An Experiment in Integrating Internet Information Sources

Alon Y. Levy
AT&T Bell Laboratories
AI Principles Research Dept.
600 Mountain Ave., Room 2C-406
Murray Hill, NJ, 07974
levy@research.att.com

Joann J. Ordille
AT&T Bell Laboratories
Computing Science Research Center
600 Mountain Ave., Room 2C-301
Murray Hill, NJ, 07974
joann@research.att.com

Introduction

The number of online information sources is growing rapidly. Though much of this information is unstructured (e.g., text, images) the number of structured information sources (e.g., databases) is also increasing. The advantage of structured information sources is that we can actually query their contents and use them to answer queries, perhaps in combination with other structured sources, rather than just viewing the whole information source. Given the variability in the information sources available, it is impractical for a user to interact with each source using its specific terminology. As a consequence, several systems (such as Nomenclator [Ordille and Miller, 1993b; Ordille, 1994], The Information Manifold [Kirk *et al.*, 1995], TSIMMIS [Chawathe *et al.*, 1994], SIMS [Arens *et al.*, 1994]) have been developed based on the notion of a *mediator* [Wiederhold, 1992]. The key idea behind a mediator is that a user interacts with one uniform model of the domain, and the mediator translates between queries posed in the domain model, and the ontologies of the specific information sources. In order to do so, the mediator needs *descriptions* of the specific information sources that relate their contents to the domain model seen by the user.

A major open issue which is crucial to the success of applying the notion of a mediator to large numbers of networked information sources is the ability to *obtain* accurate descriptions of the sources. The difficulty lies in problems such as sources using different attribute names to refer to the same intended attribute, or using the same attribute name to refer to different intended attributes. Further complications arise when the *format* of the attributes differs, even if they mean the same thing.

An important characteristic of many networked information sources is that they fall into classes that model very similar domains. For example, a large class of information sources describe data about people in an organization, such as their addresses, phones, and email addresses. Similarly, many information sources contain bibliographic citations, or product and price listings. This paper describes an approach to the problem of obtaining descriptions of information sources that is well suited for networked information sources of this type, and the first large scale experiment that we are currently conducting, aimed at obtaining such descriptions in a real world environment. Our approach is based on the assumption that obtaining descriptions of information sources is a task that will inherently involve the managers of the information sources themselves, and the key to the solution of the problem is to find ways in which they can provide the most accurate description of their sources with minimal effort.

Certain aspects of this problem have been considered in the database community as the *schema integration problem.* The main focus of that work was to design automatic or semi-automatic methods for mapping schemas of different databases to each other, or to a global schema. The automatic methods required extensive knowledge of the meaning of the information in the databases, and often placed an excessive burden on managers to express that meaning in some new language. Recently, some works have applied Machine Learning methods to the problem of schema integration [Perkowitz and Etzioni, 1995; Li, 1995]. The key idea in these methods is to look at some parts of the data itself in order to determine possible mappings. Our approach builds on that of [Perkowitz and Etzioni, 1995] in that it obtains the descriptions from examples of entries in the information source. Specifically, we ask the information source manager to map a representative entry in their database to our domain model, and to tell us how a *canonical* example that we design maps to their database. The fact that we design our canonical example enables us to reduce ambiguities in the descriptions.

We are conducting an experiment that obtains descriptions of CCSO name servers [Dorner and Pomes, 1992]. Currently, there exist approximately 300 institutions that use the CCSO protocol, each with its own database manager. The servers provide various data about people in their institutions. Aside from the validation of our approach to obtaining descriptions, the results of the experiment will be significant for two reasons. First, given the mappings we will be able to provide the first large scale white pages service that is completely based on knowledge representation methods, thereby showing the utility of knowledge representation in this problem. Second, the mappings obtained (which will be validated with the site managers) will provide a first test data set on which researchers in the field can make meaningful comparisons of their methods for obtaining descriptions.

