IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4524

KDD Toward Maturity: Issues and Challenges

Stijn Viaene and Guido Dedene
K.U.Leuven, Department of Applied Economics
Management Information Systems Group
Naamsestraat 69, B-3000 Leuven, Belgium
Phone +32 16 32.68.91, Fax +32 16 32.67.32
{Stijn.Viaene;Guido.Dedene}@econ.kuleuven.ac.be

ABSTRACT

In this article we set the stage for broadening the scope of the discussion on crossing the chasm for knowledge discovery in databases. This is a discussion on its essence, its connection to other disciplines and its viability in a real-world business context.

INTRODUCTION

Knowledge discovery in databases (KDD), commonly termed data mining, although the latter is merely a step within the broader KDD process, has been authoritatively defined by Fayyad *et al.* [14] as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."

KDD grew out of an inability to effectively utilise the nuggets of information and hidden knowledge that are believed to reside in large, diverse, complex, dirty and often highly dispersed data sources. The scale of the data management problem has numbed our human abilities of analysis and synthesis for problem solving and has made it quasi impossible to filter out the valuable from the bulk. Making sense out of the enormous amalgamation of data sources is the challenge taken up by KDD [14-15]. The increasing interest in KDD from the academia and industry closely parallels the trends of decreasing cost of massive data storage, ease of collection of this data over worldwide network architectures, increasing processing power, and decreasing cost of processing.

In this article we argue that making KDD deliver on its expectations is as much about addressing real-world issues, seeing the larger business picture, adhering to good business practice, and starting with the right mindset and attitude, as it is about technology. This extends beyond the discussion on "crossing the chasm" initiated by KDD-heavyweights such as Rakesh Agrawal [1] and Ronny Kohavi [29]. Data mining techniques are given ample attention in the domain literature. What we feel is missing, or at least grossly underemphasized, however, is a clear, well-founded and appropriately scoped discussion on the issues to be addressed for KDD to deliver in a business context. Each of the following sections covers a basic theme pertaining to the viable use of KDD in business.

BEYOND DATA MINING

KDD is to be conceived as the iterative, interactive and repetitive process (see e.g. [8,14,43]) of extracting incremental business knowledge from a myriad of data sources by using several flavours of smart technology in each process phase. Still, the KDD community continues to overly focus on the data mining phase of the process. However, there is much more to KDD than running mining algorithms. Cabena *et al.* [4] and Hirji [23], for example, acknowledge that the application of mining algorithms in a business context generally involves complex and resource-intensive subsidiary tasks of pre- and post-processing. Moreover, jumping back and forth in the process' task model is the rule rather than the exception.

The process nature has to be reflected in KDD tool support. For example, general-purpose mining tools such as SAS Enterprise Miner, SPSS Clementine or IBM Intelligent Miner adhere to the philosophy of the transformation graphical user interface (GUI) [29]. The tools are task-oriented [3]. They provide for a GUI that is organised along the lines of the phases and tasks of a KDD process model that covers extensive pre-processing, mining, as well as post-processing possibilities. It is also important that data explorers at all times be given efficient and effective access to the data via query and reporting (Q&R) and online analytical processing (OLAP) facilities supported by powerful visuals. This essential requirement for data proximity can only be realised when there is plug and play interoperability with database and data warehouse facilities.

The exploratory nature of the KDD activity¾that is, at the outset one typically has only a broad idea of what is to be discovered and how exactly one will proceed toward this end¾requires for careful bookkeeping of transformations and multiple version or scenario control. More generally, the success of KDD requires for a metaprocess that parallels the actual KDD process.

DESIGN FOR CHANGE

Design for change encompasses several proven change-based design principles and practices that are aimed at creating effective and MATURE (Maintainable, Adaptable, Transparent, User-friendly, Reliable, Efficient) systems capable of evolving with the changing business reality, and at supporting organisational transformation. These systems must satisfy the requirements of flexibility and interoperability so that they can effectively contribute to the realisation of the modern collaborative, networked organisation [47]. This must be carefully planned and designed for. Change management is as real an issue to KDD as it is to any other discipline that interfaces on a permanent basis with the evolving business reality. From the very start KDD should be conceived as a repeatable exercise. This means organising and managing the data, process, systems, system interfaces, knowledge base and applications in such a way that they can be used and reused over time.

