A Theory for a Semantic-based Search Service

Leonardo Candela Donatella Castelli Pasquale Pagano
Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Consiglio Nazionale delle Ricerche
Area della Ricerca CNR di Pisa
Via G. Moruzzi, 1 - 56124 PISA - Italy
{L.Candela|D.Castelli|P.Pagano}@isti.cnr.it

Abstract

We illustrate a new approach to the formulation and processing of queries to search digital libraries built re-using and integrating pre-existing heterogeneous information sources. This approach provides a better support for unified search by enhancing the capability of the digital library to satisfy the user needs. The paper presents the theory underlying the proposed approach and describes how we exploit semantic information in the metadata formats and controlled vocabularies produced by information sources locally, and usually discarded by current digital libraries systems.

Contents

ISTITUTO DI SCIENZA E TECNOLOGIE
DELL'INFORMAZIONE "A. FAEDO"

1 Introduction

Digital libraries (DLs) are often built by re-using and integrating information sources that were originally created by single institutions to serve their own purposes. Each institution describes its documents using specific cataloguing rules. Even when a standard metadata format is adopted, the semantic interpretation of the metadata fields and the cataloguing terms used are strongly influenced by the assumptions and terminology of the application context in which the institution operates. The content acquired by a DL from a variety of different heterogeneous sources can be used to serve a multitude of users coming from institutions that have not necessarily contributed to providing this content. However, the different cataloguing rules used at the source level are completely transparent to the DL users, who formulate queries that express their information needs in terms of the metadata format and controlled vocabularies supported by the digital library search service.

This dichotomy between the information source cataloguing environment and the search environment complicates both the formulation and the processing of user queries. As in the DL framework the users neither know how documents have been originally described nor have access to the original description format, they are not always able to formulate precisely the conditions required to retrieve documents that satisfy their needs. Most DL search services attempt to minimize this problem by automatically expanding the user query with the help of stemming and query expansion algorithms.

In order to process the user queries the system must be able to map the query conditions against the metadata descriptions of the documents provided by the different information sources. The most common solution implemented today to carry out this task is to require every DL information source provider to expose the descriptions of their documents in at least a shared common format. This format is usually also the one accepted by the DL search service language. In order to fulfill this requirement, the provider establishes a mapping between its internally used format(s) and the mandatory metadata format and then he applies this mapping to all the metadata records of its resources. The DL search service thus operates in a context where the metadata descriptions and query language are homogeneous and can process the query with traditional techniques.

Current DL systems support both query formulation and processing using techniques based on syntactic manipulations, without exploiting any of the semantic information contained in the metadata schemas and controlled vocabularies. One of the reasons for this choice is the cost of such information, which is usually manually produced, another is the lack of appropriate techniques for exploiting it successfully.

This paper presents an innovative technique for supporting query formulation and processing based on the use of appropriate query mediator services. This technique exploits the information about the semantic relationships between the concepts represented by the metadata fields and the terms of the controlled vocabularies used. These relationships are usually defined locally by each information source provider as a preliminary step before translating the local description formats into the common format. By implementing this technique, the DL search service is able to offer the choice among a range of possible different interpretations for the same query. The users can thus select the one that better satisfy their needs. This technique does not require any explicit generation of the metadata records in a pre-defined shared format.

The technique described here is based on the concept of mediators over ontology-based information sources introduced by Tzitzikas et al. in [13]. We have appropriately modified and extended the theory introduced by Tzitzikas for ontologies in order to apply it to our framework composed by metadata schemas and controlled vocabularies.

We are currently experimenting the technique illustrated as part of the activities of the OpenDLib project [7]. A new service is under development that will provide a rich query language for digital libraries built by exploiting content dynamically harvested from heterogeneous information sources.

The experience we have gained so far has confirmed the validity of the technique proposed and its simplicity.

The outline of the paper is as follows. Section 2 discusses the limitations of the current search services that exploit only syntactic relations and then it presents our proposed solution. Section 3 describes a logical application framework for our approach; Section 4 and 5 motivate the proposed technique formally by introducing the underlying theory. The paper concludes by presenting our plans for the exploitation of this technique in the context of other DL services.

2 Motivation

In experimenting digital libraries built by re-using content from heterogeneous sources, we have often encountered situations in which the users could not formulate queries that express their needs and the system was not able to process them properly. These observations have motivated the work described here. In this section, we give examples to illustrate some of problems which have convinced us of the need to exploit semantic information on information sources' content presentation model in query processing.

Let us consider a simple DL in which the provider of the information sources IS_1 publishes the following metadata records:

(Subject = "`text processing`"
Subject.ACM = unspecified
URL = "`http://www.example.doc1`")
(Subject = unspecified
Subject.ACM = "`I.7.1 Document and Text Editing`"
URL = "`http://www.example.doc2`")

According to the internal rules of the institution that maintains IS_1, the authors of the documents can describe their documents by assigning either a code extracted from the ACM Computing Classification System [3] to the field *Subject.ACM* or a free term to the more generic field *Subject*. When the free term option is used, the user assigns the term that, in his/her opinion, best describes the subject of the document. The records produced are processed by the search service which extracts the information required to process the user queries.

Imagine now that, in this information framework, the user John Smith wants to retrieve exactly those documents that have been described with *Subject* equal to "text processing". The trivial solution for satisfying this information need is to formulate the following query: "*subject = text processing*". In order to process the query the search service has only to match the query condition against the information extracted from the metadata records. The answer to this query usually returned by the search services correctly includes *doc1* and excludes *doc2*.

Let us now consider another user of the same DL, Henry Stamp, who is interested in all the documents on the topic that his community of interest calls "text processing". By exploiting the functionality offered by a traditional search service, this user cannot do anything better than formulate the same query as that expressed by John Smith. However, the result expected in this case is different. It should include: *i*) the documents retrieved under the previous more strict interpretation; *ii*) the documents whose *Subject* contains values morphologically and syntactically close to the query term, e. g. "textual processing" and "documents and text processing", and *iii*) the documents whose more specific subject, i. e. *Subject.ACM*, contains values that are semantically close to the query term. Under this interpretation the system should, therefore, return not only *doc1*, but also *doc2* since it is classified under an ACM subcategory of I.7 "Documents and Text Processing".

While the majority of DL search services support an interpretation of the query based on automatically extracted morphological and syntactic relationships, e. g. stemming and query expansion,

and are therefore able to return the documents described in *i*) and *ii*) above, they are not able to take into account the semantic relationships between the concepts represented by metadata fields terms. Referring to our example, this means that the current search services do not usually return documents, like *doc2*, which are indexed under metadata fields that are specializations of those indicated in the query, i. e. *subject.ACM* in our example.

Despite this example may seem very trivial, it must be remembered that in order to satisfy the requirements of the second user, the query must find *doc2* which has been classified using a narrower subject field but a broader classification term. When manipulating complex metadata formats and sophisticated categorization schemes, as for example those used in video and audio archives, the identification of those documents that will satisfy the query is no longer a simple task.

The limitation described above becomes more incisive in a DL information space composed by multiple information sources, each of which may have described its documents using a different metadata format. In order to achieve search interoperability over a set of heterogeneous and federated information sources, DLs often require to publish their metadata descriptions in an established shared format. This format is often the Dublin Core (DC) [1]. In order to adhere to the rules of the DL, each information source provider maps its local format into DC. This mapping is done locally by people that have a clear understanding of the semantics that has been associated with the original metadata fields during the document description phase. The DL system, however, has usually no knowledge about the mappings, it only exploits their final products, i. e. the metadata records in the shared format. The query interpretation supported is therefore defined without taking into account the local descriptive interpretations. This behavior negatively influences the quality of the DL search service.

