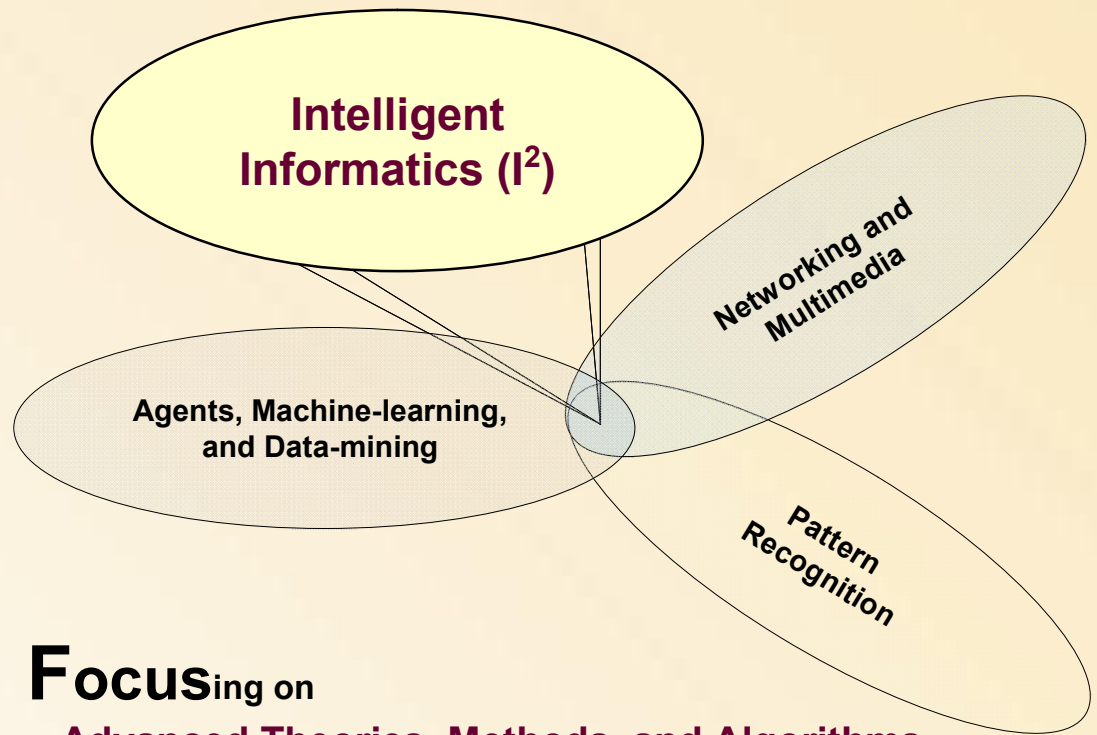


Areas of Research Strength



Focusing on

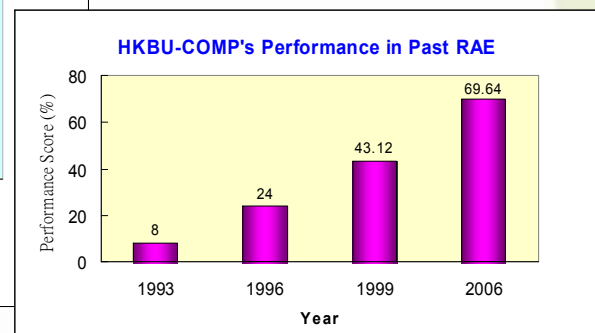
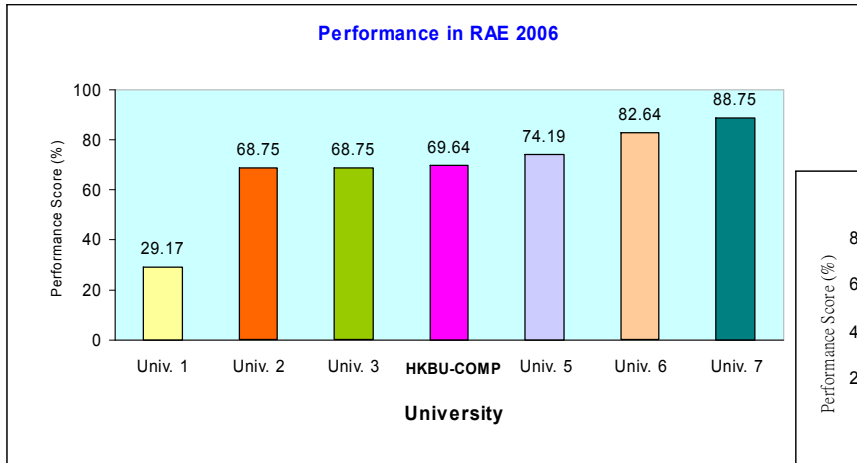
Advanced Theories, Methods, and Algorithms
of **Information Processing and Communications** for
Adaptive Interfaces, Systems, and Networks

Research & Development Laboratories

Laboratories	Expertise
1. Agents, Machine-learning, and Data-mining (AMD)	<ul style="list-style-type: none"> Autonomous Agents and Multi-agent Systems Data Mining and Machine Learning Intelligent Software Tools and Systems On-line Media and Network Analysis Web Intelligence
2. Networking and Multimedia	<ul style="list-style-type: none"> All-optical Networks IP Networks Mobile Computing Multimedia Communications and Systems Visual Information Systems Wireless Communications
3. Pattern Recognition	<ul style="list-style-type: none"> Image Processing Vision-based Human Understanding Technology Wavelet Analysis and its Applications

Research Assessment Exercise (RAE) 2006

The Department has achieved very good research performance in the past several years, as confirmed in the Research Assessment Exercise 2006 conducted by the UGC. Our research performance scores among local universities in the 2006 exercise and in the past exercises are presented as follows:



Research Outputs

Number of Refereed International **Journal** Papers

	No. of Papers	Average No. of Papers per Faculty
2005-2006	38	2.714
2004-2005	42	3.50
2003-2004	32	2.285

Monographs, Books, & Proceedings Written or Edited, 2003-2006

Monographs	5
Edited Books & Proceedings	12
Total	17





Editorships & Editorial Boards

Editors-in-Chief:

Annual Review of Intelligent Informatics (World Scientific)
International Journal of Wavelets, Multiresolution and Information Processing (World Scientific)
The IEEE Intelligent Informatics Bulletin (IEEE Computer Society)
Web Intelligence and Agent Systems (IOS Press, The Netherlands)

Editorial Boards:

IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews
International Journal of Applied Systematic Studies (Inderscience)
International Journal of Document Analysis and Recognition (Springer)
International Journal of Pattern Recognition and Artificial Intelligence (World Scientific)
Journal of Data Warehousing and Mining (IGI Global)
Journal of Embedded Computing (IOS Press)
Journal of Mobile Multimedia (Rinton Press)
Journal of Multimedia Tools and Applications (Springer)
Journal of Pervasive Computing and Communications (Troubador)
Journal of Ubiquitous Computing and Intelligence (American Scientific Publishers)
Knowledge and Information Systems (Springer)
Microprocessor and Microsystems (Elsevier)
Pattern Recognition (Elsevier)
Real-Time Systems: The International Journal of Time-Critical Computing Systems (Kluwer)

... in High-Impact Journals

Artificial Intelligence
Communications of the ACM
IEEE Computer
IEEE Intelligent Systems
IEEE Internet Computing
IEEE Journal of Lightwave Technology
IEEE Journal of Selected Areas in Communications
IEEE Multimedia
IEEE Transactions on Circuits and Systems for Video Technology
IEEE Transactions on Communications
IEEE Transactions on Computers
IEEE Transactions on Evolutionary Computation
IEEE Transactions on Image Processing
IEEE Transactions on Knowledge and Data Engineering
IEEE Transactions on Medical Imaging
IEEE Transactions on Mobile Computing
IEEE Transactions on Multimedia
IEEE Transactions on Neural Networks
IEEE Transactions on Parallel and Distributed Computing
IEEE Transactions on Pattern Analysis and Machine Intelligence
IEEE Transactions on Signal Processing
IEEE Transactions on Systems, Mans and Cybernetics
IEEE/ACM Transactions on Networking
Journal of Systems and Software
Pattern Recognition
Real-Time Systems Journal



Major Conferences Organized/Co-organized, 2003-2006

Including ...

- 2006 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2006 & IAT 2006)
- 2006 IEEE International Conference on Data Mining (ICDM 2006)
- 18th International Conference on Pattern Recognition (ICPR 2006)
- 2005 International Conference on Computational Intelligence and Security (CIS 2005)
- 2005 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2005 & IAT 2005)
- 11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA 2005)
- 2005 IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE 2005)
- 2004 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2004 & IAT 2004)
- 2004 IEEE International Conference on e-Technology, e-Commerce, and e-Service (EEE 2004)
- 2003 IEEE/WIC International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI 2003 & IAT 2003)
- 4th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL 2003)



Centre for e-Transformation Research (CTR)

(An affiliated research centre of Web Intelligence Consortium)

Mission

To be a major contributor to Hong Kong's knowledge base in researching the new fundamental roles and practical impacts of Artificial Intelligence (AI) and advanced Information Technology (IT) on the next generation of Web-empowered products, systems, services, and activities

Milestones

- 2003: Established under the Faculty of Science
- 2004: Funded by HK Research Grants Council (first IT group research project funded)
- 2007: Advanced e-Transformation Technology Research becomes a University's strategic research area

Research Team

Director:

Jiming Liu, Head and Professor
Computer Science Department, HKBU

Area Leaders:

- Qiang Yang, Professor
Computer Science Department, HKUST
- C. J. Tan, Director and Chair Professor
E-Business Technology Institute (ETI), HKU

Coordinator:

William K. Cheung, Associate Professor
Computer Science Department, HKBU



Key Members:

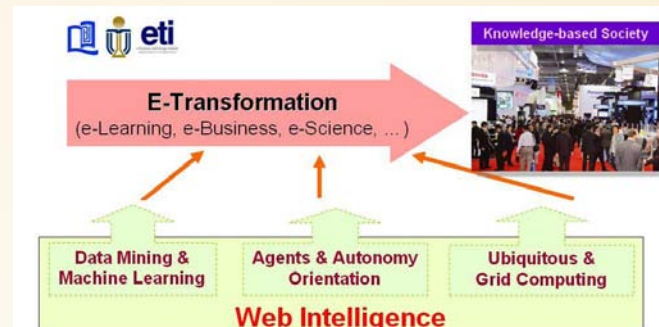
From HKBU, HKUST, and ETI-HKU

Research Focus

Basic Research	e-Transformation Applications
<ul style="list-style-type: none">- Agents and Autonomy Oriented Computing- Data Mining and Machine Learning- Grid and Service Oriented Computing	<ul style="list-style-type: none">- Community Analysis in the Networked Society- Decision Technologies for Supply Chain Management- Services Management for Computation, Data Analysis, and Knowledge Inference

Publications

- Communications of the ACM
- IEEE Internet Computing
- IEEE Transactions on Knowledge and Data Engineering
- Journal of Data Mining and Knowledge Discovery
- SIGKDD Exploration
- Proceedings of prestigious international conferences, such as AAAI, ICDM, ICML, ISCAI, and WIIAT



Research Centre for Ubiquitous Computing

Department of
Computer
Science

Mission

- To excel in some focused areas related to ubiquitous/pervasive computing
- To coordinate and develop research activities towards ubiquitous/pervasive computing
- To construct test-beds based on different positioning technologies for testing & benchmarking and to develop software which becomes the common basic building blocks for ubiquitous/pervasive applications

Research Team

Principal Investigator:

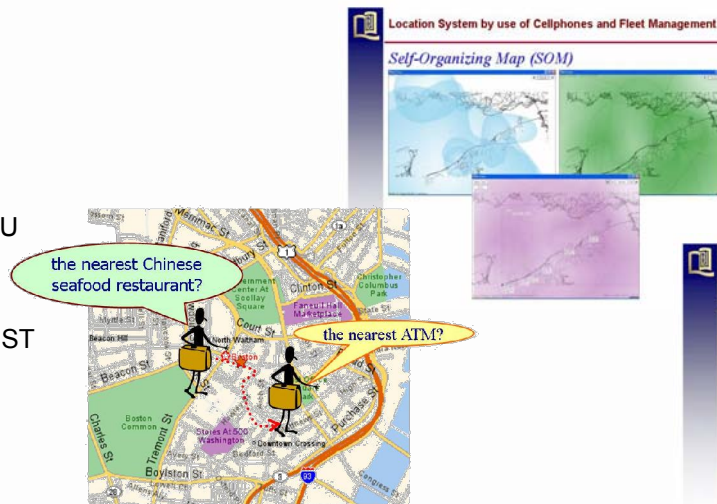
Joseph Kee-Yin Ng, Professor
Computer Science Department, HKBU

Co-Investigator:

Lionel Ni, Head and Professor
Computer Science Department, HKUST

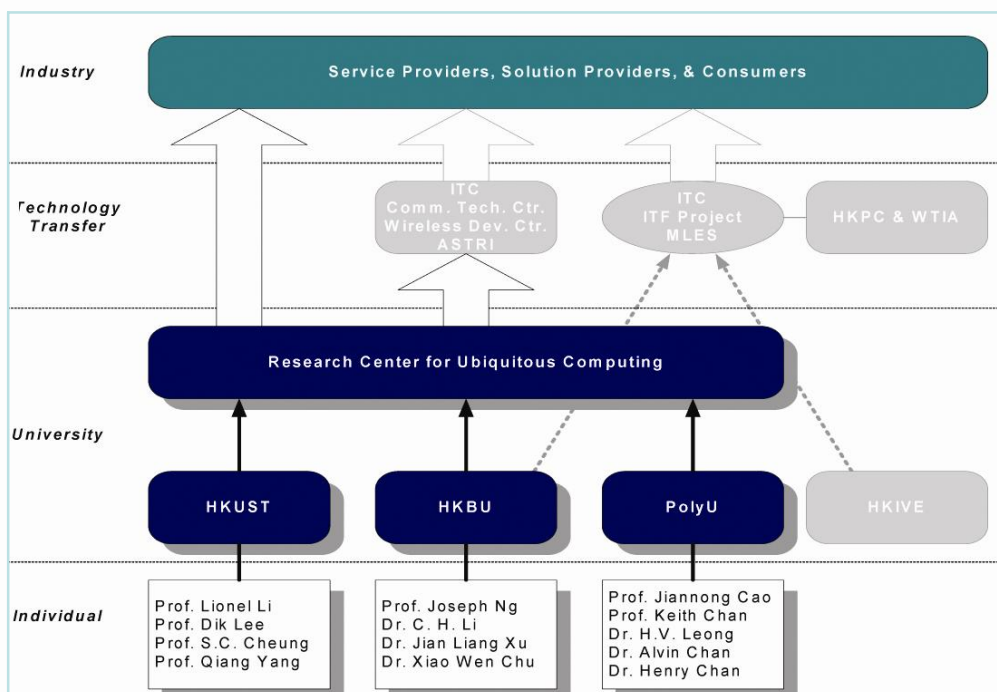
Key Members:

From HKBU, HKUST, and PolyU



Research Focus

- Location-aware computing with a focus on mobile and wireless technology
- RFID and sensor networks with a focus on positioning and data access
- The design and construct of API / Middleware for ubiquitous computing
- Software development on providing location-based services



For more information about the Research Centre for Ubiquitous Computing, please contact Prof. Joseph Ng at jng@comp.hkbu.edu.hk or (852)-3411-7864



Awards



Prof. Liu Jiming

President's Award
for Outstanding Performance 2007
in recognition of his Outstanding Performance
in Scholarly Work



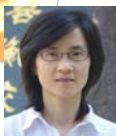
Prof. Tang Yuan-yan

First Class of Natural Science Award
of Technology Development Centre
Ministry of Education of the People's
Republic of China in 2005

Elected Fellow of the IEEE in January 2004
and Fellow of IAPR
in August 2004



President's Award
for Outstanding Performance 2003
in recognition of his Outstanding Performance
in Scholarly Work



Dr. Feng Jian

Most Cited Paper Award
for the Journal of Visual Communication and
Image Representation 2004-2006



Prof. Ng Kee-yin, Joseph

Together with his two German partners, Prof. Wolfgang Halang and
Dr. Thomas Erdner and his graduate student Mr. Stephen Ka Chun
Chan, received two **patents** issued by the German Patent Office for
their research work entitled "*Clock Synchronization in Nodes on a
Ring Bus*", and "*Real Time Capable and Fault Tolerant Data
Transfer*" granted by The Germany Patent Office, Munchen,
9 February 2006 and 22 March 2007, respectively.
Patent Registration Number: 10253533 / 10253534



IEEE Computer Society
Meritorious Service Award 2004

Together with his two students Mr. Kan Ka-ho Kenny,
Mr. Chan Ka-chun Stephen, received
"AINA2003 Excellent Paper Award"



Mr. Tong Tat Ming

under the supervision of Prof. Tang Yuan-yan

First-grade Award
in the Pan-Pearl River Delta Region
Amway Universities IT Project
Competition 2007



**Mr. Chan Ka-ho
Mr. Chan Kai-kin
Mr. Yeung Man-chung**

under the supervision of Prof. Ng Kee-yin, Joseph

2nd Runners-up
in the Microsoft Imagine Cup 2007



Mr. Yeung Man-chung

under the supervision of Prof. Ng Kee-yin, Joseph

Distinction Award
in Intersvarsity Internet Technologies Exposition
and Conference 2005

Certificate of Merit (Post-Secondary)
of the 7th IT Excellence Awards by
Hong Kong Computer Society in 2005



ON KNOWLEDGE GRID AND GRID INTELLIGENCE: A SURVEY

WILLIAM K. CHEUNG AND JIMING LIU

Computer Science Department, Hong Kong Baptist University, Kowloon Tong, Hong Kong

The last generation Web Intelligence (WI) aims at enabling users to go beyond the existing online information search and knowledge queries functionalities and to gain, from the Web, practical wisdom for problem solving. To achieve such a Web intelligence, we need to develop a Web intelligence system that integrates various techniques in Web intelligence, such as Internet, Web, Semantic, and Web, and integrates them with existing knowledge, such as a semantic and social network. In this paper, we provide an overview of recent development in WI and Semantic Knowledge Grid. Then, the development, capabilities of the Wisdom Web as well as the conceptual architecture of an intelligent Grid for supporting it are described. Technical challenges for realizing Grid Intelligence are highlighted and the future researches in related research areas are reviewed.

Key words: Wisdom Web, Grid Intelligence, Knowledge Grid, autonomy-oriented computing.

