Hui Xiong

MSIS Department
Rutgers University
180 University Avenue
Newark, NJ 07102

Voice: (973) 353-5261
Fax: (973) 353-5003
Email: hui@rbs.rutgers.edu
WWW: http://cimic.rutgers.edu/~hui

RESEARCH INTERESTS

Data Mining, Statistical Computing, Geographic Information Systems, Biomedical Informatics, and Information Security.

EDUCATION

Ph.D., Computer Science, University of Minnesota - Twin Cities, 2005.
Ph.D. Minor, Statistics, University of Minnesota - Twin Cities
M.S., Computer Science, National University of Singapore, 2000
B.E., Automation, University of Science and Technology of China, 1995

HONORS

- Excellence in Research Recognition Award, 2005, University of Minnesota - Twin Cities
- Travel Fellowship, the National Library of Medicine and the Department of Energy, 2005, Pacific Symposium on Biocomputing
- KDD Student Travel Award, 2004, ACM SIGKDD International Conference
- SDM Student Travel Award, 2004, SIAM International Conference on Data Mining
- Excellent Student Scholarship, 2000, National University of Singapore
- Excellent Student Scholarship, 1995, University of Science and Technology of China

PROFESSIONAL EXPERIENCE

Assistant Professor, Management Science & Information Systems
Rutgers University, August 2005 - Present.

Research Co-op, IBM TJ Watson Research Center, Summer 2005

Research Assistant, University of Minnesota - Twin Cities, 2000 - 2005

- Developed techniques for mining hyperclique patterns and showed that hyperclique patterns tend to capture strongly-related objects.
- Proposed a framework for efficiently identifying spatial co-location patterns.
- Developed techniques for efficiently computing all-pairs correlations.
- Proposed the concept of pattern preserving clustering.
- Demonstrated the problem of privacy leakage in database views via semi-supervised learning.

Research Intern, Lawrence Berkeley National Laboratory, Summer 2004

- Explored hyperclique pattern discovery techniques for extracting protein functional modules in large-experimentally determined protein complexes.
- Developed techniques for clique formation through transitive closure in protein networks.

Research Assistant, National University of Singapore, 1999 - 2000

- Designed and implemented server based agent for web usage mining.

Software Engineer, Elite Business Machines Manufacturing Co. Ltd (EBM), Shenzhen, Guangdong province, P.R.China, **07/1995 - 10/1998.**

- Developed a software package for computer-based POS (Point of Sales) using C combined with 8086 assembly language, designed the embedded software (Using C/8051) for Electrical Typewriters (Olympia TW88, TW89).

Research Assistant, University of Science and Technology of China, 1993 - 1995

- Developed communication components based on TCP/IP using C++ for the project titled *Simulation & Training System of TDC3000 Distributed Control System.*

TEACHING EXPERIENCE

Instructor, Rutgers University, Spring 2006

- Course: "Computer Network Applications (29:623:375)".
- Course: "Management Information Systems (29:623:220)".

Instructor, Rutgers University, Fall 2005

- Course: "Computer Network Applications (29:623:375)".
- Course: "Management Information Systems (29:623:220)".

Mentor, University of Minnesota - Twin Cities, Spring 2004

- Course: "CSci5980: Data Mining".

PUBLICATIONS[1]

Book

1. Weili Wu, Hui Xiong, Shashi Shekhar (Eds.), Clustering and Information Retrieval , ISBN: 1-4020-7682-7, Kluwer Academic Publishers, 2003.
2. Shashi Shekhar and Hui Xiong (Editor-in-Chiefs), Encyclopedia of Geographical Information Science (The book web site, http://refworks.springer-sbm.com/geograph/, maintained by Springer). In Preparation, publication by Springer is planned for July 2007.

Journal Papers

3. Hui Xiong, Shashi Shekhar, Pang-Ning Tan, Vipin Kumar, TAPER: A Two-Step Approach for All-strong-pairs Correlation Query in Large Databases, IEEE Transactions on Knowledge and Data Engineering (**TKDE**), accepted for publication as a regular paper, 2006.
4. Hui Xiong, Gaurav Pandey, Michael Steinbach, Vipin Kumar, Enhancing Data Analysis with Noise Removal, IEEE Transactions on Knowledge and Data Engineering (**TKDE**), accepted for publication as a regular paper, 2006.
5. Sam Yuan Sung, Yao Liu, Hui Xiong, Peter Ng, Privacy Preservation for Data Cubes, Knowledge and Information Systems - An International Journal (**KAIS**), to appear, 2006.
6. Yao Liu, Sam Y. Sung, Hui Xiong, A Cubic-Wise Balance Approach for Privacy Preservation in Data Cubes, **Information Sciences**, to appear, 2006.

