# In-Context Peer-to-Peer Information Filtering on the Web

**Aris M. Ouksel**
The University of lllinois at Chicago,
Dept. of Information and Decision Sciences, Chicago,
IL 60607, USA. E-Mail: aris@uic.edu

## 1. Introduction

Significant advances have been achieved in system, syntactic and structural/schematic interoperability in a distributed network. Yet, meaningful exchange of information among autonomously designed and populated, dynamic, structured and semi-structured Heterogeneous Information Sources (HIS) remains a major challenge. A pure Peer-to-Peer architecture lends itself naturally to this problem as the information sources are totally autonomous and, practically, a-priori integration cannot be assumed. While ontologies may play a role in facilitating integration, they are not the panacea [3] advocated by many researchers. Rather, integration ought to be viewed as an emergent phenomenon constructed incrementally, and its state is dependent on the frequency and the quality of interactions between the peers and their subsequent negotiations and agreements to reach common interpretations within the context of a given task.

The sources of incompatibility generically called *semantic conflicts* include differences in structural representations of data, differences in data models, mismatched domains, different naming and formatting schemes, and different interpretations. The keyword approach of most current search engines is inadequate to deal with the misinterpretations resulting from these semantic conflicts. The difficulty is exacerbated by the fact that interpretations cannot be standardized. The rise of XML as a data interchange standard does not solve this problem as different users model the same data in different ways leading to syntactic/semantic heterogeneities. Unless semantic conflicts are reconciled within the context of a task or a user request, Internet services and users may not be able to capitalize on the full potential of the World Wide Web. The alternative is the retrieval of volumes of unfiltered and irrelevant information. Clearly, there is a symbiotic relationship between context on the one hand and content and relevancy on the other, as well other environmental aspects necessary for a comprehensive view of context.

Many semantic reconciliation techniques have been proposed to facilitate a meaningful exchange of information between HIS [summaries in 1, 2, 3]. One promising semantic reconciliation approach is our SCOPES (Semantic Coordinator over Parallel Exploration Spaces), which mimics and formalizes the semantic reconciliation approach used by a typical human integrator. SCOPES integrates several automated and semi-automated techniques [4, and more recently 19, 20, 21] to facilitate an incremental construction of the knowledge, referred heretofore as *context*, necessary to translate a query posed against a local database into an equivalent one against a remote database. Context has been mainly considered in Artificial Intelligence research areas, including Natural Language [8], Computational Linguistics, Categorization [6], Knowledge Representation and Reasoning [10, 12], and also in information retrieval [9, 11]. The modeling of context and its represent have been receiving increasing attention in recent years [7]. But what exactly is context? It has been used to describe a multitude of things from descriptions, explanations and analysis.

In the sequel, we shall motivate our approach to context in SCOPES in section 2, and show that it satisfies the characteristics of a pure Peer-to-peer architecture with emergent semantics. As context is viewed as a discovery problem, we present an analysis of the complexity of the Context Discovery (CD) problem and then describe the SCOPES CD algorithm. We then provide a brief discussion of heuristics, based on results from conceptual structures [3].

