Estimating Information Loss for Multi-ontology Based Query Processing

E. Mena[1] V. Kashyap[2] A. Illarramendi[3] A. Sheth[4]
[1]*IIS depart., Univ. de Zaragoza. Spain. jibmenie@si.ehu.es*
[2]*MCC. Austin, TX 78759-6509. kashyap@mcc.com*
[3]*LSI depart., UPV. San Sebastián. Spain. http://siul02.si.ehu.es/~jirgbdat*
[4]*LSDIS Lab, Univ. of Georgia, Athens, GA 30602. http://lsdis.cs.uga.edu*

Abstract

The World Wide Web is fast becoming a ubiquitous computing environment. Prevalent keyword-based search techniques are scalable, but are incapable of accessing information based on concepts. We investigate the use of concepts from multiple, real-world pre-existing, domain ontologies to describe the underlying data content and support information access at a higher level of abstraction. It is infeasible to expect a single domain ontology to describe the vasts amounts of data on the web. In fact we expect multiple ontologies to be used as different world views and present an approach to "browse" ontologies as a paradigm for information access. A critical challenge in this approach is the vocabulary heterogeneity problem. Queries are rewritten using interontology relationships to obtain translations across ontologies. However, some translations may not be semantics preserving, leading to uncertainty or loss in the information retrieved. We present a novel approach for estimating loss of information based on navigation of ontological terms. We define measures for loss of information based on intensional information as well as on well established metrics like *precision* and *recall* based on extensional information. This measures are used to select results of the desired quality of information.

1 Introduction

The World Wide Web (WWW) has fast become a preferred information access and application support environment for a large number of computer users. In most cases, there is no centralized or controlled information management, as anyone can make a wide variety of data available on the Web. This has lead to an exponential growth in the information accessible on the Web. In distributed and federated database systems, logical integration of the schemas describing the underlying data is used to handle the structural and representational heterogeneity. In a tightly coupled federated database approach, the relationships are fixed at schema integration or definition time. In a loosely coupled federated database (or multidatabase) approach, the relationships are defined when defining the multidatabase view prior to querying the databases. Neither of these options are feasible or appealing in the much more diversified and unmanaged Web-centric environment.

Use of domain specific ontologies is an appealing approach to allow users to express information requests at a higher level of abstraction compared to keyword based access. We view ontologies as the specification of a representational vocabulary for a shared domain of discourse which may include

definitions of classes, relations, functions and other objects [10]. Since it is infeasible to expect a single ontology to describe the vast amounts of data on the web, we propose the use of multiple domain specific ontologies as different world views describing the wide variety of data and capturing the needs of a varied community of users. A critical issue that prevents wide spread use of ontologies is the labor intensive nature of the process of designing and constructing an ontological specification. In the OBSERVER[1] system, we demonstrate our approach of using multiple pre-existing real-world domain ontologies to access heterogeneous, distributed and independently developed data repositories. This enables the use of "off the shelf ontologies", thus minimizing the need of designing ontologies from scratch.

One consequence of our emphasis on ontology re-use is that they are developed independently of the data repositories and have been used to describe information content in data repositories independently of the underlying syntactic representation of the data [11]. New repositories can be easily added to the system by mapping ontological concepts to data structures in the repositories. This approach is more suitable for open and dynamic environments such as the Web and allows each data repository to be viewed at the level of the relevant semantic concepts.

We present an approach for browsing multiple related ontologies for information access. A user query formulated using terms in some user view/domain ontology is translated by using terms of other (target) domain ontologies. Mechanisms dealing with incremental enrichment of the answers are used. The substitution of a term by traversing interontological relationships like *synonyms* (or combinations of them [15, 14]) and combinations of *hyponyms* (specializations) and *hypernyms* (generalizations) provide answers not available otherwise by using only a single ontology. This, however, changes the semantics of the query. The main contribution of this paper is the use of mechanisms to estimate loss of information (based on intensional and extensional properties) in the face of possible semantic changes when translating a query across different ontologies. It may be noted that in our approach thousands of data repositories may be described by hundreds of ontologies. In general, a user may be willing to sacrifice the quality of information for a quicker response from the system, as said in [19].

Several projects that deal with the problem of interoperable systems can be found in the literature, e.g., TSIMMIS [5], SIMS [1], Information Manifold [12], InfoSleuth [17], Carnot [7], etc. The commonalities between their approach and ours can be summed up as: (a) some way of using a high level in semantic view to describe data content (ontologies); and (b) use of specialized wrappers to access underlying data repositories. However, we present in this paper techniques, not considered by the others, that would allow previous systems to estimate the loss of information incurred when translating user queries into other ontologies. This measure of the loss (whose upper limit is defined by the user) guides the system in navigating those ontologies that have more relevant information; it also provides the user with a "level of confidence" in the answer that may be retrieved. We use well-established metrics like precision and recall and adapt them to our context in order to measure the change in semantics incurred when providing an answer with a certain degree of imprecision, as opposite to measure the change in the extension, like classical Information Retrieval methods do.

