Technical Report DCS-TR-301, Department of Computer Science, Rutgers University, October 1993.

Semantics-based Information Brokering: A step towards realizing the Infocosm

Vipul Kashyap[1] and Amit Sheth[2]

[1]Department of Computer Science, Rutgers University,
New Brunswick, NJ 08903

[2]Bellcore, 444 Hoes Lane, Piscataway, NJ 08854-4182

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 propose an architecture which may facilitate meeting these needs. It consists of three main components: *information providers*, *information brokers* and *information consumers*. We also propose an approach to information brokering. 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*.

While the access-based search, and syntactic and hierarchical information organization has been adequate in the past, information brokering in presence of huge digital libraries or millions of information sources will likely require semantics and information-content based search and structuring of information. 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 context to capture the semantics in terms of the *meaning* and/or the *use* of an object. Using a partial context representation, we capture 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 In the not too distant a future ...

Merging of computers and communications, or the impending merger of TV, PC and the cable box, has created a lot of excitement. Gigabit public and private networks with huge leaps in wired (fiber) communication and high frequency low power wireless transmission capabilities, combined with fast switching (e.g., ATM) will move all types of information you wish to deliver in an eye-wink, so they say. Media servers exploiting parallelism will be able to quickly search the information they have and pump it on the network.

These developments have helped create an environment where one can have access to "any information anywhere you want in (m)any form(s)". We envision the emergence of an **infocosm** in this environment (a twist on George Gilder's "telecosm [tel93]), which we define to mean a society with ubiquitous access/exchange of tradeable information in all electronic forms. In this future society, two major classes of applications and services will likely emerge, viz. *mass market applications* and *information content sensitive applications*. We review below, at a high level two critical aspects of these applications: the organization of information and the search for information.

Mass market applications

This segment will include mass appeal applications in interactive media [int93] (video on demand, interactive TV/games, edutainment and infotainment applications, multimedia publications, and communications applications such as telephony, video conferencing, and e-mail) and simpler information marketplace [SS93] applications (home shopping and banking, enhanced on-line services). Most of the Business to Residence services identified in the Appendix belong to this category. Main approaches to the key aspect of the organization and the search for information used to support this class of applications are as follows.

- Organization of information: A popular approach used in resource discovery on the internet is to keep inverted indices on stored document contents. In another approach, entities in the directory are represented by entries in a global, hierarchical name space according to the structural relationships between the classes of information they represent. An extension of the above approach is to maintain multiple hierarchies and use hypertext links to integrate them. Some approaches maintain active catalogs to constrain the search space.

- Search for the information: In a popular approach, resource discovery is based on text search of the contents of the documents. In other approaches, a search is conducted for relevant information based on the attributes/structure of the organization in a single hierarchy. Where there are multiple hierarchies, search is implemented by following the hypertext links based on any of the hierarchies. Some approaches perform search in the list of data repositories returned by the active catalogs.

Information content sensitive applications

While the above class of applications will likely cover a segment of market enabled by the emerging technologies, the rest of market will be covered by applications that are (at least moderately) information content sensitive. A significant portion of the Business to Business services identified in the Appendix belongs to this category. While hardware technologies for both classes of (mass market and information content sensitive) applications may soon become reality, we believe there are significant software challenges that need to be addressed to enable this class of applications in future. However, the approaches and the related techniques for organization and search of information identified in the previous section may break down when we try to apply them to support the information content sensitive applications. Some of the challenges that arise are:

- **Inadequacy of structural representations**: The organization and search for information enumerated above is based on the structure and/or textual contents of the objects. The ability to represent the structure of an object is, however, inadequate in capturing the information content of the object. This is important when we are dealing with information content sensitive applications.
- **Inadequacy of hierarchical organizations**: Most of the approaches use a hierarchical organization for the meta-information. In such cases, it is efficient to search only according to the criteria used in the structuring of the hierarchy. What is required here is identification of the relevant criteria and development of an appropriate *focusing mechanism* based on these criteria.
- **Knowledge of the contents of the information sources**: In order to obtain the answer for a particular query, the user is required to have a general familiarity with the contents and structure of the information sources. This however, is a tall order in the Infocosm as it might require familiarity with a huge number of information sources.

Digital libraries, a component in the strategy to realize the National Information Infrastructure, can be an important early service in the infocosm. One vision of the digital library involves, among other things, a unified access to digital information managed by a large number of autonomous and heterogeneous information systems. Besides repositories of digitized information currently found in conventional libraries, one may also be able to access personal databases and repositories of large collections of scientific data.

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 without compromising the identity and independence of each of the components. This will require the enhancement of current information search and organization techniques by a *semantics-based* organization and brokering of information. 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" [Wie92], "knowbots" [KC88] and "software agents" [gen94]).

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. We believe that the reuse of existing ontologies – possibly ad hoc, certainly domain-specific, and if possible those already used by organizations and businesses – is the best approach for making ontologies available for the construction of contexts. A brief discussion on the ontologies used and the language for representation is presented in Section 5.3.

The mechanisms of comparing contexts to discover the information sources relevant to the query and generating the mappings to retrieve information are discussed in the later sections of this report. 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. Our approach to semantics-based organization and search for information can be summarized as follows.

- **Organization of Information**: To capture the semantic similarity between two objects, we propose that the definition of mappings between their domains be made with respect to a context. For example, we can use the *definition context* of an object to explicate the assumptions implicit in the mind of the designer about the objects in an information source (Section 3.3.1). This may be viewed as a form of value addition, i.e. an attempt to structure information about the information sources to facilitate information resource discovery.

- **Search for the Information**: The context of the search/query can be used as the focusing/arbitration mechanism. We use the *query context* to explicate the semantics of the query posed by a user looking for information (Section 3.3.2). Information focusing is modeled as the identification of the relevant source objects as the result of the comparison of the definition and the query contexts (Section 4.1). The resulting context helps to focus onto the relevant information. Information correlation can be achieved by combining the information associated with the resulting context (Section 4.2). The mechanism of context comparison is used to support information resource discovery (Section 5.2).

The rest of the report is organized as follows. In Section 2 we review an architecture in which information content sensitive services may be realized. We also analyze the needs of the information consumers and the information brokering tasks involved. In Section 3 we illustrate the representation of semantic and structural similarities and their relation to context. We also propose a partial context representation to capture the design assumptions in an information system and the semantics of a user query. 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 illustrate our approach to information discovery on the basis of context comparison. Initial thoughts on issues of language and ontology involved in the representation of context are presented. Section 6 discusses the conclusions and enumerates some of the emerging challenges.

