Distributed Knowledge Networks

Vasant Honavar
Artificial Intelligence Research Laboratory
Department of Computer Science
Iowa State University, Ames, IA, U.S.A.
honavar@cs.iastate.edu
www.cs.iastate.edu/~honavar/aigroup.html
Background and Motivation

- Many practical applications call for reactive as well as proactive information processing under tight time, resource, and performance constraints
  - computer-aided scientific discovery (e.g., computational genomics)
  - organizational decision support (e.g., intelligence data handling, situation assessment, command and control)
  - distributed design and manufacturing
  - medical decision support
  - monitoring and control of complex dynamic systems (e.g., power systems, communication networks)
Background and Motivation

- Advances in sensor, high throughput data acquisition, and digital storage technologies have made it possible to acquire and store large volumes of data.
- Advances in computers, communications, Internet and mobile computing have made it possible, in principle, for scientists to have at their disposal, data as well as analysis and decision support tools residing on multiple, distributed, heterogeneous hardware software platforms.
- Sophisticated tools are needed for selective, reactive, proactive information retrieval, knowledge discovery, distributed problem solving to translate these advances into fundamental scientific advances and technological advances (rational drug design).
Technical Challenges

- Interoperability of geographically distributed, heterogeneous, possibly mobile data and knowledge sources and clients
- Selective, proactive, reactive, context-sensitive, customizable information retrieval, extraction, and fusion, using heterogeneous, autonomous, dynamic data and knowledge sources (e.g., data acquisition devices, sensors, simulations)
- Discovery of a-priori unknown, complex relationships (e.g., predictive rules) or knowledge that characterize the specific domain (e.g. patterns of activity that predict intrusions in a distributed computer network)
- Distributed, autonomous or semi-autonomous, collaborative problemsolving
- Reliability, fault tolerance, performance, scalability, security
The Distributed Knowledge Networks Solution

- A modular, extensible and open system of multiple autonomous software agents
- Mobile software agents to support communicating applications
- Intelligent software agents for information retrieval, extraction, fusion
- Object-oriented views and mediators to support interoperability among heterogeneous data and knowledge sources
- Data-driven knowledge acquisition and knowledge discovery using machine learning from distributed, dynamic data sources
- Tools to support multi-agent problem solving through inter-agent negotiation, communication, and coordination
Mobile Agents for Communicating Applications

- A Distributed Knowledge Network consists of communicating applications that use public or private networks to locate, gather, assimilate, analyze, and use information from heterogeneous distributed data and knowledge sources
- Mobile agent infrastructure helps transform public networks into computing platforms that can support the development and deployment of communicating applications: e.g., a mobile agent can visit multiple data and knowledge sources and assist in information fusion and data-driven knowledge discovery
Mobile Agent Infrastructure

- A Mobile agent consists of program code, persistent internal state, and other attributes (e.g., travel plan, access privileges)
- A mobile agent moves from host to host as needed to accomplish its tasks - can combine knowledge and data from the client and the server to perform inference on the server (where data and computing resources are available)
- A mobile agent supports ongoing interaction without ongoing communication - reduced bandwidth requirements, robustness (e.g., when access is intermittent)
- A mobile agent enables small, lightweight mobile devices (computers, sensors, etc.) to interact with heavyweight applications running on distant hosts
Mobile Agent Infrastructure (MAI)

- MAI provides a host-independent execution environment for mobile agents
- MAI provides standard communication languages that can be used for interaction between servers and agents
- MAI provides support for creation, deployment, transportation, authentication, management of mobile agents
  - agent servers (agent migration, authentication)
  - agent brokers (agent naming, finding)
  - agent interface (agent interaction with applications)
- Examples of MAI: Voyager (available from Objectspace, Inc.), JavaMob (developed at ISU AI Lab)
Intelligent Software Agents

- Intelligent software agents are software entities that continuously sense dynamic conditions in their environment, draw inferences, choose and perform actions to accomplish specific tasks with some degree of autonomy.
- Agents can be reactive, deliberative, proactive, goal-directed, communicative, collaborative, rational, adversarial, adaptive, learning, self-replicating, persistent, …
- Agent designs build on a rich body of research in Artificial Intelligence
- DKN call for a modular, extensible, toolbox of agent designs that can be adapted for a broad range of applications and tools for rapid specification, prototyping, and validation of agent designs
Intelligent Software Agents in DKN

- Intelligent agents can automate routine tasks of proactively, reactively, context-sensitively and selectively locating, retrieving, extracting, analyzing, summarizing, fusing, updating information from multiple sources.
- Intelligent agents can customize their behavior to specific users or specific contexts by learning from interaction with users and environments.
- Intelligent agents can hide the underlying complexity and heterogeneity of the underlying data and knowledge sources.
- Intelligent agents can accomplish tasks specified at a high level by composing appropriate action sequences (plans).
- Intelligent agents can provide smart customizable interfaces to generic software (e.g., news readers, web browsers, databases).
Intelligent Software Agents in DKN

- Prototype DKN systems that we have implemented to date include intelligent agents for
  - customized text retrieval from distributed text repositories using machine learning of user profiles
  - knowledge discovery and data-driven theory refinement using machine learning from distributed structured data sources
  - information fusion using an object-oriented data warehouse
  - monitoring multiple dynamic data and knowledge sources for information of interest
  - distributed problem solving (e.g., task allocation) through inter-agent negotiation using the contract network protocol
Heterogeneous Data and Knowledge Source Interoperability

