
Research on Business Intelligence with Data Mining Applications

Jason C. H. Chen, Napoleone Piani

School of Business Administration, Gonzaga University, Spokane, USA

Email address:

chen@gonzaga.edu (J. C. H. Chen)

To cite this article:

Jason C. H. Chen, Napoleone Piani. Research on Business Intelligence with Data Mining Applications. *International Journal of Business and Economics Research*. Vol. 6, No. 2, 2017, pp. 19-24. doi: 10.11648/j.ijber.20170602.11

Received: February 7, 2017; **Accepted:** April 14, 2017; **Published:** April 21, 2017

Abstract: Business Intelligence (BI) has become an important agenda for many top executives because they have become extremely aware of its value in providing a competitive differentiator at all levels of the organizations. This paper discusses the concepts and technologies of business intelligence, specially, data warehousing and data mining and how these can positively influence and benefit a business. Review on BI frameworks and research models for developing data warehousing and data mining are presented and analyzed. The paper also illustrates a business scenario in which the Rapidminer, a data mining tool can be used to extrapolate relevant data to a small startup ski shop.

Keywords: business intelligence (BI), data mining, rapidminer

1. Introduction

In the business world, information is king. Most firms in the world strive to obtain as much information as they possibly can, while using as little resources as possible. This information can range anywhere from competitor performance, to customer preferences, to performance of the firm itself. But having information is worth nothing if it is not used and studied appropriately. Information about a firm's competitors is close to useless if it is not applied. A firm can use this information to determine what course of action their competitors might plan on taking, using public information and calculating obvious trends. By collecting and analyzing data from customers, such as what product they have bought or what their reaction to a certain promotion was, a company can assess what items are more popular, what items are commonly bought together, and the success of a promotion that was held, not solely based on profits. The information a firm can collect about itself can give them a good idea of how their current operations are proceeding, it can help them detect fraud and bottlenecks, and it can even help them combat overproduction and having excessive inventory on hand.

The rest of the paper is organized as follows. Section 2 introduces business intelligence (BI) including characteristics of data warehousing and data mining. Sections 3 and 4

present review on BI systems and research frameworks for developing BI system. In Section 5, a business example with a BI tool (data mining) is illustrated. Finally, Section 6 presents the conclusion.

2. Business Intelligence

For many years, firms have followed a process of trial and error to find out what works for them and what doesn't. While there have always existed people to develop new ideas and systems to make the performance of their increase, whether by reducing costs, or finding more efficient ways to solve problems, it was not until recently that computers became a staple in the working world. These computers are capable of storing, processing, and outputting data on demand, making them extremely useful for everyday activities. Ken Olson, the president, chairman and founder of Digital Equipment Corporation, was quoted once in 1977 saying that "There is no reason for any individual to have a computer in his home" [13]. We see how wrong that statement has been 35 years later. With what is considered a simple computer now days, massive amounts of data can be calculated every second if needed. When these systems are applied to a business environment, this massive amount of data processing can be used to the great benefit of a company.

As Turban *et al.* [11] stated “Business Intelligence is an umbrella term that combines architectures, tools, databases, analytical tools, applications, and methodologies.” While this seems like a simple description, it encompasses so many different concepts, that it is almost mind-boggling. Another definition says that “Business intelligence (BI) software is a collection of decision support technologies for the enterprise aimed at enabling knowledge workers such as executives, managers, and analysts to make better and faster decisions” [3]. Business Intelligence strives for the collection of useful data; data that can be used for a great number of applications, from airlines to ice cream shops. By providing managers and analysts constant access to this data, they can use it to weigh decisions against each other and make the correct one. This data can come in many forms, including but not limited to performance data, situational data, current data, and historical data, all data which can be used in different situations to predict different outcomes. However, this data is of no use if it is not processed into useful information. No matter how much data is collected, if it is not relevant or not mined correctly, it will turn into a useless waste of space and time. In the following sections of this paper, several common parts of a Business Intelligence environment will be covered and their impact will be presented and explained.

2.1. Data Warehousing

For data to be analyzed and converted into useful information, it needs to be collected and stored in an accessible location to all those who will be in charge of working with it. A Data Warehouse was a concept designed for this very purpose. “A data warehouse is a repository of an organization’s electronically stored data, designed to facilitate reporting and analysis” [10]. It was a large database to store historical data from every aspect of an organization, giving managers and analysts any information about the company that they required, whenever they decided it was necessary. This easy access to data about the firm’s performance proved to be very helpful, since it would allow the firm to view specifically what went wrong, what were the current circumstances in which a problem occurred, and what steps they could take to fix these issues.

