Semantics-based Information Brokering

Vipul Kashyap
Department of Computer Science
Rutgers University
New Brunswick, NJ 08903
kashyap@pepper.rutgers.edu

Amit Sheth
Distributed Information Systems Lab
Department of Computer Science, University of Georgia
415 GSRC, Athens, GA 30602-7404
amit@cs.uga.edu

Abstract

The rapid advances in computer and communication technologies, and their merger, is leading to a global information market place. It will consist of federations of very large number of information systems that will cooperate to varying extents to support the users' information needs. We discuss an approach to *information brokering* in the above environment. We discuss two of it's tasks: *information resource discovery*, which identifies relevant information sources for a given query, and *query processing*, which involves the generation of appropriate mapping from relevant but structurally heterogeneous objects. Query processing consists of *information focusing* and *information correlation*.

Our approach is based on: *semantic proximity*, which represents semantic similarities based on the *context* of comparison, and *schema correspondences* which are used to represent structural mappings and are associated with the context. The *context* of comparison of the two objects is the primary vehicle to represent the semantics for determining semantic proximity. Specifically, we use a partial context representation to capture the semantics in terms of the assumptions in the intended *use* of the objects and the intended *meaning* of the user query. Information focusing is supported by subsequent context comparison. The same mechanism can be used to support information resource discovery. Context comparison leads to changes in schema correspondences that are used to support information correlation.

1 Introduction

With concerted efforts to develop a National Information Infrastructure (NII) and the advent of the Information Super Highway, global information systems founded on the cooperation between various information systems cannot be far behind. We believe that the integration of the various systems, or the interoperability among the information systems, will have to be at a higher *semantic level* in a scalable manner. This should not however compromise the identity and independence of each of the components. We believe that representation of context-bound semantics will enable us to realize and manage digital libraries and develop "middleware software" with *information brokers* (with such better known cousins as "mediators" [23], "knowbots" [9] and "software agents" [1]).

We plan to represent the contents of the information sources and the query of the user by constructing contexts which capture their semantics. The contexts are constructed from the domain ontologies which may be known or available to the user. The mechanisms of comparing contexts to discover the information sources and resource objects relevant to the query and generating the mappings to retrieve information are illustrated in this paper. The problem of knowing the contents and structure of each of the huge number of information sources is reduced to the smaller problem of knowing (or making available) the domain ontologies relevant to a query.

We propose that mappings between domains of objects be made with respect to a context. In Section 3.3.1, we use the *definition context* of an object to make explicit the assumptions implicit about objects in an information source. This may be viewed as a form of **value addition**, i.e. an attempt to *organize information* to facilitate interoperability. In Section 3.3.2, we use the *query context* to explicate the semantics of a user query. The comparison of the definition and the query contexts provide an *arbitration mechanism* (Section 4.1) for information focusing and discovery (Section 5). The resulting context is used for *information focusing/search*.

This paper is organized as follows. In Section 2, we analyze the information brokering tasks. In Section 3 we illustrate the representation of semantic and structural similarities and their relationship to context. We also propose a partial context representation. In Section 4 we illustrate our approach to information focusing based on context comparison and information correlation on the basis of schema correspondences and their relationship to context. In Section 5 we discuss how context comparison can be used for information resource discovery. Issues of ontology involved in context representation are discussed. Section 6 discusses the conclusions and enumerates some emerging challenges.

2 An anatomy of Information Brokering Tasks

In the presence of millions of information sources, it is the information brokers which facilitate meeting the information needs of the users. Two important information brokering tasks are as follows:

Appeared in the Proceedings of the Third International Conference on Information and Knowledge Management (CIKM), Gaithersburg, MD, November, 1994.

- **Information Resource Discovery**: The first critical task is to identify the information sources with the relevant information based on the meta-information or on direct approaches involving the information itself.
- **Query Processing**: This involves getting the answer to the query posed by a user and consists of the following sub-tasks:
  - **Information Focusing**: When the relevant information sources are identified, the next critical task, which we term information focusing, is to identify a subset of the relevant information available at the relevant information sources that can be used to answer the user query.
  - **Information Correlation**: Relevant information identified by information focusing may be from semantically different but related domains (represented in different forms). These can also be correlated with each other (e.g., by developing mappings between schematically heterogeneous data) and presented in a manner which would enhance the decision-making capabilities of the user. This is the information correlation problem.

3 Similarities : Semantic and Structural

In this section, we discuss the concept of *semantic proximity* to characterize semantic similarities between objects. The *context* of comparison of the objects is the pivotal component of the semantic proximity. We discuss the concept of *schema correspondences* to represent the structural similarities between objects and associate them with the context.

