Developing Query Patterns

Panos Constantopoulos[1,2], Vicky Dritsou[1,2,⋆], and Eugénie Foustoucos[1]

[1] Dept. of Informatics, Athens University of Economics and Business, Athens, Greece
[2] Digital Curation Unit / Athena Research Centre, Athens, Greece
{panosc,vdritsou,eugenie}@aueb.gr

Abstract. Query patterns enable effective information tools and provide guidance to users interested in posing complex questions about objects. Semantically, query patterns represent important questions, while syntactically they impose the correct formulation of queries. In this paper we address the development of query patterns at successive representation layers so as to expose dominant information requirements on one hand, and structures that can support effective user interaction and efficient implementation of query processing on the other. An empirical study for the domain of cultural heritage reveals an initial set of recurrent questions, which are then reduced to a modestly sized set of query patterns. A set of Datalog rules is developed in order to formally define these patterns which are also expressed as SPARQL queries.

1 Introduction

A common feature of the informational function of most digital library systems is that they must serve users who may not be familiar with the subject matter of a domain, or, even if they are, may be ignorant of the organization of the information in a given source. This poses the requirement of supporting users to explore sources and discover information. Moreover, it is desirable for users to have access to a repertoire of model questions, complex as well as simple, which can be trusted to represent important information about objects of interest, facilitate the formulation of particular questions and efficiently implement those.

Streamlining the formulation of questions has been addressed through several approaches ([1]). In [5] authors have developed a catalog of question templates to facilitate query formulation, where users must define a corresponding scenario for each query. Similarly in [8] questions are matched to specific question templates (called patterns) for querying the underlying ontology. In [15] a template-based querying answering model is proposed. In the area of Data Mining, and more precisely in association rule mining, many studies are based on the idea of the discovery of frequent itemsets ([14,2]). The effectiveness of such approaches can be enhanced if templates accommodate frequent query patterns. Algorithms for discovering such patterns have been developed in areas with a large amount of

⋆ Vicky Dritsou was supported by E.U.-European Social Fund(75%) and the Greek Ministry of Development-GSRT(25%) through the PENED2003 programme.

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 119–124, 2009.

Table 1. From propositions to tacit questions

Proposition	Tacit question
This statue was found at the same place with the statue No 1642.	Which other objects have been found at the same place with the current one?
This object is a copy of a famous scultpure made by Kalimachos.	Who created the original object whose copy is the current one?
This object was exhibited at the museum of Aegina until 1888.	Where was this object exhibited at during other periods of time?

data, such as biology and medicine ([9,4]) and social networks ([12]), but with no separation between the syntactic and the semantic layers of these patterns.

In this paper we make the case that the "right questions" in a domain can be captured through query patterns that are characteristic of the domain of discourse. More precisely we address the development of query patterns at successive representation layers; this layered abstraction (from semantics to structure) can elegantly support the subsequent development of information search and delivery services. We ground our work on the analysis of a specific domain, namely cultural heritage, through a selected sample of collection and exhibition catalogs, where we consider each statement as an answer to a tacit question. A large, yet not vast, set of recurrent tacit questions is revealed from our study: we then focus on reducing it to a set of query patterns. These patterns determine the structure of the questions and disclose the most interesting information within the specific domain. As research has shown that digital libraries can benefit from semantic technologies and tools ([10,3]), these query patterns are conceived as RDF-like graphs and they are shown to be generated from an even smaller set of graph patterns, here called signatures. Finally, we express our query patterns into two well-known query languages, namely Datalog and SPARQL.

2 Empirical Study

In order to conduct a grounded analysis, we considered determining query patterns in the cultural heritage domain first. Such patterns, together with their relative frequency of occurrence, can be taken to represent the largest part of the information needs of users in the domain. Rather than conducting a user study, we turned to an indirect source, namely descriptive texts written by experts about objects in specific collections. These texts contain statements about objects and concepts, which we interpret as answers to tacit questions: questions that could be answered by the statements at hand. Query patterns, abstracted from this set of tacit questions, essentially capture the domain experts' perceptions of what is important to know in a given domain.

