Matthias Klusch
Sascha Ossowski
Vipul Kashyap
Rainer Unland (Eds.)

LNAI 3191

Cooperative Information Agents VIII

8th International Workshop, CIA 2004
Erfurt, Germany, September 2004
Proceedings

Springer

Matthias Klusch Sascha Ossowski
Vipul Kashyap Rainer Unland (Eds.)

Cooperative Information Agents VIII

8th International Workshop, CIA 2004
Erfurt, Germany, September 27-29, 2004
Proceedings

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Matthias Klusch
German Research Center for Artifical Intelligence,
Deduction and Multiagent Systems
Stuhlsatzenhausweg 3, 66123 Saarbrücken, Germany
E-mail: klusch@dfki.de

Sascha Ossowski
University Rey Juan Carlos, School of Engineering (ESCET)
Campus de Mostoles, Calle Tulipan s/n, 28933 Madrid, Spain
E-mail: s.ossowski@escet.urjc.es

Vipul Kashyap
Lister Hill National Center on Biomedical Communications
8600 Rockville Pike, Bethesda MD 20894, USA
E-mail: kashyap@nlm.nih.gov

Rainer Unland
University of Duisburg-Essen
Institute for Computer Science and Business Information Systems
Schützenbahn 70, 45117 Essen, Germany
E-mail: UnlandR@informatik.uni-essen.de

Library of Congress Control Number: 2004112167

CR Subject Classification (1998): I.2.11, I.2, H.4, H.3.3, H.2, C.2.4, H.5

ISSN 0302-9743
ISBN 3-540-23170-6 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11323815 06/3142 5 4 3 2 1 0

Lecture Notes in Artificial Intelligence 3191
Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Preface

These are the proceedings of the 8th International Workshop on Cooperative Information Agents (CIA 2004), held at the Fair and Congress Center in Erfurt, Germany, September 27–29, 2004. It was part of the multi-conference Net.ObjectDays 2004, and, in particular, was co-located with the 2nd German Conference on Multiagent Systems Technologies (MATES 2004).

In today's networked world of linked heterogeneous, pervasive computer systems, devices, and information landscapes, the intelligent coordination and provision of relevant added-value information at any time, anywhere, by means of cooperative information agents becomes increasingly important for a variety of applications. An information agent is a computational software entity that has access to one or multiple, heterogeneous, and geographically dispersed data and information sources. It proactively searches for and maintains information on behalf of its human users, or other agents, preferably just in time. In other words, it is managing and overcoming the difficulties associated with information overload in open, pervasive information and service landscapes. Cooperative information agents may collaborate with each other to accomplish both individual and shared joint goals depending on the actual preferences of their users, budgetary constraints, and resources available. One major challenge of developing agent-based intelligent information systems in open environments is to balance the autonomy of networked data, information, and knowledge sources with the potential payoff of leveraging them using information agents.

Interdisciplinary research and development of information agents requires expertise in relevant domains of information retrieval, artificial intelligence, database systems, human-computer interaction, and Internet and Web technology. The CIA 2004 workshop aimed at providing an interdisciplinary forum for researchers, software developers, and managers to get informed about, present, and discuss the latest high-quality results in advancements of theory and practice in information agent technology for the Internet, the Web, and the Semantic Web.

The CIA workshop series was initiated in 1997; since then the events of the series have been held annually, mainly at different locations in Europe. Each event of the series offers regular and invited talks of excellence that are given by distinguished experts, and system demonstrations, and honors innovative research and development on information agents by means of best paper awards and system innovation awards. The proceedings of the series are regularly published as volumes of the Lecture Notes in Artificial Intelligence (LNAI) series of Springer.

In keeping with its tradition, this year's workshop featured a sequence of regular and invited talks of excellence given by leading experts in the field. These talks covered a broad area of topics of interest, such as information agents for pervasive and peer-to-peer computing environments; issues of communication and cooperation; industrial applications; and recommender agents.

