

Center for Research in Wireless Mobility and Networking (CReWMaN)

Pervasively Secure Infrastructures (PSI) Grant No. IIS-0326505

NSF Third Annual Report (June 2005 - May 2006)

<http://crewman.uta.edu/psi>

Abstract

The third year of the PSI project was marked with increased effort in translating research ideas to full-fledged prototype development in overlapping domains of Wireless Sensor Networks, Computer and Network Security, Pervasive Computing, Machine Learning, and Databases. We have made significant gains in mission programming and decision making logic of sensor nodes, new routing models for mobile ad-hoc networks, and in understanding the reach of malicious programs in wireless sensor networks. In pervasive computing, we have been successful in developing an initial prototype for a middleware aimed at data fusion, aggregation and intelligent dissemination from heterogeneous sensors, RFIDs (active and passive), cellphones, PDAs, and laptops. In understanding relational information, we have extended the capability of our “Subdue” data learning and mining system to increase our ability in identifying potential threats. We have also made significant progress in expanding and perfecting the capabilities of “InfoFilter” and related pattern specification language (PSL) so as to understand complex patterns such as combinations of sequential, structural patterns, wild cards, and word frequencies. An early prototype of “InfoFilter” is available for initial evaluation. At the same time we are exploring ideas and methodologies in modeling complex biological systems, network traffic characterization, and understanding the closed form dynamics of network and system interaction; all of which supports our vision of building a Pervasively Secure Infrastructure.

1 Wireless Sensor Networks (WSNs) and Network Security.

Faculty Involved: Kalyan Basu, Sajal Das, Frank Lewis.

Student Members: H. Ammari, H. J. Choe, J. Ho, P. De, P. Ballal, S. Kundu, W. Zhang.

WebPage: <http://www.crewman.uta.edu/psi>, <http://arri.uta.edu/acs/>

Decision-Making and Fusion for Mobile Wireless Sensor Networks

Coordination and control of distributed wireless sensor networks is a challenging task. Data from sensors of different types must be collected, combined, and interpreted, and alarms must be responded to. We have developed an efficient computational framework for coordinating sensory activities of a WSN with mobile nodes that perform as sentries to respond to various alarm situations. The framework is a discrete

event (DE) controller based on matrices, and allows for efficient simulation of WSN, as well as decision-making and resource assignment on-line. The matrix DE controller provides a method for integrating decision-making of different sorts, including fuzzy logic and Dempster-Shafer, both with respect to sensor fusion and data interpretation, as well as control coordination with conflicting commands. Results are implemented on a new mobile WSN Testbed at UTA's Automation & Robotics Research Institute (ARRI).

Mission Programming for Deployable Wireless Sensor Networks

In this work, we have applied a matrix-based discrete event controller (US Patent) to wireless sensor networks for mission programming, supervision in task sequencing, and resource assignment. The needs of WSN are not the same as other discrete event systems, and we modified the controller to allow for multiple missions, mission priority interrupts, and fast assignment of sensors without blocking phenomena. The WSN consists of some fixed unattended ground sensors, and some mobile sentry nodes that can vary their location to enhance the capabilities of the WSN in terms of repairing damage, compensating for faults, providing additional sensor information, and responding to detected events. The discrete event controller assigns the mobile nodes to tasks to carry out programmed missions, and has alarm capabilities when certain events are detected.

Localization of Deployed WSN

Two new methods have been developed for localization of deployable WSN. The first method uses a mobile robot node with GPS information to localize the unattended ground sensors. Kalman filters are run on the mobile robot and on each UGS. Once localized, the UGS can help in the navigation and positioning of the mobile robot. An adaptive localization scheme was developed that allows the mobile robot to proceed as needed to localize the nodes in the fastest and most efficient manner. In the second method, a potential field approach was developed to localize a deployed WSN using only range information. A Lyapunov method based on potential fields was developed and it was proven that the method converges. An algorithm was developed for absolute positioning of the WSN when at least three nodes also know their absolute coordinates, e.g. using GPS.

Adaptive Sampling in WSN (jointly with Dan Popa)

In this work, we have developed an adaptive sampling algorithm in mobile WSN for environmental monitoring. A dynamical model is developed that describes both the robot mobility and the uncertainty due to data gathering. This allows combining navigation and measurement so that the mobile sensors are directed to make measurements at those locations that provide the greatest decrease in uncertainty. A Bayesian method based on error covariances allows prediction of the best next-sampling locations. We have built robots and implemented these algorithms in the ARRI WSN lab.

Modeling the degree of node compromise in WSNs using Epidemic Theory

We have mathematically modeled the process of node compromise spread based on Epidemic Theory and studied the effects of various node deployments on the process. We assume the nodes in a sensor network to be securely communicating with each other using secret keys shared between them. In the event of a capture of a sensor node we assume that all its keys are known by the adversary. In such a situation, we observe the process by which a captured node gradually compromises other nodes by securely communicating with it and transmitting malware. Since a sensor network is generally static in nature, the assumption of homogeneous mixing of the individual nodes, which is done normally in a differential rate equation based formulation in Epidemic Theory, is not applicable. Therefore, we approach the problem from a random graph theoretic standpoint. The key parameters that we try to identify are the exact points when the node compromise process scales into an epidemic and affects the whole network. Moreover, since a random graph theoretic approach is unable to capture the temporal effects of the spread of this node compromise, we performed simulations to do that. In our simulation study, we observed the way the node compromise process behaves with time. We studied the process

under two assumptions. One, when there is no node recovery process and second, when there is a recovery process underway. We are also studying the process of virus/malicious code spread in sensor networks piggybacking on broadcast protocols that have been developed in sensor networks. Sensor networks are especially vulnerable through the way Broadcast protocols work because of the scale and density of such networks. Therefore, careful study of the vulnerability of broadcast protocols is necessary to ultimately come up with mechanisms to make them secure.

Trust based framework for secure aggregation in WSNs

In unattended and hostile environments, node compromise can become a disastrous threat to wireless sensor networks which by nature process limited resources and hence defense capability. A compromised node often tends to completely reveal its secrets to the adversary that in turn renders purely cryptography-based approaches vulnerable. In this work, we propose an integrative, systematic framework for defending against compromised nodes to secure information gathering and aggregation. Our approach is a trust based framework rooted in sound statistics and some other distinct and yet closely coupled techniques. Specifically, by extracting the statistical characterization from the gathered information, an information theoretic concept, Kullback-Leibler (KL) distance, is introduced to evaluate the trustworthiness (reputation) of each individual sensor node. And based on the reputations, an unsupervised learning algorithm is developed to dynamically detect the compromised nodes. Moreover, whenever aggregation is performed, an opinion, a metric of the degree of belief, is generated to interpret the trustworthiness in the aggregation result. As the result is being disseminated and assembled through the routes to the sink, this opinion will be propagated and regulated by the powerful Josang's belief model. Through this way, the uncertainty within the data and aggregation results can be effectively quantified and reasoned. Our simulation results show that our trust based framework provides a powerful mechanism for detecting of malicious nodes and can purge the false data, thus, effectively accomplish robust aggregation in the presence of compromised nodes.

Aggregation survival authentication in wireless sensor networks

In wireless sensor networks (WSNs), authentication is usually implemented via message authentication code (MAC). Although MAC is effective and efficient in terms of computation and communication, it has some limitations. The main disadvantage of MAC is that it cannot stand even slightest modification after generated. However, in WSNs, in-network processing/aggregation is one primary operation for the intermediate nodes to save energy, but MAC fail to fulfill end-to-end authentication whenever in-network processing is performed. By visualizing the densely deployed WSN as an image and further employing watermarking techniques, we propose a new authentication scheme that is resistant to in-network processing in this work. In the proposed scheme, a robust, spatial watermark is superposed into the data sent by each sensor in such a way that the watermarked data can not only survive intermediate nodes' aggregation (compression) but also authenticate the origin of the data. So, our scheme realizes end-to-end authentication while coexisting with in-network processing. The analytical and simulation results show that our aggregation resistant scheme can detect the illegitimate modification and locate the position where the attack occurs, that is, it accomplishes end-to-end authentication.

Energy-efficient Data Dissemination

The design of wireless sensor networks faces several challenges, such as fault tolerance, connectivity, coverage, energy efficiency, and efficient data dissemination, to name a few. We are particularly interested in energy-efficient data dissemination with a goal to extend the network lifetime. There is an ongoing debate on whether long-hop (or single-hop) is better than short-hop (or multi-hop) data dissemination in wireless sensor networks. Believing that energy savings is the most important constraint that should be met for network lifetime elongation, we have designed an energy-efficient data dissemination protocol for wireless sensor networks based on the remaining energy of the sensors and exploiting the geometric properties of Voronoi diagram and Delaunay triangulation. Our theoretical results have proved that our proposed protocol outperforms existing ones such as BVGF and GPSR protocols, which favor data forwarding through long distances. We have also considered dynamic (or energy-aware) Voronoi diagram and sink mobility together to balance the load on all the sensors in the network and hence increase

the network lifetime. Delay is another metric that some sensing applications take into account to meet specified time constraints. Thus, we proposed a data dissemination protocol for wireless sensor networks that trades off between energy savings and delay. This protocol offers certain flexibility to network designers to meet the specific requirements of sensing applications in terms of delay and energy efficiency.

References

- [1] S. S. Ge and F.L. Lewis, "Autonomous Mobile Robots: Sensing, Control, Decision-Making, and Applications", *CRC Press*, 2006.
- [2] G. Vachtsevanos, F.L. Lewis, M. Roemer, A. Hess, B. Wu, "Intelligent Fault Diagnosis and Prognosis for Engineering Systems", *John Wiley, New York*, 2006, to appear.
- [3] S. Bogdan, F.L. Lewis, Z. Kovacic, and J. Mireles, "Manufacturing Systems Control Design: A Matrix Based Approach", *Springer-Verlag, London*, 2006, to appear.
- [4] V. Giordano, F.L. Lewis, P. Ballal, and B. Turchiano, "Supervisory control for task assignment and resource dispatching in mobile wireless sensor networks" in *Cutting Edge Robotics*, ed. V. Kordic, pp. 133-152, 2005.
- [5] D.O. Popa and F.L. Lewis, "Algorithms for robotic deployment of WSN in adaptive sampling applications", in *Wireless Sensor Networks and Applications*, ed. Y. Li, M. Thai, and W. Wu, Springer-Verlag, Berlin, 2005. Journal Papers
- [6] A. Tiwari and F.L. Lewis, "Energy-efficient wireless sensor network design and implementation for condition-based maintenance", *ACM Trans. On Sensor Networks*, 2006, to appear.
- [7] V. Giordano, P. Ballal, F.L. Lewis, B. Turchiano, J.B. Zhang, "Supervisory control of mobile sensor networks: Matrix formulation, simulation and implementation", in *IEEE Trans. Systems, Man, Cybernetics, Part B*, to appear, 2006.
- [8] V. Giordano, F.L. Lewis, J. Mireles, B. Turchiano, "Coordination control policy for mobile sensor networks with shared heterogeneous resources", in *Proc. Int. Conf. Control and Automation*, pp. 191-196, Budapest, June, 2005.
- [9] V. Giordano, F.L. Lewis, B. Turchiano, P. Ballal, V. Yeshala, "Matrix computational framework for discrete event control of wireless sensor networks with some mobile agents", in *Proc. Mediterranean Conf. Control and Automation*, Limassol, Cyprus, June 2005. This paper won an award at MED 05.
- [10] O. Kuljaca, N. Swamy, J. Gadewadikar, F.L. Lewis, "Transfer Function Illustration With Simple Electronic Circuits", in *Proc. XXVII Int. Meeting MIPRO 2005, CE, Conference on Computers in Education*, 2005.
- [11] D.O. Popa, K. Sreenath, and F.L. Lewis, "Robotic deployment for environmental sampling applications", in *Proc. Int. Conf. Control and Applics.*, pp. 197-202, Budapest, June 2005.
- [12] V. Giordano, F. Lewis, B. Turchiano, P. Ballal, and V. Yeshala, "Matrix computational framework for discrete event control of wireless sensor networks with some mobile agents", in *Proc. Mediterranean Conf. Control and Automation*, pp. 176-181, Limassol, Cyprus, June 2005.
- [13] P. De, Y. Liu, and S. K. Das, "Modeling Node Compromise Spread in Sensor Networks using Epidemic Theory", in *World of Wireless, Mobile and Multimedia Networks, WoWMoM*, 2006.
- [14] P. De, K. Basu, and S. K. Das, "RFArch : An RFID based Pervasive Architectural Framework for Object Tracking, Distribution and Recall", in *IEEE Transactions on Mobile Computing (TMC)*, under submission.
- [15] H. M. Ammari and S. K. Das, "An Energy-Efficient Data Dissemination Protocol for Wireless Sensor Networks", in *Proc 2nd IEEE Int. Workshop on Sensor Networks and Systems for Pervasive Computing (PerSenS)*, in conjunction with PerCom 2006, Pisa, Italy, March, 2006.
- [16] H. Luo, J. Luo, Y. Liu, and S. K. Das, "Routing Correlated Data with Fusion Cost in Wireless Sensor Networks", in *IEEE Trans. on Mobile Computing (TMC)*, 2006, to appear.
- [17] H. Luo, J. Luo, Y. Liu, and S. K. Das, "Adaptive Data Fusion for Energy Efficient Routing in Wireless Sensor Networks", in *IEEE Trans. on Computers*, 2006, to appear.