The rest of the paper is organized as follows[2]. Section 2 introduces the query processing strategy in OBSERVER and briefly discusses the translation mechanisms when synonym, hyponym and hypernym relationships are used for controlled and incremental query expansion to different target ontologies. Section 3 discusses the techniques used for estimating the imprecision in information retrieval. Conclusions are presented in Section 4.

[1] *Ontology Based System Enhanced with Relationships for Vocabulary hEterogeneity Resolution.*

[2] To avoid duplication and for brevity, we do not repeat much of the basic discussion on query processing approach and prototype system architecture which appears in [14, 15] and focus here only on the new contributions.

2 Query Processing in OBSERVER

The idea underlying our query processing algorithm is the following: give the first possible answer and then enrich it in successive iterations until the user is satisfied. Notice that in our context thousands of data repositories described by hundreds of ontologies could be available so users will prefer to get a good set of semantically correct data rather than waiting for a long time until all the relevant data in the Global Information System has been retrieved. Moreover, certain degree of imprecision (defined by each user) in the answer could be allowed if it helps to speed up the search of the wanted information.

We use ontologies WN and Stanford-I (see [13]) and the following example query to illustrate the main steps of our query expansion approach.

User Query: '*Get title and number of pages of books written by Carl Sagan*'

2.1 Step 1: Query Formulation over user ontology

The user browses the available ontologies (which are ordered by knowledge areas) and chooses a *user ontology* that includes the terms needed to express the semantics of her/his information needs. Then, and with the help of a GUI, the user chooses terms from the user ontology to build the constraints and projections that comprise the query.

In the example, the WN ontology is selected since it contains all the terms needed to express the semantics of the query, i.e., terms that store information about titles ('NAME'), number of pages ('PAGES'), books ('BOOK') and authors ('CREATOR').

$$\mathcal{Q} = \textit{[NAME PAGES] for (}\textbf{AND}\ \textit{BOOK (}\textbf{FILLS}\ \textit{CREATOR "Carl Sagan"))}$$

Syntax of the expression is taken from CLASSIC [2], the system based on Description Logics (DL) [3] that we use to describe ontologies.

2.2 Step 2: Access data underlying user ontology and present answer

The DL expression that comprises the query is translated, with the help of *mappings*[3] [15, 8] of the terms involved in such an expression, into several subqueries expressed in the local query language of the underlying repositories[4]. To perform that task the system uses different translators and wrappers. Different answers coming from different data sources must be translated into the "language" of the user ontology to facilitate removal of redundant objects and update of incomplete objects. Thus, the answer is correlated and presented to the user. A more detailed description of this step appears in [9].

2.3 Step 3: Controlled and Incremental Query Expansion to Multiple Ontologies

If the user is not satisfied with the answer, the query processor retrieves more data from other ontologies in the Information System to "enrich" the answer in an incremental manner. Some researchers

[3]Mappings in our approach are expressions of Extended Relational Algebra that relate terms in ontologies with the underlying data elements.

[4]In the case of relational databases, the DL expression is translated into a list of SQL sentences.

have looked into the problem of query relaxation [6, 4]. However, they have proposed techniques for query relaxation within the same schema/ontology/knowledge base. We differ with the above in two important ways: (1) we propose techniques for query relaxation *across* ontologies by using synonym, hyponym and hypernym relationships; and (2) we provide techniques to estimate the loss of information incurred.

In our system, a new component ontology which we call the *target ontology* is selected from the Global Information System. The user query must be expressed/translated into terms of that target ontology. For that task the user and target ontologies are integrated (see [14]) by using the interontology relationships defined between them. When a new ontology is made available to the Global Information System the semantic relationships between its terms and other terms in other ontologies must be defined in a module called the Interontology Relationship Manager (IRM) [15]. Thus, this module manages the *semantic properties* between terms in different ontologies. This information allows an integration of two given ontologies in the system without user intervention. The IRM is the key for managing different component ontologies without missing the semantics of each one.

As we said above, the user and target ontology are integrated automatically by the system. The user query will be rewritten and classified in the integrated ontology during this process. Two situations can occur after integration:

1. All the terms in the user query may have been rewritten by their corresponding synonyms in the target ontology. Thus the system obtains a semantically equivalent query (*full translation*) and no loss of information is incurred. In this case, the plan to obtain the answer consists on accessing the data corresponding to the translated query expression.