1. INTRODUCTION

1.1. Web Intelligence and Wisdom Web

The Web has irrevocably revolutionized the world we live in. This impact is inevitable due to the facts that the Web connectivity rapidly increases and that the online information astronomically explodes. In order not only to live with such a change but also to benefit from the information infrastructure that the Web has empowered, we have witnessed the fast development as well as applications of many Web Intelligence (WI) techniques and technologies (Zheng, Liu, and Yao 2003), which cover:

1. *Internet-level communication, infrastructure, and security protocols.* The Web is regarded as a computer-networked system. WI techniques for this level include, for instance, Web data-prefetching systems built upon Web-surfing patterns to resolve the issue of Web latency. The intelligence of the Web prefetching comes from adaptive learning based on observations of user-surfing behavior.
2. *Interface-level multimedia presentation standard.* The Web is regarded as an interface for human-Internet interaction. WI techniques for this level are used to develop the intelligent Web interfaces in which the capabilities of adaptive cross-language processing, representing the semantic contents of the Web available in machine-understandable formats for agent-based computing, such as searching, aggregation, classification, filtering, managing, mining, and discovery on the Web (Berners-Lee, Hendler, and Lassila 2001).
3. *Knowledge-level information processing and management tools.* The Web is regarded as a distributed data/knowledge base. We need to develop semantic mapping languages to represent the semantic contents of the Web available in machine-understandable formats for agent-based computing, such as searching, aggregation, classification, filtering, managing, mining, and discovery on the Web (Berners-Lee, Hendler, and Lassila 2001).
4. *Application-level ubiquitous computing and social intelligence environments.* The Web is regarded as a basis for establishing social networks that contain communities for establishing social networks that contain communities of people (or organizations or other social entities) connected by social relationships, such as friendship, coworking, or information exchange with common interests. They are Web-supported social networks or virtual communities. The study of WI concerns the important issues central to social

Address correspondence to William K. Cheung, Computer Science Department, Hong Kong Baptist University, Kowloon Tong, Hong Kong; e-mail: william@comp.hkbu.edu.hk

Note: the value of "doi" should be 10.1002/ci.20052

© 2005 Blackwell Publishing, 350 Main Street, Malden, MA 02148, USA, and 9600 Garsington Road, Oxford OX4 2DQ, UK.



Service-Oriented Distributed Data Mining

Data mining research currently faces two great challenges: how to embrace data mining services with just-in-time and autonomous properties and how to mine distributed and privacy-protected data. To address these problems, the authors adopt the Business Process Execution Language for Web Services in a service-oriented distributed data mining (DDM) platform to choreograph DDM component services and fulfill global data mining requirements. They also use the learning-from-observation methodology to achieve privacy-preserving DDM. Finally they illustrate how localized autonomy on privacy-policy enforcement plus a bidding process can help the service-oriented system self-organize.

William K. Cheung,
Xiao-Feng Zhang,
Ho-Fai Wong, and Jiming Liu
Hong Kong Baptist UniversityZong-Wai Luo
and Frank Tong
e-Business Technology Institute,
University of Hong Kong

Most data mining algorithms assume that data analysts will aggregate data extracted from production systems at a server for subsequent computational intensive data-crunching processes. However, issues such as data privacy concerns (with respect to customer information stored in bank servers, for example) and limits on data transmission bandwidth (affecting terabyte-sized data generated from remote lab instruments or supercomputers) demonstrate that aggregating data for centralized mining simply isn't possible in a growing number of cases. Instead, it's become necessary to develop methodologies for mining distributed data that must remain private.¹ In addition, being able to get the right information at the right time (with respect to real-time business intelligence, for example) is an important business strategy in today's highly dynamic

market. This real-time objective imposes additional requirements on distributed data mining (DDM), including providing on-demand and self-adaptive services so that companies can cope with heterogeneities in data sources, with respect to data privacy requirements, which aren't always known in advance.

We can address these challenges in two ways: a distributed computing architecture that supports seamless provision, integration, and coordination of just-in-time and autonomous data mining services; and a privacy-conscious DDM methodology can work on top of this architecture.

The ship methodology. In this article, we describe our recent efforts to create a novel DDM methodology known as *learning from observation* on a service-oriented platform in which the underlying processes are specified in the Business Process Execution Language for Web Services (BPEL4WS). We

Correspondence

Extended Latent Class Models for Collaborative Recommendation

Kwok-Wai Cheung, Kwok-Ching Tsui, and Jiming Liu

Abstract—With the advent of the World Wide Web, providing just-in-time, personalized product recommendations to customers now becomes possible. Collaborative recommender systems utilize correlation between customer preference ratings to identify "like-minded" customers and predict their product preference. One factor determining the success of the recommender systems is the prediction accuracy, which is mainly limited by lacking adequate ratings (the sparsity problem). Recently, the use of latent class model (LCM) has been proposed to alleviate this problem. In this paper, we first study how the LCM can be extended to handle customers and products outside the training set. In addition, we propose the use of a pair of LCMs (called *latent class model-DLGM*) instead of a single LCM, to model customer likes and dislikes separately for enhancing the prediction accuracy. Experimental results based on the *BookMovie* dataset show that DLGM outperforms both LCM and the conventional correlation-based method when the available ratings are sparse.

Index Terms—Collaborative filtering, latent class models (LCMs), personalization, recommender systems.

1. INTRODUCTION

Product recommendation is one of the most important business activities for attracting customers. With the advent of the World Wide Web, online companies can now recommend products to their customers on a one-to-one basis in real time, and more importantly, at a much lower cost. Different recommender systems have been proposed in the literature [1]–[2] and related products/services have also been released in the market (e.g., *Andromedia.com*, *Netperception.com*). Based on the underlying technology, recommender systems can be broadly categorized as *content-based* or *collaborative*.

Content-based recommender systems match customer interest profiles (e.g., revealed by their highly rated products) with the product attributes (or features) when making recommendations. Different machine learning [3], [4] and information retrieval [5], [6] algorithms have been proposed for profile representation and ratings prediction. One successful application of the content-based approach is personalized Web pages recommendation (e.g., *Leitica* [7]). In order for the approach to be effective, sufficient rich and accurate product information as well as personal profiles should be available. Besides, the product attributes have to be carefully chosen for the product and profile. Bad choices of features result in recommender systems with either low discriminating power (the *shadowy-analogy* problem) or bias in recommending the customer interest (the *over-specialization* problem) [8]. Collaborative recommender systems are based on the similarity between customer preference ratings for computing recommendations.

Manuscript received February 27, 2003; revised September 27, 2003 and July 10, 2004. This work was jointly supported by Hong Kong Baptist University via Faculty Research Grant FG03-04/05-06 and by RGC Grant HKU02/03-04/05.

The authors are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: william@comp.hkbu.edu.hk; kwokc@comp.hkbu.edu.hk; jimling@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TSAC.2003.818877

1083-4427/04/\$20.00 © 2004 IEEE

As the approach does not rely on product contents, it is free from the two problems of the content-based approach and thus has widely been used for recommending products where product descriptions are either lacking or found to be too specific to be useful. Many different techniques have been proposed for collaborative recommendation, including the most original correlation-based methods [9], [10], latent semantic indexing (LSI) [11], [12], Bayesian learning [13], [14], etc. Successful application domains include recommendation of Usenet articles [9], music [10], etc. In order for collaborative recommendation to be accurate, a large enough number of customers willing to provide preference ratings for the products are required, and the product coverage of their ratings must be significant enough. However, this may not be the case in reality because of either lacking such a large customer pool or new products being encountered (the *sparsity* problem). Applying simple clustering or some statistical cluster models to the preference ratings has been demonstrated to be able to improve the local density of the ratings and is considered to be a promising remedy for the sparsity problem [15], [16].

In this paper, we first describe a statistical cluster model—the latent class model (LCM), originally proposed by Hofmann et al. for collaborative filtering [15], and study how a properly trained LCM can also be used to handle customers and products outside the training set for recommendation. Also, we argue that the LCM is limited in terms of correctly modeling like and dislike ratings and propose a dual latent class model (DLGM) which is trained using two sets of data converted from the original ratings, one with ratings for liked items and another with those for disliked ones. This modification allows the groupings of customers with similar likes and dislikes to be captured separately and (2) improve the overall predictive power of the model (experiments based on the *BookMovie* dataset were conducted for performance evaluation). It was found that DLGM outperforms LCM and a conventional correlation-based method when the ratings are sparse.

II. COLLABORATIVE RECOMMENDER SYSTEMS

The concept of collaborative recommendation (also called the word-of-mouth approach) was first used in Goldberg et al.'s e-mail filtering system [17]. The idea was then quickly picked up for product recommendation. In this section, we further elaborate the sparsity problem and briefly survey some existing methods proposed in the literature for alleviating it.

A. Sparsity Problem

Most of the pioneering collaborative systems use the correlation-based approach for recommendation prediction. For example, in [9], the predicted rating for customer i for product j is computed as

$$r_{ij} = \bar{r}_j + \sum_k w_{ik} (r_{kj} - \bar{r}_j)$$

where r_{ij} denotes the recorded ratings of customer i for product j , \bar{r}_j denotes the expected rating of customer i over all the products, w_{ik} is the Pearson correlation coefficient (P-Corr) between the ratings of customer i and k , given as

$$w_{ik} = \frac{\sum_j (r_{ij} - \bar{r}_j)(r_{kj} - \bar{r}_j)}{\sqrt{\sum_j (r_{ij} - \bar{r}_j)^2} \sqrt{\sum_j (r_{kj} - \bar{r}_j)^2}}$$

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

Grounding Collaborative Learning in Semantics-Based Critiquing

William K. Cheung¹, Anders I. Merch², Kelvin C. Wong¹, Cynthia Lee³, Jiming Liu¹, and Masoumeh A. Lani¹¹ Department of Computer Science, Hong Kong Baptist University, Hong Kong (william, kwkwong, jimling, taeon@comp.hkbu.edu.hk)² InterMedia, University of Oslo, Norway (anders.a.merch@intermedia.uio.no)³ Language Centre, Hong Kong Baptist University, Hong Kong (clic@hkbu.hk)

ABSTRACT

In this paper we investigate the use of Latent Semantic Analysis (LSA), Critiquing Systems, and Knowledge Building to support computer-based teaching of English composition. We have built and tested an English Composition Critiquing System that makes use of LSA to analyze student essays and compute feedback by comparing their essays with teacher's model essays. LSA values are input to a critiquing component to provide a user interface for the students. A software agent can also use the critic feedback to coordinate a collaborative knowledge building session with multiple users (students and teachers). Shared feedback provides shared questions that can trigger discussion and extend reflection about the next phase of writing. We present the first version of a prototype we have built, and report the results from three experiments. We end the paper by describing our plans for future work.

Keywords: Semantic Matching; Web-based Learning; Critiquing Systems; Knowledge Building; Essay Writing.

INTRODUCTION

English is the preferred second language for many people and learning it occurs in many ways. For example, young people are quite apt in learning spoken English phrases when watching TV, browsing the Internet and communicating with peers on mobile phones (e.g., SMS). However, previous studies have shown these influences may have negative effect on vocabulary development (Rice et al 1990; Weizman & Snow 2001). As a consequence, students' reading and writing skills do not keep pace with listening, viewing and speaking. Furthermore, English composition is primarily taught in the classroom and practiced in homework assignments, supported by qualified teachers and parents. These are important but scarce resources, creating an im-

A Novel Orthogonal NMF-Based Belief Compression for POMDPs

Xin Li
William K. W. Cheung
Jiming Liu
Zhili Wu
Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, HK

LIN@COMP.HKBU.EDU.HK

WILLIAM@COMP.HKBU.EDU.HK

JIMING@COMP.HKBU.EDU.HK

VINCENT@COMP.HKBU.EDU.HK

Abstract

High dimensionality of POMDP's belief state space is one major cause that makes the underlying optimal policy computation intractable. Belief compression refers to the methodology that projects the belief state space to a low-dimensional one to alleviate the problem. In this paper, we propose a novel orthogonal non-negative matrix factorization (O-NMF) for the projection. The proposed O-NMF not only factors the belief state space by minimizing the reconstruction error, but also allows the compressed POMDP formulation to be efficiently computed (due to its orthogonality) in a value-directed manner so that the value function will take same values for corresponding belief states in the original and compressed spaces. In this paper, we propose to combine the strengths of the two approaches via a novel orthogonal non-negative matrix factorization (O-NMF). The proposed belief compression approach has a number of advantages, including: (1) O-NMF guarantees all the elements of the low-dimensional belief states to be non-negative, which is important as belief states are by themselves probability distributions; (2) O-NMF explores sparsity in belief spaces; (3) The value-directed property can be realized using Bayesian; (4) The overhead computation needed for getting the compressed POMDP formulation is carefully designed to avoid solving LP problems; and (5) The high-dimensional α -vectors (characterizing the value function) can be represented by a low-dimensional α -vectors.

1. Introduction

Partially Observable Markov Decision Process (POMDP) models how an agent acts in a stochastic environment given partial observations and feedback from the environment for a better average reward in the long run. Due to the partial observability, it is common to represent the belief state of a POMDP as a probability mass function defined over the true states. Upon each action-taking and then observation arrival, the belief state is re-evaluated using Bayesian updating. The complete set of belief states spans

Appearing in Proceedings of the 8th International Conference on Machine Learning, Corvallis, OR, 2007. Copyright 2007 by the author(s) (owner(s)).

1083-4427/04/\$20.00 © 2004 IEEE

1083-4427/04/\$20.00 © 2004 IEEE

Agents, Machine-learning and Data-mining

The Development of Successful On-Line Communities

Karen S.K. Cheung, Fion S.L. Lee, Rachael K.F. Ip, and Christian Wagner

Department of Information Systems
City University of Hong Kong
HONG KONG

Abstract

Virtual communities play an increasingly important role in economic, information, and emotional exchanges. Furthermore, some of them have become so large that their social and economic impact can be considerable. At the same time, not all virtual communities seem to flourish, and some perform better than others. This study seeks to examine the main design criteria underlying successful communities, and to identify principles for the design of successful communities. The article is based on a review of the literature, and the study of several communities. As such, it is exploratory in nature. All of the principles will require further in-depth research to test its validity and generalizability. Among the studies of the interesting observations are the (almost) non-sustainability of hub-type (1:N) communities, and the attempt of many communities to reposition themselves to take advantage of community interaction and the social capital that exists in tightly integrated communities.

1. Introduction

Over the last several years, we have been witnessing the emergence of geographically unbound communities, with vast economic power and impact. For example, by the end of 2001 AOL reached a worldwide membership of over 34 Million subscribers (8 million outside the US, with one million in France and 750,000 in Latin America), seven million more than the year before. With 34 Million subscribers, the AOL "economy" has a larger number of "citizens" than Canada, Morocco, or Peru, and as a country would rank 34th in population size (<http://blnc.census.gov/cgi-bin/ipc/dbrank.pl>).

If these 34 million citizens make behavioral changes, such as switching from the Internet Explorer Browser to Netscape, it has a huge impact on the popularity of such technologies. If they make economic decisions, such as purchasing decisions, they can have a large financial impact. If 34 million subscribers decide to support a particular cause, they present a formidable

International Journal of The Computer, the Internet and Management Vol. 13(1) (January–April, 2005) pp 71–87
71

Virtual Community Informatics: What We Know and What We Need to Know

Fion S. L. Lee
Department of Information Systems
City University of Hong Kong
Tel: (852) 2784-7538
Email: isfion@is.cityu.edu.hk

Douglas Vogel
Department of Information Systems
City University of Hong Kong
Tel: (852) 2784-7560
Email: isdoug@is.cityu.edu.hk

Moez Limayem
Department of Information Systems
City University of Hong Kong
Tel: (852) 2784-8350
Email: ismoez@is.cityu.edu.hk

Abstract

The virtual community has just recently emerged with divergent opinions on the basic understanding of it. This study aims at collecting different definitions and classifications in the virtual community, and offers a working definition. It also addresses research conducted in the field by referring to Information Systems journals. The With the exponential growth of the virtual community, more and more studies have been conducted on how virtual communities affect living standards by providing functions for relationship building and knowledge sharing (Sivolep 1997, Brown 2000, Bieber et al. 2001, Blase 2000, lack of research available on the topic).

With others. The current practice is to build web sites and allow people to register as members who can then share information or feelings virtually. However, it is doubtful whether the tools that support virtual community web sites assist in relationship building and knowledge sharing.

This paper aims at comparing the different definitions and classifications of the virtual community to achieve a more compromising agreement on these basic concepts of the virtual community. Existing research on the virtual community is identified and future research topics are proposed. A survey is conducted on internet tools used in virtual community web sites, and suggestions are provided for how these tools can provide support. It is intended that these guidelines will support the virtual community research.

research categorizes the different stages in virtual community growth to show the transition of research in this area. A survey is also conducted on the extent of the adoption of informatics in virtual community web sites.

1) Introduction

Barnsli 1997). Nevertheless, among these studies, little consensus has been reached on basic concepts such as definitions and classifications of the virtual community. Without such underlying concepts, researchers show various meanings for the same terms used. There is also a The virtual community provides access for engaging in common activities, sharing feelings, or discussing ideas

2) Definition of a Virtual Community

A generally agreed upon definition of a virtual community would be a good starting point. What we need is a working definition of the virtual community, a consensus found in the major stream of literature, a definition that understood by most of people. To achieve this goal, definitions of the virtual community proposed by various authors are compared in Table 1. Similar items found in definitions are then extracted in order to build up a working definition.

Author	Definition
Howard (1992)	social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling to form webs of personal relationships in cyberspace.

P6

JITTA
JOURNAL OF INFORMATION TECHNOLOGY THEORY AND APPLICATION

VIRTUAL COMMUNITY INFORMATICS:
A REVIEW AND RESEARCH AGENDA

FION S. L. LEE, City University of Hong Kong
Department of Information Systems, Kowloon Tong, Tel: (852) 2788-7538, Email: isfion@is.cityu.edu.hk

DOUGLAS VOGEL, City University of Hong Kong
Department of Information Systems, Kowloon Tong, Tel: (852) 2788-7560, Email: isdoug@is.cityu.edu.hk

MOEZ LIMAYEM, City University of Hong Kong
Department of Information Systems, Kowloon, Tel: (852) 2788-8350, Email: ismoez@is.cityu.edu.hk

ABSTRACT

Divergent opinions exist on the basic understanding of the concept, virtual community. This study offers a working definition by examining different definitions, and proposes adoption of virtual community classifications. It also includes a summary of research conducted in the field. The research categorizes the different stages in virtual community growth to show the transition of research in this area. The results illustrate a paucity of technology development studies. We also investigate the extent of the adoption of informatics in these communities using a survey 200 virtual communities. The results indicate that discussion forum is the most popular tool adopted in virtual communities. The integration of the research review and tool adoption survey contributes to the generation of an agenda to direct future virtual community research.

