2009 Ninth International Conference on Intelligent Systems Design and Applications

WikiArt: An Ontology-based Information Retrieval System For Arts

Roberto Pirrone Vincenzo Cannella Orazio Gambino Arianna Pipitone
Giuseppe Russo
Dep.t of Computer Science and Engineering (DINFO)
University of Palermo
Viale delle Scienze, Building 6, 90128 Palermo, Italy
{pirrone,cannella,gambino,pipitone,russo}@dinfo.unipa.it

Abstract

Retrieving and integrating contents of different kinds is a complex task. Actually the web offers few examples of systems able to carry out such a task. This paper presents WikiArt, a system able to integrate three different kinds of information sources: a database, a wiki, and an ontology. All these three elements regard arts. Finally, WikiArt is able to generate automatically thematic paths to browse its contents thanks to an ad-hoc procedural knowledge base.

1. Introduction

Wikis are nowadays more and more used to carry on information in many domains. This paper describes an ontology-based information retrieval process for the WikiArt System. WikiArt is a specialized wiki system able to show advanced results for user interested in Arts. To start up the wiki, a huge collection of documents regarding artists, works of art, artistic movements, and so on has been created. All this content has been organized into a relational database. WikiArt allows a semantic retrieving of these contents. The interaction with users is managed by an agent with a knowledge base about arts expressed through an ontology. The agent is able to supply contents to the users not only on the basis of the analysis of a query composed by tags, as in a common wiki. It is also able to plan a more complex browsing of the content of the wiki on the basis of the typologies of contents and relations described in the ontology. WikiArt is not intended to be a semantic wiki as intended in literature. WikiArt has some different characteristics with respect to a semantic wiki. In WikiArt every concept is linked to one or more pages of the wiki, and the relations between the concepts are described in a distinct ontology. In a semantic wiki concepts and relations are inserted into the text of the pages. This second approach could be added to the system as a future possible improvement and extension. The retrieving of contents from the wiki is planned on the basis of a series of possible different approaches. The browsing of the ontology is described by general exploration patterns. A pattern represents a way of browsing and collecting content to reply to a request of the user. This task can be inspired to teaching criteria.

The rest of the paper is organized as follows: next section presents briefly some related works. An overwiew of the WikiArt system is presented in the section 3. Section 4 discusses about the ontology and the mapping process between the ontology and the database. Section 5 presents the Exploration Patterns used to help users in the navigation process. The WikiArt implementation is depicted in section 6. Finally, section 7 presents some conclusions and the future work.

2. State of The Art

Many models have been proposed in literature during the years with the purpose to perform efficient information retrieval . A common characteristic is an implicit or explicit definition of a conceptual framework [2] able to underline the investigated phenomena.

In the latest years the information retrieval systems on the web have to challenge a very important aspect regarding how users perform their data access strategies. Different studies point that usual search on web is performed with a relatively poor set of terms and rarely with advanced search capabilities [14] [11] [13].

Starting from the richness and the complexity of the materials on the web a new and more interactive way of finding and retrieving documents becomes necessary [3].

A common way to enable better performances on research is to use ontologies for representing both the domain organization and the possible paths through concepts. Developing such systems regards different application areas.

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.219

The focus of the present work is on the retrieval of contents about arts. Some approaches use ontology at a low level in the retrieval framework to describe particular aspects of the involved domain such as the work presented in [5]. In this paper a common ontology is used to describe features of the art images to be retrieved. Other systems uses ontologies as enabler to perform simpler and more efficient queries over a complex and structured database regarding particular artifacts [12]. Many systems [10], [7], [6], [8] are organized like semantic wikis. The ontology instances and concepts in a semantic wiki are directly associated to a wiki page. In this way wiki links and annotations are related the concepts while pages are related with each other. The navigation of the ontology and the definition of rules allows to obtain more accurate matching and better documents and pages retrieval.