[1] A more up-to-date list of publications is available at my homepage http://cimic.rutgers.edu/∼hui

7. Yan Huang, Jian Pei, and Hui Xiong, Mining Co-location Patterns in Spatial Data Sets with Rare Events, **GeoInformatica** - An International Journal on Advances of Computer Science for Geographic Information Systems, accepted for publication as a regular paper, 2006.
8. Jieping Ye, Qi Li, Hui Xiong, Haesun Park, Ravi Janardan, and Vipin Kumar. IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition, IEEE Transactions on Knowledge and Data Engineering (**TKDE**), Special Issue - Intelligent Data Preparation, 17(9), pp. 1208-1222, September 2005.
9. Yan Huang, Shashi Shekhar, and Hui Xiong (Corresponding Author), Discovering Co-location Patterns from Spatial Datasets: A General Approach, IEEE Transactions on Knowledge and Data Engineering (**TKDE**), 16(12), pp. 1472 - 1485, December 2004.

Conference Papers

10. Yaochun Huang, Hui Xiong, Weili Wu, Sam, Y. Sung, Mining Quantitative Maximal Hyperclique Patterns: A Summary of Results, Pacific-Asia Conference on Knowledge Discovery and Data Mining (**PAKDD**), to appear, 2006. (19% accepted)
11. Gyorgy Simon, Hui Xiong, Eric Eilertson, and Vipin Kumar, Scan Detection: A Data Mining Approach, SIAM International Conf. on Data Mining (**SDM**), to appear, 2006. (16% accepted)
12. Hui Xiong, Michael Steinbach, and Vipin Kumar, Privacy Leakage in Multi-relational Databas es via Pattern based Semi-supervised Learning, ACM Conference on information and Knowledge Management (**CIKM**), 2005, to appear.
13. Hui Xiong, Xiaofeng He, Chris Ding, Ya Zhang, Vipin Kumar, Stephen R. Holbrook, Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery, in Proc. of the Pacific Symposium on Biocomputing (**PSB**), 2005. (25% accepted)
14. Hui Xiong, Shashi Shekhar, Pang-Ning Tan, and Vipin Kumar, Exploiting a Support-based Upper Bound of Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs, in Proc. of the Tenth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (**KDD**), pp. 334 - 343, Seattle, USA, 2004. (12% accepted)
15. Yaochun Huang, Hui Xiong, Weili Wu, and Zhongnan Zhang, A Hybrid Approach for Mining Maximal Hyperclique Patterns, in Proc. of the 16th IEEE Int'l Conf. on Tools with Artificial Intelligence (**ICTAI**), pp. 354 - 361, Florida, USA, 2004. (25% accepted)
16. Jieping Ye, Qi Li, Hui Xiong, Haesun Park, Ravi Janardan, Vipin Kumar. IDR/QR: An Incremental Dimension Reduction Algorithm via QR Decomposition, in Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD**), pp. 364 - 373, Seattle, USA, 2004. (12% accepted)
17. Michael Steinbach, Pang-Ning Tan, Hui Xiong, Vipin Kumar, Extending the Notion of Support, in Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (**KDD**), pp. 689 - 694, Seattle, USA, 2004. (Poster Paper, 25% accepted)
18. Hui Xiong, Michael Steinbach, Pang-Ning Tan, and Vipin Kumar, HICAP: Hierarchical Clustering with Pattern Preservation, in Proc. 2004 SIAM International Conference on Data Mining (**SDM**), pp. 279 - 290, Florida, USA, 2004. (21% accepted)
19. Hui Xiong, Shashi Shekhar, Yan Huang, Vipin Kumar, Xiaobin Ma, Jin-Soung Yoo, A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects, in Proc. 2004 SIAM International Conference on Data Mining (**SDM**), pp. 78 - 89, Florida, USA, 2004. (21% accepted)

