Query Evaluation for Mediators over Web Catalogs

Yannis Tzitzikas[1,2] Nicolas Spyratos[3]
Panos Constantopoulos[1,2]

[1] *Department of Computer Science, University of Crete, Greece*
[2] *Institute of Computer Science, ICS-FORTH*
[3] *Laboratoire de Recherche en Informatique, Universite de Paris-Sud, France*
Email : tzitzik@ics.forth.gr, spyratos@lri.fr, panos@csi.forth.gr

Abstract. The Web catalogs like Yahoo! and Open Directory are very useful for browsing and querying the Web. Although they index only a fraction of the pages that are indexed by search engines these catalogs are hand-crafted by domain experts and are therefore of high quality. We present a model for building *mediators* over Web catalogs, so as to provide users with customized *views* of such catalogs. We focus on query evaluation, specifically on the complexity of query answering by the mediator.

1 Introduction

Searching for information in various sources, such as Web catalogs, often requires the user to pose the query using a controlled vocabulary. While the use of structured and controlled vocabularies (taxonomies or *ontologies*) promises effective retrieval [4], it forces the user of the source to become familiar with the terms in the ontology. This requirement can pose a considerable burden on the user, especially when the user wants to extract information from more than one source, and the sources use *different* ontologies for indexing their objects. The need for using more than one source arises from the increasingly distributed nature of information. While it is reasonable to expect the user to be conversant with one ontology, it is quite unrealistic to expect the user to be familiar with all the ontologies that are used to index the various databases. One solution could be the use of a single standardized universal ontology. However, except for very narrowly defined subject matters, such a solution is not feasible. An alternative solution is to provide software solutions that will permit the user to pose his queries using his own terms, that are not necessarily those used to index the database being searched. The software will then translate the query into the terms of the database being searched. One way of rendering the heterogeneities of the sources transparent to the user is through the use of *mediators*.

The concept of mediator was initially proposed by Wiederhold [9]. The architecture and the functioning of a mediator is determined by the kind of the underlying sources, for example we can have mediators over relational sources (e.g. see [3, 2]), retrieval systems (e.g. see [8, 5]), etc. A model for building mediators over ontology-based sources of the kind of Web catalogs has been proposed by the authors in [6, 7]. In the present paper we study query evaluation in the model proposed in [6, 7], specifically the complexity of query answering by the mediator.

2 Sources

In this paper, a *source* consists of an *ontology* and an *interpretation*. The ontology is a pair $(T, \preceq)$ where T is a *terminology*, i.e. a set of names, or *terms*, and $\preceq$ is a *subsumption* relation over T, i.e. a reflexive and transitive relation over T. The interpretation is a function $I : T \rightarrow 2^{Obj}$ that associates each term of T with a set of objects, where Obj denotes the set of all objects of the underlying domain.

The terms of a source correspond to concepts, such as `Computer Science`, `AI`, `Databases`, that users of the source may use to formulate queries, while the subsumption relation says how these concepts are related, e.g. `AI` $\preceq$ `Computer Science`, `Databases` $\preceq$ `Computer Science`, and so on. Finally, the stored interpretation I what is currently the extension of each concept in terms of objects in the underlying domain.

A typical example of a source as defined here is a Web catalog. Indeed, a Web catalog consists of a set of terms structured by a subsumption relation and each term is associated to a set of URLs. In other words the set Obj for a Web catalog is the set of all URLs.

Figure 1 shows an example of a source. Here the stored objects are denoted by the natural numbers 1,2 and 3, dashed oriented lines are used to connect each object with the terms under which it has been indexed, solid arrows indicate subsumption, and solid non-oriented lines equivalence or *synonymy* defined as follows: $t \sim t'$ iff $t \preceq t'$ and $t' \preceq t$.

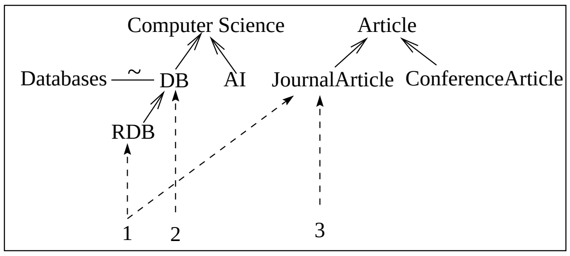

Figure 1: Graphical representation of a source

A query is a boolean expression of terms, i.e. any string that is derived by the following grammar, where t is a term of T:
$q ::= \ t \mid q \wedge q' \mid q \vee q' \mid q \wedge \neg q' \mid (q)$. Any interpretation I of T can be extended to an interpretation over the set of all queries as follows:
$I(q \wedge q') = I(q) \cap I(q'), I(q \vee q') = I(q) \cup I(q'), I(q \wedge \neg q') = I(q) \setminus I(q')$.

