Ijesm
Consulting, help, relaxation
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & MANAGEMENT

DATA MINING: INTRODUCTION, PROCESS, APPLICATION AND SURVEY REPORT

Pankaj Sahu*and Arif Khan[2]
Central India Institute of Technology, Indore, (M.P.)

Abstract

Data mining, a branch of computer science is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery. The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations. In this paper, we present the overview of data mining and data mining process and related data mining technique.

Keywords: Data fishing, Data snooping, Data dredging

Introduction

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes' theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has increased data collection, storage and manipulations. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1980s). Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns.[2] It has been used for many years by businesses, scientists and governments to sift through volumes of data such as airline passenger trip records, census data and supermarket scanner data to produce market research reports. (Note, however, that reporting is not always considered to be data mining.) [1]

*** Corresponding Author**
E-Mail: Sahu.pankaj07@gmail.com

A primary reason for using data mining is to assist in the analysis of collections of observations of behaviour. Such data are vulnerable to collinearity because of unknown interrelations. An unavoidable fact of data mining is that the (sub-) set(s) of data being analysed may not be representative of the whole domain, and therefore may not contain examples of certain critical relationships and behaviours that exist across other parts of the domain. To address this sort of issue, the analysis may be augmented using experiment-based and other approaches, such as Choice Modelling for human-generated data. In these situations, inherent correlations can be either controlled for, or removed altogether, during the construction of the experimental design.

There have been some efforts to define standards for data mining, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0). These are evolving standards; later versions of these standards are under development. Independent of these standardization efforts, freely available open-source software systems like the R Project, Weka, KNIME, RapidMiner, jHepWork and others have become an informal standard for defining data-mining processes. Notably, all these systems are able to import

and export models in PMML (Predictive Model Markup Language) which provides a standard way to represent data mining models so that these can be shared between different statistical applications.[3] PMML is an XML-based language developed by the Data Mining Group (DMG),[4] an independent group composed of many data mining companies. PMML version 4.0 was released in June 2009.[4][5][6]

Research and evolution

In addition to industry driven demand for standards and interoperability, professional and academic activity have also made considerable contributions to the evolution and rigour of the methods and models; an article published in a 2008 issue of the *International Journal of Information Technology and Decision Making* summarises the results of a literature survey which traces and analyzes this evolution.[7]

The premier professional body in the field is the Association for Computing Machinery's Special Interest Group on Knowledge discovery and Data Mining (SIGKDD).[*citation needed*] Since 1989 they have hosted an annual international conference and published its proceedings,[8] and since 1999 have published a biannual academic journal titled "SIGKDD Explorations".[9] Other Computer Science conferences on data mining include:

- DMIN – International Conference on Data Mining[10]
- DMKD – Research Issues on Data Mining and Knowledge Discovery.
- ECDM – European Conference on Data Mining.
- ECML-PKDD – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases.
- EDM – International Conference on Educational Data Mining.
- ICDM – IEEE International Conference on Data Mining[11]
- MLDM – Machine Learning and Data Mining in Pattern Recognition.
- PAKDD – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining.
- PAW – Predictive Analytics World[12].
- SDM – SIAM International Conference on Data Mining

The paper is organized as follows: section II gives the overview of process of data mining, section III gives the notable uses of data mining in various areas, section IV gives the privacy concern and ethics of data mining, section V gives marketplace survey of data mining and Section VI gives conclusion of the paper.

Process of data mining

Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns already present in the data, the target dataset must be large enough to contain these patterns while remaining concise enough to be mined in an acceptable timeframe. A common source for data is a datamart or data warehouse. Pre-process is essential to analyse the multivariate datasets before clustering or data mining.

The target set is then cleaned. Cleaning removes the observations with noise and missing data. The clean data are reduced into feature vectors, one vector per observation. A feature vector is a summarised version of the raw data observation. For example, a black and white image of a face which is 100px by 100px would contain 10,000 bits of raw data. This might be turned into a feature vector by locating the eyes and mouth in the image. Doing so would reduce the data for each vector from 10,000 bits to three codes for the locations, dramatically reducing the size of the dataset to be mined, and hence reducing the processing effort. The feature(s) selected will depend on what the objective(s) is/are; obviously, selecting the "right" feature(s) is fundamental to successful data mining.

The feature vectors are divided into two sets, the "training set" and the "test set". The training set is used to "train" the data mining algorithm(s), while the test set is used to verify the accuracy of any patterns found.

Data mining

Data mining commonly involves four classes of tasks:[13]

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.
- Regression – Attempts to find a function which models the data with the least error.
- Association rule learning – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information

for marketing purposes. This is sometimes referred to as market basket analysis.