- [18] A. Ghosh and S. K. Das, “A Distributed Greedy Algorithm for Connected Sensor Cover in Dense Sensor Networks”, in *IEEE /ACM Intl Conference on Distributed Computing in Sensor Systems (DCOSS)*, July, 2005.
- [19] H. M. Ammari and S. K. Das, “Data Dissemination to Mobile Sinks in Wireless Sensor Networks: An Information Theoretic Approach”, in *Proc. 2nd IEEE Int. Conf. on Mobile Ad-hoc and Sensor Systems (MASS)*, Washington, DC, USA, Nov. 2005.
- [20] H. M. Ammari and S. K. Das, “Trade-off between Energy Savings and Source-to-Sink Delay in Data Dissemination for Wireless Sensor Networks”, in *Proc. 8th ACM/IEEE Int. Symp. on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM)*, Quebec, Montreal, Canada, Oct. 2005.
- [21] H. M. Ammari and S. K. Das, “Data Dissemination in Sensor Networks: Trading off Energy Savings and Source-to-Sink Delay”, in *ACM Student Research Competition (ACM SRC), The 11th Annual Int. Conf. on Mobile Computing and Networking (MobiCom)*, Cologne, Germany, 2005.

2 Middleware based on Community Computing for Information Fusion in Pervasive Environments

Faculty Involved: Mohan Kumar, Behrooz Shirazi, and Sajal K. Das

Student Members: M. J. Kim, H. Alex, N. Mallesh, and M. H. Ko

WebPage: <http://www.cse.uta.edu/pico@cse/>

MidFusion: Middleware for Information Fusion

Information acquired by large number of heterogeneous sensors needs to be integrated in a proactive, intelligent, and situation-aware manner to predict the occurrence of events (including security) in the PSI framework. In this project, we investigate the applicability of sensors significantly, by deploying collaborating software agents that meet the needs of dynamic applications. Two major challenges for proactive and real time collaboration among agents are: (1) heterogeneity of sensors, information representation and granularity and (2) fusion of uncertain, redundant, complementary and time sensitive information from various sensors. We investigate the coupling of sensors and associated agents for real time information fusion and decision making in distributed and dynamic applications. The agents cooperate in real-time to make intelligent and informed decisions using Bayesian Network reasoning. We propose and demonstrate a learning based approach called LEVeL to effectively measure the confidence in cooperating agent observations. We present MidFusion, an adaptive middleware architecture to facilitate information fusion in sensor network applications. MidFusion discovers and selects the best set of sensors or sensor agents on behalf of applications (transparently), depending on the quality of service (QoS) guarantees and the cost of information acquisition. We also provide the theoretical foundation for MidFusion to select the best set of sensors using the principles of Bayesian and Decision theories. A sensor selection algorithm (SSA) for selecting the best set of sensors has been designed, and its implementation on a real sensor network is being prototyped.

Centralized fusion points have many limitations:

- Nontrivial due to large quantities of information
- Inadequate communication and processing capacities
- Vulnerability to single points of failure
- Not well suited for dynamic systems as available information sources and their arrangements can change during the operation
- Centralized reasoning about the states of the fusion itself results in massive flows of information and additional processing.

A large class of fusion problems can be decomposed into smaller fusion problems. Fusion is backward reasoning from symptoms (observable) to events (unobservable). Causal models that link symptoms to events facilitate clear problem decomposition. Complex fusion problems can be accomplished through hierarchies of simpler tasks each corresponding to partial estimation problem. Besides most information

sources are typically very much hierarchical in nature. Problem decomposition means decomposition of the fusion problem into ordered sub tasks. Multi agent systems are suitable because they allow for encapsulation of sub tasks. Simple building blocks of the distributed and hierarchical information systems can be implemented through agents of different types dynamically organized into systems.

The community computing concept for creating services developed in the Pervasive Information Communities Organization (PICO) project, will be adopted for distributed information Fusion in pervasive computing applications. We will develop algorithms that will be adapted into the middleware framework to provide mechanisms for dynamic adaptation of applications to underlying changes in the information sources and sub fusion techniques. The use of the community computing model also makes a distributed information framework suitable for various information fusion models. The high level fusion model translates to the goal of the computing communities. The community members or delegates (or intelligent delegates) are autonomous members that perform sub fusion tasks of the high level fusion model.

Temporal Fusion using Dynamic Time Warping

Traditionally sensor fusion processes only concern fusing across raw data, features or decisions at specific points of time. However recently, there is a growing interest in inferring the behavioral aspects of environments or objects that are monitored by multisensor systems, rather than just their states at specific points in time. In order to infer environmental behaviors, it may be necessary to fuse data acquired from i) geographically distributed sensors at specific points of time and ii) specific sensors over a period of time. Fusing multisensor data over a period of time (also known as Temporal fusion) is a challenging task, since the data to be fused consists of complex sequences that are multidimensional, multimodal, interacting, and timevarying in nature. Additionally, performing temporal fusion efficiently in realtime is another challenge due to the large amounts of data to be fused. To solve this, we propose a robust and efficient framework that uses Dynamic Time Warping (DTW) as the core recognizer to perform online temporal fusion on either the raw data or the features. We evaluate the performance of the online temporal fusion system on two real world datasets: 1) accelerometer data acquired from performing two hand gestures and 2) a benchmark dataset acquired from carrying a mobile device and performing the predefined user scenarios. Performance results of the DTW based system are compared with those of a Hidden Markov Model (HMM) based system. The experimental results from both datasets demonstrate that the proposed system outperforms HMM based systems, and has the capability to perform online temporal fusion efficiently and accurately in realtime.

Architecture for Deploying Services in Heterogeneous Pervasive Environments

We propose VSD (Service Discovery based on Volunteers), a service discovery architecture/protocol for heterogeneous and uncertain pervasive computing environments. The proposed architecture is centralized, but flexible. The servicing area of one directory can be overlapped with those of other directories. A small subset of the nodes called volunteers, perform the directory services willingly in the system. More precisely, the term volunteer refers to a duo a software entity performing the directory services and a node hosting the software entity (a volunteer node). Relatively stable (less mobile) and capable (resourceful) nodes serve as volunteers. Cooperation among nodes is required for service discovery systems to perform seamlessly in pervasive computing environments. In adversarial environments, malicious nodes may harm honest ones. For instance, mean service providers may pretend to provide good services and make use of naive service requestors. Therefore, there is a need to monitor each nodes behavior and distinguish adversary nodes from good ones. Recognizing this need for secure operation, trust management can be incorporated into VSD architecture. The VSD architecture is not limited to any specific routing protocols or physical network media. The volunteers may appear to have the same roles as directory agents (service proxies or brokers) in existing protocols. In our scheme, volunteer operations take place only on certain nodes volunteer nodes. Indeed, the volunteer mechanism has been introduced to exploit node heterogeneity and unevenness that prevails in most existing pervasive systems. The followings are unique features of VSD:

- Auto-configures the network with directory services without any administration or explicit leader election mechanism since any node (ideally stable and resourceful) can perform directory services.

- Provides reliability in uncertain environments through overlapped servicing areas of directories (clusters).
- Allows the establishment of trust relationships among directories, service requestors and providers in open environments without prior relationship or knowledge.

For secure interaction among participants in open networks, we present a hierarchical distributed trust management scheme tightly integrated with authentication protocols of the middleware architecture. We define trust notation and operators, and develop trust evolution processes. In the proposed trust management scheme, trust values of clients are maintained globally and consistently in their communities, resulting in the decrease in total overhead compared to the distributed approaches. In addition, two user-transparent authentication protocols are also proposed to complete the trust management efficiently within. The proposed security mechanisms operate transparently and perform autonomously through cooperation amongst nodes. Simulation studies demonstrate that the proposed trust management protocols exhibit high efficiency and performance compared to existing distributed approaches. We validate that both belief and trust models achieve the basic goal of distinguishing good and bad nodes in malicious environments. The middleware architecture developed in the PICO project can be exploited to incorporate any existing trust and security mechanisms effectively. We have demonstrated the incorporation of security mechanism within service discovery as an application. However, the proposed mechanism can be applied to any service or application that requires interactions among nodes.

Completed Tasks

- We proposed and developed the MidFusion middleware architecture to facilitate information fusion in sensor network applications.
- The proposed Bayesian network based scheme efficiently meets the challenges of heterogeneity of sensors and information fusion.
- We developed a learning based method, LEVeL to integrate ontologically mapped information exchange between the agents into their corresponding BNs.
- We are also developing a prototype on sensor nodes to evaluate the feasibility of the proposed algorithms.
- Development of collaborating agent community framework that allows heterogeneous agents to interoperate, collaborate and perform decision making about application goals.
- A technique using Dynamic time warping (DTW) for temporal fusion of real-time data in sensor systems has been developed.
- The proposed DTW based method is shown to perform better than Hidden Markov Model (HMM) based methods for temporal fusion of data in real world applications.
- Developed an architecture for deploying services in heterogeneous and dynamic pervasive computing environments.
- Novel methodologies have been developed for incorporating trust and belief models into the middleware framework that has been developed for pervasive computing.

Work in Progress

- Incorporation of multi-agent based collaborative and distributed information fusion and decision making.
- Improvisation of the MidFusion mechanism to address issues of efficient computation and scalability.
- Development of suboptimal sensor selection algorithms for real-time applications by using context information and middleware techniques.
- We will develop techniques for the creation and composition of data enabled middleware services that create communities of data sources for information fusion.
- We will investigate DTWs for recognizing more complex multimodal sequences such as interleaved sequences, sequences with gaps and missing sub-sequences.
- We will investigate techniques for packet forwarding and replication and buffer management in Delayed Tolerant Networks (DTNs) in pervasive environments.