In the appendix, we have summarized two market studies, which predict the growth of the information industry. Their forecast suggests that around 50% of the revenue is generated by information content sensitive applications. This enhances the significance of the problem we are trying to tackle.

2 Notes on Architecture, Environment, and Some Issues

2.1 Architecture

Let's quickly review an architecture in which information content sensitive services may be realized (see Figure 1). The architecture consists of three main components:

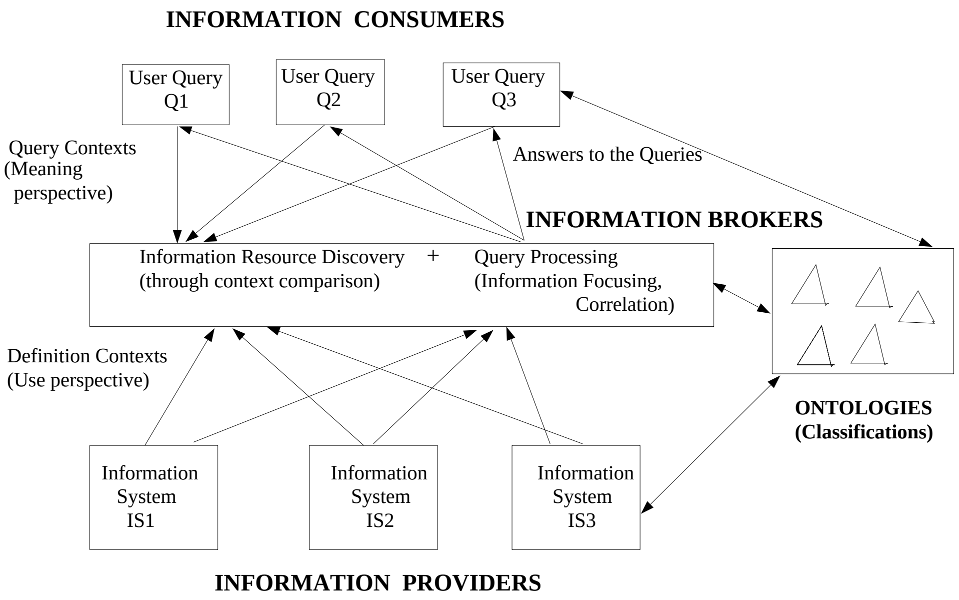
INFORMATION CONSUMERS
User Query Q1
User Query Q2
User Query Q3
Query Contexts (Meaning perspective)
Answers to the Queries
INFORMATION BROKERS
Information Resource Discovery (through context comparison)
+
Query Processing (Information Focusing, Correlation)
Definition Contexts (Use perspective)
ONTOLOGIES (Classifications)
Information System IS1
Information System IS2
Information System IS3
INFORMATION PROVIDERS

Figure 1: A high level architecture of our approach

1. **Information Providers**: This consists of the millions of **information sources** which will eventually be available to the users on various networks (private and public) by the various information agencies (viz. Dow Jones, Reuters etc.) which we call information providers. The information sources could contain information in a structured form (viz. databases, knowledge bases) or in a semi-structured form (viz. newsgroups, email, multi-media documents) or in an unstructured form (viz. unix files).

2. **Information Brokers**: The user in the infocosm would be deluged by the information available from the millions of information providers. This requires *arbitration* between the user and the information providers in the form of explicating the semantics and (re)interpretation of the information/query. This task is performed by the information brokers.

3. **Information Consumers**: We envisage millions of *consumers* utilizing various public and commercial networks and the services and the information offered by them (current online systems and the internet already support millions of non-scientific users). These consumers might be individual users on workstations or application programs running on many machines at the same time.

The distinction between the information itself and those who control and manipulate it is important from a *business* perspective. In that case, issues like the fees to be charged for the information provided and intellectual property rights become important. In this report, however, we shall limit ourselves to a *technical* perspective and ignoring the difference between the information providers and the information sources.

2.2 An informal classification of the Information Brokering approaches

We present an informal classification of the information brokering approaches that might be taken in the infocosm, based on the approach taken for the search for the information required by a user query.

1. **User directed approach**: In this group, the user is presented with menu-driven and browsing interfaces to the information system from which he can choose options which he believes might satisfy his information needs. A typical example is the support for information queries supported by many current on-line services.

2. **Syntactic keyword/attribute-based approach**: This group of queries are those which can be answered by a keyword based text search of the contents of documents. They can also be answered by an attribute based search for the information. These strategies have been primarily used in text processing and information retrieval.

3. **Descriptive semantics-based approach**: This group of queries requires making explicit (at least partially) the semantics of the query and the design assumptions of the information providers. The semantics of the query may be explicated in a semi-automatic manner with input from the user. Similarly, the design assumptions are explicated with input from the designers of the information systems. Examples of queries that can better be addressed by this approach are:

    - 'Get all the representatives and senators who have published papers on the socio-political implications of the Abortion issue.' where there may be multiple databases that store partially relevant information, and these databases are not known in advance. We illustrate the processing of this query in the latter sections of this report.
    - 'Display the Lipper's multi-media report on three highest yielding high quality intermediate corporate bonds.' Here we assume that meta data defining information such as "high quality" exists.

4. **Cognition based approach**: This group of queries are the most difficult types of queries. They require a deeper understanding of the semantics of the contents of the information systems. Here we look upon cognition as the basis of the (possibly sub-symbolic) semantics, where cognitive criteria such as perceptive and visual cues in an image, may not be amenable to a symbolic description. An example query is the following which may be issued to a multi-media database.

    - 'Get me all images from the database which contain the view of a sunset'.

We believe that the first two approaches will likely meet the requirements of most early mass market applications. We believe that the last two approaches can be of significant value in developing information content sensitive applications. In the descriptive semantics-based approach which we further explore in this report, the cost involved in partially capturing the semantics can be offset very easily as this would lead to the automation of processing a class of queries important for information content sensitive applications. We do not pursue the fourth approach as more basic research is need and our knowledge in this area is limited.