3. Overview of WikiArt

The core of WikiArt is a wiki of documents treating contents related to the arts. At the start up, the collection of documents contained pages for more then 15000 artworks and more then 4000 artists.

Moreover, the collection contains many documents treating artistic movements, and specific techniques employed by artists. Besides the documents, the system has a very huge relational database collecting many data regarding arts. The domain managed by the system has inspired the design of its schema. This allows a specialized treatment of the contents of the wiki. This approach is different from that adopted in a common wiki, where the contents are archived in the database with no particular attention to their nature, and to the elements of the treated domain. The database contains data and information together. On the one hand, in fact, it contains data regarding artists and artworks, as dates, names, locations. All these elements are pure data. On the other hand, the database is used also by the wiki to manage the pages with their versioning. The third element of the system, the knowledge base, manages the knowledge of the system about the arts. In this way, the system allows an integrated management of the first three levels of the well-known DIKW (Data, Information, Knowledge, and Wisdom) hierarchy.

The relationship diagram reporting the relations among the tables is shown in Fig.1-RIGHT. The meanings of the tables Artist, Work of Arts, Birth and Death are obvious. In some cases the author of a work of arts is anonymous, but the artist was the apprentice of an important master. For the same reason, a work of arts can be retrieved by specifying the period (intended as historical period or time range), the artistic current or movement (Impressionism, Futurism and so on), the subject (Madonna, San Gerolamo etc etc), typology (painting, sculpture...) or artistic techniques. The field values of the corresponding tables just mentioned should be constrained, but this choice should preclude the chance to define new values. The *Location* table specifies generic information (for e.g. *location_typology*: museum, *location_name*: Louvre) about the physical site where the work of arts is; while the *Place* table is referred to the geographic position of the site, as it can be seen by its fields. The *Chronology* and *Document* tables are used for the content versioning. If a less informative or "vandalic" new version has been stored by a bad editor, the document can revert to the previous one. The pages of the wiki are automatically linked by the system, according to some rules and what is recorded in the database. For instance, the page of an artist lists the artworks of that artist recorded into the database as work by him. In this way, some sections of the pages of the wiki are generated automatically by the system on the basis of the collection of data recorded into the database.

Three kinds of users can browse the WikiArt internet site: visitors, editors and administrators. The visitor is a generic internet surfer who visits WikiArt without any need of registration or authentication. A visitor can only perform a query to search a particular work of arts, information about an artist, a monument, an artistic technique or movement. The research can be done by writing a keyword in the proper form of the welcome page, by clicking on the list of all the artists stored in the system or through an advanced research. WikiArt can storage and manage documents written in many languages: Italian, English, German, French, Spanish, Chinese, Portuguese, Russian, Arabian, Albanian. Other languages will be provided for the future versions of the system, if required. Fig.1-LEFT shows the Italian version of the WikiArt welcome page. The editor can write new information or editing the existing one. For this reason, an authentication step is required for this kind of user and his account registration must be approved by an administrator, who holds all the privileges in the users hierarchy.

4. The Ontology of WikiArt

The WikiArt database is mapped in the OntoWikiArt.owl ontology defined as an OWL-DL profile. The basic concepts in the ontology are the entity tables in the database; the columns which are no primary keys are defined as datatype properties and columns which are primary keys are defined as object properties.

As an example, the following definition is an object property definition in the owl schema. The Property "hascreated" links the concept Artist to the concept Opera.