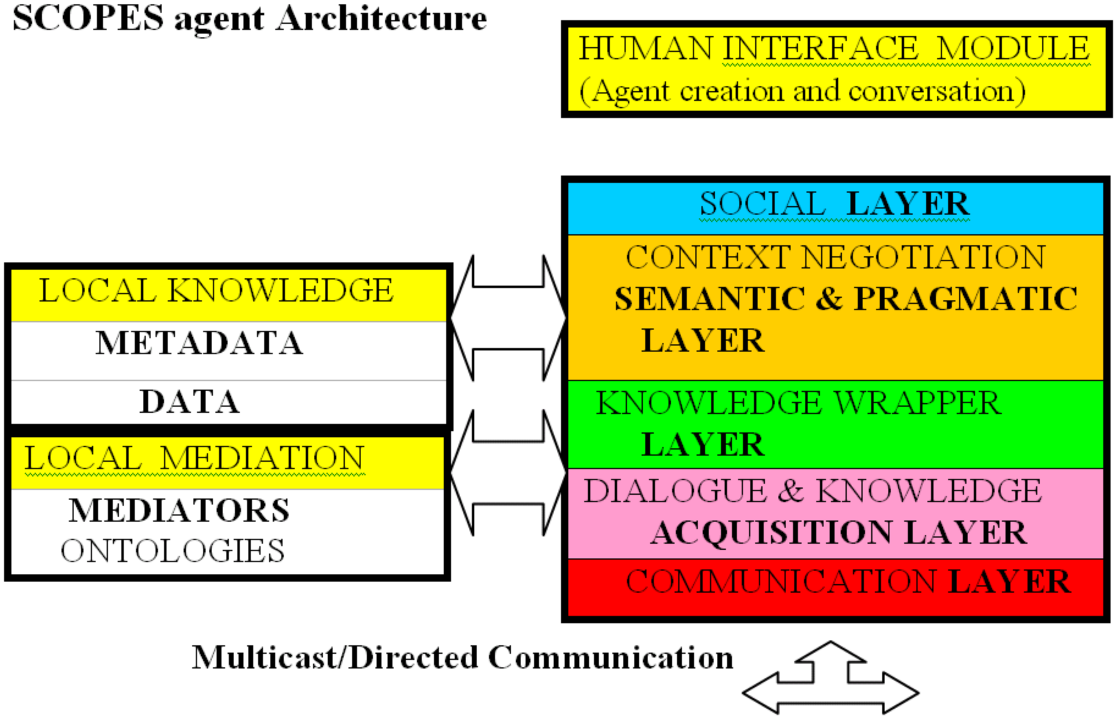
## 2. Context Construction

In theory, context in SCOPES is conceived to be constructed partly on the basis of mutually accepted propositions (beliefs). These mutual beliefs (MBs) are expected to bear on establishing shared ontologies and contribute to delineating emerging communities of interest on the Web. Before any domain-specific collaboration between communities of interest can occur, they must identify themselves to each other; mutual belief is expected to be an important diagnostic for defining communities of interest. While the metaphor of constructing a context appropriately connotes activity [17], we supplement that with another one connoting an even more dynamic development: agents negotiate contexts. Negotiation recognizes that contexts dynamically evolve as agents learn more about each other and as interests broaden or become more focused. The current SCOPES implementation reflects this fundamental concept within the semantic approach. Negotiation of context in a query-based framework is expected to favor solutions to problems of identification of domains for collaboration. One important question in a query-based program is how to structure queries for efficient construction of domains of mutual interest by way of cooperative negotiation. Another issue concerns the presuppositions of negotiation and the conditions that must be established to reasonably expect favorable outcomes. To what extent can mutual belief (MB) bear the entire burden of constructing or negotiating context? Other pragmatic elements beside (MBs) are expected to play a prominent role in establishing appropriate contexts for specific investigations. Clearly, pragmatic parameters such as domain specification (among others) will need to be determined if ambiguities are to be avoided or resolved. Related issues pertain to alternative approaches in pragmatic theory in which (MBs) plays a prominent role, principally that of P.Grice [18] (Cooperative Principle and Maxims) and that of D. Sperber and D. Wilson [11] (Relevance Theory). These theories, which depend on and also augment (MBs), introduce fundamental concepts of communicative cooperation and relevance necessary for understanding communication and inference. At the same time they offer guidelines to afford simplifications of semantic theory. SCOPES offers a useful framework in terms of which these concepts can be systematically investigated and developed. SCOPES integrates and goes well beyond the vision articulated recently in [20].

Practically, context is defined as inter-schema mappings between the schema of the local database and that of the remote database, consistent as a whole and with respect to the original query. This approach satisfies two important characteristics the pure Peer-to-Peer aspect of this approach; namely, the **autonomy** and the **dyadic interaction** characteristics. Unlike many previous methods, semantics are seen in SCOPES as a matter of continuing negotiation and evolution in the presence of inconsistent, uncertain and incomplete information. Thus, the common semantics between peers are fundamentally **emergent** from their interactions. The dyadic interaction approach supports the **scalability** property of pure Peer-to-Peer architecture. As a consequence, it is up to each peer to determine the commonality of knowledge with the other peers by factorizing the dyadic interactions, giving rise to multiple ontologies at different levels of granularity. Further, the participation of peer is not only **optional** but also **dynamic**. Several systems have been developed recently which partially satisfy these properties [22,23].