This year the CIA System Innovation Award and the CIA Best Paper Award were sponsored by Whitestein Technologies AG, Switzerland, and the Spanish Association for Artificial Intelligence (AEPIA), respectively. The DMR Decision Engineering Lab of the Universidad Rey Juan Carlos in Madrid, Spain provided limited financial support to students who were co-authors of accepted papers so that they could give their presentations at the CIA 2004 workshop.

CIA 2004 featured 3 invited and 18 regular papers selected from 59 submissions. The result of the peer-review of all contributions is included in this volume, rich with interesting, inspiring, and advanced work on research and development on intelligent information agents worldwide. All previous workshop proceedings were published by Springer as Lecture Notes in Artificial Intelligence volumes: 1202 (1997), 1435 (1998), 1652 (1999), 1860 (2000), 2182 (2001), 2446 (2002), and 2782 (2003).

The CIA 2004 workshop was organized in cooperation with the Association for Computing Machinery (ACM). In addition, we are very much indebted to our sponsors, whose financial support made this event possible. The sponsors of CIA 2004 were:

tranSIT GmbH, Germany
URJC Decision Engineering Lab (DMR), Spain
Whitestein Technologies, Switzerland
Spanish Association for Artificial Intelligence, Spain
AgentLink III, EU FP6 Coordinating Action

We are in particular grateful to the authors and invited speakers for contributing their latest and inspiring work to this workshop, as well as to the members of the program committee, and the external reviewers for their critical reviews of submissions. Finally, a deep thanks goes to each of the brave members of the local organization team from tranSIT GmbH at the Fair and Congress Center in Erfurt for their hard work in providing the CIA 2004 event with a modern, comfortable location, and a social program.

September 2004

Matthias Klusch, Sascha Ossowski,
Vipul Kashyap, Rainer Unland

Co-chairs

Matthias Klusch	DFKI, Germany, *Workshop Chair*
Sascha Ossowski	University of Rey Juan Carlos in Madrid, Spain
Vipul Kashyap	Lister Hill National Center on Biomedical Communications, USA
Rainer Unland	University of Duisburg-Essen, Germany

Program Committee

Elisabeth Andre	(University of Augsburg, Germany)
Wolfgang Benn	(TU Chemnitz, Germany)
Monique Calisti	(Whitestein Technologies, Switzerland)
Brahim Chaib-draa	(Laval University, Canada)
Yves Demazeau	(Leibniz, France)
Frank Dignum	(University of Utrecht, The Netherlands)
Tharam Dillon	(La Trobe University, Australia)
Boi Faltings	(EPFL, Switzerland)
Fausto Giunchiglia	(University of Trento, Italy)
Rune Gustavsson	(Blekinge TH, Sweden)
Heikki Helin	(TeliaSonera AB, Sweden/Finland)
Mike Huhns	(University of South Carolina, USA)
Toru Ishida	(University of Kyoto, Japan)
Catholijn Jonker	(Free University of Amsterdam, Netherlands)
Hillol Kargupta	(UMBC, USA)
Manolis Koubarakis	(Technical University of Crete, Greece)
Daniel Kudenko	(University of York, UK)
Victor Lesser	(University of Massachusetts, USA)
Beatriz Lopez	(University of Girona, Spain)
Mike Luck	(University of Southampton, UK)
Pablo Noriega	(IIIA CSIC, Spain)
Werner Nutt	(Heriot-Watt University, Edinburgh, UK)
Eugenio Oliveira	(University of Porto, Portugal)
Andrea Omicini	(University of Bologna, Italy)
George Papadopoulos	(University of Cyprus, Cyprus)
Michal Pechoucek	(TU Prague, Czech Republic)
Paolo Petta	(Austrian Research Institute for AI, Austria)
Alun Preece	(University of Aberdeen, UK)
Volker Roth	(Fraunhofer IGD, Germany)
Claudio Sartori	(University of Bologna, Italy)
Ken Satoh	(National Institute of Informatics, Japan)
Heiko Schuldt	(UMIT Innsbruck, Austria)