In that respect, data mining metadata standards such as the XML Data Mining Specification Language [30], the Predictive Model Markup Language [9], the Common Warehouse Metamodel [35], and the XML for Analysis specification language [25], data web standards such as the Data Space Transfer Protocol [33], and KDD process standards such as the Cross-Industry Standard Process for Data Mining [8] are extremely valuable contributions. These initiatives set the stage for further efforts of standardisation and formalisation. For example, at the top of the data mining standards agenda is agreeing on standards for cleaning, transforming and preparing data for data mining [17]. In addition, there is a lot to be learned from the multidisciplinary field of software and information systems engineering. Among the most important achievements we count the state-of-the-art theory and practice on object-oriented and component-based systems development (see e.g. [10,21,32]) and (meta)model-driven, architectural design of information systems (see e.g. [36,44-45]).

THINK OUTSIDE THE BOX

Making KDD work in a real-world setting requires a broad scope of techniques, expertise and knowledge. As a discipline KDD addresses the

theoretical, methodological and practical aspects of making the multidisciplinary mix of elements fit the knowledge discovery task and address its challenges. Hereto, the discipline of KDD can build on proven advancement in mature fields of research and practice such as database management, software engineering, high-performance computing, computer graphics and statistics.

The challenge of KDD increasingly forces researchers and practitioners to build bridges between disciplines that have largely been evolving in parallel universes. The promise of profitable synergies can effectively overcome the reluctance to look beyond the boundaries of the single discipline. Some of the areas and issues of potential synergy are being explicitly addressed in research and now find their way into KDD public debate fora. Take the issue of database support for efficient data mining (see e.g. [5-6,26,34,42]), or the valuable contributions of statistics to data mining (see e.g. [16,18-20,27,31]). However, there still is a lot of ground to be covered. We need to continually stimulate an attitude of openness, thinking from a broad perspective on problems, and actively search for developed ideas, theory, methodology and practice in disciplines beyond the fields of machine learning, artificial intelligence or statistics.

RISK-RETURN MANAGEMENT

However exciting the new technology may be, its viability in a business context is but guaranteed if it is purposefully applied and effectively improves profitability. This is no different for KDD (and complementary technologies) than for other projects that are competing for the same scarce resources. Essentially, KDD is not profitable until successful deployment of the discovered knowledge in the operational business setting is a fact. The discovery-driven nature of KDD makes it all the more important to adhere to proven methods of risk management for organising work.

Business managers and sponsors demand to see clear business opportunities and goals, relatively short-term expected profits and a budgeted resource plan based on a clear project plan. They demand regular progress reports, project audits, and so on. To convince the business to invest the necessary resources in KDD one should be prepared to commit to results¾for which one will eventually be held accountable¾and show that the chosen approach is feasible. Organising work along the lines of proven program and project management practice is imperative (see e.g. [2,48]). Admittedly, the exploratory character of the KDD effort is likely to introduce some of its own rationale. Nevertheless, one should watch out for too much experimentation, freewheeling and improvisation. Leave room for creativity, but always try to stay faithful to the program, the plan and the method.

The choice of the first few projects is critical: Think big, but start small. Target projects with clearly visible and quantifiable deliverables within a relatively short time frame. Start by setting realistic expectations and take into account the restrictions posed by the in-house state-of-affairs. Obviously, you need to get your data right before you start using KDD technology [22]. Here too, garbage in, garbage out applies. Consider the fact that you may not actually need data mining technology to solve your business problem. Sometimes, all you need is Q&R or OLAP on the data warehouse. Nevertheless, you should learn how to get the most out of your data. This means exploring the whole range of KDD technologies, including data mining. This requires going through a learning curve. Training and education should encompass elements of theory, methodology and best practice of KDD. It also pays to educate and train the business-side customers of the knowledge, and to show them how exactly they can make optimal use of the uncovered business intelligence.