Our study comprised five collection catalogs published by the Greek Ministry of Culture (see references in [6]). In total the descriptions of some 1250 objects were studied and a list of tacit questions underlying the object descriptions was

compiled. Examples of this question extraction process are shown in Table 1. We contend that these questions lead to information considered as important by the experts. We further contend that these questions also correspond to possible queries that interested users should be able to execute in order to retrieve important information. Therefore, the execution of such queries should be facilitated by cultural heritage digital library systems.

The initial outcome of this study was the recording of 82 distinct questions. Studying these further, we noticed that most of them recur throughout many different descriptions. This recurrence is, of course, expected of the who, where, when, what queries that stand as universal for all exhibits and all types of resources. The remaining questions show a document frequency of at least 25%.

3 Patterns

In the current work, we first set out to analyze the recurrence of queries that we noticed among our 82 queries. As a first step we partitioned the set of these queries into classes of queries that share a common structure at the schema level: this analysis revealed 8 different structures, each one represented as an RDFS graph called "signature" in what follows (see Fig. 1). As a second step we refined the previous partition into a finer one, thus producing 16 structures, each one also represented as an RDF graph called "graph pattern" in the sequel. These 16 graph patterns are given in [6]. Each of them consists of two subgraphs: the first one is a signature among the 8 ones of Fig. 1 and the second one is an "instance" of that signature. An example of such a graph pattern is given in Fig. 4. Each graph pattern describes the common structure of queries in a finer way since it captures not only their common structure at the schema level but also at the instance level. Signatures as well as graph patterns give structural patterns.

In order to better describe the structure of a query, we have to assign to each element of a graph pattern either the characterization "*known*", if it corresponds to an input element of the query, or otherwise the characterization "*unknown*" or "*searching*". Thus, starting from our 16 graph patterns, we develop all possible queries that can be expressed through them. Considering all possible combinations (on the same graph pattern) of known and unknown elements leads to structures richer than the graph pattern itself: we call these structures query patterns.

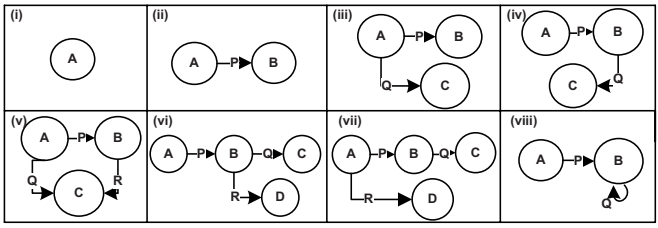

Fig. 1. Signatures of recurrent questions

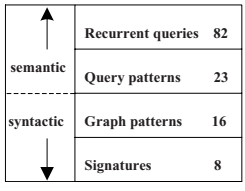

Fig. 2. Pattern types

However, not all such possible combinations represent meaningful queries in the domain of discourse; our goal then is to identify the valid query patterns, where the validity of graph/query patterns is judged by whether they correspond to at least one recurrent question recorded in our empirical study. The above process resulted in the development of 23 query patterns that are valid for the domain of cultural heritage. From now on we call query patterns only the valid ones. A schematic representation of the abstraction hierarchy of all the aforementioned pattern types and their frequencies is shown in Fig. 2.

The developed query patterns have been expressed as Datalog rules on the one hand, following the recent researches on the development of a Semantic Web layer using rules expressed in logic languages with nonmonotonic negation ([7]). On the other hand, the impact of SPARQL together with current studies that suggest translations of SPARQL to Datalog in order to express rules ([11,13]), lead us to express these query patterns as SPARQL queries as well. Due to space limitations, we do not present here the theoretical foundation of our work and we therefore refer interested readers to [6]. Instead, we study in the sequel a complicated example and show how to define its pattern.

4 Defining and Using a Query Pattern: An Example

Consider the following query: "Find all objects that depict the same figure as the current object, but which have been created by different persons" (query I). Our goal here is, first, to express this query in terms of a query pattern and, second, formulate it as a Datalog rule and as a corresponding SPARQL query.