**2.1 A Semiotics Framework for Emergent Semantics:** The construction of context in SCOPES involves several abstraction layers other than semantics. These layers are best captured within a semiotics framework. Peirce [15] founded

**Figure 1. A SCOPES agent Architecture**



semiotics as a formal theory of sign. It consists of five distinct layers: Physical, syntactics, semantics, and pragmatics, social. The social layer is added to provide a more comprehensive picture of the issues to examine in social interaction and knowledge networks [16]. According to Andersen [13], the semiotics approach to computing emphasizes the importance of integrating computers in social reality. We adapt the semiotics framework to our problem and the resulting framework is shown in Figure 1. The semiotics framework provides an extensive blueprint which provides guidelines to our research in extracting knowledge from heterogeneous knowledge sources. It allows modular development of SCOPES as our understanding of the issues relevant within this framework deepens. The architecture of a SCOPES agent is illustrated in the Figure 1 below. The *knowledge acquisition and dialogue layer* handles conversations with outside agents and acquires information relevant to the query, called **anchors** in SCOPES, possibly utilizing available ontologies and/or mediators. It includes the negotiated policies and constraints established with remote sources. It is also responsible for answering requests from other SCOPES agents in the network. The *knowledge wrapper layer* services requests from other SCOPES agents on the network about the local knowledge source (book, database, a human agent, etc...) in an agreed upon format More details are given in [14]. The context negotiation layer will be discussed in the next section.

Context construction requires several mechanisms to allow realistic interaction and structured acquisition of knowledge and reconciliation of conflicts. The layered design of the agent architecture allows a clear separation of the mechanisms and a modular construction. We identified and partially implemented [14] four mechanisms essential to context construction. We list them here without explanation: i) Designing rules of interaction or Semantic Cooperation Protocols; (ii) Reengineering pragmatics, semantics, and syntactics; (iii) Handling approximation; (iv) Coming to agreement or Negotiation Protocols.

**2.2 Inter-Schema Correspondence Assertions:** Without loss of generality the semantic reconciliation process assumes the existence of an object-oriented schema describing structured or semi-structured information sources. Our classification of semantic conflicts [1] classifies conflicts along three dimensions namely, *naming*, *abstraction,* and *levels of heterogeneity.* The Inter-Schema Correspondence Assertion (lSCA) which represents the semantic relationship between two elements or concepts of two different database schemas has the general form below, and assumes the existence of morphisms between the schemas:

**Assert** [naming, abstraction, heterogeneity] (1)

Where *naming (abstraction)* stands for a naming (abstraction) mapping between an element *x* in the local database and an element *y* in the remote database; heterogeneity indicates the structural schema description of *x* and *y* in their respective databases. This classification combines the dimensions of semantic conflicts with a structural description, thereby facilitating the process of operational integration. Along the naming dimension, the relationships between two elements *x* and *y* can be categorized as *synonyms,* denoted *syn(x,y),* which are terms having similar meaning; *homonyms*, denoted *hom(x,y),* which are similar terms representing different concepts; and *unrelated,* denoted *unrel(x,y)* which are not related along the dimension of naming, however these could be related in some other way such as functional relationships. Along the dimension of abstraction, the relationships between two elements *x* and *y* can be categorized as *class* relationship, denoted *class(x,y)*; *generalization*/*specialization* relationship, denoted *gen(x,y); aggregation* relationships, denoted *agg(x,y);* and relationships due to *computed or derived functions,* denoted *function-name(x,y).*