2.3 An anatomy of Information Brokering Tasks

In the evolving infocosm mentioned above, it is the information brokers which facilitate information trading between the information providers and the information consumers. Two important information brokering tasks are as follows:

- **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 an information consumer 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 that 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. The term object in this paper refers to an object in a model world (i.e., a representation or intensional definition in the model world, e.g., an object class definition in object-oriented models) as opposed to an entity or a concept in the real world. These objects may model information at an *attribute level* or an *entity level*.

Wood [Woo85] defines semantics to be "the scientific study of the relations between signs and symbols and what they denote or mean." It is not possible to *completely* define what an object denotes or means in the model world [SG89]. Another perspective of semantics is *the different ways signs and symbols are used*. We believe that, in general it is not possible to completely enumerate the different ways an object might be used in the model world. We take both, the meaning and use perspectives in Section 3.2 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
where D_i is domain of O_i and S_i is state of O_i.

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.

Context of the two Objects

Each object has its own context. The term context in semPro refers to the context in which a particular semantic similarity holds. This context may be related to or different from the contexts in which the objects were defined. It is possible for two objects to be semantically closer in one context than in another context. Some of the alternatives for representing a context are as follows:

- In [SM91], the context is identified as the semantics associated with an application's view of existing data and is called the **application semantic view**. They propose a rule-based representation to associate metadata with a given attribute, and use this rule based representation to define the application's semantic view of the data.

- Just as a context may be associated with an application, it can also be associated with a **database** or a group of databases (e.g., the object is defined in the context of DB1).

- When many entities participate in a relationship, the entities can be thought of as belonging to the same context, which in this case is identified as the **relationship** in which the entities participate.

- In a federated database approach, we can use a **federated schema** [SL90] to identify a context to which two objects may belong to.

- From the five-level schema architecture for a federated database system [SL90], a context can be specified in terms of an **export schema** (a context that is closer to a database) or an **external schema** (a context that is closer to an application). We can also build a context hierarchy, by considering the contexts associated with the external schemas to be subcontexts of the context associated with the appropriate federated schema.

- At a very elementary level, a context can be thought of as a **named collection** of the domains of the Objects.

- Sometimes a context can be "hard-coded" into the definition of an object. For example, when we have the two entities EMPLOYEE and TELECOMM-EMPLOYEE,

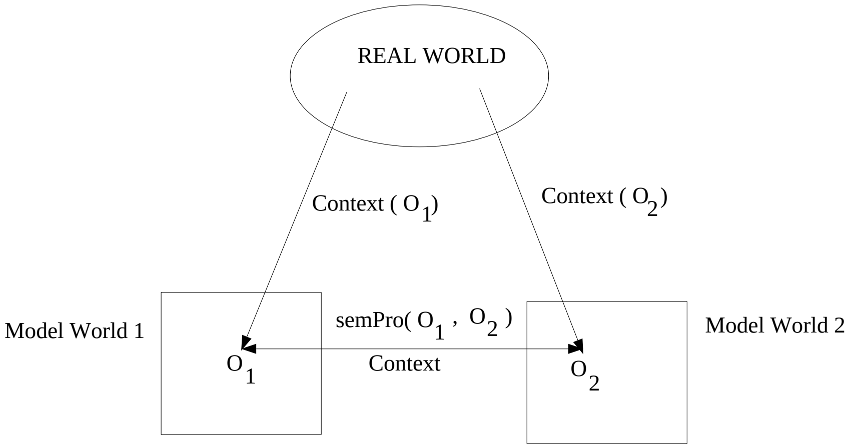

Figure 2: Semantic Proximity between two Objects

the TELECOMMUNICATIONS context is "hard-coded" in the second entity. We are interested in representing and reasoning about context as an explicit concept.

As discussed later in this report, our use of contexts is based on domain specific ontologies.

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

It has been discussed in Sheth and Gala/Kashyap [SG89][SK92][SG93] and Fankhauser et al. [FKN91], that the semantics of an object cannot be adequately captured using it's structural representations. For example, to capture the semantic similarities between two objects, the relation between the domains of the objects (in the values the objects can take) and the similarity in their structures are not enough to guarantee semantic similarity. Consider two attributes *person-name* and *department-name*. We may be able to define a mapping between the value-domains of these two attributes, but we know that they are not semantically equivalent. There should be then some way to denote their lack of semantic equivalence. We propose that this can be done by **defining the mappings between the domains of the two objects with respect to a context.** Whether the attributes are equivalent or not would then be determined by the context in which they are being compared.

In linguistics [Woo85], the interest in semantics has focused on characterizing the different meanings of the same sentence. A knowledge engineer [BW85], 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 [Tho89]. We observe a commonality in diverse fields of research when it comes to representing the meaning of an object which is that the same sentence/entity can have different meanings/descriptions. We propose that in either case, 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 [BN75]. In the area of linguistics and cognitive psychology, experiments have borne out a strong relationship between semantic similarity and contextual similarity [MC91]. 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 [MC91]. 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

While it may not be possible to precisely define the context of an object, it may be useful to simply name it at a specific level of information modeling architecture (e.g., external/export schema or federated schema). A partial context specification can be used by humans to decide whether the context for modeling of two objects is the same or different, and whether the comparison of semantic similarity of objects is valid in all known contexts or specific ones. In this section we propose such a partial representation of context. We believe that in most cases, a partial representation should suffice to judge the semantic similarity between any two objects. This representation should help explicate the semantic similarities between the two objects being compared.

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" [SSR92]. In linguistics [CMG90], 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 [VD92] and code words [ML92] in the area of text-retrieval systems may be considered analogous to meta-attributes.

Based on the above discussion,we represent context as :

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-PUBLISHED(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. In the following examples, we use the representation of context defined above to make those assumptions explicit. This approach is similar to the *assuming(p,c)* predicate in [McC93] where one can view the context as a collection of assumptions. With each object O defined, we associate the definition context $\mathbf{C}_{def}\mathbf{(O)}$ which makes explicit the assumptions behind the definition of that entity O. Since in this case we are trying to make explicit the assumptions made about the intended use of the object O, $\mathrm{C}_{def}(\mathrm{O})$ reflects the "use" perspective of semantics.

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.

[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 page 11.