2. There exist terms in the user query that can not be translated into the target ontology - they do not have synonyms in the target ontology (we called them *conflicting terms*). Then the translation is called a *partial translation.*

If the user allows the system to provide answers with a certain degree of imprecision, new plans could be generated by substituting the conflicting terms by **semantically similar** expressions that could lead to a full translation of the user query. So, each conflicting term in the user query is replaced by the intersection of its immediate parents (*hypernyms*) or by the union of its immediate children (*hyponyms*), recursively, until a translation of the conflicting term is obtained using only the terms of the target ontology. Each substitution of a term in the original query implies a certain loss of information.

This process can generate *several* possible translations of a conflicting term into a target ontology. All the possibilities are explored and the result is a list of plans for which the system will estimate the associated loss (discussed in Section 3).

2.4 Example: generation of plans

We now illustrate the computation of the plans obtained by processing our example query.

$$\mathcal{Q} = \textit{[NAME PAGES] for (}\textbf{AND}\ \textit{BOOK (}\textbf{FILLS}\ \textit{CREATOR ``Carl Sagan''))}$$

The query $\mathcal{Q}$ has to be translated into terms of the Stanford-I ontology [13] (the ontology chosen as target ontology). After the process of integrating the WN and Stanford-I ontologies, $\mathcal{Q}$ is redefined as follows:

$\mathcal{Q}$ = *[title number-of-pages] for (***AND** *BOOK (***FILLS** *doc-author-name "Carl Sagan"))*

The only conflicting term in the query is 'BOOK' (it has no translation into terms of Stanford-I). The process of obtaining the various plans corresponding to the different translations of the term 'BOOK' is not described here due to space limitation, but it results on the four following:

Plan 1: (**AND** document (**FILLS** doc-author-name "Carl Sagan"))

Plan 2: (**AND** periodical-publication (**FILLS** doc-author-name "Carl Sagan"))

Plan 3: (**AND** journal (**FILLS** doc-author-name "Carl Sagan"))

Plan 4: (**AND** UNION(book, proceedings, thesis, misc-publication, technical-report) (**FILLS** doc-author-name "Carl Sagan"))

Notice that 'BOOK' has been translated by the expressions 'document', 'periodical-publication', 'journal' or 'UNION(book, proceedings, thesis, misc-publication, technical-report)', respectively. Details of this translation process can be found in [14].

The computation of the loss of information incurred in the substitution of the user query by each plan is illustrated in the next section.

3 Estimating the Loss of Information

We now discuss the central theme of the paper, where we describe approaches to measure the change in semantics when a term in a query is replaced by an expression from another ontology (in a try of getting a full translation of the user query).

There have been approaches in the research literature for approximating query answering in situations where multiple answers may be obtained from multiple information sources. Most approaches are typically accompanied by an attempt to estimate some measure of divergence from the true answer and are based on some technique of modeling uncertainty. In the Multiplex project [16], the soundness and completeness of the results are estimated based on the intersections and unions of the candidate results. In our approach, the Information Retrieval analogs of soundness (precision) and completeness (recall) are estimated based on the sizes of the extensions of the terms. We combine these two measures to compute a composite measure in terms of a numerical value. This can then be used to choose the answers with the least loss of information. Numerical probabilistic (possibilistic) measures are on the other hand used in [TC93, DLP94], but are based on *ad hoc* estimates of the initial probability (possibility) values. In our approach we provide a set theoretic basis for the estimation of information loss measures.

In our case, the change in semantics caused by the use of hyponym and hypernym relationships must be measured not only to decide which substitution minimizes the loss of information but also to present to the user some kind of "level of confidence" in the new answer. This would enable the system to navigate those ontologies which contain more relevant information for the user needs. In this section, we define and illustrate with examples, measures for estimating the loss of information. First, we present a way of measuring the change in semantics based on intensional information, and second, a technique that measures the change in semantics based on extensional information. Both measures are presented to the user whenever a new answer is obtained.

3.1 Loss of Information measurement based on intensional information

In our context, and due to the use of DL systems for describing and querying the ontologies, loss of information can be expressed like the terminological difference between two expressions, the user query and its translation. The terminological difference between two expressions consists of those

constraints of the first expression that are not subsumed by the the second expression. The DL system is able to calculate the difference automatically[5]. Let us show an example based on the plans obtained in Section 2.4:

Original query: *Q = [NAME PAGES] for (***AND** *BOOK (***FILLS** *CREATOR "Carl Sagan"))*
Plan 1: *Q = [title number-of-pages] for (***AND** *document (***FILLS** *doc-author-name "Carl Sagan"))*

Taking into account the following term definitions[6]:

*BOOK = (***AND** *PUBLICATION (***ATLEAST** *1 ISBN)),*
*PUBLICATION = (***AND** *document (***ATLEAST** *1 PLACE-OF-PUBLICATION))*

The terminological difference is, in this case, the constraints of $\mathcal{Q}$ not considered in the plan: (**AND** (**ATLEAST** 1 ISBN) (**ATLEAST** 1 PLACE-OF-PUBLICATION))

A special problem arises when computing intensional loss due to the vocabulary differences. As the intensional loss is expressed using terms of two different ontologies, the explanation might make no sense to the user as it mixes two "vocabularies". The problem can be even worse if both ontologies are expressed in different natural languages. So, the intensional loss can help to understand the loss only in some cases.