Onn Shehory	(IBM Research, Israel)
Von-Wun Soo	(National Tsing Hua University, Taiwan)
Steven Willmott	(UPC Barcelona, Spain)
Mike Wooldridge	(University of Liverpool, UK)
Makoto Yokoo	(NTT Communication, Japan)
Franco Zambonelli	(University of Modena, Italy)
Ning Zhong	(Maebashi Institute of Technology, Japan)

External Reviewers

Ilya Zaihrayeu
Pavel Shvaiko
Jan Treur
Peter-Paul van Maanen
Henrique Lopes Cardoso

Lecture Notes in Artificial Intelligence (LNAI)

Vol. 3249: B. Buchberger, J.A. Campbell (Eds.), Artificial Intelligence and Symbolic Computation. X, 285 pages. 2004.

Vol. 3238: S. Biundo, T. Frühwirth, G. Palm (Eds.), KI 2004: Advances in Artificial Intelligence. XI, 467 pages. 2004.

Vol. 3229: J.J. Alferes, J. Leite (Eds.), Logics in Artificial Intelligence. XIV, 744 pages. 2004.

Vol. 3206: P. Sojka, I. Kopecek, K. Pala (Eds.), Text, Speech and Dialogue. XIII, 667 pages. 2004.

Vol. 3202: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Knowledge Discovery in Databases: PKDD 2004. XIX, 560 pages. 2004.

Vol. 3201: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.), Machine Learning: ECML 2004. XVIII, 580 pages. 2004.

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004.

Vol. 3192: C. Bussler, D. Fensel (Eds.), Artificial Intelligence: Methodology, Systems, and Applications. XIII, 522 pages. 2004.

Vol. 3191: M. Klusch, S. Ossowski, V. Kashyap, R. Unland (Eds.), Cooperative Information Agents VIII. XI, 303 pages. 2004.

Vol. 3187: G. Lindemann, J. Denzinger, I.J. Timm, R. Unland (Eds.), Multiagent System Technologies. XIII, 341 pages. 2004.

Vol. 3176: O. Bousquet, U. von Luxburg, G. Rätsch (Eds.), Advanced Lectures on Machine Learning. IX, 241 pages. 2004.

Vol. 3159: U. Visser, Intelligent Information Integration for the Semantic Web. XIV, 150 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004.

Vol. 3070: L. Rutkowski, J. Siekmann, R. Tadeusiewicz, L.A. Zadeh (Eds.), Artificial Intelligence and Soft Computing - ICAISC 2004. XXV, 1208 pages. 2004.

Vol. 3068: E. André, L. Dybkjær, W. Minker, P. Heisterkamp (Eds.), Affective Dialogue Systems. XII, 324 pages. 2004.

Vol. 3067: M. Dastani, J. Dix, A. El Fallah-Seghrouchni (Eds.), Programming Multi-Agent Systems. X, 221 pages. 2004.

Vol. 3066: S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse (Eds.), Rough Sets and Current Trends in Computing. XX, 853 pages. 2004.

Vol. 3065: A. Lomuscio, D. Nute (Eds.), Deontic Logic in Computer Science. X, 275 pages. 2004.

Vol. 3060: A.Y. Tawfik, S.D. Goodwin (Eds.), Advances in Artificial Intelligence. XIII, 582 pages. 2004.

Vol. 3056: H. Dai, R. Srikant, C. Zhang (Eds.), Advances in Knowledge Discovery and Data Mining. XIX, 713 pages. 2004.

Vol. 3055: H. Christiansen, M.-S. Hacid, T. Andreasen, H.L. Larsen (Eds.), Flexible Query Answering Systems. X, 500 pages. 2004.

