Editorial

P. S. Bradley
digiMine, Inc
10500 NE 8th Street
Bellevue, WA 98004
U.S.A.
paulb@digiMine.com

Usama Fayyad
digiMine, Inc
10500 NE 8th Street
Bellevue, WA 98004
U.S.A.
fayyad@acm.org

S. Sarawagi
School of Information Tech.
IIT Bombay
Powai Mumbai-400076
India
sunita@it.iitb.ernet.in

K. Shim
Computer Science Department
KAIST
373-1 Kusong-dong, Yu-song-gu
Korea
shim@cs.kaist.ac.kr

The Challenge is Growing

According to recent studies on the tremendous increase in the amount of data recorded and stored on digital media, it is clear that the gap between this digitally stored data and our ability to process it is widening. The primary driver for this phenomenon is the more aggressive relative of the familiar "Moore's Law". Moore's Law is the observation that computational power doubles approximately every 18 months. The corresponding law in the world of digital storage is even more amazing. Storage capacity, for a fixed price, appears to be doubling approximately every 9 months! [1]. In a world governed by these two empirical laws, the importance of the fields of databases and Data Mining and Knowledge Discovery are growing. From a strategic perspective, our need to navigate the rapidly growing universe of digital data will rely heavily on our ability to come up with the right systems that can scale with these growth rates, allowing us to effectively manage and mine the raw data.

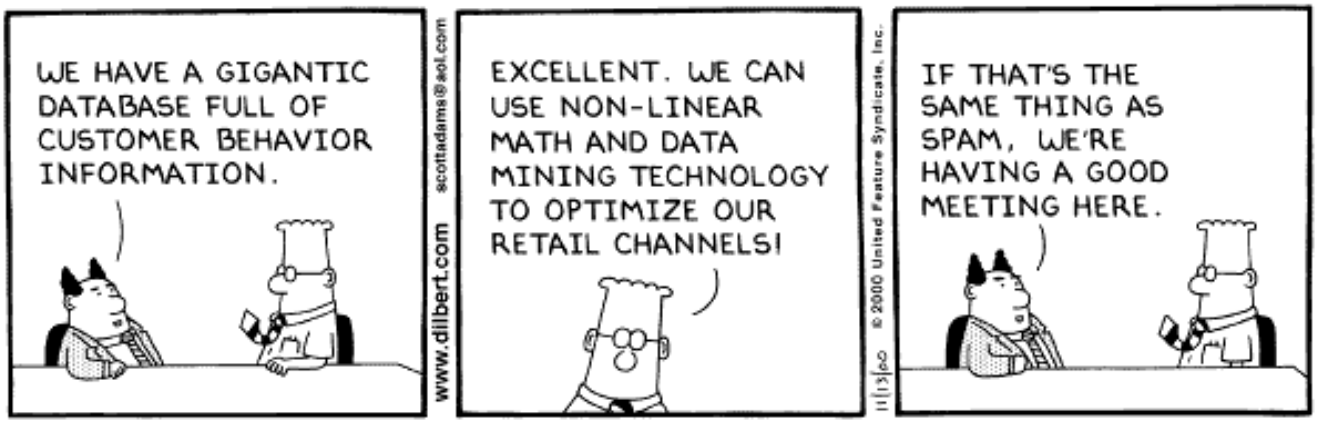
Data Mining centers about developing algorithms for extracting structure from data. This structure can take the form of statistical patterns, models, and relationships. Viewed at a high level, these "structures" are basic *data reductions* that enable us to summarize the data, to bring it down in dimensionality to spaces in which we have some hope of understanding and visualizing. These structures provide the basis for us to predict and anticipate when certain events occur. When viewed at this level, one begins to understand the fundamental importance of data mining in the future of the digital data universe. For a more futuristic outlook, see the article [2].

