lecturer at UNED. He has been involved with research and teaching at university since 1997. He received the PhD in computer science from the Universidad Nacional de Educación a Distancia (Spain) in 2008 and he got the best PhD award in the same year for his research on dynamic clustering of web search results. His research interests include multilingual and multimedia information access, information retrieval and clustering, natural language processing, visualization and interaction. His work on these areas has been published in national and international journals, conferences and workshops. He has also participated in several European and National projects as senior researcher.

Miguel Rodríguez-Artacho (PhD) is a researcher and associate professor at UNED. He is senior member of the IEEE and head of the Cooperative Learning Technologies and Systems Group (http://ltcs.uned.es) at LSI department. His research focuses on educational content and processes modelization and collaborative learning. Miguel is coordinator of the CREASE and eMadrid (http://www.emadridnet.org) projects at UNED, and shares the technical secretariat of the AEN/CTN71/SC36 group for learning technologies standards in the Spanish Agency for standardization. He has also been invited to conferences related to learning technologies and has several publications on educational content and design of collaborative learning environments.