Research Statement

AnHai Doan

The rapid spread of computers and communication networks has transformed our world into a vast information bazaar, with millions of sources providing data in every imaginable format and mode of interaction. Distributed information processing systems hold the promise of acting as crucial middlemen in this chaotic market, by interacting with data sources, translating, and combining their data in order to obtain the information requested by users. However, today this promise remains largely unfulfilled, because such systems are still very hard to build and costly to operate. They must be told in tedious detail how to interact with the data sources and understand the languages they use. With the sources in constant evolution, once deployed the systems must still be under continuous supervision and told how to deal with the changes. The laborious teaching and supervision incur huge costs and severely limit the deployment of such "middleman" systems in practice. As a consequence, the vast potential of the global information market has so far remained largely untapped.

My research seeks to unleash this potential by making distributed information processing systems much easier to use with far less need for human supervision. My ultimate goal is to achieve the widespread use of online information processing systems that take only minutes to be deployed (instead of weeks or months as is the case today), that require only minimal human coaching to rapidly reach and maintain competence, and that continuously improve over time, in terms of both performance and capabilities.

Toward this goal, I begin by studying an important and representative class of distributed information processing systems: data integration ones. Such systems provide a *uniform* query interface to a multitude of data sources, thereby freeing the users from the tedious job of manually selecting the relevant sources, querying them, and combining their data to obtain the answers. Most recent research (including some of my own [9, 10]) has addressed only the modeling and query processing aspects of data integration systems. I now plan to apply techniques from a variety of fields – most notably databases and machine learning – to significantly reduce the complexity of building and managing such systems. Specifically, I will focus on the following research areas:

Learning Source Descriptions To process user queries, a data integration system must know the descriptions of data sources. Today, such descriptions are created manually. I will develop techniques to automatically learn source descriptions. Toward this goal, my Ph.D. thesis has addressed the problem of learning the semantic mappings between a source schema and the query interface of the system [2, 5, 4]. The thesis shows that machine learning techniques can be applied to produce such mappings with high accuracy. Specifically, it describes a multi-strategy learning approach that applies multiple learners to predict mappings, then combines the learners' predictions using a meta-learner. This approach subsumes and generalizes most previous approaches on schema mapping, which employ only a single mapping strategy. The thesis also shows that data instances can be utilized effectively to generate mappings. This is in contrast to most previous works, which utilize only schema information.

While research to date on schema mapping has been very promising, it is only the first step. None of the current approaches can automatically learn the complex, non-one-to-one mappings that occur frequently in practice. I have already begun to investigate this issue [3]. A second challenge is to develop well-founded notions for semantic mapping. Such notions should help to communicate the meaning of mappings and to leverage specialized techniques for the mapping process. A third challenge is to develop a unified framework for schema mapping that combines in a principled, seamless, and efficient way all the relevant information (e.g., user feedback, mappings from a different application) and techniques (e.g., machine learning, heuristics). My recent work [11] suggests that mappings can be given well-founded definitions based on probabilistic interpretations, and that a unified mapping framework can be developed by leveraging probabilistic representation and reasoning methods such

as Bayesian networks. I plan to further investigate these issues. Learning other source characteristics such as schema, reliability, and query-processing capability also raises many fascinating challenges that I plan to pursue. For example, suppose we want to build a data integration system over all C.S. department web sites in the U.S. What would be the schema of such a department web site? How do we characterize and learn it automatically? I believe this problem can be cast as a schema mapping problem, and hence it will benefit from the mapping techniques that I have developed.

Dealing with Changes in Source Descriptions In dynamic and autonomous environments (e.g., the Internet) sources often undergo modifications with respect to schema, data, and query-processing capabilities. Hence, the operators of a data integration system must constantly monitor the component sources to detect and deal with changes. Clearly, manual monitoring is very expensive and not scalable. My goal therefore is to develop techniques to automate the monitoring and updating process. I believe an effective solution to the monitoring problem is to sample source data periodically, then use machine learning techniques to compare the current sample with previous source samples to detect changes. For example, the problem of detecting if the semantic mappings are still valid can be recast as a schema mapping problem that involves two consecutive source samples. A key challenge in the updating process will be updating the system's query interface to reflect changes in a source schema. This problem is similar to schema integration, a well-known and difficult problem. Several areas in databases and AI, including schema management, ontology merging, and model management, have addressed different aspects of this problem. I plan to build on techniques in these areas to develop an effective solution for updating.

Matching Objects across Sources The problem of deciding if two objects in two sources refer to the same real-world entity lies at the heart of the data integration enterprise. Previous solutions to this problem are unsatisfactory: they are largely ad-hoc and as a result have limited applicability. I believe this problem has close resemblances with the schema mapping problem. Hence, I plan to develop a solution to object matching that builds upon recent advances in schema mapping (including my own work [2, 11]) and utilizes techniques from machine learning and probabilistic reasoning.

