Towards Semantics-Enabled Distributed Cyberinfrastructure for Knowledge Acquisition[1]

Vasant Honavar
Department of Computer Science
226 Atanasoff Hall
Iowa State University
Ames, Iowa 50011-1040
honavar@cs.iastate.edu

Doina Caragea
Department of Computing & Information Sciences
234 Nichols Hall
Kansas State University
Manhattan, KS 66506
dcaragea@cis.ksu.edu

1 Introduction

Recent development of high throughput data acquisition technologies in a number of domains (e.g., biological sciences, atmospheric sciences, space sciences, commerce) together with advances in digital storage, computing, and communications technologies have resulted in the proliferation of a multitude of physically distributed data repositories created and maintained by autonomous entities (e.g., scientists, organizations). The resulting increasingly data rich domains offer unprecedented opportunities in computer assisted data-driven knowledge acquisition in a number of applications including in particular, data-driven scientific discovery in bioinformatics (e.g., characterization of protein sequence-structure-function relationships in computational molecular biology), environmental informatics, and health informatics. Machine learning algorithms offer some of the most cost-effective approaches to knowledge acquisition (discovery of features, correlations, and other complex relationships and hypotheses that describe potentially interesting regularities) from large data sets. However, the applicability of current approaches to machine learning in emerging data rich applications in practice is severely limited by a number of factors:

- Data repositories are large in size, dynamic, and physically distributed. Consequently, it is neither desirable nor feasible to gather all of the data in a centralized location for analysis. In other domains, the ability of autonomous organizations to share raw data may be limited due to a variety of reasons (e.g., privacy considerations [Clifton et al., 2003]). In both cases, there is a need for efficient algorithms for learning from multiple distributed data sources without the need to transmit large amounts of data.
- Autonomously developed and operated data sources often differ in their structure and organization (relational databases, flat files, etc.). Furthermore, the data sources often limit the operations that can be performed (e.g., types of queries – relational queries, restricted subsets of relational queries, statistical queries, execution of user-supplied code to compute answers to queries that are not directly supported by the data source;). Hence, there is a need for effective strategies for efficiently obtaining the information needed for learning under the operational constraints imposed by the data sources, and theoretical guarantees about the performance of the resulting classifiers relative to the setting in which the learning algorithm has unconstrained access to a centralized data set.
- Autonomously developed data sources differ with respect to data *semantics*. The Semantic Web enterprise [Berners-Lee et al., 2001] is aimed at making the contents of the Web machine interpretable. Data and resources on the Web are annotated and linked by associating metadata that make *explicit*, the ontological commitments of the data source providers or in some cases, the shared ontological commitments of a small community of users. The increasing need for information sharing between organizations, individuals and scientific communities has led to several community-wide efforts aimed at the construction of ontologies in several domains. Explicit specification of the ontology associated with a data repository helps standardize the semantics to an extent. Collaborative scientific discovery applications often require users to be able to analyze data from multiple, semantically disparate data sources from different perspectives in different contexts. In particular, there is no single privileged perspective that can serve all users, or for that matter, even a single user, in every context. Hence, there is a need for methods that can efficiently obtain from a federation of autonomous, distributed, and semantically heterogeneous data sources, the information needed for learning (e.g., statistics) based on user-specified semantic constraints between user ontology and data-source ontologies..

Against this background, we consider the problem of data driven knowledge acquisition from autonomous, distributed, semantically heterogeneous, data sources.