We re-visit plans found in the example of Section 2.4 and enumerate the intensional loss incurred in the case of Plan 1 and Plan 4. Intensional loss of information for plans 2 and 3 are equivalent to the one for Plan 1.

Original query: *Q = [NAME PAGES] for (***AND** *BOOK (***FILLS** *CREATOR "Carl Sagan"))*

- Plan 1= [title number-of-pages] for (**AND** document (**FILLS** doc-author-name "Carl Sagan"))
  Loss="Instead of books written by Carl Sagan, OBSERVER is providing all the documents (even if they do not have an ISBN and place of publication)[7] written by Carl Sagan."

- Plan 4= [title for number-of-pages] for (**AND** UNION(book, proceedings, thesis, misc-publication, technical-report) (**FILLS** doc-author-name "Carl Sagan"))
  Loss="Instead of books written by Carl Sagan, OBSERVER is providing the books[8], proceedings, theses, misc-publication and technical manuals written by Carl Sagan. Any book not included in this group is not retrieved."

In addition to the vocabulary problem, an intensional measure of the loss of information can make it hard for the system to decide between two alternatives, in order to execute first that plan with less loss. Thus, some numeric way of measuring the loss should be explored.

[5]If the concrete DL system used lacks of that feature the terminological difference could be calculated anyway with the help of its *subsumption mechanism* (see [3]).

[6]The terminological difference is calculated between the extended definitions.

[7]'(ATLEAST 1 ISBN)' and '(ATLEAST 1 PLACE-OF-PUBLICATION)' are constraint in the description of 'BOOK' with no translation into Stanford-I and they were ignored in all the plans; see example in Section 2.4.

[8]This is the problem commented at the beginning of this section. The sentence makes no sense for the user since they are homonyms.

3.2 Loss of Information measurement based on extensional information

We also measure the loss of information based on the number of instances of terms involved in the substitutions performed on the query. Since the measure depends on the sizes of the term extensions, we first discuss techniques to estimate the extensions of complex expressions based on set theoretic operations such as unions and intersections of terms. Second, we briefly describe a composite measure combining measures like *precision* and *recall* [18] that is used to estimate the information loss when a term is substituted by an expression. The composite measure defined takes into account the bias of the user ("is precision more important or recall ?"). And third, we show our proposal for adapting these measures based on semantic relationships between the various expressions. We give priority to semantic relationships before resorting to extensional information because, for instance, it can happen that a term in one ontology is more general **semantically** than another term in another ontology; at the same time the subsumer term can have less instances than the subsumed term because they belong to different ontologies and take instances from different data repositories (not necessarily related). In Section 3.3 real examples of these cases are shown and above techniques are illustrated by evaluating the loss of information for the plans presented in Section 2.4.

3.2.1 Estimating the size of the extension of an expression

Given an expression considered as a translation of a conflicting term, we approximate the size of its extension. The expression is a combination of unions and intersections of terms in the target ontology since at each step, the system substituted conflicting terms by the intersection of its parents or by the union of its children. The estimate is an interval with an upper (|Ext(Expr)|.high) and lower (|Ext(Expr)|.low) bound. It is computed as follows:

- The intersection of two sets can be empty at the least (no overlap). At the most, the intersection of two sets can only be the smaller of the two sets (maximum overlap).
  $|Ext(Subexpr1) \cap Ext(Subexpr2)|.low = 0$
  $|Ext(Subexpr1) \cap Ext(Subexpr2)|.high = \min [\ |Ext(Subexpr1)|.high, |Ext(Subexpr2)|.high\ ]$
- The union of two sets can be at the least the bigger of the two sets (maximum overlap). At the most the size of the union can be the sum of the sizes of the two sets (no overlap).
  $|Ext(Subexpr1) \cup Ext(Subexpr2)|.low = \max [\ |Ext(Subexpr1)|.high, |Ext(Subexpr2)|.high\ ]$
  $|Ext(Subexpr1) \cup |Ext(Subexpr2)|.high = |Ext(Subexpr1)|.high + |Ext(Subexpr2)|.high$

As trivial case, when "Subexpr" is the name of a term, both bounds are equal to the size of the extension of such a term. Intervals when calculating the extension implies intervals for the resulting precision, recall and loss of information.