We distinguish between the *real world*, and the *model world* which is a representation of the real world. Wood [24] defines semantics to be "the scientific study of the relations between signs and symbols and what they denote or mean." Another perspective of semantics is "the different ways signs and symbols are used". It is not possible to *completely* define what an object denotes or means [17] or enumerate the ways it may be used in the model world. We take both, the *meaning* and *use* perspectives to explain the need for identification and representation of context.

3.1 Semantic Proximity

Given two objects O_1 and O_2, the *semantic proximity* between them is defined by the 4-tuple

semPro(O_1, O_2)
=<Context, Abstraction, (D_1, D_2), (S_1, S_2)>, where

- A context of an object is the primary vehicle to capture the semantics of the object. Thus, the respective contexts of the objects, and to a lesser extent the abstraction used to map the domains of the objects, help to capture the semantic aspect of the relationship between the two objects.
- Various alternatives for identifying/representing a context may be metadata [20], database, relationship, federated schema, external schema,export schema (refer to the schema architecture of [19]), collection of object domains, and hard-coded. (See [11] for a detailed discussion.)
- D_i is the domain of object O_i (i = 1,2).

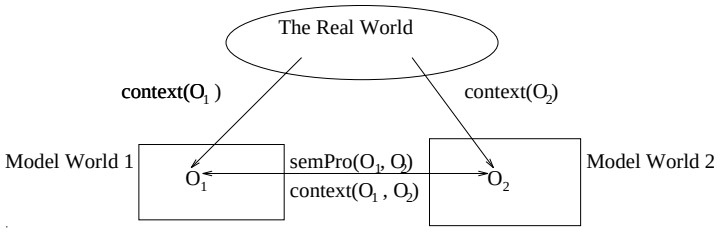

Figure 1: Semantic Proximity between two Objects

- S_i is the state of object O_i (i = 1,2).

3.2 Perspectives on Semantics : Meaning, Use and Necessity of representing Context

It has been discussed in Sheth and Gala/Kashyap [17][18] and Fankhauser et al. [7], that the semantics of an object cannot be adequately captured using it's structural representations. In [17], it is shown how we may be able to define a mapping between the value-domains of two attributes which are not equivalent semantically. We propose that the mappings between the domains of the two objects be defined with respect to a context [18]. Whether the attributes are equivalent or not would then be determined by the context in which they are being compared.

In linguistics [24], the interest in semantics has focused on characterizing the different meanings of the same sentence. A knowledge engineer [4], on the other hand, is usually interested in a (semantic) description that represents partial knowledge about an entity and accommodates multiple descriptions of the entity from different viewpoints. In a multidatabase environment, the contents of a database can be meaningful in a given context and the meaning/significance can be looked at in terms of an interpretation in the context [21]. One commonality observed in the above diverse fields of research is that the same sentence/entity can have different meanings/descriptions. We propose that it is the **context** which determines the applicable meaning/descriptor/ assumption. The query context defined in Section 3.3.2 reflects this perspective.

One view suggested in AI is that one memory schema refers to another only through the use of a description which is dependent on the context of the original reference [3]. In the area of linguistics and cognitive psychology, experiments have borne out a strong relationship between semantic similarity and contextual similarity [14]. This has led to the belief that semantic similarity is a function of the contexts in which an object is used and that the contextual representation of an object is the knowledge of how that object is used. The contextual representation is visualized as an abstract cognitive structure that accumulates the attributes common to all the contexts in which an object is used [14]. We propose that **context** can be used as a tool for characterizing the intended usage of the objects. The definition context defined in Section 3.3.1 reflects this perspective.

3.3 A partial representation of Context

Attempts have been made to represent context in diverse areas of research, such as linguistics, text-retrieval and multidatabases. In the area of multidatabases an attempt has been made to represent context based on "semantic values" [16]. In linguistics [5], criteria for selection of "contextual coordinates" to represent context are suggested. We consider

these approaches as a variant of the basic approach where context is represented as a collection of meta-attributes. The concepts of thematic roles [22] and code words [15] in the area of text-retrieval systems may be considered analogous to meta-attributes. A (partial) representation is :

Context = $\{(c_i, v_i) \mid c_i$ is a contextual coordinate, v_i is the value of $c_i\}$

We give below an example that involves a query that can be processed using two databases found to be relevant as a result of information resource discovery. We will use this example throughout the paper to explain our approach. Information resource discovery, while not explicitly demonstrated, can be supported by applying a strategy similar to information focusing and is discussed briefly later.