HUMAN-CENTRED PROCESS

Despite all the technology involved, KDD is still rather human-intensive: We have humans that set out the course of discovery, humans at the wheel, in the navigator's seat, in support, in the jury and, most importantly, humans to be pleased. Some tasks are notoriously hard to automate (e.g. pre-processing tasks and validation/appreciation of the mined results). Moreover, knowledge extracted in the KDD process is eventually meant for deployment in systems or applications that are operated by humans or assist humans in the decision making process. KDD is so closely intertwined with technology that it sometimes is difficult to remember its true objective, which is to leverage creative, innovative human intelligence and entrepreneurship. Thus, until further notice, human creativity, human problem-solving abilities and entrepreneurship are key to its business success.

Brachman and Anand [3], for example, are strong proponents of a human-centred view on KDD. This means looking at the KDD problem, process and environment from a careful understanding of the interactions between the human actors involved in KDD and the data. A human-centred view takes human characteristics and human abilities as a point of departure. Obviously, this design rationale applies for the systems that are part of the actual KDD environment as well as for the operational and management systems in which the discovered knowledge is deployed. Swift and agile data accessibility and visualisation support for exploration and use are merely a start. We should be thinking in terms of multimedia support, speech recognition, animated simulation, or even virtual reality engaging all the senses. Another example is the institutionalisation of a computer-supported platform for disseminating interesting knowledge or training programs to help people use this knowledge effectively.

One of the core issues in data mining is optimising the interplay between the mining algorithm and the prior domain knowledge residing with the human expert. Much of the sense-making process depends on this prior domain knowledge and the act of human interpretation. The feasibility of the idea of integrating human expert knowledge into an algorithm essentially depends on how successful one is at eliciting that knowledge, codifying it and consolidating it in a formal knowledge base that is conceptually compatible with the algorithm. The idea is that algorithms could then be designed to tap into the codified expert knowledge base and produce only incremental knowledge¾that is, beyond the fool's gold. The question remains how successful we will be at eliciting and codifying expert knowledge. *What about tacit expert knowledge?* (see e.g. [40-41]). Everyone involved in organisational knowledge management¾including the people involved in KDD¾needs to develop a profound appreciation for the intangible assets resident in the minds and experiences of knowledge workers, as well as for the social nature of the decision making process.

CUSTOMER TRUST

Concerns about the preservation and invasion of customer privacy have given rise to a rather polemic discussion on the merits and perils of the use of KDD technology (see e.g. [7,13,37,46]). May 1st, 1999 The Economist spoke out on "The End of Privacy," [12]. The fear of the customer is grounded in the loss of control over personal information and the perception that the legal system is unable to appropriately protect privacy. Policy makers in Europe and the USA have moved the issue way up the political agenda (see e.g. [11,24]). Balancing technological advancement with the very real issue of privacy protection is considered by the KDD community as one of the key challenges for the 21st century [38].

Technically, the potential is immense, but the possibilities of technology reach far beyond what is provided for in the law. Still, about the worst thing a company could do is to deny itself this opportunity and wait for legal matters on the privacy issue to get settled. At the same time, we should not underestimate the potential of the public to react either. Ill-considered obtainment or careless usage of customer data could easily ruin a company's reputation. In an economy where customer loyalty is built on the very foundations of companies staging positive customer experiences [39], failure to take the issue seriously can ruin the whole setup.

The fear for misuse of private information forms a real threat to the customer-company trust relationship. On the other hand, take away that fear and you will likely have gained yourself a comparative advantage. The trick is then to explicitly address the issue and appropriate yourself a clear, concise and transparent corporate privacy protection policy endorsed by the voice of the customer [28]. Companies

making a clear commitment to honour and build upon a trust relationship with individual customers may go a long way toward assuaging customer fears. Moreover, customers may be willing to share more if they feel that they are given greater control over the disclosure and the use of information, and feel that the data is carefully shielded from abuse by internal or external parties.

