Management of Unspecified Semi-Structured Data in Multi-Agent Environment

Yosi Ben-Asher
Computer Science Department
University of Haifa, Israel
yosi@cs.haifa.ac.il

Shlomo Berkovsky
Computer Science Department
University of Haifa, Israel
slavax@cs.haifa.ac.il

Yaniv Eytani
Computer Science Department
University of Haifa, Israel
ieytani@cs.haifa.ac.il

ABSTRACT
Amounts of available heterogeneous semi-structured data grow rapidly on the Web and other data repositories. This raises the need to provide simple and universal ways to access this data. To provide such an interface, we propose to exploit the notion of "unspecified ontologies", describing the data objects as a list of attributes and their respective values. In order to facilitate an efficient management of the unspecified data objects we use a multi-agent channeled multicast communication platform. The data objects are stored distributively, such that each attribute is assigned a designated channel. This allows performing efficient searches by parallel querying of the relevant channels only, and aggregating the partial results. Moreover, the multi-agent platform facilitates advanced data management through extracting meta-data from the data objects. We implemented a prototype system and experimented with a corpus of real-life E-Commerce advertisements. Our results demonstrate scalability of the proposed approach and the accuracy of the extracted meta-data.

General Terms
Algorithms, Measurement, Performance, Experimentation

Keywords
Data Management, Meta-data Extraction, Multi-Agent Systems.

1. Introduction
Nowadays, the amount of available semi-structured data, which is naturally highly dynamic, distributed among multiple repositories, and represented heterogeneously, grows rapidly. This aggravates the users in finding and accessing the relevant data objects and emerges a need to devise a universal mechanism to access the data. Uniform data access mechanism will provide the users a generic access interface and will free them from the limiting requirement of having a-priory semantic data about the way the data objects are structured (referred as schema or ontology [10]).

The issue of accessing heterogeneous semantically enriched data has been addressed from different angles. Mariposa [14] implemented distributed data sharing over heterogeneous resources (but not ontologies). Research works in the Semantic Web [6] community focus on highly expressive knowledge representation languages, facilitating creation of a universal ontology. Distributed query processing assuming a known global ontology has been intensively studied [13]. However, none of these techniques addresses the key challenge in accessing heterogeneous data: the large number and the dynamicity of distinct ontologies

Alternatively, research efforts in the Data Integration domain aim at resolving this restriction by merging different ontologies and generating a global centralized ontology to allow data access in a uniform manner. However, a central ontology must be shared by all the repositories and users accessing the data objects. This may well obstruct it from being expanded in short-time intervals. In addition, since data providers could potentially violate the mappings to the global ontology by significantly changing their local ontologies, updates of the global ontology should be done by a central administration point. As a result, data integration systems provide limited support for large-scale data sharing.

This work proposes use an alternative approach of "UNSpecified Ontologies" (UNSO) for accessing the data objects [2]. UNSO assumes that the ontologies are not fully defined, leaving parts of it to be dynamically specified by the data providers. Thus, instead of basing the data description on a set of a-priori known predefined ontologies, providing an explicit formal specification of the data, the data objects are described in a relatively flexible form of an unspecified list of attributes and their values.

The contributions this work are two fold. First, the proposed approach, stores the data objects in a distributed, multi-agent communication platform, where each unspecified attribute is assigned a single logical communication channel. This allows to pose search queries against a dynamically evolving mechanism capturing the descriptions of the data objects. The queries are routed to the relevant channels only, and the query result is aggregated from the partial results received from the channels.

In addition, implementing the proposed mechanism over channeled multicast communication platform facilitates extraction of domain meta-data. This functionality is achieved through autonomously collecting statistic properties of the unspecified descriptions of the data objects. Note that the extracted meta-data reflects the dynamicity of the data objects descriptions and it can be exploited for advanced data management functionalities, such as descriptions' correction, pro-active assistance and others.

We implement a prototype system using LoudVoice [4], a multi-agent platform based on the channeled multicast communication model. Multi-agent environment facilitates a decentralized sharing and administration of the data. Every agent is capable of providing new data, relating it to the existing semantic concepts, and defining new concepts that for further use by other agents.

SAC'06, April, 23-27, 2006, Dijon, France.

Experiments were conducted over a set of real-life E-Commerce advertisements (ads) from different application domains. The experiments demonstrate efficient data management and search capabilities, and high scalability of the proposed approach. We also show successful extraction of accurate domain ontologies basing on the descriptions of the data objects.