Example : Let us consider two databases that model information from different domains:

- **UnivDB** : A typical University Database consisting of the following entities :
  - EMPLOYEE(SS#, Name, SalaryType, Dept, Affiliation, ...).
  - PUBLICATION(Id, Title, Journal, ...).
  - HAS-PUBLICATION(SS#, Id).
- **GovtDB** : A typical Government Database consisting of the following entities :
  - WORKER(SS#, Name, Salary, ...).
  - POSITION(Id, Title, Dept, Type, ...).
  - HOLDS-POSITION(SS#, Id).

Let us consider a user query Q :

Get all the representatives and senators who have published papers on the socio-political implications of the Abortion issue.

With the help of the above example we demonstrate the following in Sections 3.3.1 and 3.3.2 :
A1. Context representation reflecting the usage of an object.
A2. Context representation reflecting the meaning of an object.
A3. Context representation reflecting the semantics by a combination of domains and by establishing dependencies between the domains.
A4. Recursive context representation, i.e., a value of a contextual coordinate might have a context associated with it at arbitrary levels of nesting.

3.3.1 The Definition Context

When a database is designed, the implicit assumptions in the mind of the designer are reflected in the design of the database. With each object O defined, we associate the definition context $\mathbf{C_{def}(O)}$ which makes explicit the assumptions behind the definition of that entity O. Since these assumptions are about the intended use of the object O, C_{def}(O) reflects the *use* perspective of semantics. This approach is similar to the *assuming(p,c)* predicate in [13] where one can view the context as a collection of assumptions.
Consider the entities defined above and the assumptions behind their definitions :

- Assumptions in the definition of the entity EMPLOYEE [A1][1] :
  - An employee either works for a department or is doing a dissertation in the department.
  - The employee works either as a teacher, a researcher or a non-teaching staff.
  - The different possibilities of non-teaching staff are not relevant.
  - The employee could be paid a salary or an honorarium.

  Note that the person defining the context can refer to pre-existing ontologies in the federation for choosing the contextual coordinates (e.g. affiliation, etc.) and their values (e.g. teaching, research, etc.). Please refer to Section 5 for a detailed discussion.

  C_{def}(EMPLOYEE) = ((employer Deptypes[2]$\cup${restypes}), (affiliation {teacher, research, non-teaching}), (reimbursement {salary, honorarium}))
- Assumptions in the definition of the entity PUBLICATION [A1]:
  - Various publications at a university are in the research areas corresponding to the departments established in the university.

  C_{def}(PUBLICATION) = ((researchArea Deptypes))
- Assumptions in the definition of the relationship HAS-PUBLICATION [A1]:
  - All published articles have been written by various employees of the University who are affiliated with it as researchers. (Faculty members are considered researchers.)
  - There is a semantic dependency between the domains of EMPLOYEE and PUBLICATION [A3].
  - The value of the contextual coordinate author (EMPLOYEE) has a context associated with it [A4].

  C_{def}(HAS-PUBLICATION) = ((author EMPLOYEE (affiliation {research})), (article PUBLICATION))
- Assumptions in the definition of the entity WORKER [A1]:
  - A worker can work for either of the Judicial, Executive or Legislative branches of the Government.
  - A worker can be paid either a salary or an honorarium.

  C_{def}(WORKER) = ((employer {judiciary, executive, legislative}), (reimbursement {salary, honorarium}))

[1] The tag in a square bracket, e.g., [A1], indicates that this discussion illustrates the feature A1 given in a preceding box, e.g., the box on the left.

[2] The domain of Deptypes contains all departments of the university. We assume that such domain information is available as metadata to the mechanisms discussed in the report.

- Assumptions in the definition of the entity POSITION [A1]:
  - A position is either an elected or nominated position.

  C_{def}(POSITION) = ((appt {elected, nominated}))

- Assumptions in the definition of the relationship HOLDS-POSITION [A1]:
  - All positions are held by the workers.
  - There is a semantic dependency between the domains of WORKER and POSITION [A3].

  C_{def}(HOLDS-POSITION)
  = ((designee WORKER), (appt POSITION))

3.3.2 The Query Context

Here, we try to make explicit the meaning of the query posed by a user. With a query Q we associate the query context $\mathbf{C}_Q$ which makes explicit the (partial) semantics of Q and thus reflects the *meaning* perspective of semantics.

Consider the example query Q on page 3 [A2,A4].
C_Q = ((author self), (designee self[3]),
(employer {legislative, restypes}), (post ((appt elected))),
(article ((title "*abortion*"))),
(researchArea {socialSciences, politics}))

The user gets the values from the domain of a database object. We assume for the purpose of this paper that the domains are incorporated into a pre-existing ontology (see Section 5).