Query expression I comprises four main concepts, namely *Object*, *Depiction*, *Figure*, *Creator*; these four concepts are particular instances of the unspecified concepts A, B, C, D respectively. Moreover, it can be checked that A, B, C, D are interrelated according to signature (vii) of Fig. 1. Thus signature (vii) gives the structural pattern of our query at the schema level; by properly instantiating signature (vii) we obtain the RDFS graph of Fig. 3, which constitutes the semantical pattern of the query at the schema level. A finer structural pattern is given in Fig.4; this graph pattern is obtained by enriching signature (vii) with an instance level.

From our current query we produce the graph pattern of Fig.4 which has two copies at the instance level. In order to understand how we produced this

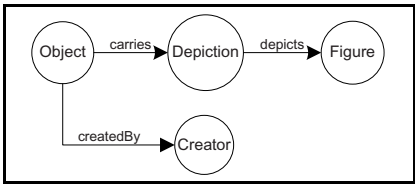

Fig. 3. RDFS graph of sign. (vii)

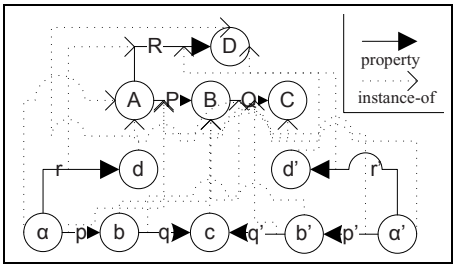

Fig. 4. Graph Pattern I

complex graph pattern from query I, consider the following reformulation of the query: "For the given object a [the current object] in $A : Object$, for the given object b [depiction of the current object] in $B : Depiction$, for the given object c [figure of the depiction of the current object] in $C : Figure$, for the given object d in $D : Creator$, and for the given properties $p : P : carries$, $q : Q : depicts$, $r : R : createdBy$, find all objects a' in $A : Object$, a' different from a, such that a' $createdBy$ d' in $D : Creator$, d' different from d, and a' $carries$ the depiction b' which $depicts$ c." Clearly, the known variables are $A, P, B, Q, C, R, D, a, p, b, q, c, r, d$, the unknown are a', p', b', q', r', d' and the searching variable is a'. That reformulation makes explicit the semantical pattern of the query: it constitutes an "instance" of the graph pattern of Fig. 4. If we remove the semantics from the above reformulation of our query, we obtain its query pattern. Therefore, the Datalog rule that retrieves the desired answer of query I (and of every query having the same query pattern) is:

$$S(x_A) \leftarrow P_{AB}(a,p,b), Q_{BC}(b,q,c), R_{AD}(a,r,d), P_{AB}(x_A,x_P,x_B),$$
$$x_A \neq a, Q_{BC}(x_B,x_Q,c), R_{AD}(x_A,x_R,x_D), x_D \neq d$$

This Datalog rule can also be easily translated in SPARQL as follows:

```
SELECT   ?x_A
WHERE { ?x_A   ns:P      ?x_B;     rdf:type   ns:A.
        ?x_B   rdf:type  ns:B;     ns:Q       ns:c.
        ?x_A   ns:R      ?x_D.
        ?x_D   rdf:type  ns:D.
        FILTER (?x_A ≠ ns : a && ?x_D ≠ ns : d). }
```

where ns is the namespace prefix of the ontology.

5 Testing and Discussion

In an attempt to test the adequacy of the query patterns to express arbitrary queries submitted by users, we used a set of dialogs between a robot in the role of an online guide of a museum and its online visitors. These dialogues were collected during Wizard of Oz experiments of project INDIGO (see `http://www.ics.forth.gr/indigo/`) and a total of 23 distinct tours were examined, each one of them containing from 10 to 50 questions. The interesting result is that 93% of the recorded user questions are covered by our developed list of query patterns and the respective Datalog rules and/or SPARQL queries, giving evidence in support of the validity and adequacy of the query pattern development process that we employed.