Ken Peffers acted as senior editor for this paper.

Lee, F.S.L., D. Vogel, and M. Limayem, "Virtual Community Informatics: A Review and Research Agenda," *The Journal of Information Technology Theory and Application (JITTA)*, 5:1-2003, 47-61.



Available online at www.sciencedirect.com

ScienceDirect

Journal of Computer and System Sciences *** (2007) ***

JOURNAL of
COMPUTER
AND SYSTEM
SCIENCES

www.elsevier.com/locate/jcss

Topological analysis of AOCD-based agent networks and
experimental results

Hao Lan Zhang*, Clement H.C. Leung, Gitesh K. Raikundalia

School of Computer Science and Mathematics, Victoria University, PO Box 14428, Melbourne City, MC 8001, Australia
Received 15 June 2006; received in revised form 31 October 2006

Abstract

Topological analysis of intelligent agent networks provides crucial information about the structure of agent distribution over a network. Performance analysis of agent network topologies helps multi-agent system developers to understand the impact of topology on system efficiency and effectiveness. Appropriate topology analysis enables the adoption of suitable frameworks for specific multi-agent systems. In this paper, we systematically classify agent network topologies and propose a novel hybrid topology for distributed multi-agent systems. We compare the performance of this topology with two other common agent network topologies—centralized and decentralized topologies—within a new multi-agent framework, called *Agent-based Open Connectivity for DSS (AOCD)*. Three major aspects are studied for estimating topology performance, which include (i) transmission time for a set of requests; (ii) waiting time for processing requests; and (iii) memory consumption for storing agent information. We also conduct a set of AOCD topological experiments to compare the performance of hybrid and centralized agent network topologies and illustrate our experimental results in this paper.

© 2007 Elsevier Inc. All rights reserved.

Keywords: Agent network topology; Agent network performance; Intelligent agents; AOCD

1. Introduction

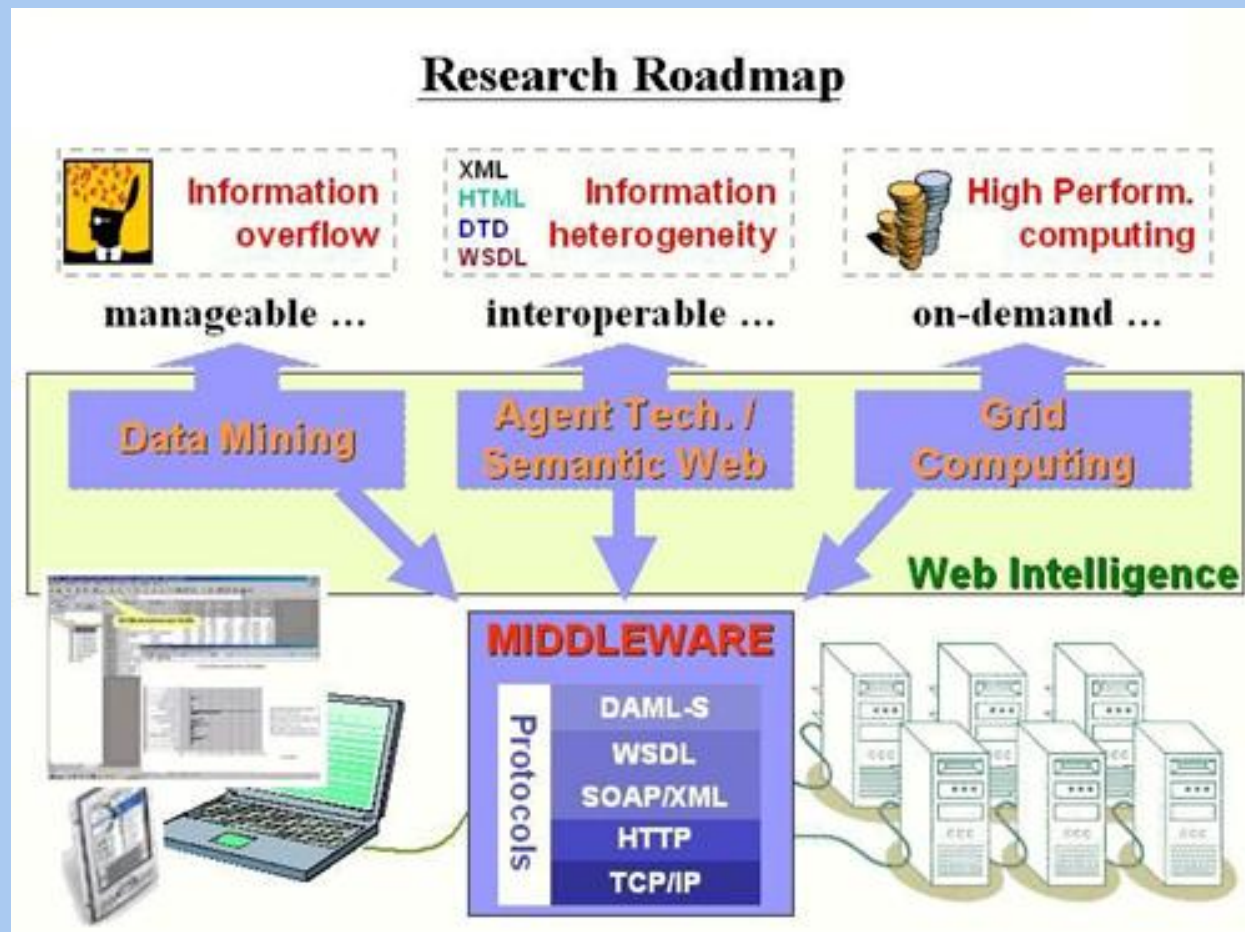
Applications of multi-agent systems have been arising in many areas. Agent-based systems provide a way of conceptualizing complex software applications. These applications often face problems involving multiple and distributed sources of knowledge [1]. However, current research work with respect to topological theory in the area of intelligent agents is inadequate. This situation has led to a set of important research problems concerning how an agent network should be designed to perform efficiently and effectively. The agent network topology issue becomes one of the key issues in solving problems of agent cooperation and communication. An appropriate agent network topology design helps to enhance agent communication efficiency and network mobility.

An agent network topology represents the information of agent distribution over an agent network, which incorporates agent mobility and intelligence aspects into the process of arranging and configuring an agent network. The term,

* Corresponding author.
E-mail addresses: haolan@sci.vu.edu.au (H.L. Zhang), clement.leung@vu.edu.au (C.H.C. Leung), gitesh.raikundalia@vu.edu.au (G.K. Raikundalia).

0022-0000/\$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jcss.2007.04.006

Please cite this article in press as: H.L. Zhang et al., Topological analysis of AOCD-based agent networks and experimental results, J. Comput. System Sci. (2007), doi:10.1016/j.jcss.2007.04.006

Feature Selection with Transductive Support
Vector Machines

ZhiLi Wu¹ and Chunhung Li²

¹ Department of Computer Science, Hong Kong Baptist University
vincent@comp.hkbu.edu.hk

² Department of Computer Science, Hong Kong Baptist University
chli@comp.hkbu.edu.hk

Summary. SVM-related feature selection has shown to be effective, while feature selection with transductive SVMs has been less studied. This paper investigates the use of transductive SVMs for feature selection, based on three SVM-related feature selection methods: filtering scores (SVM wrapper), recursive feature elimination(RFE) and multiplicative updates(MU). We show transductive SVMs can be tailored to feature selection by embracing feature scores for feature filtering, or acting as wrappers and embedded feature selectors. We conduct experiments on the feature selection competition tasks to demonstrate the performance of Transductive SVMs in feature selection and classification.

1 Introduction

SVMs have been studied to work with different divisions of feature selection methods, like wrappers, embedded methods, and filters. For example, the SVM as a wrapper can be used to select features, while the feature set quality is indicated by the performance of the trained SVM (Yu and Chu, 2003). Feature selection can also be conducted during the process of training SVMs. This induces some embedded feature selection methods (Frank et al., 2003; Guyon et al., 2002; Weston et al., 2003, 2000; Guyon et al., 2003; Parkins et al., 2003). SVMs are also useful for filter methods. Although filters often score or rank features without utilizing a learning machine, they can utilize a SVM as the final predictor for classification. Moreover, filters are often integrated as a preprocessing step into wrappers and embedded methods, such that they can help feature selection of wrappers or embedded methods. The development of SVM-related feature selection methods motivates us to explore the role of Transductive SVMs (TSVMs) in feature selection. Transductive SVMs are a type of transductive learning (Vapnik, 1998) machine. They aim to build SVM models upon both labeled and unlabeled data. Transductive SVMs, as generalized from inductive SVMs, inherit the advantages of inductive SVMs such as the large margin (Boser et al., 1992) and regularization formulation as well as kernel mapping. Besides, TSVMs are suitable for the tasks we consider here, which are characterized

Applied Intelligence 22, 37–46, 2005
© 2005 Springer Science + Business Media, Inc. Manufactured in The Netherlands.

Guided Cluster Discovery with Markov Model*

CH. LI

Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong
chli@comp.hkbu.edu.hk

Abstract. Cluster discovery is an essential part of many data mining applications. While cluster discovery process is mainly unsupervised in nature, it can often be aided by a small amount of labeled data. A probabilistic model on the clustering structure is adopted, and a novel unified equation for clustering that incorporates both labeled data and unlabeled data is introduced. This formulation is inspired by a force-field model integrating labeling constraint on labeled data and similarity information on unlabeled data for joint estimation. Experimental results show that good clusters can be identified using small amount of labeled data.

Keywords: clustering semi-supervised learning, Markov model

1. Introduction

In machine learning for classification problems, there are two distinct approaches to learning or analyzing data: the supervised learning and unsupervised learning. The supervised learning deals with problem where a set of data are labeled for training and another set of data would be used for testing. The unsupervised learning deals with problem where none of the labels of the data are available. Unsupervised clustering can be broadly classified into whether the clustering algorithm is hierarchical or non-hierarchical. Hierarchical methods often model the data to be clustered in the form of a tree, or a dendrogram [1]. The lowest level of the tree is usually each datum as a cluster. A dissimilarity measure is defined for merging clusters at a lower level to form a new cluster at a higher level in the tree. The hierarchical methods are often computationally intensive for large number of samples and is difficult to analyze data where the web pages can be represented by two independent representations [7]. The drawback of this co-training approach is that not all data have two independent representations and the algorithm is thus

ple to its cluster center. The k-means algorithm, also known as Forgy's method [2] or MacQueen [3] algorithm is a classical algorithm for non-hierarchical unsupervised clustering. However, the k-means algorithm tends to cluster data into even populations and rare abnormal samples in medical problems can be properly extracted as individual clusters. Recent progress in clustering includes the modeling of proximity structure [4], the dynamic programming approach to hierarchical clustering using graphs [5] and spectral method to clustering [6]. However, these methods do not make use of prior knowledge on dataset such as possible labels or possible structures within the dataset.

In recent years, important data mining tasks have emerged with enormous volume of data. The labeling of a significant portions of the data for training is either infeasible or impossible. Sufficient labeled data for training are often unavailable in data mining, text categorization and web page classification. A number of approaches have been proposed to combine a set of labeled data with unlabeled data for improving the classification rate. The co-training approach has been proposed to solve the problem of web page classification where the web pages can be represented by two independent representations [7]. The drawback of this co-training approach is that not all data have two independent representations and the algorithm is thus

418

A Novel Fractal Image Watermarking

Ming Hong Pi, Chun Hung Li, Member, IEEE, and Hua Li, Member, IEEE

Abstract—A novel watermarking method is proposed to hide a binary watermark into image files compressed by fractal block coding. This watermarking method utilizes a special type of orthogonalization fractal coding method where the fractal affine transform is determined by the range block mean and contrast scaling. Such orthogonalization fractal decoding is a mean-invariant iteration. In contrast, the fractal parameters of classical fractal compression are very sensitive to any change of domain block pool and to common signal and geometric distortion. Hence, it is impossible to directly place a watermark in fractal parameters. The proposed watermark embedding procedure inserts a permuted pseudo-random binary sequence into the quantized range block means. The watermark is detected by computing the correlation coefficient between the original and the extracted watermark. Experimental results show that the proposed fractal watermarking scheme is robust against common signal and geometric distortion such as JPEG compression, low-pass filtering, rotation, and clipping.

Index Terms—Detector response, fractal block coding, orthogonalization fractal transform, watermarking.

1. INTRODUCTION

WITH THE extensive distribution of multimedia data such as text, image, video, and audio, there is a strong demand for ownership management and copyright protection. Digital watermarking is a process of hiding a watermark in multimedia object without perceptual degradation so that the watermark can be detected or extracted later for copyright ownership identification. Early watermarking techniques are directly implemented in the pixel domain. For instance, Schnyder et al. [1], [2] proposed inserting a watermark into the least significant bits (LSB) of an image by bit-plane manipulation of LSBs or adding the watermark and the image. The watermark is detected by computing the correlation coefficient between the original *m-sequence* and the watermarked image. Matsui and Tanaka [3] applied linear predictive coding into watermarking, and hid a watermark to make the watermark resemble quantization noise. However, in general, a digital watermark embedded in the LSBs is highly sensitive to noise and susceptible to be destroyed.

Many watermarking techniques embed a watermark in the transform domain, such as Discrete Cosine Transform (DCT),

Discrete Fourier Transform (DFT), and Discrete Wavelet Transform (DWT). Cox et al. [4] asserted that a watermark should be placed in the perceptually significant components to deter intentional and unintentional attacks. A watermark length of 1000 was added into the 1000 largest coefficients of the DCT (except DC value). Lin et al. [5] proposed a watermarking algorithm robust to rotation, scaling and translation, where the watermark is embedded into a one-dimensional signal, a cumulative function of the Fourier magnitudes along the log-radius axis. Zhu et al. [6] inserted a watermark into all high-pass wavelet coefficients, and multiresolution detection can be allowed because of the pyramid structure. A spectrum of watermarking techniques have been surveyed in [7].

Over the past decade, fractal block coding has mainly been exploited for the purpose of image compression [15]–[17]. Recently, fractal block coding is investigated for watermarking. Based on the fractal codes, several watermarking techniques have been proposed [8]–[11].

Most existing watermarking techniques using fractal codes are developed based on the classification of the fractal codes and a watermark is hidden in a small part of fractal codes [8]–[11]. Davern and Scott [8] divided the domain block pool into two halves, the watermark is hidden in a set of selected range blocks according to which half the best-pair domain block belongs to. Similarly, Pate and Jordan [9] proposed hiding a 32-bits binary signature into the fractal codes with a redundancy V , based on the principle that the original and decoded images possess the same fractal codes. Each range block is encoded by searching the local search region (LSR), which is defined as a square region around the range block. LSR is divided into two sub-regions: A and B, and all range blocks are classified into two categories according to whether the best-pair domain belongs to A or B. Each bit is hidden with V randomly selected range blocks so that the embedded bit can survive even though the fractal codes of some range blocks are altered by the attacks. The analogous fractal watermarking technique was proposed by Li and Wang [10] and Wu and Chang [22], but each bit is hidden by the isometric transforms, instead of the geometric position of the best-pair domain block [8]. The eight isometric transforms are divided into two subgroups: I_0 and I_1 . If a bit is zero, the corresponding range block is encoded using I_0 ; otherwise, the corresponding range block is encoded using I_1 . In addition, Bas et al. [11] proposed a binary watermarking technique based on classification of the fractal codes (or the maps). A part of original maps are replaced with modified ones. As a result, a binary watermark is placed into the image.

Because the contrast scaling and luminance offset depend on the domain block pool, the contrast scaling and luminance offset obtained from the attacked watermarked image are likely different from those obtained from the original image as illustrated in Fig. 1. As a result, the watermark is susceptible to tampering.

888

IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 18, NO. 7, JULY 2008

Agent-Based Load Balancing on
Homogeneous Minigrids: Macroscopic
Modeling and Characterization

Jiming Liu, Senior Member, IEEE, Xiaolong Jin, and Yuanshi Wang

Abstract—In this paper, we present a macroscopic characterization of agent-based load balancing in homogeneous minigrids environments. The agent-based load balancing is regarded as agent distribution from a macroscopic point of view. We study two quantities on minigrids: the number and size of teams where agents (tasks) reside. In macroscopic modeling, the load balancing mechanism is characterized using differential equations. We show that the load balancing we concern always converges to a steady state. Furthermore, we show that load balancing with different initial distributions converges to the same steady state. Finally, we prove that the steady state becomes an even distribution if and only if agents have complete knowledge about agent teams on minigrids. Utility gains and efficiency are introduced to measure the quality of load balancing. Through numerical simulations, we discuss the utility gains and efficiency of load balancing in different cases and give a series of analysis. In order to maximize the utility gain and the efficiency, we theoretically study the optimization of agents' strategies. Finally, in order to validate our proposed agent-based load balancing mechanism, we develop a controlling platform, called Simulation System for Grid Task Distribution (SSGTD). Through experimentation, we note that our experimental results in general confirm our theoretical proofs and numerical simulation results from the proposed equation system. In addition, we find a very interesting phenomenon, that is, agent-based load balancing mechanism is topology independent.

Index Terms—Homogeneous minigrids, load balancing, task distribution, agents, macroscopic modeling, steady states, convergence, grid simulation

1 BACKGROUND

IN order to meet the increasing demand of large-scale scientific computation in the fields of life sciences, physics, and astronomy, the notion of "computational grid" was proposed in mid 1990s [1], [2], [3], [4]. It has been observed that computers (such as PCs, workstations, and clusters) in the Internet are idle. Grid computing aims to integrate idle computational power over the Internet and provide powerful computation capability for users all over the world [1], [2], [3], [5]. Since a grid connects numerous geographically distributed computers, and tasks are submitted to grid nodes in a distributed fashion, an important issue is how to evenly distribute submitted tasks to nodes. This is a load balancing problem, one of the scheduling problems on the grid. By solving this problem, we can optimally utilize computational resources of the grid. In this paper, we will propose an agent-based load balancing mechanism.

1.1 Scheduling on Grids

The scheduling problem on grids has been widely studied [6], [7], [8]. Many schedules for grid computing have been developed [9], such as Applan [10], [11], Nimrod-G [12],

J. Liu and X. Jin are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.
E-mail: liujm@comp.hkbu.edu.hk, jinxl@comp.hkbu.edu.hk.