The rest of the paper is organized as follows. Section 2 briefly describes the basics of unspecified data management. Section 3 elaborates on the unspecified data management and access over multi-agent multicast communication platform and details the meta-data extraction process. Section 4 presents our experimental results. Section 5 surveys prior studies on data management in multi-agent environments. Finally, section 6 presents our conclusions and discusses future research directions.

2. Unspecified Data Management

A key concept in semantic data management is ontology, i.e., a formal shared conceptualization of a particular domain of interest [10]. It is a standardized reference model, providing both human-understandable and machine-processable semantic mechanism, allowing enterprises and application systems to collaborate efficiently.

For example, consider simple predefined ontology for MS-Office documents, described by a list containing three attributes [*size* | *application* | *modified*]. Let us consider each attribute having the following range of predefined values [*{below 100K, 100k–1M, above 1M}* | *{Word, Excel, Powerpoint}* | *{last_week, last_month, last_year}*]. However, predefined ontology requires centralized management, obstructing dynamic changes and updates, and does not allow objects descriptions, mentioning attributes or features that are not anticipated.

Conversely, UNSO [2] proposes a flexible way overcoming the restrictions associated with sharing of a global, a-priori defined ontology. UNSO assumes that the ontology is not fully specified and that parts of it can be dynamically specified in the data objects descriptions. It allows the users to provide relatively free semi-structured description of data objects as a list of <*attribute:value*> pairs. Thus, in UNSO, the data objects are represented by a list of attributes, where the range of the values for each attribute corresponds its possible values. Although the description of data objects with attribute-value pairs may not be applicable for complex objects, it is sufficient for description of simple objects, e.g., files, computing resources, ads, and so on

Clearly, different UNSO users might provide different syntactic descriptions for the same data objects. To eliminate the ambiguity and to enhance the precision, UNSO incorporates WordNet [8] that performs simple semantic standardization of the data objects descriptions.

2.1 Mapping UNSO to Implicit Organizations

To facilitate efficient data management of objects represented by a list of <*attribute:value*> pairs, we use the idea of logical *implicit organizations* [5]. An implicit organization is a group of entities playing the same role and willing to coordinate their actions for service delivering. The term implicit stresses the fact that there is no need for an explicit group formation phase as joining the organization is the matter of sharing the same functionality with other members of the organization.

In our case, implicit organizations reflect the attributes mentioned in the data objects descriptions, such that each attribute is assigned a single implicit organization. The resulted set of implicit organizations facilitates dynamic management and access to the underlying data objects. However, neither the attributes nor their values are anticipated by any predefined ontology and the users can independently insert descriptions of the data objects in a flexible and decentralized way.

3. UNSO over Multi-Agent Platform

In our system, the descriptions of the existing data objects are represented as a list of <*attribute:value*> pairs. The descriptions are partitioned between different agents, mimicking a real-life situation, where each agent is responsible for a certain set of tasks. We assume that both the descriptions of the data objects can be distinctively identified. Thus every description is assigned a unique $obectj_{id}$ and every agent is assigned a unique $agent_{id}$.

Every unique attribute mentioned in the data objects descriptions is assigned a single implicit organization. Thus, each agent is connected only to a subset of organizations, matching the attributes mentioned in the descriptions stored by the agent. The above mapping of data objects descriptions to the implicit organizations through the agents is illustrated in figure *1*.

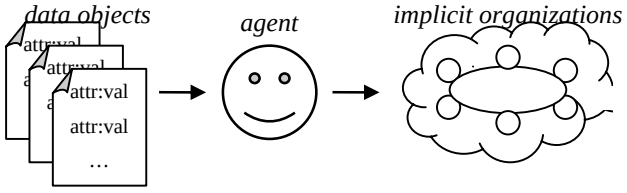

Figure 1. Mapping of the data objects descriptions to the implicit organizations.

The chosen communication platform, LoudVoice [4], has been designed to support implicit organization of agents. Each implicit organization is represented in LoudVoice by a communication channel and the agents join the organization through connecting the relevant channel. As the number of available communications channels is limited, and it is lower than the number of different attributes in data objects descriptions (for large corpora), the attributes are mapped to LoudVoice channels using hashing mechanism. For example, consider the following unspecified description of MS-Office document [*author: John* | *size:23.45Kb* | *type:PPT*]. As a result, the agent storing the description is connected to channels *hash(author), hash(size)* and *hash(type)*.

