Appears in the **Communications of the ACM,**
Volume 45, Number 8, Aug 2002, pages 45-48
Note: This is the final draft prior to copy-editing done by CACM

Emerging Trends in Business Analytics

Ron Kohavi
Blue Martini Software
2600 Campus Drive
San Mateo, CA 94403
ronnyk@cs.stanford.edu

Neal J. Rothleder
DigiMine, Inc.
10500 NE 8th St., Suite 1300
Bellevue, WA 98004
nealr@digimine.com

Evangelos Simoudis
Apax Partners
2100 Geng Road
Palo Alto, CA 94303
evangelos.simoudis@apax.com

Introduction

The field of business analytics has improved significantly over the last few years, providing business users with better insights, particularly from operational data stored in transactional systems. As an illustrative example, analysis of e-commerce data has recently come to be considered a killer-app for data mining [1,2]. The data sets created by integrating clickstream records generated by web sites with demographic and other behavioral data dwarf, in size and complexity, the largest data warehouses of a few years ago [3], creating massive databases that require a mix of automated analysis techniques and human effort in order to provide business users with critical insight about the activity on the site and the characteristics of the site's visitors and customers. With many millions of clickstream records being generated on a daily basis and aggregated to records with hundreds of attributes, there is a clear need for automated techniques to find patterns in the data. In this paper we discuss the technology and enterprise-adoption trends in the area of business analytics.

The key consumer of these analytics is the *business user*, a person whose job is not directly related to analytics per-se (e.g., a merchandiser, marketer, salesperson), but who typically must use analytical tools to improve the results of a business process along one or more dimensions (e.g., profit, time to market). Fortunately, data mining[1], analytic applications, and business intelligence systems are now being better integrated with transactional systems creating a closed loop between operations and analyses that allows data to be analyzed faster and the analysis results to be quickly reflected in business actions. Mined information is being deployed to a broader business audience, which is taking advantage of business analytics in everyday activities. Analytics are now regularly used in multiple areas, including sales, marketing, supply chain optimization, and fraud detection [4, 5].

[1] Note that the terms *data mining* and *analytics* are used interchangeably here to denote the general process of exploration and analysis of data to discover and identify new and meaningful patterns in data. This definition is similar to those presented in [4] and [5] (under the term knowledge discovery).

The Business Users and their Challenges

Despite these advances in analytic systems, it continues to be the case that the business user, while an expert in his area, is unlikely also to be an expert in data analysis and statistics. To make decisions based on the data enterprises collect, the business user must either rely on a data analyst to extract information from the data, or employ analytic applications that blend data analysis technologies with task-specific knowledge. In the first case, the business user must impart domain knowledge to the analyst, then wait while the analyst organizes the data, analyzes it, and communicates back the results. These results typically raise further questions and hence several iterations are necessary before the business user can start acting on the analysis. In the second case, analytic applications must not only incorporate a variety of data mining techniques, but also provide recommendations to the business user of how to best analyze data and present the extracted information.

Business users are expected to better utilize the extracted information and improve performance along multiple metrics. Unfortunately, the gap between the relevant analytics and the critical needs of the intended business users still remains significant. The following challenges highlight characteristics of this gap:

1. The time to perform the overall cycle of collecting, analyzing, and acting on enterprise data must be reduced. While business constraints may impose limits on reducing the overall cycle time, business users want to be empowered and rely less on other people to help with these tasks.
2. Within this cycle, the time and analytic expertise necessary to analyze data must be reduced.
3. Clear business goals and metrics must be defined. In the past, unrealistic expectations about data mining "magic" led to misguided efforts without clear goals and metrics.
4. Data collection efforts must have clear goals. Once metrics are identified, organizations must strive to collect the appropriate data and transform it. In many situations, data analysis is often an afterthought, restricting the possible value of any analysis.
5. Analysis results must be distributed to a wide audience. Most analysis tools are designed for quantitative analysts, not for the broader base of business users who need the output to be translated into language and visualizations that are appropriate for the business needs.
6. Data must be integrated from multiple sources. The extract-transform-load (ETL) process is typically complex and its cost and difficulty are usually underestimated.