Y. Wang is with the School of Mathematics and Computational Science, Zhongshan University, Guangzhou, China. E-mail: wywang@zhu.edu.cn.

Manuscript received 28 Feb. 2008; revised 14 June 2008; accepted 30 Sep. 2008; published online 20 May 2009.
For information on obtaining reprints of this article, please send e-mail to: tpds@computer.org and reference IEEECS Log Number TPDS-08-019-003.

Published by the IEEE Press

TOWARD NATURE-INSPIRED COMPUTING

NIC-based systems utilize autonomous entities that self-organize to achieve the goals of systems modeling and problem solving.

by JIMING LIU and K.C. TSUI

Nature-inspired computing (NIC) is an emerging computing paradigm that draws on the principles of self-organization and complex systems. Here, we examine NIC from two perspectives. First, as a way to help explain, model, and characterize the underlying mechanisms of complex real-world systems by formulating computing models and testing hypotheses through controlled experimentation. The end product is a potentially deep understanding or at least a better explanation of the working mechanisms of the modeled systems. And second, as a way to reproduce autonomous (such as lifelike) behavior in solving computing problems. With detailed knowledge of the underlying mechanism(s), simplified abstracted autonomous lifelike behavior can be used as a model in practically any general-purpose problem-solving strategy or technique.

Neither objective is achievable without formulating a model of the factors underlying the system. The modeling process can begin with a theoretical analysis from either a macroscopic or microscopic view of the system. Alternatively, the application developer may adopt a blackbox or whitebox approach. Blackbox approaches (such as Markov models and artificial neural networks) normally do not reveal much about their working mechanisms. On the other hand, whitebox approaches (such as agents with bounded rationality) are more useful for explaining behavior.

The essence of NIC formulation involves conceiving a computing system opened by population(s) of autonomous entities. The rest of the system is referred to as the environment. An autonomous entity consists of a detector (or set of detectors), an effector (or set of effectors), and a repository of local behavior rules (see Figure 1) [3, 8].

A detector receives information related to its neighborhood and to the environment. For example, in a simulation of a flock of birds, this information would include the speed and direction the birds are heading and the distance between the birds in question. The



Characterizing Web Usage Regularities with Information Foraging Agents

Jiming Liu, Senior Member, IEEE, Shiwu Zhang, and Jie Yang

Abstract—Researchers have recently discovered several interesting, self-organized regularities from the World Wide Web, ranging from the structure and growth of the Web to the access patterns in Web surfing. What remains to be a great challenge in Web log mining is how to realize user behavior understanding observed Web usage regularities. In this paper, we will address the issue of how to characterize the strong regularities in Web surfing in terms of user navigation strategies, and present an information foraging agent-based approach to describing user behavior. By experimenting with the agent-based decision model of Web surfing, we aim to explain how some Web design factors as well as user cognitive factors may affect the overall behavioral patterns in Web usage.

Index Terms—Web log, Web mining, power law, regularities, user behavior, decision models, information foraging, autonomous agents, agent-based simulation.

1 INTRODUCTION

This contents and services on the World Wide Web (or the Web) have been growing at a very rapid rate. Until now, there may have existed over one billion Web sites on the Web at anytime, if projected based on the studies reported in [1], [2]. Viewing the Web as a large directed graph of nodes (i.e., Web pages) connected with links (i.e., hyperlinks), Huberman et al. [3] proposed a random-walk model to simulate certain regularities in user navigation behavior and suggested that the probability distribution of surfing depth (step) follows a two-parameter inverse Gaussian distribution. They conjectured that the probability of finding a group starting at a given level scales inversely in proportion to its depth, i.e., $P(d) \propto 1/d^2$.

In order to further characterize user navigation regularities as well as to understand the effects of user interests, motivation, and content organization on the user behavior, in this paper we will present an information foraging agent-based model that takes into account the interest profiles, motivation aggregation, and content selection strategies of users and, therefore, predicts the emerged regularities in user navigation behavior.

1.1 Organization of the Paper

The remainder of this paper is organized as follows: In Section 2, we will provide a survey of the existing work in Web mining with a special focus on studies that deal with the regularities on the Web. This is followed by Section 3 which states the problems as well as important issues to be dealt

with in our present study. Section 4 presents the detailed formulation of our proposed information foraging agent model. Section 5 shows several experimental results on characterizing Web usage regularities. Section 6 discusses the effects on the emerged regularities under different conditions in our model. Finally, Section 7 concludes the paper by summarizing the key contributions and findings of this study.

2 RELATED WORK

This section provides an overview of research work related to Web mining. Generally speaking, Web mining is aimed to study the issues of 1) where and how information can be efficiently found on the Web and 2) how and why users behave in various situations when dynamically accessing and using the information on the Web.

2.1 Web Mining for Pattern-Oriented Adaptation

The first major task in Web mining may be called Web mining for pattern-oriented adaptation; that is, to identify the interrelationships among different Web sites, either based on the analysis of the contents in Web pages or based on the discovery of the access patterns from Web log files. By understanding such interrelationships, we aim to develop adaptive Web search tools that help facilitate or personalize Web surfing operations.

This task is certainly justified as studies have shown that 85 percent of users use search engines to locate information [4]. Even though good search engines normally index only about 16 percent of the entire Web [2], an adaptive utility can still be useful to filter or rank thousands of Web pages that are often returned by search engines. For instance, some researchers have developed efficient search techniques that detect authorities, i.e., pages that offer the best resources of the information on a certain topic and topic, i.e., pages that are collections of links to authorities [5], [6]. When it is difficult to directly find relevant information from search engines, navigating from one page to another

J. Liu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong.
E-mail: jliu@hkbu.edu.hk.
S. Zhang and J. Yang are with the Department of Precision Machinery and Instrumentation, University of Science and Technology of China, 230026, Hefei, P.R. China.
E-mail: shiwu@ustc.edu.cn, jyang@ustc.edu.cn.
Manuscript received 18 Dec. 2005; revised 21 Oct. 2006; accepted 17 Mar. 2006.
For information on obtaining reprints of this article, please send e-mail to: tdm@computer.org, and reference IEEECS Log Number 155851.

1043-4446/06/05 0204 IEEE

Published by the IEEE Computer Society

Equilibria, Prudent Compromises, and the “Waiting” Game

Kwang Mong Sim

Abstract—While evaluation of many e-negotiation agents are carried out through empirical studies, this work supplements that by analyzing the performance of agents in terms of designing market-driven agents (MDAs) in terms of equilibrium points and stable strategies. MDAs are negotiation agents designed to make prudent compromises taking into account factors such as time preference, outside option, and rivalry. This work shows that 1) in a given market situation, an MDA negotiates optimally because it makes minimally sufficient concessions, and 2) by modeling negotiation of MDAs as a game of incomplete information, it is shown that the strategies adopted by MDAs are stable. In a bilateral negotiation, it is proven that the strategy pair of two MDAs forms a sequential equilibrium for Γ . In a multilateral negotiation, it is shown that the strategy profile of MDAs forms a market equilibrium for Γ .

Index Terms—Automated negotiation, negotiation agent, sequential equilibrium.

I. INTRODUCTION

AUTOMATED negotiation [1] among software agents is becoming increasingly important because automated interactions between agents can in many different contexts, and research on engineering e-negotiation agents [2]–[4] has received a great deal of attention in recent years (see [2], [3] for a survey). Whereas a substantial portion of the existing work evaluates e-negotiation by experimentation and stochastic simulation, this work supplements and complements the literature by providing a game-theoretic analysis of market-driven agents (MDAs) [10]–[15]. MDAs are e-negotiation agents that make minimally sufficient concessions (1), taking into account time preference, deadline, outside option, differences in proposals, and market rivalry (see Section II-B). The impetus of this work is analyzing MDAs in terms of sequential equilibrium [16], [17], market equilibrium [17], [18], and stable (and dominant) strategies. While [19]–[21] have proven sequential equilibrium on the strategies of their agents in bilateral negotiations (see Section VII), this work shows that the strategies of MDAs are in sequential equilibrium (see Section V-A) and market equilibrium (see Section V-B) for bilateral and multilateral negotiations, respectively. Proving market equilibrium and considering the influence of market conditions are essential in existing applications such as e-commerce because when traders enter/leave an e-market, the conditions for deliberation change

as new opportunities/threats are constantly being introduced. For future applications such as grid computing, it is envisioned that agents negotiating for resources in future computational grids [22] must take into consideration the dynamics of a grid-computing environment because it is expected that resources and services are constantly being added/removed from a grid [23]. Whereas empirical results derived from stochastic simulations discussed in the author's previous work [12], [14], [15] verified some of the desirable properties (e.g., high average utility and relatively high success rate of reaching consensus) of (enhanced) MDAs, this work considers and further supports previous contributions by showing that 1) for a given market situation, an MDA negotiates optimally by making minimally sufficient concessions (see Section IV), and 2) in a bilateral as well as multilateral negotiations, the strategies of MDAs are stable (see Sections V-A and B); this rests on the fact that there is a dominant strategy (see Section III) for an MDA in both bilateral and multilateral negotiations (see also Section IV). Results from this work show that 1) MDAs are designed to optimize their utilities in a given market situation and trading constraints, and 2) since the strategies are stable, MDAs are motivated to behave in a desired manner, even though it is assumed that MDAs do not have complete information about their opponents (see Section II-A). In addition to being able to optimize utility, stability is also an essential criterion for evaluating the design of negotiation mechanisms [18]. Additionally, Section IV discusses how the “sit-and-wait” strategy [19] can be modeled in MDAs.

II. NEGOTIATION MECHANISM

This section presents the negotiation protocol (see Section II-A) and the market-driven negotiation strategy (see Section VII). A market-driven agent is an e-negotiation agent with the distinguishing feature that determines the amount of concession using the time, opportunity, and competition decision. Section II-C discusses the collection of strategies for a range of negotiation situations.

A. Negotiation Protocol

To set the stage for specifying the negotiation protocol and the market-driven strategy in Section II-B, some assumptions are given as follows:

- 1) Agents do not have information about the deadline, reserve price, strategy, and time preference of other agents.

1043-4446/06/05 0204 IEEE

Continuous-Time Negotiation Mechanism for Software Agents

Bo An, Kwang Mong Sim, Liang Gui Tang, Shuang Qing Li, and Dai Jie Cheng

Abstract—While there are several existing mechanisms and systems addressing the crucial and difficult issues of automated one-to-many negotiation, this paper develops a flexible one-to-many negotiation mechanism for software agents. Unlike the existing general one-to-many negotiation mechanism, in which an agent should wait until it has received proposals from all its trading partners before generating counterproposals, in the flexible one-to-many negotiation mechanism, an agent can make a proposal in a flexible way during negotiation, i.e., negotiation is conducted in continuous time. To decide when to make a proposal, two strategies based on fixed waiting time and a fixed waiting ratio are proposed. Results from a series of experiments suggest that, guided by the two strategies for deciding when to make a proposal, the flexible negotiation mechanism achieved more favorable trading outcomes as compared with the general one-to-many negotiation mechanism. To determine the amount of concession, negotiation agents are guided by four mathematical functions based on factors such as time, trading partners' strategies, negotiation situations of other threads, and competition. Experimental results show that agents guided by the four functions react to changing market situations by making prudent and appropriate rates of concession and achieve generally favorable negotiation outcomes.

Index Terms—Automated negotiation, negotiation agents, one-to-many negotiation.

I. INTRODUCTION

AUTOMATED negotiation [19], [21] among software agents is becoming increasingly important because automated interactions between agents [4], [3], [29] can occur in many different contexts (e.g., negotiation for resources [9]). In terms of the number of agents participating in negotiations, agent-based automated negotiation can be divided into three cases [6], namely: 1) one-to-one negotiation (bilateral negotiation); 2) many-to-many negotiation; and 3) one-to-many negotiation. Compared with auction mechanisms [18], one-to-many interactive negotiation is more flexible. For example, agents can adopt different negotiation strategies with different trading partners (alternatives), and negotiations can be taken under different negotiation environments and protocols.

In one-to-many negotiation (take the negotiation between a buyer and several sellers as an example), there are two alternatives: 1) buyer negotiates sequentially with all the sellers and 2) buyer negotiates concurrently with these sellers. Generally,

the buyer gets more desirable negotiation outcomes when it negotiates concurrently with all the sellers in competitive situations in which there are information uncertainty and deadlines [16], [17]. In this paper, we assume that an agent negotiates concurrently with its trading partners.

Let a negotiation cycle be the time spent in a round of negotiations and the reaction time of a trading partner be the time from an agent's proposing to its receiving a counterproposal from the trading partner. In existing general one-to-many negotiation mechanisms and systems (e.g., [1], [7], [20], and [30]), taking the negotiation between a buyer and several sellers as an example, the buyer's negotiation with the set of sellers is divided into several rounds (indexed by $(0, 1, 2, \dots)$), i.e., negotiation is conducted in discrete time. A problem with the general one-to-many negotiation mechanism [7], [12], [16], [20] is that during negotiation, no matter how long an agent has to wait and how many proposals have been received, the agent cannot propose until it has received proposals from all its trading partners. In actual negotiation environments, as agents may have different negotiation strategies, reasoning mechanisms, communication time, constraints, and preferences, the agents generally receives its trading partners' proposals at different times in each round after it sent proposals to all its trading partners at the same time (i.e., different trading partners have different reaction times).¹ Therefore, the general one-to-many negotiation mechanism is not flexible enough when negotiation agents are of different reaction times.

To overcome the limitation of the general one-to-many negotiation mechanism, this research focuses on developing a more flexible mechanism than the general one, where an agent can decide when to make a proposal according to the synchronization situations of negotiation, which are determined by the reaction time of each trading partner.

There are three critical issues in designing a flexible one-to-many negotiation mechanism.

- 1) How to coordinate all the subnegotiation threads (Section II): One-to-many negotiation can be treated as a series of subnegotiation threads, and different subnegotiation threads have different negotiation situations. A coordination strategy concerns issues such as whether all the subnegotiation threads are interactive or not and

Manuscript received July 11, 2005; revised December 10, 2005 and January 7, 2006. This paper was recommended by Associate Editor G. Skarmas. B. An, L. G. Tang, S. Q. Li, and D. J. Cheng are with the College of Computer Science, Chongqing University, Chongqing 400044, China (e-mail: boan@cqu.edu.cn).

K. M. Sim is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: kmsim@hkbu.edu.hk).

Digital Object Identifier 10.1109/TSMC.2006.1746851

¹The reason that bring about different trading partners with different reaction times vary. For example, agents located in different positions in the network may have different communication distances. The factors affecting communication quality of service (QoS), e.g., bandwidth, congestion, and network failure, may result in different communication delays. Agents with different negotiation mechanisms and preferences may have different processing speeds. According to strategies, agents may decide to propose sending proposals, even though the proposals have already been generated successfully.

1043-4446/06/05 0206 IEEE

Grid Commerce, Market-Driven G-Negotiation, and Grid Resource Management

Kwang Mong Sim

Abstract—Although the management of resources is essential for realizing a computational grid, providing an efficient resource allocation mechanism is a complex undertaking. Since Grid providers and consumers may have independent goals, negotiation among them is necessary. The contribution of this paper is showing that market-driven agents (MDAs) are appropriate tools for Grid resource negotiation. MDAs are e-negotiation agents designed with the flexibility of: 1) making adjustable amounts of concession taking into account market rivalry, outside options, and time preferences; and 2) relating bargaining terms in the face of intense pressure. A heterogeneous testbed consisting of several types of e-negotiation agents to simulate a Grid computing environment was developed. It compares the performance of MDAs against other e-negotiation agents (e.g., Kasbah) in a Grid-commerce environment. Empirical results show that MDAs generally achieve: 1) higher budget efficiency in many market situations than other e-negotiation agents in the testbed and 2) higher success rates in acquiring Grid resources under high Grid loads.

Index Terms—Grid commerce, Grid resource allocation, negotiation, resource management, software agent.

I. INTRODUCTION

GRID COMPUTING is distinguished from conventional distributed computing because it focuses on large-scale resource sharing [1, p. 200]. Hence, a resource management system is central to the operation of a Grid [2, p. 133]. A Grid is a very large-scale network computing system that can potentially scale to Internet size, and the network computing system can be viewed as a virtual computer consisting of a networked set of heterogeneous machines (powered by multiple organizations) that agree to share their local resources with each other [2, p. 135]. Due to its scale, and because resource owners and consumers may have different goals, preferences, interests, and policies, providing an efficient resource management and coordination mechanism in the Grid is a complex undertaking. Hence, automatic scheduling programs are needed to replicate computing resources because of both the complexity of the resource allocation problem and the dynamically changing performance of the Grid resources [3, p. 747]. Agents (or autonomous problem solvers) that can act flexibly in dynamic environments can provide supportive efforts for

a computational Grid [4, p. 8]. Sim [5] argued that software agents, in particular e-negotiation agents, can play an essential role in realizing the Grid vision. Adopting a market-driven approach [6], this work attempts to address some of the issues raised in [3] by designing and building negotiation agents that participate in Grid-commerce (G-commerce) [3], [7], [8] in a market-oriented Grid [9]–[11].

1) G-commerce and Market-Oriented Grid: In [3], [7], and [8], Wolski coined the term G-commerce to refer to computational economies for controlling the resource allocation in computational Grid environments. The Grid can be viewed as a network of computations [12], and computations can be viewed in economic terms [13, p. 133]. It was noted in [7] and [8] that existing Grid resource allocation mechanisms are not flexible enough to address the issues raised in [3].

First, the utilization of Grid resources is not free [3]. In a market-oriented Grid, providers can receive royalties for the (computing and storage) resources and services they provide, whereas Grid users can attempt to mold the Grid systems to their needs by exercising their market powers as Grid consumers. In a Grid economy [14, p. 699], resource management systems should provide the tools and mechanisms for both providers and consumers to express their requirements and facilitate the realization of their goals. A Grid economy not only helps regulate the supply and demand for Grid resources, but also provides the incentives for providers to contribute resources and benefit from doing so and offers an efficient mechanism for managing resources [14, p. 699].

Second, there is an enormous literature on economic theories and principles for explicating and understanding the emergent behavior of the Grid and its constituents (participants). It was also noted in [13, p. 134] that using market models as an economic organization for computation is effective in promoting efficient and cooperative interactions among entities with different goals and knowledge.

