Editorial

P. S. Bradley
Microsoft Research
One Microsoft Way
Redmond, WA 98052
U.S.A.
bradley@microsoft.com

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

Usama Fayyad
DigiMine, Inc
11250 Kirkland Way, Suite 201
Kirkland, WA 98033
U.S.A.
fayyad@acm.org

With this issue of *SIGKDD Explorations* we launch our second volume of the newsletter. With our continued growth and the continued growth of the field we need to grow *Explorations* accordingly. We are pleased to bring in two changes to the editorial staff and format of your newsletter. We have invited Paul Bradley to join the editorial staff as Associate Editor. We have also asked him to serve as Guest Editor of this issue. This represented the second change in editorial format: future issues will have themes and special topics with guest editors invited to cover each special topic. We shall of course continue to include contributed articles on topics of interest and conference and workshop reports. In addition to serving as Guest Editor on this issue, Paul has also taken a major load of additional work on producing this issue in his new role as Associate Editor. Having welcomed Paul aboard, we'll let him introduce this issue's special focus.

The special issue section of this issue *SIGKDD Explorations* is focused on "Internet Data Mining". The "Internet" continues to provide the general population of computer users with innumerable benefits (and possibly some drawbacks). It also provides the data mining community with large amounts of data and great challenges through new applications areas like customer relationship management, e-commerce and web search.

As the Internet expands, there is a greater need to effectively index and classify content to support efficient search and navigation. In addition, the Internet provides businesses with an operational platform for interaction with their customers around the clock without geographic or physical boundaries. According to a Forrester report, "worldwide net commerce – both B2B and B2C – will hit $6.8 trillion in 2004" [3]. Databases currently are (and will continue to be) critical components in the e-commerce arena. The data generated by this domain provide "all the right ingredients for successful data mining" [1]. In addition to providing some of the key foundations in internet content indexing, text mining techniques may prove very useful in helping the end-user deal with the information overload streaming down to her or him by processing email and other electronic communication and prioritizing messages and tasks [2].

While such enthusiasm and opportunity for data mining in the Internet arena is exciting and intriguing, the serious issue of privacy needs to be effectively handled. Although it is not only an issue with data mining and the Internet, data mining practitioners in this space need to be constantly aware of the implications of tracking and analysis technologies on privacy. Without properly addressing the issue of privacy on the Internet, this abundant tap of data may eventually flow much slower due to regulations, and other corrective or preventive restrictions.

Given the opportunities, challenges and issues touched upon above, we chose to focus this issue of *SIGKDD Explorations* on the topic of "Internet Data Mining". On this topic, five articles are included in the special focus section. Raymond Kosala and Hendrik Blockeel contribute a survey on web mining research, providing an overview of sub-areas and problems encompassed by web mining. Yiming Ma, Bing Liu and Ching Kian Wong propose a web-browser-based graphical user interface for displaying, navigating and annotating a set of discovered rules. Sunita Sarawagi and Sree Hari Nagaralu discuss the notion, use and implication of providing data mining models as services on the web. Stanley Loh, Leandro Krug Wives and José Palazzo de Oliveira propose a concept-based approach to analyzing texts extracted from the web using an interesting collaboration between human and computer to obtain the final result. José Borges and Mark Levene present a heuristic for speeding up a previous system based upon a hypertext probabilistic grammar for extracting browsing behavior from web log data.

Additional articles addressing general topics of interest to the SIGKDD community include the introduction of a simple new scalable k-means clustering algorithm proposed and evaluated by Frederik Farnstrom, James Lewis and Charles Elkan. We hope that this article prompts further discussion on scalable data mining algorithms in general. A survey and overview of algorithms discovering association rules appears by Jochen Hipp, Ulrich Güntzer and Gholamreza Nakhaeizadeh, including several runtime experiments. Alex Freitas provides a position paper outlining differences between the classification task and that of discovering association rules and raises many interesting points, hopefully prompting further discussion on the subject from the SIGKDD community.Reports from workshops and meetings of interest include a report from Jeanne Behnke, Elaine Dobinson and others on the NASA Workshop on Issues in the Application of Data Mining to Scientific Data. Coming from one of the early organizations with a definite need to analyze large data sets, the topics discussed in this report provide interesting insight into the relationship between the data analyst and the domain specialist in scientific areas of interest to NASA. Dan Suciu and Gottfried Vossen have kindly provided a report from WebDB'2000 following the ACM PODS/SIGMOD conferences. Dimitrios Gunopulos and Rajeev Rastogi have, similarly, kindly provided a report on the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery.

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 contacts us with any ideas for future topics of interest, and any feedback on *Explorations.*

We conclude with a reminder that the premier event in the SIGKDD community is quickly approaching. KDD-2000 is taking place in Boston, MA, beginning Sunday, August 20th and concluding on August 23rd. See the conference website for details (http://www.acm.org/sigkdd/kdd2000). We look forward to seeing you in Boston!

REFERENCES

[1] Kohavi, R. Mining E-commerce Data: Challenges and Stories from the Trenches. Presentation at DIMACS/IBM Workshop on Data Mining in the Internet Age. May 2, 2000.

[2] Markoff, J. Microsoft Sees Software 'Agent' as Way to Avoid Distractions. New York Times, Technology Section, July 17, 2000.

[3] Sanders, M. R., and Temkin, B. D. Global eCommerce Approaches Hypergrowth. The Forrester Brief, April 18, 2000. http://www.forrester.com.

About the editors:

Paul Bradley is a Researcher in the Data Management, Exploration and Mining Group at Microsoft Research. Research interests include classification and clustering algorithms; underlying mathematical problem formulations; and issues related to scalability. After receiving the Ph.D. degree from the University of Wisconsin-Madison in 1998 on the topic of mathematical programming and data mining, he joined Microsoft Research where he works on research in developing new data mining algorithms as well as on shipping data mining components in Microsoft products such as 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 at 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.

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.