2 Learning from Distributed Data

Given a data set D, a hypothesis class H, and a performance criterion P, an algorithm L for learning (from centralized data D) outputs a hypothesis $h \in H$ that optimizes P. In pattern classification applications, h is a classifier (e.g., a decision tree.) The data D consists of a (multi)set of training examples. Each training example is an ordered tuple of attribute values, where one of the attributes corresponds to a class label and the remaining attributes represent inputs to the classifier. The goal of learning is to produce a hypothesis that optimizes the performance criterion (e.g., minimizing classification error on the training data) and the complexity of the hypothesis. In a distributed setting, a data set D is distributed among the sites 1,...,n containing data set fragments $D_1,\ldots,D_n$. Two simple (and common) types of data fragmentation are: horizontal fragmentation and vertical fragmentation. More generally, the data may be fragmented into a set of relations (as in the case of tables of a

[1] This research is supported in part by NSF grants IIS 0219699, IIS 0639230 and IIS 0711356.

relational database, but distributed across multiple sites). We assume that the individual data sets $D_1,\ldots,D_n$ collectively contain (in principle) all the information needed to construct the data set D.

The distributed setting typically imposes a set of constraints Z on the learner that are absent in the centralized setting. For example, the constraints Z may prohibit the transfer of raw data from each of the sites to a central location while allowing the learner to obtain certain types of statistics from the individual sites (e.g., counts of instances that have specified values for some subset of attributes), or in the case of knowledge discovery from clinical records, Z might include constraints designed to protect the privacy of patients.

The problem of learning from distributed data can be stated as follows: Given the fragments $D_1,\ldots,D_n$ of a data set D distributed across the sites $1,\ldots,n$, a set of constraints Z, a hypothesis class H, and a performance criterion P, the task of the learner L_d is to output a hypothesis that optimizes P using only operations allowed by Z. Clearly, the problem of learning from a centralized data set D is a special case of learning from distributed data where n=1 and Z=∅. Having defined the problem of learning from distributed data, we proceed to define some criteria that can be used to evaluate the quality of the hypothesis produced by an algorithm L_d for learning from distributed data relative to its centralized counterpart. We say that an algorithm L_d for learning from distributed data sets $D_1,\ldots,D_n$ is *exact* relative to its centralized counterpart L if the hypothesis produced by L_d is identical to that obtained by L from the data set D obtained by appropriately combining the data sets $D_1,..,D_n$.

Proof of exactness of an algorithm for learning from distributed data relative to its centralized counterpart ensures that a large collection of existing theoretical (e.g., sample complexity, error bounds) as well as empirical results obtained in the centralized setting carry over to the distributed setting.

A General Strategy for Transforming Centralized Learners into Distributed Learners: Our general strategy for designing an algorithm for learning from distributed data that is provably exact with respect to its centralized counterpart (in the sense defined above) follows from the observation that most of the learning algorithms use only some *statistics* computed from the data D in the process of generating the hypotheses that they output. (Recall that a statistic is simply a function of the data.) This yields a natural decomposition of a learning algorithm into two components:

a. an information extraction component that formulates and sends a statistical query to a data source and
b. a hypothesis generation component that uses the resulting statistic to modify a partially constructed hypothesis (and further invokes the information extraction component as needed).

A statistic $s(D)$ is called a sufficient statistic for a parameter θ if $s(D)$, loosely speaking, provides all the information needed for estimating the parameter from data D [Fisher, 1922]. Thus, sample mean is a sufficient statistic for the mean of a Gaussian distribution.