3.4 Schema Correspondences and Context

We propose a uniform formalism to represent the mappings which are generated to represent the structural similarities between objects having schematic differences and some semantic similarity (see [10] for a detailed discussion.

Given two objects O_1 and O_2, the *schema correspondence* between them can be represented as

schCor(O_1, O_2)
= < O_1, attr(O_1), O_2, attr(O_2), Ψ >, where

- O_1 and O_2 are objects in the model world. They are representations or intensional definitions in the model world (e.g., an object class definition in object-oriented models).

- The objects enumerated above may model information at different levels of representation. If an object O_i models information at the entity level, then attr(O_i) denotes the representation of the attributes of the entity modeled by O_i. If O_i models objects at the attribute level, then attr(O_i) is an empty set.

- Ψ is a mapping (first order or second order) expressing the correspondences between objects, their attributes and the values of the objects/attributes.

[3] "self" refers to the answer expected from the query Q. This is analogous to the arguments of the select clause in an SQL statement.

Each information system exports the definition contexts of the objects it manages. The exported context partially explicates the semantics of the object. In our approach we consider structure to be a part of semantics. This is achieved by the association between the exported definition contexts and the objects defined in the database. We use schema correspondences to express these associations. We assume that for each object O in the database, there exists a virtual object O_F, associated with C_{def}(O). We assume that the attributes of O_F are the contextual coordinates of the definition context, i.e. coord(C_{def}(O)). The modified schema correspondence can then be used to relate one or more contextual coordinates in the definition context with the database object(s) and can be defined as

schCor(O_F, O)
= < O_F, coord(C_{def}(O)), O, attr(O), Ψ >

Consider the object EMPLOYEE as defined in the example on page 3. Let the object corresponding to the definition context C_{def}(EMPLOYEE) be $EMPLOYEE_F$.

The schema correspondence associated with the context C_{def}(EMPLOYEE) is
schCor($EMPLOYEE_F$, EMPLOYEE)
= <$EMPLOYEE_F$, {employer, affiliation, reimbursement},
EMPLOYEE, {Dept, Affiliation, SalaryType},Ψ >
where Ψ is a mapping given by:

```
select SS#, Name, reimbursement = SalaryType,
       employer = Dept, affiliation = Affiliation
from EMPLOYEE
where Dept in [{restypes} U Deptypes]
and Affiliation in {teacher, research, non-teaching}
and SalaryType in {salary, honorarium}
```

4 Semantics-based Query Processing

In this section we illustrate with the help of an example, how query processing is accomplished. The mechanism of context comparison is used to support information focusing. Information correlation is achieved by appropriately manipulating the schema correspondences.

4.1 Information Focusing using context comparison

The *definition context* of an object (Section 3.3.1) may be viewed as a form of **value addition**, i.e. an attempt to structure information about the information sources. However, this additional sophistication is achieved at the cost of extra effort in providing context information. For complex queries like the one in the example on page 3, this sophistication and extra work is necessary and worthwhile because of the following reasons.

- The value addition introduced facilitates information focusing and discovery.

- The contexts are constructed from the domain ontologies which may be known or available to the user. Mechanisms for discovering information relevant to the query and for generating mappings for retrieving the information use these contexts. The problem of knowing the contents and structure of each of the huge number of resource objects is now reduced to the smaller problem of knowing (or making available) the domain ontologies relevant to a query.