Beyond the Giants and the Monsters

A recent study conducted at UC-Berkeley indicates that humankind seems to be producing over two exabytes of information per year [3]. This study confirms the persistence of the original trends first uncovered in [4]. The amount of information generated in 1999 constitutes more than 12% of the aggregate amount of information generated by humankind in all of recorded history. We conjecture that the rate is more dramatic for the year 2000. It is interesting that an overwhelming majority of collected data is never seen by humans [4].

As recently as three years ago, people were wondering how they would manage terabyte databases. Jim Gray, who is familiar with systems and scalability issues when dealing with large amounts of data, used to refer to these as "Terror Byte" databases. It is interesting that we have now gone beyond gigabytes and terabytes, and are now worried about petabytes and exabytes. Interestingly, "Giga" derives from the Latin term for "Giant", and "Tera" derives from the Greek term for "Monster". A gigabyte is 10^9 bytes, while a terabyte is 10^{12} bytes. The higher units have less colorful names: petabyte and exabyte. The next term is the zettabyte (deriving from the last letter of the Greek alphabet), then the yottabyte, indicating a backtracking from the last letter.

In the Data Mining field, it is fairly safe to think that we are wrestling with the issues in dealing with giants (gigabytes). We certainly have not stepped up to the monsters yet. It is worrisome that within this plethora of names we have few ideas on how to approach the monsters. The terabytes are still indeed the "terror bytes". However, given the driving forces of technology out there, it is with urgency that we must start to see progress in scaling to



DILBERT reprinted by permission of United Feature Syndicate, Inc. [2000].

larger databases. With each year, the challenges will grow larger, and so will the corresponding pressure for breakthrough advancement in data mining technologies. Data mining solution providers will be measured along the lines of successful real world applications. Integrating with new generation business practices is essential.

Corporations currently are and have been collecting business data. Furthermore, with the growth of E-commerce and the World Wide Web, online stores are automatically collecting huge amounts of valuable data. For example, web servers typically generate large volumes of daily access logs, capturing page sequences for millions of users. Sites can easily generates tens of gigabytes of data per day. Thus, we easily witness the wide availability of huge amounts of data; data that was hardly there three or four years ago. And the volume of this data grows significantly every year.

However, it is very difficult to turn such data into useful information due to the limited availability of scalable data mining algorithms. For algorithms to be scalable, the execution time should increase linearly, or preferably sub-linearly, in proportion to the size of the input data. Traditional data analysis algorithms generally do not scale well with limited system resources (e.g. main memory). Therefore, we are now faced with exciting opportunities and great challenges for data mining to develop efficient, scalable algorithms. This "scalability challenge" goes beyond algorithms to include: systems issues, the delivery of services, and the delivery of mining results. Results need to be made usable and useful for businesses and organizations knowing little about data mining. This challenge must be addressed for data mining to cross into its next level of integration within mainstream business and daily operations of organizations.

Special Issue on "Scalable Data Mining Algorithms"

We focus this issue of SIGKDD Explorations on the topic of Scalable Data Mining Algorithms, and we welcome **Kyuseok Shim** as **Guest Editor**. On this challenging and exciting topic, eleven articles are included in the section on scalable data mining algorithms.