In order to exemplify this point, let us add another information source, IS_2 to our example. Let us assume that it maintains a set of audio-video documents of university courses described as in the following example:

(CourseArea = "`Computing Methodologies`"
CourseTopic = "`Text processing`"
AudioVideoSubject = "`Document Management`"
URL = "`http://www.example.doc3`")

where *AudioVideoSubject* is the subject of the specific audio-video document, i. e. the subject of a specific course lecture, *CourseTopic* is the topic of the whole course, and *CourseArea* is the framework research area addressed in the course.

Let us now suppose that the common metadata format is Dublin Core. The institution that maintains IS_1 maps both *Subject* and *Subject.ACM* to *dc:subject*, whereas the institution that maintains IS_2 maps only *AudioVideoSubject* to this DC field.

Under the above hypothesis, any query interpretation, provided by the search service that operates on the DL, is unable to return *doc3* as a result of the query presented at the beginning of this section even if the query term exactly matches the subject of the course whose the video is a part of.

The situations exemplified above, and many others observed while experimenting the use of DLs, have convinced us that the search services implemented so far are too strict. In order to have search services that could better satisfy the user needs we decided to study approaches able to exploit semantic information about the document description terminologies. As our focus was on approaches that could be implemented with reasonable costs, we decided to exploit existing semantic information, as far as possible, i. e. the semantic information provided by metadata fields that are produced locally by the information sources providers and usually discarded by current digital library systems.

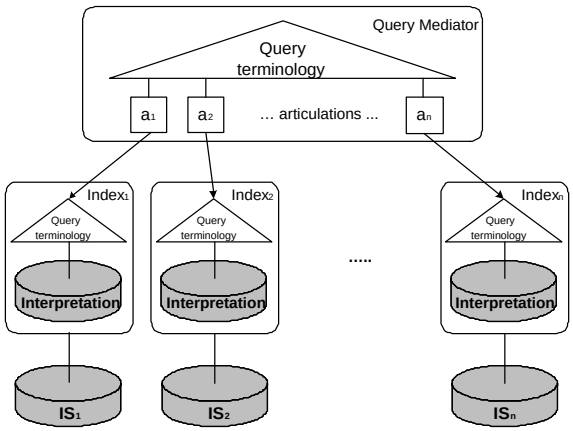

Figure 1: The architectural framework.

3 The Architectural Framework

Figure 1 shows a logical DL architectural framework for our approach. The content of the DL is given by a number of independent heterogeneous Information Sources IS_1,IS_2, $\ldots IS_n$ that disseminate metadata records of their resources in one or more formats. These records are indexed by specific services, the Index Services. For simplicity, we assume that records of different Information Sources in different formats are indexed by separate Index services[1]. An Index processes queries formulated according to the same terminology, i. e. metadata format and controlled vocabularies, used for the indexed records. We assume that this terminology and the corresponding semantic descriptions are known to the Index, i. e. the Index has access to the schemas that specify the metadata format and the controlled vocabularies associated with the metadata fields. For simplicity, we also assume that all the Index services accepts the same query structure and relational operators.

An Index service supports different interpretations of the same query condition. Each interpretation is characterized by a different level of precision given to the condition. For example, the different intended semantics given by John Smith and Henry Stamp in the query "*subject = text processing*" mentioned in the previous section are two different interpretations of this condition.

The DL user queries are actually not directly evaluated by the Index services but are first processed by the Query Mediator service. The role of this service is to hide the heterogeneity of the underlying information space. The Query Mediator serves search operations formulated in terms of the query terminology that is shown to the user[2]. It first maps the queries received by the user into queries formulated in the terminology of the underlying information sources, then it dispatches them to the Index services and, finally it merges the results received. The mapping is done by exploiting the knowledge of specific semantic relationships between the handled terminology and the local indexed terminologies. These relationships are defined by the Information Source providers and they are stored by the corresponding Index services[3]. The Query Mediator,

[1]This assumption is only given for simplicity of exposition, it does not compromise the generality of the solution.

[2]A DL can also offer search operations defined on more than one terminology. This situation can be handled by introducing a Query Mediator for each of these terminologies

[3]Protocols, like OAI-PHM, require that any information source provides at least a common DC metadata description of its items. In order to adhere to this protocol, each information source provider must first define the mapping between its local metadata format and DC, and then generate the DC records. Our approach makes less demand on the information source providers than OAI-PHM since it only requires the mapping, it does not need the explicit generation of the records in the agreed common format.

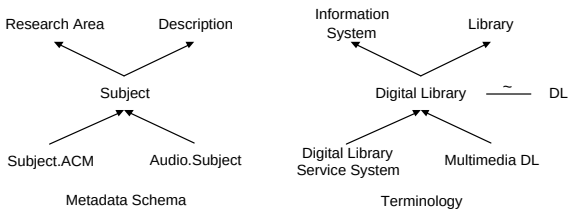

Figure 2: A metadata schema and a terminology.

similarly to the Index, can support different mapping modalities. The choice of which mapping to apply depends on the query interpretation that is required by the user.

To conclude this section, let us stress that the framework described is given to provide a context for the technique proposed. It must not be intended in any way as a design solution. The approach presented can certainly operate in many other compliant frameworks, e. g. a framework where a single service provides both the functionality of the Information Source and the functionality of the Index service.

The next two sections introduce our approach from the theoretical point of view. The solution proposed is based on the theory introduced by Tzitzikas et al. in [13]. Following the terminology introduced by Tzitzikas, we propose a theory that applies to our framework composed by metadata schemas and controlled vocabularies. In particular, we specify the different query interpretations that can be supported by the Index and Query Mediator services and how they are obtained by the existing terminology mappings.

4 The Index

Each information source uses a metadata schema to describe its own documents. This metadata schema is a pair $(\mathcal{F}, \leq_{\mathcal{F}})$ where $\mathcal{F}$ is a set of schema fields and $\leq_{\mathcal{F}}$ is a *subsumption* relation over $\mathcal{F}$[4] that models the existing specialization relationship among these fields. For example in Figure 2, `Subject.ACM` $\leq_{\mathcal{F}}$ `Subject` means that `Subject.ACM` is a more specialized property than `Subject`. Each field f of the schema is populated via an appropriate *terminology* defined as a pair $(\mathcal{V}_f, \leq_{\mathcal{V}_f})$ where $\mathcal{V}_f$ is a set of terms and $\leq_{\mathcal{V}_f}$ is a subsumption relation over $\mathcal{V}_f$ that models the existing specialization relationship among these terms. For example, in Figure 2 `Multimedia DL` $\leq_{\mathcal{V}_f}$ `DL` means that `Multimedia DL` is a more specialized term than `DL`. In certain cases the latter assumption is too *strong*. A field is often populated via free terms or free text. In these cases, the terminology can easily and automatically be obtained considering that each term is in relation only with itself or, if we are going to use stemming, we can assume that the term t is subsumed by the stemmed term t'.

Combining the metadata schema with the set of terminologies $\mathcal{V}_f$[5] that the Index uses, one for each field of the schema, we can define the *query terminology* that the Index "speaks" as a pair $(\mathcal{C}, \leq_{\mathcal{C}})$, where $\mathcal{C}$ is a set of *conditions* or pairs (f, v) such that $f \in \mathcal{F}$, $v \in \mathcal{V}_f$. The latter pair models the boolean condition "field f equals term v". For example, a valid condition for the Index in Figure 2 is (`Subject`, `Digital Library`) representing the information need "the documents whose *Subject* is *Digital Library*".