3.2.2 A Composite Measure combining Precision and Recall

Precision and Recall have been very widely used in Information Retrieval literature to measure loss of information incurred when the answer to a query issued to the information retrieval system contains some proportion of *irrelevant* data [18]. The measures are defined, and adapted to our context, as follows:
C-Term = conflicting term to be translated into the target ontology
Ext(C-Term) = extension underlying C-Term = relevant objects[9] (RelevantSet)

[9] This extensional information will be retrieved, stored and updated periodically by the system.

Expression = "lossy" translation of the term
Ext(Expression) = extension underlying Expression = retrieved objects (RetrievedSet)

Precision = *proportion of the retrieved objects that are relevant* = Probability(Relevant|Retrieved)
$= \frac{|RetrievedSet \cap RelevantSet|}{|RetrievedSet|} = \frac{|Ext(C-Term) \cap Ext(Expression)|}{|Ext(Expression)|}$
Recall = *proportion of relevant objects that are retrieved* = Probability(Retrieved|Relevant) =
$\frac{|RetrievedSet \cap RelevantSet|}{|RelevantSet|} = \frac{|Ext(C-Term) \cap Ext(Expression)|}{|Ext(C-Term)|}$

Based on the above we use a composite measure [20] which combines the precision and recall to estimate the loss of information. We seek to measure the extent to which the two sets do not match. This is denoted by the shaded area in Figure 1. The area is, in fact, the symmetric difference:
RelevantSet Δ *RetrievedSet* = *RelevantSet* $\cup$ *RetrievedSet* - *RelevantSet* $\cap$ *RetrievedSet*

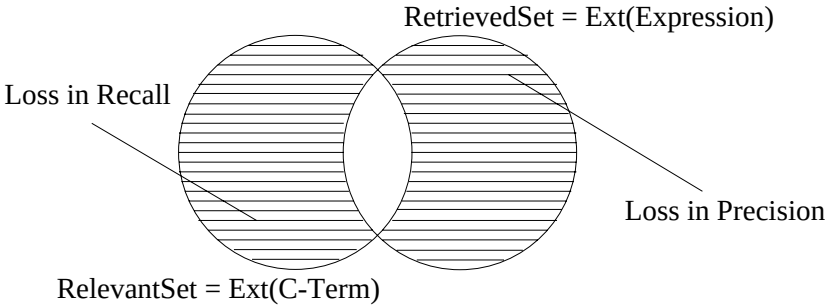

Figure 1: The mismatch between the RetrievedSet and Relevant Set

We are interested in the proportion (rather than the absolute number) of relevant and non-relevant objects retrieved, so a normalization of the measure gives: Loss = $\frac{|RelevantSet \Delta RetrievedSet|}{|RelevantSet|+|RetrievedSet|}$
In terms of precision and recall we have: Loss = 1 - $\frac{1}{\frac{1}{2}(\frac{1}{Precision})+\frac{1}{2}(\frac{1}{Recall})}$
In an open and dynamic environment, it is critical to satisfy the information needs of a widely varying cross-section of users. The users may have widely varying preferences when it is necessary to choose between precision and recall. We introduce a parameter α ($0 \leq \alpha \leq 1$) to capture the preference of the user where α denotes the importance attached by a user to precision. The modified composite measure may now be given as: Loss = 1 - $\frac{1}{\alpha(\frac{1}{Precision})+(1-\alpha)(\frac{1}{Recall})}$

3.2.3 Extensional Information vs. Semantic Relationships: Semantic adaptation for precision and recall measures

Techniques on estimating precision appear in Information Retrieval literature, but our work differs in the following important aspect: **we give higher priority to semantic relationships than those suggested by the underlying extensions**. Only when the semantics are not available, the system resorts to the use of extensional information. Since the system has translated a term from one ontology into an expression with terms from another different ontology with different underlying repositories, the extensional relationships may not reflect the semantic relationships. For instance a term in a user ontology which semantically[10] subsumes a term in the target ontology may have a smaller extension than the child term. This is reflected in the proposed measures.

[10]The interontology relationships used in integration of the ontologies are semantic and not extensional relationships.