Results validation

The final step of knowledge discovery from data is to verify the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set, this is called overfitting. To overcome this, the evaluation uses a test set of data which the data mining algorithm was not trained on. The learnt patterns are applied to this test set and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish spam from legitimate emails would be trained on a training set of sample emails. Once trained, the learnt patterns would be applied to the test set of emails which it had not been trained on, the accuracy of these patterns can then be measured from how many emails they correctly classify. A number of statistical methods may be used to evaluate the algorithm such as ROC curves.

If the learnt patterns do not meet the desired standards, then it is necessary to reevaluate and change the preprocessing and data mining. If the learnt patterns do meet the desired standards then the final step is to interpret the learnt patterns and turn them into knowledge.

Notable uses

Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called tablebases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened up. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully have the required high level of abstraction in order to be applied successfully. Instead, extensive experimentation with the tablebases, combined with an intensive study of tablebase-answers to well designed problems and with knowledge of prior art, i.e. pre-tablebase knowledge, is used to yield insightful patterns. Berlekamp in dots-and-boxes etc. and John Nunn in chess endgames are notable examples of researchers doing this work, though they were not and are not involved in tablebase generation.

Business

Data mining in customer relationship management applications can contribute significantly to the bottom line.[*citation needed*] Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimise resources across campaigns so that one may predict which channel and which offer an individual is most likely to respond to—across all potential offers. Additionally, sophisticated applications could be used to automate the mailing. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this "sophisticated application" can either automatically send an e-mail or regular mail. Finally, in cases where many people will take an action without an offer, uplift modeling can be used to determine which people will have the greatest increase in responding if given an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

Businesses employing data mining may see a return on investment, but also they recognise that the number of predictive models can quickly become very large. Rather than one model to predict how many customers will churn, a business could build a separate model for each region and customer type. Then instead of sending an offer to all people that are likely to churn, it may only want to send offers to customers. And finally, it may also want to determine which customers are going to be profitable over a window of time and only send the offers to those that are likely to be profitable. In order to maintain this quantity of models, they need to manage model versions and move to *automated data mining*.

Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees. Information obtained, such as universities attended by highly successful employees, can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.[14]

Another example of data mining, often called the market basket analysis, relates to its use in retail sales.

If a clothing store records the purchases of customers, a data-mining system could identify those customers who favour silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical or inexact rules may also be present within a database. In a manufacturing application, an inexact rule may state that 73% of products which have a specific defect or problem will develop a secondary problem within the next six months.

Market basket analysis has also been used to identify the purchase patterns of the Alpha consumer. Alpha Consumers are people that play a key role in connecting with the concept behind a product, then adopting that product, and finally validating it for the rest of society. Analyzing the data collected on this type of users has allowed companies to predict future buying trends and forecast supply demands.

Data Mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich history of customer transactions on millions of customers dating back several years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

Related to an integrated-circuit production line, an example of data mining is described in the paper "Mining IC Test Data to Optimize VLSI Testing."[15] In this paper the application of data mining and decision analysis to the problem of die-level functional test is described. Experiments mentioned in this paper demonstrate the ability of applying a system of mining historical die-test data to create a probabilistic model of patterns of die failure which are then utilised to decide in real time which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products.

Science and engineering

In recent years, data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

In the area of study on human genetics, an important goal is to understand the mapping relationship between the inter-individual variation in human DNA sequences and variability in disease susceptibility. In lay terms, it is to find out how the changes in an individual's DNA sequence affect the risk of developing common diseases such as cancer. This is very important to help improve the diagnosis, prevention and treatment of the diseases. The data mining technique that is used to perform this task is known as multifactor dimensionality reduction.[16]

In the area of electrical power engineering, data mining techniques have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on the insulation's health status of the equipment. Data clustering such as self-organizing map (SOM) has been applied on the vibration monitoring and analysis of transformer on-load tap-changers(OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for exactly the same tap position. SOM has been applied to detect abnormal conditions and to estimate the nature of the abnormalities.[17]

Data mining techniques have also been applied for dissolved gas analysis (DGA) on power transformers. DGA, as a diagnostics for power transformer, has been available for many years. Data mining techniques such as SOM has been applied to analyse data and to determine trends which are not obvious to the standard DGA ratio techniques such as Duval Triangle.[17]

A fourth area of application for data mining in science/engineering is within educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning[18] and to understand the factors influencing university student retention.[19] A similar example of the social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalised and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate Institutional memory. Other examples of applying data mining technique applications are biomedical data facilitated

by domain ontologies,[20] mining clinical trial data,[21] traffic analysis using SOM,[22].