$< owl : ObjectPropertyrdf : ID = "hascreated" >$
$< rdfs : domainrdf : resource = "\sharp Artist" / >$
$< owl : inverseOf >$
$< owl : FunctionalPropertyrdf : ID = "createdby" / >$
$< /owl : inverseOf >$

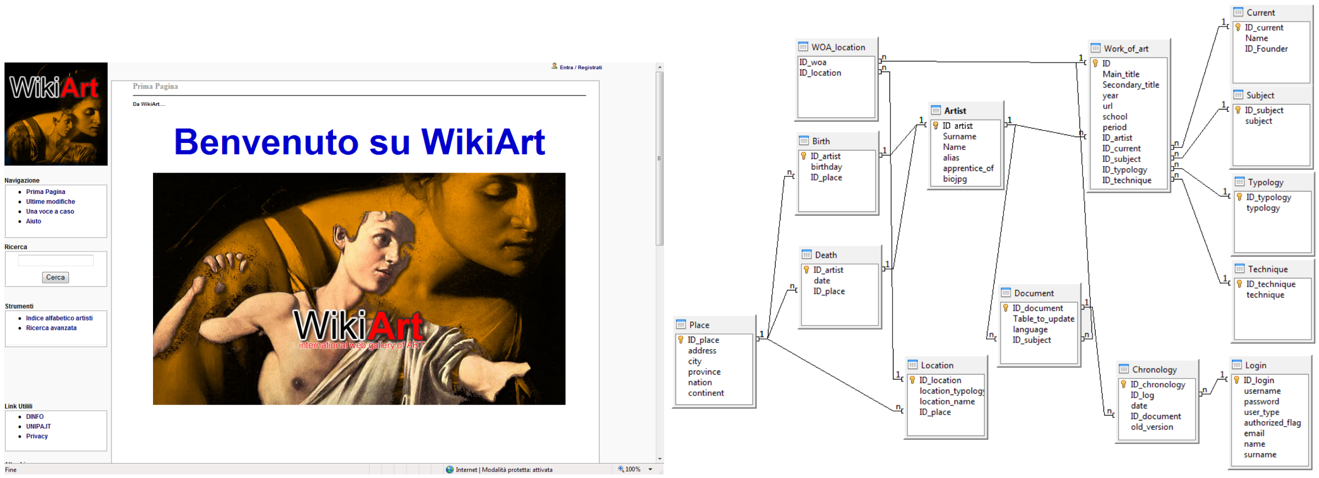

Figure 1. LEFT: The welcome page of WikiArt. RIGHT: The relationship diagram of WikiArt Database about the domain. Service tables are discarded

$< rdfs : rangerdf : resource = "\sharp Opera" / >$
$< /owl : ObjectProperty >$

The ontology database mapping is obtained trough a proper transformation able to transform ontology concepts into to a proper selection of significative subsets of data responding to a conceptualization in the domain. To this aim a methodology based on the utilization of an XML file for the definition of a correspondence has been defined. XML is widely accepted as a universal and neutral language for data exchange over the web. This methodology simply allows to set a correspondence between a concept and a set of tables representing a portion of the database. To this aim a proper DTD for the transformation has been defined. A concept is related to a set of tables. For each table is possible to define the set of related attributes, the values of these attribute and how the attributes have to be used (attributes can be related with the OR, AND and NOT relation). The level of granularity that can be reached is very precise. Furthermore the instances on the database are obtained simply by querying the tables mapped to the concept. For a more accurate definition of the methodology the reader is referred to a previous work of some authors [9].

5. The Ontology Exploration Patterns

The content of WikiArt can be consulted like any other common wiki. Moreover WikiArt is able to propose thematic paths about specific subjects. A thematic path can be defined as a sequence of subjects presented to the user one after another, when the user asks for a particular subject. Each subject in the thematic path is joined to a specific document in the wiki regarding that subject. For instance, each step of the path needs to deepen, or detail a particular aspect of the main subject.

The thematic path can also represent the time-line of the development of a particular element in the history of arts, as, for instance, the artistic experience of an artist, an artistic movement, or a particular artistic technique with its improvements. Each step of the thematic path has a particular meaning, represented by a informational goal. When a user asks a subject, the system offers some possible thematic paths related to that subject. These thematic paths are generated automatically. To this aim, we have defined a set of thematic path generation rules. Such rules dictate how to choose the subjects, and how to arrange them. According to this rules, the system browses the arts ontology, following relations between concepts, and establishes the best concepts for the path. The thematic path generation is in fact tightly related to the direct exploration of the ontology by the system.