Coming to terms with privacy issues is not easy. For one, the exploratory nature of the KDD activity makes it quasi impossible to anticipate all future usage of the data or knowledge from the outset. Moreover, companies are continually exploring their boundaries. The boundaries set by technological advance may not always coincide with the boundaries set by the trust relationship. Previous collisions between data mining and privacy are likely to be just the beginning. Unless the issue is effectively addressed in concert by all players involved, we can expect to see further legal challenges to the use of KDD.

CONCLUSION

The aim of this article was to synthesise from our experience in using KDD in a continuously changing real-world context, a context in which KDD is increasingly held accountable for keeping its end of the bargain. We argued that making KDD deliver on its expectations is as much about addressing real-world issues, seeing the larger business picture, adhering to good business practice, and starting with the right mindset and attitude, as it is about technology. Specifically, we addressed the following themes: (1) beyond data mining; (2) design for change; (3) thinking outside the box; (4) risk-return management; (5) human-centred process; and (6) customer trust.

REFERENCES

[1] Agrawal, R., August 1999. Data mining: Crossing the chasm. Invited talk. In: Fifth ACM SIGMOD International Conference on Knowledge Discovery and Data Mining. San Diego, California, USA. http://www.almaden.ibm.com/cs/quest/papers/kdd99_chasm.ppt

[2] Bartlett, J., 1997. Managing programmes of business change. PMT Books.

[3] Brachman, R., Anand, T., 1996. The process of knowledge discovery in databases: A human-centered approach. In: Fayyad, U., Piatetsky-Shapiro, G., Uthurusamy, R. (Eds.), Advances in knowledge discovery and data mining. AAAI/MIT Press.

[4] Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., Zanasi, A., 1997. Discovering data mining: From concept to implementation. Prentice Hall.

[5] Chaudhuri, S., 1998. Data mining and database systems: Where is the intersection? Data Engineering Bulletin 21 (1), 4–8.

[6] Chen, M.-S., Han, J., Yu, P., 1996. Data mining: An overview from a database perspective. IEEE Transactions on Knowledge and Data Engineering 8 (6), 866–883.

[7] Cranor, L. (Ed.), 1999. Special issue on internet privacy. Vol. 42 (2) of Communications of the ACM.

[8] CRISP-DM Consortium, 2002. Cross-industry standard process for data mining. http://www.crisp-dm.org/

[9] Data Mining Group, 2002. Predictive model markup language. http://www.dmg.org/

[10] D'Souza, D., Wills, A., 1998. Objects, components, and frameworks with UML: The Catalysis approach. Addison-Wesley.

[11] Dumortier, J., March 2001. Legal aspects of electronic business. E-Business Chair Lectures, K.U.Leuven, Department of Applied Economics, Leuven, Belgium. http://www.econ.kuleuven.ac.be/leerstoel/e-business/downloads/LegalAspects.zip

[12] Economist (print edition), May 1st, 1999. The end of privacy.

[13] Estivill-Castro, V., Brankovic, L., Dowe, D., 1999. Privacy in data mining. Privacy Law & Policy Reporter 6 (3), 33–35.

[14] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: An overview. In: Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy, R. (Eds.), Advances in knowldge discovery and data mining. AAAI/MIT Press.

[15] Frawley, W., Piatetsky-Shapiro, G., Matheus, C., 1992. Knowledge discovery in databases: An overview. AI Magazine 13 (3), 57–70.

[16] Glymour, C., Madigan, D., Pregibon, D., Smyth, P., 1997. Statistical themes and lessons for data mining. Data Mining and Knowledge Discovery 1 (1), 11–28.

[17] Grossman, R., Hornick, M., Meyer, G., 2002. Data mining standards initiatives. Communications of the ACM 45 (8), 59–61.

[18] Hand, D., 1998. Data mining: Statistics and more? American Statistician 52 (2), 112–118.

[19] Hand, D., 1999. Statistics and data mining: Intersecting disciplines. SIGKDD Explorations 1 (1), 16–19.

[20] Hand, D., 2000. Data mining: New challenges for statisticians. Social Science Computer Review 18 (4), 442–449.

[21] Henderson-Sellers, B., Bulthuis, A., 1997. Object-oriented metamethods. Springer.