- 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 5.3 for a detailed discussion.

$$\begin{aligned} \mathrm{C}_{def}(\mathrm{EMPLOYEE}) = (&(\mathrm{employer\ Deptypes}^{2} \cup \{\mathrm{restypes}\}),\\ &(\mathrm{affiliation}\ \{\mathrm{teacher,\ research,\ non\text{-}teaching}\}),\\ &(\mathrm{reimbursement}\ \{\mathrm{salary,\ honorarium}\})) \end{aligned}$$

- 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.

$$\mathrm{C}_{def}(\mathrm{PUBLICATION}) = ((\mathrm{researchArea\ Deptypes}))$$

- Assumptions in the definition of the relationship HAS-PUBLISHED [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].

$$\begin{aligned} \mathrm{C}_{def}(\mathrm{HAS\text{-}PUBLISHED}) = (&(\mathrm{author\ EMPLOYEE\ (affiliation}\ \{\mathrm{research}\})),\\ &(\mathrm{article\ PUBLICATION})) \end{aligned}$$

- 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.

$$\begin{aligned} \mathrm{C}_{def}(\mathrm{WORKER}) = (&(\mathrm{employer}\ \{\mathrm{judiciary,\ executive,\ legislative}\}),\\ &(\mathrm{reimbursement}\ \{\mathrm{salary,\ honorarium}\})) \end{aligned}$$

- Assumptions in the definition of the entity POSITION [A1]:

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

- 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. Since in this case we are trying to make explicit the meaning of a query, C_Q reflects the "meaning" perspective of semantics. Users can refer to pre-existing ontologies in the federation for choosing the contextual coordinates and their values (see Section 5.3).

Consider the example query Q on page 11 [A2,A4].
C_Q = ((author self), (designee self),
(employer {legislative, restypes}), (post ((appt elected))),
(article ((title "*abortion*"))), (researchArea {socialSciences, politics})) where,
"self" refers to the answer expected from the query Q. This is analogous to the arguments of the select clause in an SQL statement.

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).

3.4 Schema Correspondences as a uniform formalism

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. As described in detail in [KS93], this formalism is a generalization of the concept of *connectors* used to augment the relational model in [CRE87].

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

$\mathbf{schCor(O_1, O_2) = < O_1, attr(O_1), O_2, attr(O_2), \Psi >}$, 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.

The concept of dynamic attributes has been proposed in [LA86] to specify the mappings between different attributes. Various ways of implementing the mappings are proposed (viz. mathematical formulae, tables, programs). However, we here focus on the specification of mappings at a conceptual level between the domains of attributes and objects. In Sections 4.2 we will discuss how these schema correspondences support the information correlation between information systems.

3.5 Schema Correspondences and Context

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.

The association between the definition contexts and the objects in the database might be implemented in different ways by various component systems. 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

$$\mathbf{schCor(O_F,\ O) = <\ O_F,\ coord(C_{def}(O)),\ O,\ attr(O),\ \Psi\ >}$$

Consider the object EMPLOYEE as defined in the example on page 10. Let the object corresponding to the definition context C_{def}(EMPLOYEE) be EMPLOYEE_F.

The schema correspondences associated with the context C_{def}(EMPLOYEE) are :

- $<$ EMPLOYEE_F, {employer}, EMPLOYEE, {Dept}, Ψ $>$
  where Ψ : Deptypes $\cup$ {restypes} $\leftrightarrow$ Dept

- $<$ EMPLOYEE_F, {affiliation}, EMPLOYEE, {Affiliation}, Ψ $>$
  where Ψ : {teacher, research, non-teaching} $\leftrightarrow$ Affiliation

- $<$ EMPLOYEE_F, {reimbursement}, EMPLOYEE, {SalaryType}, Ψ $>$
  where Ψ : {salary, honorarium} $\leftrightarrow$ SalaryType

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

As illustrated earlier (Section 3.3.1), we use the *definition context* of an object to explicate the assumptions implicit in the mind of the designer about the objects in an information source. This 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 10, this sophistication and extra work is necessary and worthwhile because of the following reasons.

- The value addition introduced (as discussed above) facilitates the information discovery process through context comparison. This is illustrated in the next section.
- 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.

We assume here that the information sources relevant to the user query have been identified (see Section 5.2). 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 10 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.

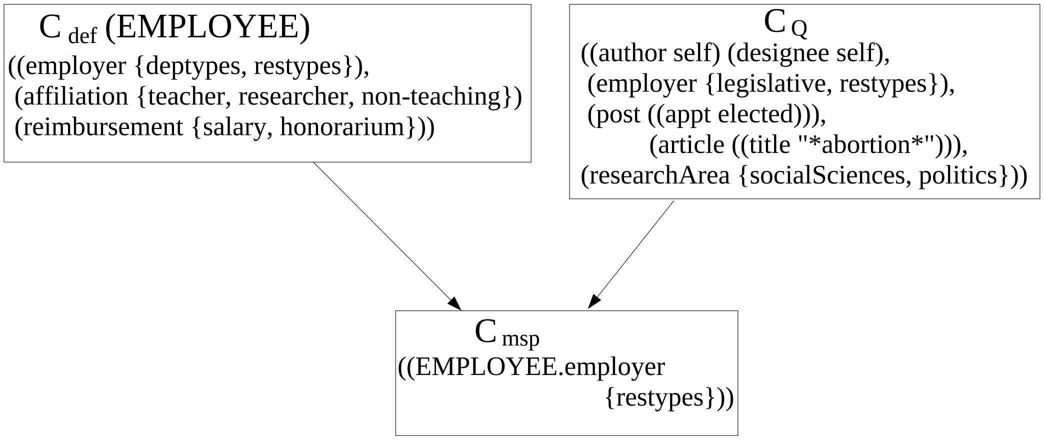
C def (EMPLOYEE)
((employer {deptypes, restypes}),
(affiliation {teacher, researcher, non-teaching})
(reimbursement {salary, honorarium}))
C Q
((author self) (designee self),
(employer {legislative, restypes}),
(post ((appt elected))),
(article ((title "*abortion*"))),
(researchArea {socialSciences, politics}))
C msp
((EMPLOYEE.employer
{restypes}))

