IST 659

Sam Edelstein

Best Practices Paper: Data Mining

November 26, 2014

Introduction

Imagine a world where a business had access to a wealth of information that had the opportunity to make or break the company, and the amount of **data** was set to increase by hundreds of times by the end of the decade. The International Data Corporation (IDC) reports that from the year 2005 through 2020, the "digital universe will grow by a factor of 300, from 130 **extabytes** to 40,000 extabytes." (Gantz and Reinsel, 2012) The possibilities surrounding this data are endless and limitless, and yet almost all of it sits without anything happening. Why would this be? The vast majority of this data is **unstructured**. Unlike **structured** data, which are traditionally stored in tabular form and contain information like dates, names, and more, unstructured data can contain images, long strings of text, and other information that may not fit easily into a traditional database, and does not come to the user as a clean, and easily analyzed data set (Hoffer, Ramesh, & Topi, 2013, 419). Proper analysis, through **data mining**, of this unstructured data will be a key tool for businesses in the future and those companies that master the technique will have a competitive advantage. Data mining is "a term that refers to the use of algorithms and computers to discover novel and interesting patterns within data." (Stanton, 2013, 172) The best practice for data mining is that all organizations should invest in this technique and stay up to date on how and where to

mine data, but should also know that it is not the ultimate key to success. Instead it serves as an important piece of the equation when determining customer interests and behavior.

Best Practice

The ability to gain insights into customer data, paired with the amount of data available in one form or another makes pursuing data mining as a business strategy an important consideration. There are many different ways to build the algorithms that interpret data – through coding software like **R** or **NoSQL** or by hiring one of the many consulting companies that can do the work and provide the results. The number of available systems to do this work are constantly increasing, and they are becoming cheaper to implement (Waxer, 2013).

Additionally, the sources available for data mining constantly increase. A decade ago, sites like Twitter, YouTube, and Facebook either didn't exist, or barely existed. Online shopping was not as common as it is today. This serves as both a blessing and a curse. The opportunities for collecting data for analysis are nearly endless, but this also means it is difficult to know where to start. Organizations that decide to invest in data mining must stay current on the most efficient and effective ways to do data mining. This means at some point, an organization may not need to hire their own data scientist to do the work. Organizations also must pay attention to the data sources that can provide them the best information. A decade from today, Facebook and Twitter may be very different, or might not exist. A different, much more powerful service, may be created. It will be critical for businesses to know when to start mining different data sources, or else they will be left behind.

But, even if the organization is able to keep current with appropriate technologies and data sources, it still needs to maintain the appropriate understanding of where the findings fit within their larger strategies. Just knowing a married man in his mid-30s is likely to buy diapers and beer during an evening on a Wednesday is important, but the organization – in this case a grocery store – also needs to put this information into use appropriately. The grocery store still needs to provide a positive shopping experience, needs to be clean, and needs to remain a place where other shoppers will find products they want, too. Otherwise, the grocery store will likely go out of business, despite knowing a lot about a certain segment of its potential customer base.

Assessment

Both well-known as well as smaller organizations have used data mining to help build their business. Amazon is likely the best known, and one where many people have seen data mining at work first-hand. When browsing for products on Amazon, customers are shown other products that they are likely to want. This is based on huge amounts of data that Amazon has collected and analyzed (Simon, 2014). This success is revolutionary, but also becomes a self-fulfilling prophecy. Because Amazon is so good at selling products, more people buy, which gains Amazon more information about what people like to buy, which helps them sell products better. Not all companies have this market-share on sales and information. Additionally, Amazon's shopping experience is fast and easy. Without an easy shopping experience, it would not matter how much information Amazon had about potential customers, because they would not come to shop at Amazon. Amazon has put data mining best practices to work, heavily investing in

the proper algorithms, but also finding the correct ways to use the information as a part of their bigger business strategy.

A small business, a women's fashion retailer called Sway, also recognized the need to know more about their customers. Sway had recently launched a new website and were sending periodic newsletters that were very general. Without much feedback or many conversions, the managers at Sway invested in data mining their customer information and found a payoff of 300% in increased revenues. This was done by splitting their customers into segments based on when they open the Sway newsletter, when they visit the Sway website, and more. Based on this information, Sway sent out emails with different messaging to each of these segments. The results were positive (Waxer, 2013). Just investing in data mining, though, would not have solved Sway's problems. The newsletter would still need to be well written, and Sway would still need to offer products that were appealing to its customers.