The subsumption relation over $\mathcal{C}$, $\leq_{\mathcal{C}}$, models the specializations among these conditions and is formally defined as follows:

[4]Each subsumption relation $\leq$ is a *reflexive* and *transitive* relation over the reference universe. We write $o_1 \sim o_2$ meaning that the two objects are *equivalent* w. r. t. $\leq$ if both $o_1 \leq o_2$ and $o_2 \leq o_1$.

[5]We will use $\mathcal{V}_f$ instead of $(\mathcal{V}_f, \leq_{\mathcal{V}_f})$ were no confusion arises.

Table 1: A stored interpretation

Field	Value	Documents
Subject	Digital Library	$\{d_1\}$
	DL	$\{d_2\}$
	Information System	$\{d_5\}$
	Library	$\{d_6\}$
Subject.ACM	DLSS	$\{d_3\}$
Audio.Subject	Information System	$\{d_4\}$
	Library	$\{d_4\}$
Research Area	DL	$\{d_7\}$
	DLSS	$\{d_8\}$
	Information System	$\{d_9\}$
	Library	$\{d_9\}$
Description	Digital Library	$\{d_{10}\}$
	DL	$\{d_7, d_{11}\}$
	Multimedia DL	$\{d_8\}$
	Information System	$\{d_9\}$
	Library	$\{d_9\}$

Definition 4.1 (Subsumption relation) *Let $(\mathcal{F}, \leq_{\mathcal{F}})$ be a metadata schema. Let $(\mathcal{V}_f, \leq_{\mathcal{V}_f})$ be the terminology for the field f of the schema. Given $c_1, c_2 \in \mathcal{C}$ where $c_i = (f_i, v_i)$, $f_i \in \mathcal{F}$ and $v_i \in \mathcal{V}_{f_i}$ we define $c_1 \leq_{\mathcal{C}} c_2 \iff f_1 \leq_{\mathcal{F}} f_2 \wedge v_1 = v_2$.*

Considering Figure 2 we are saying, for example, that (`subject.ACM`, `DLSS`) $\leq_{\mathcal{C}}$ (`Subject`, `DLSS`) or that (`Audio.Subject`, `Library`) $\leq_{\mathcal{C}}$ (`Research Area`, `Library`) meaning that the first condition is a specialization of the second one in each of the example cases.

As $\leq_{\mathcal{C}}$ is a subsumption relation over $\mathcal{C}$ we can define the equivalence relation w. r. t. $\leq_{\mathcal{C}}$ as $c_1 \leq_{\mathcal{C}} c_2$ and $c_2 \leq_{\mathcal{C}} c_1$ and we will write $c_1 \sim_{\mathcal{C}} c_2$. Theorem 4.1 follows from Definition 4.1:

Theorem 4.1 (Equivalence among conditions) *For each $c_1, c_2 \in \mathcal{C}$ where $c_i = (f_i, v_i)$*

$$c_1 \sim_{\mathcal{C}} c_2 \iff f_1 \sim_{\mathcal{F}} f_2 \wedge v_1 = v_2$$

Proof. For each pair of objects $o_1, o_2 \in U$, where U is a generic set of object, and subsumption relation $\leq$ over U hold that $o_1 \sim o_2 \iff o_1 \leq o_2 \wedge o_2 \leq o_1$. As a consequence $c_1 \sim_{\mathcal{C}} c_2 \iff c_1 \leq_{\mathcal{C}} c_2 \wedge c_2 \leq_{\mathcal{C}} c_1$. From Definition 4.1 follows that $c_1 \leq_{\mathcal{C}} c_2 \iff f_1 \leq_{\mathcal{F}} f_2 \wedge v_1 = v_2$ and $c_2 \leq_{\mathcal{C}} c_1 \iff f_2 \leq_{\mathcal{F}} f_1 \wedge v_2 = v_1$. Observing that $f_2 \leq_{\mathcal{F}} f_1 \wedge f_1 \leq_{\mathcal{F}} f_2 \iff f_1 \sim_{\mathcal{F}} f_2$ we conclude $c_1 \leq_{\mathcal{C}} c_2 \wedge c_2 \leq_{\mathcal{C}} c_1 \iff f_1 \sim_{\mathcal{F}} f_2 \wedge v_1 = v_2$. ∎

In the remaining text we will write $c_i \nsim_{\mathcal{C}} c_j$ meaning that c_i is not equivalent to c_j, i. e. $\neg(c_i \sim_{\mathcal{C}} c_j)$.

A query for the Index is either a simple condition or a combination of conditions using the classical connectives $\wedge, \vee, \neg$ and is formally defined as follows:

Definition 4.2 (Query) *Let $\mathcal{C}$ be a query terminology, and $c \in \mathcal{C}$. A query is any expression derived by the following BNF grammar:*

$$Q ::= c \mid Q \wedge Q \mid Q \vee Q \mid \neg Q$$

For example, a simple query can be (`subject`, `Digital Library`) $\vee$ (`Description`, `Library`).

Definition 4.3 (Interpretation) *An interpretation I of a query terminology $\mathcal{C}$ is a function $I : \mathcal{C} \to 2^{Obj}$ that associates each condition of $\mathcal{C}$ with a set of objects of the domain.*

Each Index has an *interpretation* I that is the result of the indexing phase. Table 1 presents an interpretation of the Index presented in Figure 2[6].

The interpretation that an Index uses for query evaluation must comply with the structure of the query terminology (i. e. $\leq_{\mathcal{C}}$). This requirement is expressed by introducing the notion of *model*.

[6]For simplicity, we will use the same terminology to populate all the schema fields.

Definition 4.4 (Model) *An interpretation I is a model of a query terminology if* $\forall\ c_1, c_2 \in \mathcal{C}$ *where* $c_i = (f_i, v_i)$, $c_1 \leq_{\mathcal{C}} c_2 \Rightarrow I(c_1) \subseteq I(c_2)$ *and* $f_1 = f_2 \wedge v_1 \leq_{\mathcal{V}_f} v_2 \Rightarrow I(c_1) \subseteq I(c_2)$.

For example, suppose that an Index has indexed a set of documents under the condition c_1 and another set of documents under the condition c_2 and no documents under the condition c that subsumes the previous two conditions. This interpretation is acceptable as we can "respect" the structure of $\leq_{\mathcal{C}}$ by defining the interpretation of c as the union of the set of documents indexed under c_1 and those indexed under c_2. Note that you can always generate a model form an interpretation by extending the interpretation of the conditions that do not comply with the terminology. The smallest model generated by an interpretation is the one used to answer queries.

For technical reasons we assume that every query terminology $\mathcal{C}$ contains two special queries, the *top query* $\top_{\mathcal{C}}$ and the *bottom query* $\bot_{\mathcal{C}}$. These two queries have the following properties: the *top query* subsumes every other query, i. e. $\forall c \in \mathcal{C} : c \leq_{\mathcal{C}} \top_{\mathcal{C}}$, while the *bottom query* is strictly subsumed by every other query different from $\top_{\mathcal{C}}$ and $\bot_{\mathcal{C}}$, i. e. $\forall c \in \mathcal{C} : c \neq \top_{\mathcal{C}} \wedge c \neq \bot_{\mathcal{C}} \Rightarrow \bot_{\mathcal{C}} <_{\mathcal{C}} c$. Moreover we assume that every model I of $\mathcal{C}$ satisfies the condition $I(\bot_{\mathcal{C}}) = \emptyset$. For the same reason, we assume that (a) every metadata schema $\mathcal{F}$ contains the special fields *top field* $\top_{\mathcal{F}}$ and *bottom field* $\bot_{\mathcal{F}}$, and (b) every terminology $\mathcal{V}_f$ contains the same special fields *top term* $\top_{\mathcal{V}}$ and *bottom term* $\bot_{\mathcal{V}}$.