Third, it was pointed out in [3, p. 748] that many economic systems and some of their assumptions seem to be familiar (e.g., many people can associate price to the supply-and-demand patterns of resources). Moreover, these economic principles also extend to artificial decision-making agents in general [13, p. 134], including software agents and entities.

Finally, it was noted in [14, p. 699] that some of the economic models for resource allocation include: commodity market models, auction models, tendering or contract-net models, and bargaining or negotiation models.

2) Market-Driven G-Negotiation: Whereas [3], [7], and [8] focused on both commodity markets and auction formalizations

A Pilot Study on the Impact of the Web-based Essay Critiquing System on Writing at the Tertiary Level

Kevin C.K. Wong*, Fiona S.L. Lee*, Cynthia F.K. Lee*, William K.W. Cheung*, Anders I. Mørch*, and Jiming Liu*

*Department of Computer Science, Hong Kong Baptist University, Hong Kong, Language Centre, Hong Kong Baptist University, Hong Kong, *InterMedia, University of Oslo, Norway.
*Corresponding Author: kckwong@comp.hkbu.edu.hk

ABSTRACT

As English is widely used worldwide, it is the preferred second language in Hong Kong. Many students find essay writing stressful because they do not have sufficient ideas to fully cover the topic of the essay. To alleviate learning barriers while writing, a web-based essay critiquing system was developed using Latent Semantic Analysis (LSA). LSA is an automatic text analysis technique for providing just-in-time feedback to students. The feedback takes two forms: new sub-themes suggested to be included, and the visualization of the existing sub-themes' organization. In this paper, we present our findings on students' performance and their perception of the usefulness of this system.

Keywords: Essay Writing, Latent Semantic Analysis, Critiquing System

1. INTRODUCTION

Acquiring good essay writing skills is acknowledged as important, and yet challenging for students in Hong Kong. Many of them find the essay-writing task stressful because they are short of ideas to fully discuss the essay topic and complete the essay. If it is an in-class writing exercise, it will be difficult for the teacher to give immediate context-specific hints to each individual student. If it is a take-home writing exercise, getting immediate feedback from the teacher is simply impossible. Students usually make multiple drafts before finalizing their essays. Even if the teacher can afford to provide feedback to each draft, the turn-around time will often be in days at least. This makes the learning process quite ineffective.

To alleviate the aforementioned learning barriers, a computer-supported critiquing system was developed using Latent Semantic Analysis (LSA) (Lafayette and Laham, 1998) is an automatic text analysis technique, for providing just-in-time feedback to students. The feedback takes two forms:

- new sub-themes suggested to be included, and
- the visualization of the existing sub-themes' organization.

Design of Node Configuration for All-Optical Multi-Fiber Networks

Yiu-Wing Leung, Senior Member, IEEE, Gaotai Xiao, Member, IEEE, and Kwok-Wah Hung

Abstract—It is cost-effective to install multiple fibers in each link of an all-optical network, because the cost of fibers is relatively low compared with the installation cost. The resulting network can provide a large capacity for good quality of service, future growth, and fault tolerance. If a node has more incoming/outgoing fibers, it requires larger optical switches. Using the current photonic technology, it is difficult to realize large optical switches. Even if they can be realized, they are expensive. To overcome this problem, we design a node configuration for all-optical networks. We exploit the flexibility that, to establish a lightpath across a node, can select any one of the available channels in the incoming link and any one of the available channels in the outgoing link. As a result, the proposed node configuration requires significantly smaller optical switches while it can result in nearly the same blocking probability as the existing one. We demonstrate that a good network design is to adopt the proposed node configuration and slightly more fibers in each link, so that the network requires small optical switches while it has a small blocking probability.

Index Terms—All-optical networks, blocking probability, node configuration, optical switches.

I. INTRODUCTION

WAVELENGTH-DIVISION MULTIPLEXING (WDM) can effectively exploit the enormous bandwidth of an optical fiber [1]. Consequently, WDM networks can support high data rates and provide large network capacity. If a WDM network delivers information in the optical domain within the network, it is known as an *all-optical network* [2]–[16]. It can avoid many overheads within the network, such as O/E and E/O conversion, processing and buffering.

Before all-optical networks can be used for real-world applications, many issues have to be solved. In the literature, there are studies on routing and wavelength assignment [2]–[7], analysis of blocking probability [8]–[10], wavelength conversion [3], [8]–[10], analysis of wavelength converters [11], distributed network control [12], design of logical topology [13], and design of physical topology [14].

Recently, there is a growing interest to study all-optical networks with multiple fibers per link [15], [16]. It is cost-effective to install multiple fibers in every link, because the cost of fibers is relatively low compared with the cost of installing fibers in every link.

Paper approved by W. C. Kwong, the Editor for Optical Communications of the IEEE Communications Society. Manuscript received September 12, 2000; revised April 26, 2001.

Y. W. Leung is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: jwg@comp.hkbu.edu.hk).

G. Xiao and K. W. Hung are with the School of Professional Education and Executive Development, Hong Kong Polytechnic University, Hung Hom, Hong Kong.

Publisher Item Identifier S 0090-6778/02/00016-0.

0090-6778/02/00016-0 © 2002 IEEE

design and/or undersea. With multiple fibers per link, the network can provide a large capacity for good quality of service, future growth, and fault tolerance. If a node has more incoming/outgoing fibers, it requires larger optical switches. If a node has L incoming/outgoing links, each link has F fibers and each fiber has C channels at wavelengths $\lambda_1, \lambda_2, \dots, \lambda_C$, then this node has $N = LF$ incoming/outgoing fibers. Using the existing node configuration, the node requires an $N \times N$ optical switch for each wavelength. Fig. 1 shows an example. When each link has more fibers, each node has more incoming/outgoing fibers and hence it requires larger optical switches. For example, if $L = 5$ and $F = 100$, a node requires 500×500 optical switches; if F is increased to 1000, the node requires 5000×5000 optical switches. This results in the following two drawbacks.

- 1) It is difficult to realize large optical switches because of various technological constraints (e.g., the constraints imposed by insertion loss and control complexity). At the time of revising this paper, the largest all-optical switch in shipment is 1024×1024. In this case, if a node has more than 1024 incoming/outgoing fibers, the existing node configuration is not applicable.
- 2) Large optical switches are much more expensive than small optical switches. For example, an optical switch is composed of a cross-connect and other parts for various functions such as I/O and control, and the prices of a 8×8 cross-connect and a 1024×1024 cross-connect are about two thousand and four million, respectively.

In this paper, we design a node configuration that requires small optical switches even when the node has many incoming/outgoing fibers. We demonstrate that the proposed node configuration has several advantages.

- Compared with the existing node configuration, the proposed design requires significantly smaller optical switches while it can result in nearly the same blocking probability as the existing one. To design an all-optical network or upgrade an existing one with spare fibers, we can adopt the proposed node configuration and slightly more fibers in each link so that the network requires small optical switches while it has a small blocking probability.
- Using the existing node configuration, the nodes with different number of incoming/outgoing fibers require optical switches of different sizes. Using the proposed node configuration, all the nodes can use optical switches of the same size. This can simplify network implementation and management.

Some Results on the Self-Similarity Property in Communication Networks

Shubin Song, Joseph Kee-Yin Ng, Senior Member, IEEE, and Bihai Tang

Abstract—Due to the strong experimental evidence that packet network traffic is self-similar in nature, it is important to study the problems to see whether the superposition of self-similar processes retains the property of self-similarity, and whether the service of a server changes the self-similarity property of the input traffic. In this letter, we first discuss some definitions and superposition properties of self-similar processes. We obtain some good results about the property of self-similarity of the output process. Then we present a model of a single server with infinite buffer and prove that when the queue length has finite second-order moment, the input process, being strong asymptotically second-order self-similar (sa-s), is equivalent to the output process which also bears the sa-s property.

Index Terms—Long-range dependent, packet networks, self-similar, short-range dependent.

I. INTRODUCTION

SEVERAL empirical studies on the local-area network (LAN), the variable bit rate (VBR) video traffic, the Integrated Services Digital Network (ISDN), and other communication systems indicate that this traffic is self-similar in nature. For instance, Leland et al. [7] have demonstrated the self-similar nature of Ethernet traffic by a statistical analysis of the Ethernet traffic measurements at Bell-Core; Beran et al. [2], [3], [5] have demonstrated long-range dependence in samples of VBR video traffic generated by a number of different codecs; and Paxson and Floyd [11] have concluded the presence of long-range dependence in TELNET and other wide-area network traffic.

In the light of this strong experimental evidence, it is important to examine in more detail the possible implications that self-similar traffic may have on the design and performance of network systems. For example, real-time communications require the network to provide end-to-end delay guarantee. In order to analyze the delay of networks with self-similar traffic, we need to know the property of queueing systems with self-similar input traffic. In particular, there are two important questions we need to study (as shown in Fig. 1): whether the superposition of self-similar processes retains the self-similarity properties.

Paper approved by M. Hamdi, the Editor for Network Architecture of the IEEE Communications Society. Manuscript received May 2, 2003; revised December 3, 2003. This work was supported in part by the Research Grants Council under Research Grants HKR02/00-01/00 and by the Faculty Research Grant program under FRG/97-98/HK-76 and FRG/98-99/HK-68.

S. Song is with the Department of Risk Management and Insurance, Lingnan College, Chinese University of Hong Kong, Shatin, New Territories, Hong Kong (e-mail: ling@comp.hkbu.edu.hk).

J. K.-Y. Ng is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: jkg@comp.hkbu.edu.hk).

B. Tang is with the Department of Quantitative Economics, Guangzhou University of Economics, Guangzhou 510600, China (e-mail: bhtang@guet.edu.cn).

Digital Object Identifier 10.1109/TCOMM.2004.833136.

0090-6778/04/10020-00 © 2004 IEEE

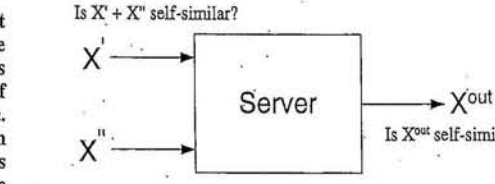


Fig. 1. Self-similarity property.

and whether a server mechanism will change the self-similarity nature of the traffic.

In [13] and [14], Tsybakov and Georgakakos point out that the superposition of two uncorrelated self-similar processes retain some asymptotic self-similarity property. Varnakos and Anantharam [15] consider a special case of a leaky bucket system with long-range-dependent input traffic, and prove that the output (departure) process is also long-range dependent.

In our previous work [12], we take a first look at the traffic characteristics of the output process. In this letter, we carry on with our previous work and focus on the superposition of self-similar processes, and further explore the property of the output process from a server with self-similar input. The rest of the letter is organized as follows. Section II discusses the concepts of self-similar processes and presents the self-similar definitions and their relationships. In Section III, we present the superposition property of self-similar processes. We obtain the result that the superposition of two or more uncorrelated strong asymptotically self-similar processes (or long-range-dependent processes) is strong asymptotically self-similar (or long-range-dependent). Since traffic arrival to a switch is multiplexed from many connections, this superposition property is very important for the analysis of queueing systems. In Section III, we also discuss the superposition of two correlated self-similar processes, and the superposition of a short-range-dependent process with a self-similar process.

Section IV considers a model of a single server with infinite buffer, and proves that when the second-order moment of a queue-length process is finite, the strong asymptotically second-order self-similar (sa-s) properties of the input process and that of the output process are equivalent, which means that the self-similarity will neither be removed nor added by any server mechanism with a finite second-order moment of queue length. And finally, we conclude our paper in Section V.

II. DEFINITIONS OF SELF-SIMILAR

In this section, we present some definitions of self-similar processes which are based on a second-order-stationary random stochastic process.

We begin with the introduction of $X = (X_1, X_2, \dots)$, a semi-infinite segment of a second-order-stationary real-number stochastic process of discrete argument (time) $t \in \mathbb{N} \triangleq \{1, 2, \dots\}$.

Assignment of Movies to Heterogeneous Video Servers

Yiu-Wing Leung, Senior Member, IEEE, and Ricky Yuen-Tan Hou, Member, IEEE

Abstract—A video-on-demand (VOD) system provides an electronic video rental service to geographically distributed users. It can adopt multiple servers to serve many users concurrently. As a VOD system is being used and evolved, its servers probably become heterogeneous. For example, if a new server is added to expand the VOD system or replace a failed server, the new server may be faster with a larger storage size. This paper investigates how to assign movies to heterogeneous servers in order to minimize the blocking probability. It is proven that this assignment problem is NP-hard, and a lower bound is derived on the minimal blocking probability. The following approach is proposed for assignment: 1) *problem relaxation*—a relaxed assignment problem is formulated and solved to determine the ideal load that each server should handle, and 2) *good scalability*—an assignment and reassignment are performed iteratively while fulfilling all the constraints so that the load handled by each server is close to the ideal one. This approach is generic and applicable to many assignment problems. This approach is adopted to design two specific algorithms for movie assignment with and without replication. It is demonstrated that these algorithms can find optimal or close-to-optimal assignments.

Index Terms—Assignment, server system, video-on-demand (VOD).

I. INTRODUCTION

A. Background

A video-on-demand (VOD) system provides an electronic video rental service to geographically distributed users [1], [2]. Using this service, users can select and watch movies at their convenient time and places, and they may interact with the movies using interactive operations such as fast forward, rewind, and pause. VOD is considered to be a promising Internet service [3]. Fig. 1 shows a typical VOD system. A server system stores a collection of movies. When a user requests to watch a movie, the server system retrieves this movie and delivers it to the user through a communication network.

VOD service is usually open to many users and so the server system should be able to serve multiple users concurrently. To serve each user, the server system retrieves and delivers video at the video playback rate (e.g., 1.5 Mbit/s for MPEG video). Consequently, the server system should have a large capacity to handle a large volume of bits. To satisfy this requirement, the server system can adopt one of two configurations [4].

Manuscript received February 23, 2002; revised April 23, 2004. This project was supported by the RGC Research Grant HKBU 200/01-02. This paper was recommended by Associate Editor C. Hu.

The authors are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: ywleung@comp.hkbu.edu.hk; rickyhou@comp.hkbu.edu.hk).

Digital Object Identifier 10.1109/TSM.2003.851158.

1083-4427/03/0005-0000 © 2003 IEEE

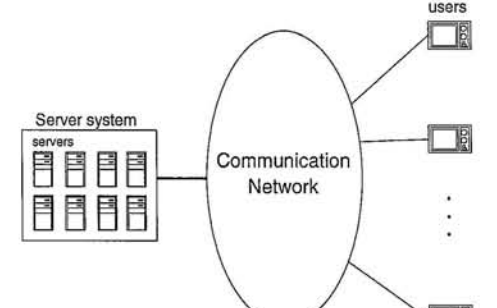


Fig. 1. Typical video-on-demand (VOD) system.

- 1) *Centralized server configuration* [Fig. 2(a)]—The server system uses a high-end computer (e.g., a multi-processor computer) as a video server.
- 2) *Distributed server configuration* [Fig. 2(b)]—The server system uses multiple low-end computers (e.g., personal computers) as video servers. Each server stores some of the movies, and it can serve multiple users concurrently. Overall, the server system can serve many users concurrently.

Serpanos and Bouloutas [4] made a comprehensive comparison between these two configurations. In particular, the distributed server configuration is attractive in three aspects.

- 1) *Good scalability*—The system can easily be scaled up by adding more video servers.
- 2) *High availability*—The system can still provide service when some servers fail or are under preventive maintenance.
- 3) *Competitive performance-to-price ratio*—Personal computers are currently fast and cheap.

The distributed server configuration is adopted by a commercial VOD system called iTV system [5].

B. Relevant Results in the Literature

Using the distributed server configuration, it is necessary to assign each movie to one or more video servers. This assignment problem has been investigated in the literature [6]–[8].

- Little and Venkatesh [6] formulated the assignment problem for a server system with identical servers. Their objective was to minimize the *blocking probability*, which is the probability that a request for VOD service is blocked (rejected) because the server system does

Design of an Interactive Video-on-Demand System

Yiu-Wing Leung, Senior Member, IEEE, and Tony K. C. Chan

Abstract—We design an interactive video-on-demand (VOD) system using both the client-server paradigm and the broadcast delivery paradigm. Between the VOD warehouse and the customers, we adopt a client-server paradigm to provide an interactive service. Within the VOD warehouse, we adopt a broadcast delivery paradigm to support many concurrent customers. In particular, we exploit the enormous bandwidth of optical fibers for broadcast delivery, so that the system can provide many video programs and maintain a small access delay. In addition, we design and adopt an interleaved broadcast delivery scheme, so that every video stream only requires a small buffer size for temporary storage. A simple proxy is allocated to each outgoing customer, and it retrieves video from the optical channels and delivers the video to the customer through an information network. The proposed VOD system is suitable for large-scale applications with many customers, and it has several desirable features: 1) it can be scaled up to serve more concurrent customers and provide more video programs; 2) it provides interactive operations; 3) it only requires point-to-point communication between the VOD warehouse and the customer and it does not involve any network control; 4) it has a small access delay; and 5) it requires a small buffer size for each video stream.

Index Terms—Broadcast delivery paradigm, client-server paradigm, video-on-demand.

I. INTRODUCTION

AN INTERACTIVE video-on-demand (VOD) system provides an electronic video rental service to geographically distributed customers [1]. It retrieves video programs from its storage and delivers them to the customers through an information network. The customers can select and watch video programs at their convenient time and places, and they can interact with the programs via interactive operations such as pause, fast-forward, and rewind.

A VOD system has to serve multiple customers concurrently, and therefore it must have a large enough capacity to provide multiple video streams. Many designs have been proposed for this purpose and they can be categorized into two categories: *client-server design* and *broadcasting design*.

A. Client-Server Design

The client-server design adopts the *client-server paradigm*. The system is composed of one or more servers [2], [3]. It maintains a dedicated video stream for each outgoing customer. When the customer performs an interactive operation, the system retrieves and delivers the corresponding video for him. This design

Manuscript received February 23, 2002; revised May 30, 2001. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Thomas L. Givens.

The authors are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: ywleung@comp.hkbu.edu.hk; tonychan@hkbu.edu.hk).

Digital Object Identifier 10.1109/TMM.2003.808181.

1520-9101/03/0001-0000 © 2003 IEEE

sign can provide an ideal interactive service to the customers, but it needs dedicated resources (such as I/O bandwidth) to maintain a video stream for each outgoing customer. For large-scale applications with many customers, this design requires large amount of resources.