In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents.[23] Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.[24]

Spatial data mining

Spatial data mining is the application of data mining techniques to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasises the importance of developing data driven inductive approaches to geographical analysis and modeling.

Data mining, which is the partially automated search for hidden patterns in large databases, offers great potential benefits for applied GIS-based decision-making. Recently, the task of integrating these two technologies has become critical, especially as various public and private sector organisations possessing huge databases with thematic and geographically referenced data begin to realise the huge potential of the information hidden there. Among those organisations are:

- offices requiring analysis or dissemination of geo-referenced statistical data
- public health services searching for explanations of disease clusters
- environmental agencies assessing the impact of changing land-use patterns on climate change
- Geo-marketing companies doing customer segmentation based on spatial location.

Challenges

Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components, that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management.[25] Related to this is the range and diversity of geographic data formats, that also presents unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data such as imagery and geo-referenced multi-media.[26]

There are several critical research challenges in geographic knowledge discovery and data mining. Miller and Han [27] offer the following list of emerging research topics in the field:

- **Developing and supporting geographic data warehouses** – Spatial properties are often reduced to simple aspatial attributes in mainstream data warehouses. Creating an integrated GDW requires solving issues in spatial and temporal data interoperability, including differences in semantics, referencing systems, geometry, accuracy and position.
- **Better spatio-temporal representations in geographic knowledge discovery** – Current geographic knowledge discovery (GKD) techniques generally use very simple representations of geographic objects and spatial relationships. Geographic data mining techniques should recognise more complex geographic objects (lines and polygons) and relationships (non-Euclidean distances, direction, connectivity and interaction through attributed geographic space such as terrain). Time needs to be more fully integrated into these geographic representations and relationships.
- **Geographic knowledge discovery using diverse data types** – GKD techniques should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia, as well as dynamic data types (video streams, animation).

Surveillance

Previous data mining to stop terrorist programs under the U.S. government include the Total Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II)), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE[28]), and the Multi-state Anti-Terrorism Information Exchange (MATRIX).[29] These programs have been discontinued due to controversy over whether they violate the US Constitution's 4th amendment, although many programs that were formed

under them continue to be funded by different organisations, or under different names.[30]

Two plausible data mining techniques in the context of combating terrorism include "pattern mining" and "subject-based data mining".

Pattern mining

"Pattern mining" is a data mining technique that involves finding existing patterns in data. In this context *patterns* often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behaviour in terms of the purchased products. For example, an association rule "beer ⇒ potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise."[31][32][33] Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search techniques.

Subject-based data mining

"Subject-based data mining" is a data mining technique involving the search for associations between individuals in data. In the context of combatting terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."[32]

Applications

- Customer analytics
- Data Mining in Agriculture
- National Security Agency
- Police-enforced ANPR in the UK
- Quantitative structure-activity relationship
- Surveillance / Mass surveillance (e.g., Stellar wind (code name))

Privacy concerns and ethics

Some people believe that data mining itself is ethically neutral.[34] It is important to note that the term data mining has no ethical implications. The term is often associated with the mining of information in relation to peoples' behavior. However, data mining is a statistical technique that is applied to a set of information, or a data set. Associating these data sets with people is an extreme narrowing of the types of data that are available in today's technological society. Examples could range from a set of crash test data for passenger vehicles, to the performance of a group of stocks. These types of data sets make up a great proportion of the information available to be acted on by data mining techniques, and rarely have ethical concerns associated with them. However, the ways in which data mining can be used can raise questions regarding privacy, legality, and ethics.[35] In particular, data mining government or commercial data sets for national security or law enforcement purposes, such as in the Total Information Awareness Program or in ADVISE, has raised privacy concerns.[36][37]

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed.[38] This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous. It is recommended that an individual is made aware of the following before data are collected:

- the purpose of the data collection and any data mining projects,
- how the data will be used,
- who will be able to mine the data and use them,
- the security surrounding access to the data, and in addition,
- How collected data can be updated.[38]

In the United States, privacy concerns have been somewhat addressed by their congress via the passage of regulatory controls such as the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA requires individuals to be given "informed consent" regarding any information that they provide

and its intended future uses by the facility receiving that information. According to an article in Biotech Business Week, "In practice, HIPAA may not offer any greater protection than the longstanding regulations in the research arena, says the AAHC. More importantly, the rule's goal of protection through informed consent is undermined by the complexity of consent forms that are required of patients and participants, which approach a level of incomprehensibility to average individuals." [39] This underscores the necessity for data anonymity in data aggregation practices.