20. Yao Liu, Sam Y. Sung, Hui Xiong, Peter Ng, Data Declustering with Replication, in Proc. of the 9th International Conference on Database Systems for Advanced Applications (**DASFAA**), pp. 682 - 693, Korea, 2004. (22% accepted)
21. Hui Xiong, Pang-Ning Tan, and Vipin Kumar, Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution, In Proc. of the Third IEEE Int'l Conf. on Data Mining (**ICDM**), pp. 387-394, Melbourne, Florida, USA, 2003. (12% accepted)
22. Jinmei Xu, Hui Xiong, Sam Y. Sung, and Vipin Kumar, A New Clustering Algorithm for Transaction Data Via Caucus, in Proc. 2003 Pacific-Asia Conf. on Knowledge Discovery and Data Mining (**PAKDD**), pp. 551-562, Korea 2003. (18% accepted)
23. Yan Huang, Hui Xiong, Shashi Shekhar, and Jian Pei, Mining Confident Co-location Rules without a Support Threshold, in Proc. 2003 ACM Symposium on Applied Computing (**ACM SAC**), pp. 497-501, Melbourne, FL, March 2003. (26% accepted).
24. Hui Xiong, Sam Y. Sung, and S. Huang, Adapting the Right Web Pages to the Right Users, in Proc. of the First SPIE Int'l Conf. on Data Mining and Knowledge Discovery, USA, 2000.

PROFESSIONAL ACTIVITIES

Professional Affiliations

- Member of ACM, ACM SIGKDD, IEEE, the IEEE computer society, and Sigma Xi.

Conference Organizers

- Publicity Chair, the Sixth IEEE International Conference on Data Mining (ICDM), 2006.
- Program Committee Vice Chair, the Ninth Pacific Rim International Conference on Artificial Intelligence (PRICAI), 2006

Program Committees

- The Sixth SIAM International Conference on Data Mining (SDM), USA, 2006.
- The 10th Pacific-Asia International Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore, 2006.
- ICDM Workshop on Temporal Data Mining: Algorithms, Theory and Applications, 2005.
- ISPRS Workshop on Spatial/Spatio-Temporal Data Mining and Learning, Turkey, 2005.

Presentations

- Identification of Functional Modules in Protein Complexes via Hyperclique Pattern Discovery, the Pacific Symposium on Biocomputing, (PSB 2005), the Fairmont Orchid, Big Island of Hawaii, USA, 2005.
- Exploiting a Support-based Upper Bound of Pearson's Correlation Coefficient for Efficiently Identifying Strongly Correlated Pairs, the Tenth ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2004), Seattle, USA, 2004.
- Hyperclique Pattern Discovery and Its Application to Protein Functional Module Extraction, Lawrence Berkeley National Laboratory (LBL), CA, USA, July 2004.
- HICAP: Hierarchical Clustering with Pattern Preservation, the Fourth SIAM International Conference on Data Mining (SDM), Lake Buena Vista, Florida, USA, April 2004.

- A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects, the Fourth SIAM International Conference on Data Mining (SDM), Lake Buena Vista, Florida, USA, April 2004.
- Mining Strong Affinity Association Patterns in Data Sets with Skewed Support Distribution, the Third IEEE Int'l Conf. on Data Mining (ICDM), Florida, USA, November 2003.

Referee

- Conferences
  SIGKDD, SDM, ICDM, ICML, ACMGIS, ICDE, PAKDD, VECPAR, ICPP, WI, ICCSA
- Journals
  IEEE Transactions on Knowledge and Data Engineering (TKDE)
  Data Mining and Knowledge Discovery Journal
  Knowledge and Information Systems (KAIS)
  International Journal on Digital Libraries (IJDL)
  Data and Knowledge Engineering
  Geoinformatica
  International Journal of Image and Graphics
  Multimedia Tools and Applications Journal
  Journal of Computer science and Technology
  Journal of Zhejiang University Science
  Journal of Systems Science and Systems Engineering

- NSF ITR/IDM Proposal Review

RESEARCH PROJECTS

Hyperclique Pattern Discovery (Data Mining)

- A hyperclique pattern is a type of association pattern containing objects that are highly affiliated with each other; that is, every pair of objects within a hyperclique pattern is guaranteed to have a cosine similarity (uncentered correlation coefficient) above a certain level. Discovering such patterns is extremely useful for a variety of applications such as identifying low-support-high-confidence association rules, rare event detection, and protein functional module extraction. A key idea central to the discovery of hyperclique patterns is the use of an association measure called h-confidence. The h-confidence measure possesses an anti-monotonic property that allows us to incorporate the measure directly into the mining process, rather than using it during post-processing. Furthermore, the h-confidence measure has another important property, called the cross-support property, which allows dramatic pruning of the exponential search space by eliminating patterns involving items from different levels of support. Due to these two properties, hyperclique patterns can be found even when the support threshold is set to zero. A summary of the preliminary work has been published in *the Third IEEE International Conference on Data Mining (ICDM 2003)* and the extended version of this work has been submitted to the journal *Data Mining and Knowledge Discovery*.

