Distributed Knowledge Networks

Vasant Honavar, Les Miller and Johnny Wong
Artificial Intelligence Research Laboratory, Department of Computer Science
226 Atanasoff Hall, Iowa State University, Ames, IA 50011. U.S.A.
http://www.cs.iastate.edu/~honavar/aigroup.html

Abstract— **Distributed Knowledge Networks (DKN) provide some of the key enabling technologies for translating recent advances in automated data acquisition, digital storage, computers and communications into fundamental advances in organizational decision support, data analysis, and related applications. DKN include computational tools for accessing, organizing, transforming, and analyzing the contents of heterogeneous, distributed data and knowledge sources and for distributed problem solving and decision making under tight time, resource, and performance constraints. This paper presents an overview of the DKN project in the Iowa State University Artificial Intelligence Laboratory.**

I. Introduction

Advanced scientific research (e.g., the genome project), military applications (e.g., intelligence data handling, situation assessment, command and control), law enforcement (e.g., terrorism prevention), crisis management, design and manufacturing systems, and medical information infrastructure, power systems, communications infrastructure, offer numerous examples of scenarios which call for reactive as well as proactive decision making, often under tight time, resource, and performance constraints to accomplish the desired objectives. Recent advances in sensor, high throughput data acquisition, and digital information storage technologies, have made it possible to acquire and store large volumes of data in digital form. Advances in computers and communications, Internet, and mobile computing have made it possible for decision makers, to have at their disposal, at least in principle, large volumes of data as well as analysis and decision support tools residing on multiple, geographically distributed, heterogeneous, hardware and software platforms around the globe that are connected to the Internet. However, in order to translate the advances in our ability to acquire, store, and analyze data in increasing volumes and at increasing rates into gains in our understanding of the respective domains and new capabilities for effective decision-making, sophisticated tools are needed for information retrieval, knowledge discovery, and distributed problem solving and decision making.

The following example motivates some of the problems that need to be addressed: Military commanders and intelligence analysts, in order to be able to effectively perform their responsibilities, need to have critical information in a timely fashion. Day-to-day operations involve intelligence data gathering and analysis, situation monitoring and assessment, looking for potentially interesting patterns in data as it is being gathered (e.g., relationship between troop movements and significant political developments in a region). This information can be extremely valuable to decision makers in taking both proactive as well as reactive measures designed to safeguard a nation's security concerns. Having the right information at the right time at the disposal of the decision makers will often mean that potential hotspots can be dealt with and major crisis prevented. However, given the complex dynamics of global military and political developments, it is impossible to completely avoid crisis situations. In a crisis, it is of utmost importance to make available to the decision makers, accurate information at the right time without overwhelming them with large volumes of irrelevant data. Examples of data sources include: satellite images, intelligence reports, etc. The availability of data in digital form also presents opportunities for discovery of otherwise hidden regularities or patterns in specific domains. The knowledge so gathered can be useful in the decision-making process. Advanced scientific, commercial, law enforcement applications also share very similar information processing needs.

Translating recent advances in our ability to gather, store, and analyze a wide variety of data on multiple, geographically distributed, heterogeneous and often autonomously owned and operated hardware and software platforms into significant breakthroughs in distributed problem solving and decision making in scientific, military, commercial, and industrial applications presents some challenges in several areas of computer and information science and technology. This paper proposes a modular and and scalable approach to addressing these challenges by building distributed knowledge networks (DKN). The rest of the paper is organized as follows: Section II briefly describes the key technical problems that need to be addressed in the design and implementation of DKN. Section III describes our approach to the design of DKN and briefly discusses the current state of the implementation and applications of DKN. Section IV concludes with a summary and discussion of some directions for future research.

II. Technical Problems

Some of the key technical problems that need to be addressed in the design implementation, and success-

Vasant Honavar's research was partially supported by the National Science Foundation (through grant IRI-9409580) and the John Deere Foundation.

ful deployment of distributed knowledge network tools for practical applications are summarized below.

In many instances, the data sources are geographically distributed. This calls for the use of *information assistants* or *software agents* for intelligent, selective, and context-sensitive data gathering and data assimilation prior to large scale data analysis. Hence, DKN include tools for monitoring different data sources and routing the appropriate information selectively to relevant sites or specific users. Since the information of interest is user and context-dependent, the tools are customizable to specific users and information contexts.

Given the large volumes of data involved, it is desirable to perform as much analysis as feasible at the sites where the data is located and transmit only the results of analysis rather than flooding the network with data. Hence, DKN use *mobile* software agents that can transport themselves to appropriate sites, carry out the computation on site, and return with useful results.