When data mining best practices are not followed, the results can be devastating. J.C. Penney's former CEO Ron Johnson decided to invest heavily in data analysis. The results told him to change the products being offered in the store. He deemphasized ecommerce and also ended many of the product lines unique to J.C. Penney. Johnson did not consider the whole picture, though, and the company quickly lost money. Johnson was also fired (Healey, 2013). Since so much data exists, there can also be many findings that potentially look significant. This best practice is critical because managers and decision makers need to find the **signals** in all the **noise**, and also need to know their business well so their decisions based on the data are well-informed and ultimately correct.

Conclusions and Recommendations

Data mining techniques are available to all kinds of organizations, not just for-profit institutions. Non-profits looking to fundraise can use these techniques to determine their donor's interests, helping to make a fundraising appeal more meaningful. Even government can use these techniques to determine how constituents are feeling about the direction of the community or country. No matter what, though, the idea behind this best practice remains: while data mining holds the potential for a large amount of benefit and should be used as a business strategy, it is critical to analyze the data while thinking about the business as a whole. Failure to do so could result in catastrophe.

Data mining best practices should be implemented in six steps. (1) First, an organization needs to determine its resources available. A large corporation may be able to hire a team of data scientists who can customize results to particular business units within the company. A smaller company may need to hire a consulting firm for short-term data mining work. A non-profit might find an intern with some machine learning experience to give them a sense of what kind of insights they might find with more resources dedicated to the task. (2) Next, appropriate mediums for analysis should be determined and acquired, and a plan acquisition of data should be created. This would include what kind of software is needed, and where the data should come from. (3) Once a round of analysis is completed, bring all stakeholders together to consider the findings. Some stakeholders may have insights about why the analysis will not be effective in for this company. (4) Gain buy-in from stakeholders to follow the data. Some employees will not trust the data, and would instead want to trust their gut. While this instinct may be effective in certain cases, while implementing a data-driven strategy, it is critical that

decisions are made based on the data, not based on feelings. (5) Evaluate constantly to see how the data-driven decision is performing, and then change the algorithms based on what is learned. (6) Finally, review the methods for doing data mining analysis and investigate other opportunities that may supplement the analysis. Using these strategies will ensure best practices are used when data mining.

Glossary

Source: Many definitions taken directly from Modern Database Management, 11th Edition

Data: Stored representation of objects and events that have meaning and importance in the user's environment.

Data mining: Knowledge discovery, using sophisticated blend of techniques from traditional statistics, artificial intelligence, and computer graphics.

Extrabytes: A multiple of the unit, byte, one quintillion bytes, or one billion gigabytes.

Noise: The information that provides no value and can often mislead

NoSQL: Also known as Not Only SQL, this allows for storage and retrieval of data found in places other than traditional relational databases

R: A software programming language for statistics and visualizing information

Signal: The pieces of information that have value

Structured: Data that is often contained in a relational database or spreadsheet, and sits in a fixed field within a record.

Unstructured: No pre-defined data model, and is not organized in a pre-defined manner.

References

Gantz, J., & Reinsel, D. (2012, December 1). THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East.

Healey, M. (2013, August 16). Don't Be Blinded By Big Data - InformationWeek. Retrieved November 18, 2014, from http://www.informationweek.com/big-data/big-data-analytics/dont-be-blinded-by-big-data/d/d-id/1111199?

Hoffer, J., & Ramesh, V. (2013). *Modern database management* (11th ed., pp. 55-56, 63). Boston: Pearson.

Simon, S. (2014, July 26). Happy Birthday To Amazon, And Its Data Mining. Retrieved November 22, 2014, from http://www.npr.org/2014/07/26/335404545/happy-birthday-to-amazon-and-its-data-mining

Stanton, J. (2013). *An Introduction to Data Science* (Vol. 3). Syracuse.

Waxer, C. (2013, October 28). How data mining can boost your revenue by 300%. Retrieved November 19, 2014, from http://money.cnn.com/2013/10/28/smallbusiness/data-mining/