Figure 3: Context Comparison : Focusing on the relevant employees

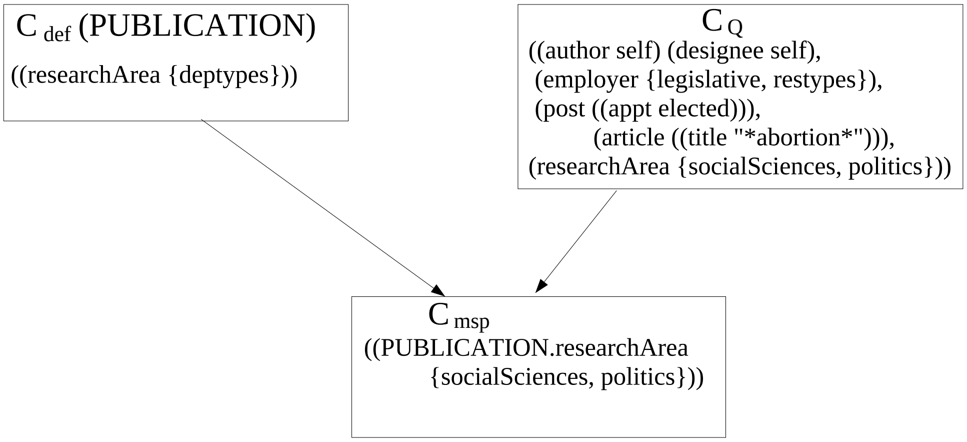
C def (PUBLICATION)
((researchArea {deptypes}))
C Q
((author self) (designee self),
(employer {legislative, restypes}),
(post ((appt elected))),
(article ((title "*abortion*"))),
(researchArea {socialSciences, politics}))
C msp
((PUBLICATION.researchArea
{socialSciences, politics}))

Figure 4: Context Comparison : Focusing on the relevant research areas

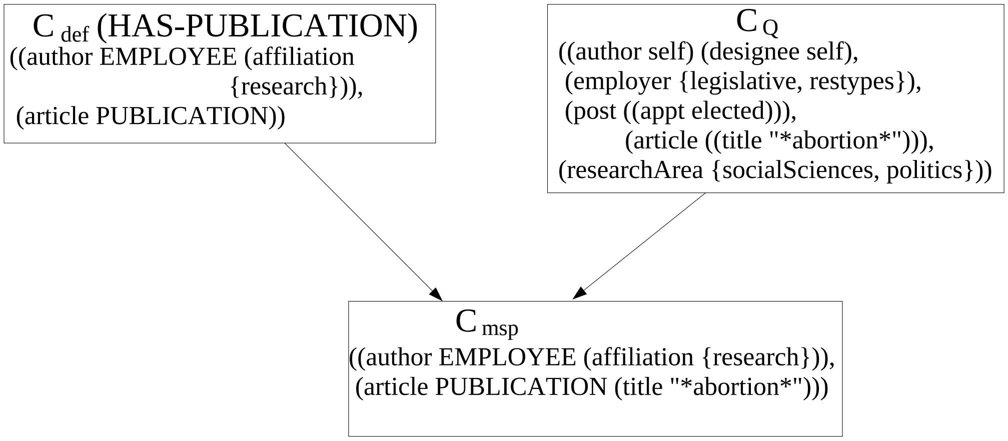
C def (HAS-PUBLICATION)
((author EMPLOYEE (affiliation
{research})),
(article PUBLICATION))
C Q
((author self) (designee self),
(employer {legislative, restypes}),
(post ((appt elected))),
(article ((title "*abortion*"))),
(researchArea {socialSciences, politics}))
C msp
((author EMPLOYEE (affiliation {research})),
(article PUBLICATION (title "*abortion*")))

Figure 5: Context Comparison : Focusing on the relevant publications

In Figure 3, 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 4, 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 5, 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 3.3.1 we demonstrated how the implicit design assumptions are represented as definition contexts at each database site. We assume that the associations between the definition contexts and the corresponding objects are stored at each database site. The associations are expressed by using modified schema correspondences as illustrated in Section 3.5.

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 old schema correspondences 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 continue with our example started on page 10 to illustrate the process of *conditioning* the schema correspondences at the database site *wrt* to the C_{msp} at that site and determine the modified schema correspondences. We assume the existence of mappings between the various contextual coordinates and the objects in the databases as illustrated in Section 3.5. 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

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-PUBLISHED) 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-PUBLISHED}, {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-PUBLISHED
  where <$Q_{1,1}$.author.SS#, article> in (select * from HAS-PUBLISHED)
  $M_1 = M_{1,1} \circ M_{1,2}$, 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

In this section we enumerate various approaches for information resource discovery, explain how our approach is different, and how we can add value to the previous approaches. We adapt the mechanism of context comparison to propose an approach to information resource discovery.

5.1 Previous Approaches

A common text retrieval approach used in resource discovery on the internet is primarily based on text search of the contents of the documents. In the WAIS project [KM91], database servers keep complete inverted indices on stored document contents and execute full text searches on them. In the Archie project [ED92], files are currently located by their names. Names and descriptions of software packages and documents are also stored. Entries can also be text strings consisting of keywords and associated descriptions.

In the X.500 directory service [CCI91], entities in the directory are represented by entries in a global, hierarchical name space called the Directory Information Tree (DIT) according to the structural relationships between the classes of information they represent. In the CORBA [OMG93] specification, it is the Implementation Repository which contains the information about object implementations and a structure similar to the DIT in X.500 has been suggested as an implementation of the above.

In the Gopher project [McC92], objects are identified by type, user-visible name, server's host name and port number and the object's absolute path name within the system. Keyword searches on the contents of documents and boolean pattern matches are also performed. In the Netfind project [Sch90], the white pages directory tool tries to locate information about an Internet user given the user's name and organization. It returns information such as the user's address and e-mail address.

A "partly semantic" approach is based on classifying the information in a hierarchical manner and by using attribute based searches. The World Wide Web [BL$^+$92] has three discovery trees which classify information according to subject, server type and organization. It merges information discovery and hypertext techniques. In Nomenclator [OM93], attribute-based naming is implemented on top of other naming systems. An active catalog constrains the search space for a query by returning a list of data repositories where the answer to the query is likely to be found. In Semantic File Systems [S$^+$91], associative access is achieved by providing with an attribute extraction and a query interface.

Our approach to Information Resource Discovery

In our approach we extend the "partly semantic" approach by incorporating more semantics in the form of context information. Association of structural information with the context information enables us to use database techniques to get answers in an efficient manner. We believe that our approach would add significant value to the text based and partly semantic approaches enumerated above. The advantages of our approach can be summarized as follows:

- The approach followed in X.500 [CCI91], Gopher [McC92], Nomenclator [OM93] and CORBA [OMG93] is based on the structure of the objects. The structure of an object, however is inadequate to represent the information content of an object. Our approach incorporates the semantics of the object in an attempt to capture the information content.