Note that hashing mechanism inserts some extent of uncertainty, as hashing collisions may cause different attributes attr1 and attr2 to be mapped to a single LoudVoice channel. The following subsection dealing with the search of data objects describes a way to overcome this issue.

In addition to the implicit organization, LoudVoice inherently supports the notion of channeled multicast. Messages are sent on a channel and received by all agents that are connected to it. Channeled multicast reduces the amount of communication needed when more than two agents are involved in a task, and allows overhearing, i.e., the ability to listen to messages addressed to others. Overhearing, in turn, enables functionalities such as the meta-data extraction, pro-active assistance, and monitoring without interfering with the existing communication protocols.

Except the agents storing the data objects containing the relevant attribute, one arbitrary agent on each LoudVoice channel serves as a *mediating agents* vis-a-vis the other channels. These agents are connected both to their original channel and to the inter-organization communication channel. The mediating agents transfer the extracted meta-data between different channels and facilitate advanced data management functionalities (will be described later). For example, consider the structure of two LoudVoice channels "channel A" and "channel B", and their respective mediating agents, as illustrated in figure *2*.

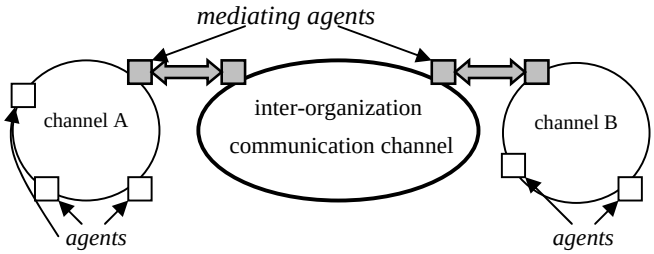

Figure 2. Organization of LoudVoice channels.

To address the problem that object descriptions may contain different syntactic terms with the same semantic meaning (synonyms), <*attribute:value*> pairs mentioned in the descriptions are standardized using WordNet [8]. In WordNet, English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. For each concept, the set of synonyms can be sorted according to the frequency of usage. To eliminate possible ambiguity and improve the precision, the terms mentioned by the users in the unspecified data objects descriptions, undergo semantic standardization. It is done by substituting the original mentioned terms with its most frequent synonyms. Thus, similar, but not identical terms are replaced by a single 'representative' term.

3.1 Management and Search

One of the functionalities the system provides is efficient storage, management and search of the data objects. As described earlier, the agents storing the data objects descriptions connect only to the channels representing the attributes mentioned in the stored descriptions. Due to the overhearing of LoudVoice channels, the agents are able to receive all the messages sent over a channel, and they autonomously decide whether to respond to the received message.

We assume that the search queries are also structured as a list of <*attribute:value*> pairs. The search query is transmitted as follows: (1) a user defines the search query in form of unspecified list, (2) the agent of the user initiating the search query, opens a temporary connection to the relevant channels matching the mentioned attributes, (3) the agent launches queries for single <*attribute:value*> pairs over the relevant channels, (4) only the agents that are connected to the channels (and store descriptions of the data objects containing this attribute) receive the query.

Upon receiving the query, each agent autonomously decides whether the stored descriptions satisfy the requirement posed by the query. This is done by comparing the *value* received in the query with the values stored in the descriptions of the data objects. Only the agents, satisfying the precise *value* requirement of the relevant *attribute*, respond the query.

The response message to the query is formulated as <$agent_{id}$, {$object_{id}$}>, where $agent_{id}$ is the unique identity of the agent, and {$obectj_{id}$} is the set of data objects identities, whose descriptions contain the relevant <*attribute:value*> pair. Thus, every agent responding to the query sends a single response message, regardless of the number of data objects satisfying the requirement posed by the query.

The above hashing-based mapping of attributes to LoudVoice channels may lead to hashing collisions and to false search results. However, the values posed in the query will usually discriminate between the real attribute being searched and the attribute which was accidentally mapped to the same channel. Our experimental results strengthen this observation.

In order to obtain the results of a multi-attribute query, responses received over different channels are intersected. The intersection is computed by the agent of the user initiating the query, and the resulting set of $object_{id}$ comprises only those descriptions of data objects, that exactly match the query. Thus, the system inherently guarantees high values of precision and recall.

3.2 Meta-Data Extraction

Additionally to the search capability, the proposed structure of channels facilitates autonomous generation of domain meta-data. Any agent connected to a channel, and in particular the mediating agent, receives all the messages transmitted over the channel. Thus, it is able to collect meta-data, i.e., data referring to the statistical properties (e.g., distribution, values, frequency etc…) of a given attribute and its known values.