Since the data sources are often autonomously owned and operated, and reside on heterogeneous hardware and software platforms, their effective use requires a sufficient degree of interoperability among the different data sources (despite their heterogeneity). For example, military applications have to access and utilize data from multiple sensors, intelligence sources, etc. Hence, DKN make use of intelligent software agents to provide seamless access to such data sources.

The data sources contain multiple types of data (text, images, relational databases, sequence data, spectrograms, protein structures, etc.) Hence, DKN provide tools for extracting, transforming, and assimilating relevant information from heterogeneous data sources into a *data warehouse* where it can be further analyzed to facilitate knowledge discovery.

The data sources are dynamic (i.e., they change quite rapidly over time as data items are added or modified). Hence, DKN include software agents that detect and propagate the changes and trigger the necessary updates in the affected data and knowledge repositories. For example, in military applications, interpretation of data from satellite images might be influenced by the contents of independently gathered intelligence reports.

The large volumes of data, the range of potentially relevant and useful complex interrelationships that need to be discovered, and the diversity of data sources challenge state-of-the-art approaches to *data mining* and *knowledge discovery*. Hence, DKN adapt and extend current statistical and artificial intelligence tools to support data-driven knowledge acquisition and incremental theory refinement from multiple heterogeneous, structured as well as semi-structured data and knowledge sources (including multiple types of sensor data, text, images, etc.)

Design of complex information systems in general, and knowledge networks in particular, in order to be feasible, often requires modular design which involves the decomposition of the overall task into more manageable subtasks. Hence, DKN consist of *multi-agent systems* that are made of multiple, more or less autonomous agents, each of which is responsible for a data source (e.g., an independently managed database) or analysis capability (e.g., a knowledge discovery tool). Modular design also lends itself to being adapted and extended for a broader class of knowledge network applications. In order to ensure satisfactory operation of multi-agent systems, DKN include mechanisms for coordination and control of collections of agents.

In addition to the technical problems listed above, additional issues that need to be addressed have to deal with reliability, fault-tolerance, performance, and security of the infrastructure that is needed to provide the necessary connectivity among the distributed data and knowledge sources and users. Discussion of these issues is beyond the scope of this paper.

III. Design of Distributed Knowledge Networks

Our current design of DKN consists of the following components: a mobile agent infrastructure; intelligent agents for information retrieval, information extraction, assimilation, and knowledge discovery; and coordination and control mechanisms for multi-agent systems. The system consists of modular and extensible, object-oriented toolkits for rapid design and prototyping of multi-agent systems for different applications.

Mobile agent technology [1], facilitated by recent advances in computers, communications, and artificial intelligence, provides an attractive framework for the design and implementation of *communicating applications* in general, and distributed networks in particular [1], [2]. A mobile agent is a named object which contains code, persistent state, data and a set of attributes (e.g., movement history, authentication keys) [1] and can move about or transport itself from one host to another as needed for accomplishing its tasks. Mobile agents provide a potentially efficient framework for performing computation in a distributed fashion at sites where the relevant data is available instead of expensive shipping of large volumes of data across the network.

There is considerable ongoing research on mobile agent infrastructures (MAI) [1], [3]. Most MAI designs consist of at least three components: agent servers, agent interface, and agent brokers (service directory). Agent servers support basic agent migration mechanisms, authentication, and sometimes provide other services. Agent brokers provide addresses of agent servers and support mechanisms for uniquely naming agents and agent servers. The agent interface is used by application programs to create and interact with agents. Recently, a consortium of several companies and research groups has proposed standards (MAF) for key aspects of mobile agent infrastructure to facilitate interoperability among different mobile agent systems (despite their different architecture, design, and implementation choices).

We have explored DKN prototypes using the commercially available Voyager mobile agent infrastructure. We have also designed and implemented an MAF-compliant, platform independent, MAI [3] in Java using *Common Object Request Broker Architecture* (CORBA) for managing distributed objects. Experiments with DKN designs using different mobile agent platforms are aimed at evaluating the portability and interoperability of DKN across multiple MAIs.

Intelligent agents – software entities that perform specific tasks on behalf of users with varying degrees of autonomy and intelligence [4], [2] – offer an attractive approach to the design of DKN components. Of particular interest are reactive agents which respond reactively to changes that they perceive in their environment, deliberative agents that plan and act in a goal-directed fashion, utility-driven agents that act in ways designed to maximize a suitable utility function, learning agents which modify their behavior as a function of experience, and agents that combine different modes of behavior [5], [2].

The prototype DKN systems that we have implemented include intelligent agents for customized information retrieval, information assimilation, and knowledge discovery functions. Customizable information retrieval agents acquire user preferences using machine learning techniques [6], [7], [8], [9] and have been successfully incorporated into mobile agents for selective retrieval of journal articles, news articles, email messages, etc. from remote sites [9], [8]. Current work is aimed at extending the capability of customizable information assistants for monitoring multiple, heterogeneous data sources (e.g., traditional databases, image data, scientific databases (e.g., genome and protein databanks), etc.) for different applications.