Development/Prototype

A prototype comprising large number of heterogeneous sensors, RFIDs (active and passive), cellphones, PDAs and laptops is under development. The prototype will depict a real-life situation and demonstrate the applicability of the proposed concepts.

References

- [1] Alex H, Kumar M and Shirazi B, MidFusion: An Adaptive Middleware for Information Fusion in Sensor Network Applications, Information Fusion Journal, Special Issue on Information Fusion in Distributed Sensor Networks, In Press.
- [2] Kim M, Kumar M, and Shirazi B, Service Discovery using Volunteer Nodes in Heterogeneous Pervasive Computing Environments, Elseviers Pervasive and Mobile Computing, Accepted for Publication.
- [3] Ko MH, West G, Venkatesh V, and Kumar M, Online Temporal Fusion in Multisensor Systems using Dynamic Time Warping, Under 2nd review, Information Fusion Journal.
- [4] H. Alex, M. Kumar, B. Shirazi, Collaborating Agent Communities for Information Fusion and Decision Making, International Conference on Knowledge Integration and Multi Agent Systems, Boston, USA, April 2005.
- [5] H. Alex, M. Kumar, B. Shirazi, MidFusion: Middleware for Information Fusion in Sensor Network Applications, International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP), Melbourne, Australia, December 2004.
- [6] H. Shen, M. Kumar, S. K. Das, and Z. Wang, "Energy-Efficient Data Caching and Prefetching for Mobile Devices Based on Utility," *ACM/Springer Journal on Mobile Networks and Applications* (Special Issue on Mobile Services, Guest Eds: Q. H. Mahmoud and U. Varshney), Vol. 10, pp. 475-486, 2005.
- [7] N. Roy, S. K. Das, K. Basu, M. Kumar, "Enhancing Availability of Grid Computational Services to Ubiquitous Computing Applications," *Proceedings of 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, Denver, Colorado, Apr 2005.
- [8] N. Banerjee, A. Acharya and S. K. Das, "SIP-based Mobility Architecture for Next Generation Wireless Networks," *Proceedings of 3rd IEEE Symposium on Pervasive Computing and Communications (PerCom)*, Kauai Island, Hawaii, Mar 2005.

3 Machine Learning, Data Mining and Computational Intelligence

Faculty Involved : Diane Cook, Lawrence Holder

Student Members : J. Potts, N. Ketkar, B. Eberle, S. Inavolu, S. Cannykr, and R. Hawes

WebPage: A.I Lab <http://ailab.uta.edu/> & <http://www.subdue.org>

Mining Graph Data

The ability to learn concepts from relational data has become a crucial challenge in many security-related domains. For example, the U.S. House and Senate Intelligence Committees' report on their inquiry into the activities of the intelligence community before and after the September 11, 2001 terrorist attacks revealed the necessity for "connecting the dots", that is, focusing on the relationships between entities in the data, rather than merely on an entity's attributes. The ability to discover relationship-driven patterns can impact our ability to prevent future attacks and ensure national security. A natural representation for relational information is a graph, and the ability to discover previously-unknown patterns in such information could lead to a significant improvement in our ability to identify potential threats. This research investigated approaches to improve the scalability and effectiveness of the Subdue

graph-based relational learning system. In this work we also extended Subdue to perform not only unsupervised discovery of patterns from graph data, but also supervised learning from graph-based examples, hierarchical clustering, and graph grammar learning from graph data.

Substructure Discovery

Subdue accepts as input directed or undirected graphs with labeled vertices (nodes) and edges (links). As an unsupervised discovery algorithm, Subdue searches for a substructure, or subgraph of the input graph, that best compresses the input graph. Once the search terminates and Subdue returns the list of best substructures, the graph can be compressed using the best substructure. The Subdue algorithm can be invoked again on this compressed graph, generating a hierarchical description of discovered substructures. To allow slight variations between instances of a discovered pattern, Subdue applies an inexact graph match between the substructure definition and potential instances. Subdue's run time is polynomial in the size of the input graph. Substructure discovery using Subdue has yielded expert-evaluated significant results in domains including predictive toxicology, network intrusion detection, earthquake analysis, web structure mining, and protein data analysis. Subdue can also be used to learn concepts that distinguish examples of different classes. New examples that contain the discovered substructures are classified as positive examples, otherwise they are classified as negative examples.

Application of Subdue to Security Data

As part of the U.S. Air Force program on Evidence Assessment, Grouping, Linking and Evaluation (EAGLE), a domain has been built to simulate the evidence available about terrorist groups and their plans prior to their execution. The simulator was developed based on feedback from intelligence analysts. The data we use represents the activities of terrorist organizations as they attempt to exploit vulnerable targets and consists of a number of concepts, including threat and non-threat actors, threat and non-threat groups, targets, exploitation modes (vulnerability modes are exploited by threat groups, productivity modes are exploited by threat and non-threat groups), capabilities, resources, communications, visits to targets, and transfer of resources between actors, groups and targets. As part of this project, we are testing Subdue's capabilities on EAGLE data. The results have indicated the success of this algorithm in discovering patterns that are valuable for security-related applications.

Recent Activities

- Processing incremental data. A challenge that arises in applying Subdue to security data is processing structural data that arrives in incremental blocks. Instead of preprocessing the accumulated graph after each increment, we provide summary statistics for each increment and process the new block independently. This actually results in overall runtime improvement for Subdue and allows near-real-time handling of streaming structural data.
- Learning from supervised graphs. We have explored a new representation for supervised learning that allows instances of all classes to be embedded in one connected graph. The examples themselves are of varied size and can overlap. We have shown with NASA data and with EAGLE data that distinguishing concepts can be learned with this representation and that the accuracy of the result and efficiency of the algorithm is superior to the case where examples are extracted from the single graph.
- Tighter integration with relational database management system (RDBMS) technologies. Since most EAGLE-related data is stored in RDBMSs, a tighter integration of Subdue with RDBMS technologies will allow improved scalability by reducing the computational and cognitive overhead of converting the database into the graph representation required by Subdue. The previous automated method requires a significant portion of the data to be extracted from the database and converted to a graph representation. We have been investigating a more incremental method whereby the database adapter requests data from the database only when prompted by Subdue during the pattern learning process. We are also investigating an even tighter integration in which the data will remain within the RDBMS and not be converted into graph form. This will involve an integration of Subdue with the RDBMS at the algorithm level. The approach developed in this task infers the entity-relationship model for the database using the table descriptions along with primary and

foreign key constraints and generates the instances of the graphs from the populated instances of the relations. As part of this task the following issues were addressed: i) representation of an individual relations as a graph (template), ii) representation of multiple relations as a graph (based on foreign key constraints), iii) alternative representations for ii), and iv) sequence in which to process the relations to generate graph instances. The choice of graph representation is important as the size of the input for mining may increase significantly (e.g., can double) based on the representation chosen. This will have a critical impact on the processing time and main memory requirements. Also, choice of a representation that captures the semantics of the transactions is important. Otherwise, it may be difficult to interpret the results of mining. Finally, the sequence in which relations are processed for generating the graph determines the amount of memory used for generating the graph and the time complexity of the algorithm. XRDB2Graph algorithm developed for this task is an efficient algorithm that converts any relational database into a graph form that can be used by Subdue. XRDB2Graph is loosely coupled with the database platform (using the JDBC bridge) in order to accommodate any relational database (MySQL, Oracle, MS SQLServer, etc.). We have used only SQL92 so that compatibility and portability is guaranteed across widely used databases. XRDB2Graph also minimizes the maximum memory requirement while transforming relational database to Graph domain; hence larger databases can be transformed into Graphs without need for large amounts of main memory. Even with SQL92, as there are some differences between create tables of different databases, the system is modularized. Only a small portion of the system needs to be customized for a particular database.

- Improved performance. We are continuing to improve the efficiency and effectiveness of the Subdue algorithm. First, we are developing faster approaches to graph isomorphism, which is one of the core functions used in Subdue, accounting for about 80% of the runtime. We are also developing ways to avoid unnecessary calls to the graph isomorphism test. Third, we are investigating ways to reduce the redundancy of patterns considered by Subdue.

We released a new version of Subdue (5.1.5) on December 20, 2005 that contained significant performance improvements, at times achieving order of magnitude speedups. The two main improvements were further refinements to the instance mapping technique that avoids calls to the graph isomorphism check when two matching graphs are extended in similar ways, and termination of the search on substructures with only one instance, as extending them further would not improve their value.

- We are continuing our research on graph-based anomaly detection by detecting changes in global graph metrics between normal graphs and those containing anomalies. We are focusing on structural metrics (e.g., average degree, clustering coefficient) and have generated results on several artificial domains. Results show that different metrics are good at detecting different types of anomalies, but no one metric captures them all. We are investigating combinations of metrics to provide a more robust detection method.
- We have completed a comparison of Subdue to a graph kernel approach for support vector machines. While the graph kernel methods consistently achieve better results in terms of classification accuracy and learning time, they do not provide a comprehensible learned pattern. These results were obtained with a random-walk kernel, which constructs features based on various random walks through the graphs being compared and then attempts linearly learning to discriminate the graphs based on a vector of these features. We found there are rare cases in which the random walk approach can be fooled (i.e., when the positive and negative examples differ by small changes in large common structures, e.g., cliques versus near cliques), and that in cases where numerous random walks are needed, computation time can exceed that of Subdue.

We ran old and new version of Subdue on EAGLE data. Before making the improvements described here, Subdue achieved 82% and processed 80,000 entities+links per hour. The most recent experiment on the same data yielded 84% and processed 80,000 entities+links per hour. As this demonstrates, we were able to improve our running time by approximately 50%.

References

- [1] Mining Graph Data, (D. J. Cook and L. B. Holder, editors), John Wiley and Sons, to appear in September 2006.

- [2] Data Mining of Complex Data, (S. Bandyopadhyay, U. Maulik, L. B. Holder, and D. J. Cook, editors), Springer, September 2005.
- [3] J. Coble, D. J. Cook, and L. B. Holder, Structure Discovery in Sequentially-Connected Data Streams, to appear in International Journal on Artificial Intelligence Tools, 2006.
- [4] L. Holder, D. Cook, J. Coble and M. Mukherjee, Graph-based Relational Learning with Application to Security, to appear in Fundamenta Informaticae Special Issue on Mining Graphs, Trees and Sequences, 2006.
- [5] J. Potts, D. J. Cook, and L. B. Holder, "Learning from Supervised Graphs", to appear in Applied Graph Theory (M. Last, A. Kandel, and H. Bunke, editors), 2006.
- [6] V. Jakkula, M. Youngblood, and D. Cook, Identification of Lifestyle Behavior Patterns with Prediction of the Happiness of an Inhabitant in a Smart Home, to appear in Proceedings of the AAAI Workshop on Computational Aesthetics, 2006.
- [7] J. Kukluk, L. Holder and D. Cook, Inference of node replacement recursive graph grammars, Proceedings of the SIAM Conference on Data Mining, April 2006.
- [8] C. D. Corley, D. J. Cook, L. B. Holder, and K. P. Singh, Graph-based data mining in epidemia and terrorism data, to appear in Proceedings of the Conference on Quantitative Methods and Statistical Applications in Defense and National Security, 2006.
- [9] N. Ketkar, L. B. Holder, and Diane J. Cook, Comparison of Graph-based and Logic-based Multi-Relational Data Mining, SIGKDD Explorations Special issue on Link Mining 7(2), 2005.
- [10] N. Ketkar, L. Holder and D. Cook, Qualitative Comparison of Graph-based and Logic-based Multi-Relational Data Mining: A Case Study, Proceedings of the ACM KDD Workshop on Multi-Relational Data Mining, August 2005.
- [11] N. Ketkar, L. Holder, D. Cook, R. Shah and J. Coble, Subdue: Compression-based Frequent Pattern Discovery in Graph Data, Proceedings of the ACM KDD Workshop on Open-Source Data Mining, August 2005.
- [12] J. Coble, D. J. Cook, R. Rathi, and L. B. Holder, Structure Discovery from Sequential Data, International Journal of Artificial Intelligence Techniques, 14(1-2), 2005.
- [13] D. Cook, L. Holder, J. Coble and J. Potts, Graph-based Mining of Complex Data, S. Bandyopadhyay, U. Maulik, L. Holder and D. Cook (eds.), Advanced Methods for Knowledge Discovery from Complex Data, Springer, 2005.
- [14] J. Kukluk, L. B. Holder, and D. J. Cook, Algorithm and Experiments in Testing Planar Graphs for Isomorphism, Journal of Graph Algorithms and Applications, 8(3), 2005.
- [15] L. B. Holder and D. J. Cook, Graph-based Data Mining, J. Wang (ed.), Encyclopedia of Data Warehousing and Mining, Idea Group Publishing, 2005.
- [16] J. Potts, L. B. Holder, D. J. Cook, and J. Coble, Learning Concepts from Intelligence Data Embedded in a Supervised Graph, Proceedings of the International Conference on Intelligence Analysis, 2005.
- [17] J. Coble and D. J. Cook, Structure Discovery in Sequentially Connected Data, Proceedings of the Florida Artificial Intelligence Research Symposium, 2005.
- [18] R. Rathi and D. J. Cook, A Serial Partitioning Approach to Scaling Graph-Based Knowledge Discovery, Proceedings of the Florida Artificial Intelligence Research Symposium, 2005.
- [19] J. Potts, D. J. Cook, and L. B. Holder, Learning from Examples in a Single Graph, Proceedings of the Florida Artificial Intelligence Research Symposium, 2005.