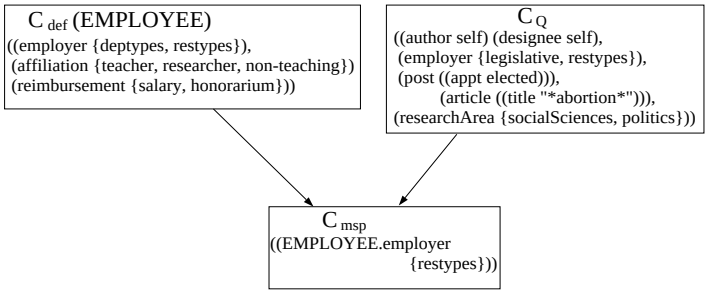

Figure 2: Context Comparison : Focusing on the relevant employees

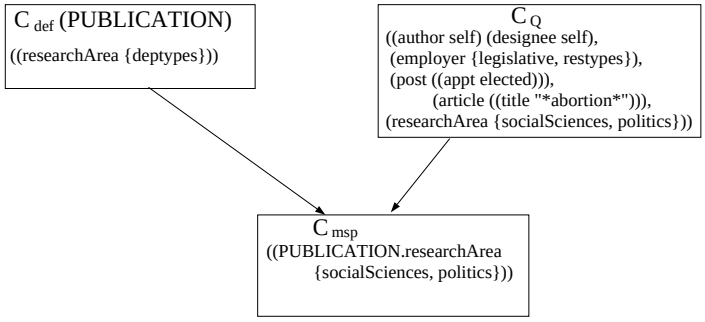

Figure 3: Context Comparison : Focusing on the relevant research areas

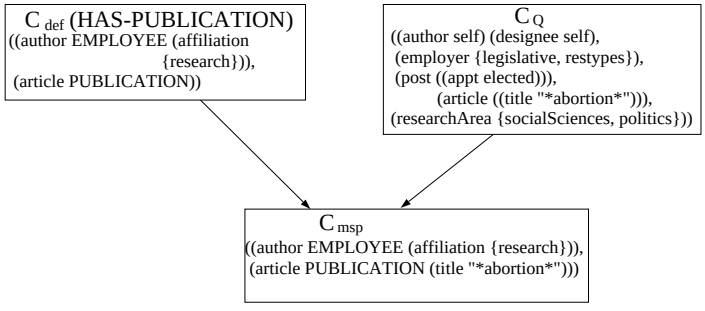

Figure 4: Context Comparison : Focusing on the relevant publications

We assume here that the information sources relevant to the user query have been identified (see Section 5). However, each information source may have thousands of resource objects. We need to identify the subset of objects relevant to the user query. This is called *information focusing*. Continuing our example that started on page 3 we illustrate the process of context comparison and illustrate how it supports information focusing. The resulting most specific context computed at the information source is called C_{msp}.

In the rest of this section we consider the query and its context discussed in Section 3.3.2 and demonstrate the following :
B1. The comparison of the query context with the definition contexts of the resource objects.
B2. Identification of the relevant resource objects and the resulting focusing of information.
B3. Use of contextual coordinates to focus on information at deeper levels of nesting or to associate a context with the value of a coordinate.

In Figure 2, we compare the definition context of the entity EMPLOYEE with the query context [B1]. This helps us to identify an employee who is doing dissertation as relevant to the user query [B2].

In Figure 3, we compare the definition context of the entity PUBLICATION with the query context [B1]. This helps us identify the publications relating to the areas of Social Sciences and Politics as relevant to the user query [B2].

In Figure 4, we compare the definition context of the relationship HAS-PUBLICATION with the query context [B1]. This helps us identify the publications having the substring "abortion" in their title as relevant to the user query [B3].

Thus the most specific context computed at the UnivDB site is given by :
C_{msp}(Q, UnivDB)
= ((author EMPLOYEE ((affiliation {research}))),
(article PUBLICATION ((title "*abortion*"))),
(EMPLOYEE.employer {restypes}),
(PUBLICATION.researchArea {socialSciences, politics}))

Using a procedure similar to the one described above, the comparison of C_Q with C_{def}(WORKER) and C_{def}(HOLDS-POSITION) at the GovtDB side leads to the following :
C_{msp}(Q, GovtDB)
= ((WORKER.employer {legislative}), (designee WORKER),
(post POSITION (appt {elected})))

4.2 Information Correlation using schema correspondences

In Section 4.1 we demonstrated how C_{msp} is computed at each site. The values of the contextual coordinates of C_{msp} as a result of this process are likely to be different from those of the original definition contexts. New schema correspondences expressing the associations between the new values and the data items can be computed by the *conditioning* of the modified schema correspondences (Section 3.4) by the new values. The final answer is then computed by the *composition* of these conditioned schema correspondences.

In the rest of this section we demonstrate how information mapping can be achieved by :
C1. Determining the conditioned schema correspondences with respect to C_{msp}.
C2. Composition of the schema correspondences within and across databases.

4.2.1 Conditioning of the Schema correspondences

We illustrate the process of *conditioning* the schema correspondences at the database site *wrt* to the C_{msp} at that site and determine the new schema correspondences. At each database, we post query objects which will contain the information relevant to the query at that site. We then determine the schema correspondences between them and the objects in the database.