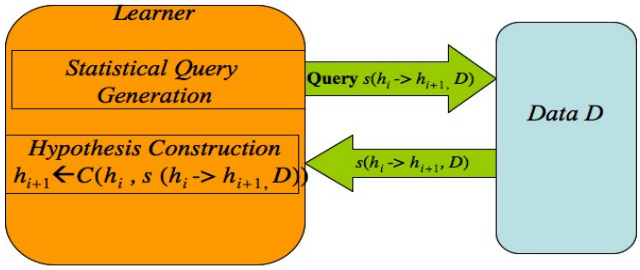

Figure 1: Learning = Statistical Query Answering + Hypothesis Construction

Inspired by theoretical work on PAC learning from statistical queries [Kearns, 1998], we have generalized this notion of a sufficient statistic for a parameter θ into a sufficient statistic $s_{L,h}(D)$ for learning a hypothesis h using a learning algorithm L applied to a data set D [Caragea et al., 2004a; 2005]. Trivially, the data D and the hypothesis h are both sufficient statistics for learning h using L. We are typically interested in statistics that are minimal or at the very least, substantially smaller in size (in terms of the number of bits needed for encoding) than the data set D. In some simple cases, it is possible to extract a sufficient statistic $s_{L,h}(D)$ for constructing a hypothesis h in one step (e.g., by querying the data source for a set of conditional probability estimates when L is the standard algorithm for learning a Naive Bayes classifier). In a more general setting, h is constructed by L by interleaving information extraction (statistical query) and hypothesis construction operations. Thus, a decision tree learning algorithm would start with an empty initial hypothesis h_0, obtain the sufficient statistics (expected information concerning the class membership of an instance associated with each of the attributes) for the root of the decision tree (a partial hypothesis h_1), and recursively generate queries for additional statistics needed to iteratively refine h_1 to obtain a succession of partial hypotheses h_1 , h_2 … culminating in h (See **Figure 1**). In this model, the only interaction of the learner with the repository of data D is through queries for the relevant statistics. Information extraction from distributed data entails decomposing each statistical query q posed by the information extraction component of the learner into sub queries $q_1,...q_n$ that can be answered by the individual data sources $D_1,..,D_n$ respectively, and a procedure for combining the answers to the sub queries into an answer for the original query q (See **Figure 2**).

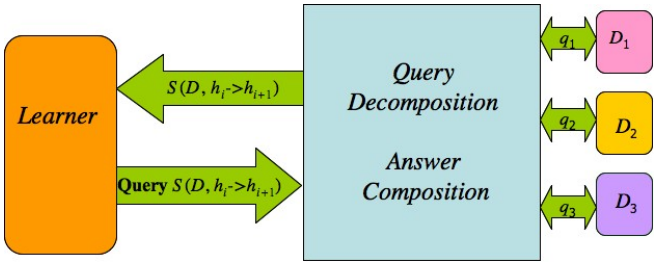

Figure 2: Learning from Distributed Data = Statistical Query Answering + Hypothesis construction

We have shown that this general strategy for learning classifiers from distributed data is applicable to a broad class of algorithms for learning classifiers from data [Caragea et al., 2004a]. Consequently, for these algorithms, we can devise a strategy (plan) for computing h from the data D using sufficient statistics. When the learner's access to data sources is subject to constraints Z, the resulting plan for information extraction has to be executable without violating the constraints Z. The *exactness* of the algorithm L_d for learning from distributed data relative to its centralized counterpart follows from the correctness of the query decomposition and answer composition procedures.

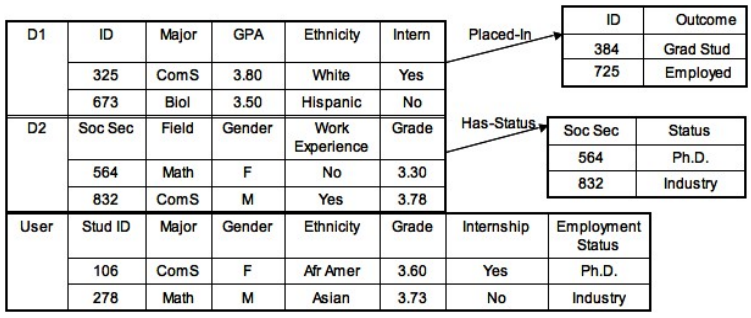

Figure 3: Student data collected by two departments from a statistician's perspective.

More precisely, we have applied the general framework described above for construction of algorithms for learning classifiers from distributed data to design provably exact algorithms for learning Naïve Bayes, Nearest Neighbor, Bayes Network, Neural Network, and Decision Tree classifiers from distributed data under horizontal and vertical data fragmentation [Caragea, 2004a], and Support Vector Machine (SVM) classifiers under horizontal data fragmentation (at the expense of multiple passes through the distributed data) [Caragea et al., 2004; Honavar & Caragea, 2008]. We have also established the precise conditions under which the proposed algorithms offer significant savings in bandwidth, memory, and/or computation time (relative to their centralized counterparts) [Caragea, 2004; Honavar & Caragea, 2008].