- Types of heterogeneity
  - Heterogeneous hardware and software platforms
  - Multiple data sources (relational data, object-oriented data, text, images, sensor signals, etc.) with different structure and semantics
- Approaches to interoperability of multiple data sources
  - Relational views and tools like JDBC offers support for interoperability among traditional relational databases
  - Object-oriented views to incorporate rich semantics of common data types and create a uniform interface to multiple data sources
  - Intelligent agents (mediators) offer a framework for bridging mismatch between different data sources and clients
Heterogeneous Data and Knowledge Source Interoperability

- The current design of DKN takes a pragmatic approach to heterogeneous data and knowledge source interoperability
  - Heterogeneity of platforms is handled using mobile and static agents or simple mediators supported by MAI
  - Heterogeneity of data structure and semantics is handled using extended object-oriented views embedded in software agents
  - Extended object-oriented views can extract and represent relevant information in a structured form (e.g., slot and filler representation of text, feature vector representation of images, etc.) to facilitate further analysis
  - Local or distributed object-oriented data warehouses are used to support information extraction, transformation, and assimilation
Data-Driven Knowledge Discovery in DKN

- Machine learning currently offers the most cost-effective and practical approach to data-driven knowledge acquisition.
- The choice of specific machine learning algorithm depends on the objectives of the knowledge discovery task, the nature and amount of data and prior knowledge available, etc.
- DKN call for a wide variety of machine learning algorithms for knowledge discovery from heterogeneous structured, semi-structured, and unstructured data and knowledge sources.
- The current design of DKN includes a variety of learning algorithms for induction of neural networks, decision trees, rules, grammars, feature subset selection and feature construction.
- Current focus is on algorithms for *incremental* data-driven theory refinement and transformations for handling unstructured data.
Multi-Agent Coordination in DKN

- Effective working of DKN requires mechanisms for inter-agent coordination and communication, activation and deactivation, selection, adaptation, and dynamic self-organization of autonomous or semi-autonomous agents or groups of agents
- Approaches to coordination and control of multi-agent systems
  - Inter-agent communication using shared syntax and semantics (e.g., KQML and related approaches)
  - Inter-agent negotiation (e.g., contract net protocol)
  - Agent organization (e.g., structured agent hierarchies)
  - Adaptation
- DKN prototypes that we have built and experimented with to date have used simple agent organizations and inter-agent negotiation using the contract net protocol
Sample Applications of Distributed Knowledge Networks

● Current work on DKN systems in our lab targets
  ■ automated selective information retrieval, extraction, fusion, and knowledge discovery in computational molecular biology where there is an abundance of heterogeneous, distributed data sources (e.g., genome data, protein data, etc.)
  ■ monitoring and control of complex distributed systems (e.g., intrusion detection and counter-measures in computer networks, security assessment of power systems)

● Other applications of interest include
  ■ organizational decision support systems
  ■ healthcare information systems
  ■ intelligence data handling and situation assessment
Distributed Knowledge Networks for Intrusion Detection

● Increased reliance on networks and growth in connectivity between computers has made critical systems more susceptible to attacks by intruders.

● The different approaches developed to counter attacks are
  ■ Authentication by means of user name and password.
  ■ Using firewalls.
  ■ Detecting intrusions and using countermeasures.

● An intrusion can be defined as a set of actions by an *intruder* or an authorized person that attempt to compromise the integrity, security, confidentiality, or availability of a resource.
Distributed Knowledge Networks for Intrusion Detection

- An intrusion detection system monitors the system and attempts to detect any activity that may be undesirable for the system from a security perspective
  - Anomaly Detection.
  - Misuse Detection.
Distributed Knowledge Networks for Intrusion Detection

- Anomaly detection system establishes a profile of normal system activity. Deviations from the normal pattern get flagged as potential intrusions
  - Able to detect unknown intrusions and can adaptively learn the behavior of users and modify the profile but .. they can gradually be trained by intruders so that eventually, intrusive events are considered normal.

- Misuse detection systems typically have a knowledge base of known problems (e.g. security loopholes in mail programs or network protocols) and look for event sequences that indicate intrusions
Distributed Knowledge Networks for Intrusion Detection

- Most existing intrusion detection systems (e.g., Haystack, NADIR, IDES) are designed to look for specific types of problems at specific sites
- Recent efforts have focused on the development of multi-host network based intrusion detection systems (e.g., DIDS)
- Coordinated attacks (e.g., the Langley Cyber attack) over large networks call for intrusion detection systems that can monitor a host of distributed, heterogeneous sensors, system log files, etc. in real time and look for patterns of activity that indicate intrusions
- Distributed Knowledge Networks offer a promising approach to robust and responsive distributed system monitoring and protection
Distributed Knowledge Networks for Intrusion Detection

- Static and Mobile agents gather data from multiple distributed sensors and log files
- Data fusion agents assimilate the data into a data warehouse to provide a global picture of the situation
- Knowledge discovery agents analyze the data as it is being gathered to acquire useful predictive rules using a variety of machine learning techniques
- Detection and response agents look for evidence of attacks, correlate activities that are distributed over space and time, and alert the relevant staff
- Prototype system has been implemented and tested on intrusions involving certain abuses of sendmail and network protocols
Acknowledgements

● Students
  + Chun-Hsien Chen, Ph.D. 1997
  + Rajesh Parekh, Ph.D., 1998
  + Karthik Balakrishnan, Ph.D., 1998
  + Jihoon Yang, Ph.D., 1998
  + Rushi Bhatt
  + Doina Caragea
  + Fajun Chen
  + Taru Sharma
  + Adrian Silvescu
  + Kent van der Velden
  + Guy Helmer

● Collaborators
  + Leslie Miller
  + Johnny Wong
Acknowledgements

This research was partially supported by grants from:

- National Science Foundation (9409580)
- John Deere Foundation
- Department of Defense
- Department of Energy
- Carver Foundation
- IBM
- Iowa State University Graduate College