Modeling Multiple Information Sources

To provide access to multiple information sources, a system that is built on the concept of a mediator requires a description of the contents of the information sources available. The description is needed for the mediator to (1) decide which information sources are relevant to a given query, and (2) formulate subqueries to the information source and understand its answers.

We illustrate the concepts associated with site descriptions using the *people domain* which we also use in our experiment. To distinguish different domain models, we use this font to denote relations in the domain model of the mediator (i.e., that in which the user poses queries), which we refer to as the *world-view.* We use `type-writer` font to refer to relations stored in specific information sources.

In the people domain we have the unary relation person, whose extension is a set of unique identifiers for people in the domain (most name servers have a unique identifier for people in their databases, and therefore unique identifiers can be viewed as the juxtaposition of the server name and the unique identifier given by the server). The properties of the person are given by binary relations, which we refer to as *attributes*, such as first-name, middle-name, last-name, email, office-phone, office-address, office-city, office-state. Each binary relation associates the unique identifier of a person with the value of an attribute for that person.

Information sources, on the other hand, use their own specific attribute names. For example, one source may use the attribute `full-name`, while another may use `nom`, to refer to the person's name (i.e. to the concatenation of first-name, middle-name, and last-name). Some sources may distinguish `internal-email` and `external-email`. Finally, some sources may not have the attribute corresponding to office-city, because all people in the organization work in the same city.

We focus on three kinds of information that are given in a description of an information source:[1]

1. Relation and attribute mappings: how the relations and attributes used in the information source relate to those used in the world view.
2. Format mappings: mappings of the formats in which the values are represented in the information source to those in the world view.
3. Integrity constraints on the source: constraints that are guaranteed to hold on *all* entries in the information source.

Formally, a description of an information source is composed of the following three kinds of logical sentences: (the R's represent relations in the world view and the E's represent relations in the site relations. The first argument of every binary relation is assumed to be the unique identifier for a person.)

1. $\mathsf{R}(x,y) \equiv \mathtt{E}_1(x,x_1)\wedge\ldots\mathtt{E}_n(x,x_n)\wedge y = f(x_1,\ldots,x_n)$.
   The $\mathtt{E}_i$'s are all relations in *same* information source. The function f can play two roles. First, it can translate between the formats used in the world view and the information source. Second, it can compose a world view attribute from several attributes in the information source or vice versa. For example, it may concatenate attributes or select a substring of an attribute.
   In the above sentence, the attribute of the world view is composed of several attributes in the information source.
2. $\mathsf{R}_1(x,x_1) \wedge \ldots \mathsf{R}_n(x,x_n) \wedge y = f(x_1,\ldots,x_n) \equiv \mathtt{E}(x,y)$.
   In this sentence, an attribute of the information source is composed of several attributes of the world view.
3. $\mathtt{E}(x,y) \Rightarrow C(y)$.
   The relation C denotes an interpretable constraint (e.g., order constraint, membership in a given set of values). This sentence expresses an integrity constraint on the values in an information source.

The following are a few examples of descriptions:

- first-name(x,y) $\equiv$ `uiuc:first-name(x,y)`: there is a simple 1-1 mapping between the relations in the world view and those in the information source.
- first-name(x,y) $\wedge$ last-name(x,z) $\equiv$ `uiuc:full-name`(x,Append(y,z)): The information source does not store the first and last name as two separate attributes, but as one field.
- phone-number(x,y) $\equiv$ `BL-phone-number(x,z)` & y=Append("582",z): there is 1-1 mapping, however the actual values in the information source contain some implicit information.
- home-state(x,y) $\equiv$ `local-home-state(x,z)` & y=State-translation-table(z): Here too there is a 1-1 mapping, but the actual values in the information field need to be processed through a translation table to obtain their representation in the world view.
- `uiuc:Person(x)` $\wedge$ office-city(x,y) $\Rightarrow$ y=Urbana: All the people in the UIUC database work in Urbana.

A major open issue which is crucial to the success of applying the notion of a mediator to large numbers of networked information sources is the ability to obtain these descriptions accurately. In the next section we

[1] Other kinds of information that are present in a description but are irrelevant to our current discussion can be found in [Levy *et al.*, 1995].

describe a method for obtaining such descriptions, and then we describe a first large scale experiment aimed at obtaining such descriptions in a real world environment.