Relative to the large body of work on learning classifiers from distributed data (surveyed in Park & Kargupta, 2002), the distinguishing feature of this approach is a clear separation of concerns between hypothesis construction and extraction of sufficient statistics from data. This makes it possible to explore the use of sophisticated techniques for query optimization that yield optimal plans for gathering sufficient statistics from distributed data sources under a specified set of constraints describing the query capabilities of the data sources, operations permitted by the data sources, and available computation, bandwidth, and memory resources. It also opens up the possibility of exploring algorithms that learn from distributed data a hypothesis h_ε whose error is small relative to the error of a hypothesis h (obtained in the setting when the learner has unrestricted access to D), in scenarios where the constraints Z make it impossible to guarantee *exactness* in the sense defined above. The proposed approach also lends itself to adaptation to learning from semantically heterogeneous data sources.

3 Learning from Semantically Heterogeneous Data

In order to extend our approach to learning from distributed data (which assumes a common ontology that is shared by all of the data sources) into effective algorithms for learning classifiers from *semantically heterogeneous* distributed data, techniques need to be developed for answering the statistical queries posed by the learner in terms of the learner's ontology O from the heterogeneous data sources (where each data source D_i has an associated ontology O_i). Thus, we have to solve a variant of the problem of integrated access to distributed data repositories – the data integration problem [Levy, 2000; Calvanese & De Giacomo, 2005] in order to be able to use machine learning approaches to acquire knowledge from semantically heterogeneous data.

This problem is best illustrated by an example: (**Figure 3**). Consider two academic departments that independently collect information about their students. Suppose a data set D_1 collected by the first department is organized in two tables: *Student*, and *Outcome*, linked by a *Placed-In* Relation using ID as the common key. Students are described by *ID*, *Major*, *GPA*, *Ethnicity* and *Intern.* Suppose a data set D_2 collected by the second department has a *Student* table and a *Status* table, linked by *Has-Status* relation using *Soc Sec* as the common key. Suppose *Student* in D_2 is described by the attributes *SocSec*, *Field*, *Gender*, *Work-Experience* and *Grade.* Consider a user, e.g., a university statistician, interested in constructing a predictive model based on data from two departments of interest from his or her own perspective, where the representative attributes are *Student ID*, *Major*, *Gender*, *Ethnicity*, and *Grade, Internship* and *Employment Status.* For example, the statistician may want to construct a model that can be used to infer whether a typical student (represented as in the entry corresponding to D_U in **Figure 3**) is likely go on to get a *Ph.D.* This requires the ability to perform queries over the two data sources associated with the departments of interest from the user's perspective (e.g., *fraction of students with internship experience that go onto Ph.D*). However, because the structure (schema) and data semantics of the data sources differ from the statistician's perspective, he/she must establish the correspondences between the schema attributes and their values.

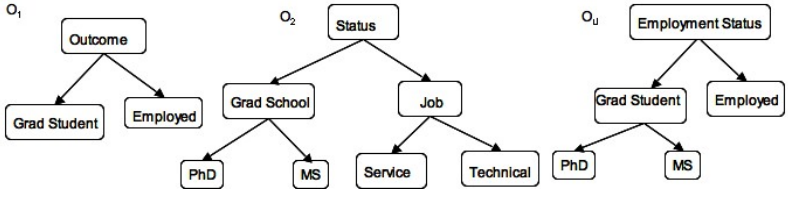

Figure 4: Attribute value taxonomies (ontologies) O_1 and O_2 associated with the attributes *Outcome* and *Status* in two data sources of interest. O_U is the ontology for *Employment Status* from the user's perspective.