Protein Functional Module Extraction (Bioinformatics)

- Proteins usually do not act in isolation but function within complicated cellular pathways, interacting with other proteins either in pairs or as components of larger complexes. While many protein complexes have been identified by large-scale experimental studies, due to a large number of false-positive interactions existing in current protein complexes, it is still difficult to obtain an accurate understanding of functional modules, which encompass groups

of proteins involved in common elementary biological function. In this project, we present a hyperclique pattern discovery approach for extracting functional modules (hyperclique patterns) from protein complexes. The analysis of hyperclique patterns using the Gene Ontology suggests that proteins within the same hyperclique pattern more likely perform the same function and participate in the same biological process. More interestingly, the 3-D structural view of proteins within a hyperclique pattern reveals that these proteins physically interact with each other. In addition, we observe that several hyperclique patterns corresponding to different functions can participate in the same protein complex as independent modules; and a hyperclique pattern can be involved in different complexes performing different higher-order biological functions, although the pattern corresponds to a specific elementary biological function. Finally, the results also indicate that our method can facilitate the functional annotation of uncharacterized proteins. (More detailed results are available at the project web site: http://www.cs.umn.edu/∼huix/pfm/pfm.html). A summary of this work has been published in the *Pacific Symposium on Biocomputing 2005 (PSB 2005).*

Spatial Co-location Patterns (Spatial Databases/Spatial Data Mining)

- Given a collection of Boolean spatial features, the co-location pattern discovery process finds the subsets of features frequently located together. Co-location patterns are important for many application domains involving large geographical datasets such as location-based E-services, epidemiology, and NASA's climatologic project. With no notion of transaction in continuous geographic space, traditional measures and mining algorithms are difficult to apply. Instead, we proposed the concept of user-specified neighborhoods in place of transactions to specify groups of items. New interest measures were also proposed which are robust in the face of potentially infinite overlapping neighborhoods. The full version of this work has been published in *the IEEE Transaction on Knowledge and Data Engineering (TKDE).* In addition, we proposed a novel measure called the maximal participation index and developed efficient algorithms to find high-confidence-low-prevalence co-location rules. This work was published in *the ACM Symposium on Applied Computing (ACM SAC 2003)* and the extended version of this paper has been accepted for publication in *the GeoInformatica, An International Journal on Advances of Computer Science for Geographic Information Systems.* Finally, we are continuing work on this project and extending the concept of co-locations from point features to extended spatial objects (e.g. polygons and line strings). This research has resulted in the paper "A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects," published in *the Fourth SIAM International Conference on Data Mining (SDM 2004).*

Pattern Preserving Clustering (Data Mining)

- This project proposes a new approach for clustering—pattern preserving clustering—which produces more easily interpretable and usable clusters. This approach is motivated by the following observation: while there are usually strong patterns in the data—patterns that may be key for the analysis and description of the data—these patterns are often split among different clusters by current clustering approaches. This is perhaps not surprising since clustering algorithms have no built in knowledge of these patterns and may often have goals that are in conflict with preserving patterns, e.g., minimize the distance of points to their nearest cluster centroid. Also, patterns are typically overlapping, i.e., may involve some of the same objects, and if the clustering algorithm produces disjoint clusters, then some patterns must be split when the objects are clustered. In this project we describe a technique for pattern preserving clustering that first finds patterns composed of tightly connected groups of objects or attributes and then, starting from these patterns, performs agglomerative clustering using the Group Average (UPGMA) technique. We present the results of some experiments on doc-

ument data that compare our approach, HIerarchical Clustering with PAttern Preservation (HICAP), to two other clustering techniques: bisecting K-means and traditional UPGMA. These results show that, despite the extra constraint of pattern preservation, HICAP has performance very much like traditional UPGMA with respect to the cluster evaluation criteria of entropy and F-measure. More importantly, we also illustrate how patterns, if preserved, can aid cluster interpretation. A summary of the preliminary work has been published in *the Fourth SIAM International Conference on Data Mining (SDM 2004).*

Server Based Agent for Web Usage Mining (Web Mining)