Vol. 3040: R. Conejo, M. Urretavizcaya, J.-L. Pérez-de-la-Cruz (Eds.), Current Topics in Artificial Intelligence. XIV, 689 pages. 2004.

Vol. 3035: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XII, 326 pages. 2004.

Vol. 3034: J. Favela, E. Menasalvas, E. Chávez (Eds.), Advances in Web Intelligence. XIII, 227 pages. 2004.

Vol. 3030: P. Giorgini, B. Henderson-Sellers, M. Winikoff (Eds.), Agent-Oriented Information Systems. XIV, 207 pages. 2004.

Vol. 3029: B. Orchard, C. Yang, M. Ali (Eds.), Innovations in Applied Artificial Intelligence. XXI, 1272 pages. 2004.

Vol. 3025: G.A. Vouros, T. Panayiotopoulos (Eds.), Methods and Applications of Artificial Intelligence. XV, 546 pages. 2004.

Vol. 3020: D. Polani, B. Browning, A. Bonarini, K. Yoshida (Eds.), RoboCup 2003: Robot Soccer World Cup VII. XVI, 767 pages. 2004.

Vol. 3012: K. Kurumatani, S.-H. Chen, A. Ohuchi (Eds.), Multi-Agnets for Mass User Support. X, 217 pages. 2004.

Vol. 3010: K.R. Apt, F. Fages, F. Rossi, P. Szeredi, J. Váncza (Eds.), Recent Advances in Constraints. VIII, 285 pages. 2004.

Vol. 2990: J. Leite, A. Omicini, L. Sterling, P. Torroni (Eds.), Declarative Agent Languages and Technologies. XII, 281 pages. 2004.

Vol. 2980: A. Blackwell, K. Marriott, A. Shimojima (Eds.), Diagrammatic Representation and Inference. XV, 448 pages. 2004.

Vol. 2977: G. Di Marzo Serugendo, A. Karageorgos, O.F. Rana, F. Zambonelli (Eds.), Engineering Self-Organising Systems. X, 299 pages. 2004.

Vol. 2972: R. Monroy, G. Arroyo-Figueroa, L.E. Sucar, H. Sossa (Eds.), MICAI 2004: Advances in Artificial Intelligence. XVII, 923 pages. 2004.

Vol. 2969: M. Nickles, M. Rovatsos, G. Weiss (Eds.), Agents and Computational Autonomy. X, 275 pages. 2004.

Vol. 2961: P. Eklund (Ed.), Concept Lattices. IX, 411 pages. 2004.

Vol. 2953: K. Konrad, Model Generation for Natural Language Interpretation and Analysis. XIII, 166 pages. 2004.

Vol. 2934: G. Lindemann, D. Moldt, M. Paolucci (Eds.), Regulated Agent-Based Social Systems. X, 301 pages. 2004.

Vol. 2930: F. Winkler (Ed.), Automated Deduction in Geometry. VII, 231 pages. 2004.

Vol. 2926: L. van Elst, V. Dignum, A. Abecker (Eds.), Agent-Mediated Knowledge Management. XI, 428 pages. 2004.

Vol. 2923: V. Lifschitz, I. Niemelä (Eds.), Logic Programming and Nonmonotonic Reasoning. IX, 365 pages. 2004.

Vol. 2915: A. Camurri, G. Volpe (Eds.), Gesture-Based Communication in Human-Computer Interaction. XIII, 558 pages. 2004.

Vol. 2913: T.M. Pinkston, V.K. Prasanna (Eds.), High Performance Computing - HiPC 2003. XX, 512 pages. 2003.

Vol. 2903: T.D. Gedeon, L.C.C. Fung (Eds.), AI 2003: Advances in Artificial Intelligence. XVI, 1075 pages. 2003.

Vol. 2902: F.M. Pires, S.P. Abreu (Eds.), Progress in Artificial Intelligence. XV, 504 pages. 2003.

Vol. 2892: F. Dau, The Logic System of Concept Graphs with Negation. XI, 213 pages. 2003.