4 Mobile Database

Faculty Involved : Ali Hurson

Student Members : G. Jahn, J. Ghaznavi, J. Ploskonka, J. Sustersic, B. Yang, X. Gao, M. Ontang, A. Tangpong

WebPage: <http://www.cse.psu.edu/gis/>

Research Topics Investigated

The proposed PSI framework is a two-tier network topology. The first-tier (front-end) is a collection of smart sensors performing security missions in pervasive fashion. This level deals with dynamic data and control signals generated by pervasively implanted surveillance smart devices heterogeneous sensors (fixed or mobile), actuators, RFID tags and detectors, monitors, and the computing and networking (including wireless) infrastructure. The second tier (back-end) is a collection of networked possibly heterogeneous and autonomous information resources that will be used for information extraction, data mining, and profiling operations to aid seamless decision making process. Our research interest in this collaborative project mainly lies to the operations at this tier with software agents communicating between these two tiers. Generated software agents at the front-end tier will be roaming through the backbone infra structure to:

- Locate the relevant information intelligently, efficiently, and transparently.
- Extract, process, and integrate relevant information efficiently and securely.
- Interpret and communicate the processed information intelligently and seamlessly.

Summary of Current Research Work and Results:

- **On Demand Based Services:** Database systems play important roles in information storing and sharing. They are widely used in business, military, and research fields. However, since they are developed, evolved, and applied in isolation over a relatively long period of time, the inevitable heterogeneity becomes an indispensable characteristic of any information-sharing environment. Moreover, for many practical and performance purposes, the creation of databases is usually close to the application domains. Consequently, information resources are distributed in nature. The heterogeneity and distribution of information worsens the problem of global information sharing. To overcome the obstacles brought by the local database heterogeneity, two possible solutions have been studied in the literature: (i) Redesign the existing databases to form a homogeneous information-sharing system, or (ii) Develop a global system on top of the heterogeneous local databases to provide a uniform information access method (multidatabase system). High cost associated with the first choice prohibits it from becoming a feasible solution in many practical cases. On the other hand, the latter approach offers a more practical solution to share information globally. Typically, a multidatabase system consists of a global component and a collection of local components. The global component hides the underlying heterogeneity and provides users a uniform global information access method. In order to preserve local data autonomy while providing full database functionality, the global component usually maintains a global schema that contains local schema information. Problems arise as the size of the multidatabase and the global schema grows. Maintaining and manipulating multiple copies of large global schema in a distributed environment is problematic. Within the scope of the multidatabases, we proposed an elegant solution (The Summary Schemas Model SSM) for large-scale organization that addresses the problems associated with global schema approaches. The SSM is designed to support the identification of semantically similar/dissimilar data entities. The model maintains a hierarchical meta-data based on the semantics of the access terms exported from underlying local databases. This meta-data is used to intelligently resolve name differences using word relationships defined in a standard thesaurus such as the Roget's Thesaurus. Users can submit imprecise queries at any site without knowing the location of requested information and/or the method to access the information. Based on the data semantics, the SSM maps imprecise query terms with precise access terms found at local databases.

As mobile technology advances, more and more database users demand anywhere, any time data access from their mobile devices. In this mobile data access system (MDAS) environment, the

SSM faces new challenges: mobile devices usually have limited resource, users are moving, and the wireless media is low in bandwidth and unreliable in nature. Traditional client-server based SSM implementations cannot satisfy these challenges because the success of such systems relies on reliable communication.

Mobile agent technology has been developed as a software design paradigm that provides special support for mobile computing. One of the prominent advantages of mobile agents is their execution autonomy. After submission, the mobile agent can roam the network and collaborate with other agents without the owners intervention. They make decisions according to the preset execution plan and information obtained at the current execution environment. Therefore, the user only needs to maintain a network connection during the agent submission and retrieval. This significantly relaxes the constraint on network connectivity as compared to client-server implementations. Many research projects have demonstrated the advantages of mobile agents in designing distributed information access systems, electronic trading systems, virtual enterprises, and so on.

After witnessing the success of many mobile agent applications, we proposed and implemented a new infrastructure MAMDAS Mobile Agents within the framework of Mobile Data Access System. MAMDAS combines the merits of SSM and mobile agents in building a distributed large-scale information access systems. It aims to achieve higher performance, while providing special support for mobile users. Our experimental results have shown that MAMDAS is 6 times faster than the client-server based SSM prototype, because of the reduced network traffic. Moreover, MAMDAS demonstrated great scalability, portability, and robustness. Within the scope of MAMDAS our research in the past concentrated on security, power management, and construction of the application specific thesauri.

Achievements: Initially, our infrastructure assumed traditional flat files and databases. The scope of MAMDAS was extended to accommodate multimedia data bases (image data bases). Conceptual frame work was developed to represent and manipulate image data sources by our intelligent search engine. The scope of the local nodes in MAMDAS was extended to allow a community of wireless devices with ad hoc connectivity to share information among themselves and the background infrastructure.

- **Quality of Service Based Coherence Protocol for Internet:** Caching has long been employed in computer system architectures to improve performance in terms of reduced memory access times and latencies, and hence to improve system throughput at the expense of additional complexity in memory organization and managing multiple copies of shared data. In traditional large-scale computer architectures, the cache coherency schemes developed permit very large, cache-coherent non-uniform memory access (CC-NUMA) to shared memory spaces. However, these approaches fail when applied to the vastness of the Internet and the growing complexities introduced by mobility. While caching has been successfully employed in specific, limited applications in modern Internet implementations, there has been no general-purpose approach to cache coherency on the Internet. A quality-of-service (QoS) approach is ideally suited for such a general-purpose cache coherence protocol, providing strong consistency for those data items that require it (online transaction processing, e-commerce, etc.) while permitting weaker consistency for less critical data. A statistical analysis of the read/write behavior of typical Internet data will be used to suggest a low overhead, inexpensive QoS solution to cache coherency issues on the Internet. An experimental framework for QoS coherence implementation is being studied to demonstrate its potential as a viable solution to cache coherence on both wired and wireless Internet applications.

Achievements: The growing application of caching in Internet applications have heretofore relied largely on qualitative observation and empirical data on the update behaviour of Internet data in their design. While it is empirically known that the update behaviour of such data is distinctly bimodal, much less is known about the details of these behaviours and the processes that drives them. A detailed study of the composition of modern web sites that includes in-depth analyses of the changes made to the data of which those sites are composed offers a large potential benefit not only to Internet caching protocol developers but also to a broad cross-section of information technology professionals and researchers. A variety of popular Internet web sites were selected and monitored for changes in both composition and content over a period of several weeks. Data collected in this study is then analyzed using traditional stochastic methods, including maximum likelihood

distribution fits for data with 2 or more observed updates. The results of this investigation were summarized and reported in a published paper.

- **Broadcast based services:** Many applications are directed towards public information that are characterized by i) the massive number of users, ii) the similarity and simplicity in the requests solicited by the users, and iii) the fact that data is modified by a few. The reduced bandwidth attributed to the wireless environment places limitations on the rate and amount of communication. Broadcasting is a potential solution to this limitation. In broadcasting, information is generated and broadcast to all users on the air channels. Mobile users are capable of searching the air channels and pulling their required information. The main advantage of broadcasting is due to the fact that it scales up as the number of users increases, eliminating the need to multiplex the bandwidth among users accessing the air channel. In addition, broadcast channel can be considered as an additional storage available over the air for the mobile clients. Finally, it is shown that pulling information from the air channel consumes less power than pushing information to the air channel. Broadcasting is an attractive solution, because of the limited storage, processing capability, and power sources of the mobile unit. Within the scope of broadcasting one needs to address three issues: (i) Broadcast contents, Network latency, and (ii) Power consumption of the mobile unit. The employment of broadcasting in the mobile-computing environment motivates the need to study the proper organization of objects along the air channel(s). Due to the natural differences between the serial air channel and the random-access disk, one has to look at different and efficient methodologies to organize and cluster objects on the air channels in order to reduce the response time. In addition, the network latency (response time) is the major source of power consumption at the mobile unit. The reduction in response time translates into the reduced amount of time a mobile unit spends accessing the channel(s) and thus, it has its impact on conserving energy at the mobile unit. The necessity of minimizing power consumption and network latency lies in the limitation of current technology the expected increase of the capacity of batteries is at much lower rate than the increase of the chip density. The hardware of the mobile units has been designed to overcome this limitation by operating in various operational modes such as active, doze, sleep, nap, etc. to conserve energy. A mobile unit can be in active mode (maximum power consumption) while it is searching or accessing data objects; otherwise, it can be in doze mode (reduced power consumption) when the unit is not performing any computation. Along with the architectural and hardware enhancements, efficient power management and energy aware algorithms can be devised to (i) organize and cluster related data entities on broadcast channel(s) and (ii) schedule data retrieval from broadcast channel(s).

Achievements: Within the scope of broadcasting our research concentrated on: (i) Effective data organization on the broadcast channel, and (ii) Efficient data retrieval from the broadcast channel with intension to reduce the access latency and power consumption. Several issues including: (i) Application of indexing on single and parallel channels, Object organization on single and parallel channels, and (ii) Scheduling of object retrieval from parallel broadcast channels at the presence of conflicts.

On-going Activity and Planned Work for year 2006-2007:

Security in MAMDAS: We herein propose to implement a secure framework for the execution of Aglets; presently, to the best of our knowledge such framework is non-existent. In order to achieve our goal of providing a secure framework for Aglets, we will address the three main facets of Aglet security and provide some mechanisms to enforce security restrictions aimed at.