Through the inter-organization communication channel, the mediating agents can also collect meta-data regarding the set of attributes describing the objects. For example, a mediating agent of a given channel can manage a list of other attributes that are mentioned in the descriptions jointly with the current attribute. In addition to the list of other attributes, meta-data regarding the possible values of an attribute can be collected by the mediating agent. This meta-data is also referred as the domain ontology.

This capability can be exploited for the purposes of advanced data management. For example, it facilitates interactive search process through presenting the users a set of attributes describing the domain objects, or a set of possible values of the current attribute. It can also serve as a correcting mechanism, suggesting the user to add some attributes to the unspecified description of the data objects, or to replace the mentioned value with another similar, but more popular (and hopefully, more exact) value.

The extracted meta-data can be also organized in a hierarchical manner using the statistic properties of the attributes and their values. The hierarchical meta-data facilitates differentiating between significant and insignificant attributes and suggesting corrections to the queries. For example, if a query lacks a significant attribute, the user can be proposed to refine it by inserting the missing attribute.

Although these functionalities are important and valuable, they fall beyond the scope of this work, which only validates the initial meta-data extraction functionality.

4. Experimental Results

We implemented a prototype version of the proposed approach using the LoudVoice multi-agent communication platform based on the channeled multicast communication model. We

incorporated the publicly available version 2.1 of WordNet for the purposes of simple semantic standardization.

The experiments were conducted in the realm of E-Commerce advertisements (ads). A corpus of *1272* supply ads from different domains was gathered from *www.recycler.com*. To achieve *<attribute:value>* representation, each ad was manually converted to the form of unspecified description. For example, "*Nokia 5190 phone, with charger and leather case, in good condition, 125$*" ad was converted to *<manufacturer:Nokia, model:5190, charger:included, case:leather, condition:good, price:125>*. Conversions were done as close as possible to the original to mimic behavior of naive users. A set of demand ads was built by changing a subset of the attributes and values in the supply ads.

The first experiment was designed to evaluate two traditional Information Retrieval metrics of the proposed system: *precision* and *recall* [15]. In context of E-Commerce ads, precision is computed as the number of retrieved relevant ads divided by the total number of retrieved ads. Similarly, recall is computed as the number of retrieved relevant ads divided by the total number of relevant ads in the system. For example, consider *100* mobile phones ads inserted to the system. User looking for a mobile phone launches a query, and receives *80* ads. *60* out of them are mobile phone ads, while the rest are irrelevant. In this case the precision is *60/80=0.75*, and the recall is *60/100=0.6*.

The values of precision and recall are measured for gradually increasing from *1* to *10* number of available LoudVoice channels (N_C). For each value of N_C we compute the average recall and precision among launching the *64* demand queries. Both the precision and the recall are measured twice: for the original terms mentioned in the ads (the brighter curves), and after standardizing the *<attribute:value>* pair with WordNet (the darker curves). Figure 3 shows the average values of precision and recall as a function of the number of channels N_C. The dashed curves show the precision values, while the continuous curves the recall.

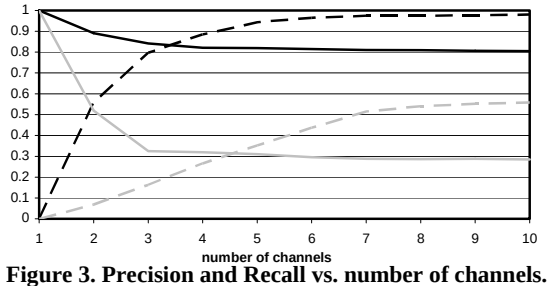

Figure 3. Precision and Recall vs. number of channels.

In general, both the probability of hash collisions and the number of irrelevant ads, accidentally mapped to a channel, decrease with the number of channels N_C. Thus, the precision increases with the N_C. It can be seen from the chart that the standardized results outperform the original results. This holds for any value of N_C, and the standardized precision values asymptotically converge to *1* (maximal value, when all the ads are relevant) starting from relatively low value of N_C*=4*. The standardized precision values are higher, as WordNet substitutes synonyms with a representing term. Thus, the number of different attributes and the number of ads accidentally mapped to a channel decreases, whereas the precision increases.