We adopt a federated, query-centric approach to answering statistical queries from semantically heterogeneous data sources, based on ontology-extended relational algebra [Bonatti et al., 2003]. Specifically, we associate explicit ontologies with data sources to obtain *ontology extended relational data sources* (OERDS). An OERDS is a tuple $\mathcal{D}=\{D,S,O\}$, where D is the actual data set in the data source, S the data source schema and O the data source ontology [Caragea, 2004; Caragea et al., 2005; Honavar & Caragea, 2008]. A relational *data set* D is an instantiation $I(S)$ of a schema S. The *ontology* O of an OERDS D consists of two parts: *structure ontology*, O_S, that defines the semantics of the data source schema (entities, and attributes of entities that appear in data source schema S); and *content ontology*, O_I, that defines the semantics of the data instances (values and relationships between values that the attributes can take in instantiations of schema S). Of particular interest are ontologies that take the form of *is-a* hierarchies and *has-part* hierarchies. For example, the values of the *Status* attribute in data source D_2 are organized in an *is-a* hierarchy. A *user's view of data sources* $\mathcal{D}_1, \mathcal{D}_2 .. \mathcal{D}_n$ *is specified* by user schema S_U, user ontology O_U, together with a set of semantic *constraints IC*, and the associated set of *mappings* from the user schema S_U to the data source schemas $S_1,...,S_n$ and from user ontology O_U to the data source ontologies $O_1,\ldots,O_n$ [Caragea et al, 2005]. **Figure 4** shows examples of ontologies that take the form of *is-a* hierarchies over attribute values. **Figure 5** shows some simple examples of user-specified semantic constraints between the user perspective and the data sources D_1 and D_2, respectively.

How can we answer a statistical query in a setting in which autonomous data sources differ in terms of the levels of abstraction at which data are described? For example: Consider the data source ontologies O_1 and O_2 and the user ontology O_U shown in **Figure 4**. The attribute *Status* in data source D_2 is specified in greater detail (lower level of abstraction) than the corresponding attribute *Outcome* is in D_1. That is, data source D_2 carries information about the precise status of students after they graduate (specific advanced degree program e.g., *Ph.D.*, *M.S.* that the student has been accepted into, or the type of employment that the student has accepted), whereas data source D_1 makes no distinctions between the types of graduate degrees or types of employment. Suppose we want to answer the query: What fraction of the students in the two data sources got into a *Ph.D.* program? Answering this query is complicated by the fact that the *Outcome* of students in data source D_1 are only *partially specified* [Zhang et al., 2003; 2006] with respect to the ontology O_U. Consequently, we can never know the precise fraction of students that got into a Ph.D. program based on the information available in the two data sources. In such cases, answering statistical queries from semantically heterogeneous data sources requires the user to supply not only the mapping between the ontology and the ontologies associated with the data sources but also *additional assumptions of a statistical nature* (e.g., that grad program admits in D_1 and D_2 can be modeled by the same underlying distribution). The validity of the answer returned depends on the validity of the assumptions and the soundness of the procedure that computes the answer based on the supplied assumptions.

$O_1 \rightarrow O_U$	$O_2 \rightarrow O_U$
ID:O_1=Stud ID:O_U	Soc Sec:O_2=Stud ID:O_U
Major:O_1=Major:O_U	Field:O_2=Major:O_U
GPA:O_1=Grade:O_U	Grade:O_2=Grade:O_U
Ethnicity:O_1=Ethnicity :O_U	
	Gender:O_2=Gender:O_U
Intern:O_1=Internship:O_U	Work-Experience:O_2=Internship:O_U
Outcome:O_1=Employment-Status:O_U	Status:O_2=Employment-Status:O_U

Figure 5: An example of user-specified semantic correspondences between the user ontology O_U and data source ontologies O_1 and O_2 (from **Figure 4**)

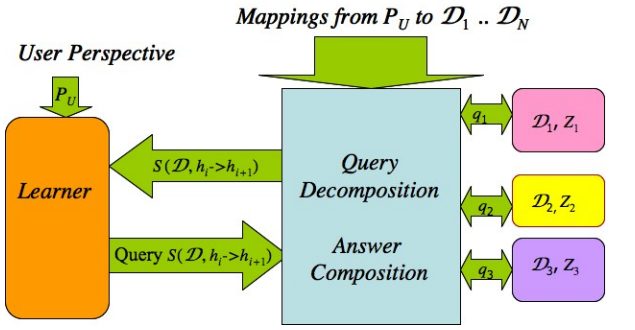

Figure 6: General Framework for learning classifiers from semantically heterogeneous distributed data.

