Data Mining and Knowledge Discovery, 1, 5–10 (1997)

Editorial

USAMA M. FAYYAD fayyad@microsoft.com
Microsoft Research

Looming atop a wide variety of human activities are the menacing profiles of ever-growing mountains of data. These mountains grew as a result of great engineering successes that enabled us to build devices to generate, collect, and store digital data. With major advances in database technology came the creation of huge efficient data stores. Advances in computer networking have enabled the data glut to reach anyone who cares to tap in. Unfortunately, we have not witnessed corresponding advances in computational techniques to help us *analyze* the accumulated data. Without such developments, we risk missing most of what the data have to offer.

Be it a satellite orbiting our planet, a medical imaging device, a credit-card transaction verification system, or a supermarket's checkout system, the human at the other end of the data gathering and storage machinery is faced with the same problem: *What to do with all this data?* Ignoring whatever we cannot analyze would be wasteful and unwise. Should one choose to ignore valuable information buried within the data, then one's competition may put them to good use; perhaps to one's detriment. In scientific endeavours, data represents observations carefully collected about some phenomena under study, and the race is on for who can explain the observations best. In business endeavours, data captures information about the markets, competitors, and customers. In manufacturing, data captures performance and optimization opportunities, and keys to improving processes and troubleshooting problems.

The value of raw data is typically predicated on the ability to extract higher level information: information useful for decision support, for exploration, and for better understanding of the phenomena generating the data. Traditionally, humans have done the task of analysis. One or more analysts get intimately familiar with the data and with the help of statistical techniques provide summaries and generate reports. In effect, analysts determine the *right* queries to ask and sometimes even act as sophisticated query processors. Such an approach rapidly breaks down as the volume and dimensionality of the data increase. Who could be expected to "understand" millions of cases each having hundreds of fields? To further complicate the situation, the data grow and change at rates that would quickly overwhelm manual analysis (even if it were possible). Hence tools to aid in at least the partial automation of analysis tasks are becoming a necessity.

Why Data Mining and Knowledge Discovery?

Knowledge Discovery in Databases (KDD) is concerned with extracting useful information from databases (see [3] for basic definitions and a more detailed discussion). The term *data mining* has historically been used in the database community and in statistics (often in the latter with negative connotations to indicate improper data analysis). We take the view that any algorithm that enumerates patterns from, or fits models to, data is a *data mining algorithm*. We further view data mining to be a *single step in a larger process that we call the*

KDD process. See [3] for more details on the various steps of the process which include data warehousing, target data selection, cleaning, preprocessing, transformation and reduction, data mining, model selection (or combination), evaluation and interpretation, and finally consolidation and use of the extracted "knowledge". Hence data mining is but a step in this iterative and interactive process. We chose to include it in the name of the journal because it represents a majority of the published research work, and because we wanted to build bridges between the various communities that work on topics related to data mining.

KDD's goal, as stated above, is very broad, and can describe a multitude of fields of study. Statistics has been preoccupied with this goal for over a century. So have many other fields including database systems, pattern recognition, artificial intelligence, data visualization, and a host of activities related to data analysis. So why has a separate community emerged under the name "KDD"?

The answer: new approaches, techniques, and solutions have to be developed to enable analysis of large databases. Faced with massive data sets, traditional approaches in statistics and pattern recognition collapse. For example, a statistical analysis package (e.g. K-means clustering in your favorite Fortran library) assumes data can be "loaded" into memory and then manipulated. What happens when the data set will not fit in main memory? What happens if the database is on a remote server and will never permit a naïve scan of the data? How do I sample effectively if I am not permitted to query for a stratified sample because the relevant fields are not indexed? What if the data set is in a multitude of tables (relations) and can only be accessed via some hierarchically structured set of fields? What if the relations are sparse (not all fields are defined or even applicable to any fixed subset of the data)? How do I fit a statistical model with a large number of variables?