The original recall values of the system are relatively low, approximately *0.29*. This is explained by the observation that without the standardization the users use different terms in the unspecified descriptions. Consequently, the searches find only the ads that mentioned the exact term (or the ads accidentally mapped to the same channel). Using WordNet, the recall values are significantly higher, roughly *0.8*. Note that initially the recall values decrease when the number of channels N_C increases. This happens due to the fact that the probability of similar ads mentioning different attributes to be accidentally mapped to the same channel is relatively high for low values of N_C.

The next experiment was designed to measure the communication overhead of the proposed structure. For a constant number of channels N_C*=25*, we gradually increased the number of inserted ads (chosen randomly), and launched an identical set of *64* search queries on the inserted ads. For each number of ads we measured the number of *live* channels (i.e., channels with agents connected to them), the number of ads per processing of a single query, and the average size of a message. Each experiment was repeated *1000* times. Figure *4* shows the results of the experiment (the average message size is scaled down to appear in the chart)[*].

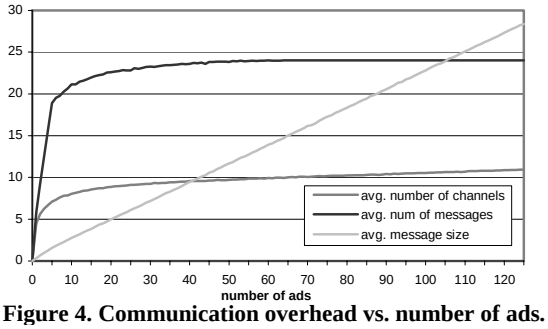

Figure 4. Communication overhead vs. number of ads.

The figure shows that both the number of live channels and the number of messages in a single search converge quickly. It can be seen that both of them reach approximately *90%* of their maximal values when inserting only *10%* of the ads. Thus, the system stabilizes fast in terms of the number of channels and the number of messages per one query. Since the response messages are formulated as $<agent_{id}, \{object_{id}\}>$, the average message size increases with the number of ads. However, it increases linearly. These observations indicate on a good scalability of the proposed approach and allows us to hypothesize that it will perform well for a large corpora of data objects descriptions, e.g., on the Web.

The last experiment was designed to examine the domain meta-data extracted by the system. We hypothesize that objects are mostly described using a set of commonly used attributes (i.e., "common sense" assumption). This was verified by recording the frequencies of the attributes mentioned in the ads form a single domain. The results are illustrated in figure *5*[*].

The results show that the distribution of the attributes' frequencies is logarithmic. There is a small set of dominating attributes (*price*, *manufacturer*, *model*), while many other attributes (*case*, *manual*, *SIM*, *headset*, *age*) appear in a very few object descriptions. This supports our idea of extracting hierarchical meta-data that can be

[*] These experiments were conducted over a corpus of *130* mobile phones ads. In other domains the results are similar and due to the lack of space we present these results only.

exploited for advanced data management functionalities, such as descriptions' correction, pro-active assistance and others.

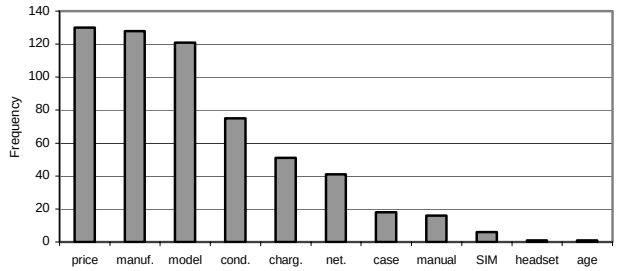

Figure 5. Frequency of domain attributes.

5. Related Works

The issue of accessing heterogeneous semantically enriched data has been addressed from different angles. Mariposa [14] implemented distributed data sharing over heterogeneous resources. Data objects are reformulated in microeconomic terms and queries are administered by a broker, which obtains bids for pieces of a query from various sites. All Mariposa clients and servers have an account and an allocated budget. The goal of the system is to solve the query within the allotted budget

Research works on the Semantic Web [6] focus on treating the Web as a knowledge base defining concepts and relationships. In particular, developed knowledge representation languages allowed representing meanings of the concepts relating them within custom ontologies for different domains, thus facilitating reasoning about the concepts.

Research in the Data Integration [3] domain aims to address the key challenge in accessing heterogeneous data: the large number and the dynamicity of distinct ontologies. This restriction is resolved by merging different ontologies. Thus, by generating a global centralized ontology it allows data access in a uniform manner. In [11], a middleware system designed to integrate data from a broad range of data sources with very different query capabilities, is presented. This is done by defining generic rules for the middleware and using wrapper-provided rules to encapsulate the capabilities of each data source.