As there may be several models of $\mathcal{C}$, we assume that each Index is able to process queries from one or more models of its interpretation. In this paper, we will use two families of models for query processing, the *sure evaluation models* and the *possible evaluation models*. In order to define these models formally we need two preliminary definitions: the first one allows us to follow the subsumption relation over the fields of the metadata schema, while the second one follows the subsumption relation over the terminologies.

Definition 4.5 (Tail and Head) *Given a condition* $c \in \mathcal{C}$, $c = (f, v)$, *we define*

$$\begin{aligned} tail(c) &= \{c' \in \mathcal{C} | c' \leq_{\mathcal{C}} c\} \\ head(c) &= \{c' \in \mathcal{C} | c \leq_{\mathcal{C}} c'\} \end{aligned}$$

Intuitively, $tail(c)$ and $head(c)$ contains c and, respectively, all the conditions that are stricter than c and wider than c according to the query terminology and, in particular, to the subsumption relations over the schema fields. For example, considering Figure 2, $tail$(`subject`, `DL`)={(`subject`, `DL`), (`subject.ACM`, `DL`), (`Audio.subject`, `DL`)} while $head$(`subject`, `DL`)={(`subject`, `DL`), (`Research Area`, `DL`), (`Description`, `DL`)}.

These definitions can be reformulated considering the Definition 4.1 as follows:

$$\begin{aligned} tail(c) &= \{c' \in \mathcal{C} | f' \leq_{\mathcal{F}} f \wedge v' = v\} \\ head(c) &= \{c' \in \mathcal{C} | f \leq_{\mathcal{F}} f' \wedge v = v'\} \end{aligned}$$

Definition 4.6 (Value models) *Given an interpretation I of $\mathcal{C}$ and a condition $c \in \mathcal{C}$, $c = (f, v)$, we define three kinds of value models for c generated by I as follows:*

$$\begin{aligned} I^{\mathcal{V}}_{\sim}(c) &= \bigcup \{I(c') | f = f' \wedge v' \sim_{\mathcal{V}_f} v\} \\ I^{\mathcal{V}}_{\leq}(c) &= \bigcup \{I(c') | f = f' \wedge v' \leq_{\mathcal{V}_f} v\} \\ I^{\mathcal{V}}_{\geq}(c) &= \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f = f' \wedge v \leq_{\mathcal{V}_f} v' \wedge v \nsim_{\mathcal{V}_f} v'\} \end{aligned}$$

The above interpretations correspond to three different ways in which the Index can evaluate a condition that involves the field f using the stored interpretations and the semantic information about the terminology. These interpretations correspond to the set of documents considered indexed under conditions involving the field f and, respectively, the value v or values equivalent to v ($I^{\mathcal{V}}_{\sim}$), the value v or values subsumed by v ($I^{\mathcal{V}}_{\leq}$), and all the values that subsume v ($I^{\mathcal{V}}_{\geq}$).

Theorem 4.2 (Relationship among value models) *If I is a model for a query terminology then $I^{\mathcal{V}}_{\sim}$, $I^{\mathcal{V}}_{\leq}$ and $I^{\mathcal{V}}_{\geq}$ are models and $I^{\mathcal{V}}_{\sim} \subseteq I^{\mathcal{V}}_{\leq} \subseteq I^{\mathcal{V}}_{\geq}$.*

PROOF. The proof that $I^{\mathcal{V}}_{\sim}$, $I^{\mathcal{V}}_{\leq}$, $I^{\mathcal{V}}_{\geq}$ are models is trivial and follows from Definition 4.6.
Let $c_1 = (f_1, v_1)$ and $c_2 = (f_2, v_2)$ and $c_1 \leq_{\mathcal{C}} c_2$, i. e. $f_1 \leq_{\mathcal{F}} f_2 \wedge v_1 = v_2$:
$I^{\mathcal{V}}_{\sim}(c_1) \stackrel{def}{=} \bigcup \{I(c') | f_1 = f' \wedge v' \sim_{\mathcal{V}_f} v_1\} \subseteq \bigcup \{I(c') | f_2 = f' \wedge v' \sim_{\mathcal{V}_f} v_2\} \stackrel{def}{=} I^{\mathcal{V}}_{\sim}(c_2)$ as $f_1 \leq_{\mathcal{F}} f_2$.
$I^{\mathcal{V}}_{\leq}(c_1) \stackrel{def}{=} \bigcup \{I(c') | f_1 = f' \wedge v' \leq_{\mathcal{V}_f} v_1\} \subseteq \bigcup \{I(c') | f_2 = f' \wedge v' \leq_{\mathcal{V}_f} v_2\} \stackrel{def}{=} I^{\mathcal{V}}_{\leq}(c_2)$ as $f_1 \leq_{\mathcal{F}} f_2$.
$I^{\mathcal{V}}_{\geq}(c_1) \stackrel{def}{=} \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f_1 = f' \wedge v_1 \leq_{\mathcal{V}_f} v' \wedge v_1 \nsim_{\mathcal{V}_f} v'\} \subseteq \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f_2 = f' \wedge v_2 \leq_{\mathcal{V}_f} v' \wedge v_2 \nsim_{\mathcal{V}_f} v'\}$
$\stackrel{def}{=} I^{\mathcal{V}}_{\geq}(c_2)$ as $f_1 \leq_{\mathcal{F}} f_2$.

Let $c_1 = (f_1, v_1)$ and $c_2 = (f_2, v_2)$ and $f_1 = f_2 \wedge v_1 \leq_{\mathcal{V}_f} v_2$:
$I^{\mathcal{V}}_{\sim}(c_1) \stackrel{def}{=} \bigcup \{I(c') | f_1 = f' \wedge v' \sim_{\mathcal{V}_f} v_1\} \subseteq \bigcup \{I(c') | f_2 = f' \wedge v' \sim_{\mathcal{V}_f} v_2\} \stackrel{def}{=} I^{\mathcal{V}}_{\sim}(c_2)$ as $v_1 \leq_{\mathcal{V}_f} v_2$.
$I^{\mathcal{V}}_{\leq}(c_1) \stackrel{def}{=} \bigcup \{I(c') | f_1 = f' \wedge v' \leq_{\mathcal{V}_f} v_1\} \subseteq \bigcup \{I(c') | f_2 = f' \wedge v' \leq_{\mathcal{V}_f} v_2\} \stackrel{def}{=} I^{\mathcal{V}}_{\leq}(c_2)$ as $v_1 \leq_{\mathcal{V}_f} v_2$.
$I^{\mathcal{V}}_{\geq}(c_1) \stackrel{def}{=} \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f_1 = f' \wedge v_1 \leq_{\mathcal{V}_f} v' \wedge v_1 \nsim_{\mathcal{V}_f} v'\} \subseteq \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f_2 = f' \wedge v_2 \leq_{\mathcal{V}_f} v' \wedge v_2 \nsim_{\mathcal{V}_f} v'\}$
$\stackrel{def}{=} I^{\mathcal{V}}_{\geq}(c_2)$ as $v_1 \leq_{\mathcal{V}_f} v_2$.