- In the WAIS project [KM91] a complete inverted index of the text of all the documents is stored. This approach, however would not scale well in the infocosm consisting of millions of information sources. Since a full text index is comparable to the size of the document it references, we would need to build a structure of the same size as all the documents. We believe that a structure based on the semantic content of the document would be of a much smaller size.

- In X.500 [CCI91] the information is organized in a hierarchical manner. This approach would not scale well because it would be inefficient to search this hierarchy based criteria other than the ones used to develop it. We believe that an identification of the relevant criteria captured in a context representation could act as a *focusing mechanism* as illustrated in the examples in Section 4.1.

- In the approaches enumerated in the previous section, there is no sensitivity to environmental conditions. A partial exception is the World Wide Web, where three different classifications are represented in Discovery Trees. We believe that sensitivity to the environmental conditions can be easily captured in our representation of context.

5.2 Information Resource Discovery based on context comparison

The possibility 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 will be different from the definition context of a resource object in an information source. We now make the following observations.

- The definition context of the information source may, in the worst case, be the union of the contexts of all the resource objects in the information source. This will, however, make the process of information resource discovery inefficient as the query context will have to be compared with the definition contexts of the thousands of resource objects.

- 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. In case the information source is determined as relevant to the query, these abstractions can be used in information focusing (Section 4.1).
  - 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. In this case, we would need an inference mechanism which would identify these abstractions in the ontology and use them appropriately in determining the relevance of the information source to the query.
- 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 **C_{res}(Query, InformationSource)** at each site (see Figure 6). 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 [McC93], 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.

5.3 Issues of language and ontology in context representation

In this section we discuss the issues of a language in which the contexts can be best represented. We also discuss issues of ontology, i.e. the vocabulary used by the language to represent the contexts.

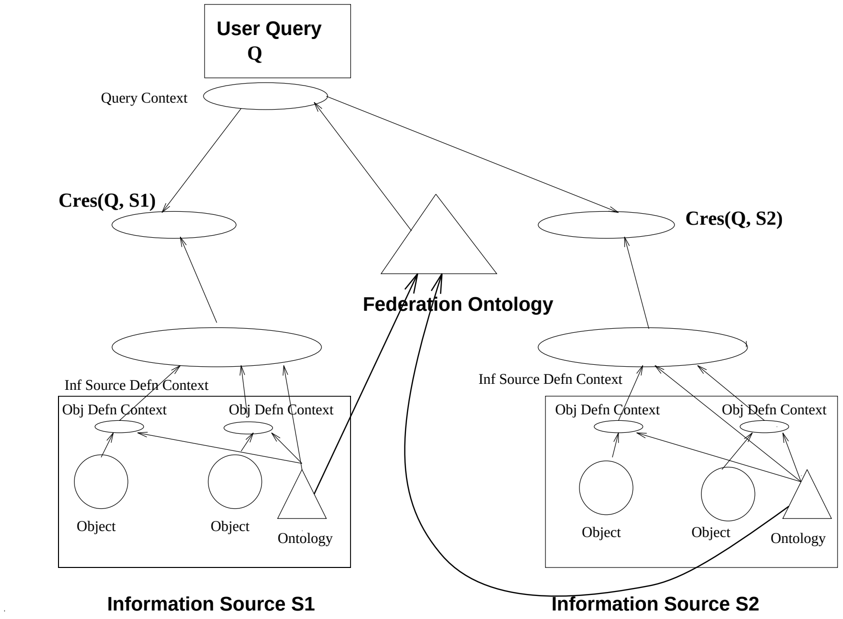

Figure 6: Information resource discovery using context comparison

5.3.1 Language for context representation

We envisage the context as the medium for information exchange between the information providers and the information broker on one hand, and the information consumer and the information broker on the other. The special features required in a language to represent the contextual coordinates in the context representation illustrated in Sections 3.3.1, 3.3.2 are:

- The language should have the ability to describe what kinds of sentences it can describe, i.e. it should be self-describing. This enables us to represent:
  - the definition contexts of the resource objects and that of the information source in the same uniform language. This is because the information source definition context might contain the definition contexts of the resource objects.
  - nesting of contexts to any arbitrary level. This is because the definition context of a resource object might contain the definition contexts of one or more resource objects.
- It should have a core feature which would hold all the contextual coordinates together. This would help express the notion of a semantic dependency between the contextual coordinates.
- The other feature would help express the context as a collection of contextual coordinates (meta-attributes), each describing a specific aspect of information in information systems.

- The language should have primitives for context manipulation (viz., determining the most specific subtype of two types, pattern matching, etc.) in the model world, which might be useful in the task of context comparison.

- The language should have primitives for performing inference on the ontology to identify the abstractions related to the ontological objects in the query context or the definition contexts of the information resource objects and the information source. We view ontology as the symbolic layer closest to the concepts in the real world.

We are looking into the possibility of the Knowledge Interchange Format [GF92] as the language for representing context.

5.3.2 The Ontology Problem

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 [Gru93]. 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 $C_{def}(O)$.

Another critical issue in designing an ontology for the federation is the values associated with the contextual coordinates. As proposed in Section 3.3 these values could be from a pre-existing ontology or types or 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.

We assume that the each information source has available to it an ontology corresponding to a specific domain. The definition contexts of the resource objects take their terms and values from this ontology. However in designing the definition contexts of the information sources and the query context, the issues of combining the various ontologies arise. Another issue is of presenting these ontologies to the user in order for him to construct the query context appropriately.

We now enumerate various approaches one might take in building ontologies for a federation of information sources. Other than the ontological commitment, a critical issue in designing ontologies is the **scalability** of the the ontology as more information sources enter the federation.

- **The Common Ontology approach:**

  - One approach has been to build an extensive global ontology. A notable example of global ontology is Cyc [LG90] consisting of around 30,000 objects. In Cyc, the mapping between each individual information resource and global ontology is accomplished by a set of *articulation axioms* which are used to map the entities of an information resource to the concepts (viz. frames, slots) in Cyc's existing ontology.

  - Another approach has been to exploit the semantics of a single problem domain (viz. transportation planning) [ACHK93]. The domain model is a declarative description of the objects and activities possible in the application domain as viewed by a typical user. The user formulates queries using terms from the application domain.