The open problems are not restricted to scalability issues of storage, access, and scale. For example, a problem that is not addressed by the database field is one I like to call the "query formulation problem": *what to do if one does not know how to specify the desired query to begin with?* For example, it would be desirable for a bank to issue a query at a high level: "give me all transactions whose likelihood of being fraudulent exceeds 0.75." It is not clear that one can write a SQL query (or even a program) to retrieve the target. Most interesting queries that arise with end-users of the data are of this class. KDD provides an alternative solution to this problem. Assuming that certain cases in the database can be identified as "fraudulent" and others as "known to be legitimate", then one can construct a *training sample* for a data mining algorithm, let the algorithm build a predictive model, and then retrieve records that the model triggers on. This is an example of a much needed and much more natural interface between humans and databases. Issues of inference under uncertainty, search for patterns and parameters in large spaces, and so on are also fundamental to KDD. While these issues are studied in many related fields, approaches to solving them in the context of large databases are unique to KDD. I outline several other issues and challenges for KDD later in this editorial, and I am sure future pages of this journal will unveil many problems we have not thought of yet.

Related Fields

Many research communities are strongly related to KDD. For example, by our definition, all work in classification and clustering in statistics, pattern recognition, neural networks, machine learning, and databases would fit under the data mining step. In addition to exploratory data analysis (EDA), statistics overlaps with KDD in many other steps including data selection and sampling, preprocessing, transformation, and evaluation of extracted knowledge. The Database field is of fundamental importance to KDD. The efficient and reliable storage and retrieval of the data, as well as issues of flexible querying and query optimization, are important enabling techniques. In addition, contributions from the database research literature in the area of data mining are beginning to appear. On-line Analytical Processing (OLAP) is an evolving field with very strong ties to databases, data warehousing, and KDD. While the emphasis in OLAP is still primarily on data visualization and query-driven exploration, automated techniques for data mining can play a major role in making OLAP more useful and easier to apply.

Other related fields include optimization (in search), high-performance and parallel computing, knowledge modeling, the management of uncertainty, and data visualization. Data visualization can contribute to effective EDA and visualization of extracted knowledge. Data mining can enable the visualization of patterns hidden deep within the data and embedded in much higher dimensional spaces. For example, a clustering method can segment the data into homogeneous subsets that are easier to describe and visualize. These in turn can be displayed to the user instead of attempting to display the entire data (or a global random sample of it) which usually results in missing the embedded patterns.

In an ideal world, KDD should have evolved as a proper subset of statistics. However, statisticians have not focused on considering issues related to large databases. In addition, historically, the majority of the work has been primarily focused on hypothesis-verification as the primary mode of data analysis (which is certainly no longer true now). The decoupling of database issues (storage and retrieval) from analysis issues is also a culprit. Furthermore, compared with techniques that data mining draws on from pattern recognition, machine learning, and neural networks, the traditional approaches in statistics perform little search over models and parameters (again with notable recent exceptions). KDD is concerned with formalizing and encoding aspects of the "art" of statistical analysis and making analysis methods easier to use by those who own the data, regardless of whether they have the pre-requisite knowledge of the techniques being used. A marketing person interested in segmenting a database may not have the necessary advanced degree in statistics to understand and use the literature or the library of available routines. We do not dismiss the dangers of blind mining and that it can easily deteriorate to *data dredging*. However, the strong need for analysis aids in the data-overloaded society need to be addressed.

In response to the growing need, the KDD community emerged. The first KDD workshop [7] was held in Detroit in August of 1989. Workshops followed in 1991, 1993, with the last workshop in 1994 [5]. The workshops evolved into an annual international conference in 1995 [4]. KDD-96: the Second International Conference on Knowledge Discovery and Data Mining [8] attracted over 500 people. The field is still growing at a dramatic pace. Our intent in launching this journal is to provide a forum for the publication of high quality work relevant to KDD. It is our sincere hope that we will get participation from the various relevant communities. This first issue is a good example, drawing articles from the statistics, databases, and AI fields.