Vol. 2891: J. Lee, M. Barley (Eds.), Intelligent Agents and Multi-Agent Systems. X, 215 pages. 2003.

Vol. 2882: D. Veit, Matchmaking in Electronic Markets. XV, 180 pages. 2003.

Vol. 2871: N. Zhong, Z.W. Raś, S. Tsumoto, E. Suzuki (Eds.), Foundations of Intelligent Systems. XV, 697 pages. 2003.

Vol. 2854: J. Hoffmann, Utilizing Problem Structure in Planing. XIII, 251 pages. 2003.

Vol. 2843: G. Grieser, Y. Tanaka, A. Yamamoto (Eds.), Discovery Science. XII, 504 pages. 2003.

Vol. 2842: R. Gavaldá, K.P. Jantke, E. Takimoto (Eds.), Algorithmic Learning Theory. XI, 313 pages. 2003.

Vol. 2838: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Knowledge Discovery in Databases: PKDD 2003. XVI, 508 pages. 2003.

Vol. 2837: N. Lavrač, D. Gamberger, L. Todorovski, H. Blockeel (Eds.), Machine Learning: ECML 2003. XVI, 504 pages. 2003.

Vol. 2835: T. Horváth, A. Yamamoto (Eds.), Inductive Logic Programming. X, 401 pages. 2003.

Vol. 2821: A. Günter, R. Kruse, B. Neumann (Eds.), KI 2003: Advances in Artificial Intelligence. XII, 662 pages. 2003.

Vol. 2807: V. Matoušek, P. Mautner (Eds.), Text, Speech and Dialogue. XIII, 426 pages. 2003.

Vol. 2801: W. Banzhaf, J. Ziegler, T. Christaller, P. Dittrich, J.T. Kim (Eds.), Advances in Artificial Life. XVI, 905 pages. 2003.

Vol. 2797: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. XII, 281 pages. 2003.

Vol. 2792: T. Rist, R.S. Aylett, D. Ballin, J. Rickel (Eds.), Intelligent Virtual Agents. XV, 364 pages. 2003.

Vol. 2782: M. Klusch, A. Omicini, S. Ossowski, H. Laamanen (Eds.), Cooperative Information Agents VII. XI, 345 pages. 2003.

Vol. 2780: M. Dojat, E. Keravnou, P. Barahona (Eds.), Artificial Intelligence in Medicine. XIII, 388 pages. 2003.

Vol. 2777: B. Schölkopf, M.K. Warmuth (Eds.), Learning Theory and Kernel Machines. XIV, 746 pages. 2003.

Vol. 2752: G.A. Kaminka, P.U. Lima, R. Rojas (Eds.), RoboCup 2002: Robot Soccer World Cup VI. XVI, 498 pages. 2003.

Vol. 2741: F. Baader (Ed.), Automated Deduction – CADE-19. XII, 503 pages. 2003.

Vol. 2705: S. Renals, G. Grefenstette (Eds.), Text- and Speech-Triggered Information Access. VII, 197 pages. 2003.

Vol. 2703: O.R. Zaïane, J. Srivastava, M. Spiliopoulou, B. Masand (Eds.), WEBKDD 2002 - MiningWeb Data for Discovering Usage Patterns and Profiles. IX, 181 pages. 2003.

Vol. 2700: M.T. Pazienza (Ed.), Extraction in the Web Era. XIII, 163 pages. 2003.

Vol. 2699: M.G. Hinchey, J.L. Rash, W.F. Truszkowski, C.A. Rouff, D.F. Gordon-Spears (Eds.), Formal Approaches to Agent-Based Systems. IX, 297 pages. 2002.

Vol. 2691: V. Mařík, J.P. Müller, M. Pechoucek (Eds.), Multi-Agent Systems and Applications III. XIV, 660 pages. 2003.

Vol. 2684: M.V. Butz, O. Sigaud, P. Gérard (Eds.), Anticipatory Behavior in Adaptive Learning Systems. X, 303 pages. 2003.