Use of data from heterogeneous data sources residing on multiple hardware platforms and operating systems at different geographical locations requires a robust and flexible framework for interoperability between the various data sources and clients. Interoperability among multiple hardware and software platforms while using standard relational databases has become easier to manage with the availability of platform-independent tools such as the JDBC and CORBA. However, DKN have to be able to provide seamless access to data that is distributed over multiple relatively autonomous databases, as well as data that is heterogeneous in form and/or content.

Approaches to processing heterogeneous data sources can be broadly classified into two categories: *multidatabase systems* [10] and *mediator based* systems [11]. The multidatabase systems [10] apply traditional database techniques to bridge the mismatch between the underlying data sources. Our work has approached this task using object-oriented views [12] which exploits the underlying object structure for incorporating the rich semantics of the common data types. It creates a uniform interface to multiple databases to hide the heterogeneity and the distributed nature of the underlying data sources. Mediator based systems [11] offer an approach to bridging the mismatch between heterogeneous data sources. In a mediator-based system, new data sources can be added by simply formulating a set of rules that define the new data source. Although no current implementation offers the full range of capabilities envisioned in [11], several implementations are underway.

The current design of DKN takes a pragmatic approach to heterogeneous database interoperability. It borrows from both the multidatabase as well as the mediator-based approaches to design and implement an object-oriented *data warehouse* [13], [14] based on object-oriented views using knowledge-based software agents. Data source, domain, and application-specific knowledge is used to extract, transform, assimilate, and organize information from multiple heterogeneous data sources in one or more data warehouse(s) [15] in a form that is suitable for selecting information using sophisticated queries as well as further analysis (e.g., using data-driven knowledge acquisition and theory refinement). Current work is aimed at building on and extending this system to incorporate a broad range of tools for data transformation and information extraction from different types of data. These tools will include software agents for extracting information from different types of data including text (using domain-specific text analysis) and images (using domain-specific image analysis).

Given the large volume, diversity, and variety of data to be analyzed in specific applications and the range of scientifically relevant but complex relations that are likely to exist among them, DKN have to include sophisticated tools for data-driven knowledge discovery. *Machine learning* currently offers one of the most practical and cost-effective approaches to automated or semi-automated data-driven knowledge discovery and theory refinement. A variety of approaches to machine learning [16], [17] including artificial neural networks, statistical methods syntactic methods, rule induction and evolutionary techniques are available. The nature of the knowledge acquisition and discovery process and the choice of specific algorithms or tools to be used for the task depends on a number of factors [17], [16] such as: Overall objectives of the knowledge acquisition task (e.g., pattern classification, prediction, theory refinement, control); The nature and amount of *a priori* domain knowledge that is available; Choice of data (e.g., feature vectors, text, images) and knowledge representation (e.g., decision trees, neural networks, rules); The amount as well as the quality of data that is available; and Whether all of the necessary data is available for learning from data or whether there is a need for *incremental* knowledge refinement as new data is gathered.

Prototype DKN designs that have been implemented to date include representative algorithms from several machine learning paradigms which are part of a modular and and extensible knowledge discovery toolkit. They have been used to customize information retrieval agents [9], [8], discover patterns that characterize intrusion in computer systems [18], and various data-driven knowledge discovery applications [19], [20]. Current work is aimed at extending the machine learning tools discovery and theory refinement to handle heterogeneous data sources (including images, text, and different types of sensor data).

Multi-agent systems are natural consequences of a modular approach to designing complex distributed knowledge networks. In such multi-agent systems, satisfactory completion of the tasks at hand depend critically on effective communication and coordination among the agents. In order to harvest the potential power of such systems in practical applications, especially when the individual agents might be autonomous (e.g., the agents associated with independently managed data sources as is often the case when dealing with distributed data sources in practical applications), it is essential that suitable mechanisms be devised to exercise adequate control over the behavior of such systems. In multi-agent systems, the notion of control suggests such functions as coordination among agents, synchronization among multiple agents, activation and deactivation of individual agents or groups of agents, selection among agents, creation of new agents when needed, elimination of agents that are no longer needed, adaptation of individual agents and agent populations to changes in the environments or task demands, learning (both at the individual as well as group levels) from experience, and (at a much slower timescale) evolution of agent populations toward more desirable behaviors. Both natural and artificial systems offer rich sources of examples of a wide variety of coordination and control mechanisms that can be beneficially incorporated into DKN [21], [22].