The heterogeneity dimension includes the *object* level, the *attribute* level and the *instance* level of the database schema. Semantic conflicts due to naming and abstraction can occur at any of these levels. It requires a pair of values, one for each element *x* and *y,* as represented in its corresponding schema. Each value is denoted either *att(x,O,DB),* where *x* is the element considered in the assertion, *O* the object to which it is attribute, *DB* the database in which it appears; or *obj(x,DB),* where *x* is an object in *DB; inst(x,O,DB),* where is an instance of object *0* in *DB.*

One important advantage of this classification is the partitioning of semantic conflicts into 12 disjoint classes based on naming and abstraction. Some of these classes are transient in that the classification is incomplete due to lack of evidence, whereas the other classes may occur in both static and transient. In [1], the classification is analyzed and shown to capture the fundamental semantic conflicts identified in the literature on heterogeneous conceptual schemas.

**2.3 Inference Engine:** In SCOPES evidence to support assertions may be obtained using a variety of knowledge sources including ontologies [3], lexicons, reconciliation techniques, general or domain specific knowledge repositories, metadata specifications, general rules derived from conceptual structures. The strength of this evidence is determined by its source. The reconciliation techniques and knowledge sources are coordinated using the following simple interface template:

r: **If** C(*p*) **Then** consequent. [*p.q*]

where C::=E /Assertion /Assumption/E and C /Assertion and C/Assumption and C, and C may be quantified over domains of variables. Thus C is a complex typed predicate expression constructed from either a directly elicited piece of evidence E, available knowledge, a reasonable assumption, or a combination thereof. In the above template "consequent" is a disjunction of assertions about two objects O1 and O2, *p* represents the degree of belief in all the assertions in C, and *q* the belief in rule r if *p* = 1. The measures of belief are a recognition that context construction is emergent since supporting schematic evidence is gathered incrementally during exchanges between two peers.

## 3. Context Discovery (CD) Problem

SCOPES is designed and implemented as a semantic reconciliation system, which assists the human integrator in discovering the context. The **Context Discovery (CD) Problem** can be stated informally as follows: Given a query **Q**, expressed in SQL or Xpath, against a local database with schema S. Find one or more sets of consistent inter-schema mappings within which **Q** may be translated into an equivalent query **Q'** against a remote database with schema S'. A set of consistent mappings is said to represent a *satisficing* interpretation if it is consistent with the schematic values of the data given in the query. Note that this set may be incomplete.

**3.1 Complexity:** For each term in a local query, a typical human integrator first tries to establish matching terms (or anchors) in the remote database, exploiting lexical ontologies like *WordNet* or other domain ontologies and available initial similarity values. Each term may have several anchors. Let *q* be the number of terms in a query, $T_{local} = \{t_1, t_2, t_3 \dots t_q\}$, and *r* matching terms $T_{remote} = \{t'_1, t'_2, t'_3, \dots t'_r\}$ in the remote database. Assume that each term in $T_{local}$ maps to each of the *r* terms in $T_{remote}$ with some probability (or a similarity value), thus forming *r* anchors for each of the query terms. An initial reconciliation of $T_{local}$ may randomly select one anchor for each term in $T_{local}$. For example given $T_{local} = \{t_1,t_2,t_3\}$ and $T_{remote} = \{t'_1,t'_2,t'_3,t'_4\}$. Let $A^u = \{(t_1,t'_4), (t_2,t'_3),(t_3,t'_2)\}$ be the set of anchors considered initially. If this reconciliation fails, another set may be randomly selected to continue reconciliation.