Obtaining Site Descriptions

Several aspects of the problem of obtaining site descriptions have been considered in the area of interoperable databases and multidatabases (see [Batini *et al.*, 1986; Hurson *et al.*, 1994] for surveys) as the problem of schema integration. The main focus in those works has been on developing automatic or semi-automatic methods for obtaining mappings between schemas of different databases. The automatic methods required extensive knowledge of the meaning of the information in the databases, and often placed an excessive burden on database managers to express that meaning in some new language. Our methods eliminate this burden by allowing managers to specify mappings by example. Moreover, there are several assumptions that have been made in most of these works that are not applicable in our context. First, the number of databases which were to be incorporated is significantly smaller than the number considered here. Methods for schema integration have not been shown to scale up to large numbers of databases. Second, it is assumed that all the databases belong to one organization, or a small set of highly cooperative organizations. In our context, the organizations involved are autonomous, and not necessarily fully cooperative (note the difference between autonomy of the databases, which is assumed also in previous work, and autonomy of the organization to which the database belongs). Finally, much of the previous work has focused on obtaining *complete* mappings between database schemas, i.e., mappings in which we do not lose any information. Such mappings are important when we want to enable the user to make updates to the databases through the world-view. In our application, we start out assuming that some information will be lost, and we are trying only to map attributes that are considered of wide interest.

Several recent works have tried to apply machine learning techniques to the problem of obtaining source descriptions. These methods have the advantage that they do not require cooperation of the managers of the databases systems, and that they are likely to scale up to a large number of sources. However, these methods are very preliminary and it is not clear yet how useful they will be in real applications.

The overall goal of our work is to build a system that provides knowledge-based access to a large number of information sources, using the query processing methods described in [Ordille and Miller, 1993a; Kirk *et al.*, 1995; Levy *et al.*, 1995]. As a first goal, we aim to provide a *white-pages* service that integrates Internet name servers, currently provided by a growing number of organizations. To build such a service, we require descriptions of the information sources.

Our main assumption is that such descriptions can only be obtained with the cooperation of the source managers (every organization has a designated name server manager who is familiar with its schema and its contents). Our approach is, therefore, to obtain descriptions of the information sources by asking the information source managers to answer a set of queries *about* their database. The key issue that we need to address is how to design a small set of queries to pose to the managers, in order to minimize their effort and encourage cooperation. Our approach also exploits the fact that the information sources are modeling very similar domains. This leads us (in the spirit of [Perkowitz and Etzioni, 1995]) to obtain descriptions by considering *canonical* examples of database entries.

We obtain the descriptions by asking the database managers two sets of queries. In the first set, $Q1$ they are requested to select a representative entry from their database, and tell us how it would be represented in the world-view. That is, for every attribute in the world-view, they should provide the information corresponding to the entry they chose. In the second set, $Q2$, we design a canonical example of a database entry using the world-view, and the site managers are requested to return a set of attribute-value pairs using their local schema, describing how the canonical example would be represented in their database.

Informally, the answers to these two sets of queries are used as follows. We consider the answers for $Q1$ (i.e., an example from the database mapped to the world view) and compare them to the actual database entry corresponding to the example (the actual database entry is obtained directly from the database). For every attribute of the world view, we search for occurrences of its value in the database entry, and use several heuristics to rank the possible mappings between attributes in the world view and attributes in the site. For example, if we find that the value occurs *alone* and in a *single* attribute, we postulate a 1-1 mapping between the world-view attribute and the site attribute in which it was found. Alternatively, we search for mappings in which a value may have to be *extracted* from one of the fields in the information source (e.g., the first name may be the first string in the attribute corresponding to name), or mappings in which several attributes of the world view schema are found together in one entry in the database or vice versa.