Let $Q_{i,j}$ be a temporary query object j at site i. The schema correspondences at the UnivDB site are as follows :

- Schema correspondence induced by the contextual coordinates **author** and **EMPLOYEE.employer** :
  <$Q_{1,1}$, {author}, EMPLOYEE, {SS#, Name}, $M_{1,1}$ >
  where $M_{1,1}$ is a mapping given by :

```
select author = <SS#, Name>
from EMPLOYEE
where employer = "restypes"
and affiliation = "research"
```

- Schema correspondence induced by the contextual coordinates **article** and **PUBLICATION.researchArea** :
  <$Q_{1,2}$, {article}, PUBLICATION, {Id, Title, Journal}, $M_{1,2}$ >
  where $M_{1,2}$ is a mapping given by :

```
select article = Id
from PUBLICATION
where Journal of {"socialSciences", "politics"}
and substring("abortion", Title)
```

The schema correspondences at the GovtDB site are :

- The schema correspondence induced by the contextual coordinates **WORKER.employer** and **designee** :
  <$Q_{2,1}$, {designee}, WORKER, {SS#, Name}, $M_{2,1}$ >
  where $M_{2,1}$ is a mapping given by :

```
select designee = <SS#, Name>
from WORKER
where employer = "legislative"
```

- The schema correspondence induced by the contextual coordinate **post** :
  <$Q_{2,2}$, {post}, POSITION, {Id}, $M_{2,2}$ >
  where $M_{2,2}$ is a mapping given by :

```
select post = Id
from POSITION
where appt = "elected"
```

4.2.2 Composition of the schema correspondences

In this section, we illustrate how information can be combined using the composition of schema correspondences.

Intra-database composition

In some cases, schema correspondences at the same database site are combined because of the dependencies introduced by a definition context of an object at the database. This is called *intra-database composition*.

- The dependency between the contextual coordinates **author** and **article** introduced by C_{def}(HAS-PUBLICATION) at UnivDB leads to the composition of $M_{1,1}$ and $M_{1,2}$ defined in Section 4.2.1 :
  <Q_1 , {author}, {$Q_{1,1}$, $Q_{1,2}$, HAS-PUBLICATION}, {author, article, SS#, Id}, M_1 >
  where M_1 is a mapping given by :
  select author = $Q_{1,1}$.author
  from $Q_{1,1}$, $Q_{1,2}$, HAS-PUBLICATION
  where <$Q_{1,1}$.author.SS#, article>
  in (select * from HAS-PUBLICATION)
  $M_1 = M_{1,1} \circ M_{1,2}$, where $\circ$ denotes the composition of the mappings.

- The dependency between the contextual coordinates **designee** and **post** introduced by C_{def}(HOLDS-POSITION) at GovtDB leads to the composition of $M_{2,1}$ and $M_{2,2}$ defined in Section 4.2.1 :
  <Q_2 , {designee}, {$Q_{2,1}$, $Q_{2,2}$, HOLDS-POSITION}, {designee, post, SS#, Id}, M_2 >
  where M_2 is a mapping given by :
  select designee = $Q_{2,1}$.designee
  from $Q_{2,1}$, $Q_{2,2}$, HOLDS-POSITION
  where <$Q_{2,1}$.designee.SS#, post> in
  (select * from HOLDS-POSITION)
  $M_2 = M_{2,1} \circ M_{2,2}$

Inter-database composition

In some cases the schema correspondences at different database sites are combined because two (or more) contextual coordinates having the value *self* in the query context are associated with objects in different databases. This is called *inter-database composition*.

There is a dependency between the contextual coordinates **designee** and **author** as they have the value **self** in C_Q. This leads to the composition of M_1 and M_2 defined in the previous section:
<Q, {name}, {Q_1, Q_2}, {designee, author}, M>
where M is a mapping given by :
select name
from Q_1, Q_2
where SS# in (select UnivDB.author.SS# from Q_1)
and in (select GovtDB.designee.SS# from Q_2)
$M = M_1 \circ M_2$

5 Information Resource Discovery based on context comparison

The likelihood of an information system containing the information relevant to a user query can be gauged by comparing the semantics of the user query and the design assumptions made by an information system. In Section 4.1, we identified the resource objects relevant to a query by comparing the definition contexts of the objects to the query context. However, we need to identify the relevant information sources before we can proceed to identify the relevant resource objects at that information source. Thus, we need

to solve the *information resource discovery* problem before the *information focusing* problem.

We plan to adapt the mechanism of context comparison (Section 4.1) for the information resource discovery problem. However, the definition context of an information source may be different from the definition context of a resource object in the following ways.

- The definition context of the information source might be a union of the definition contexts of all the objects in the information source.
- The definition context may contain information about the resource objects at a higher level of abstraction.
  - The ontological objects in the definition context of the information source might be abstractions (aggregations/generalizations) of the ontological objects in the definition contexts of the resource objects.
  - The ontological objects in the query context might be abstractions (aggregations/generalizations) of the ontological objects in the definition context of the information source or vice versa.
- The definition context of the information source might contain information about the information source as a whole (viz. guidelines, purpose, formats, protocols). This type of meta-information is typically not captured by the definition contexts of the resource objects.
- The definition context of the information source might contain parts of the definition contexts of the resource objects incorporated in an appropriate manner.