Vol. 2682: R. Meo, P.L. Lanzi, M. Klemettinen (Eds.), Database Support for Data Mining Applications. XII, 325 pages. 2004.

Vol. 2671: Y. Xiang, B. Chaib-draa (Eds.), Advances in Artificial Intelligence. XIV, 642 pages. 2003.

Vol. 2663: E. Menasalvas, J. Segovia, P.S. Szczepaniak (Eds.), Advances in Web Intelligence. XII, 350 pages. 2003.

Vol. 2661: P.L. Lanzi, W. Stolzmann, S.W. Wilson (Eds.), Learning Classifier Systems. VII, 231 pages. 2003.

Vol. 2654: U. Schmid, Inductive Synthesis of Functional Programs. XXII, 398 pages. 2003.

Vol. 2650: M.-P. Huget (Ed.), Communications in Multiagent Systems. VIII, 323 pages. 2003.

Vol. 2645: M.A. Wimmer (Ed.), Knowledge Management in Electronic Government. XI, 320 pages. 2003.

Table of Contents

Recommender Agents

Information Agents and Mobile Computing

Industrial Applications

Cooperation in Open Environments

Multi-agent Systems and Distributed Data Mining

Chris Giannella, Ruchita Bhargava, and Hillol Kargupta

Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250 USA
{cgiannel,ruchital,hillol}@cs.umbc.edu

Abstract. Multi-agent systems offer an architecture for distributed problem solving. Distributed data mining algorithms specialize on one class of such distributed problem solving tasks—analysis and modeling of distributed data. This paper offers a perspective on distributed data mining algorithms in the context of multi-agents systems. It particularly focuses on distributed clustering algorithms and their potential applications in multi-agent-based problem solving scenarios. It discusses potential applications in the sensor network domain, reviews some of the existing techniques, and identifies future possibilities in combining multi-agent systems with the distributed data mining technology.

Keywords: multi-agent systems, distributed data mining, clustering

1 Introduction

Multi-agent systems (MAS) often deal with complex applications that require distributed problem solving. In many applications the individual and collective behavior of the agents depend on the observed data from distributed sources. In a typical distributed environment analyzing distributed data is a non-trivial problem because of many constraints such as limited bandwidth (*e.g.* wireless networks), privacy-sensitive data, distributed compute nodes, only to mention a few. The field of Distributed Data Mining (DDM) deals with these challenges in analyzing distributed data and offers many algorithmic solutions to perform different data analysis and mining operations in a fundamentally distributed manner that pays careful attention to the resource constraints. Since multi-agent systems are also distributed systems, combining DDM with MAS for data intensive applications is appealing.

This paper makes an effort to underscore the possible synergy between multi-agent systems and distributed data mining technology. It particularly focuses on distributed clustering, a problem finding increasing number of applications in sensor networks, distributed information retrieval, and many other domains. The paper discusses one of these application domains, illustrates the ideas, and reviews existing work in this area. Although, the power of DDM is not just restricted to clustering, this paper chooses to restrict the scope for the sake of brevity.

The paper is organized as follows. Section 2 provides the motivation behind the material presented in this paper. Section 3 introduces DDM and presents an overview of the field. Section 4 focuses on a particular portion of the DDM literature and takes

M. Klusch et al. (Eds.): CIA 2004, LNAI 3191, pp. 1–15, 2004.

an in-depth look at the distributed clustering literature. Section 5 considers distributed clustering algorithms in the context of sensor networks that are drawing an increasing amount of interest from the multi-agent systems community. Finally, Section 6 concludes this paper.