Using the classification of semantic conflicts [1], let the sets of ISCAs corresponding to anchors $(t_1,t'_4)$, $(t_2,t'_3)$, and $(t_3,t'_2)$ be sets $\mathbf{ISCA}_{(t1,\, t'4)} = \{a_1, a_2,\dots a_{12}\}$, $\mathbf{ISCA}_{(t2,\, t'3)} = \{b_1, b_2,\dots b_{12}\}$ and $\mathbf{ISCA}_{(t3,\, t'2)} = \{c_1, c_2,\dots c_{12}\}$.Each element of $\mathbf{ISCA}_{(t1,\, t'4)}$, $\mathbf{ISCA}_{(t2,\, t'3)}$ and $\mathbf{ISCA}_{(t3,\, t'2)}$, is of the form in (1). Without any additional semantic knowledge from the remote database, any of the ISCAs for each anchor is plausible unless refuted by contradictory evidence. A ***context*** in this case is constructed by selecting one one ISCA each of $\mathbf{ISCA}_{(t1,\, t'4)}$, $\mathbf{ISCA}_{(t2,\, t'3)}$ and $\mathbf{ISCA}_{(t3,\, t'2)}$ such that the ISCAs together are consistent. This forms a *consistent* (or non-contradictory) and satisficing interpretation for the query. In the absence of complete knowledge, each combination set resulting from the Cartesian product of sets $\mathbf{ISCA}_{(t1,\, t'4)}$, $\mathbf{ISCA}_{(t2,\, t'3)}$ and $\mathbf{ISCA}_{(t3,\, t'2)}$ represents one plausible set of assertions. For example the combination set $\{a_1, b_2, c_9\}$ represents a plausible set of assertions. However not all of these combination sets may be consistent (or non-contradictory) with respect to the assertions contained within. Theoretically in the worst case scenario the total number of sets of plausible inter-schema correspondence assertions which may be examined is: $(12r)^q$. The problem of determining the TRUE context given the above state search space is NP hard. This provides the motivation to explore general rules and heuristics. The CD algorithm, partitions the state search space by utilizing the best-first search approach, exploring the most promising context(s) using available heuristicsavailable heuristic information. The SCOPES prototype, implemented using Multi-Agent Systems [4] approach, is based on the following algorithm:

```
While reconciliation is not terminated
 {For each query term and for the anchor
  with the next highest similarity value
     Assert[(x,y),naming,abstraction, hetrogeneity] ;
  For each anchor selected above, infer ISCAs
  based on available knowledge
  (Metadata, data, etc …);
  Partition the ISCAs into consistent subsets;
  Apply heuristics to prune search space;
  If TRUE context is not found
  then Re_Validate Anchors ( find anchors with
  the next highest similarity value)
  else Terminate reconciliation}
```

## 4. Search Space Pruning Methods

A list of general rules used by the inference engine in SCOPES is specified in [1], which are used to prune the search space. For each term in $T_{local}$ the anchor with the highest similarity may be considered in the first round of reconciliation. Instead of considering all twelve plausible ISCAs for each anchor, the set is split in two. The *un-retracted* set comprises all ISCAs that are synonyms on *naming*, and the *retracted* space, those ISCAs that are either homonyms or unrelated. The latter set is maintained for the backtracking purposes in the event an initial assumption of homonymy (or unrelatedness) is later overturned. Thus SCOPES considers combination sets generated from only those ISCAs in the initial un-retracted space of anchors $A^u$. If the resulting set is unacceptable to the user, the process continues with only those ISCAs acceptable to the user. Using this approach, the search space is reduced to $(4r)^q$.

According to the information modeling principles a database schema is derived from conceptual modeling techniques and embodies an inherent structure, which represents various conceptual relationships among schema contents [4]. A schema consists of objects, their properties, domain values, and cardinality relationships between different objects. Key- or non-key properties of an object are also identified in the schema, and relationships such as generalization/ specialization, aggregation and functional relationships may be derived. Heuristics based on this knowledge are crucial in pruning the search space. Below we present two examples without details:

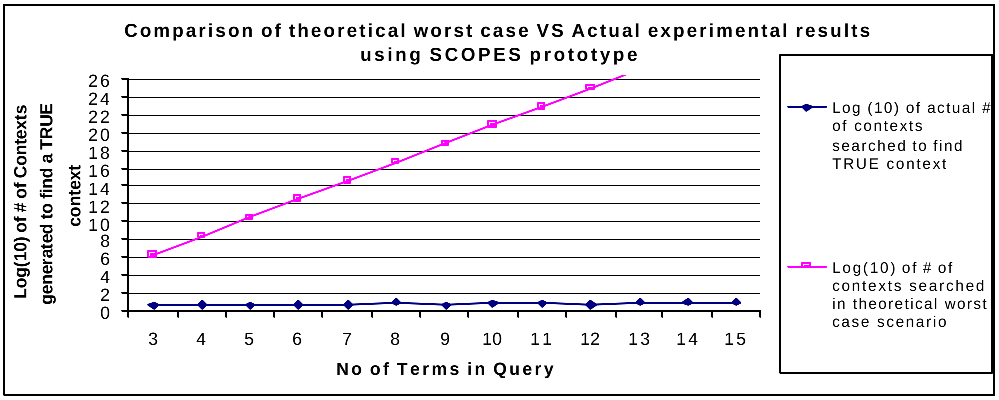
**Heuristic 1:** *Let $O_1$ be the single object in a query represented by $T_{local}$. Let P denote the number of terms in $T_{local}$ including object $O_1$, its key-attribute(s), and the value(s) of the key-attribute(s) specified in the query. Then in the worst case scenario the overall search space is no larger than:*

$$4^{[q-(P-1)]}$$

Let $T_{local}$ = {$t_1$, $t_2$, $t_{3,}$ $t_4$, $t_{5,}$} and $T_{remote}$ = {$t'_1$, $t'_2$, $t'_3$, $t'_4$, $t'_5$ }. Assume that terms $t_1$, $t_{2,}$ $t_3$ represent an object, its key-attribute, and the value of the key-attribute; and terms $t_4$, $t_5$ are non-key attributes of $t_1$. Let $A^u$ ={($t_1$,$t'_3$), ($t_2$,$t'_1$),($t_3$,$t'_4$), ($t_4$,$t'_2$), ($t_5$,$t'_5$)}. SCOPES considers semantics interpretation simultaneously across all terms to ensure consistency and to avoid contradictory interpretations. Any semantic relationship that holds between $t_1$, $t_2$ and $t_3$ is also very likely to hold between $t'_3$, $t'_1$ and $t'_4$, especially if the concept represented by term $t'_1$ uniquely identifies the concept represented by term $t'_3$. It follows that any semantic relationship, such as synonymy and generalization that holds between $t_1$ and $t'_3$ must also hold between $t_2$ and $t'_1$ and between $t_3$ and $t'_4$. Hence it is unlikely to have an ISCA representing a synonymy and aggregation relationship between $t_1$ and $t'_3$, an ISCA representing a synonym and generalization relationship between $t_2$ and $t'_{1,}$ and an ISCA representing synonym and computed-function relationship between $t_3$ and $t'_4$ as part of the context. SCOPES therefore does not generate any combination sets, which contain contradictory or inconsistent knowledge with respect to pairs ($t_1$,$t'_1$), ($t_2$,$t'_2$) and ($t_3$,$t'_3$). Hence in this example q=5, P = 3, the worst case is reduced to $(4)^{[5-(3-1)]}$ = 64.

**Heuristic 2:** *Let $O_1$, $O_2$ … $O_n$, be 'n' objects in query $T_{local}$. Let $P_i$ (for i from 1 to n) denote the number of terms in $T_{local}$ corresponding to object $O_i$, its key-attribute(s), and the value(s) of the key-attribute(s). Let M be the number of terms in $T_{local}$ that are non-key attributes of objects $O_1$, $O_2$ … $O_n$ and have their value(s) in $T_{local}$. Let K be the number of terms in $T_{local}$ that are domain value(s) of these non-key attribute(s) in $T_{local}$. If K > M, then the worst case reduces to:* $4^{[q-\{(P_1-1)+(P_2-1)+...+(P_n-1)+(M)+(K-M)\}]}$ *Else to:* $4^{[q-\{(P_1-1)+(P_2-1)+...+(P_n-1)+(M)\}]}$

**Empirical Results and Analysis:** The SCOPES prototype was developed to reconcile semantic conflicts between several pairs of heterogeneous databases from different organizations in health care. Queries ranging from three up to fifteen terms were posed. The more schematically inter-related



are the terms in the query, the better is the expected performance. The figure below depicts some of the results.