In order to prove that $I^{\mathcal{V}}_{\sim} \subseteq I^{\mathcal{V}}_{\leq}$ we can just observe that $\forall c = (f, v), \{c' | f = f' \wedge v' \sim_{\mathcal{V}_f} v\} \subseteq \{c' | f = f' \wedge v' \leq_{\mathcal{V}_f} v\}$ and that I is a model.

Let us prove that $I^{\mathcal{V}}_{\leq} \subseteq I^{\mathcal{V}}_{\geq}$. $I^{\mathcal{V}}_{\geq}(c) \stackrel{def}{=} \bigcap \{I^{\mathcal{V}}_{\leq}(c') | f = f' \wedge v \leq_{\mathcal{V}_f} v' \wedge v \nsim_{\mathcal{V}_f} v'\}$. As $I^{\mathcal{V}}_{\leq}$ is a model, it holds that $\forall c', I^{\mathcal{V}}_{\leq}(c) \subseteq I^{\mathcal{V}}_{\leq}(c')$, so we can conclude $I^{\mathcal{V}}_{\leq} \subseteq I^{\mathcal{V}}_{\geq}$. ∎

By exploiting the definitions given above, we can now define the *sure evaluation model* and the *possible evaluation model* of the stored interpretation I. These are obtained taking into account the subsumption relations among the schema fields and the subsumption relations among terminologies.

Definition 4.7 (Sure models) *Given an interpretation I of $\mathcal{C}$ we define three kinds of sure evaluation models of $\mathcal{C}$ generated by I as follows:*

$$
\begin{aligned}
I^{-}_{\sim}(c) &= \bigcup \{I^{\mathcal{V}}_{\sim}(c') | c' \in tail(c)\} \\
I^{-}_{\leq}(c) &= \bigcup \{I^{\mathcal{V}}_{\leq}(c') | c' \in tail(c)\} \\
I^{-}_{\geq}(c) &= \bigcup \{I^{\mathcal{V}}_{\geq}(c') | c' \in tail(c)\}
\end{aligned}
$$

Theorem 4.3 (Relationship among sure models) *If I is a model then $I^{-}_{\sim}$, $I^{-}_{\leq}$ and $I^{-}_{\geq}$ are models and $I^{-}_{\sim} \subseteq I^{-}_{\leq} \subseteq I^{-}_{\geq}$.*

PROOF. The proof that the sure evaluation models I^{-}_{*} are models is quite trivial. Let $c_1 = (f_1, v_1)$ and $c_2 = (f_2, v_2)$ and $c_1 \leq_{\mathcal{C}} c_2$, i. e. $f_1 \leq_{\mathcal{F}} f_2 \wedge v_1 = v_2$:
$I^{-}_{*}(c_1) \stackrel{def}{=} \bigcup \{I^{\mathcal{V}}_{*}(c') | c' \in tail(c_1)\} \subseteq \bigcup \{I^{\mathcal{V}}_{*}(c') | c' \in tail(c_2)\} \stackrel{def}{=} I^{-}_{*}(c_2)$ as $tail(c_1) \subseteq tail(c_2)$.

Let $c_1 = (f_1, v_1)$ and $c_2 = (f_2, v_2)$ and $f_1 = f_2 \wedge v_1 \leq_{\mathcal{V}_f} v_2$:
$I^{-}_{*}(c_1) \stackrel{def}{=} \bigcup \{I^{\mathcal{V}}_{*}(c') | c' \in tail(c_1)\} \subseteq \bigcup \{I^{\mathcal{V}}_{*}(c') | c' \in tail(c_2)\} \stackrel{def}{=} I^{-}_{*}(c_2)$ as $tail(c_1) \subseteq tail(c_2)$.

The proof that $I^{-}_{\sim} \subseteq I^{-}_{\leq} \subseteq I^{-}_{\geq}$ is a trivial consequence of their definitions and of Theorem 4.2 that states $I^{\mathcal{V}}_{\sim} \subseteq I^{\mathcal{V}}_{\leq} \subseteq I^{\mathcal{V}}_{\geq}$. ∎

Table 2: Interpretations of an information source index

Condition	I	$I^-_\sim$	$I^-_\leq$	$I^-_\geq$	$I^+_\sim$	$I^+_\leq$
$\bot_\mathcal{C}$	∅	∅	∅	∅	∅	∅
(Subject,Digital Library)	{1}	{1,2}	{1,2,3}	{1,2,3,4}	{1,2,7}	{1,2,3,7,8}
(Subject,DL)	{2}	{1,2}	{1,2,3}	{1,2,3,4}	{1,2,7}	{1,2,3,7,8}
(Subject,Info. Sys.)	{5}	{4,5}	{1,2,3,4,5}	{1,2,3,4,5,6}	{4,5,9}	{1,2,3,4,5,7,8,9}
(Subject,Library)	{6}	{4,6}	{1,2,3,4,6}	{1,2,3,4,5,6}	{4,6,9}	{1,2,3,4,6,7,8,9}
(Subject.ACM,DLSS)	{3}	{3}	{3}	{3}	{3}	{3}
(Audio.Subject,Info. Sys.)	{4}	{4}	{4}	{4}	{4,5}	{1,2,3,4,5}
(Audio.Subject,Library)	{4}	{4}	{4}	{4}	{4,6}	{1,2,3,4,6}
(Research Area,DL)	{7}	{1,2,7,10}	{1,2,3,7,8,10}	{1,2,3,7,8,9,10}	{1,2,7,10,11}	{1,2,3,7,8,10,11}
(Research Area,DLSS)	{8}	{3,8}	{3,8}	{3,8}	{3,8}	{3,8}
(Research Area,Info. Sys.)	{9}	{4,5,9}	{1,2,3,4,5,7,8,9,10}	{1,2,3,4,5,6,7,8,9,10}	{4,5,9}	{1,2,3,4,5,7,8,9,10,11}
(Research Area,Library)	{9}	{4,6,9}	{1,2,3,4,6,7,8,9,10}	{1,2,3,4,5,6,7,8,9,10}	{4,6,9}	{1,2,3,4,6,7,8,9,10,11}
(Research Area,Dig. Lib.)	{10}	{1,2,7,10}	{1,2,3,7,8,10}	{1,2,3,4,7,8,9,10}	{1,2,7,10,11}	{1,2,3,7,8,10,11}
(Description,Multimedia DL)	{8}	{8}	{8}	{1,2,3,7,8,11}	{8}	{8}
(Description,DL)	{7,11}	{1,2,7,11}	{1,2,3,7,8,11}	{1,2,3,4,7,8,9,11}	{1,2,7,10,11}	{1,2,3,7,8,10,11}
(Description,Info. Sys.)	{9}	{4,5,9}	{1,2,3,4,5,7,8,9,11}	{1,2,3,4,5,6,7,8,9,11}	{4,5,9}	{1,2,3,4,5,7,8,9,10,11}
(Description,Library)	{9}	{4,6,9}	{1,2,3,4,6,7,8,9,11}	{1,2,3,4,5,6,7,8,9,11}	{4,6,9}	{1,2,3,4,6,7,8,9,10,11}

Definition 4.8 (Possible models) *Given an interpretation I of $\mathcal{C}$ we define three kinds of possible evaluation models of $\mathcal{C}$ generated by I as follows:*

$$
\begin{aligned}
I^+_\sim(c) &= \bigcap \{I^-_\sim(c') | c' \in head(c) \wedge c' \nsim_\mathcal{C} c\} \\
I^+_\leq(c) &= \bigcap \left\{I^-_\leq(c') | c' \in head(c) \wedge c' \nsim_\mathcal{C} c\right\} \\
I^+_\geq(c) &= \bigcap \left\{I^-_\geq(c') | c' \in head(c) \wedge c' \nsim_\mathcal{C} c\right\}
\end{aligned}
$$

Theorem 4.4 (Relationship among possible models) *If I is a model then $I^+_\sim$, $I^+_\leq$ and $I^+_\geq$ are models and $I^+_\sim \subseteq I^+_\leq \subseteq I^+_\geq$.*

Proof. In order to proof that the possible evaluation models I^+_* are models we can just observe that $\forall c_1 = (f_1, v_1), c_2 = (f_2, v_2)$, if $c_1 \leq_\mathcal{C} c_2 \vee (f_1 = f_2 \wedge v_1 \leq_{\mathcal{V}_f} v_2)$ then $head(c_1) \subseteq head(c_2)$.