  In our opinion both the above attempts are lacking in scalability of their approaches. In the Cyc example, the maintenance of the ontology *wrt* the consistency of the various concepts is a difficult process. In the second example ontologies of only one domain are modeled. The query processing techniques are geared for only a single domain and will not scale up for answering queries requiring correlation of information between different domains.

- **Reuse of Existing Ontologies:** Given our assumption that there will be numerous information systems participating in the federation, it is unrealistic to expect any one existing ontology or classification to suffice. We propose a re-use of various existing classifications viz. ISBN classification for publications, botanical classification for plants etc.

  These ontologies can then be combined in different ways and made available to the federation. A critical issue in combining the various ontologies is determining the overlap between them. One possibility is two 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 enunciated the concept of *infocosm* as a society with ubiquitous access/exchange of information as a "tradeable" commodity. Two major classes of applications that are likely in this future society are *mass market applications* and *information content sensitive applications*. We propose a conceptual architecture in which the applications may be realized. The three main components of the architecture are: *information providers/sources*, *information brokers* and *information consumers*.

We present an informal classification of the various information brokering approaches possible in the infocosm, viz. *user directed approach*, *syntactic keyword/attribute based approach*, *descriptive semantics based approach* and *cognition based approach*. We advocate a semantics based approach, especially for the information content sensitive applications.

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 language and 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. Addressing the challenges arising from the semantics-based approach will be necessary (but not sufficient) to exploit fully the tremendous possibilities in the emerging infocosm. An important aspect not addressed in this report is that of efficient query processing, including the issues of design of efficient indices and access structures, efficient search strategies based on caching of data and meta-data, handling system and database heterogeneities, etc.

Acknowledgments

We thank Dr. Saul Amarel for suggesting the interesting query on page 11 that led us to explore several of the issues discussed in this paper. We also thank Luke Kane for his valuable comments on the introduction and the information provided in the appendix. Dr. Gio Wiederhold pointed us to the need for combining ontologies.

References

[ACHK93] 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.

[BL$^+$92] T. Berners-Lee et al. World-Wide Web : The Information Universe. *Electronic Networking : Research, Applications and Policy*, 1(2), 1992.

[BN75] D. Bobrow and D. Norman. Some principles of Memory Schemata. In *Representation and Understanding*. New York : Academic Press, 1975.

[BW85] D. Bobrow and T. Winograd. An overview of KRL, a Knowledge Representation Language. In *Readings in Knowledge Representation*. Morgan Kaufmann, 1985.

[CCI91] CCIT. The Directory - Overview of Concepts, Models and Services, CCIT X.500 Series Recommendations. Technical report, CCIT, December 1991.

[CMG90] G. Chierchia and S. McConnell-Ginet. *Meaning and Grammar : An Introduction to Semantics*, chapter 6. MIT Press Cambridge MA, 1990.

[CRE87] B. Czejdo, M. Rusinkiewicz, and D. Embley. An approach to Schema Integration and Query Formulation in Federated Database Systems. In *Proceedings of the 3rd IEEE Conference on Data Engineering*, February 1987.

[ED92] A. Emtage and P. Deutsch. Archie : An Electronic Directory Service for the Internet. In *Proceedings of the Winter 1992 Usenix Conference*, 1992.

[FKN91] 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.

[gen94] Software 'Agents' will make life easy. In Fortune, January 1994.

[GF92] M. Genesereth and R. Fikes. Knowledge interchange format, version 3.0 reference manual. Technical Report Logic-92-1, Computer Science Department, Stanford University, 1992.

[Gru93] 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.

[int93] An Interactive Life. In Newsweek, May 1993.

[KC88] 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.

[KM91] B. Kahle and A. Medlar. An Information System for Corporate Users : Wide Area Information Servers. *Connexions - The Interoperability Report*, 5(11), November 1991.

[KS93] 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.

[LA86] W. Litwin and A. Abdellatif. Multidatabase Interoperability. *IEEE Computer*, 19(12), December 1986.

[LG90] D. Lenat and R. V. Guha. *Building Large Knowledge Based Systems : Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company Inc, 1990.

[mar93] World Multimedia Applications. In Market Intelligence, 1993.

[MC91] G. A. Miller and W. G. Charles. Contextual Correlates of Semantic Similarity. *Languauge and Cognitive processes*, 1991.

[McC92] M. McCahill. The Internet Gopher Protocol : A Distributed Server Information System. *Connexions - The Interoperability Report*, 6(7), July 1992.

[McC93] J. McCarthy. Notes on formalizing Context. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1993.

[ML92] S. H. Myaeng and M. Li. Building Term Clusters by acquiring Lexical Semantics from a Corpus. In *Proceedings of the CIKM*, 1992.

[OM93] J. Ordille and B. Miller. Distributed Active Catalogs and Meta-Data Caching in Descriptive Name Services. In *Proceedings of the 13th International Conference on Distributed Computing Systems*, May 1993.

[OMG93] OMG. The Common Object Request Broker : Architecture and Specification. Technical report, Object Management Group, Inc., December 1993.

[S$^+$91] M. Sheldon et al. Semantic File Systems. In *Proceedings of the 13th ACM Symposium on Operating System Principles*, 1991.

[Sch90] M. Schwartz. Experience with a Semantically Cognizant Internet White Pages Directory Tool. *Internetworking Research and Experience*, 1(2), December 1990.

[SG89] A. Sheth and S. Gala. Attribute relationships : An impediment in automating Schema Integration. In *Proceedings of the NSF Workshop on Heterogeneous Databases*, December 1989.

[SG93] A. Sheth and S. Gala. On automatic Reasoning for Schema Integration. *International Journal on Intelligent and Cooperative Information Systems*, 2(1), March 1993.

[sim93] On-line Services: 1993 Review, Trends, and Forecast. By SIMBA Information Inc., 1993.

[SK92] 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.

[SL90] A. Sheth and J. Larson. Federated Database Systems for managing Distributed, Heterogeneous and Autonomous Databases. *ACM Computing Surveys*, 22(3), September 1990.

[SM91] M. Siegel and S. Madnick. A Metadata Approach to resolving Semantic Conflicts. In *Proceedings of the 17th VLDB*, September 1991.

[SS93] J. Sheth and R. Sisodia. The information mall. 1993.

[SSR92] E. Sciore, M. Siegel, and A. Rosenthal. Context Interchange using Meta-Attributes. In *Proceedings of the CIKM*, 1992.