## References

1. C. Naiman, and A.M. Ouksel, "A Classification of Semantic Conflicts," *Journal of Organizational Computing*. 5(2), pp. 167-193, 1995.
2. A. M. Ouksel, C. Naiman, "Coordinating Context Building in Heterogeneous Information Systems," *Journal of Intelligent Information Systems* 3, pp. 151-183 , 1994
3. A. M. Ouksel and I. Ahmed, "Ontologies are not the Panacea in Data Integration: A Flexible Coordinator to Mediate Context Construction," *International Journal of Distributed and Parallel Databases*, January 1999.
4. J. Sowa, "Conceptual Structures: Information Processing in Mind and Machine," *The Systems Programming Series*, Addision Wesley, 1984.
5. Baker & D. Chuhan: JAFMAS: http://www.ececs.uc.edu/~abaker/JAFMAS/
6. J. Barwise and Perry J, "The rights and wrongs of natural regularity". In *Philosophical Perspectives 8: Logic and Language*. (Ed. J. Tomberlin). (Ridgeview Press: Atascadero, CA), 1992.
7. P. Brezillon, "Context in Artificial Intelligence: I. A survey of the literature," *Computer & Artificial Intelligence* 18, 321-340, 1999.
8. H. Clark and TB. Carlson, "Context Comprehension. In *Attention and Performance IX* (Eds J. Long and A. Baddeley). (Lawrence Erlbaum Associates. Hillsdale, NJ.). 1981.
9. S. P. Harther, "Psychological relevance and information science," *Journal of the American Society for Information Science* 43, 602-615, 1992.
10. J. McCarthy, "Generality in Artificial Intelligence," *Communications of the ACM* 30, 1030-1035, 1987.
11. D. Sperber and D. Wilson, "Relevance: Communication and Cognition," (Basil Blackwell: Oxford, UK). 1986.
12. Y. Shoham, "Varieties of Context". In *Artificial Intelligence and Mathematical Theory of Computation: Papers in Honor of John McCarthy*. (Ed. V. Lifshitz), Academic Press: Boston, MA, pp.393-408, 1991.
13. P. Andersen. "A Theory of Computer Semiotics," In *Semiotics Approaches to Construction Assessment of Computer Systems*. Cambridge University Press, Cambridge, Mass., 1990.
14. A. M. Ouksel, "SCOPES: Technical Specifications. CISORS Lab" *Working Paper* 1, 1998.
15. C. Peirce, "Collected Papers of Ch. S. Peirce". Hartshorne and WeiSS (Eds.), Cambridge, MA, 1990.
16. Y. Shoham et al. "On social laws for artificial intelligence societies: Off-line design," *AI Journal* 1993.
17. L. Wittgenstein, "Philosophical Investigations," *Macmillan Publishing*, New York, NY 1953.
18. P. Grice, "Studies in the way of words," *Harvard University Press*, Cambridge (MA), 1989.
19. K. Aberer, P. Cudre-Mauroux, and M. Hauswirth, "*The Chatty Web: Emergent Semantics Through Gossiping*," Proc. Of the 12th International World Wide Web Conf., 2003.
20. P. A. Bernstein, F. Giunchiglia, A. Kementsietsidis, J. Mylopoulous, L. Serafini, and I. Zaihrayeu, "*Data management for peer-to-peer computing: A vision*," Workshop on the Web and Databases (WebDB), 2002.
21. A. Kementsietsidis, M. Arenas, R. J. Miller, "*Mapping Data in Peer-to-Peer Systems: Semantics and Algorithmic Issues*," SIGMOD 2003.
22. A. Halevy, et al., "*Schema Mediation in Peer Data Management Systems*," ICDE, 2003.
23. S. Bergamaschi, F. Guerra, "*Peer to Peer Paradigm for a Semantic Search Engine*," Proc. of the Intern. Workshop on Agents and Peer-to-Peer Computing, Bologna, 15 July 2002, LNCS 2530, Springer.