The proof that $I^+_\sim \subseteq I^+_\leq \subseteq I^+_\geq$ is a trivial consequence of their definitions and of Theorem 4.2 that states $I^\mathcal{V}_\sim \subseteq I^\mathcal{V}_\leq \subseteq I^\mathcal{V}_\geq$. ∎

Theorem 4.5 (Relationship among sure and possible models) *If I is a model then the following relationships holds between sure and possible models:*

$$I^-_\sim \subseteq I^+_\sim \qquad\qquad I^-_\leq \subseteq I^+_\leq \qquad\qquad I^-_\geq \subseteq I^+_\geq$$

Proof. All the relationship will be proved in the same way. First of all, we can observe that possible models are defined in terms of sure models, i. e. $I^+_*(c) = \bigcap \{I^-_*(c') | c' \in head(c) \wedge c' \nsim_\mathcal{C} c\}$. For each $c' \in \{c' | c' \in head(c) \wedge c' \nsim_\mathcal{C} c\}$ holds that $I^-_*(c) \subseteq I^-_*(c')$ as $c \leq_\mathcal{C} c'$, so $I^-_*(c) \subseteq I^+_*(c)$. ∎

Table 2[7] shows the interpretation models of our Index that use the terminology in Figure 2 and the stored interpretation in Table 1.

We have stated that an Index stores its interpretation I. Our approach allow us to observe that, even if the indexing phase is correct, certain documents may not have been indexed under all the conditions that could apply to them. So, given a simple query c, we may want the source

[7] In this table we have used i referring to d_i of Table 1, e. g. 1 is d_1.

to be able to answer including either all the documents that are known to be indexed under c or all the documents that are possibly indexed under c. In the first case we are considering the sure evaluation model while in the latter case we are considering the possible evaluation model.

Referring to Definition 4.2, we define the query answering as follows:

Definition 4.9 (Sure and Possible Query answering) *Let q be a query over $\mathcal{C}$ and let I be an interpretation of $\mathcal{C}$. The sure answer $I^-_{\leq}(q)$ and the possible answer $I^+_{\leq}(q)$ are defined as follows:*

$$\begin{aligned}
I^-_{\leq}(c) &= \bigcup \{ I^{\mathcal{V}}_{\leq}(c') | c' \in tail(c) \} \\
I^-_{\leq}(q \wedge q') &= I^-_{\leq}(q) \cap I^-_{\leq}(q') \\
I^-_{\leq}(q \vee q') &= I^-_{\leq}(q) \cup I^-_{\leq}(q') \\
I^-_{\leq}(\neg q) &= \overline{I^-_{\leq}(q)} \\
I^+_{\leq}(c) &= \bigcap \left\{ I^-_{\leq}(c') | c' \in head(c) \wedge c' \nsim_{\mathcal{C}} c \right\} \\
I^+_{\leq}(q \wedge q') &= I^+_{\leq}(q) \cap I^+_{\leq}(q') \\
I^+_{\leq}(q \vee q') &= I^+_{\leq}(q) \cup I^+_{\leq}(q') \\
I^+_{\leq}(\neg q) &= \overline{I^+_{\leq}(q)}
\end{aligned}$$

where we use $\overline{I}$ to indicate the set-complement operation on the set I. All the other sure and possible answers for the other models, i. e. $I^-_{\sim}$, $I^-_{\geq}$, $I^+_{\sim}$ and $I^+_{\geq}$, are defined in a similar way.

Each of the above query answering modes represent a modality of query processing. Note that the sure answer is appropriate for users that focus on *precision* while the possible answer is for users that focus on *recall*. Moreover, in both the family of sure answers and that of possible answers, we can distinguish more precision-oriented responses, i. e. $I^-_{\sim}$, versus more recall-oriented response, i. e. $I^-_{\geq}$. An Index that stores an interpretation, like the one given in Table 1, and that has access to the semantics of the metadata schema and its controlled vocabularies, can thus potentially offer a range of additional interpretations, like the ones given in Table 2, to any of its clients to express their information needs more precisely.

For example, expressing the query (`Subject`, `DL`) a user could be interested in those documents that have been described using the field `Subject`, or a more specialized one, and the term `DL` or an equivalent term, so this user is asking for $I^-_{\sim}$. Another user expressing the same query could be interested, instead, in those documents that have been described using the field `Subject`, or a more generic field, and the term `DL` or an equivalent term, so this user is asking for $I^+_{\sim}$. In the case of Table 2, the Index will return the set of documents $\{d_1, d_2\}$ to the first user and the set of documents $\{d_1, d_2, d_7\}$ to the second user. Note that while d_1 and d_2 are indexed under the condition (`subject`, `DL`) and (`subject`, `Digital Library`) respectively, the document d_7 is indexed under a pair of conditions, (`Research Area`, `DL`) and (`description`, `DL`), more general but pertinent to the one expressed by the user.

5 The Query Mediator

The previous section has described which are the potential query evaluation choices of an Index service that exploits semantic information. Having clarified this point, we can now examine the more general problem of understanding which query evaluation choices can be supported by a Query Mediator service. In what follows we will assume that such kind of mediator dispatches queries to Index services that behaves as described in the previous section.

Abstractly a Query Mediator service can be considered as an Index service that *virtually stores* all the objects of the underlying sources and supplies a query language that satisfies the needs of its users community.

However, there is an important difference between a Query Mediator and an Index: the Query Mediator does not store explicitly any interpretation of the information space. Such interpretations are maintained by the Index services. The Query Mediator only stores an *articulation* for each source, i. e. a set of relationships among the terminology of the Mediator and the terminology of the Index.

A Query Mediator is formally defined as follows:

Definition 5.1 (Query Mediator) *A Query Mediator over n Index services $I_1, \ldots, I_n$, such that $I_i = (\mathcal{C}_i, \leq_{\mathcal{C}_i})$, consists of:*

1. *a query terminology $(\mathcal{C}_M, \leq_{\mathcal{C}_M})$ and*

2. *a set of articulations a_i, one for each Index I_i; each articulation a_i is a subsumption relation over $\mathcal{C}_M \cup \mathcal{C}_i$ which contains:*

   - *a subsumption relation, $\preceq^i_{\mathcal{F}}$, over $\mathcal{F}^M \cup \mathcal{F}^i$, i. e. a set of relationships among the Mediator metadata schema and the Index metadata schema,*
   - *a set of subsumption relations, $\preceq^i_{\mathcal{V}_f}$, over $\mathcal{V}^M_f \cup \mathcal{V}^i_{f'}$, i. e. a set of relations among each field terminology of the Mediator and the corresponding ones in the Index. There exists one of such relation for each pair of (Mediator field terminology, Index field terminology).*

For simplicity, we introduce a special subsumption relation between Mediator and Index field terminologies, Π_f, that is a short-cut to indicate that every term of the first terminology is mapped into the *same* term of the second terminology. In such case we impose that $\mathcal{V}^i_{f'} = \mathcal{V}^M_f$, i. e. the terminology of the Index is the *same* as that of the Mediator, and $\preceq^i_{\mathcal{F}}$ is defined such that for each $v \in \mathcal{V}^i_{f'}$ and $v' \in \mathcal{V}^M_f$, $v \sim^i_{\mathcal{V}_f} v'$ if and only if $v = v'$, i. e. the term on the Mediator is *equivalent* to the same term of the Index w. r. t. the articulation.