We accomplish information resource discovery by comparing the definition context and the query context to compute the resulting context $\mathbf{C_{res}(Query, InformationSource)}$ at each site (see Figure 5). If C_{res}(Query, InformationSource) is empty, then that information source does not contain the relevant information (or at least we are not able to find any relevant information) for the query. Otherwise the C_{res}(Query, InformationSource) identifies the information source as being relevant to the query. This approach may be considered as one way of achieving *transcendence*. In [13], transcendence is defined as the ability to move a proposition from one context to another which relaxes or changes some assumptions of the old context. We can view context comparison as a means of transcending from the context defined for the information source to the query context.

Issues of ontology in context representation

An ontology may be defined as the specification of a representational vocabulary for a shared domain of discourse which may include definitions of classes, relations, functions and other objects [8]. In constructing the contexts as illustrated in Sections 3.3.1 and 3.3.2, the choice of the contextual coordinates and the values assigned to them is very important. There should be *ontological commitments*, i.e. agreements about the ontological objects used between the users and the information system designers.

In our case this corresponds to an agreement on the terms used for the contextual coordinates and their values by a user in formulating the query context C_Q and a designer for formulating the definition context C_{def}(O). As

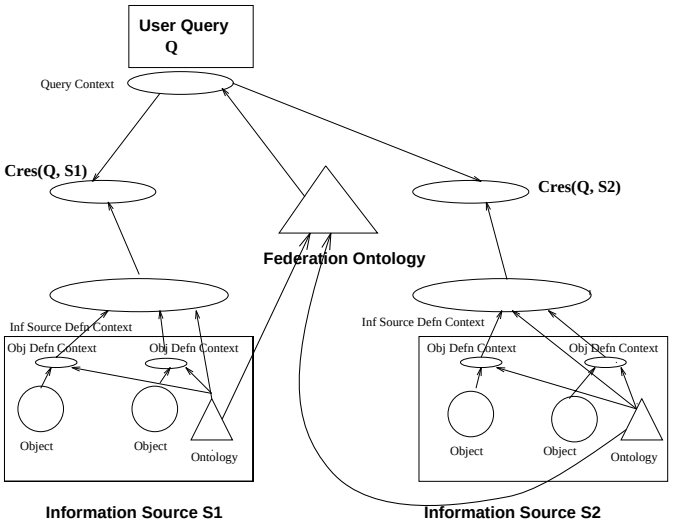

Figure 5: Information resource discovery using context comparison

proposed in Section 3.3 the values of the contextual coordinates could be from a pre-existing ontology of types and objects from the database. In Section 3.3.2 we used the values "socialSciences" and "politics" which belong to the domain of the type Deptypes in the UnivDB database. We assume that the domains of the types defined in the database are incorporated in the ontology associated with that information source.

Various approaches have been taken for building and using ontologies for a federation of information sources. A notable example of a global ontology is Cyc [12], where a set of *articulation axioms* is used to map the entities of an information source to concepts in the Cyc ontology [6]. Another approach has been to exploit the semantics of a single problem domain (viz. transportation planning in [2]). We propose a re-use of various existing classifications viz. ISBN classification for publications, botanical classification for plants, etc.

However in designing the definition contexts of the information sources and the query context, issues of combination of the various ontologies and their presentation arise. This must be done in a manner to enable the user to construct the query context with ease. A critical issue in combining the various ontologies is determining the overlap between them. One approach is to define the "intersection" and "mutual exclusion" points between the various ontologies. Identifying "intersection" would be similar to the identification of the various concepts which are synonyms of each other. Identifying "mutual exclusion" would be similar to the identification of concepts which are homonyms of each other. This process would require the input and coordination of the various domain experts. Also important are issues of presenting the "intersections" and "mutual exclusions" to the user.

6 Conclusions and Future Work

We advocate a semantics based approach for information brokering. The conceptual bases of our approach are *semantic proximity*, which represents semantic similarities based on the *context* of comparison, and *schema correspondences* which are used to capture the structural similarities. The schema correspondences are associated with the context as a

component of the semantic proximity. Semantics is captured from two perspectives: *meaning* and *use*. Using a partial representation, we use the context to capture the meaning of a user query as the *query context*, intended use of a resource object as *object definition context* and the purpose and intended use of an information source as *information source definition context*. Issues of ontology that arise in context representation are also discussed.