We use the answers of the second set of queries to disambiguate common overlaps between various fields. For example, it is often the case that the email and the last name of a person are the same. The canonical example is designed such that these fields are different, and therefore when the database managers map these into the local schema, they must indicate the corresponding field names. Note that the set of queries $Q2$ essentially removes the need to consider several training example as done in [Perkowitz and Etzioni, 1995]. The second set of queries is important for several other reasons. First, the example given by the local database manager may only capture common cases in the database and miss some valuable information that exists only in uncommon cases. For example, she may choose an entry that does not have a value for the attribute `WWW-homepage`, while some entries in the database do have them. Mapping our canonical example requires that the database manager consider all the possible attributes that are of interest to us. Second, the answers produced by the first set of queries will use the format from the information source. For example, if we consider the date of birth, the value given for the date of birth will be mapped to the attribute in the world-view, but its format will be that of the local database. Using our canonical example, we give the database managers the data in our format, and they will map it to their representation. The first set of queries, on the other hand, is essential because the canonical example used in the second set may not have data dependencies that the local database enforces. For example, it may be the case that the email address is always *determined* by the last name of the person, while in our canonical example, they were designed to be different. In such a case, the database manager will point out that some attributes of the canonical example cannot be mapped appropriately to their local database.

The Experiment

We are conducting an experiment to test techniques for collecting source descriptions in a real environment, the CCSO name servers [Dorner and Pomes, 1992] on the global Internet. In this section, we describe the CCSO servers and their heterogeneity. We then describe our ongoing experiment.

The CCSO name server is a relational-style name server developed and distributed free by the University of Illinois Computing and Communications Services Office. Approximately 300 different organizations use the CCSO server to provide public white pages services on the Internet. More than 100 new organizations began using the server in the last year. The Internet Engineering Task Force (IETF) follows the growing success of the CCSO servers with interest and supports our project to integrate the servers.

The CCSO servers together provide one of the most extensive bodies of information about people available on the Internet; however, that information is not integrated nor are sophisticated cross-server searches possible. The schema for each server differs, because each organization separately administers its server and tailors the schema of the server to the organization's needs. The mappings in our source descriptions will allow us to integrate the servers, and the constraints in the source descriptions will allow us the speed cross-server searches for information.

We have conducted a study of the CCSO servers to determine the difficulty of integrating the servers. In our study, we collected and analyzed the schemas for 90% of the CCSO name servers.[2] We report two measures of the heterogeneity of the servers: the number of distinct attributes and a cumulative distribution (see Figure 1) of the number of attributes that occur in at least a certain percentage of the servers. There were 908 distinct attributes in the servers. The 60 most frequent attributes are graphed in the cumulative distribution in Figure 1. For example, the graph shows that 20 attributes occur in at least 56 percent of the servers. The distribution has a long tail of attributes (not shown) that occur in less than four percent of the servers.

The difficulty of our integration task is indicated by the large number of different attribute names. We seek to provide information about people in a much smaller set of attributes. Our perusal of the schema files and queries to the CCSO servers indicates that the large number of attributes results from a large diversity in the representation of the same information. Figure 1 shows that very few attributes (if any) are common across the different servers. In fact, only `name` and `email` occur in more than 90 percent of the servers. Figure 1 shows that most world-view attributes will need to be integrated from different attributes in the different servers. As our experiment continues, we will generate additional measurements to characterize the heterogeneity of the servers.

We are collecting our example mappings and integrity constraints through the World-Wide Web at

http://www.cs.att.com/csrc/nomen/nomenclator.html.

We have currently collected 40 descriptions in our ongoing experiment. Administrators have found the technique simple to use. Our heuristics for generating mappings seem to work well in practice, but a more thorough evaluation is still needed. We plan to verify our final mappings with the administrators to ensure accuracy. The validated mappings will be used in a white pages service sponsored by InterNIC Directory and Database Services. Together with the schemas, the validated mappings will be available to other researchers for studies of information integration techniques.