The current implementation of DKN includes a modular and extensible implementation of a general framework for inter-agent negotiation inspired by the *contract net protocol* (CNP) [23], [24], [25] which provides an attractive framework for negotiation and coordination among self-interested rational agents. Within this framework, each agent can announce tasks, make bids, evaluate bids made by other agents to complete the tasks, and offer contracts. Prototype DKN systems that utilize CNP have been implemented and tested on information retrieval and knowledge discovery applications. Current research is aimed at systematic evaluation of the performance of such coordination mechanisms as well as additional structures e.g. hierarchies and other modular organizations.

IV. Summary

We have presented a design for DKN. Several prototype systems have been implemented using this framework. Current work is aimed at refinements and extensions to handle a broader range of real-world applications of DKN.

References

[1] J. White, "Mobile agents," in *Software Agents*, J. Bradshaw, Ed. MIT Press, Cambridge, MA, 1997.

[2] V. Honavar, "Intelligent agents," in *Encyclopedia of Information Technology*, J. Williams and K. Sochats, Eds. Marcel Dekker, New York, 1998, To appear.

[3] J. Wong, V. Honavar, L. Miller, and V. Naganathan, "Design and implementation of mobile agent infrastructure based on mobile agent interoperability facilities (maf)," Submitted for publication.

[4] J. Bradshaw, *Software Agents*, MIT Press, Cambridge, MA, 1997.

[5] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[6] P. Maes, "Agents that reduce work and information overload," in *Software Agents*, J. Bradshaw, Ed. MIT Press, Cambridge, MA, 1997.

[7] T. Joachims, D. Freitag, and T. Mitchell, "Webwatcher: A tour guide for the world wide web," in *International Joint Conference on Artificial Intelligence*, 1997.

[8] J. Yang, P. Pai, V. Honavar, and L. Miller, "Mobile intelligent agents for document classification and retrieval: A machine learning approach," in *14th European Meeting on Cybernetics and Systems Research. Symposium on Agent Theory to Agent Implementation*, Vienna, Austria, 1998.

[9] J. Yang, V. Honavar, L. Miller, and J. Wong, "Intelligent mobile agents for information retrieval and knowledge discovery from distributed data and knowledge sources," in *IEEE Information Technology Conference*, Syracuse, NY, 1998.

[10] A. Sheth and *et al.*, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Computing Survey*, 1990.

[11] G. Wiederhold, "The conceptual basis for mediation services," *IEEE Expert*, vol. 12, no. 5, Sep-Oct 1997.

[12] C. Yen and L. Miller, "An extensible view system for multidatabase integration and interoperation," *Integrated Computer-Aided Engineering*, vol. 2, no. 2, pp. 97–123, 1995.

[13] W. Inmon, "The data warehouse and data mining," *Communications of the ACM*, vol. 39, no. 11, pp. 49–50, 1996.

[14] H. Gupta, "Selection of views to materialize in a data warehouse," in *Proceedings of the International Conference on Database Theory*, 1997.

[15] L. Miller, V. Honavar, and J. Wong, "Object-oriented data warehouses for information fusion from heterogeneous distributed data and knowledge sources," in *IEEE Information Technology Conference*, Syracuse, NY, 1998.

[16] T. Mitchell, *Machine Learning*, McGraw Hill, New York, 1997.

[17] V. Honavar, "Machine learning: Principles and applications," in *Encyclopedia of Electrical and Electronics Engineering*, J. Webster, Ed. Wiley, New York, 1998, To appear.

[18] G. Helmer, J. Wong, V. Honavar, and L. Miller, "Intelligent agents for intrusion detection and countermeasures," in *IEEE Information Technology Conference*, Syracuse, NY, 1998.

[19] R. Parekh and V. Honavar, "Constructive theory refinement in knowledge based neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, Anchorage, Alaska, 1998.

[20] K. Balakrishnan and V. Honavar, "Intelligent diagnosis systems," *International Journal of Intelligent Systems*, 1998, In press.

[21] V. Honavar and L. Uhr, "Coordination and control structures and processes: Possibilities for connectionist networks," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 2, pp. 277–302, 1990.

[22] L. Uhr, *Algorithm-Structured Computer Arrays and Networks: Architectures and Processes for Images, Percepts, Models, Information*, Academic Press, New York, 1984.

[23] R. Smith, "The contract net protocol: High-level communication and control in a distributed problem solver," *IEEE Transactions on Computers*, vol. C-29, no. 12, pp. 1104–1113, December 1980.

[24] T. Sandholm, "Contract types for satisficing task allocation: I theoretical results," in *AAAI 1998 Spring Symposium: Satisficing Models*, 1998.

[25] J. Yang, R. Havaldar, V. Honavar, L. Miller, and J. Wong, "Coordination of distributed knowledge networks using contract net protocol," in *IEEE Information Technology Conference*, Syracuse, NY, 1998.