MAMDAS and Pervasive Computing: The use of MAMDAS can be extended to the pervasive computing environment, where computers and databases are pushed to the background and services are provided to users without being specifically requested. Smart classrooms that can automatically load lecture slides for professors according to the course schedule and syllabus exemplifies the idea of pervasive computing. We envision that the SSM can be used as the backbone knowledge base and autonomous mobile agents can act as user representatives who actually perform tasks on users behalf. Several difficulties must be addressed in this environment: (i) seamlessly integrate heterogeneous networks, (ii) implant human intelligence in agents, and (iii) introduce user incentives.

MAMDAS and Sensor network: The fast growth of sensor networks has attracted lots of research attention. Several prominent challenges in sensor networks include i) sensors have extremely scarce resources and short lifetime, ii) sensors have almost no physical protection, and iii) sensor network topology changes frequently. We believe the execution autonomy and decision making capability of mobile intelligent agents can alleviate the aforementioned problems.

QoS-based relaxed consistency model: The Quality of Service (QoS)-based relaxed consistency model for Internet caching has been proposed to permit a user-defined level of coherence on a per-request basis. However, the complex nature of the Internet and the multitude of underlying factors have limited the acceptance of the proposed model despite extensive simulation and analytical results. Consequently, a trace-driven simulation and analysis is necessary to better support the practicality of the proposed model. To this end, an extensive effort to obtain significant application-level data from the Penn State University's fiber-channel border router has been proposed to permit the most realistic simulations and analyses possible short of actual implementation. This proposed trace data collection project would capture IP and TCP headers for all network packets traversing the optical fiber links between the Penn State domain and all other Internet domains. Additionally, specific HTTP headers will be captured for those TCP/IP packets that contain HTTP headers. As a large Academic community geographically isolated from other large population centers, Penn State offers unique opportunity in which the actual network traffic generated by a large community may be observed at single network point.

Image Retrieval in ad hoc network: Mobile ad hoc networks have gained more and more research attentions by provisions of wireless communications without location limitations and pre-built fixed infrastructure. Because of the absence of any static support structure, ad hoc networks are prone to several limitations such as bandwidth, connectivity, and power. Multimedia retrieval is a challenging task in wireless ad hoc networks because of the multiple limitations. We will investigate that data content distribution can be employed to facilitate content-based multimedia retrieval in ad hoc networks. Motivated by this data organization methodology, we will propose a logic-based content summary framework that is able to represent semantic contents of multimedia data using concise logic terms. Furthermore, we will build a virtual infrastructure to cluster mobile nodes according to the semantic contents. The proposed framework will be simulated and analyzed based on various performance metrics.

Location dependent data processing: The scope of our infrastructure will be extended to accommodate location dependent data processing, location aware data processing, and continuous query processing in a mobile environment with an eye toward to the improvement of performance metrics such as access time, power consumption, etc.

Education and Outreach

Based on the aforementioned activities:

- One full day tutorial entitled, Heterogeneous and mobile databases was presented at the CSICC 2006 conference.
- Research results presented in a panel entitled, Hot Topics in Computing and Information Technology in the Next Decade at the CSICC 2006 conference.
- Research results presented in various ACM/IEEE/International conferences.
- Guest Co-Editor Journal of Pervasive and Mobile Computing, Special issue on security.

References

- [1] Hurson, A. R., Muoz-Avila, A.M., Orchowski, N., Shirazi, B., and Jiao, Y., Power-Aware Data Retrieval Protocols for Indexed Broadcast Parallel Channels, *Journal of Pervasive and Mobile Computing*, vol. 2, No. 1, 2006, pp. 85-107.
- [2] Yang, B. and Hurson, A.R., Similarity-Based Clustering Strategy for Mobile Ad Hoc Multimedia Databases, *Journal of Mobile Information Systems*, Vol. 1, No. 4, 2005, pp. 253-273.
- [3] Yang, B. and Hurson, A.R., Hierarchical Semantic-Based Index for Ad Hoc Image Retrieval, *Journal of Mobile Multimedia*, Special Issue on Mobile Multimedia Computing and Communications, Vol. 1, No. 3, 2005, pp. 235-254.

- [4] Hurson, A.R., Jiao, Y., and Shirazi, B., Broadcasting a Means to Disseminate Public Data in a Wireless Environment Issues and Solutions, *Advances in Computers*, Vol. 67, 2006, pp.1-85.
- [5] Jiao, Y., Hurson, A.R., and Potok, T., *Mobile Agent-Based Information Systems and Security*, *Encyclopedia of Information Science and Technology*, 2nd edition, 2006.
- [6] Jiao, Y., Hurson, A.R., Potok, T.E., and Beckerman, B.G., *Integrating Mobile-Based Systems with Health Care Databases, Web Mobile-Based Applications for Healthcare Management*, 2006.
- [7] Yang, B. and Hurson, A.R., Location-Aware Caching for Spatial Queries in Dynamic Environments, *IEEE Wireless Communications and Networking Conference, WCNC 2006*.
- [8] Ayyagari, P., Mitra, P., and Hurson A.R., Efficient Object Retrieval from Parallel Air Channels in the Presence of Replicated Objects, *International Conference on Mobile Data Management, MDM, 2006*.
- [9] Gao, X., Sustersic, J., and Hurson A.R., Window Query Processing with Adaptive Proxy Cache, *International Conference on Mobile Data Management, MDM, 2006*.
- [10] Yang, B. and Hurson, A.R., On the Content Predictability of Cooperative Image Caching in Ad Hoc Networks, *International Conference on Mobile Data Management, MDM, 2006*.
- [11] Yang, B. and Hurson, A.R., Multimedia Semantics Integration Using Linguistic Model, *International Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2006, pp 679-688.
- [12] Yang, B. and Hurson, A.R., Content-Initiated Organization of Mobile Image Repositories, *IEEE/IFIP WONS2006 (The Third Annual Conference on Wireless On demand Network Systems and Services)*, 2006.
- [13] Sustersic, J., Hurson A.R., and Nickel, R. M. An Analysis of Internet Data Update Behaviors, *IEEE AINA2006 International Conference*, 2006, pp. 773-778.
- [14] Yan, P., Jiao, Y., Hurson, A.R., and Potok, T.E., Semantic-Based Information Retrieval of Biomedical Data, *ACM Symposium on Applied Computing, SAC 2006*.
- [15] Yang, B. and Hurson, A.R., Similarity Search in Ad Hoc Networks Using Semantic-Based Caching, *IEEE Conference on Local Computer Networks (LCN)*, 2005, pp. 115-122.
- [16] Yang, B. and Hurson, A.R., Content-Navigated Multimedia Search in Ad Hoc Networks, *ACM/IEEE MSWiM 2005, Montreal, 2005*, pp. 103-110.
- [17] Jiao, Y., Hurson, A.R., and Shirazi, B., Online Adaptive Application-Driven WLAN Power Management, *IEEE Global Telecommunication Conference*, 2005, pp. 2663-2668.
- [18] Yang, B. and Hurson, A.R., A Content-Aware Multimedia Accessing Model in Ad Hoc Networks. *International Conference on Parallel and Distributed Systems, ICPADS 2005*, pp. 613-619.
- [19] Yang, B., Hurson, A.R., Jiao, Y., and Potok, T.E., Multimedia Correlation Analysis in Unstructured Peer-to-Peer Networks ,under review.
- [20] Ploskonka, J.A., and Hurson, A.R., Self-Monitoring Security in Ad Hoc Routing, under review.
- [21] Jiao, Y., and Hurson, A.R., Energy-Efficient Wireless Information Retrieval, under review.
- [22] Yang, B., and Hurson, A.R., Semantic-Aware and QoS-Aware Image Caching in Wireless Ad Hoc Networks, under review.
- [23] Yang, B., and Hurson, A.R., *Mobile Multimedia: Representation, Indexing, and Retrieval*, Book Chapter.
- [24] Ontang, M. and Hurson, A.R., Agent-based Transaction Management for Mobile Multidatabase, Book Chapter.

5 InfoFilter: A Content-Based Information Filtering System

Faculty Involved : Sharma Chakravarthy

Student Members: R. Adaikkalavan, B. Kendai, A. Telang, C. H. H. Subramanian

WebPage: <http://itlab.uta.edu/>

Introduction

Applying appropriate searching mechanisms to retrieve only relevant information becomes critical to avoid information overload (or retrieving very large number or portions of documents). *Information Retrieval* is the process of extracting relevant or useful portions of documents from a relatively static collection of documents based on a stream of incoming user patterns (or queries). In information retrieval, expressiveness of pattern (or query) specification by a user and its detection (or matching) play a significant role. In other words, in order to extract useful or meaningful information, users need to have the ability to specify complex patterns. A critical limitation of current search engines is that they can support only simple patterns or a simplistic combination of patterns using Boolean operators. Thus, current query languages are quite restrictive in their expressive power and need to be extended and generalized to address the specification of meaningful complex user patterns. On the other hand, ability to specify complex patterns will not be meaningful or effective without a correct and efficient mechanism for their detection in real-time. In this report we summarize, InfoSearch, a novel approach for expressive pattern matching over stored data. It allows users to specify complex patterns and matches these patterns over stored data. Complex patterns such as combinations of sequential, structural patterns, wild cards, word frequencies, proximity, Boolean operators and synonyms are formulated using the expressive pattern specification language, PSL.

Pattern Specification Language

PSL, an expressive pattern specification language, allows the specification of complex patterns. Occurrence of a *Pattern P* is a Boolean function whose domain is an offset interval and range is TRUE or FALSE, depending upon whether the specified pattern occurs in that interval. According to the semantics of PSL, a pattern is classified into a simple and composite pattern.

Simple patterns form the basic building block of the PSL. A *simple pattern* is either a word such as *filtering*, a phrase such as *information filtering systems* or a simple regular expression (regular expression on a single word) such as *info**. A simple pattern is denoted by $P[Os, Oe]$, where $Os = Oe$ (i.e., the starting and ending offset of the pattern is the same). A simple pattern occurrence is an atomic occurrence of a simple pattern. It occurs over an interval $[Os, Oe]$ and it is detected at the end of the interval (i.e., Oe). PSL supports two types of simple patterns, system-defined (e.g., BeginDoc, EndDoc, BeginPara, EndPara) and user-defined (single word, phrase and regular expression). A *composite pattern* is an expression constructed using simple patterns, previously constructed composite patterns, PSL operators and options. PSL provides a comprehensive set of operators, OR, non-occurrence (NOT/N), sequential (FOLLOWED BY/N), structural (WITHIN/N), frequency (FREQUENCY/N), proximity (NEAR/N) and the option synonyms (SYN) that allow users to compose *composite patterns*.