A client-server design can use a *batching policy* [4]–[8] to serve more concurrent customers. The main idea is that the system waits for a time interval (called *batching window*) to collect a batch of requests for a video program. Then the system creates one video stream for this program and multicasts it to a batch of customers. In this manner, one video stream can serve multiple customers simultaneously. However, the customers have to wait before starting a VOD session (the waiting time is called *access delay*) and they cannot perform (or can only perform some constrained) interactive operations. Several batching policies have been proposed in the literature and they are as follows.

- Dan et al. [5] proposed that when the system can establish a new video stream, it selects the batch with the largest number of waiting customers and creates a video stream to serve all the customers in this batch. This batching policy can minimize the mean access delay, but some customers may experience long access delay.
- Dan et al. [6] proposed to choose a shorter batch window for the more popular video programs. They developed an analytical model and determined the window size for each video program.
- Almeroth et al. [7] proposed a batching policy that supports some constrained interactive operations by buffering a certain portion of the video program or joining the customers to another existing video stream.
- Liao and Li [8] proposed a batching policy called *split-and-merge*. When a customer performs an interactive operation, the system splits him from his original video stream and attempts to create a new video stream for him. If this is not possible, the customers wait. Once the interaction is done, the system attempts to merge this customer back to an existing video stream via buffering. If this is not possible, the customers wait.

B. Broadcasting Design

The broadcasting design adopts the *broadcast delivery paradigm* [9], [10] to serve many concurrent customers. There are three broadcasting designs for VOD [11]–[13]. The first design is called *periodic broadcasting* [11]. It broadcasts multiple streams of the same video program at staggered times periodically. To watch a video program, a customer waits until a video stream for this program (A) is broadcasted and then he receives this stream. The system can serve many concurrent customers because many customers can receive the same video stream from a broadcast channel simultaneously. However, it has a long mean access delay or

Location Estimation via Support Vector Regression

Zhi-Hi Wu, Chun-hung Li, Member, IEEE, Joseph Kee-Yin Ng, Senior Member, IEEE, and Karl R.P.H. Leung, Senior Member, IEEE

Abstract—Location estimation using the Global System for Mobile communication (GSM) is an emerging application that infers the location of the mobile receiver from multiple signals measurements. While geometrical and signal propagation models have been deployed to tackle this estimation problem, the terrain factors and power fluctuations have confined the accuracy of such estimation. Using support vector regression, we investigate the missing value location estimation problem by providing theoretical and empirical analysis on existing and novel approaches. A novel synthetic experiment is designed to assess the performance of the proposed estimation approaches. The proposed support vector regression approach shows promising performances, especially in terrains with local variations in environmental factors.

Index Terms—Location estimation, support vector regression, statistical estimation, Global System for Mobile communication.

1. INTRODUCTION TO GSM POSITIONING

LOCATION estimation service is very important for mobile location ubiquitous computing. Its range of applications includes logistics, field worker deployment, resource management, personal safety, and personal. Although the global positioning system (GPS) has been in service for many years, its wide-spread application has been limited by its reliance on special hardware. Furthermore, in metropolitan areas, the access to GPS signals is often limited.

On the other hand, location estimation technology based on GSM technology has advanced rapidly in recent years. In 1996, the Federal Communications Commission (FCC) directed the rules that wireless service companies should provide location identification for wireless emergency calls, and the accuracy of location estimation should be within 125 meters [2]. Since 2000, the FCC has stipulated separate requirements for both terminal-based and network-based positioning methods, and the FCC rules revised that 67 percent of estimations should be less than 50 meters and 100 meters in terminal-based and network-based solutions, respectively [6]. A terminal-based positioning solution based on modifications of handsets, SIM cards, or both, means that subscribers have to upgrade their mobile phones or renew SIM cards to enable this location service. By contrast, a network-based solution estimates location by upgrading the operator network instead of the handset; therefore, it can benefit all mobile phone subscribers. In July of 2002, these wireless Enhanced 911 rules were revised again to improve the effectiveness and reliability of 911 service by reporting the telephone number

of a wireless 911 caller and the location of the cellular phone within 50 to 100 meters in most cases [4]. Hence, there have been a lot of investigations on location estimations for satisfying the rules of Enhanced 911 from the FCC.

In general, the terminal-based method can provide a more accurate estimate than the network-based method, which is also more (server-to-proxy) bandwidth consumption. However, this poses significant challenges to caching. Most conventional rate adaptation mechanisms are executed during the encoding process (e.g., adjusting quantizers [8], [22]) and, hence, they are difficult to apply to cached videos. There have been research efforts to combine proxy caching with video layering or transcoding [10], [14], [18], [20], but these adaptive systems suffer from either coarse adaptation granularity (due to the inflexible structures of existing video coders) or a high computation overhead (due to the transcoding operations).

In this paper, we propose a novel video caching framework to achieve low-cost and fine-grained rate adaptation. The network employs the MPEG-4 fine-grained scalable (FGS) video with bit-plane coding, which enables post-encoding rate control by partitioning the video stream at specific rates [7]. The fine-grained rate control operations can be efficiently implemented at the server or the proxy, resulting in a low computational cost and a fast response. The proposed framework is both network-aware and media adaptive: clients can benefit from heterogeneous access bandwidths, and adaptive FGS videos are used to meet the clients' bandwidth conditions and control the backbone bandwidth consumption.

Z.-H. Wu, C.-H. Li, and J.-K.-Y. Ng are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: fioncwai, chli, jkg@comp.hkbu.edu.hk).

K.R.P.H. Leung is with the Department of Information and Communications Technology, Hong Kong Institute of Vocational Education (Ting Yi), Tsim Yi Island, Hong Kong (e-mail: leung@hkiv.edu.hk).

Manuscript received 8 June 2004; revised 6 June 2005; accepted 12 Sept. 2005; published online 16 Jan. 2007.

For information on obtaining reprints of this article, please send e-mail to: tmm@computer.org, and reference IEEECS Log Number TMM-05-0604.

1520-9101/07/0003-0000 © 2007 IEEE

But, for the GSM network, to implement time-involved estimate algorithms such as TDOA, TDOA, and E-OTD, expensive clocks for precise time synchronization should be installed. Collocate is an example of a TDOA-driven system that uses GPS time synchronization. Their field tests, based on real-world data, resulted in an accuracy in between 187 and 287 m [7]. Even though a GPS receiver can be used for synchronization purposes, extra hardware would add a heavy burden in terms of installation costs for the network operator.

Scheduling real-time requests in on-demand data broadcast environments*

Victor C. S. Lee · Xiao Wu · Joseph Kee-Yin Ng

Published online: 15 May 2006
© Springer Science + Business Media, LLC 2006

Abstract On-demand broadcast is an attractive data dissemination method for mobile and wireless computing. In this paper, we propose a new online preemptive scheduling algorithm, called PRDS that incorporates urgency, data size and number of pending requests for real-time on-demand broadcast system. Furthermore, we use pyramid preemption to optimize performance and reduce overhead. A series of simulation experiments have been performed to evaluate the real-time performance of our algorithm as compared with other previously proposed methods. The experimental results show that our algorithm substantially outperforms other algorithms over a wide range of workloads and parameter settings.

Keywords Real-time scheduling algorithm · On-demand broadcast · Preemption · Mobile computing

1. Introduction

With the increasing availability of high bandwidth links as well as the popularity of portable wireless devices, such as notebook computers and PDA, mobile and wireless computing

*The work described in this paper was partially supported by grants from CityU (Project No. 7001841) and RGC CERG Grant No. HKBU 217/03E.

†This paper is an extended version of the paper "A preemptive scheduling algorithm for wireless real-time on-demand data broadcast" that appeared in the 11th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications.

V. C. S. Lee (✉) · X. Wu
Department of Computer Science, City University of Hong Kong,
83 Tat Chee Avenue, Kowloon, Hong Kong
e-mail: vcslee@cityu.edu.hk

X. Wu
e-mail: wuxiao@cs.cityu.edu.hk

J. K.-Y. Ng
Department of Computer Science, Hong Kong Baptist University,
Kowloon Tong, Hong Kong
e-mail: jkg@comp.hkbu.edu.hk

A QoS-Enabled Transmission Scheme for MPEG Video Streaming*

JOSEPH KEE-YIN NG jkg@comp.hkbu.edu.hk

Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

KARL R.P.H. LEUNG krl@comp.hkbu.edu.hk

Department of Information & Communications Technology, Hong Kong Institute of Vocational Education (Ting Yi), Tsim Yi Island, Hong Kong

CALVIN KIN-CHEUNG HUI kchu@comp.hkbu.edu.hk

Department of Computer Science, Hong Kong Baptist University, Kowloon, Hong Kong

Published online: 9 June 2005

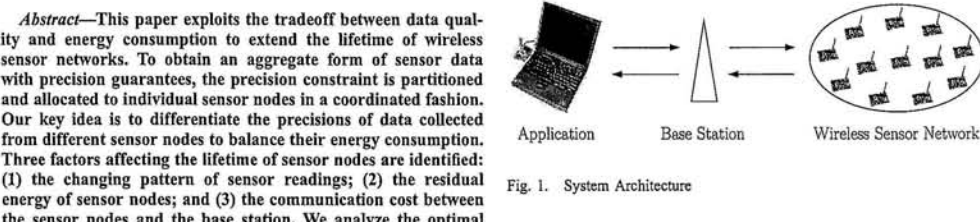
Abstract. While MPEG is the *de facto* encoding standard for video services, online video streaming service is becoming popular over the open network such as the Internet. As the performance of open network is non-predictable and uncontrollable, the tuning of the quality of service (QoS) for on-line video streaming service is difficult. In order to provide better QoS for the delivery of videos, there are proposals of new encoding formats or new transmission protocols for on-line video streaming. However, these results are not compatible with popular video services or network protocols and hence these approaches are so far not very successful. We use another approach which tries to bypass these problems. We designed a QoS Tuning Scheme and QoS-Enabled Transmission Scheme for transmitting MPEG videos from video servers to clients. According to the traffic characteristics between the video server and each individual client, the QoS Tuning Scheme tunes the QoS to be delivered to each individual client on the fly. Furthermore, our QoS-Enabled Transmission Scheme can be applied over any protocol, such as HTTP which is the most popular protocol over the open network. With our transmission scheme, bandwidth can be better utilized by reducing transmitted frames which would have missed their deadlines and would eventually be discarded by the clients. This is achieved by sending frames according to their impact on the QoS in the playback under the allowed throughput. With these schemes, users can enjoy video streaming through their favorite video players and with the best possible QoS. In order to facilitate the real time QoS tuning, a metric, QoS-GFS, is developed. This QoS-GFS is extended from the QoS-Index, another metric which has taken human perspective in the measurement of video quality. Hence QoS-GFS is better than the common metrics which measure QoS by means of rate of transmission of bytes or MPEG frames. We designed and implemented a middleware to perform empirical tests of the proposed transmission scheme and QoS tuning scheme. Experiment results show that our schemes can effectively enhance the QoS for online MPEG video streaming services.

Keywords: MPEG video streaming, transmission scheme, quality of services, QoS control, video on demand

1. Introduction

Optimizing Lifetime for Continuous Data Aggregation with Precision Guarantees in Wireless Sensor Networks

Xueyan Tang, Member, IEEE, and Jianliang Xu, Member, IEEE



Abstract—This paper explores the tradeoff between data quality and energy consumption to extend the lifetime of wireless sensor networks. To obtain an aggregate form of sensor data with precision guarantees, the precision constraint is partitioned and allocated to individual sensor nodes in a coordinated fashion. Our key idea is to differentiate the precision of data collected from different sensor nodes to balance their energy consumption. Three factors affecting the lifetime of sensor nodes are identified: (1) the changing pattern of sensor readings; (2) the residual energy of sensor nodes; and (3) the communication cost between the sensor nodes and the base station. We analyze the optimal precision allocation in terms of network lifetime and propose an adaptive scheme that dynamically adjusts the precision constraints at the sensor nodes. The adaptive scheme also takes into consideration the topological relation among sensor nodes and the effect of in-network aggregation. Experimental results using real data traces show that the proposed scheme significantly improves network lifetime compared to existing methods.

Index Terms—data aggregation, data accuracy, energy efficiency, network lifetime, sensor network.

1. INTRODUCTION

WIRELESS sensor networks are used in a wide range of applications to capture, gather and analyze live environmental data [1], [2]. A wireless sensor network typically consists of a base station and a group of sensor nodes (see Figure 1). The sensor nodes are responsible for continuously sampling physical phenomena such as temperature and humidity. They are also capable of communicating with each other and the base station through radios. The base station, on the other hand, serves as a gateway for the sensor network to exchange data with applications to accomplish their missions. While the base station can have continuous power supply, the sensor nodes are usually battery-powered. The batteries are inconvenient and sometimes even impossible to replace. When a sensor node runs out of energy, its coverage is lost. The mission of a sensor application would not be able to continue if the coverage loss is remarkable. Therefore, the practical at different magnitudes and frequencies, the sensor nodes may report data at different rates. Second, the wireless communication cost depends on the transmission distance [10], [11]. Due to the geographically distributed nature of sensor networks, the sensor nodes are likely to differ significantly in the energy cost of sending a message to the base station. Even if all sensor nodes report data at the same rate, their energy consumption can be highly unbalanced, thereby reducing network lifetime. In addition to reporting local sensor readings, the intermediate nodes in a multi-hop network are also responsible for relaying

Manuscript received October 24, 2006; revised February 23, 2007; accepted by IEEE/ACM TRANSACTIONS ON NETWORKING Editor N. Shah. This work was supported in part by a grant from Hong Kong Baptist University (Project No. B04709). Jianliang Xu's work was supported in part by a grant from the Research Grants Council of the Hong Kong SAR, China (Project No. HKR211/03B). A preliminary report of this work was presented at IEEE INFOCOM 2006.

X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798 (e-mail: xytang@nus.edu.sg).

J. Xu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (e-mail: xujianliang@hkbu.edu.hk).

0000-0000/06/0000-0000 © 2007 IEEE

An Error-Resilient and Tunable Distributed Indexing Scheme for Wireless Data Broadcast

Jianliang Xu, Member, IEEE, Wang-Chien Lee, Member, IEEE, Xueyan Tang, Member, IEEE, Qing Gao, and Shaping Li

Abstract—Access efficiency and energy conservation are two critical performance concerns in a wireless data broadcast system. We propose in this paper a novel parameterized index called the exponential index that has a three-way distributed structure for wireless data broadcast. Based on two tuning knobs, index base and chunk size, the exponential index can be tuned to optimize the access latency with the tuning time bounded by a given limit, and vice versa. The client access algorithm for the exponential index under unreliable broadcast is described. A performance analysis of the exponential index is provided. Extensive simulation and experimental results are conducted to evaluate the performance under various link error probabilities. Simulation results show that the exponential index substantially outperforms the state-of-the-art indexes. In particular, it is more resilient to link errors and achieves more performance advantages from index caching. The results also demonstrate its great flexibility in trading access latency with tuning time.

Index Terms—index structure, data broadcast, energy conservation, mobile computing.

1. INTRODUCTION

WIRELESS data broadcast has received a lot of attention from industries and academia in recent years. It has been available as commercial products for many years (e.g., StarBand [20] and Hughes Network [21]). In particular, the recent announcement of the smart personal objects technology (SPOT) by Microsoft [16] further highlights the industrial interest in and feasibility of utilizing broadcast for wireless data services. With a continuous broadcast network (called DirectBand Networks) using FM radio subcarrier frequencies, SPOT-based devices (e.g., PDAs and watches) can continuously receive timely information such as stock information, airline schedules, local news, weather, and traffic information.

Access efficiency and energy conservation are two main performance issues for the clients in a wireless data broadcast system. Access efficiency concerns how fast a request is satisfied, and energy conservation concerns how to reduce a mobile client's energy consumption when it accesses the data of interest. While access efficiency is a constantly tackled issue in data systems and database research, energy conservation is very critical due to the

• J. Xu and Q. Gao are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. E-mail: xujianliang@hkbu.edu.hk.
• W.-C. Lee is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802. E-mail: wlee@psu.edu.
• X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798. E-mail: xytang@nus.edu.sg.
• S. Li is with the College of Computer Science, Zhejiang University, Hangzhou 310027, China. E-mail: sheli@zhu.edu.cn.

Manuscript received 6 Jun. 2005; revised 6 June 2005; accepted 8 Sept. 2005; published online 11 Jan. 2006.

For information on obtaining reprints of this article, please send e-mail to: tkd@computer.org, and reference IEEECS Log Number TKDE-0407-0505.

1041-4348/06/0000-0000 © 2006 IEEE

Published by the IEEE Computer Society

Top-k Monitoring in Wireless Sensor Networks

Minji Wu, Student Member, IEEE, Jianliang Xu, Member, IEEE, Xueyan Tang, Member, IEEE, and Wang-Chien Lee, Member, IEEE

Abstract—Top-k monitoring is important to many wireless sensor applications. This paper explores the semantics of top-k query and proposes an energy-efficient monitoring approach called FLA. The basic idea is to install a filter at each sensor node to suppress unnecessary sensor updates. Filter setting and query evaluation upon updates are two fundamental issues to the correctness and efficiency of the FLA approach. We develop a query evaluation algorithm that is capable of handling concurrent sensor updates. In particular, we present optimization techniques to reduce the probing cost. We design a skewed filter setting scheme, which aims to balance energy consumption and prolong network lifetime. Moreover, two filter update strategies, namely, eager and lazy, are proposed to favor different application scenarios. We also extend the algorithm to several variants of top-k query, that is, order-insensitive, approximate, and value monitoring. The performance of the proposed FLA approach is extensively evaluated using real data traces. The results show that FLA substantially outperforms the existing TAG-based approach and range caching approach in terms of both network lifetime and energy consumption under various network configurations.

Index Terms—Sensor network, data management, energy efficiency, top-k, continuous query.

1. INTRODUCTION

RECENT advances in signal processing, microelectronics, and wireless communications have enabled the deployment of large-scale sensor networks for many applications such as habitat and environment monitoring [28]. A wireless sensor network typically consists of a base station and a group of sensor nodes (see Fig. 1). The base station serves as a gateway for the sensor network to exchange data with external users. The sensor nodes, on the other hand, are responsible for sensing and collecting data from their local environments. They are also capable of processing sensed data and communicating with their neighbors and the base station.

Monitoring aggregate forms of sensed data is important to many sensor applications and has drawn a lot of research attention [16], [17], [14], [19], [29], [30]. Among those aggregates, a top-k query requests the list of k sensor nodes with the highest (or lowest) readings. For example:

- **Environmental Monitoring.** Consider an environment-monitoring sensor network. A top-k query is issued to find out the nodes and their corresponding areas with the highest pollution indexes for the purpose of pollution control or research study.
- **Network Management.** Power supply is critical to the operation of a wireless sensor network. Thus, a

• M. Wu and J. Xu are with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. E-mail: xujianliang@hkbu.edu.hk.
• Z. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798. E-mail: xytang@nus.edu.sg.
• W.-C. Lee is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802. E-mail: wlee@psu.edu.