The answering of queries is based on the stored interpretation I and the subsumption relation $\preceq$. Specifically, each source answers queries from a model of its terminology. An interpretation I of T is a *model* of $(T, \preceq)$ if for all t, t' in T, if $t \preceq t'$ then $I(t) \subseteq I(t')$. Here, we consider that each source answers queries from the *minimal* model which is *greater* than I, denoted by $\bar{I}$, which can be computed as follows:

$$\bar{I}(t) = \bigcup \{ \ I(s) \mid s \preceq t \}$$

3 Mediators

A *mediator* M over k sources $S_1 = (T_1, \preceq_1, I_1)$,...,$S_k = (T_k, \preceq_k, I_k)$ consists of:

1) an ontology $(T, \preceq)$, and

2) a set of *articulations* a_i, one for each source S_i; each articulation a_i is a subsumption relation over $T \cup T_i$.

Figure 2 shows an example of a mediator over two sources that provide access to electronic products. The articulation a_2 shown in this figure is the following set of subsumption relationships:

$$
\begin{aligned}
a_2 \quad = \quad & \{\texttt{Products} \preceq \texttt{Electronics},\ \texttt{SLRCams} \preceq \texttt{Reflex}, \\
& \texttt{VideoCams} \preceq \texttt{MovingPictureCams}, \\
& \texttt{MovingPictureCams} \preceq \texttt{VideoCams}\}
\end{aligned}
$$

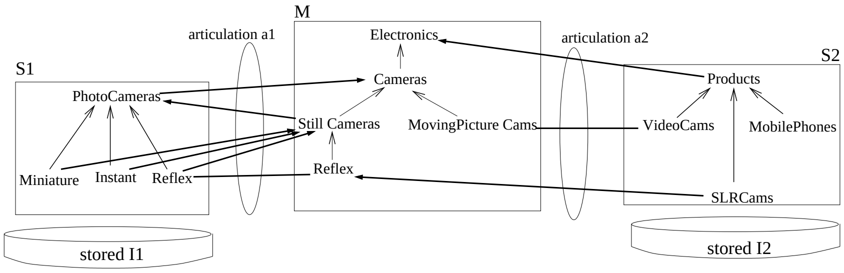

Figure 2: A mediator over two catalogs of electronic products

The mediator receives queries (boolean expressions) over its own terminology T. As it does not have a stored interpretation of T, the mediator answers queries using an interpretation of T obtained by *querying* the underlying sources. However, as the mediator and the sources have different terminologies, for computing the interpretation of a term $t \in T$, the mediator sends to each source S_i a query that *approximates* the term t, and then it takes the *union* of the answers returned by the sources. The definition of approximations is based on the articulations of the mediator. Specifically there are *two* possible approximations of a term t with respect to an articulation a_i: the *lower* approximation, denoted by t_l^i, and the *upper* approximation, denoted by t_u^i. These approximations are defined as follows:

$$
\begin{aligned}
t_l^i \quad &= \quad \bigvee \{s \in T_i \mid s a_i t\} \\
t_u^i \quad &= \quad \begin{cases} \bigwedge \{u \in T_i \mid t a_i u\}, & \text{if} \quad \{u \in T_i \mid t a_i u\} \neq \emptyset \\ t_l^i, & \text{otherwise} \end{cases}
\end{aligned}
$$

Here are some examples of approximations for the mediator shown in Figure 2:

$\texttt{StillCameras}_l^1 = \texttt{Miniature} \vee \texttt{Instant} \vee \texttt{Reflex}$
$\texttt{StillCameras}_u^1 = \texttt{PhotoCameras}$
$\texttt{Reflex}_l^1 = \texttt{Reflex}$
$\texttt{Reflex}_u^1 = \texttt{Reflex} \wedge \texttt{PhotoCameras}$

In this way, the mediator can obtain *two* interpretations I_l and I_u of T. The first is obtained by sending to each source S_i the lower approximation of each term of T, while the second by sending the upper approximation. Thus we can write:

$$I_l(t) = \bigcup_{i=1}^{k} \bar{I}_i(t_l^i), \quad \text{and} \quad I_u(t) = \bigcup_{i=1}^{k} \bar{I}_i(t_u^i)$$

The evaluation of queries can be based on the interpretations I_l and I_u. Specifically, the answer to a query q is either the set of objects $\bar{I}_l(q)$ or the set $\bar{I}_u(q)$. The user can choose the desired answer according to his information need. A user who does not want to retrieve objects which are not relevant to his information need will prefer $\bar{I}_l(q)$, while a user who does not want to miss objects which are relevant to his information need will prefer $\bar{I}_u(q)$.