We now enumerate the various cases that arise depending on the relationship between the conflicting term and its translation and present measures for estimating the information loss. We assume that a **Term** is translated into an **Expression** in the integrated ontology. The critical step here is to estimate the extension of **Expression** based on the extensions of the terms in the target ontology. Precision and recall are adapted as follows:

1. Precision and recall measures for the case where a term subsumes its translation. Semantically, we do not provide an answer irrelevant to the term, as Ext(Expression) $\subseteq$ Ext(Term) (by definition of subsumption). Thus, as Term subsumes Expression $\Rightarrow$ Ext(Term) $\cap$ Ext(Expression) = Ext(Expression). Therefore:
   Precision = 1,
   Recall = $\frac{|Ext(Term) \cap Ext(Expression)|}{|Ext(Term)|} = \frac{|Ext(Expression)|}{|Ext(Term)|}$
   Since terms in *Expression* and *Term* are from a different ontology, the extension of *Expression* can be bigger than the extension of *Term*, although *Term* subsumes *Expression* semantically. In this case we consider the extension of *Term* to be: |Ext(Term)| = |Ext(Term) $\cup$ Ext(Expression)|. Thus recall can be defined as:
   Recall.low = $\frac{|Ext(Expression)|.low}{|Ext(Expression)|.low + |Ext(Term)|}$, Recall.high = $\frac{|Ext(Expression)|.high}{max[|Ext(Expression)|.high, |Ext(Term)|]}$

2. Precision and recall measures for the case where a term is subsumed by its translation. Semantically, all elements of the term extension are returned, as Ext(Term) $\subseteq$ Ext(Expression) (by definition of subsumption). Thus, as Expression subsumes Term $\Rightarrow$ Ext(Term) $\cap$ Ext(Expression) = Ext(Term), The calculus of precision is like the one for recall in the previous case. Therefore:
   Recall = 1,
   Precision.low = $\frac{|Ext(Term)|}{|Ext(Expression)|.high + |Ext(Term)|}$, Precision.high = $\frac{|Ext(Term)|}{max[|Ext(Expression)|.low, |Ext(Term)|]}$

3. *Term* and *Expression* are not related by any subsumption relationship. The general case is applied directly since intersection cannot be simplified. In this case the interval describing the possible loss will be wider as *Term* and the *Expression* are not related semantically[11].
   Precision.low = 0, Precision.high = max[$\frac{min[|Ext(Term)|, |Ext(Expression)|.high]}{|Ext(Expression)|.high}$, $\frac{min[|Ext(Term)|, |Ext(Expression)|.low]}{|Ext(Expression)|.low}$]
   Recall.low = 0, Recall.high = $\frac{min[|Ext(Term)|, |Ext(Expression)|.high]}{|Ext(Term)|}$

Two special cases can arise in which the substitution of a term by an expression does not imply any loss:

1. Substituting a term by the intersection of its immediate parents implies no loss of information if it was defined as *exactly* its definition[12], i.e., the term and the intersection of its parents are semantically equivalent. For instance, in the example 'BOOK' was defined as exactly '(**AND** PUBLICATION (**ATLEAST** 1 ISBN))' and therefore the substitution of 'BOOK' by its immediate parents implies no loss.

2. Substituting a term by the union of its children implies no loss of information if there exists an extensional relationship saying that the term is covered extensionally by its children (total generalization).

[11] As we change in numerator and denominator we do not know which option is greater.
[12] In DL systems they are called *defined terms*.

3.3 Example of translation and measurement of the extensional loss

We now illustrate the computation of precision, recall and loss of information for plans 1, 2 and 4 presented in Section 2.4. The computation of the loss for plan 3 is similar to the one for plan 2. As the only conflicting term in the translation was 'BOOK' (the only one with no synonym), we explore the different translations for this term (no loss was incurred until replacing 'BOOK'). For the discussions, we assume α=0.5 (equal importance to precision and recall)[13] and the maximum loss allowed is 50%. The extensional values have been obtained from the underlying data repositories.

- Plan 1. The loss of information incurred on substitution of 'BOOK' by 'document' is as follows; it is an example of case 2 explained in Section 3.2.3 since 'BOOK' is subsumed by 'document'.
  |Ext(BOOK)|=1105, |Ext(document)|=24570
  Precision.low=$\frac{|Ext(BOOK)|}{|Ext(BOOK)|+|Ext(document)|}$=0.043, Precision.high=$\frac{|Ext(BOOK)|}{max[|Ext(BOOK)|,|Ext(document)|]}$=0.044,
  Recall=1,
  Loss.low=1-$\frac{1}{\frac{\alpha}{Precision.high}+\frac{(1-\alpha)}{Recall.high}}$=0.91571, Loss.high=1-$\frac{1}{\frac{\alpha}{Precision.low}+\frac{(1-\alpha)}{Recall.low}}$=0.91755