Kristin P. Bennett and Colin Campbell contributed a nice tutorial article on support vector machines that have become popular tools for data mining applications. Jiawei Han and Jian Pei present a new methodology called frequent pattern growth that avoids candidate generation in Apriori-style algorithms. This technique can be an alternative solution for mining frequent patterns when the candidate generation and test cost is very expensive in applications with long and numerous patterns. Laks. V.S. Lakshmanan, Carson Kai-Sang Leung, and Raymond Ng describe a data structure called segment support map that boosts the performance of frequent itemset mining algorithms by computing sharper upper bounds on the support and better exploitation of constraints. Wei Wang, Jiong Yang and Philip S. Yu propose several models to address different types of noise and provide scalable algorithms for each model. George Forman and Bin Zhang present a technique for developing parallel versions of centroid-based clustering algorithms that communicate only through sufficient statistics. Approaches for exploiting and pushing user-defined model constraints to speed up the mining process are discussed by Minos Garofalakis and Rajeev Rastogi. Such techniques can help to improve the system performance significantly by focusing on only interesting patterns by users. Thus, it is worthy of investigation for many existing data mining algorithms. Claudia Diamantini and Maurizio Panti show that compression capabilities of the Vector Quantizer architecture can be exploited to find simple and effective classification rules when it is trained using a stochastic gradient algorithm. William A. Maniatty and Mohammed J. Zaki address desirable design features of parallel data mining algorithms and investigate a migration path from sequential algorithms by considering an integrated hardware and software approach. Yves Bastide et al. present the PASCAL algorithm, which improves Apriori-style algorithms based on pattern counting inference and can reduce database access significantly. Gregory Piatetsky-Shapiro and Sam Steingold propose a new measure called L-quality that helps to compare predictive models by looking at the entire lift curve. Stephen D. Bay et al. describe new data sets that are available in the UCI KDD archive to help in evaluating scalable data mining algorithms. Considering it is very difficult to get large real-life data sets to conduct performance studies of scalable algorithms, this addition of large data sets to UCI KDD Archive will stimulate and foster systematic progress significantly in developing scalable data mining algorithms.

KDD-2000 Reports

Also included in this issue of Explorations are a series of reports from KDD-2000: The Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD-2000 took place in Boston, MA on August 20 – 23, 2000. The organizing committee expected 700 attendees. After the final numbers were tabulated, a total of 920 people attended! Enthusiasm and interest in the field is rapidly increasing.

The KDD Cup 2000 annual data mining competition consisted of five questions to be answered based on web-browsing data collected from an online retailer. Ron Kohavi, Carla Brodley, Brian Frasca, Llew Mason, and Zijan Zheng provide an excellent report on the competition including summary reports by the winners of the five question areas.

KDD-2000 conference highlights included workshops and panel presentations. Marko Grobelnik, Dunja Mladenic, and Natasa Milic-Frayling have provided a report on the KDD-2000 Workshop on Text Mining. Simeon Simoff and Osmar Zaïane report on MDM/KDD2000: The 1st International Workshop on Multimedia Data Mining. A summary of the WEBKDD 2000 workshop was provided by Myra Spiliopoulou, Jaideep Srivastava, Ron Kohavi and Brij Masand. A report on the Workshop on Distributed and Parallel Knowledge Discovery by Hillol Kargupta, Joydeep Ghosh, Vipin Kumar, and Zoran Obradovic is included. Ivan Bruha and A. Famili have provided a report on the Workshop on Postprocessing in Machine Learning and Data Mining. Alexander Tuzhilin summarized the Panel on Personalization and Data Mining.

We thank the organizers for the time and effort required to provide these reports.

Concluding Remarks

We hope that SIGKDD Explorations continues to meet your expectations on covering timely topics and events in this growing field. Our shift in format towards special focus topics is not intended to limit the scope of our coverage of the field. We still welcome articles and contributions, including position and idea papers as well as controversial topics that stimulate discussion and

deliberation. We hope the special focus sections will allow us to continue the general breadth of coverage while providing some additional depth along the focus topic.

We urge interested readers and contributors to propose special topics. We welcome guest editors from within the field and from outside it to help us cover related focus topics. Please contact us with ideas for future topics of interest, and any feedback on *Explorations*.

KDD-2001 in San Francisco, California, USA

We conclude with a reminder that the premier event in the SIGKDD community, KDD-2001, will take place in San Francisco, CA, beginning Sunday, August 26th and concluding on August 29th. Conference details are available at http://www.acm.org/sigkdd/kdd2001. We are looking forward to seeing you in San Francisco!

Acknowledgement: The guest editor would like to thank Paul Bradley, Usama Fayyad and Sunita Sarawagi for the opportunity of contributing to SIGKDD Explorations. Without the support of Yesook Shim, it would have been impossible to complete the special focus sections of this issue.

References

[1] J. Gray and P. Shenoy. Rules of thumb in data engineering. *in Proceedings of 16th International Conference on Data Engineering*, pages 3-12. IEEE, 2000.