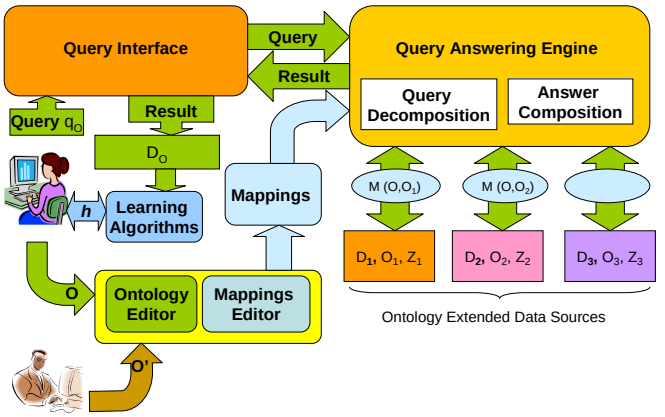

Figure 7: The INDUS System

Given a means of answering statistical queries from semantically heterogeneous data, we can devise a general framework for learning predictive models from such data (See **Figure 6**). Based on this framework, we have implemented a prototype of an *intelligent data understanding system* (INDUS) that supports: execution of statistical queries against semantically heterogeneous ontology extended data sources, and the construction of predictive models (e.g., classifiers) from such data sources (See **Figure 7**).

4 Research in Progress

Our current work is focused on the development of a semantics-enabled infrastructure for data-driven knowledge

acquisition for a broad range of e-science applications (e.g., bioinformatics). Work in progress is aimed at:

(a) Further development of algorithms with provable performance guarantees (in terms of accuracy of results, bandwidth and computational efforts), relative to their centralized counterparts, for learning predictive models (including their multi-relational counterparts) from semantically heterogeneous, distributed data sources, under a variety of constraints on the operations supported by the data sources. Of particular interest are resource-bounded approximations of answers to statistical queries generated by the learner (e.g., using sampling strategies); approximation criteria for evaluation of the quality of classifiers obtained in the distributed setting under a given set of resource constraints and query capabilities relative to that obtained in the centralized setting with or without such constraints. This is especially important in application scenarios in which it is not feasible to obtain exact answers to statistical queries posed under the access and resource constraints imposed by the distributed setting

(b) Reformulation of ontology extended data sources using description logics based ontology and mapping languages (RDF, OWL, etc.) and their distributed counterparts such as P-DL [Bao et al., 2007a]. Of particular interest are federated algorithms for reasoning with distributed ontologies [Bao et al., 2006] to support querying semantically disparate data sources and privacy-preserving reasoning algorithms [Bao et al., 2007b].

(c) Extending INDUS to support learning of predictive models from alternately structured data such as images, sequences, text, and graphs.

(d) Development of benchmark data sources (including data, associated ontologies, and mappings) to facilitate comparison of alternative approaches to building predictive models from distributed data.

(e) Experimental evaluation of the resulting semantics-enabled algorithms for learning predictive models from semantically heterogeneous data along several important dimensions including characteristics of data sources (complexity of data source schema, ontologies, and mappings; data source query and processing capabilities, size of the data sets, prevalence of partially missing attribute values as a consequence of integration of data described at multiple levels of granularity), errors or inconsistencies in semantic interoperation constraints and mappings; characteristics of algorithms (e.g., types of statistics needed for learning), and performance criteria (quality of results produced relative to the centralized counterparts, computational resource, bandwidth, and storage usage).