Manuscript received 3 Mar. 2006; revised 13 Oct. 2006; accepted 11 Jan. 2007; published online 7 Mar. 2007.

For information on obtaining reprints of this article, please send e-mail to: tkd@computer.org, and reference IEEECS Log Number TKDE-0318-0606.

Digital Object Identifier no. 10.1109/TKDE.2006.0000000

1041-4348/06/0000-0000 © 2006 IEEE

top-k query may be issued to continuously monitor the sensor nodes with the least residual energy so that these sensor nodes can be instructed to adapt themselves (for example, reducing sampling rates) to extend network lifetime.

A naive implementation of monitoring top-k query is to use a centralized approach in which all sensor readings are periodically collected by the base station, which then computes the top-k result set. To reduce network traffic in data collection, an in-network aggregation technique, known as TAG, has been proposed [14]. In this approach, a routing tree rooted at the base station is first established, and the data are then aggregated and collected along the routing tree to the base station. Consider an example shown in Fig. 2a, where sensor nodes A, B, and C form a routing tree. The readings of these sensor nodes at three successive sampling instances t_1 , t_2 , and t_3 are shown in the tables in Fig. 2a. Suppose we are monitoring a top-k query. Employing TAG, at each sampling instance, nodes B and C send their current readings to the parent (that is, node A), which compares the data received with its own reading and sends the highest reading to the base station. In this example, a total of nine messages are sent. Node B is involved in the message exchange at each sampling instance though the top-k result is always node C. Therefore, this approach incurs unnecessary updates in the network and is not energy efficient.

In this paper, we exploit the semantics of top-k query and propose a novel filter-based monitoring approach called FLA. The basic idea is to install a filter at each sensor node to suppress unnecessary sensor updates. The base station also keeps a copy of the filter setting to maintain an (error bounded) approximate view of each node's reading. A sensor node updates its reading with the base station only when the reading passes the filter. The correctness of the top-k result is ensured if all sensor nodes perform updates according to their filters. Fig. 2b shows an

Time-Critical On-Demand Data Broadcast: Algorithms, Analysis, and Performance Evaluation

Jianliang Xu, Member, IEEE, Xueyan Tang, Member, IEEE, and Wang-Chien Lee, Member, IEEE

Abstract—On-demand broadcast is an effective wireless data dissemination technique to enhance system scalability and deal with dynamic user access patterns. With the growth of time-critical information services in emerging applications, there is an increasing need for the system to support timely data dissemination. This paper investigates online scheduling algorithms for time-critical on-demand data broadcast. We propose a novel scheduling algorithm called SIN-a that takes the urgency and number of outstanding requests into consideration. An efficient implementation of SIN-a is presented. We also analyze the theoretical bound of request drop rate when the request arrival rate rises toward infinity. Trace-driven experiments show that SIN-a significantly outperforms existing algorithms over a wide range of workloads and approaches the analytical bound at high request rates.

Index Terms—Mobile computing, on-demand data broadcast, scheduling, content delivery, time constraint.

1. INTRODUCTION

THE ever-growing popularity of the Internet and the resultant slow responses perceived by users have given rise to vast research efforts on improving the performance of Web accesses. As the system scale and user base continue to grow, there is an increasing demand for information providers to be capable of concurrently delivering a large amount of information to a huge number of users, especially in popular events such as elections and Olympic games. As a result, innovative delivery technologies, including satellite communications (e.g., StarBand [26] and DISCOVER [27]), cable networks, and wireless networks (e.g., 2.5G and 3G), have been developed and deployed to provide shared broadband Internet accesses. Different from traditional networks, a distinguished feature of these new technologies is that they naturally support broadcast. In contrast to unicast, where a data item of interest to multiple clients must be sent individually to each client, broadcast satisfies all outstanding requests for the same item by a single transmission. This leads to a more efficient use of shared bandwidth, hence improving the system throughput and user-perceived response time [19], [20]. In general, there are two data broadcast approaches [8], [20]: *Push-based broadcast* computes the broadcast program based on historical access statistics; *on-demand broadcast*

1. A third approach is hybrid broadcast that combines on-demand broadcast with push-based broadcast.

• J. Xu is with the Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong. E-mail: xujianliang@hkbu.edu.hk.
• X. Tang is with the School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798. E-mail: xytang@nus.edu.sg.
• W.-C. Lee is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802. E-mail: wlee@psu.edu.

Manuscript received 2 Nov. 2004; revised 13 Mar. 2005; accepted 26 Apr. 2005; published online 28 Nov. 2005.

For information on obtaining reprints of this article, please send e-mail to: tkd@computer.org, and reference IEEECS Log Number TPDS-0308-1104.

1041-4348/06/0000-0000 © 2006 IEEE

schedules broadcast items on the fly based on current outstanding requests. While push-based broadcast is useful for certain applications (e.g., a small set of data items with stable access patterns), on-demand broadcast is more widely used for dynamic, large-scale data dissemination like that in the Internet.

With the rapid growth of time-critical information services and business-oriented applications, there is an increasing demand to support quality of service (QoS) in content distribution [15], [24], [28]. In many situations, users require are associated with time constraints as a measure of QoS. These constraints can be imposed either by the users or the applications. For example, in wireless financial services, many users are interested in the up-to-minute (or even "second") stock quotes in order to react to dynamic and rapid market developments. As another example, in wireless location-based services [18], the queried information (e.g., the local theaters) is valid only within a local area. When the mobile user moves away from the area, the information becomes invalid. In addition, a service level agreement (SLA) between a content/service provider and its users usually specifies the desired performance for Web requests, e.g., the response time of requests for CNN.com should not exceed 5 seconds [24]. In all the above cases, a *deadline* is associated with each request beyond which the serving of the request is useless (or less useful).

This paper focuses on on-demand data broadcast with time constraints, which we shall refer to as *time-critical on-demand broadcast*. A key issue in the design of an on-demand data broadcast system is the scheduling algorithm used to select and broadcast requested items from outstanding requests. While there has been significant work on developing on-demand broadcast scheduling algorithms (e.g., [15], [24], [10], [29]), none of them has considered the time constraints associated with requests. On the other hand, although some time-critical scheduling algorithms have been proposed for unicast-based real-time systems and push-based broadcast systems (e.g., [11], [14]), they are not applicable or not effective to on-demand broadcast

schedules broadcast items on the fly based on current outstanding requests. While push-based broadcast is useful for certain applications (e.g., a small set of data items with stable access patterns), on-demand broadcast is more widely used for dynamic, large-scale data dissemination like that in the Internet.

With the rapid growth of time-critical information services and business-oriented applications, there is an increasing demand to support quality of service (QoS) in content distribution [15], [24], [28]. In many situations, users require are associated with time constraints as a measure of QoS. These constraints can be imposed either by the users or the applications. For example, in wireless financial services, many users are interested in the up-to-minute (or even "second") stock quotes in order to react to dynamic and rapid market developments. As another example, in wireless location-based services [18], the queried information (e.g., the local theaters) is valid only within a local area. When the mobile user moves away from the area, the information becomes invalid. In addition, a service level agreement (SLA) between a content/service provider and its users usually specifies the desired performance for Web requests, e.g., the response time of requests for CNN.com should not exceed 5 seconds [24]. In all the above cases, a *deadline* is associated with each request beyond which the serving of the request is useless (or less useful).

This paper focuses on on-demand data broadcast with time constraints, which we shall refer to as *time-critical on-demand broadcast*. A key issue in the design of an on-demand data broadcast system is the scheduling algorithm used to select and broadcast requested items from outstanding requests. While there has been significant work on developing on-demand broadcast scheduling algorithms (e.g., [15], [24], [10], [29]), none of them has considered the time constraints associated with requests. On the other hand, although some time-critical scheduling algorithms have been proposed for unicast-based real-time systems and push-based broadcast systems (e.g., [11], [14]), they are not applicable or not effective to on-demand broadcast

Although many projects and researches have been conducted on online distance learning, the issues of security have only been studied recently (Cheung *et al.*, 1999a; Cheung and Hui, 1999; Furnell *et al.*, 1998, 1999). In fact, there are quite a number of security concerns in this type of education system, for example, user authentication and access control, non-repudiation for critical actions like course registration, course tuition fee payment, confidentiality of user personal information, course material copyright protection, etc. For more information on what security issues are involved in online learning system may consider, one can refer to the security framework given by Furnell *et al.* (1998); and for more information on the problem of user authentication and access control, one can refer to Cheung *et al.* (1999a) and Cheung and Hui (1999). In particular, Cheung *et al.* (1999) provides a security model such that a legitimately registered student cannot easily share the account with non-registered students.

Depending on the type of courses offered by an organization, the security concerns may differ slightly. There is one security problem, the copyright protection problem, which is important to all kinds of e-courses, especially for type (1) and (3) of courses mentioned above. Typical scenarios include the following: registered students infringing the copyrights of the course materials by passing the materials to non-registered students. Usually, the organization providing the course materials depends on the registration fee to maintain the operation of the organization. This copyright infringement severely jeopardizes the income of the organization.

Trustworthy Browsing – A Secure Web Accessing Model

Joe C.K. Yau¹, Lucas C.K. Hui¹, Bruce S.N. Cheung², S.M. Yiu¹, Y. Woo¹, K.W. Lau¹, Eric H.M. Li²¹Department of Computer Science, The University of Hong Kong (jcklau, hui, smyiu, ywoo, kwlau}@cs.hku.hk²School of Professional And Continuing Education, The University of Hong Kong (bruce, ehmlil}@hkuspace.hku.hk

Abstract

The web technology we are enjoying now is insecure, especially for accessing sensitive information. There is no solution that provides highly reliable user authentication to prove the identity of the information requester to the server, nor a solution that securely protects the browsed information from being stolen.

To solve this problem, the Trustworthy Browsing system, based on a special browsing paradigm, is designed and developed. It employs a hardware-software hybrid solution and the key elements include the use of cryptographic hardware token for ensuring the user's identity, a customized browser to protect the contents, and an authentication protocol for verifying HTTP requests from the browser. The proposed solution can be integrated into applications for e-Education providers, e-Books publisher as well as electronic artwork publishers. In this paper, a detailed description and an in-depth security analysis for this system are given.

1. Introduction

Nowadays, Internet has become one of the essentials to life. A research done by Nielsen/NetRatings that 73% Americans have Internet access [1], and the global Internet population is growing at a rate of 4% annually [2]. But as Internet becomes mature, users are becoming more concerned about security issues related to web access. From content provider's perspective, it is important that the content be delivered to eligible users only; and copyrighted contents

must be carefully protected from illegal copying. We need a web access model that is trustworthy.

To ensure that the content is accessible only by eligible users, most systems require users to authenticate themselves using password before accessing the information. But if the content is confidential, authenticating by password would not be secure enough.

Another concern is that most of the web browsers available today allow users to save the content that they are viewing. But there are situations where content owners want to protect the contents from being copied, and only allow users to view the content online. Examples of such contents could be entertainment materials (e.g., movie or music), or information that is highly sensitive (e.g., a confidential document). With the popular web browsers we currently have, making copies of the content is easy. It is even possible for an adversary to dig into the browser's disk cache and retrieve the information. That brings threats to the content owners.

To better protect the content, we want to improve the security of web accessing. Specifically, we want a solution that has the following characteristics:

- (1) the content owner can be more certain of users' identities who access the content; and
 - (2) security protection to content being accessed.
- To solve this problem, we have designed and developed a hardware-software hybrid solution that can provide a secure browsing environment. In this paper, we introduce our solution – *trustworthy browsing*. In Section 2, possible applications for our solution are discussed. The problem is formally defined in Section 3. Section 4 gives a detailed description of our system, and Section 5 gives an in-depth security analysis of it. Some concluding remarks and future research directions for this project are given in Section 6.

This research is supported in part by the UGC A&E Scheme (AoE/E-01/99), a RGC grant (HKU714403E), and an ITF grant (ITS17001).

Towards a Secure Copyright Protection Infrastructure for e-Education Material: Principles Learned from Experience*

Joe Cho-Ki Yau¹, Lucas Chi-Kwong Hui¹, Shu-Ming Yiu¹ and Bruce Siu-Nang Cheung²

(Corresponding author: Joe Cho-Ki Yau)

¹Department of Computer Science, The University of Hong Kong

Pokfulam Road, Hong Kong. (Email: jcklau, hui, smyiu@cs.hku.hk)

²School of Professional and Continuing Education, The University of Hong Kong

Pokfulam Road, Hong Kong. (Email: bruce@hkuspace.hku.hk)

(Received July 7, 2004; revised and accepted Aug. 9, 2005)

Abstract

Copyright of e-Education material is valuable. The need for protecting it is prominent. In the past two years, we have developed an infrastructure called e-Course eXchange (eCX) for protecting the copyrights of e-Courses, right from its development phase to its delivery phase. It has been adopted by an education institute, with a user-base of over 70,000 students, and has been receiving positive feedback from students. To design a secure and effective copyright protection infrastructure is not trivial. In particular, for efficiency purposes, one may allow students to retain a local copy of the e-Course material in their own computers; on the other hand, we should make it difficult for them to make illegal copies of the material. Only storing the material in encrypted form is not enough to protect the material. In this paper, we summarize some principles and knowledge we have gained through this project that should be observed for designing a secure copyright protection system. We believe that these principles would be useful to developers and researchers for designing and developing such a system.

Keywords: Copyright protection, software protection, reverse engineering, e-Education, e-Learning

1. Introduction

With the advent of the digital age, e-Education has become one of the most important channels for students to acquire knowledge. Students of different levels are one way or the other making use of this user-oriented learning, and researchers are actively working on this area to make

the best use of it. However, as pointed out by Furnell [8, 10], little attention has been devoted to the security concerns of e-Education. Among these security concerns, the copyright protection problem for the e-Education material is one of the most important concerns that are vital to the operation of e-Education institutes.

Almost in all e-business contexts, the intellectual property of material or content could be the most valuable asset of the business itself, and this is undoubtedly true for the e-Education sector. Many organizations rely on the income generated from students studying e-Courses. Registered students infringing the copyrights of the course materials by passing the materials to non-registered students that severely jeopardize the income of the organization. Hence, copyrights of e-Course materials must be securely protected.

About three years ago, we initiated a study of the problem of protecting the copyright of e-Education materials. We proposed an infrastructure, called e-Course eXchange (eCX [26, 27, 28, 29]). This infrastructure has been developed and deployed to a large group of users. In this paper, we will use eCX as a case study, and discuss the lessons we have learned from it. In Sections 2 and 3, we will discuss the design of eCX. In Section 4, we will discuss the issues and design principles we learned from our experience with eCX. Section 5 then concludes the paper. Although eCX may seem to be a solution specific to the e-Education sector, the lessons we have learned are general enough and applicable to other copyright protection systems.

*This paper represents the personal opinion of the authors and does not represent the official standpoint of HKU SPACE.

Pattern Recognition

IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 12, NO. 1, JANUARY 2003

Color Image Indexing Using BTC

Guoping Qiu

Abstract—This paper presents a new application of a well-studied image coding technique, namely block truncation coding (BTC). It is shown that BTC can not only be used for compressing color images, it can also be conveniently used for content-based image retrieval from image databases. From the BTC compressed stream (without performing decoding), we derive two image content description features, one termed the block color co-occurrence matrix (BCCM) and the other block pattern histogram (BPH). We use BCCM and BPH to compute the similarity measures of images for content-based image retrieval applications. Experimental results are presented which demonstrate that BCCM and BPH are comparable to similar state of the art techniques.

Index Terms—Block truncation coding (BTC), color quantization, content-based image retrieval, image coding, image database.

1. INTRODUCTION

THE rapid expansion of the Internet and fast advancement in color imaging technologies have made digital color images more and more readily available to professional and amateur users. The large amount of image collections available from a variety of sources (digital camera, digital video, scanner, the Internet, etc.) have posed increasing technical challenges to computer systems to store/transmit and index/manage the image data effectively and efficiently to make such collections easily accessible.

The storage and transmission challenge is tackled by image coding/compression, which has been studied for more than 30 years and significant advancements have been made. Many successful, efficient and effective image-coding techniques have been developed and the body of literature on image coding is huge. Well-developed and popular international standards, e.g., [6], on image coding have also long been available and widely used in many applications.

The challenge to image indexing/management is studied in the context of image database, which has also been actively researched by researchers from a wide range of disciplines including those from computer vision, image processing, and traditional database areas for over a decade [14]. One particularly promising approach to image database indexing and retrieval is the query by image content (QBIC) method [1], whereby the visual contents of the images, such as color distribution (color histogram), texture attributes and other image features are extracted from the image using computer vision/image processing techniques and used as indexing keys. In an image database, these visual keys are stored along with the actual imagery data and

image retrieval from the database is based on the matching of the models visual keys with those of the query images. Because extra information has to be stored with the images, traditional approach to QBIC is not efficient in terms of data storage. Not only is it inefficient, it is also inflexible in the sense that image matching/retrieval can only be based on the pre-computed set of image features.

Many image-coding methods developed over the years are essentially based on the extraction and retention of the most important (visual) information of the image. The retained important information, such as the DCT coefficients of JPEG can be used for image indexing and object recognition [2]. However, since JPEG and other similar methods are not explicitly designed for image indexing purpose, models and features have to be derived from the transform coefficients, which generally involves complicated and complex computation and also leads to an expansion of data. For example, it has been demonstrated that color is an excellent cue for image indexing [3], however, it is difficult to explicitly exploit color information from the transform coefficients without decoding. On the other hand, nontransform based image coding can have image features such as color more easily available. For example, recent work on color image coding using vector quantization has demonstrated that color as well as pattern information can be readily available in the compressed image stream (without performing decoding) to be used as image indices for effective and efficient image retrieval [4].