About the First Issue

At the heart of the problem of inference from data is the field of statistics. Both on the side of hypothesis validation and on the side of exploratory data analysis, statistical techniques are of fundamental importance. The article by Glymour, Madigan, Pregibon, and Smyth gives a perspective from statistics and identifies a wealth of statistical results that KDD can benefit from as well as some caveats and directions for their proper use.

The paper by Gray et al presents the datacube, a database approach to summarizing essential information in the form of aggregate sums from a database. This technique could provide the infrastructure to enable efficient exploration and analysis on the data, as well as data mining support.

Two problems of fundamental importance to KDD are classification and density estimation. The article by Friedman gives a new perspective on the old bias-variance tradeoff problem and presents results of importance to evaluating a classification model. Heckerman provides a thorough tutorial article on Bayesian statistical techniques and graphical models, a powerful technique for representing a joint probability distribution. While Bayesian networks have proven to be convenient and effective in encoding prior knowledge of a problem and reasoning about it under uncertainty, Heckerman focuses on aspects of estimating parameters as well as extracting graph structure from data.

Our journal also includes a special section on short application summaries. This section is intended to present very short summaries of deployed KDD systems, or KDD results that have impacted a field of engineering or somehow affected the public. The article by Bhandari et al plays this role in the first issue, describing impact on an unusual (and very public) application: the coaching of several basketball teams in the NBA.

Future Prospects and Challenges

Successful KDD applications continue to appear, driven mainly by a glut in databases that have clearly grown to surpass raw human processing abilities. For examples of success stories in applications in industry see [1] and in science analysis see [2]. More detailed case studies are found in [3]. Driving the healthy growth of this field are strong forces (both economic and social) that are a product of the data overload phenomenon. I view the need to deliver workable solutions to pressing problems as a very healthy pressure on the KDD field. Not only will it ensure our healthy growth as a new engineering discipline, but it will provide our efforts with a healthy dose of reality checks; insuring that any theory or model that emerges will find its immediate real-world test environment.

The fundamental problems are still as difficult as they always were, and we need to guard against building unrealistic expectations in the public's mind. The challenges ahead of us are formidable. Some of these challenges include:

1. Develop mining algorithms for classification, clustering, dependency analysis, and change and deviation detection that scale to large databases. There is a tradeoff between performance and accuracy as one surrenders to the fact that data resides primarily on disk or on a server and cannot fit in main memory.

2. Develop schemes for encoding "metadata" (information about the content and meaning of data) over data tables so that mining algorithms can operate meaningfully on a database and so that the KDD system can effectively ask for more information from the user.

3. While operating in a very large sample size environment is a blessing against overfitting problems, data mining systems need to guard against fitting models to data by chance. This problem becomes significant as a program explores a huge search space over many models for a given data set.

4. Develop effective means for data sampling, data reduction, and dimensionality reduction that operate on a mixture of categorical and numeric data fields. While large sample sizes allow us to handle higher dimensions, our understanding of high dimensional spaces and estimation within them is still fairly primitive. The curse of dimensionality is still with us.

5. Develop schemes capable of mining over nonhomogenous data sets (including mixtures of multimedia, video, and text modalities) and deal with sparse relations that are only defined over parts of the data.

6. Develop new mining and search algorithms capable of extracting more complex relationships between fields and able to account for structure over the fields (e.g. hierarchies, sparse relations); i.e. go beyond the flat file or single table assumption.

7. Develop data mining methods that account for prior knowledge of data and exploit such knowledge in reducing search, that can account for costs and benefits, and that are robust against uncertainty and missing data problems. Bayesian methods and decision analysis provide the basic foundational framework.

8. Enhance database management systems to support new primitives for the efficient extraction of necessary sufficient statistics as well as more efficient sampling schemes. This includes providing SQL support for new primitives that may be needed (c.f. the paper by Gray et al in this issue).