InfoSearch Architecture

InfoSearch [2] allows the user to specify complex queries and returns information about every occurrence of the pattern specified in the query. The scope of the search is a pre-indexed document collection (e.g., root of a Web site), and the information returned is the document (e.g., Web page) in which the pattern occurs, and the position of every occurrence of the pattern within each document. InfoSearch accepts complex queries from the user, searches an index built on a collection of documents, and returns a list of documents that contain the specified pattern. It also returns the starting and ending position of each pattern occurrence within a document. The system can be broadly divided into two components: First, an expressive query language through which the user can specify patterns involving term frequency, sequences, proximity and containment is required. Second, a pattern detection engine capable of getting the required information from the index, and processing this information to generate results in response

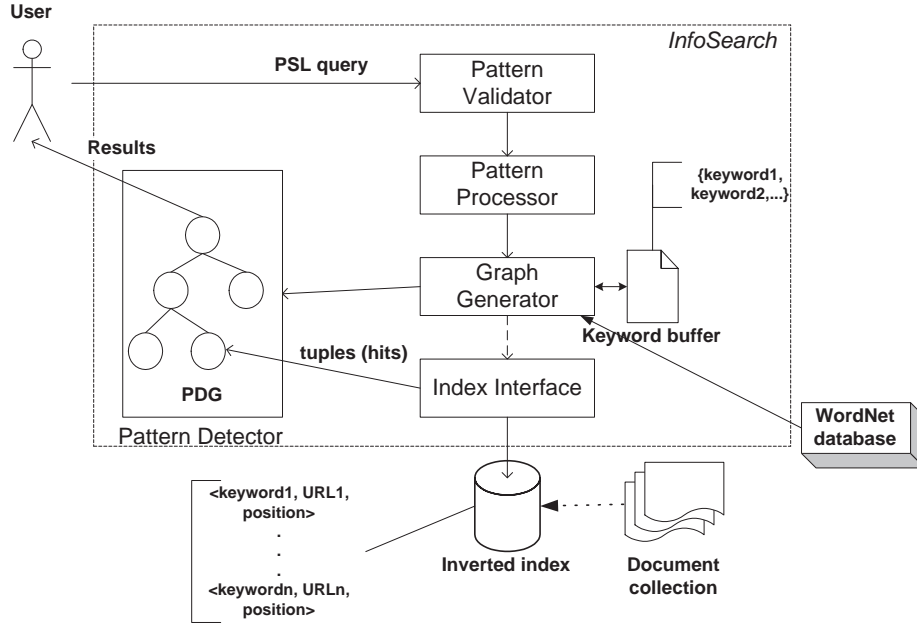


Figure 1: System Architecture

to a user query. This is a two step process in which the first step involves forming a Pattern Detection Graph (PDG) corresponding to the query. The next step involves a lookup for the index to detect the occurrences of the pattern (including complex patterns) that must be detected over the PDG to generate the query result.

The InfoSearch modules are specific to index-based retrieval and search, and the operators use algorithms that have been modified from InfoFilter [1] to detect patterns over data retrieved from an index. Figure 1 illustrates the architecture of our system. Below, we discuss various modules of our system briefly.

Pattern Validator accepts the input patterns in infix notation from users according to the BNF [1] of PSL. Once the patterns are validated they are decomposed into tokens. The tokens in a pattern can be keywords, phrases, system defined patterns, operators and other delimiters allowed by the language. These tokens are sent to the *pattern processor*, which accepts these patterns or tokens and converts them to postfix notation. Infix notation is easier to specify and the default operator precedence can be altered by using parenthesis. To evaluate the pattern, with emphasis on the operator precedence and minimization of the use of parentheses, the infix notation is converted to postfix notation. Patterns in postfix notation are easier to evaluate as the operands precede the operators. It then sends the patterns in postfix notation to the graph generator.

Graph Generator generates pattern detection graphs from user-specified patterns by invoking APIs from the pattern detector library. The graph generator constructs the PDGs corresponding to the patterns in the pattern detector, and interacts with the WordNet Database tool to extract the synonyms of single words if specified. *WordNet* [3] is used to determine the synonyms of the extracted keywords, if the synonym option is specified. The graph generator sends the keywords to WordNet to extract their synonyms. Once the synonyms are extracted, the graph generator stores them. While generating the graph, the graph generator stores the keywords specified in the query in a keyword buffer. Once the PDG is generated, the graph generator queries the index for each of the keywords it has stored in its buffer. This is done through the index interface. The index interface module is responsible for retrieving the *hits* for each keyword from the index. These hits are wrapped into a set of *tuples* and passed on to the leaf node that represents the keyword. Thus, a monotonically decreasing set of data propagates up the PDG, and the output of the root node is the answer to the query which is returned to the user.

Pattern Detection Graph (PDG): A user pattern is internally represented as a PDG. It maintains

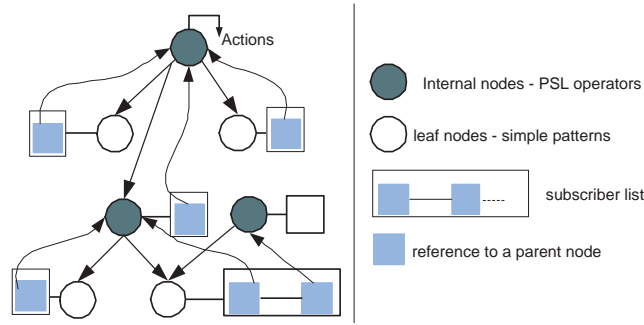


Figure 2: Pattern Detection Graph

the flow of the pattern occurrences (or maintaining partial histories). For each pattern or sub-pattern, a corresponding PDG is constructed. Many PDGs are combined to form a complex pattern as shown in Figure 2. The PDG is constructed recursively in a bottom-up fashion. The leaf nodes are created first, followed by the operators defined on these leaf nodes. When a parent node is created, it subscribes to its children. Essentially, this means that the reference of the parent is given to the children, so that data can be passed from a child to its parent. Thus, every node has a subscriber list that has a reference to each of its parents. To optimize space requirements, the graph generator shares PDG nodes wherever possible. This is achieved by keeping a single, common PDG or sub-PDG for a common expression or sub-expression. This avoids creation of a new PDG, if a PDG has already been created for a previous expression or sub-expression.

Pattern Detector: The detection of a simple pattern is straightforward since the start and end offset of a pattern is the same. However, detection of a composite pattern is complex since it involves the detection of simple or composite sub-patterns. One approach to detection is by using *Histories* (Global and Composite). The other approach for detection is usage of *Detection Modes* (Restricted and Unrestricted). These approaches are explained in detail with the help of examples in [1, 2]. The pattern detector is a library that provides APIs to construct the PDGs for the user patterns. Once simple patterns are detected in the index lookup (for stored data), pattern detector passes the notifications to the corresponding leaf nodes in the PDG leading to the computation of the associated complex pattern occurrences. The pattern occurrences flow up the tree until they reach the topmost node. That node triggers the actions associated with that PDG. Initially, when the PDG is constructed, the topmost internal node of the PDG is mapped to the action required once the entire pattern is detected. The action is typically alerting the notifier to send messages to the user providing him/her with the stream information where the pattern has been detected. The document information is present in the pattern occurrences. When a leaf node in the pattern detector receives a *set of tuples* from the Index Interface, it passes a reference to this set to its parent nodes. When internal nodes merge their input set of tuples to create a merged output set, they pass a reference to this set to their parents. The root node has a rule associated with it, which essentially directs the output of the pattern detection to the user through e-mail or other forms of notification. Even though the set of operators used for both stream data [1] and stored data [2] are same, the computation at each node (or operator) are different.

Index Interface: The detection engine of InfoSearch is designed to be generic and capable of working with any kind of index. The index interface cleanly separates the InfoSearch system and the index which is being used. It is the only module which is specific to the index being used. In other words, the index interface is the only module that needs to be changed when InfoSearch needs to be integrated with a different kind of index. The index interface accepts simple patterns from the graph generator and queries the inverted index. It is responsible for wrapping the result returned by the index in a set of $\langle \text{docID}, \text{start offset}, \text{end offset} \rangle$ tuples. The InfoSearch operators need their input in the form of tuples with offsets, and hence wrapping the output of the index into tuple sets is an important function of the index interface. The set of tuples generated as the index lookup for a keyword is then passed to the leaf node in the PDG that corresponds to that keyword.

Experiments

The primary reason for developing operators to detect complex patterns over indexed data was to support efficient searching of stored documents for complex patterns. For streaming data, it does not make sense to store, index, and process them disregarding the quality of service (QoS) requirement of near real-time pattern detection. On the other hand, it does not make sense to *stream already stored documents* to use InfoFilter [1] for detecting patterns. Hence, both approaches are needed for different scenarios. Our effort separates these two approaches by a few specific modules and keeps the bulk of the system and the architecture common. This is an important and desirable characteristic of our system. It is true that the index-based approach will perform better for larger volume of documents where as stream-based approach will be better for smaller volumes that are not static. The crossover point, in terms of number of words in the document collection, where the indexed approach perform better than the streaming approach was the first parameter of interest. A set of 20 documents of around 1.5 KB each were selected from the Reuters data set. The total number of words in this collection was around 2600. The documents were artificially converted into a stream, fed to InfoFilter and patterns involving all the operators were detected over this stream. The time taken to process the stream was noted. Subsequently, the same documents were indexed, and InfoSearch was used to detect the same patterns as in the previous case. In this case, the time taken for the result set to reach the root node was also noted. It was observed that using an index to detect complex patterns outperforms the streaming approach even for relatively small document collections. Of course, the time taken to index the documents is not considered in the above comparison, but that can be considered as a pre-processing step, and does not come into consideration once the documents are indexed and the index is brought online. The results of the experiments can be found with detailed analysis in [2].

Conclusions and Future Work

In this report, we discussed InfoSearch, a system designed to search complex patterns over a document collection. InfoSearch is an attempt to facilitate searching of complex patterns involving proximity, frequency, containment and sequences over a given document collection. One alternative for detecting complex patterns over stored documents is to read the entire data source, tokenize it, and feed it into the Pattern Detection Graph to detect the pattern. This has been explored earlier in [1]. Although this is useful in certain domains, such as news feeds and other applications where data is generated inherently as a stream, there are other domains where artificial streaming is inherently inefficient, especially when the source does not change frequently or accuracy can be traded off for response time. For such relatively static documents, it is much more efficient to use an index over the document collection (as has been demonstrated by IR and Web Search Engines) to search for patterns. There are two areas where a search system can always be improved upon: quality of the results, and performance. Presently, InfoSearch gives the result of the query in sorted order of document ID. A ranking strategy can be developed, which delivers the results in descending order of how “closely” the results match the specified query. To improve performance, strategies to selectively cache the result sets can be developed. The operator algorithms need to be more rigorously analyzed to assess their complexity, and improved, if possible, for efficiency.

References

- [1] L. Elkhalfa, “InfoFilter: Complex Pattern Specifications and Detection over Text Streams”, Master’s thesis, Department of Computer Science and Engineering, The University of Texas at Arlington, 2004. Online at <http://itlab.uta.edu/ITLABWEB/Students/sharma>.
- [2] N. Deshpande, “Infosearch: A system for searching and retrieving documents using complex queries”, Master’s thesis, The University of Texas at Arlington, 2006. [Online]. Available: <http://itlab.uta.edu/ITLABWEB/Students/sharma/theses/Des05MS.pdf>.
- [3] C. Fellbaum, “Wordnet: An electronic lexical database”, in MIT Press, 1998.
- [4] R. Baeza-Yates and B. Ribeiro-Neto, “Modern Information Retrieval”, New York: ACM Press/Addison-Wesley, 1999.
- [5] S. Chakravarthy, V. Krishnaprasad, E. Anwar and S. K. Kim, “Composite Events for Active Databases: Semantics, Contexts and Detection”, 1994, pp 606-617.

6 Mobile AdHoc Networks

Faculty Involved: Mukhesh Singhal, Raphael Finkel, Sajal Das

Student Members: V. Giruka, H. Li, R. Bai, and S. Chakrabarti

Major Research and Education Activities of the Project

The main objective of the project is to design and evaluate protocols for secure and efficient routing among a group of users in pervasive and mobile computing systems. The main objective is to insure secure information exchange among a group of users communicating over a wireless network. Research activities focused on investigating several key problems in insuring secure and efficient information exchange among a group of users communicating in ad hoc networks and mobile computing environments. Education activities included involving graduate students in the research project and developing curricula that incorporated the most recent research results. Specifically the project funded four Ph.D. students as Research Assistants. These students actively participated in the project and developed and evaluated the techniques.