2 Motivation

Agents in MAS need to be pro-active and autonomous. Agents perceive their environment, dynamically reason out actions based on conditions, and interact with each other. In some applications the knowledge of the agents that guide reasoning and action depend on the existing domain theory. However, in many complex domains this knowledge is a result of existing domain theory and also the outcome of empirical data analysis. Scalable analysis of data may require advanced data mining for detecting hidden patterns, constructing predictive models, and identifying outliers, among others. In a multi-agent system this knowledge is usually collective. This collective "intelligence" of a multi-agent system must be developed by distributed domain knowledge and analysis of distributed data observed by different agents. Such distributed data analysis may be a non-trivial problem when the underlying task is not completely decomposable and computing resources are constrained by several factors such as limited power supply, poor bandwidth connection, and privacy sensitive multi-party data, among others.

For example, consider a defense related application of monitoring a terrain using a sensor network that has many tiny mote-type [15] sensors for measuring vibration, reflectance, temperature, and audio signals. Let us say the objective is to identify and track a certain type of vehicle (*e.g.* pick-up trucks). The sensors are battery-powered. Therefore, in the normal mode they are designed not be very active. However, as soon as someone detects a possible change in scenario, the sensors must wake up, observe, reason, and collaborate with each other in order to track and identify the object of interest. The observations are usually time-series data sampled at a device specific rate. Therefore, collaboration with other sensor-nodes would require comparing data observed at different nodes. This usually requires sending a window of observations from one node to another node. This distributed problem solving environment appears to fit very well with the multi-agent framework since the solution requires semi-autonomous behavior, collaboration and reasoning among other things. However, regardless of how sophisticated the agents are, from the domain knowledge and reasoning perspective, they must perform the underlying data analysis tasks very efficiently in a distributed manner. The traditional framework of centralized data analysis and mining algorithms does not really scale very well in such distributed applications. For example, if we want to compare the data vectors observed at different sensor nodes the centralized approach will be to send the data vectors to the base station (usually connected through a wireless network) and then compare the vectors using whatever metric is appropriate for the domain. This does not scale up in large sensor networks since data transmission consumes a lot of battery power and heavy data transmission over limited bandwidth channel may produce poor response time. Distributed data mining technology offers more efficient solutions in such applications. The following discussion illustrates the power of DDM algorithms using a simple randomized technique for addressing this sensor network-related problem.

Given vectors $\boldsymbol{a} = (a_1, \ldots, a_m)^T$ and $\boldsymbol{b} = (b_1, \ldots, b_m)^T$ at two distributed site A and B, respectively, we want to approximate the Euclidean distance between them using a small number (compared to m) of messages between A and B. Note that the problem of computing the Euclidean distance between a pair of data tuples $\boldsymbol{a}$ and $\boldsymbol{b}$ can be represented as the problem of computing the inner products between them as follows:

$$d_e^2(\boldsymbol{a}, \boldsymbol{b}) =< \boldsymbol{a}, \boldsymbol{a} > + < \boldsymbol{b}, \boldsymbol{b} > -2 < \boldsymbol{a}, \boldsymbol{b} >$$

where $d_e^2(\boldsymbol{a}, \boldsymbol{b})$ denotes the Euclidean distance between $\boldsymbol{a}$ and $\boldsymbol{b}$; $< \boldsymbol{a}, \boldsymbol{b} >$ represents the inner product between $\boldsymbol{a}$ and $\boldsymbol{b}$, defined as $\sum_{i=1}^{m} a_i b_i$. Therefore, the core challenge is to develop an algorithm for distributed inner product computation. One can approach this problem in several ways. Table 1 shows a simple communication efficient randomized technique for computing the inner product between two vectors observed at two different sites.

Algorithm 2.0.1 Distributed Dot Product Algorithm($\boldsymbol{a}, \boldsymbol{b}$)

1. A sends B a random number generator seed. **[1 message]**
2. A and B cooperatively generate $k \times m$ random matrix R where $k \ll m$. Each entry is generated independently and identically from some fixed distribution with mean zero and variance one. A and B compute $\hat{a} = R\boldsymbol{a}$, $\hat{b} = R\boldsymbol{b}$, respectively.
3. A sends $\hat{a}$ to B. B computes $\hat{a}^T\hat{b} = \boldsymbol{a}^T R^T R\boldsymbol{b}$. **[k messages]**
4. B computes $D = \frac{\hat{a}^T\hat{b}}{k}$.