Query patterns can be exploited to provide user guidance in the difficult task to express a question correctly (a) in terms of the underlying schema and (b) in a given formal query language. Moreover, patterns reveal features of the domain that the user may have not been previously aware of, or filter the available information in such a way, as to highlight the most interesting features. From the developer's perspective, systems can be developed in such a way, so as to provide quick results to these frequent questions. Additionally, the recurrent questions we

have established can be considered as predetermined views over the conceptual models of different information sources, calling for the implementation of special mappings and shortcuts. Efficient shortcut definition is a field we are currently exploring. Finally, we believe that query patterns for other domains, sharing certain commonalities, can be developed in analogous ways and we intend to investigate this further.

References

1. Andrenucci, A., Sneiders, E.: Automated Question Answering: Review of the Main Approaches. In: Proc. of the 3rd IEEE International Conference on Information Technology and Applications (ICITA 2005), Sydney, Australia (2005)
2. Bayardo, R.J.: Efficiently mining long patterns from database. In: Proc. of the 1998 ACM SIGMOD Conference, Seattle, USA (1998)
3. Bloehdorn, S., Cimiano, P., Duke, A., Haase, P., Heizmann, J., Thurlow, I., Völker, J.: Ontology-Based Question Answering for Digital Libraries. In: Kovács, L., Fuhr, N., Meghini, C. (eds.) ECDL 2007. LNCS, vol. 4675, pp. 14–25. Springer, Heidelberg (2007)
4. Borgelt, C., Berthold, M.: Mining Molecular Fragments: Finding Relevant Substructures of Molecules. In: Proc. of the 2002 ICDM Conference, Japan (2002)
5. Clark, P., Chaudhri, V., Mishra, S., Thoméré, J., Barker, K., Porter, B.: Enabling domain experts to convey questions to a machine: a modified, template-based approach. In: Proc. of the K-CAP 2003 Conference, Florida, USA (2003)
6. Constantopoulos, P., Dritsou, V., Foustoucos, E.: Query Patterns: Foundation and Analysis. Technical Report AUEB/ISDB/TR/2008/01 (2008), `http://195.251.252.218:3027/reports/TR_2008_01.pdf`
7. Eiter, T., Ianni, G., Polleres, A., Schindlauer, R., Tompits, H.: Reasoning with Rules and Ontologies. In: Barahona, P., Bry, F., Franconi, E., Henze, N., Sattler, U. (eds.) Reasoning Web 2006. LNCS, vol. 4126, pp. 93–127. Springer, Heidelberg (2006)
8. Hao, T., Zeng, Q., Wenyin, L.: Semantic Pattern for User-Interactive Question Answering. In: Proc. of the 2nd International Conference on Semantics, Knowledge, and Grid, Guilin, China (2006)
9. Hu, H., Yan, X., Huang, Y., Han, J., Zhou, X.J.: Mining coherent dense subgraphs across massive biological networks for functional discovery. Bioinformatics 21(1), 213–221 (2005)
10. Jones, K.S.: Information Retrieval and Digital Libraries: Lessons of Research. In: Proc. of the 2006 International Workshop on Research Issues in Digital Libraries, Kolkata, India (2006)
11. Polleres, A.: From SPARQL to rules (and back). In: Proc. of the 16th International Conference on World Wide Web (WWW 2007), Alberta, Canada (2007)
12. Sato, H., Pramudiono, I., Iiduka, K., Murayama, T.: Automatic RDF Query Generation from Person Related Heterogeneous Data. In: Proc. of the 15th International World Wide Web Conference (WWW 2006), Scotland (2006)
13. Shenk, S.: A SPARQL Semantics Based on Datalog. In: Proc. of the 30th Annual German Conference on Artificial Intelligence, Osnabrück, Germany (2007)
14. Shrikant, R., Agrawal, R.: Mining quantitative association rules in large relational tables. In: Proc. of the 1996 ACM SIGMOD Conference, Quebec, Canada (1996)
15. Sneiders, E.: Automated Question Answering: Template-Based Approach. PhD thesis, Stockholm University, Sweden (2002)