The Mediator query terminology is defined similarly to the Index terminology, i. e. $\mathcal{C}_M$ is a set of pairs (f, v) such that $f \in \mathcal{F}^M$, $v \in \mathcal{V}^M_f$, and $\leq_{\mathcal{C}_M}$ is a subsumption relation over $\mathcal{C}_M$. Moreover each $\mathcal{V}^M_f$ is a terminology, i. e. a pair $(\mathcal{V}^M_f, \leq_{\mathcal{V}_f})$ where $\leq_{\mathcal{V}_f}$ is a subsumption relation over $\mathcal{V}^M_f$.

Figure 3 shows an example of a Query Mediator that operates over two Indexes. This mediator uses the DC metadata schema and the ACM Computing Classification System as controlled vocabulary for the field `subject`[8]. The Index services in Figure 3 are $Index_1$, that has been introduced in the previous section, and $Index_2$, an Index service that uses the LOM metadata schema [2] and free terms to populate the fields shown in the figure.

The query interpretations that are supported by the Query Mediator are defined in terms of both the interpretations stored by the Index services and the existing articulations. In order to identify these interpretations we proceed in the following way:

1. we define a query c^i for I_i as a translation of each $c \in \mathcal{C}_M$ obtained using a_i, $i = 1, \ldots, n$;

2. we evaluate c^i at I_i, $i = 1, \ldots, n$;

3. finally, we define $I(c)$ as the union of the answers to queries c^i returned by the Index services.

[8]For brevity the example shows only a partial view of the Query Mediator.

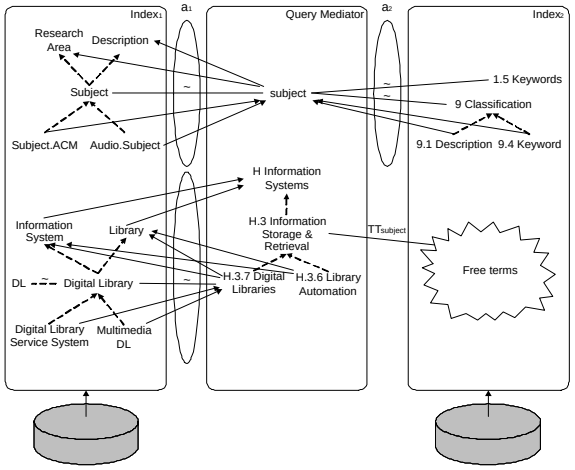

Figure 3: A Query Mediator over two Indexes.

Several possible translations can be applied. In the following we show the possible ways to perform the translation. In order to define these translations formally we need some preliminary definitions, i. e. the concepts of *head*, *body* and *tail* w. r. t. an articulation and the concept of approximations over values.

Definition 5.2 (Tail, Head and Body w. r. t. an Articulation) *Given a condition $c \in \mathcal{C}_M$ where $c = (f, v)$ and an articulation a_i we define*

$$\begin{aligned} tail^i(c) &= \left\{(f', v') | f' \preceq^i_{\mathcal{F}} f \wedge v = v' \wedge f' \in \mathcal{F}^i\right\} \\ body^i(c) &= \left\{(f', v') | f' \sim^i_{\mathcal{F}} f \wedge v = v' \wedge f' \in \mathcal{F}^i\right\} \\ head^i(c) &= \left\{(f', v') | f \preceq^i_{\mathcal{F}} f' \wedge v = v' \wedge f' \in \mathcal{F}^i\right\} \end{aligned}$$

Intuitively, $tail^i(c)$, $body^i(c)$, and $head^i(c)$ contain, respectively, all the conditions in the Index query terminology that are narrower than c, equivalent to c and broader than c w. r. t. the articulation. The conditions above involve Index metadata fields that are, respectively, subsumed by, equivalent to or that subsume the field used on the Query Mediator w. r. t. the semantic mapping among the Mediator metadata schema and Index metadata schema.

Definition 5.3 (Value Approximations w. r. t. an Articulation) *Given a condition $c \in \mathcal{C}_M$ where $c = (f, v)$ and an articulation a_i. Let*

$$\begin{aligned} tilde^i(c) &= \left\{c' \in \mathcal{C}_i | f' = f \wedge v \sim^i_{\mathcal{V}_f} v'\right\} \\ lower^i(c) &= \left\{c' \in \mathcal{C}_i | f' = f \wedge v \preceq^i_{\mathcal{V}_f} v'\right\} \\ upper^i(c) &= \left\{c' \in \mathcal{C}_i | f' = f \wedge v' \preceq^i_{\mathcal{V}_f} v \wedge v' \nsim_{\mathcal{V}_f} v\right\} \end{aligned}$$

we define three kinds of approximations over values:

$$\begin{aligned} c^i_{\sim} &= \bigvee tilde^i(c) \\ c^i_{\leq} &= \bigvee lower^i(c) \\ c^i_{\geq} &= \bigwedge \left\{c'^i_{\leq} | c' \in upper^i(c)\right\} \end{aligned}$$

The above approximations correspond to three different ways in which a condition on a field can be reformulated into a set of conditions that take into account the semantic relationships among the Query Mediator field terminology and the Index field terminology.

Now we are able to define formally the *precise approximations*, the *lower approximations* and the *upper approximations* of a conditions $c_i \in \mathcal{C}_M$. Roughly speaking the *precise approximations* of c_i w. r. t. a_j is the disjunction of all the conditions in $\mathcal{C}_j$ that are *equivalent* to c_i in a_j, $c^i_\sim$; the second one, $c^i_\leq$, is the disjunction of all the conditions in $\mathcal{C}_j$ that c_i subsume in a_j; while the last one, $c^i_\geq$, is the conjunction of all the conditions that subsume c_i in a_j. These families of approximation are formally defined as follows:

Definition 5.4 (Precise Approximations w. r. t. an Articulation) *Let $M = (\mathcal{C}_M, \leq_{\mathcal{C}_M}, a_1, \ldots, a_n)$ be a mediator over sources $S_1, \ldots, S_n$. Given a condition $c = (f, v) \in \mathcal{C}_M$ we define three kinds of precise approximations of c w. r. t. a_i as:*

$$\begin{aligned} c^i_{p_\sim} &= \bigvee \{c^i_\sim | c' \in body^i(c)\} \\ c^i_{p_\leq} &= \bigvee \{c^i_\leq | c' \in body^i(c)\} \\ c^i_{p_\geq} &= \bigvee \{c^i_\geq | c' \in body^i(c)\} \end{aligned}$$

Definition 5.5 (Lower Approximations w. r. t. an Articulation) *Let $M = (\mathcal{C}_M, \leq_{\mathcal{C}_M}, a_1, \ldots, a_n)$ be a mediator over sources $S_1, \ldots, S_n$. Given a condition $c = (f, v) \in \mathcal{C}_M$ we define three kinds of lower approximations of c w. r. t. a_i as:*

$$\begin{aligned} c^i_{l_\sim} &= \bigvee \{c^i_\sim | c' \in tail^i(c)\} \\ c^i_{l_\leq} &= \bigvee \{c^i_\leq | c' \in tail^i(c)\} \\ c^i_{l_\geq} &= \bigvee \{c^i_\geq | c' \in tail^i(c)\} \end{aligned}$$