Major Findings Resulting from these Activities

- We developed a new routing model, Way Point Routing (WPR), which maintains active routes hierarchically for Mobile Ad Hoc Networks (MANETs). WPR divides an active route into segments by selecting a number of nodes on the route as waypoint nodes. An inter-segment routing protocol runs globally, while an intra-segment routing protocol runs locally. Thus route maintenance can be localized to one or a few neighboring segments. One distinct advantage of WPR is that when a route is broken because of node mobility or node failure, instead of discarding the whole route and discovering a new route from the source to the destination, only the two waypoint nodes of the broken segment need to find a new segment. This will have a clear performance advantage in terms of routing overhead, end-to-end delay, etc. We developed an instantiation of WPR termed as DSR Over AODV (DOA). In DOA, DSR is used for inter-segment routing, and AODV is used for intra-segment routing. This is the first work to combine DSR and AODV, two well-known on-demand routing protocols, in a hierarchical manner. We also developed two novel techniques for route maintenance in DOA: a multi-target route discovery and an efficient loop detection method. We conducted extensive simulations to evaluate the performance of DOA and compare DOA with AODV and DSR. Simulation results show that DOA scales well for networks with more than 1000 nodes, routing overhead is significantly reduced while other metrics are better or comparable to AODV and DSR.
- In on-demand routing protocols for mobile ad hoc networks, usually there are a considerable number of RREQ packets in transit in the network, and a node cooperatively relays RREQs for other nodes at a very high rate. These RREQs provide opportunities for intermediate nodes to discover routes for their own, i.e., a source node may discover a route by piggybacking its route request on the RREQs it relays, instead of flooding the network with a new RREQ message. This has potential to improve the performance by reducing the overhead, congestion, and power consumption at nodes. We developed a new route discovery scheme called multiple-target route discovery (MTRD). A RREQ message in MTRD may contain multiple targets and discover these targets simultaneously. When a source needs to perform a route discovery, instead of immediately broadcasting a new RREQ message, it first tries to piggyback its request on the existing RREQ packets that it relays. This has potential to improve the routing performance by reducing the control overhead, congestion, and power consumption at nodes. We conducted simulations to evaluate the performance of MTRD. The results confirmed the effectiveness of MTRD by showing that MTRD significantly reduces the control overhead, improves the packet delivery ratio, and reduces the end-to-end delay when the network load is medium to heavy.
- Route reply (RREP) messages are important in on-demand routing protocols for ad hoc networks. The loss of RREPs causes serious impairment to the routing performance because the cost of a RREP is very high. Typically a RREP is obtained after coding the entire or a part of the network with route request (RREQ) messages. This process is expensive and time-consuming - tens, may be hundreds of transmissions are needed for a route discovery. If a RREP message is lost, a large

amount of route discovery effort will be wasted. Furthermore, the source node may have to initiate another round of route discovery to establish a route to the destination. We developed the idea of salvaging route reply (SRR), which attempts to salvage an undeliverable RREP in two possible ways. First, the node (salvor) looks up its own route cache for an alternate path to the source. Second, the node runs a one-hop SRR route discovery to find a path to the source. The SRR route discovery succeeds most of the time because neighboring nodes recently propagated the RREQ originated from the source and learned a reverse path to the source. We conducted extensive simulation study to evaluate the performance of SRR and compared it with AODV routing protocol. The results show that SRR significantly improves the performance of AODV in all metrics, namely, packet delivery ratio, control overhead and end-to-end delay.

- In ad-hoc networks, where nodes use realistic radios, links in the network may be unidirectional. Unidirectional links may exist in the network due to radio-interference or due to differences in transmission power of nodes. The presence of unidirectional links may violate assumptions made by geographic routing protocols, which may lead to persistent routing failures. Further, the presence of unidirectional links may also hinder the convergence of location service protocols. We developed a geographic routing protocol (GRPu) for ad-hoc networks with unidirectional links. GRPu is a unified routing and (implicit) location service protocol that inherits the best of three well-known techniques for routing in ad-hoc networks, viz., reactive route discovery, greedy forwarding, and geographic source-routing. In GRPu a source node establishes an on-demand loop-free geographic path to the destination, by carefully handling unidirectional links. Unlike node ID-based paths used by DSR, geographic paths decouple node ID's from the path. Thus any node along the geographic path can forward packets to the destination. Further, to utilize geographic paths in sparser networks, nodes use a path-healing mechanism. Path-healing helps geographic paths adapt according to the node mobility, and greatly mitigates path breaks. To evaluate the performance of GRPu protocol, we conducted extensive simulations using GloMoSim 2.03. Simulation results show that the GRPu protocol achieves a high percentage packet delivery, low control overhead, and low average hop count for a wide range of network scenarios.
- We developed a self-healing on-demand geographic path-based routing protocol (OGPR) for ad-hoc networks. Unlike other position-based routing protocols, which typically assume a location service protocol, OGPR uses an implicit location service to find the position of the destination. OGPR protocol establishes geographic paths that avoid dead-ends due to greedy forwarding, in static networks. To make use of geographic paths even in sparser or mobile networks, OGPR uses a path-healing mechanism that adapts geographic paths according to the network topology. If the destination is unreachable, an intermediate node along the geographic path gives a feedback to the source node in the form of a path error message. Such feedbacks are important to avoid congestion/network-wide flooding in an attempt to establish path when the network is partitioned. Further, we presented extensions to OGPR to cope with unidirectional links. To evaluate the performance of OGPR protocol and to compare its performance with AODV, DSR, and GPSR+GLS protocols, we conducted extensive simulations using GloMoSim. Simulation results show that OGPR achieves higher packet delivery rate and lower control overhead than AODV, DSR, and GPSR+GLS protocols, under a wide range of network scenarios.
- The nature of ad hoc networks makes them vulnerable to attacks, especially in the routing protocol. How to protect an ad hoc routing protocol is an important research topic. We developed a secure routing protocol for ad hoc networks with a shared group key as the sole assumption. The key security measures in this protocol are distributed authentication and Message Authentication Code. We developed a Distributed Authentication Model, with which different nodes can authenticate each other. Integrity is ensured by Message Authentication Code, which is calculated by using the shared group key or pair-wise shared secret keys. A node establishes shared secret keys only with its trustworthy neighbors rather than all network nodes. The protocol can prevent or detect most of the attacks common to ad hoc routing protocols. Our secure routing protocol can detect attacks, such as the modification of RREQ and the modification of sequence number. The protocol is also able to exclude attackers timely and proactively. Moreover the protocol is capable of discovering multiple routes existed between two nodes and is also appropriate for dynamically changing network topology.
- We developed Anchor-based Routing Protocol with Cell ID Management System ARPC, a scalable

routing protocol for ad hoc networks. It is a hybrid routing protocol, which combines the table-based routing strategy with the geographic routing strategy. However, GPS (Global Positioning System) support is not needed. ARPC consists of a location-based clustering protocol, an intra-cell routing protocol, an inter-cell routing protocol, and a Cell ID Management System. The location-based clustering protocol divides the network region into various cells. Each node knows the cell ID of the cell it is present in. The intra-cell routing protocol routes packets within one cell. The inter-cell routing protocol is used to route packets between nodes in different cells. The Cell ID Management System manages the cell IDs of all the nodes in the network. The combination of intra-cell and inter-cell routing protocols makes ARPC highly scalable, since each node needs to only maintain routes within the cell it is present in. The inter-cell routing protocol establishes multiple routes between different cells, which makes ARPC reliable and efficient. We evaluated the performance of ARPC using the ns2 simulator. Simulation results show that ARPC is efficient and scales well to large networks. ARPC combines the advantages of multi-path routing strategy and geographic routing strategy—efficiency and scalability, and avoids the burden—GPS support.

Contributions

- We developed a new routing model, Way Point Routing (WPR), which divides an active route into segments by selecting a number of nodes on the route as waypoint nodes. An inter-segment routing protocol runs globally, while an intra-segment routing protocol runs locally. A distinct advantage of WPR is that when a route breaks because of node mobility or node failure, instead of discarding the whole route and discovering a new route from the source to the destination, only the two waypoint nodes of the broken segment need to find a new segment, significantly reducing the routing overhead and end-to-end delay. We developed an instantiation of WPR termed as DSR Over AODV (DOA). In DOA, DSR is used for inter-segment routing, and AODV is used for intra-segment routing. Simulation results show that DOA scales well for networks with more than 1000 nodes, routing overhead is significantly reduced while other metrics are better or comparable to AODV and DSR.
- We developed a new route discovery scheme called multiple-target route discovery (MTRD). A RREQ message in MTRD may contain multiple targets and discover these targets simultaneously. When a source needs to perform a route discovery, instead of immediately broadcasting a new RREQ message, it first tries to piggyback its request on the existing RREQ packets that it relays. This has potential to improve the routing performance by reducing the control overhead, congestion, and power consumption at nodes. The results of a simulation study confirmed the effectiveness of MTRD by showing that MTRD significantly reduces the control overhead, improves the packet delivery ratio, and reduces the end-to-end delay when the network load is medium to heavy.
- If the RREP message is lost during a route discovery, a large amount of route discovery effort will be wasted. We developed the idea of salvaging route reply (SRR), which attempts to salvage an undeliverable RREP in two possible ways. First, the node (salvor) looks up its own route cache for an alternate path to the source. Second, the node runs a one-hop SRR route discovery to find a path to the source. The SRR route discovery succeeds most of the time because neighboring nodes recently propagated the RREQ originated from the source and learned a reverse path to the source. The results of a simulation study show that SRR significantly improves the performance of AODV in all metrics, namely, packet delivery ratio, control overhead and end-to-end delay.
- 4. We developed a geographic routing protocol (GRPu) for ad-hoc networks with unidirectional links. GRPu inherits the best of three well-known techniques for routing in ad-hoc networks, viz., reactive route discovery, greedy forwarding, and geographic source-routing. In GRPu, a source node establishes an on-demand loop-free geographic path to the destination by carefully handling unidirectional links. Unlike node ID-based paths used by DSR, geographic paths decouple node ID's from the path. Thus any node along the geographic path can forward packets to the destination. Further, to utilize geographic paths in sparser networks, nodes use a path-healing mechanism. Path-healing helps geographic paths adapt according to the node mobility, and greatly mitigates path breaks. Simulation results show that the GRPu protocol achieves a high percentage packet delivery, low control overhead, and low average hop count for a wide range of network scenarios.

- We developed a self-healing on-demand geographic path-based routing protocol(OGPR) for ad-hoc networks. Unlike other position-based routing protocols, which typically assume a location service protocol, OGPR uses an implicit location service to find the position of the destination. OGPR protocol establishes geographic paths that avoid dead-ends due to greedy forwarding, in static networks. Simulation results show that OGPR achieves higher packet delivery rate and lower control overhead than AODV, DSR, and GPSR+GLS protocols, under a wide range of network scenarios.
- To protect an ad hoc routing protocol, we developed a secure routing protocol for ad hoc networks with a shared group key as the sole assumption. The key security measures in this protocol are distributed authentication and Message Authentication Code. We developed a Distributed Authentication Model, with which different nodes can authenticate each other. Integrity is ensured by Message Authentication Code, which is calculated by using the shared group key or pair-wise shared secret keys. A node establishes shared secret keys only with its trustworthy neighbors rather than all network nodes. The protocol can prevent or detect most of the attacks common to ad hoc routing protocols.
- We developed Anchor-based Routing Protocol with Cell ID Management System ARPC, a scalable routing protocol for ad hoc networks. It is a hybrid routing protocol, which combines the table-based routing strategy with the geographic routing strategy. However, GPS (Global Positioning System) support is not needed. ARPC consists of a location-based clustering protocol, an intra-cell routing protocol, an inter-cell routing protocol, and a Cell ID Management System. The location-based clustering protocol divides the network region into various cells. Each node knows the cell ID of the cell it is present in. Simulation results show that ARPC is efficient and scales well to large networks.