- Plan 2. The loss of information incurred on substitution of 'BOOK' by 'periodical-publication' is the following; it is an example of case 3 in Section 3.2.3 since 'BOOK' and 'periodical-publication' are not related (none of them subsumes each other).
  |Ext(BOOK)|=1105, |Ext(periodical-publication)|=34
  Precision.low = 0, Precision.high = max [$\frac{min[|Ext(Term)|,|Ext(Expression)|.high]}{|Ext(Expression)|.high}$, $\frac{min[|Ext(Term)|,|Ext(Expression)|.low]}{|Ext(Expression)|.low}$] = 1
  Recall.low = 0, Recall.high = $\frac{min[|Ext(Term)|,|Ext(Expression)|.low]}{|Ext(Term)|}$ = 0.03076
  Loss.low=1-$\frac{1}{\frac{\alpha}{Precision.high}+\frac{(1-\alpha)}{Recall.high}}$=0.94031, Loss.high=1-$\frac{1}{\frac{\alpha}{Precision.low}+\frac{(1-\alpha)}{Recall.low}}$=1

- Plan 4. The loss of information incurred by considering the children of 'BOOK' in the integrated ontology is as follows; 'BOOK' subsumes the union since subsumes each of them separately, although the extension of 'BOOK' (1105) is smaller than the extension of the union (between 14199 and 14237). It is an example of case 1 in Section 3.2.3.
  |Ext(BOOK)|=1105, |Ext(book)|=14199, |Ext(proceedings)|=6, |Ext(thesis)|=0,
  |Ext(misc-publication)|=31, |Ext(technical-report)|=1
  Ext-union.low=max[|Ext(book)|, |Ext(proceedings)|, ...]=14199, Ext-union.high=sum[|Ext(book)|, |Ext(proceedings)|, ...]=14237
  Ext-expr.low=$\frac{Ext-union.low}{|Ext(BOOK)|+Ext-union.low}$=0.92780, Ext-expr.high=$\frac{Ext-union.high}{|Ext(BOOK)|+Ext-union.high}$=0.92798,
  Precision=1
  Recall.low=$\frac{Ext-expr.low}{Ext-expr.low+|Ext(BOOK)|}$=0.92780, Recall.high=$\frac{Ext-expr.high}{max[|Ext(BOOK)|,Expr-ext.high]}$=[14] 1
  Loss.low=1-$\frac{1}{\frac{\alpha}{Precision.high}+\frac{(1-\alpha)}{Recall.high}}$=0, Loss.high=1-$\frac{1}{\frac{\alpha}{Precision.low}+\frac{(1-\alpha)}{Recall.low}}$=0.07220

The four possible plans and the respective losses for the user query '(**AND** BOOK (**FILLS** doc-author-name "Carl Sagan"))' are illustrated in Table 1. Only the fourth case results in the loss below the *user-max-loss* (50%) and is hence chosen. That means that the chosen translation into Stanford-I of the original user query '*[NAME PAGES] for (***AND** *BOOK (***FILLS** *CREATOR "Carl Sagan"))*' is '*[title number-of-pages] for (***AND** *UNION(book, proceedings, thesis, misc-publication, technical-report) (***FILLS** *doc-author-name "Carl Sagan"))*'. The answer does not contain incorrect data in the best case (which is possible) but around a 7% of the ideal answer may be missed (substituted by irrelevant data or not accessed) in the worst case.

4 Conclusions

As the Web becomes the predominant environment for more and more people to create applications, and export or share information, syntactic approaches for navigation and keyword based searches

[13] Calculation of loss is measured as a fraction but presented to the user as a percentages value.

[14] If the higher bound is 1 or the lower bound is 0 no new information has been obtained.

Plan	Loss Of Information
(**AND** document (**FILLS** doc-author-name "Carl Sagan"))	91.57%<loss<91.75%
(**AND** periodical-publication (**FILLS** doc-author-name "Carl Sagan"))	94.03%<loss<100%
(**AND** journal (**FILLS** doc-author-name "Carl Sagan"))	98.56%<loss<100%
(**AND** UNION(book, proceedings, thesis, misc-publication, technical-report) (**FILLS** doc-author-name "Carl Sagan"))	0%<loss<7.22%

Table 1: The various plans and the respective Loss Of Information

are becoming increasingly inadequate. We present a novel approach based on the use of multiple, possibly pre-existing, real world domain ontologies as views on the underlying data repositories. Thus an information request can now be expressed using terms from these ontologies and a system can now browse multiple domain ontologies as opposed to users browsing individual heterogeneous repositories or web pages correlated based on statistical information.

The main contribution of this paper is the characterization of the *loss of information* when a translation results in a change of semantics of the query. Measures to estimate loss of information based on terminological difference as well as on standard and well accepted measures such as *precision* and *recall* were also presented. As far as we know our work is the first one that deal with the problem of measuring the imprecision of answers in the context of managing multiple distributed and heterogeneous data repositories.