So instead of sending a m-dimensional vector to the other site, we only need to send a k-dimensional vector where $k \ll m$ and the dot product can still be estimated accurately. Indeed, it can be shown that the expected value of D is $< \boldsymbol{a}, \boldsymbol{b} >$ and Table 1 shows some experimental results concerning accuracy.

This algorithm illustrates a simple communication-efficient-way to compare a pair of data vectors observed at two different nodes. It potentially offers a building block to support the collaborative object identification and tracking problem in sensor networks where the centralized solution does not work because of limited bandwidth and power supply for the sensor nodes.

Privacy of the data can be another reason for adopting the DDM technology. In many applications, particularly in security-related applications, data are privacy-sensitive. When the data are multi-party and privacy-sensitive, centralizing the data is usually not acceptable. Therefore, many data mining applications in such domains must analyze data in a distributed fashion without having to first download everything to a single site. There exists a growing number of DDM algorithms that address many data mining problems for distributed environments. The following section presents an overview.

3 Distributed Data Mining: A Brief Overview

Data mining [9], [10], [11],and [31] deals with the problem of analyzing data in scalable manner. Distributed data mining is a branch of the field of data mining that offers a

Table 1. Relative errors in computing the dot product between two synthetic binary vectors each with 10000 elements. k is the number of randomized iterations. k is also represented as the percentage of the size of the original vectors. Each entry of the random matrix is chosen independently from U(1,−1)

k	Mean	Var	Min	Max
100(1%)	0.1483	0.0098	0.0042	0.3837
500(5%)	0.0795	0.0035	0.0067	0.2686
1000(10%)	0.0430	0.0008	0.0033	0.1357
2000(20%)	0.0299	0.0007	0.0012	0.0902
3000(30%)	0.0262	0.0005	0.0002	0.0732

framework to mine distributed data paying careful attention to the distributed data and computing resources.

In the DDM literature, one of two assumptions is commonly adopted as to how data is distributed across sites: homogeneously (horizontally partitioned) and heterogeneously (vertically partitioned). Both viewpoints adopt the conceptual viewpoint that the data tables at each site are partitions of a single global table. In the homogeneous case, the global table is horizontally partitioned. The tables at each site are subsets of the global table; they have exactly the same attributes. In the heterogeneous case the table is vertically partitioned, each site contains a collection of columns (sites do not have the same attributes). However, each tuple at each site is assumed to contain a unique identifier to facilitate matching. It is important to stress that the global table viewpoint is strictly conceptual. It is not necessarily assumed that such a table was physically realized and partitioned to form the tables at each site. Figures 1 and 2 illustrate the homogeneously distributed case using an example from weather data. Both tables use the same set of attributes. On the other hand, Figures 3 and 4 illustrate the heterogeneously distributed case. The tables have different attributes and tuples are linked through a unique identifier, *Timestamp*.

The development of data mining algorithms that work well under the constraints imposed by distributed datasets has received significant attention from the data mining community in recent years. The field of DDM has emerged as an active area of study. The bulk of DDM methods in the literature operate over an abstract architecture which includes multiple sites having independent computing power and storage capability. Local computation is done on each of the sites and either a central site communicates with each distributed site to compute the global models or a peer-to-peer architecture is used. In the latter case, individual nodes might communicate with a resource rich centralized node, but they perform most of the tasks by communicating with neighboring nodes by message passing over an asynchronous network. For example, the sites may represent independent sensor nodes which connect to each other in an ad-hoc fashion.

Some features of a distributed scenario where DDM is applicable are as follows.

1. The system consist of multiple independent sites of data and computation which communicate only through message passing.
2. Communication between the sites is expensive.