(f) Design, implementation, and dissemination of a modular, open source implementation of a suite of ontology-based inference, query rewriting, and learning algorithms (to be implemented as *services* that can be invoked on a collection of networked ontology-extended data sources).

References

1. Berners-Lee, T., Hendler, J. and Lassila, O. (2001). The semantic web. Scientific American, May 2001.
2. Bao, J., Caragea, D., and Honavar, V. (2006). A Tableau-based Federated Reasoning Algorithm for Modular Ontologies. ACM / IEEE / WIC Conference on Web Intelligence, Hong Kong, pp. 404-410.
3. Bao, J., Slutzki, G. and Honavar, V. (2007a). Privacy-Preserving Reasoning on the Semantic Web. ACM/WIC/IEEE International Conference on Web Intelligence, In press.
4. Bao, J., Slutzki, G., and Honavar, V. (2007b). A Semantic Importing Approach to Knowledge Reuse from Multiple Ontologies, In: Proceedings of the 22nd Conference on Artificial Intelligence (AAAI-2007). pp. 1304-1310.
5. Bonatti, P., Deng, Y., and Subrahmanian, V. (2003). An ontology-extended relational algebra. In: Proceedings of the IEEE Conference on Information Integration and Reuse, IEEE Press., pp. 192-199.
6. Calvanese, D. and De Giacomo, D. (2005). Data integration: A logic-based perspective, AI Magazine, 26: 59-70.
7. Caragea, D. (2004). Learning classifiers from Distributed, Semantically Heterogeneous, Autonomous Data Sources. Ph.D. Thesis. Dept. of Computer Science. Iowa State Univ.
8. Caragea, D., Silvescu, A., and Honavar, V. (2004a). A Framework for Learning from Distributed Data Using Sufficient Statistics and its Application to Learning Decision Trees. International Journal of Hybrid Intelligent Systems. Vol. 1, No. 2. pp. 80-89
9. Caragea, D., Zhang, J., Bao, J., Pathak, J., and Honavar, V. (2005). Algorithms and Software for Collaborative Discovery from Autonomous, Semantically Heterogeneous, Distributed, Information Sources. In: Proc. of the Conference on Algorithmic Learning Theory. LNCS. Vol. 3734. Berlin: Springer-Verlag. pp. 13-44.
10. Clifton, C,, Kantarcioglu, M., Vaidya, J., Lin, X., and Zhu, M. (2003) Tools for Privacy Preserving Distributed Data Mining, ACM SIGKDD Explorations, vol. 4, no. 2, 2003.
11. Fisher, R. A. On the Mathematical Foundations of Theoretical Statistics, Philosophical Transactions of the Royal Society A *222* (1922): 309--368
12. Honavar, V. and Caragea, D. Honavar, V., and Caragea, D. (2008) Collaborative Knowledge Acquisition from Semantically Disparate, Distributed, Autonomous Data Sources. Berlin: Springer. In press.
13. Kearns, M. 1998. Efficient Noise Tolerant Learning from Statistical Queries. Journal of the ACM. Vol. 45, pp. 983-1006.
14. Levy, A., 2000. Logic-based techniques in data integration. In: Logic Based Artificial Intelligence, Jack Minker (Ed). New York: Kluwer.
15. Park, B. and Kargupta, H. (2002) Distributed Data Mining: Algorithms, Systems, and Applications. In: Data Mining Handbook, Nong Ye (Ed.), IEA, pp. 341--358
16. Zhang, J., and Honavar, V. (2003). Learning from Attribute-Value Taxonomies and Partially Specified Instances. In: Proceedings of the International Conference on Machine Learning. Washington, DC. AAAI Press. pp. 880-887.
17. Zhang, J., Silvescu, A., Kang, D-K., and Honavar, V. (2006). Learning Compact and Accurate Naive Bayes Classifiers from Attribute Value Taxonomies and Partially Specified Data. Knowledge and Information Systems. 9:157-179.