Block truncation coding (BTC) is a relatively simple image coding technique developed in the early years of digital imaging more than 20 years ago [5]. Although it is a simple technique, BTC has played an important role in the history of digital image coding in the sense that many advanced coding techniques have been developed based on BTC or inspired by the success of BTC. Even though the compression ratios achievable by BTC have long been surpassed by many newer image-coding techniques such as DCT (JPEG) [6] and wavelet [7], the computational simplicity of BTC has made it and BTC-like image coding techniques attractive in applications whereby real time fast implementation is desirable. Furthermore, with the rapid advancement in processor speed, storage device technology and faster network connection, low bit rate coding is no longer a critical factor in many practical applications. Based on a certain tradeoff, higher bit rate and complexity can be acceptable. On the other hand, with very large image collections becoming more and more common, effectively managing large image database, making images easily accessible have become a challenge. Modern imaging systems not only require efficient coding, but also easy manipulation, indexing and retrieval, the so-called "fourth criterion" in image coding [18].

In this paper we shall show another attractive feature of BTC-types image coding methods. We shall demonstrate that BTC



Available online at www.sciencedirect.com

ScienceDirect

Pattern Recognition 40 (2007) 1711–1721

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/pr

Visual guided navigation for image retrieval

Guoping Qiu*, Jeremy Morris, Xunli Fan

School of Computer Science, University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, UK

Received 12 April 2006; accepted 27 September 2006

Abstract

In this work, we are interested in technologies that will allow users to actively browse and navigate large image databases and to retrieve images through interactive fast browsing and navigation. The development of a browsing/navigation-based image retrieval system has at least two challenges. The first is that the system's graphical user interface (GUI) should intuitively reflect the distribution of the images in the database in order to provide the users with a mental picture of the database content and a sense of orientation during the course of browsing/navigation. The second is that it has to be fast and responsive, and be able to respond to users' actions at an interactive speed in order to engage the users. We have developed a method that attempts to address these challenges of a browsing/navigation-based image retrieval systems. The unique feature of the method is that we take an integrated approach to the design of the browsing/navigation GUI and the indexing and organization of the images in the database. The GUI is tightly coupled with the algorithms that run in the background. The visual cues of the GUI are logically linked with various parts of the repository (image clusters of various particular visual themes) thus providing intuitive correspondences between the GUI and the database contents. In the backend, the images are organized into a binary tree data structure using a sequential maximal information coding algorithm and each image is indexed by an n -bit binary index thus making response to users' action very fast. We present experimental results to demonstrate the usefulness of our method both as a pre-filtering tool and for developing browsing/navigation systems for fast image retrieval from large image databases.

© 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Image database; Image retrieval; Browsing/navigation; Entropy; Information theory; Color

1. Introduction

Managing large image database and providing effective tools for users to quickly find image items they are looking for is still a very challenging problem. In the past decade or so, the content-based image indexing and retrieval (CBIR) paradigm has dominated the research community, e.g. Refs. [1–6]. There are a number of intrinsic weaknesses associated with the traditional CBIR paradigm, which have hindered progress. First, the objectives of CBIR are ill defined. Although the idea of using visual examples to find similar images is sound and intuitive, it is problematic in practice. It is not clear in what circumstances/application scenarios users would want/prefer to search images by example. The definition of CBIR is too broad and vague; retrieval by example can mean different things to different users and in

different applications. Even finding a starting query example can be problematic because some visual forms are intrinsically hard to describe precisely. Second, current state of the art technologies are not yet mature enough to realize the ideal of CBIR. In particular, it is difficult to compute image similarity measures that match the perceptual differences between images. Automatically retrieved images are often not what the users expected, hence there is a gap between the retrieval results and users' expectation. Third and fundamentally, the CBIR paradigm puts the user in a passive position in the sense that retrieval results are largely determined by the computational algorithms; the user cannot actively control the retrieval results. The immaturity of the enabling computational algorithms has only made the task even more difficult.

Recent trends in CBIR have been to introduce browsing, navigation and relevant feedback facilities to enable users to interact with the retrieval system and to engage the users, examples include Refs. [7–12]. The advantages of a

* Corresponding author. Tel.: +44115 8466507; fax: +44115 9514254. E-mail address: qiu@cs.nott.ac.uk (G. Qiu).

0031-3203/\$30.00 © 2006 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patrec.2006.09.020



Available at www.ElsevierComputerScience.com

POWERED BY SCIENCE @ DIRECT®

Pattern Recognition 37 (2004) 2177–2193

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patrec

Compressing histogram representations for automatic colour photo categorization

Guoping Qiu*, Xia Feng¹, Jianzhong Fang

School of Computer Science, The University of Nottingham, Jubilee Campus, Nottingham NG8 1BB, UK

Received 12 September 2003; received in revised form 8 March 2004; accepted 8 March 2004

Abstract

Organizing images into semantic categories can be very useful for searching and browsing through large image repositories. In this work, we use machine learning to associate low level colour representations of digital colour photos with their high level semantic categories. We investigate the redundancy and performance of a number of histogram-based colour image content representations in the context of automatic colour photo categorization using support vector machines. We use principal component analysis to reduce the dimensionality of (high dimensional) histogram based colour descriptors and use support vector machines to learn to classify the images into various high level categories in the histograms subspaces. We present experimental results to demonstrate the usefulness of such an approach to organizing colour photos into semantic categories. Our results show that the colour content descriptors constructed in different ways perform quite differently and the performances are data dependent hence it is difficult to pick a "winning" descriptor. Our results demonstrate conclusively that all descriptors studied in this paper are highly redundant and that regardless of their performances, the dimensionality of these histogram based colour content descriptors can be significantly reduced without affecting their classification performances.

© 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

Keywords: Color histogram; Content-based indexing and retrieval; PCA; SVM

1. Introduction

Fast advancement in digital imaging technology has resulted in the exponential increase of image data in both professional archives and personal leisure collections. Effectively managing large image repositories and making them easily accessible poses significant technical challenges. In the past decade, there has been significant research effort in content-based image retrieval (CBIR) [1]. In a CBIR system, a user can query the image repositories with a visual example and the system will return an ordered list of images that are similar to the query in some visual sense. Traditionally, image similarity is measured by some forms of

distance metrics in the feature space. However, similarity is a subjective concept, it is therefore not surprising and perhaps inevitable that there is a "gap" between a similarity measured by the CBIR systems and a similarity perceived by human observers. How to reduce this gap, often referred to as the "semantic gap" in content-based image retrieval has received much attention in recent years, and technologies that provide solutions to reduce the semantic gap of CBIR are likely to play a pivotal role in content-based image indexing and retrieval.

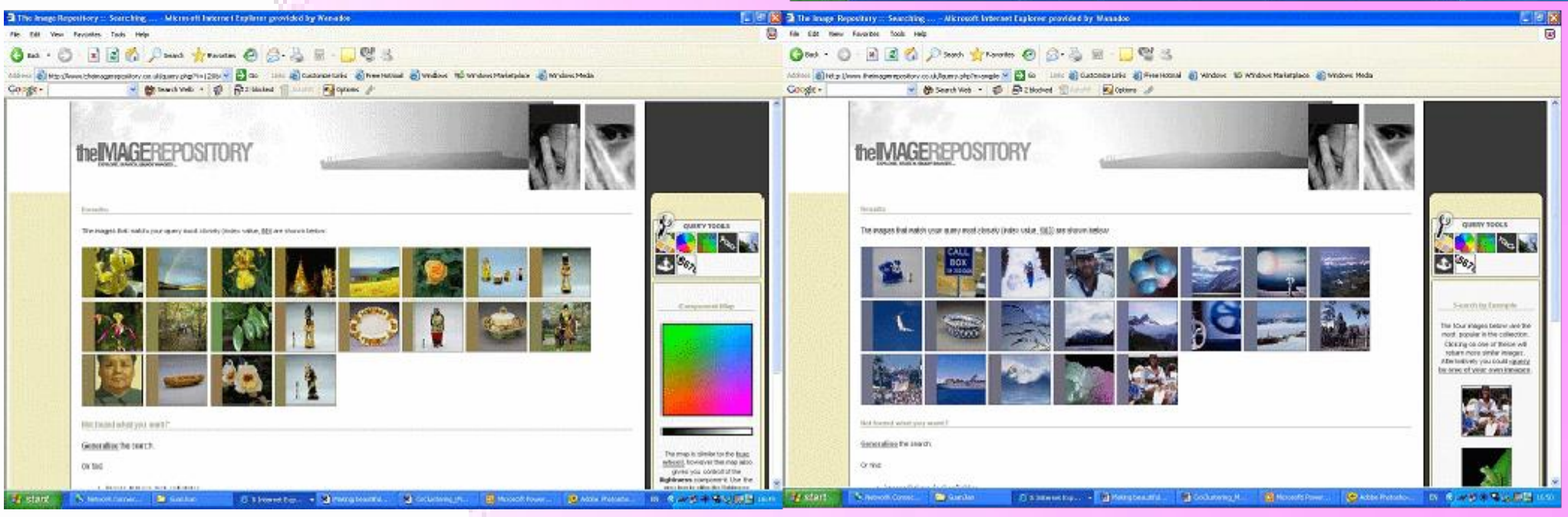
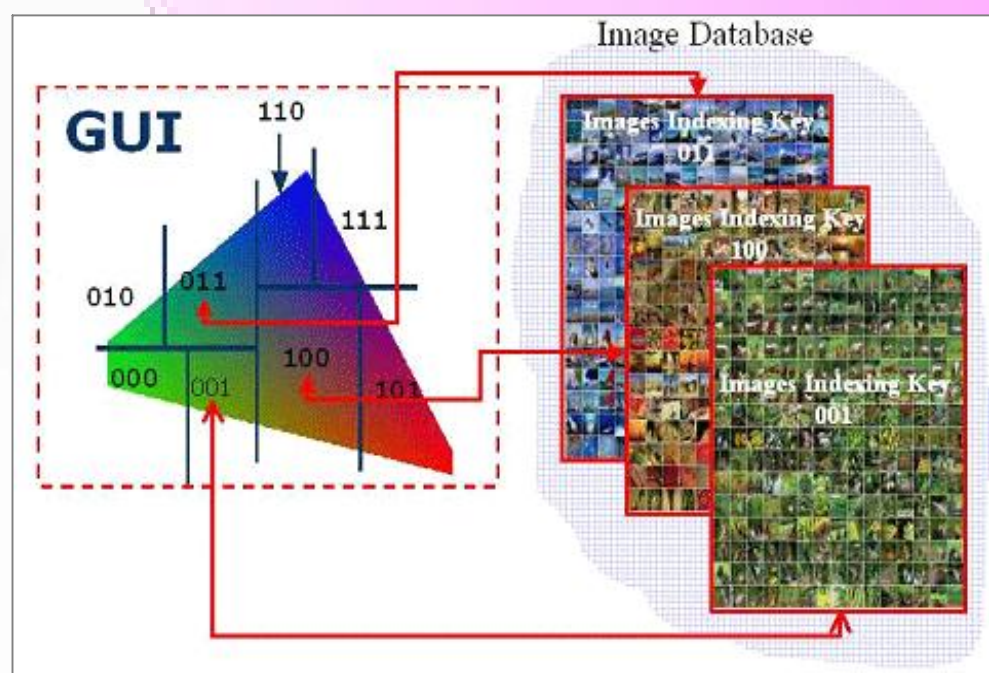
One direction pursued by researchers to narrow the semantic gap is based on the "learning by example" principle. Machine learning is an integral and essential part of human efforts to build intelligent machines, and the subject has been studied extensively in various disciplines of scientific and engineering purposes. In the context of CBIR, machine learning has been applied to classify collections of images into categories or classes of various descriptions [2]. In particular, image classification techniques based on various

* Corresponding author. Tel.: +44115 8466507; fax: +44115 9514254. E-mail address: qiu@cs.nott.ac.uk (G. Qiu).

¹ xia.feng@nott.ac.uk (X. Feng), jf.feng@nott.ac.uk (J. Fang).

² On leave from The Civil Aviation University of China and sponsored by the China Scholarship Council.

0031-3203/\$30.00 © 2004 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patrec.2004.03.006



Conservation laws for two (2+1)-dimensional differential-difference systems

Guo-Fu Yu*

Institute of Computational Mathematics and Scientific Engineering Computing, AMSS, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, P.R. CHINA
Graduate School of the Chinese Academy of Sciences, Beijing, P.R. CHINA
Hon-Wah Tam¹
Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, P.R. CHINA

Abstract

Two integrable differential-difference equations are considered. One is derived from the discrete BKP equation and the other is a symmetric (2+1)-dimensional Lotka-Volterra equation. An infinite number of conservation laws for the two differential-difference equations are deduced.

1. Introduction

In recent years, much attention has been paid to studying a variety of intrinsic features shared by discrete integrable systems. Various methods have been developed to search for new discrete integrable systems, Lax pairs, soliton solutions, symmetries and conservation laws (CLs) etc. (See, e.g. [1]–[4] and references therein). Conservation laws play an important role in mathematics and engineering as well. The general approach to the search for a conservation law for a real-life problem is the variational method, where the Hamiltonian of the problem denotes a conservation law. It is easy to establish a variational formulation for a differential system by the semi-inverse method, but it is difficult to establish variational formulations for differential-difference systems [5]. Applications of the semi-inverse method can be found in the references [6]–[7]. Concerning CLs of differential-difference equations, many results have been achieved by using some successful methods. (See, e.g. [12]–[18].) However, to our knowledge, it seems that most examples of CLs of integrable differential-difference equations given in the literature are just (1+1) or (1+2)-dimensional (one discrete and two continuous) cases [8]–[9]. Comparatively less (2+1)-dimensional (two discrete and one continuous) differential-difference equations have been considered for their CLs.

In this paper, we will derive CLs for the following two (2+1)-dimensional (two discrete and one continuous) differential-difference equations. The first one derives from the famous discrete BKP equation, which in the Hirota bilinear form is expressed by

$$[z_1 \exp(D_1) + z_2 \exp(D_2) + z_3 \exp(D_3) + z_4 \exp(D_4)] f \cdot f = 0, \quad (1)$$

where D_1, D_2, D_3, D_4 and z_1, z_2, z_3, z_4 are bilinear operators and constants, respectively, satisfying

$$D_1 + D_2 + D_3 + D_4 = 0, \quad z_1 + z_2 + z_3 + z_4 = 0.$$

If we choose

$$D_1 = \frac{1}{2}(\delta D_2 + \epsilon D_2), \quad z_1 = 1, \quad D_2 = \frac{1}{2}(\delta D_2 - \epsilon D_2), \quad z_2 = -1 + \beta \delta \epsilon, \\ D_3 = -D_4 = \frac{1}{2}(\delta D_2 + \epsilon D_2), \quad z_3 = -\alpha \epsilon - \beta \delta \epsilon, \quad D_4 = D_2 = \frac{1}{2}(\delta D_2 - \epsilon D_2), \quad z_4 = \alpha \epsilon,$$

*Electronic mail: gyu@lsec.cc.ac.cn
¹Electronic mail: tam@comp.hkbu.edu.hk

INSTITUTE OF PHYSICS PUBLISHING
J. Phys. A: Math. Gen. 39 (2006) 3367–3373
doi:10.1088/0305-4470/39/13/014

JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL
doi:10.1088/0305-4470/39/13/014

On the nonisospectral Kadomtsev–Petviashvili equation

Guo-Fu Yu^{1,2} and Hon-Wah Tam³

¹ Institute of Computational Mathematics and Scientific Engineering Computing, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, PO Box 2719, Beijing 100080, People's Republic of China
² Graduate School of the Chinese Academy of Sciences, Beijing, People's Republic of China
³ Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong, People's Republic of China

Received 8 December 2005
Published 15 March 2006
Online at stacks.iop.org/JPhysA/39/3367

Abstract

In this paper, we first present the Gramman determinant solutions to the nonisospectral Kadomtsev–Petviashvili (KP) equation. Then, by using the Pfaffianization procedure of Hirota and Ohta, an integrable coupled system is generated. Moreover, Gramm-type Pfaffian solutions to the Pfaffianized system are proposed.

PACS numbers: 02.30.Ik, 02.30.Jr, 05.45.Yv
Mathematics Subject Classification: 35Q58, 37K40

1. Introduction

In the early 1990s, Hirota and Ohta [1, 2] developed a procedure for generalizing equations from the Kadomtsev–Petviashvili (KP) hierarchy to produce coupled systems of equations, which we now call Pfaffianization. These Pfaffianized equations appear as coupled systems of the original equations and have soliton solutions expressed by Pfaffians. Such a procedure has been successfully applied to the DS equations [3], the discrete KP equation [4], the self-dual Yang–Mills equation [5], the two-dimensional Toda lattice [6], the semi-discrete Toda equation [7], the differential-difference KP equation [8], etc.

In [9], Wronskian solutions of the nonisospectral KP equation [10]
$$4u_t + y(u_{x+2} + 6u_{xy} + 3y^{-1}u_{yy}) + 2xu_y + 4y^{-1}u_y = 0 \quad (1)$$
are derived by the Hirota method and the Wronskian technique. The solutions of the bilinear nonisospectral KP equation are expressed in Wronskian determinants. Since there are both Wronskian determinant and Gramman determinant solutions for the KP equation [11], we expect that there also exist Gramm-type expressions for (1). On the other hand, all the Pfaffianization procedures above are applied to isospectral systems. Hence it would be very interesting to consider Pfaffianization for nonisospectral systems.

Spectral radius analysis of matrices and their association with integrable systems*

Honwah Tam¹ and Yufeng Zhang^{1,2}

¹ Department of Computer Science, Hong Kong Baptist University, Hong Kong, P.R.China;

² Information School, Shandong University of Science and Technology, Qingdao Huangdao 266510, P.R.China

Abstract

This paper begins with an isospectral problem and analyzes the spectral radius of its corresponding spectral matrix. This work enlightens us to set up a higher-dimensional isospectral problem whose compatibility condition gives rise to a (2+1)-dimensional zero curvature equation. From this equation a (2+1)-dimensional Lax integrable soliton equation hierarchy with constraints of potential functions, along with 5 parameters, is generated. The reduced cases of this hierarchy give three (2+1)-dimensional integrable systems, namely, the AKNS hierarchy, the Levi hierarchy and the D-AKNS hierarchy. Extending the above Lie algebra into more complicated ones, two integrable couplings of the (2+1)-dimensional hierarchy are derived. One of the two couplings has Hamiltonian structure by employing the quadratic-form identity. The corresponding integrable couplings of the reduced systems are also obtained. Finally, as a comparison study for generating expanding integrable systems, an antisymmetric Lie algebra and its corresponding loop algebra are constructed. From this Lie algebra a great many enlarging integrable systems can be generated. In addition, their Hamiltonian structures can be computed by the trace identity.

Keywords: spectral radius analysis, integrable couplings, Lie algebra, Hamiltonian structure

*This work was supported by the Hong Kong Research Grant Council under grant number HKBU2016/05P and the National Science Foundation of China under grant number 10471139.