One may additionally modify the data so that they are anonymous, so that individuals may not be readily identified.[38] However, even de-identified data sets can contain enough information to identify individuals, as occurred when journalists were able to find several individuals based on a set of search histories that were inadvertently released by AOL.[40]

Marketplace surveys

Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners. Some of these reports include:

- Forrester Research 2010 Predictive Analytics and Data Mining Solutions report.[41]
- Annual Rexer Analytics Data Miner Surveys.[42][43][44]
- Gartner 2008 "Magic Quadrant" report.[45]
- Robert Nisbet's 2006 Three Part Series of articles "Data Mining Tools: Which One is Best For CRM?"[46]
- Haughton et al.'s 2003 Review of Data Mining Software Packages in *The American Statistician*.[47]

Conclusion

Data mining is the major research area in many aspects. Data mining is the process to extract useful information from database and also creates useful pattern. In this paper we present the detailed overview of data mining.

References

1. Clifton, Christopher (2010). "Encyclopedia Britannica: Definition of Data Mining". http://www.britannica.com/EBchecked/topic/1056150/data-mining. Retrieved 2010-12-09.
2. Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons. ISBN 0471228524. OCLC 50055336.
3. Alex Guazzelli, Wen-Ching Lin, Tridivesh Jena. PMML in Action: Unleashing the Power of Open Standards for Data Mining and Predictive Analytics. CreateSpace, 2010**.**
4. The Data Mining Group (DMG). The DMG is an independent, vendor led group which develops data mining standards, such as the Predictive Model Markup Language (PMML).
5. Y. Peng, G. Kou, Y. Shi, Z. Chen (2008). "A Descriptive Framework for the Field of Data Mining and Knowledge Discovery". *International Journal of Information Technology and Decision Making, Volume 7, Issue 4* **7**: 639–682. doi:10.1142/S0219622008003204.
6. Proceedings, International Conferences on Knowledge Discovery and Data Mining, ACM, New York.
7. Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases". http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf. Retrieved 2008-12-17.
8. Xingquan Zhu, Ian Davidson (2007). *Knowledge Discovery and Data Mining: Challenges and Realities*. Hershey, New Your. p. 18. ISBN 978-159904252-7.
9. Yudong Chen, Yi Zhang, Jianming Hu, Xiang Li. "Traffic Data Analysis Using Kernel PCA and Self-Organizing Map". *Intelligent Vehicles Symposium, 2006 IEEE*.
10. Norén GN, Bate A, Hopstadius J, Star K, Edwards IR. Temporal Pattern Discovery for Trends and Transient Effects: Its Application to Patient Records. *Proceedings of the Fourteenth International Conference on Knowledge Discovery and Data Mining SIGKDD 2008*, pages 963–971. Las Vegas NV, 2008.
11. Healey, R., 1991, Database Management Systems. In Maguire, D., Goodchild, M.F., and Rhind, D., (eds.), Geographic Information Systems: Principles and Applications (London: Longman).
12. Government Accountability Office, *Data Mining: Early Attention to Privacy in Developing a Key DHS Program Could Reduce Risks*, GAO-07-293, Washington, D.C.: February 2007.
13. R. Agrawal et al., *Fast discovery of association rules*, in *Advances in knowledge discovery and data mining* pp. 307–328, MIT Press, 1996.
14. Stephen Haag et al. (2006). *Management Information Systems for the information age*.

Toronto: McGraw-Hill Ryerson. p. 28. ISBN 0-07-095569-7. OCLC 63194770.

15. William Seltzer. *The Promise and Pitfalls of Data Mining: Ethical Issues*. http://www.amstat.org/committees/ethics/linksdir/Jsm2005Seltzer.pdf.
16. K.A. Taipale (December 15, 2003). "Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data". *Columbia Science and Technology Law Review* **5** (2). SSRN 546782 / OCLC 45263753. http://www.stlr.org/cite.cgi?volume=5&article=2.
17. *Think Before You Dig: Privacy Implications of Data Mining & Aggregation*, NASCIO Research Brief, September 2004.
18. Biotech Business Week Editors. (June 30, 2008). BIOMEDICINE; HIPAA Privacy Rule Impedes Biomedical Research. Biotech Business Week. Retrieved 17 Nov 2009 from LexisNexis Academic.
19. *AOL search data identified individuals*,* SecurityFocus, August 2006.
20. James Kobielus (1 July 2008) *The Forrester Wave: Predictive Analytics and Data Mining Solutions, Q1 2010*, Forrester Research.
21. Gareth Herschel (1 July 2008) *Magic Quadrant for Customer Data-Mining Applications*, Gartner Inc.