4 Query Evaluation

The mediator evaluates queries by sending queries to the sources. The complexity measure that we use for query evaluation, is the *number of queries* that the mediator sends to the sources in order to answer a user's query. We consider this to be a reasonable measure of complexity as the mediator spends a lot of time waiting the answers of the sources.

If the query is a single term $t \in T$, then $\bar{I}_l(t)$ and $\bar{I}_u(t)$ can be evaluated as follows:

$$\bar{I}_{ap}(t) = \bigcup_{i=1..k} \bar{I}_i(q_{ap}^i(t))$$

where $q_{ap}^i(t) = \bigvee\{s_{ap}^i \mid s \preceq t\}$ and $ap = l, u$. Thus the mediator will send at most one query to each source.

Table 1 shows the maximum number of queries that the mediator has to send with respect to the form that a query can have. This analysis shows that a query in Conjunctive Normal Form (CNF) is preffered to a query in Disjunctive Normal Form (DNF), as its evaluation requires sending a smaller number of queries to the sources. Specifically a query in CNF, i.e. a query of the form $d_1 \wedge ... \wedge d_m$ where $d_j = t_{j1} \vee ... \vee t_{jn_j}$, is evaluated as follows:

$$\bar{I}_{ap}(q) = \bigcap_{j=1..m} (\bigcup_{i=1..k} \bar{I}_i(q^i_{ap}(t_{j1}) \vee ... \vee q^i_{ap}(t_{jn_j})))$$

while a query in DNF, i.e. a query of the form $c_1 \vee ... \vee c_m$ where $c_j = t_{j1} \wedge ... \wedge t_{jn_j}$, is evaluated as follows:

$$\bar{I}_{ap}(q) = \bigcup_{j=1..m} (\bigcap_{h=1..n_j} (\bigcup_{i=1..k} \bar{I}_i(q^i_{ap}(t_{jh}))))$$

For this reason, the mediator first converts the user query in CNF and then it evaluates the CNF query by sending queries to the sources. Recall that any query that contains the logical connectives $\wedge$, $\vee$ can be converted to DNF or CNF by using one of the existing algorithms (e.g. see [1]).

Query Form		*Max. num. of calls (k sources)*
single term	t	k
disjunction	$t_1 \vee ... \vee t_n$	k
conjunction	$t_1 \wedge ... \wedge t_n$	$k * n$
CNF	$d_1 \wedge ... \wedge d_m$ where $d_j = t_{j1} \vee ... \vee t_{jn_j}$	$k * m$
DNF	$c_1 \vee ... \vee c_m$ where $c_j = t_{j1} \wedge ... \wedge t_{jn_j}$	$k * \sum_{j=1..m} n_j$

Table 1: The complexity of query evaluation at the mediator

5 Conclusion

We described a mediator model which can be used for defining user views over several Web catalogs and we analyzed the complexity of query evaluation in this model.

One can easily see that a mediator could also have a stored interpretation of its terminology. This drives to a network of *mutually articulated* sources. In the environment of the Web, this allows building a system where each user can use his own terminology in order to *index*, *browse* and *query* the pages of the Web.

References

[1] Antony Galton. *"Logic for Information Technology"*. John Wiley & Sons, 1990.

[2] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. *"Database System Implementation"*, chapter 11. Prentice Hall, 2000.

[3] Alon Y. Levy, Divesh Srivastava, and Thomas Kirk. "Data Model and Query Evaluation in Global Information Systems". *Journal of Intelligent Information Systems*, 5(2), 1995.

[4] G. Salton. *"Introduction to Modern Information Retrieval"*. McGill 1983, 1983.

[5] Yannis Tzitzikas. "Democratic Data Fusion for Information Retrieval Mediators". In *ACS/IEEE International Conference on Computer Systems and Applications*, Beirut, Lebanon, June 2001.

[6] Yannis Tzitzikas, Nicolas Spyratos, and Panos Constantopoulos. "Mediators over Ontology-based Information Sources". In *Second International Conference on Web Information Systems Engineering, WISE 2001*, Kyoto, Japan, December 2001.

[7] Yannis Tzitzikas, Nicolas Spyratos, and Panos Constantopoulos. "Query Translation for Mediators over Ontology-based Information Sources". In *Second Hellenic Conference on Artificial Intelligence, SETN-2002*, Thessaloniki, Greece, April 2002.

[8] E. Vorhees, N. Gupta, and B. Johnson-Laird. "The Collection Fusion Problem". In *Proceedings of the Third Text Retrieval Conference (TREC-3)*, Gaithersburg, MD, 1995.

[9] G. Wiederhold. "Mediators in the Architecture of Future Information Systems". *IEEE Computer*, 25:38–49, 1992.