[tel93] George Gilder's TELECOSM: The New Rule of Wireless and TELECOSM: Digital Darkhorse Newspapers. In Forbes ASAP, 1993.

[Tho89] J. P. Thompson. *Data with Semantics : Data Models and Data Management.* Van Nostrand Reinhold - New York, 1989.

[VD92] D. A. Voss and J. R. Driscoll. Text Retrieval using a Comprehensive Lexicon. In *Proceedings of the CIKM*, 1992.

[Wie92] G. Wiederhold. Mediators in the Architecture of Future Information Systems. *IEEE Computer*, 25(3), March 1992.

[Woo85] J. Wood. What's in a link ? In *Readings in Knowledge Representation.* Morgan Kaufmann, 1985.

Appendix

The merging of computers and communications technology with the advent of the high bandwidth optic fiber networks is expected to bring about a qualitative sea-change in the consumer oriented markets of today. It is expected that more and more *information* will become a commodity to be traded and exchanged in the market place. Various companies and business forecasting agencies have been trying to size up and predict this future evolving market. We review two such market studies below.

The first is a DEC study [mar93] which reports the market growth as shown in Figure 7. It further states that only 10% of the information used to make corporate decisions is captured in the alphanumeric information systems today. The report concluded that 90% could be captured tomorrow in multimedia databases. We expect that a significant portion of the new information will be captured by representing the semantics of the information as a part of the meta-data. Our semantics-based information brokering techniques will utilize that meta-data.

The second is a market survey done in [sim93] which suggests that the mass market or "Business to Residence" market for information services is half of the "Business to Business" market. The forecast in their projections as enumerated in the table below is somewhat conservative. It is a linear extrapolation of the current 9% annual growth rate. Even if the mass market application segment were to explode, there is a significant market

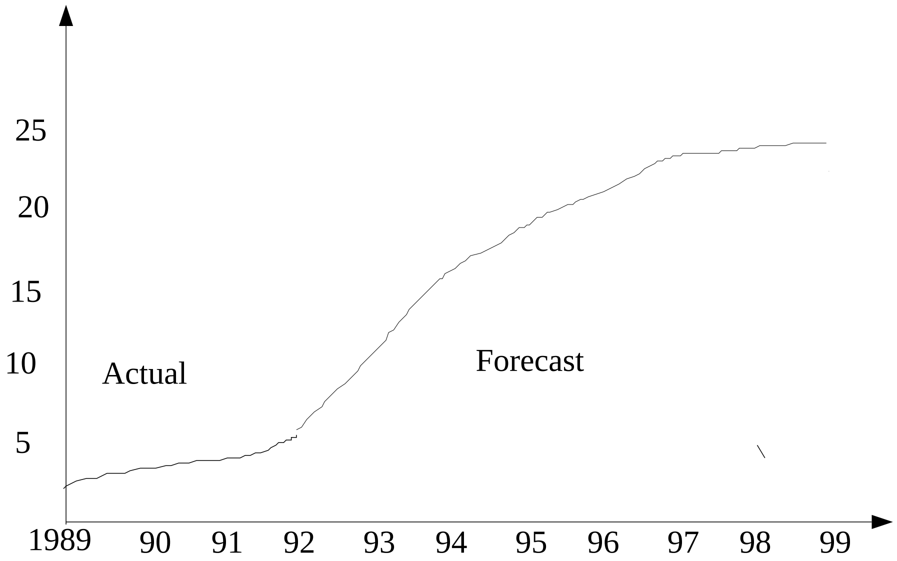

Figure 7: Predicted growth of the market for information systems

for Business to Business services which typically involve information content sensitive applications. We claim that these applications can be served well by the semantics-based techniques.

Business to Business	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**
Brokerage	4474[3]	4810	5185	5594	6042	6537
Credit	1900	2031	2142	2249	2362	2480
Financial news/Research	1850	2055	2280	2520	2765	3020
Legal/Regulatory	756	803	855	913	974	1041
Marketing	40	50	65	80	94	110
Professional	661	710	761	818	881	949
Subtotal	**9680**	**10458**	**11289**	**12175**	**13117**	**14137**
Business to Residence						
General Interest	440	619	720	821	921	1032
Individual Investor	19	26	30	35	39	44
Telephone Company Gateway	5	7	8	9	10	11
Subtotal	**463**	**651**	**758**	**865**	**970**	**1087**
Grand Total	**10144**	**11110**	**12047**	**13040**	**14087**	**15224**

Table 1: Projections of Business Revenues for On-Line services

Related Work

So Far (Schematically) yet So Near (Semantically)

Amit Sheth[1] and Vipul Kashyap[2]
(invited paper) Proceedings of the IFIP DS-5 Conference on Semantics of Interoperable Database Systems, Lorne, Australia, November 1992; In IFIP Transaction A-25, North Holland, 1993.

Abstract

In a multidatabase system, schematic conflicts between two objects are usually of interest only when the objects have some semantic affinity. In this paper we try to reconcile the two perspectives. We first define the concept of semantic proximity and provide a *semantic taxonomy*. We then enumerate and classify the *schematic and data conflicts*. We discuss *possible semantic similarities* between two objects that have various types of schematic and data conflicts. Issues of uncertain information and inconsistent information are also addressed.

Schema Correspondences between Objects with Semantic Proximity

Vipul Kashyap[2] and Amit Sheth[1]
Technical Report DCS-TR-301, Department of Computer Science, Rutgers University, October 1993.

Abstract

In a multidatabase system, schematic conflicts between two objects are usually of interest only when the objects have some semantic similarity. In this paper we try to reconcile the schematic and semantic perspectives. We introduce a uniform formalism called *schema correspondences* to represent structural similarities between the objects. We represent the semantic similarities between the objects using the concept of *semantic proximity.* We show how the reconciliation is achieved by illustrating the association of the schema correspondence(s) with and as component(s) of the semantic proximity. We also provide a data model independent *semantic taxonomy* on the basis of the semantic proximity defined. We then enumerate and classify the *schematic and data conflicts.* The association between the schema correspondences and semantic proximity helps represent the *possible semantic similarities* between two objects having these conflicts. One representation of *uncertain information* using semantic proximity as the basis is explored. Issues of *inconsistent information* are also discussed in the framework of semantic proximity.

[1] Bellcore, 444 Hoes Lane, Piscataway, NJ 08854-4182

[2] Department of Computer Science, Rutgers University,
New Brunswick, NJ 08903