[2] U. Fayyad. "The digital physics of Data Mining", *Communications of the ACM*, March 2001.

[3] P. Lyman, H. R. Varian, J. Dunn, A. Strygin, K. Swearingen, 2001. "How Much Information"*, University of California, Berkeley, Technical Report.*

[4] M. Lesk. "How much information is there in the world?" *Technical report*, http://www.lesk.com/mlesk/ksg97/ksg.html 1997.

About the Editors

Editor-in-Chief:

Usama Fayyad is President & CEO of digiMine, Inc. He received his Ph.D. from The University of Michigan, Ann Arbor in 1991. He is an Editor-in-Chief of *Data Mining and Knowledge Discovery,* the primary technical journal in the field, and has chaired several past KDD conferences. Prior to digiMine, he was at Microsoft where he founded and led Microsoft Research's Data Mining & Exploration (DMX) Group. His work with Microsoft product groups included the development of data mining prediction components that ship with Microsoft Site Server (Commerce Server 3.0 and 4.0) and developing scalable algorithms for mining large databases and architecting their fit with server products such as Microsoft SQL Server and OLAP Services. In addition to managing the DMX research group, he managed the core part of the development team that is building providers in the SQL Server product group. He was also a driving force behind establishing a new industry standard in data mining based on Microsoft's OLE DB API. Prior to joining Microsoft, Usama was at the Jet Propulsion Laboratory (JPL), California Institute of Technology (1989-1995) where he founded and headed the Machine Learning Systems Group and developed data mining systems for the analysis of large scientific databases. For this work he received the most distinguished excellence awards from Caltech/JPL and a U.S. Government Medal from NASA. He remained affiliated with JPL as Distinguished Visiting Scientist after he moved to Microsoft.

Associate Editors:

Paul Bradley is a Data Mining Engineer at digiMine, Inc. His focus is on the integration and architecture of data mining services. He received his Ph.D. from the University of Wisconsin in 1998 on the topic of mathematical programming and data mining. Prior to joining digiMine, he was a researcher in the Data Management, Exploration and Mining Group at Microsoft Research. Research interests include classification and clustering algorithms; underlying mathematical programming formulations; and issues related to scalability. While at Microsoft Research, he worked on data mining components shipping with SQL Server and Commerce Server.

Sunita Sarawagi is assistant professor at the Indian Institute of Technology, Bombay. She received her Ph.D. degree in 1996 from the University of California at Berkeley and was a research staff member with the data mining group of IBM Almaden Research Center before joining IIT Bombay in 1999. Her research interests include database mining, OLAP and extended relational database systems. She has several publications in international conferences including a best paper award at the 1998 ACM SIGMOD conference. She has served as program committee member for SIGMOD, VLDB and ICDE conferences and is an associate editor of the ACM SIGKDD newsletter and IEEE Data Engineering Bulletin.

Guest Editor:

Kyuseok Shim is currently an Assistant Professor at Korea Advanced Institute of Science and Technology (KAIST) in Korea. Before joining KAIST, he was a member of technical staff (MTS) in Bell Laboratories in which he initiated and worked for the Serendip data mining project with Rajeev Rastogi. Before that, he was a research staff member of Quest Data Mining project at IBM Almaden Research Center. He also worked as a summer intern for two summers at Hewlett Packard Laboratories. He received B.S. degree in Electrical Engineering from Seoul National University in 1986, and the MS and Ph.D. degrees in Computer Science from University of Maryland, College Park, in 1988 and 1993 respectively. He has been working in the area of data mining, data warehousing, query processing and query optimization, XML and semi-structured data. He has been an Advisory Committee Member for ACM SIGKDD. He has also served as a program committee member on international conferences that include ICDE'97, KDD'98, SIGMOD'99, SIGKDD'99, VLDB'00 and SIGKDD'01. He did a data mining tutorial at CIKM'99, ICDE'99 and SIGKDD'99 conferences. He was co-chair of the 1999 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.