This task must be inspired to a series of criteria. A specific collection of browsing criteria is called *Exploration Pattern (EP)*. These criteria must detail how the graph must be explored, which nodes must be selected, how the nodes of the path must be ordered, and many other aspects.

In general, the ontology browsing can be compared to a graph browsing. In fact, an ontology can be viewed as a graph, where concepts are nodes in the graph, and relations are arcs. A graph can be explored in many manners, all of them well investigated. Possible examples are the deep-search first and the breadth-first search. The search criterion is not enough to establish the path. When the system must choose through which out-going arc to go, other criteria must be given. In particular, a preference order for arc types can be defined. Each arc of the ontology is tagged

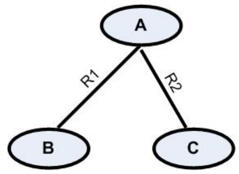

Figure 2. Concepts A and B are joined by a R1 relation, and concepts A and C are joined by the R2 relation. If $R1 < R2$, the system visits first B, and then C.

with the name of a relation between two concepts. If all involved relations are strictly partially ordered, the choice of the next arc can be dictated by this ordering. Consider the simple graph in the image 2.

If according to the order, R1 is less then R2, the system chooses first the first arcs, visiting the B node. When the system is on the A node again, and it must to make a next choice, the only R2 arc will be left. Establishing an order for relations is similar to give a weight to them. An arc preceding another one in the order has the smaller weight. Obviously, the order between relations defines a transitive relation. If $R1 < R2$, and $R2 < R3$, then $R1 < R3$. This last assertion could be inferred too. Two relations of those employed to build a thematic path must not to be always comparable. When there is no order between two relations, the system chooses randomly which arc to follow. Similarly, concepts can be ordered too. According to a specific order, a concept can have priority over another concept. In this case, when the system must to choice which child node to visit, it could choose the node with highest priority and not yet visited. Concept ordering must be transitive too. Transitivity could allow to infer other orderings between not explicitly ordered concepts. Concepts must not to be always comparable. If there is no explicit order between two concepts, the system chooses randomly. A thematic path can be made up only by concepts in ascending (at least not descending) order.

Both criteria guide our system during the exploration of the ontology. Given an order for relations, and an order for concepts, the system browses thematic paths reflecting those orders.

This can cause some problems. An order between relations defines, indirectly, an order between concepts. For instance, in the Figure image 2, the defined order between relations R1 and R2 implies an order between the concepts B and C. If the system knows, or can infer, nothing about them, this inference is valid. On the contrary, the knowledge of the system can not be contradicted. The knowledge of the system grows during the building of the thematic path. During the exploration of the ontology the system can make inferences only on the basis of what just visited in the ontology, and of the constraints on concepts and relations. This choice is useful to avoid inconsistencies that could be produced with inferences by the system. The most important properties of the thematic path generation procedure have been described. A last element is the stopping criterion. We have implemented two possible criteria: the maximum number of visited nodes in the ontology, and the maximum depth of the search. These two criteria can be used together.

An *EP* is defined as a structure made up by five slots. Each slot describes a particular facet of the ontology browsing: the exploration modality, the ordering of concepts and relations, the maximum number of nodes in the thematic path, and the maximum depth of the search.

The knowledge base of the system contains several instances of *EPs* defined manually by programmers. An *EP* can be defined according to different criteria. A concept can be preparatory for another concept, and should be always included before another one in a thematic path. In a timeline presentation, the birth of an artist should precede his death. Similar consideration can be made for relations. An EP could be developed manually by a developer, and selected explicitly by the users. Designing manually a pattern allows to produce thematic paths with certain characteristics for particular thematic goals. An EP defines the organization and the structure of a generic presentation, employed to build presentations with different contents, on the basis of which part of the ontology will be explored, or which specific instances of the concepts will be used. In the first case, in fact, a similar presentation structure is used to define presentation with different starting subjects. Starting the exploration of the ontology from different points leads to different portions of the ontology. On the other hand, an exploration of the ontology can generate different presentations on the basis of which specific subject has been asked by the user. For instance, a presentation can be structured starting from the concept Artist, collecting related biographical data and works of art. Obviously, presentations with identical structure are different if they treat different instances of the concept Artist, as Michelangelo and Leonardo.