Definition 5.6 (Upper Approximations w. r. t. an Articulation) *Let $M = (\mathcal{C}_M, \leq_{\mathcal{C}_M}, a_1, \ldots, a_n)$ be a mediator over sources $S_1, \ldots, S_n$. Given a condition $c = (f, v) \in \mathcal{C}_M$ we define three kinds of upper approximations of c w. r. t. a_i as:*

$$c^i_{u_\sim} = \begin{cases} \bigwedge \{c'^i_{l_\sim} | c' \in head^i(c) \wedge c' \nsim c\} & \text{if } head^i_\leq(c) \setminus c \neq \emptyset \\ c^i_{l_\sim} & \text{otherwise} \end{cases}$$

The other upper approximations $c^i_{u_\leq}$ and $c^i_{u_\geq}$ are defined in a similar way changing accordingly the kind of lower approximation to use.

Here are reported some examples of approximations for the mediator shown in Figure 3[9]:

$$\begin{aligned} (\texttt{DC.subject}, \texttt{H.3.7})^1_{p_\sim} = \; & (\texttt{subject}, \texttt{Digital Library}) \vee \\ & (\texttt{subject}, \texttt{DL}) \\ (\texttt{DC.subject}, \texttt{H.3.7})^2_{l_\sim} = \; & (\texttt{1.5}, \texttt{H.3.7}) \vee (\texttt{9}, \texttt{H.3.7}) \vee \\ & (\texttt{9.1}, \texttt{H.3.7}) \vee (\texttt{9.2}, \texttt{H.3.7}) \end{aligned}$$

The approximations are just queries to the information source S_i and can have sure (three kinds) or possible (three kinds) answer as shown in Section 4. For this reason we can define

[9]For brevity we have used the code of the fields or the code of a terminology terms instead of the whole value as no confusion arise. Clearly the abbreviated terms must be replaced by the whole term.

at least 54 possible interpretations I for the mediator[10]. We denote these interpretations using this formalism $I_{a,b}$ where a is the kind of approximation that mediator use and b is the kind of answer from the source, e. g. $I_{u_{\leq},+_{\leq}}$ means that the mediator uses the upper approximation with $\leq$, while the sources reply following the possible model $I^{+}_{\leq}$. Note that these approximations are defined as the set union over the source interpretations w. r. t. the mediator approximation, e. g. $I_{u_{\leq},+_{\leq}}(c) = \bigcup_{i=1}^{n} I^{+}_{\leq}(c^{i}_{u_{\leq}})$.

As the mediator can be considered an information source it can give either one of the three sure answer or one of the three possible answer for each of the above interpretations, i. e. we can have 324 possible modes under which the mediator can operate. We denote these operation modes using this formalism $I^{c}_{a,b}$ where a is the kind of approximation that mediator use, b is the kind of answer from the source and c is the answer that the mediator produce, e. g. $I^{+_{\leq}}_{u_{\leq},+_{\leq}}$ means that the mediator use the upper approximation with $\leq$ and reply following the possible model with $\leq$ while the sources reply following the possible model $I^{+}_{\leq}$.

6 Conclusion

This paper has presented a new approach to query formulation and processing that exploits semantic links between different terminologies used to describe documents. The paper has also introduced the theory underlying the proposed approach and has illustrated how this theory can be exploited in real application contexts.

Much work, especially in the area of information retrieval, has been done in order to better satisfy the search requirements of the user. The technique proposed is not intended as an alternative to the current well consolidated search processing techniques, but as complementary. By exploiting the semantic information given directly by the information source providers, who know how the documents have been classified, and by asking the users to select the query interpretation that best meets their search requirements, user satisfaction is enhanced.

One of our objectives in defining this approach was to come out with a low-cost solution. This solution proposed requires information source providers to specify only the mapping between their local document description terminology, i. e. metadata schema fields and controlled vocabularies used, and the terminology of the Query Mediator service. Unlike other approaches does not require the generation of descriptive records in a shared format.

The complexity of the Index and Query Mediator services that exploit the technique presented partly depends on the number of query processing options that are supported. Certainly, some of them are intuitively useful, while others are less useful and probably have only a theoretic value. As discussed in the paper, we have selected a number of these options for experimentation. We expect to have concrete results on this experimentation very soon.

The fuller exploitation of semantic information in query processing is not only useful to enhance the quality of the search service itself, but also to improve the quality any other service that queries the DL in order to implement its functionality. For example, it can be useful for a service that provides a virtual view of the DL collections or for a recommender service. One of our next steps will be certainly to study the impact that the proposed approach may have on the quality of these other DL services. We are firmly convinced that the exploitation of semantic information can have a very positive effect on these "user-centered" services.

References

[1] Dublin Core Metadata Initiative. `http://dublincore.org`.

[10] For simplicity we assume that all the Indexes respond using the same kind of answer.

[2] IEEE Standard for Learning Object Metadata. `http://ltsc.ieee.org/wg12/-par1484-12-1.html`.

[3] The ACM Computing Classification System. `http://www.acm.org/class/1998`.

[4] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[5] James P. Callan, Fabio Crestani, and Mark Sanderson, editors. *Distributed Multimedia Information Retrieval, SIGIR 2003 Workshop on Distributed Information Retrieval, Toronto, Canada, August 1, 2003, Revised Selected and Invited Papers*, volume 2924 of *Lecture Notes in Computer Science*. Springer, 2004.

[6] Donatella Castelli, Carlo Meghini, and Pasquale Pagano. Foundations of a Multidimensional Query Language ffor Digital Libraries. In *Proceedings of the* 6^{th} *European Conference on Digital Libraries (ECDL2002)*, pages 251–265. Springer-Verlag, 2002.

[7] Donatella Castelli and Pasquale Pagano. OpenDLib: A Digital Library Service System. In *Proceedings of the* 6^{th} *European Conference on Digital Libraries (ECDL2002)*, pages 292–308. Springer-Verlag, 2002.

[8] Chen-Chuan K. Chang and Héctor García-Molina. Mind your vocabulary: Query mapping across heterogeneous information sources. pages 335–346, 1999.

[9] Sudarshan Chawathe, Hector Garcia-Molina, Joachim Hammer, Kelly Ireland, Yannis Papakonstantinou, Jeffrey D. Ullman, and Jennifer Widom. The TSIMMIS project: Integration of heterogeneous information sources. In 16^{th} *Meeting of the Information Processing Society of Japan*, pages 7–18, Tokyo, Japan, 1994.

[10] Carl Lagoze and Herbert Van de Sompel. The open archives initiative: building a low-barrier interoperability framework. In *Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 54–62. ACM Press, 2001.

[11] Erhard Rahm and Philip A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases*, 10(4):334–350, 2001.

[12] Yannis Tzitzikas. *Collaborative Ontology-based Information Indexing and Retrieval*. PhD thesis, Department of Computer Science, University of Crete, September 2002.

[13] Yannis Tzitzikas, Panos Constantopoulos, and Nicolas Spyratos. Mediators over ontology-based information sources. In *WISE (1)*, pages 31–40, 2001.

[14] Gio Wiederhold. Mediators in the architecture of future information systems. *Computer*, 25(3):38–49, 1992.