Trends in Business Analytics

The emerging trends and innovations discussed in this paper embody approaches to these business challenges. Indeed, it is a very healthy sign for this field that regardless of the form of the solution—process, technology, system integration, user interface, etc.–the driving force is the *business* problem.

3.1 Verticalization

In order to reduce discovery cycle time, facilitate the definition and achievement of business goals, and deploy analysis results to wider audiences, developers of analytical solutions started

verticalizing their software. The first step in this verticalization was the incorporation of task-specific knowledge. Examples include knowledge about how to analyze customer data to determine the effectiveness of a marketing campaign, knowledge on how to analyze clickstream data generated by a web site to reduce shopping cart abandonment and improve ad effectiveness, knowledge about how an investment bank consolidates its general ledger and is able to produce various types of forecasts, or how an insurance company is able to analyze data in order to provide an optimally-priced policy to an existing customer.

In the process of incorporating industry-specific knowledge, companies are also able to optimize the performance of their applications for the specific verticals. For example, a company that developed an analytic application for budgeting and forecasting targeted at the financial services industry determined that its OLAP engine's execution speed could be optimized by limiting the number of dimensions handled by the engine to nine, a number deemed as sufficient for the particular application in that industry.

The use of industry-specific knowledge is not limited to the data mining components of analytic applications but also affects how the extracted information is presented and accessed. For example, enterprises from the financial services, retail, manufacturing, utilities, and telecommunications industries increasingly expect their field personnel to have access to business analytic information through wireless devices that they carry. Analytic applications companies are now working on technologies that will automatically detect the type of wireless device and its form factor, and will automatically tailor analysis results to fit the capabilities of a particular device. For example, if the information is to be displayed on a phone supporting the Wireless Access Protocol (WAP) (implying that the screen is small) it may be necessary to automatically summarize text, abbreviate several words, and limit the use of graphics by automatically selecting only the most pertinent figures.

3.2 Comprehensible Models and Transformations for Insight

With the need to let business users analyze data and provide insight quickly, and with the goal of reducing the reliance on data mining experts, comprehensible models are more commonly used than opaque models. For example, in the KDD-CUP 2000 [2], a data mining competition in which insight was important, the use of decision trees, generally accepted as relatively easy to understand, outnumbered other methods by more than two to one.

Business users do not want to deal with advanced statistical concepts. They want straightforward visualizations and task-relevant outputs. Consider Figure 1, which summarizes a Naïve-Bayes model for predicting which people earn more than $50,000 in yearly salary. Instead of the underlying log conditional probabilities that the model actually manipulates, the visualization uses bar height to represent evidence for each value of a contributing factor listed on the left of the figure and color saturation to signify confidence of that evidence [6]. For example, evidence for higher salaries increases with age, until the last age bracket, where it drops off; evidence for higher salaries increases with years of education, with the number of hours worked, and with certain marital statuses and occupations. Note also that the visualization shows only a few attributes that were determined by the mining algorithm to be the most important ones, highlighting to the business users the most critical attributes from a larger set.