6. The implementation of WikiArt System

The WikiArt System is a rule-based system. The architecture of a typical rule-based system has three components:

- The inference engine with the two subcomponents: the pattern matcher and the agenda
- The working memory (WM)
- The execution engine

The pattern matcher in the inference engine applies the rules in the rule-base to the facts in working memory to construct the agenda. The execution engine fires the rules from the agenda, which changes the contents of working memory and restarts the cycle. The system is developed using JessTab [4] [1], a plugin for Protègè, that allows to use Jess and Protègè together. Using JessTab, it is possible to manage Protègè knowledge bases directly, and to write rules that fire on patterns in the knowledge base. It is possible to use JessTab as an object-oriented extension to Jess by defining Protègè classes and instantiating them. Because, Protègè has explicit metaclasses we can map them to facts as well and perform pattern matching on properties of classes. Usually JessTab systems are organized in modules, a sets of JessTab rules with a specific scope. The WM in the system has its initial facts that establish the order between concepts and the order between relations. Precedence between two entities (relations or concepts) is defined as a couple: the first entity in the couple has the precedence on the latter one. We establish that the concepts' ordering has priority over the relations' ordering. Listing order of initial facts dictates this priority. The systems has 3 basic modules: INITIAL-FACTS, PATTERNS and PATHS modules. The INITIAL-FACTS module asserts the initial facts, and the order between concepts and relations filling the template prior. For example, the following assertion establishes the priority of concepts Born on concept Depth in a template prior with its two slots.

$(deftemplateINITIAL - FACTS :: prior(slotfirst)$
$(slotsecond))$
$(assert(INITIAL - FACTS :: prior(firstBorn)$
$(secondDepth)))$

The PATTERNS module contains rules that execute the query on the ontology to build the EP. Finally, PATHS module defines the template path computed with the patterns rules. The template path is defined in the following manner:

$(deftemplatePATHS :: path(slothead)$
$(multislotrest))$

where multislot is a list of nodes to visit in the path, starting from head slot. This list will be filled by PATTERNS module.

The following rule defines an example of EP that uses previous template. Suppose that an initial concept for the exploration is A (like Artist) and that this concept has two childs, B and C which are computed by querying the ontology. This pattern defines that if the child B has a priority on the child C, and if the subgraph must be explore in profundity way with max 4 nodes to visit, then the path that will be generated must be

$A \rightarrow B \rightarrow priorchildofB \rightarrow C$

If the system does not find all nodes, it returns a different list with possible nodes or nothing because there will be other rules that will fire. Initial values (number of nodes and childs) for the first step are assigned.

$(bind?n2);$
`(not include A and C nodes)`
$(bind?bB)$
$(bind?cC)$

Then the fact for the new path which start with node A is assigned, keeping track for this fact with its ID.

$(bind?fact_id(fact - id(assertPATH :: path$
$(headA))))$

Finally, the rule modifies the path filling the slot rest until variable $?n$ is null.

$(defrulePATTERNS :: pattern_1(declare(auto - focusTRUE))$
`;exist a priority of ?b on some node`
$(INITIAL - FACTS :: prior(first\$?b)(rest\$?rest))$
`;exist a priority of ?b on some node`

The previous its a strict vinculum, but if some priority exists, It must be keep into account. To this aim this rule is repeated in the module without this test.

