Report on the Workshop on Research Issues in Data Mining and Knowledge Discovery Workshop (DMKD 2001)

Roberto Bayardo
IBM Almaden Research Center San Jose, CA
bayardo@alum.mit.edu

Johannes E. Gehrke
Cornell University
johannes@cs.cornell.edu

ABSTRACT

This short article summarizes the program of the Sixth Workshop on Research Issues in Data Mining and Knowledge Discovery Workshop (DMKD 2001).

Keywords

Databases, data mining

1. INTRODUCTION

For the last five years, members of the database community have organized and held the Data Mining and Knowledge Discovery (DMKD) workshop in conjunction with the ACM SIGMOD conference. Over this time, the data-mining community has grown at an incredible rate, along with the number of avenues available for disseminating research results in the field. Consider this year, for example, during which two new data-mining conferences have been hosted by IEEE and SIAM. These are in addition to already popular data-mining conferences such as ACM SIGKDD, PAKDD, and a variety of other workshops and more general (but data-mining friendly) conferences such as ICML, SIGMOD, VLDB, ICDE, and AAAI/IJCAI. This incredible growth led us to rethink the purpose of the DMKD workshop, which in the past has provided an excellent avenue for data-mining topics now readily covered by most of the major conferences listed above (e.g. algorithms and scaling issues, often pertaining to secondary storage). Our reexamination led us steer the workshop away from a mini-conference for regular data mining papers and towards a discussion on the next generation of data mining research. We specifically solicited ”forward looking” research visions, and papers on challenging applications, since we see applications as driving the next generation of data-mining tools.

2. A BRIEF PROGRAM OVERVIEW

The workshop opened with an invited talk by Brian Lent, who is currently busy founding Intelligent Results, a data-mining startup. Brian spoke of his experiences while heading Amazon.com's data-mining lab. One interesting lesson from his talk came from an anecdote regarding stability of their customer segmentation model. After rerunning a clustering algorithm on new data, the resulting model contained several more clusters than models produced previously. This sudden change in the model, which happened without any explanation, shook the faith others outside Brian's group had in data-mining technology. In addition to performance and accuracy, model stability (or at least the ability to explain changes in a model) should also be a metric along which algorithms are commonly judged.

The first paper session opened with a talk by Padhraic Smyth entitled ”Breaking Out of the Black-Box: Research Challenges in Data Mining”. Padhric's paper takes a step back from the usual algorithmic view of data-mining, looking instead at what different segments of the community do with their data in order to identify opportunities and challenges for future research. The challenges he identifies include self-tuning algorithms, data-mining software environments that support the entire mining process (exploring and preprocesing, modeling, mining, evaluating, and deploying), better modeling of the time dimension (e.g. periodicity, seasonality, non-stationary, etc.), personalization from sparse data through, e.g., the use of Bayesian hierarchical models, scalable exploration (not just mining), models and languages for spatio-temporal data, exploiting prior knowledge in pattern-finding, and determining ”local structure” as opposed to identifying global models. Jian Pei presented the next and final paper of the session, entitled Fault-Tolerant Frequent Pattern Mining: Problems and Challenges, authored by Pei, A. Tung, and J. Han. Jian's talk emphasized the point that real-world data tends to be dirty, and then proposed that data-mining algorithms should accomodate this characteristic by way of ”fault tolerance”. In the context of mining frequent patterns, Jian argues that a fault tolerant algorithm should not require an exact matching of items in the itemset to a transaction, but rather a partial match. He presented an algorithm, FT-Apriori, that implements this idea, and evaluated the effect of fault tolerance on performance as well as on the rules mined by the algorithm.

In the second paper session, Baohua Gu presented the paper ”Efficiently Determine the Starting Sample Size for Progressive Sampling” by Gu, B. Liu, F. Hu, and H. Liu. Progressive sampling involves learning models using increasingly larger sample sizes until model accuracy stabilizes. In order to avoid the performance hit resulting from an improperly sized initial sample, they propose a technique for determining what they term a ”statistical optimal sample size (SOSS)” which can be computed in one pass of the data. The statistical optimal sample size attempts to predict the size of the initial sample which would maximize the performance of progressive sampling. Next in the session came Xiong Wang's presentation on "Mining Protein Surfaces". This application paper introduces the concept of an alpha-surface intended to capture the three-dimensional surface structure of a protein. Xiong presented an algorithm for extracting alpha-surfaces, and showed that patterns extracted from alpha-surfaces were useful features for protein classification tasks. The third paper titled "Motion Mining: Discovering Spatio-Temporal Patterns in Databases of Human Motion" was presented by George Kollios in joint work with S. Sclaroff and M. Betkewas. Kollios described the rapidly growing availability of data about human motion, and he laid out a research vision of indexing and mining motion data with applications such as computer-assisted physical rehabilitation, occupational safety, and sports training, medicine, and diagnosis. The final paper in the session was presented by Qin Ding. The paper entitled "On Mining Satellite and Other Remotely Sensed Images" by Ding, W. Perizo, Q. Ding, and A. Roy introduces the problem of mining rules from images. Each pixel in the image has associated attributes and can thus be interpreted as an input record to a mining problem. Ding then describes an efficient algorithm for mining quantitative association rules and classification rules based on a hierarchical data representation called the Peano Count Tree. Experiments show that the proposed algorithm works well in practice and outperforms Apriori and FP-tree based rule induction algorithms.

After lunch, the workshop continued with an invited talk "An Industry Perspective on DMKD Research Issues" by Samy Uthurusamy from General Motors. Samy talked about his experience as a technology visionary within GM and he discussed several research directions that he considered especially promising, among others work on scale-invariant clustering, research into web aggregators, querying and processing data streams, the semantic web, and the incorporation of background knowledge into data mining. He also mentioned that management and personalization of resources in a big company is a very important problem; he mentioned the existence of 220,000 desktop computers within GM as a data point.

The third paper session started with a presentation by Minos Garofalakis who introduced techniques developed in the NEMESIS project at Lucent Bell Labs. In his presentation titled "Data Mining Meets Network Management: The NEMESIS Project" (joint with Rajeev Rastogi) Minos described a semantic compression algorithm that exploits data correlation discovered by decision tree models constructed between different attributes in the data. Minos mentioned event correlation and root cause analysis as promising research directions. Afterwards, Pedro Domingos presented his vision on mining data streams. His presentation titled "Catching Up with the Data: Research Issues in Mining Data Streams" (joint with G. Hulten) motivated the transition from mining static datasets to the continuous mining of high-speed data streams, and he described design criteria and possible techniques for constructing such systems. Pedro also outlined approaches for mining time-changing data streams where the underlying distribution that generates the data changes over time. Jochen Hipp introduced in the last paper of the session the issues of data quality for data mining. In this presentation "Data Quality Mining - Making a Virtue of Necessity" (joint work with Ulrich Guntzer and Udo Grimmer) Jochen argued that in many cases the data used for mining is not perfect, and that the results obtained through mining can often help to identify and correct deficient data. He then described an algorithm that discovers association rules in a dataset to score potentially deficient records in the original dataset.

All full papers accepted at the workshop – including two submissions that were accepted as position papers – are published on the workshop homepage:

http://www.cs.cornell.edu/johannes/dmkd2001.htm.

3. ACKNOWLEDGEMENTS.

We thank the members of the DMKD 2001 Program Committee for their commitment, SIGMOD 2001 General Chair Jianwen Su for his help with the logistics, and previous chairs Dimitrios Gunopulos, Jiawei Han, Ramakrishnan Srikant, and Rajeev Rastogi for their guidance. We also thank IBM Research for their financial support of the workshop and the student travel scholarships.