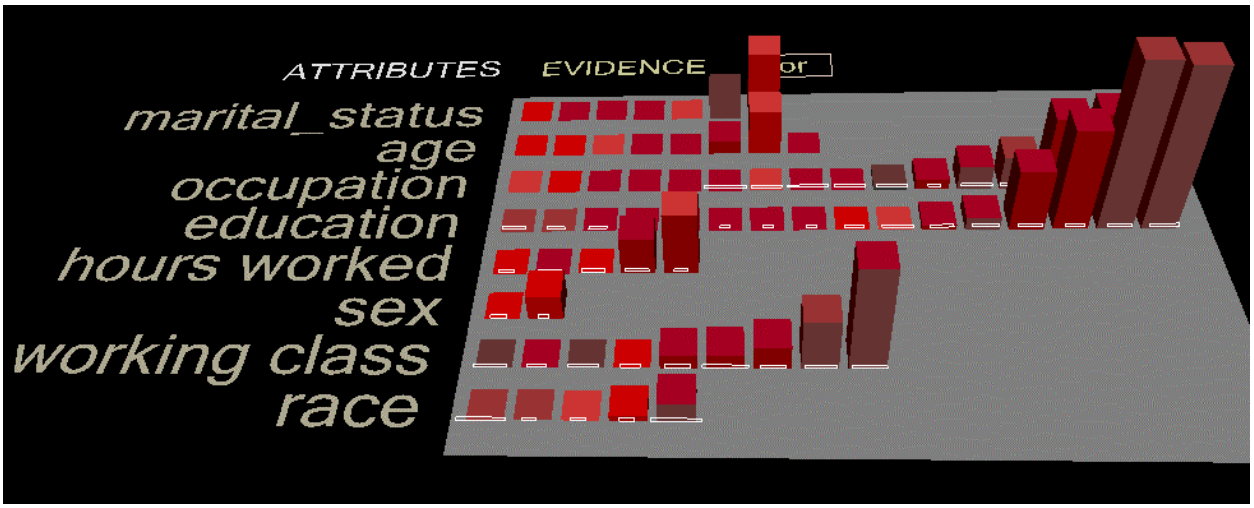

Figure 1 - A visualization of Naive Bayes

Several additional examples of visualizing data and data mining models are presented in [7] and [8].

3.3 Analytics are Part of the Larger System

The needs of data analysis are now being designed into systems, instead of being an afterthought. Specifically, the following areas are routinely addressed:

1. Data collection. You cannot analyze what you do not collect, so collecting rich data is critical. For example, e-commerce systems can collect attributes ranging from the user local time, screen resolution (useful to determine quality of images to send), and network bandwidth.
2. Generation (and storage) of unique identifiers. In order to help merge information from several records and remove duplicate records, systems must generate unique keys to join data and store them. For example, all clickstream records in the same session should store the session id so that they can later be joined to the session record stored at another table.
3. Integration with multiple data sources. Analysis is more effective when data is available from multiple sources. For example, in customer analytics it is important to merge data from multiple touchpoints, such as the web, call center, physical stores, wireless, and campaigns (both direct and online). Behavioral data can be more powerful when overlaid with demographic and socioeconomic data from other sources.
4. Hardware sizing. Analysis requires significant hardware to deal with large amounts of data. Companies have traditionally underestimated the need for sophisticated IT infrastructure and powerful hardware to make analysis feasible in a timely manner.

3.4 Analytics in New Areas

Over the past three years the analysis of customer data has attracted the most attention and has provided success in reducing customer attrition, improving customer profitability, increasing the value of e-commerce purchases, and increasing the response of direct mail and e-mail marketing campaigns. This has paved the way for new applications of business analytics to emerge. Of these new areas, three applications are particularly promising: supply chain visibility, price optimization, and workforce analysis.

Organizations have automated significant portions of their supply chain. In the process they have enabled the collection of significant data about inventory, the performance of suppliers, logistics, etc. New applications are now able to analyze this data to provide insights about the performance of suppliers and partners, material expenditures, accuracy of sales forecasts to better control materials inventory, accuracy of production plans, the accuracy of plans for order delivery, etc.

The wide adoption of CRM and Supply Chain Management software has allowed enterprises to fully interface/integrate their demand and supply chains. Based on this integration, enterprises are now able to capture up-to-the-minute data about the demand of a particular product, as well as data of similar granularity about the corresponding data's supply. By analyzing these two data streams corporations are able to optimize the price of a particular product along several dimensions so that the demand will meet the available supply. For example, the price of a product may be different through one channel, e.g., the Web, than through another, e.g., in the retail store. Such price optimizations allow corporations to maximize the profit margin of each item sold while reducing their inventory.

Once corporations have been able to analyze data about their customers and their suppliers, it is only natural for them to begin analyzing data about their employees. A new generation of analytic applications allows enterprises to identify workforce trends, such as attrition rates, and perform tasks such as compensation and benefits analyses. Companies whose cost or revenue model is dependent on hourly models, e.g., contact centers or systems integrators, are able to use this new generation of employee-centered analytics to optimize staffing levels and skills requirements in order to minimize the number of employees that are not able to bill.