References

- [1] Venkata Giruka and Mukesh Singhal, “Location Service Protocols for Wireless Ad-Hoc Networks”, to appear in *Pervasive and Mobile Computing*, Elsevier.
- [2] Rendong Bai and M. Singhal, “DOA: DSR Over AODV routing for Mobile Ad-Hoc Networks”, to appear in *IEEE Transactions on Mobile Computing*.
- [3] Huaizhi Li and M. Singhal, “ARPC: Anchor-based Routing Protocol for Mobile Ad Hoc Networks With Cell ID Management System”, to appear in *International Journal of Ad Hoc and Sensor Wireless Networks*.
- [4] Karl Persson, D. Manivannan and M. Singhal. “Bluetooth Scatternet Formation: Criteria, Models and Classification”, *Ad Hoc Networks*, 3(6):777-794, November 2005, Elsevier Science.
- [5] Venkata C. Giruka and Mukesh Singhal, “A Self-healing On-demand Geographic Path-based Routing Protocol for Mobile Ad-hoc Networks”, to appear in *Ad Hoc Networks*.
- [6] Huaizhi Li and M. Singhal, “A Secure Routing Protocol for Wireless Ad Hoc Networks”, in Proc. of 39th Hawaii International Conference on System Sciences (Minitrack on Security and Survivability of Unbounded Networked Systems), January 2006.
- [7] Huaizhi Li and M. Singhal, “An Anchor-Based Routing Protocol with Cell ID Management System for Ad Hoc Networks”, in the Proc. of ICCCN, San Diego, Nov 2005, pp. 215-222.
- [8] R. Bai and M. Singhal, “Salvaging Route Reply for On-Demand Routing Protocols in Mobile Ad-Hoc Networks”, in the Proc. of the 8th ACM/IEEE International Symposium on Modeling, Analysis and Simulation of Wireless and Mobile Systems, (MSWiM 2005), October 10-13, 2005, pp. 53-62, Montreal, Canada.
- [9] Rendong Bai and Mukesh Singhal, “Multiple-Target Route Discovery in Mobile Ad-Hoc Networks”, submitted to the 3rd IEEE International Conference on Mobile Ad-Hoc and Sensor Systems (MASS-2006), October 9-12, 2006, Vancouver, Canada.
- [10] Venkata C. Giruka, Mukesh Singhal and Saikiran, “Geographic Routing in Presence of Unidirectional Links for Wireless Ad-hoc Networks”, submitted to the 9th ACM-IEEE International Symposium on Modeling, Spain, 2006.

7 Intellectual Property

- **Algorithm for seamless handoff in WLAN:** We have developed an advanced WLAN algorithm (filed with the USA Patent office) that can be integrated into smart devices (laptops, mobile phones, wireless PDAs), opening up the opportunity of enjoying, creating, and, delivering rich real-time services (multimedia, voice services) over traditional IEEE 802.11 networks. Mobile clients can roam freely within the same or different subnetworks belonging to same or different service providers while avoiding the delays and latencies associated with traditional Wi-Fi roaming. Our approach also supports all existing and future security functionalities approved by the IEEE 802.11 standard with additional provision for embedded security profiling; mobile clients need not worry about packet level security as the underlying network (or provider) is transparently changed.
- **Algorithm for Intelligent Rate Control in WLANs** We have filed an invention disclosure with the UTA IP committee for the the design and development of a new algorithm for that is capable of rate-adaptation in IEEE 802.11 networks. The approach is primarily aimed towards improving the performance of delay sensitive applications like Voice-over-IP (VoIP) and streaming multimedia applications.

8 CrewNet: Secured Wireless Sensor Network Testbed Providing Web-based Control

Location : CReWMaN

Faculty : Sajal Das, Kalyan Basu, Younghe Liu CrewNet is a secured wireless sensor network testbed for indoor/outdoor environmental monitoring. It provides web-based application-/task-specific control interface as well as query interface. The ultimate goal of our testbed is to provide secured framework while providing application-specific unified or optimized functions. CrewNet establishes and enforces security using the following mechanisms:

- **Key establishment and management:** the challenge is how to develop an efficient mechanism to deploy keys to all the sensors with desirable scalability under the stringent resource limitation with a goal to achieve how to design effective security protocols to distribute, establish and maintain the keys among the sensors.
- **Secure routing:** With the goal to prevent malicious nodes from launching attacks that either tries to change the topology (routing information) of the network or deplete the resource of legal nodes.
- **Information security:** Once some nodes get compromised, the goal is to detect the compromised nodes and minimize the damage caused by them.
- **Energy efficient low layer protocols:** Approaches to reduce extra energy consumption caused by low layer communication when a network uses broadcast message.

9 Direct Impact of PSI

Project Coordination and Interaction

The PSI project has fostered excellent working relation and collaboration among the PIs and Co-PIs from UTA, UKY and PSU. There is intense exchange of ideas between the various personnel so as to develop an understanding of the mechanism and architecture on how to transform useful research into meaningful working prototype.

At UTA, there are regular monthly face-to-face meetings with students and faculty working in various project groups and subgroups. Two workshops (2004, 2005) on PSI have already been successfully conducted at UTA; a third one is planned during the summer of 2006. The workshops provide a platform that encourage the faculty and students to share research findings, ideas, and discuss plans for further collaborative research.

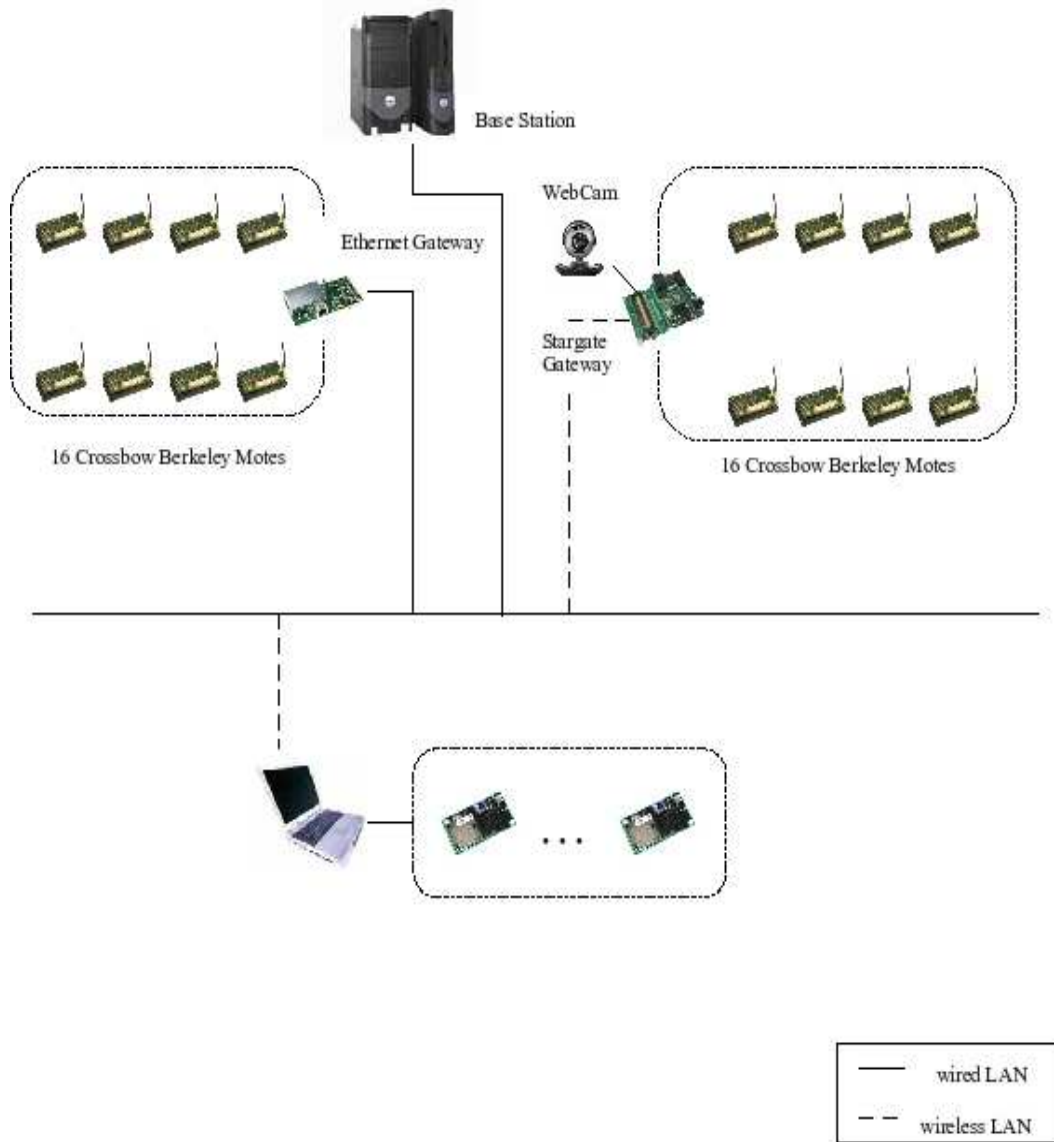


Figure 3: CrewNet configuration: Each 16 MICAz motes shown in the figure, can detect environmental phenomena of two different rooms and 4 MICAz motes connected to GPS sensor board sense outdoor environmental values. Data is finally sent to a base station as raw data or aggregated form.

Project Outcome - To Date

The PSI project has already resulted in eight PhD dissertations from the University of Texas at Arlington (UTA) and two from Pennsylvania State University (PSU). At the same time, nine Masters theses were awarded from UTA and two from PSU. To date, there are about twenty PhD dissertations being pursued at UTA. This project has supported five PhD students at UKY and five PhD students at PSU. Two undergrad students have been trained through REU supplements at UTA. Four new graduate level courses on have been introduced at the three collaborating universities. In particular, one graduate course on Wireless Security and another on Advances in Sensor Networking have been designed and taught at UTA in Spring 2005; a graduate course on Wireless Networks Security was taught at UKY in

Fall 2004; and an advanced graduate level course in the area of Heterogeneous and Mobile Data Bases was developed and offered at PSU in Spring 2005. These courses will be offered at regular intervals in respective universities. Furthermore, the course materials will be shared and offered at the collaborative universities. The objectives of such courses include the dissemination of basic and advanced concepts in those emerging topics, and encouraging research among students. Students are gaining hand-on experience in experimenting with sensor network testbed developed at UTA. Finally, students have had the opportunity to publish their research work at high quality conferences and journals.

Project Leverage

Multiple security projects out of this ITR project. They include SafetyNet: project for border security including perimeter control, airport/harbor security, Content Based Routing for imemediated notification of security events, High performance packet classifier for anomaly detection in traffic streams, and video sensor networks for surveillance in highly uncertain dynamic environments. They were submitted to NSF, Federal Earmark funding and Intel Corporation. Additionally, there are two proposals pending with the Cybertrust program and with NOSS.

A new research proposal aimed, at building a multi-layer security framework, got funded in May 2006 by the Texas Advanced Research Program. Additionally, an NSF MRI grant for training undergraduates in network and pervasive security also got funded in 2006. This is in addition to the NSF MRI proposal that got funded in 2004 to help procure equipment needed to develop a prototype of the reseach proposed in this ITR project. This ITR project has also resulted in a spin-off in Bioinformatics and System Biology. This has further led to collaboration with UT Southwestern Medical center in Dallas.