[22] Hipp, J., Güntzer, U., Grimmer, U., May 2001. Data quality mining - Making a virtue of necessity. In: Sixth ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Santa Barbara, California, USA.

[23] Hirji, K., 2001. Exploring data mining implementation. Communications of the ACM 44 (7), 87–93.

[24] Hurewitz, B., November 2001. Current developments in US privacy regulations. Harvard Information Infrastructure Project Seminars, Harvard University, John F. Kennedy School of Government, Cambridge, Massachusetts, USA. http://www.ksg.harvard.edu/iip/seminars/HurewitzPresentation112601_files/frame.htm

[25] Hyperion Solutions, Microsoft Corporation, SAS, 2002. XML for analysis. http://xmla.org/

[26] Imielinski, T., Mannila, H., 1996. A database perspective on knowledge discovery. Communications of the ACM 39 (11), 58–64.

[27] Jensen, D., 2000. Data snooping, dredging and fishing: The dark side of data mining. SIGKDD Explorations 1 (2), 52–54.

[28] Keen, P., Ballance, C., Chan, S., Schrump, S., 1999. Electronic commerce relationships: Trust by design. Prentice Hall.

[29] Kohavi, R., July 1998. Crossing the chasm: From academic machine learning to commercial data mining. Invited talk. In: Fifteenth International Conference on Machine Learning. Madison, Wisconsin, USA. http://robotics.stanford.edu/~ronnyk/chasm.pdf

[30] Kotásek, P., 2002. XML data mining specification language. http://www.fee.vutbr.cz/~kotasekp/xdmsl/.iso-8859-1

[31] Lambert, D., May 2000. What use is statistics for massive data? Invited talk. In: Fifth ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. Dallas, Texas, USA.

[32] Meyer, B., 1997. Object-oriented software construction. Prentice Hall.

[33] National Center for Data Mining, 2002. Data space transfer protocol. http://www.dataspaceweb.net/dstp.htm

[34] Netz, A., Chaudhuri, S., Fayyad, U., Bernhardt, J., April 2001. Integrating data mining with SQL databases: OLE-DB for data mining. In: Seventeenth IEEE International Conference on Data Engineering. Heidelberg, Germany.

[35] Object Management Group, 2002a. Common warehouse metamodel. http://www.omg.org/technology/cwm/

[36] Object Management Group, 2002b. OMG model driven architecture. http://www.omg.org/mda/

[37] Piatetsky-Shapiro, G. (Ed.), 1995. Knowledge discovery in personal data vs. privacy: A mini-symposium. Vol. 10(2) of IEEE Expert.

[38] Piatetsky-Shapiro, G., 2000. Knowledge discovery in databases: 10 years after. SIGKDD Explorations 1 (2), 59–61.

[39] Pine II, B., Gilmore, J., 1999. The experience economy. Harvard Business School Press.

[40] Polanyi, M., 1974. Personal knowledge: Towards a post-critical philosophy. University of Chicago Press.

[41] Reber, A., 1995. Implicit learning and tacit knowledge: An essay on the cognitive unconscious. Oxford University Press.

[42] Sarawagi, S., Thomas, S., Agrawal, R., 2000. Integrating mining with relational database systems: Alternatives and implications. Data Mining and Knowledge Discovery 4 (2-3), 89–125.

[43] SAS, 1996. Data mining with the SAS system: From data to

business advantage. http://www.sas.com/

[44] Snoeck, M., Dedene, G., Verhelst, M., Depuydt, A.-M., 1999. Object-oriented enterprise modelling with MERODE. Leuven University Press.

[45] Sowa, J., Zachman, J., 1992. Extending and formalizing the framework for information systems architecture. IBM Systems Journal 31 (3), 590–616.

[46] Thearling, K., March 17th, 1998. Data mining and privacy: A conflict in the making? http://www.tgc.com/dsstar/98/0317/100128.html

[47] Viaene, S., Dedene, G., May 2003. Coming to terms with the new economics of information. In: Fourteenth IRMA International Conference. Philadelphia, Pennsylvania, USA.

[48] Wideman, R., 1992. Project and program risk management: A guide to managing project risks and opportunities. Project Management Institute.