- To perform web usage analysis, unique user sessions must be identified. There are heuristics that can help identify user sessions on web logs from the traditional web server log mechanism; however the identified-user sessions using heuristics are often incomplete and inaccurate. Instead, we designed a server-based agent to capture user sessions explicitly at the server end and construct a new web log, which is more suitable for web usage mining tasks. A summary of the preliminary work has been published in *the First SPIE International Conference on Data Mining and Knowledge Discovery.*

Privacy Leakage in Databases via Semi-supervised Learning (Database Security)

- In multi-relational databases, a view, which is a context- and content-dependent subset of one or more tables (or other views), is often used to preserve privacy by hiding sensitive information. However, recent developments in data mining present a new challenge for database security even when traditional database security techniques, such as data access control via a view, are employed. This project presents a data mining framework using semi-supervised learning that demonstrates the potential for privacy leakage in multi-relational databases. Many different types of semi-supervised learning techniques, such as K-nearest neighbor (KNN) methods, can be used to demonstrate privacy leakage. However, we also introduce a new approach to semi-supervised learning, hyperclique pattern based semi-supervised learning (HPSL), which differs from traditional semi-supervised learning approaches in that it considers the similarity among groups of objects instead of only pairs of objects. Our experimental results show that both the KNN and HPSL methods have the ability to compromise database security, although HPSL is better at this privacy violation (has higher accuracy) than KNN methods. A summary of this work has been published in the *ACM Conference on information and Knowledge Management (CIKM 2005).*

All-Pairs Correlation Computing (Databases/Data Mining/Statistical Computing)

- With the wide spread use of statistical techniques for data analysis, it is expected that many such techniques will be made available in a database environment where users can apply the techniques more flexibly, efficiently, easily, and with minimal mathematical assumptions. The motivation of this project is directed towards developing such techniques. More specifically, given a user-specified minimum correlation threshold θ and a market basket database with N items and T transactions, an all-strong-pairs correlation query finds all item pairs with correlations above the threshold θ. However, when the number of items and transactions are large, the computation cost of this query can be very high. In this project, we identify an upper bound of Pearson's correlation coefficient for binary variables. This upper bound is not only much cheaper to compute than Pearson's correlation coefficient but also exhibits a special monotone property which allows pruning of many item pairs even without computing their upper bounds. A Two-step All-strong-Pairs corrElation queRy (TAPER) algorithm is proposed to exploit these properties in a filter-and-refine manner. Furthermore, we provide an algebraic cost model which shows that the computation savings from pruning is independent or improves when the number of items is increased in data sets with common Zipf or linear

rank-support distributions. Experimental results from synthetic and real data sets exhibit similar trends and show that TAPER can be several orders of magnitude faster than brute-force alternatives. A summary of the preliminary work has been published in *the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)* and the extended version of this work has been accepted for publication in the *IEEE Transactions on Knowledge and Data Engineering (TKDE).*

Systematic Development and Exploration of Data Cleaning Techniques for Enhancing Data Analysis (Data Mining/Bioinformatics)

- Data cleaning is an essential tool for ensuring the reliable analysis of noisy data sets. Most existing data cleaning methods primarily focus on the detection and/or correction of low-level data errors that result from an imperfect data gathering process. However, the presence of data objects that are irrelevant or only weakly relevant can significantly hinder data analysis, and these objects should also be considered as noise. Thus, there is a need for a systematic development and evaluation of data cleaning techniques that also address data sets with irrelevant or weakly relevant objects. This is particularly important for data sets with large amounts of noise. To that end, this paper studies the performance of four techniques intended for data cleaning to enhance data analysis in the presence of high noise levels. Three of these methods are based on traditional outlier detection techniques: distance-based, clustering based, and an approach based on the Local Outlier Factor (LOF) of an object. The other technique, which is a new method that we are proposing, is a hyperclique-based data cleaner (HCleaner). We also present a framework for measuring data cleaning performance in terms of its impact on the subsequent data analysis. We use this framework to evaluate the four techniques we described in terms of their impact on clustering and association analysis for document and gene expression data sets. Our experimental results show that using HCleaner generally provides better clustering performance and significantly higher quality association rules as compared to the outlier based data cleaning alternatives. The other techniques sometimes performed as well or slightly better for clustering, but their performance was not as consistent. For instance, the clustering based technique had good performance only when the number of clusters specified matched the actual number of classes in the data. However, this limitation significantly restricts the usefulness of this method. A summary of this work has been accepted for publication in the *IEEE Transactions on Knowledge and Data Engineering (TKDE).*

REFERENCE UPON REQUEST