3.5 Integration of Analytics with Action and Measurement

With business users' increasing understanding and experience in analytics, they are becoming more demanding and discerning, particularly in the areas of *action* and *return on investment* (ROI) [9]. Increasingly, analytics users are asking two key questions, "How do I turn discovered information into action," and "How do I know the effect of each action?" While in the past, success stories of data mining *ended* with a novel analytical result, it is increasingly necessary that solutions use the analytic results as a *starting* point towards the critical next steps–action and measurement. It is no longer enough for cluster-discovery algorithms, for example, to uncover interesting groups of customers. The successful analytic solution must make it easier for the user to grasp the significance of these clusters in the context of the business action plan: "Here are people with a propensity to purchase new fashions." Achieving these results requires non-trivial transformations from the base statistical models. Traditionally to achieve these results necessitated the use of an expert analyst.

Integration with other existing systems is a key to both action and measurement. For example, if the analytic application can identify customers likely to respond to a promotion, but it takes a cadre of IT specialists to incorporate the relevant data into the advertising system to run and execute the promotion, then the results will not be used, as IT specialists are in short supply within the enterprise. Similarly, if promotion targeting solutions enable distribution of catalogs with optimized promotions, but the order submission system isn't tied closely back into the analytics, the resulting lag in the ROI reports will prevent a timely adjustment in the next catalog mailing. These types of integration between operation and analytics systems have seen major initiatives in the last five years, including entire products whose value proposition is precisely the optimization of the collect-analyze-act-measure cycle.

Summary and Conclusions

The innovations and trends of business analytics spanning the areas of process, new technologies, user interface design, and system integration, are being driven by *business value*. In this paper, business value is measured as progress towards bridging the gap between the needs of the business user and the accessibility and usability of analytic tools. In an effort to make analytics more relevant and tangible to the business user, solutions are focusing on specific vertical applications and tailoring the results and interfaces towards the business audience, so they are comprehensible and provide human-level insight. For ease of use, simpler and effective deployment, and optimum value, analytics are being embedded in larger systems. Consequently, issues such as data collection, storage, and processing specific to analytics are increasingly considered important issues in overall system design. In efforts to broaden the effectiveness of analytics in the business process, solutions are emerging that go beyond the customer facing applications, reaching "behind the scenes" to applications in sales, marketing, supply chain visibility, price optimizations, and workforce analysis. Finally, in order to achieve full impact and value, an increasing number of analytic solutions are making results actionable and measurement of changes key components.

Acknowledgements

The authors would like to thank Michael Berry and Gregory Piatetsky-Shapiro for their comments on an early draft of this paper.

Bibliography

[1] Ron Kohavi and Foster Provost. Applications of data mining to electronic commerce. Data Mining and Knowledge Discovery, 5(1/2), 2001. http://robotics.Stanford.EDU/users/ronnyk/ecommerce-dm.
[2] Ron Kohavi, Carla Brodley, Brian Frasca, Llew Mason, and Zijian Zheng. KDD-Cup 2000 organizers' report: Peeling the onion. SIGKDD Explorations, 2(2):86-98, 2000. http://www.ecn.purdue.edu/KDDCUP.
[3] Ralph Kimball and Richard Merz. The Data Webhouse Toolkit: Building the Web-Enabled Data Warehouse. John Wiley & Sons, 2000.
[4] Michael J. A. Berry and Gordon S. Linoff. Mastering Data Mining. John Wiley & Sons, Inc, 2000.
[5] Usama M. Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery: An overview. In Advances in Knowledge Discovery and Data Mining, chapter 1, pages 1-34. AAAI Press and the MIT Press, 1996.

[6] Barry Becker, Ron Kohavi, and Dan Sommerfield. Visualizing the simple Bayesian classifier. In Information Visualization in Data Mining and Knowledge Discovery, chapter 18, pages 237-249. Morgan Kaufmann Publishers, San Francisco, 2001.
[7] Kurt Thearling, Barry Becker, Dennis DeCoste, Bill Mawby, Michel Pilote, and Dan Sommerfield. Visualizing data mining models. In Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers, San Francisco, 2001.
[8] Juhnyoung Lee, Mark Podlaseck, and Edith Schonberg and Robert Hoch. Visualization and analysis of clickstream data of online stores for understanding web merchandising. Data Mining and Knowledge Discovery, 5(1/2), 2001.
[9] Randy Souza, Harley Manning, and Katharine M. Gardiner. How to measure what matters. The Forrester Report, May 2001.