$(INITIAL - FACTS :: prior(first?child - of - b))$
$(test(neq?nnil)) \Rightarrow$
$(bind?temp - ch(create))$
`;here we insert child ?b in a list`
$(bind?temp - ch(insert\$?temp - ch1\$?b))$
`;here we fill multislot with list`
$(modify?fact - id(rest?temp - ch)))$
`;here we decrese variable ?n`
$(bind?n(-?n1)$
`;here we apply rule to childs-of ?b`
$(bind\$?b?child - of - b))$
`;end rule`

7. Conclusions and Future Work

The paper presents WikiArt, a new system integrating three distinct types of contents about the art: data, information, and knowledge, to generate automatically thematic paths to consult all its contents. WikiArt is a wiki, allowing to manage cooperatively documents about artists, artworks, artistic movements or techniques, and so on. It is also an expert system, provided with an ontology about arts, with which it is able to plan possible different ways of consulting and browsing its contents. This ability is made possible by a second part of the ontology of the system, describing a collection of criteria regarding how to plan thematic paths, and by a set of rules followed by the expert system to carry out this task. WikiArt is not a semantic wiki, because the ontology has not been employed to tag semantically the documents by the authors. But only their subjects. Our efforts are now devoted to the extension of the system to make it a semantic wiki too.

References

[1] Jesstab: Integrating protégé and jess. http://www.ida.liu.se/ her/JessTab.

[2] D. Engelbart. *Augmenting human intellect: a conceptual framework*. Stanford Research Institute. (Summary report AFOSR-3233), Menlo Park, CA:, 1962.

[3] E. García and M.-Á. Sicilia. User interface tactics in ontology-based information seeking. *Psychology Journal*, 1(3):242–255, 2003.

[4] E. F. Hill. *Jess in Action: Java Rule-Based Systems*. Manning Publications Co., Greenwich, CT, USA, 2003.

[5] S. Jiang, T. Huang, and W. Gao. An ontology-based approach to retrieve digitized art images. In *WI '04: Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 131–137, Washington, DC, USA, 2004. IEEE Computer Society.

[6] K. Kawamoto, Y. Kitamura, and Y. Tijerino. Kawawiki: A semantic wiki based on rdf templates. In *Web Intelligence and Intelligent Agent Technology Workshops, 2006. WI-IAT 2006 Workshops. 2006 IEEE/WIC/ACM International Conference on*, pages 425–432, Dec. 2006.

[7] F. D. Paoli and M. Loregian. On integrating a semantic wiki in a knowledge management system. In *SWAP*, 2006.

[8] L. Quan, J. Yu-ying, and C. Jing. Uimwiki: An enhanced semantic wiki for user information management. In *IT in Medicine and Education, 2008. ITME 2008. IEEE International Symposium on*, pages 930–934, Dec. 2008.

[9] G. Russo, A. Gentile, R. Pirrone, and V. Cannella. Xml-based knowledge discovery for the linguistic atlas of sicily (als) project. *Complex, Intelligent and Software Intensive Systems, International Conference*, 0:98–104, 2009.

[10] S. Schaffert. Ikewiki: A semantic wiki for collaborative knowledge management. In *In 1st International Workshop on Semantic Technologies in Collaborative Applications (STICA06*, pages 388–396, 2006.

[11] A. Spink, D. Wolfram, B. Jansen, and T. Saracevic. Searching the web: the public and their queries. *Journal of the American Society for Information Science and Technology*, 53(3):226–234, 2001.

[12] K. N. Vavliakis, A. L. Symeonidis, G. T. Karagiannis, and P. A. Mitkas. Eikonomia-an integrated semantically aware tool for description and retrieval of byzantine art information. *Tools with Artificial Intelligence, IEEE International Conference on*, 2:279–282, 2007.

[13] P. Wang, M. W. Berry, and Y. Yang. Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8):743–758, 2003.

[14] J. Zhang, D. Wolfram, and P. Wang. Analysis of query keywords of sports-related queries using visualization and clustering. *Journal of the American Society for Information Science and Technology*, 2009.