In [1] a cost-based optimization technique that caches statistics is suggested. In addition a novel invariants mechanism is employed, in which semantic information about data sources is used to discover cached query results of interest. Piazza [12] offers a language for mediating between data sources on the Semantic Web, which maps both the domain structure and document structure. Mappings are provided at a local scale between small sets of nodes, and a query answering algorithm is able to chain sets mappings to obtain relevant data from across the network.

Other venue of research, highly correlated with data integration, is schema matching, i.e. the task of matching between concepts describing the meaning of data in heterogeneous, distributed data sources. Basic concepts and a model are proposed in COMA [7]. Various schema matchers differ mainly in the similarity metrics they exploit, yielding different similarity degrees. These metrics can be arbitrarily complex, and may use various techniques e.g., name matching, domain matching, structure matching, etc. The work in [9] provides an array of matching and filtering algorithms and a framework for developing new schema matchers which can be plugged-in and used.

6. Conclusions and Future Research

In the conducted experiments we verified the scalability of our approach. The number of generated channels and the number of messages in the query processing stabilizes after a relatively low number of inserted ads. We verified our hypothesis that in a given domain there is a small number of dominating attributes that capture the communalities in the underlying objects descriptions.

In future works, we intend to collect statistical meta-data to facilitate the use of domain-designated recommending agents. Such agents will be connected to channels representing the most significant attributes of a domain. They will accumulate meta-data about specific users through the queries launched by their agents. This will facilitate identifying the queries that the user would have like to launch for notifying the user about matching objects.

7. References

[1] S.Adali, K.Candan, Y.Papakonstantinou, V.Subrahmanian, "*Query Caching and Optimization in Distributed Mediator Systems*", in Proceedings of the SIGMOD Conference, Montreal, 1996.

[2] Y. Ben-Asher, S. Berkovsky, "*UNSO: Unspecified Ontologies for P2P E-Commerce Applications*", in Proceedings of the ICI Conference, Cesme, 2004.

[3] P.A.Bernstein, S.Melnik, "*Meta Data Management*", in Proceedings of the ICDE Conference, Boston, 2004.

[4] P.Busetta, A.Dona, M.Nori, "*Channeled Multicast for Group Communications*", in Proceedings of the AAMAS Conference, Bologna, 2002.

[5] P.Busetta, M.Merzi, S.Rossi, F.Legras, "*Intra-Role Coordination Using Group Communication: A Preliminary Report*", in Proceedings of the ACL Workshop, Melbourne, 2003.

[6] M.Dean, D.Connolly, F.van Harmelen, J.Hendler, I.Horrocks, D.McGuinness, P.Patel-Schneider, L.Stein, "*OWL Web Ontology Language*", W3 Consortium, 2002.

[7] H.H.Do, E.Rahm, "*COMA - a System for Flexible Combination of Schema Matching Approaches*", in Proceedings of the VLDB Conference, Hong-Kong, 2002.

[8] C.Fellbaum, "*WordNet - An Electronic Lexical Database*", MIT Press, 1998.

[9] A.Gal, G.Modica, H.M.Jamil, A.Eyal, "*Automatic Ontology Matching Using Application Semantics*", in AI Magazine, vol. 26(1), 2005.

[10] T.R.Gruber, "*A Translation Approach to Portable Ontology Specifications*", in Knowledge Acquisition Journal, vol. 6(2), 1993.

[11] L.Haas, D.Kossmann, E.Wimmers, J.Yang, "*Optimizing Queries Across Diverse Data Sources*", in Proceedings of the VLDB Conference, Athens, 1997.

[12] A.Y.Halevy, Z.G.Ives, P.Mork, I.Tatarinov, "*Piazza: Data Management Infrastructure for Semantic Web Applications*", in Proceedings of WWW Conference, Budapest, 2003.

[13] D.Kossmann, "*The State of the Art in Distributed Query Processing*", in ACM Computing Surveys, vol. 32(4), 2002.

[14] I.M.Stonebraker, P.M.Aoki, W.Litwin, A.Pfeffer, A.Sah, J.Sidell, C.Staelin, "*Mariposa: A Wide-Area Distributed Database System*", in VLDB Journal, vol. 5(1), 1996.

[15] I.H.Witten, A.Moffat, T.C.Bell, "*Managing Gigabytes: Compressing and Indexing Documents and Images*", Morgan Kaufmann Publishers, 1999.