Incorporating User Feedback User feedback is critical to many tasks during the construction and maintenance of a data integration system, because of the inherent subjectivity of the tasks and the imperfection of learning techniques. My experience in dealing with user feedback [2] suggests that, unless handled properly, it can quickly become a serious bottleneck in building and maintaining a system. My goal therefore is to develop techniques to minimize necessary user feedback while maximizing the impact of the feedback. My approach is to build a *single* feedback loop from the user to the system, instead of a *separate* loop from the user to each of the tasks that requires user supervision. Users would give feedback only on the correctness of the answers to a selected set of queries. The system would then use the feedback to verify the correctness of the semantic mappings, the source schemas, and so on. I will investigate techniques from active learning and intelligent user interfaces for this purpose.

I plan to validate the above research ideas by applying them to the construction and maintenance of large-scale data integration systems on the Internet and in specific application domains, such as medicine, astronomy, and biology. To ensure the success of this project, I intend to work closely with researchers in the application domains, drawing on my substantial experience in interdisciplinary collaboration (applying AI planning techniques [14, 1, 6, 13, 12] to medical diagnosis problems [15, 8, 7, 16]).

While the above research agenda focuses on data integration, it should have implications well beyond that context. Many of the problems that it investigates, such as schema mapping, object identification, schema integration, and user interaction, are fundamental issues in numerous data

management and data mining applications. The agenda also necessitates a strong emphasis on extending current machine learning techniques to deal with novel learning problems brought about by data integration. For example, my work on schema mapping has extended classification methods to handle semi-structured data [2], and developed efficient techniques based on relaxation labeling to classify entities that are interrelated in complex ways [11]. Such techniques should also find application in many database and machine learning problems. Hence, in parallel with pursuing my research on data integration, I also plan to investigate its implications for other problems. For example, I have applied my thesis work to the problem of translating between ontologies on the Semantic Web [11], and plan to apply it to the problem of information extraction from text.

Further into the future, I intend to build on my work in data integration to investigate systems with more sophisticated capabilities, such as those that integrate online services, and those that perform peer-to-peer data sharing. Such distributed information processing systems should play an important role in transforming the global information bazaar into a vast knowledge base for humankind, unleashing a revolution of new possibilities. The goal of my research is to see this vision realized.

References

[1] A. Doan. Modeling probabilistic actions for practical decision-theoretic planning. In *Proc. of the 3rd Int. Conference on AI Planning Systems (AIPS)*, 1996.

[2] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine learning approach. In *Proc. of the ACM Conference on Management of Data (SIGMOD)*, 2001.

[3] A. Doan, P. Domingos, and A. Halevy. Learning complex mappings between database schemas. 2002. To be submitted to the Conference on Very Large Databases (VLDB).

[4] A. Doan, P. Domingos, and A. Levy. Learning mappings between data schemas. In *Proc. of the AAAI-2000 Workshop on Learning Statistical Models from Relational Data*, 2000.

[5] A. Doan, P. Domingos, and A. Levy. Learning source descriptions for data integration. In *Proc. of the Third Int. Workshop on the Web and Databases (WebDB)*, 2000.

[6] A. Doan and P. Haddawy. Sound abstraction of probabilistic actions in the constraint mass assigment framework. In *Proc. of the 12th Nat. Conference on Uncertainty in AI (UAI)*, 1996.

[7] A. Doan, P. Haddawy, and C. Kahn. Decision-theoretic refinement planning: A new method for clinical decision analysis. In *Proc. of the 19th AMIA Annual Symposium on Computer Applications in Medical Care (SCAMC)*, 1995.

[8] A. Doan, P. Haddawy, and C. Kahn. Decision-theoretic planning for clinical decision analysis. In *Proc. of the Annual AI in Medicine Spring Symposium*, 1996.

[9] A. Doan and A. Halevy. Efficiently ordering query plans for data integration. In *Proc. of the 18th IEEE Int. Conference on Data Engineering (ICDE)*, 2002. To appear.

[10] A. Doan and A. Levy. Efficiently ordering query plans for data integration. In *Proc. of the IJCAI-99 Workshop on Intelligent Information Integration*, 1999.

[11] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the semantic web. 2002. Submitted to the World-Wide Web Conference (WWW).

[12] V. Ha, A. Doan, V. Vu, and P. Haddawy. Geometric foundations for interval-based probabilities. *Annals of Mathematics and Artificial Intelligence*, 24, 1998.

[13] P. Haddawy and A. Doan. Abstracting probabilistic actions. In *Proc. of the 10th Conference on Uncertainty in AI (UAI)*, 1994.

[14] P. Haddawy, A. Doan, and R. Goodwin. Efficient decision-theoretic planning: Techniques and empirical analysis. In *Proc. of the 11th Nat. Conference on Uncertainty in AI (UAI)*, 1995.

[15] P. Haddawy, A. Doan, and C.E. Kahn. Decision-theoretic refinement planning in medical decision making: Management of acute deep venous thrombosis. *Journal of Medical Decision Making*, 1996.

[16] C. Kahn, A. Doan, and P. Haddawy. Management of acute deep venous thrombosis of the lower extremities (abstract). In *American Roentgen Ray Society Meeting*, 1996.