The task of information brokering is defined to consist of two arbitration tasks – *information resource discovery*, to identify the information sources that might have data relevant to a query, and *query processing*, to retrieve the specific data items from relevant information sources to satisfy the query. Query processing involves *information focusing* to identify specific data items of interest within the known relevant information sources and *information correlation*, to correlate semantically related but schematically heterogeneous data. We illustrate how information focusing can be performed by comparing the query context and the object definition contexts at an information source. Context comparison leads to changes in the associated schema correspondences. Information correlation is performed by computing these changes and combining the schema correspondences in an appropriate manner. We propose using the same mechanism as information focusing for information resource discovery, but with context information of the information sources (rather than that of the data items in an information source).

Several challenges need to be addressed related to the semantics-based approach we have proposed. Notable among them are: capturing the semantics of the information sources in a context-bound manner; the relationship between semantics, context and uncertainty; the semantics of context comparison and manipulation; and issues of language and ontology for context representation.

Acknowledgments

We thank Dr. Saul Amarel for suggesting the interesting query on page 3 that led us to explore several of the issues discussed in this paper. Dr. Gio Wiederhold pointed us to the need for determining the overlap between the ontologies.

References

[1] Software 'Agents' will make life easy. In Fortune, January 1994.

[2] Y. Arens, C. Chee, C. Hsu, and C. Knoblock. Retrieving and Integrating Data from Multiple Information Sources. *International Journal of Intelligent and Cooperative Information Systems*, 2(2), June 1993.

[3] D. Bobrow and D. Norman. Some principles of Memory Schemata. In *Representation and Understanding*. New York : Academic Press, 1975.

[4] D. Bobrow and T. Winograd. An overview of KRL, a Knowledge Representation Language. In *Readings in Knowledge Representation*. Morgan Kaufmann, 1985.

[5] G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar: An Introduction to Semantics*, chapter 6. MIT Press Cambridge MA, 1990.

[6] C. Collet, M. Huhns, and W. Shen. Resource Integration using a Large Knowledge Base in Carnot. *IEEE Computer*, December 1991.

[7] P. Fankhauser, M. Kracker, and E. Neuhold. Semantic vs. Structural resemblance of Classes. *SIGMOD Record, special issue on Semantic Issues in Multidatabases*, A. Sheth, ed., 20(4), December 1991.

[8] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition, An International Journal of Knowledge Acquisition for Knowledge-Based Systems*, 5(2), June 1993.

[9] R. Kahn and V. Cerf. An open architecture for a Digital Library System and a plan for it's development. Technical report, Corporation for National Research Initiatives, March 1988.

[10] V. Kashyap and A. Sheth. Schema Correspondences between Objects with Semantic Proximity. Technical Report DCS-TR-301, Department of Computer Science, Rutgers University, October 1993.

[11] V. Kashyap and A. Sheth. Semantic similarities between Objects in Multiple Databases. In A. Elmagarmid, M. Rusinkiewicz, and A. Sheth, editors, *Heterogeneous Distributed Databases*, chapter 3. Morgan Kaufmann, 1995. (in preparation).

[12] D. Lenat and R. V. Guha. *Building Large Knowledge Based Systems : Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company Inc, 1990.

[13] J. McCarthy. Notes on formalizing Context. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1993.

[14] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. *Languauge and Cognitive processes*, 1991.

[15] S. H. Myaeng and M. Li. Building Term Clusters by acquiring Lexical Semantics from a Corpus. In *Proceedings of the CIKM*, 1992.

[16] E. Sciore, M. Siegel, and A. Rosenthal. Context Interchange using Meta-Attributes. In *Proceedings of the CIKM*, 1992.

[17] A. Sheth and S. Gala. Attribute relationships : An impediment in automating Schema Integration. In *Proceedings of the NSF Workshop on Heterogeneous Databases*, December 1989.

[18] A. Sheth and V. Kashyap. So Far (Schematically), yet So Near (Semantically). *Invited paper in Proceedings of the IFIP TC2/WG2.6 Conference on Semantics of Interoperable Database Systems, DS-5*, November 1992.

[19] A. Sheth and J. Larson. Federated Database Systems for managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys*, 22(3), September 1990.

[20] M. Siegel and S. Madnick. A Metadata Approach to resolving Semantic Conflicts. In *Proceedings of the 17th VLDB Conference*, September 1991.

[21] J. P. Thompson. *Data with Semantics : Data Models and Data Management*. Van Nostrand Reinhold - New York, 1989.

[22] D. A. Voss and J. R. Driscoll. Text Retrieval using a Comprehensive Lexicon. In *Proceedings of the CIKM*, 1992.

[23] G. Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3), March 1992.

[24] J. Wood. What's in a link ? In *Readings in Knowledge Representation*. Morgan Kaufmann, 1985.