Approaches for modeling imprecision and measures for uncertain information have been proposed in the literature. The novelty of our approach is that we provide a set theoretic basis for an extensional measure of semantic information loss. The user is provided with a means to control the quality of information based on his preference of more precision or more recall, and the limit of the total loss incurred. On the other hand a qualitative description of information loss using intensional term descriptions is also presented and illustrated with the help of examples. Based on the estimates of information loss, the system chooses that translation which minimizes the loss of information. We thus establish vocabulary heterogeneity as the basis for identifying and measuring the quality of information, a very useful feature for information processing in open and dynamic environments.

The above ideas have been implemented as a part of the OBSERVER system which uses real world ontologies to provide access to real world heterogeneous data repositories (more information about OBSERVER can be found at http://siul02.si.ehu.es/~jirgbdat/OBSERVER/). We have synthesized techniques to estimate imprecision in answers based on vocabulary heterogeneity with internet computing to design and implement a system, thus demonstrating concrete progress in developing multiple ontology based access to information on the web.

References

[1] Y. Arens, C.A. Knoblock, and W. Shen. Query reformulation for dynamic information integration. *Journal of Intelligent Information Systems*, 6(2-3):99–130, 1996.

[2] A. Borgida, R.J. Brachman, D.L. McGuinness, and L.A. Resnick. CLASSIC: A structural data model for objects. In *Proceedings ACM SIGMOD-89, Portland, Oregon*, 1989.

[3] R. Brachman and J. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9:171–216, 1985.

[4] S. Chaudhuri. Generalization and a framework for Query Modification. In *Proceedings of the sixth International Conference on Data Engineering*, February 1990.

[5] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogeneous information sources. In *Proc. of the 10th IPSJ, Tokyo, Japan*, 1994.

[6] W. W. Chu, H. Yang, K. Chiang, M. Minock, G. Chow, and C. Larson. Cobase: A Scalable and Extensible Cooperative Information System. *Journal of Intelligent Information Systems*, 6(2-3), 1996.

[7] C. Collet, M. N. Huhns, and W. Shen. Resource integration using a large knowledge base in CARNOT. *IEEE Computer*, pages 55–62, December 1991.

[8] A. Goñi, J.M. Blanco, and A. Illarramendi. Connecting knowledge bases with databases: a complete mapping relation. In *Proc. of the 8th ERCIM Workshop. Trondheim, Norway*, 1995.

[9] A. Goñi, E. Mena, and A. Illarramendi. Querying heterogeneous and distributed data repositories using ontologies. In *Proceedings of the 7th European-Japanese Conference on Information Modelling and Knowledge Bases (IMKB'97), Toulouse (France)*, May 1997.

[10] 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.

[11] V. Kashyap and A. Sheth. Semantic and Schematic Simililarities between Databases Objects: A Context-based approach. *The VLDB Journal*, 5(4), December 1996.

[12] A.Y. Levy, D. Srivastava, and T. Kirk. Data model and query evaluation in global information systems. *Journal of Intelligent Information Systems*, 5(2):121–143, September 1995.

[13] E. Mena. http://siul02.si.ehu.es/~jirgbdat/OBSERVER/ontologies.html.

[14] E. Mena, V. Kashyap, A. Illarramendi, and A. Sheth. Domain specific ontologies for semantic information brokering on the global information infrastructure. In *Proc. of the International Conference on Formal Ontologies in Information Systems (FOIS'98). Trento (Italy)*, June 1998.

[15] E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies. In *Proc. of the First IFCIS International Conference on Cooperative Information Systems (CoopIS'96), Brussels (Belgium), June*. IEEE Computer Society Press, 1996.

[16] A. Motro. Multiplex: A formal model of multidatabases and its implementations. Technical report, Technical Report ISSE-TR-95-103, Department of Information and Software Systems Engineering, George Mason University, Fairfax, Virginia, March 1995.

[17] R. Bayardo, W. Bohrer, R. Brice, A. Cichocki, G. Fowler, A. Helai, V. Kashyap, T. Ksiezyk, G. Martin, M. Nodine, M. Rashid, M. Rusinkiewicz, R. Shea, C. Unnikrishnan, A. Unruh, and D. Woelk. Infosleuth: Semantic integration of information in open and dynamic environments. In *Proceedings of the 1997 ACM International Conference on the Management of Data (SIGMOD), Tucson, Arizona.*, May 1997.

[18] G. Salton. *Automatic text processing*. Addison-Wesley, 1989.

[19] A. Silberschatz and S. Zdonik. Database systems - breaking out of the box. *SIGMOD Record*, 26(3), September 1997.

[20] C. J. van Rijsbergern. Information retrieval. http://dcs.glasgow.ac.uk/Keith/Chapter.7/Ch.7.html.