References

Arens, Yigal; Chee, Chin Y.; Hsu, Chun-Nan; and Knoblock, Craig A. 1994. Retrieving and integrating data from multiple information sources. *International Journal on Intelligent and Cooperative Information Systems.*

Batini, C.; Lenzerini, M.; and Navathe, S. B. 1986. A comparative analysis of methodologies for database schema integration. *ACM Computing Surveys* 18 (4):323–364.

[2]The remaining 10% were unavailable, because the organizations do not export descriptions of their schema, the server deviates from the CCSO protocol, or the server is chronically unavailable.

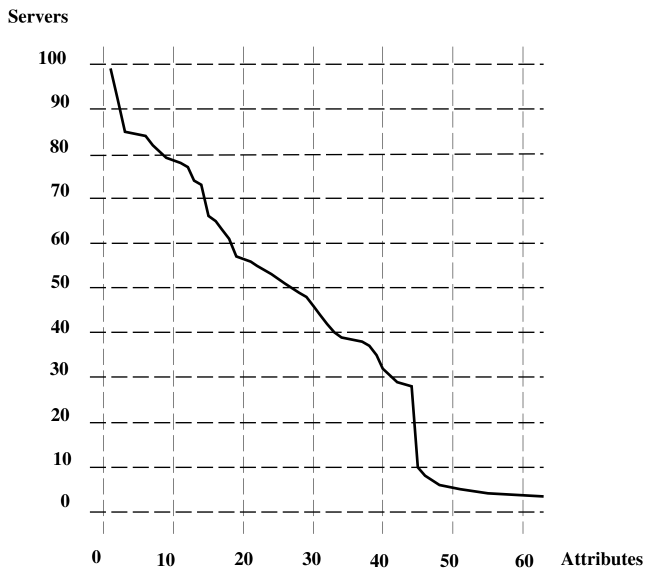

Figure 1: Number of Attributes in Percentages of Servers. The graph represents the number of attributes that occur in at least a certain percentage of the servers.

Chawathe, Sudarshan; Garcia-Molina, Hector; Hammer, Joachim; Ireland, Kelly; Papakonstantinon, Yannis; Ullman, Jeffrey; and Widom, Jennifer 1994. The TSIMMIS project: Integration of heterogenous information sources. In proceedings of IPSJ, Tokyo, Japan.

Dorner, Steven and Pomes, Paul 1992. The ccso nameserver a description. Computer and Communications Services Office Technical Report, University of Illinois, Urbana, USA, Available from ftp://vixen.cso.uiuc.edu/pub/qi.tar.gz.

Hurson, A. R.; Bright, M. W.; and Pakzad, S. H., editors 1994. *Multidatabase Systems: An Advanced Solution for Global Information Sharing.* IEEE Computer Society Press, Los Alamitos, California.

Kirk, Thomas; Levy, Alon Y.; Sagiv, Yehoshua; and Srivastava., Divesh 1995. The information manifold. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments.* American Association for Artificial Intelligence.

Levy, Alon Y.; Srivastava, Divesh; and Kirk, Thomas 1995. Data model and query evaluation in global information systems. *Journal of Intelligent Information Systems, Special Issue on Networked Information Discovery and Retrieval.* 5 (2).

Li, Wen-Syan 1995. Knowledge gathering and matching in heterogeneous databases. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments.* American Association for Artificial Intelligence.

Ordille, Joann J. and Miller, Barton P. 1993a. Database challenges in global information systems. In *Proceedings of ACM SIGMOD 1993 International Conference on Management of Data.* 403–407.

Ordille, Joann J. and Miller, Barton P. 1993b. Distributed active catalogs and meta-data caching in descriptive name services. In *Proceedings of the 13th International IEEE Conference on Distributed Computing Systems, Pittsburgh, USA.* 120–129.

Ordille, Joann J. 1994. Information gathering and distribution in nomenclator. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments, Stanford, California, USA.* 138–141.

Perkowitz, Mike and Etzioni, Oren 1995. Category translation: Learning to understand information on the internet. In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous Distributed Environments.* American Association for Artificial Intelligence.

Wiederhold, Gio 1992. Mediators in the architecture of future information systems. *IEEE Computer* 38–49.