9. Scale methods to parallel databases with hundreds of tables, thousands of fields, and terabytes of data. Issues of query optimization in these settings are fundamental.

10. Account for and model comprehensibility of extracted models; allow proper tradeoffs between complexity and understandability of models for purposes of visualization and reporting; enable interactive exploration where the analyst can easily provide hints to help the mining algorithm with its search.

11. Develop theory and techniques to model growth and change in data. Large databases, because they grow over a long time, do not typically grow as if sampled from a static joint probability density. The question of how does the data grow? needs to be better understood (see articles by P. Huber, by Fayyad & Smyth, and by others in [6]) and tools for coping with it need to be developed.

KDD holds the promise of an enabling technology that could unlock the knowledge lying dormant in huge databases, thereby improving humanity's collective intellect: a sort of amplifier of basic human analysis capabilities. Perhaps the most exciting aspect of the launch of this new journal is the possibility of the birth of a new research area properly mixing statistics, databases, automated data analysis and reduction, and other related areas. While KDD will draw on the substantial body of knowledge built up in its constituent fields, it is my hope that a new science will inevitably emerge. A science of how to exploit massive data sets, how to store and access them for analysis purposes, and how to cope with growth and change in data. I sincerely hope that future issues of this journal will address some of the challenges and chronicle the development of theory and applications of the new science for supporting analysis and decision making with massive data sets.

Acknowledgments

I would like to thank the Heikki Mannila and Gregory Piatetsky-Shapiro for their help over the past year and for comments on this editorial. Others who commented include: Surajit Chaudhuri, Jim Gray, David Hand, David Heckerman, David Madigan, Chris Meek, Daryl Pregibon, Padhraic Smyth, and R. Uthurusamy. I am grateful to the Editorial Board, to the contributing (and submitting) authors, and to the reviewers for working under harsh tight deadlines to put the first issue together.

References

Brachman, R., Khabaza, T., Kloesgen, W., Piatetsky-Shapiro, G. and Simoudis, E. Industrial Applications of Data Mining and Knowledge Discovery, Communications of ACM, Nov. 1996.

Fayyad, U., Haussler, D. and Stolorz, P. Mining Science Data, Communications of ACM, Nov. 1996.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. "From Data Mining to Knowledge Discovery: An Overview." In Advances in Knowledge Discovery and Data Mining, Fayyad et al (Eds.) MIT Press, 1996.

Fayyad, U. and Uthurusamy, R. (Eds.). Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), AAAI Press, 1995.

Fayyad, U. and Uthurusamy, R. (Eds.). Proceedings of the AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94), AAAI Press, 1995.

Kettenring, J. and Pregibon, D. (Eds.). Statistics and Massive Data Sets, Report to the Committee on Applied and Theoretical Statistics, National Research Council, Washington, D.C. 1996.

Piatetsky-Shapiro, G. and Frawley, W. (Eds). Knowledge Discovery in Databases, MIT Press 1991.

Simoudis, E., Han, J. and Fayyad, U. (Eds.). Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), AAAI Press, 1996.

Usama Fayyad is a Senior Researcher at Microsoft Research. His research interests include knowledge discovery in large databases, data mining, machine learning, statistical pattern recognition, and clustering. After receiving the Ph.D. degree in 1991, he joined the Jet Propulsion Laboratory (JPL), California Institute of Technology JPL where (until 1996) he headed the Machine Learning Systems Group and developed data mining systems for automated science data analysis. He remains affiliated with JPL as Distinguished Visiting Scientist. He received the 1994 NASA Exceptional Achievement Medal and the JPL 1993 Lew Allen Award for Excellence in Research. He is a co-editor of Advances in Knowledge Discovery and Data Mining (AAAI/MIT Press, 1996). He was program co-chair of KDD-94 and KDD-95 (the First International Conference on Knowledge Discovery and Data Mining) and a general chair of KDD-96.