As technology progressed, and systems could store and process more information more quickly and cheaply, the possibility of storing current data became available. With data collection becoming more responsive and effective every day, companies could realistically attempt to solve problems merely days after they had happened. But with the demand of more current information increasing exponentially, systems had to be put in place to collect and process data almost immediately, creating a real-time Data Warehouse, containing up to the minute information about the firm it was being used in. For example, the Atlanta airport for its location and the time of year, the airport can board over 100 million passengers per year, which gives an average of 260 thousand passengers every 24 hours [14]. The logistics of coordinating this amount of movement without a catastrophe occurring is an extremely impressive one. By

keeping real time data of everything that is happening around the airport, most customers can be served as they require. If an airplane is delayed, whether for weather, or for technical difficulties, or any other reason, the airline in conjunction with the airport can make on-the-fly adjustments to their itineraries. And while changing gates and moving across an airport may seem like a hassle to the customer, they are put on a flight where they want to go, and get there safely.

The amount of data that can be stored and processed in a group of databases is incredibly impressive and useful. As was mentioned before, access to this data allows a company to fix problems as they arise, and not much after they have occurred. This data can be used as a competitive advantage over a firm’s competitors. The more knowledgeable a firm is about the market, the more opportunities they have to develop competitive advantages against their competitors.

2.2. Data Mining

While the importance of having information available at your fingertips is well known, what is done with this information once it is obtained is just as important, or in some cases even more important than just having data available. “Data Mining is a field of study located between Computer Science and Statistics that seeks to discover patterns and relationships within large databases” [4]. The computer science aspect of data mining emerges in the programming of software to process the data obtained rapidly and efficiently, and in the algorithms involved in doing this job. While the statistics side comes in in calculating the relationships within the information that is processed, and sits at the core of the algorithms developed for the data mining programs.

As was mentioned in the definition above from Hastie, *et al.* the purpose of data mining is to discover patterns in large databases (big data). Primarily, it seeks associations, predictions, clusters, and sequential relationships. Finding associations entails finding those data entries that commonly occur together, for example a part of a plane that malfunctions every time a sequence of buttons is pressed on the console. Predictions inform users about future events that are taken from the data, such as how many watermelons may be sold on a cold Sunday afternoon in December. Clustering helps assign different data sets to a respective group, for example sorting customers by consumer demographics, seeing as Hispanic people might differ in choices from North American and European people, this could be useful information to have. Finally sequential relationships, that similarly to predictions, seek to determine future events that might happen. How they differ from predictions, is they are based on a sequence of causality, for example trying to determine when the best moment to offer the sale of a boat to a person who just bought a lake house is.

An example of association mining is what is called a Market Basket Analysis, which is generally used in a store environment to determine what items are commonly bought together. A Market Basket Analysis is what is called an unsupervised data mining model. It receives this name from

the fact that there is no preconceived notion of what the output might be. Its purpose is not to actively predict a value, but rather it focuses on finding associations (patterns) between existing data points. This process for mining associations can be extremely useful, but seeing as programs seek to find every possible relationship between items; it requires complicated processes and time intensive. The fact that when item A is bought, item B is bought as well, is a different data entry than saying that when item B is bought, item A is bought as well, making the amount of combinations increase exponentially as more items and transactions are added. But even with these drawbacks, a Market Basket Analysis can be a powerful tool if used correctly. It can be used to determine what items should be sold together, what is store layout, and when it should be done.

Safeway store is illustrated as a simple example of a market basket analysis. Since it is located near by a college campus, there is the knowledge that customers are going to buy alcohol. If the relationship was not known by the store, it could be determined with a market basket analysis that when college students buy beer, they will very likely also purchase ping pong balls. Safeway can also take advantage of students receiving their student loans early in the semester, and raise their prices on some items to make a larger profit. Then it would have to be determined when the prices need to be lowered again. In the next section, a short example of a Market Basket Analysis is presented, which shows what can be done with this analysis if implemented properly.

3. Review on BI Frameworks

Frameworks of developing business intelligence applications on data warehousing and data mining have been

widely introduced and implemented [2, 5, 6, 11]. Rahman et al. [8] proposed an open multi-tier architecture for high-performance using SOA. The model is a scalable, extensible, interoperable, distributed and easy-to-use data mining applications and has proved to be challenging. In their paper [1], Berbel and Gonzal proposed a schema and processes describing an effective solution for recommending textual OLAP over data warehousing environments. Whereas, their recommendation process is based on text semantics and query personalization to improve the relevant of the retrieved results. Victor and Lopez [12] also illustrated a big data processing framework for developing a privacy models for big data as big data cannot be processed using traditional systems.

Biere [2] argued that business intelligence solutions have absolutely no direction on the enterprise's success if there are biases built in from a technology perspective. The BI applications should always be driven by corporate objectives or missions and business impact. Therefore, two platform options are available: 1) mainframe environment and 2) the distributed or client/server model. A number of server consolidation from distributed back to the mainframe for large enterprise applications, which offers greater security, greater availability, and lower costs for power and support. Research frameworks for developing BI systems is presented in the next section.

4. Research Models for Developing BI Systems

4.1. An Architecture of Data Mart and Real-Time Data Warehouse

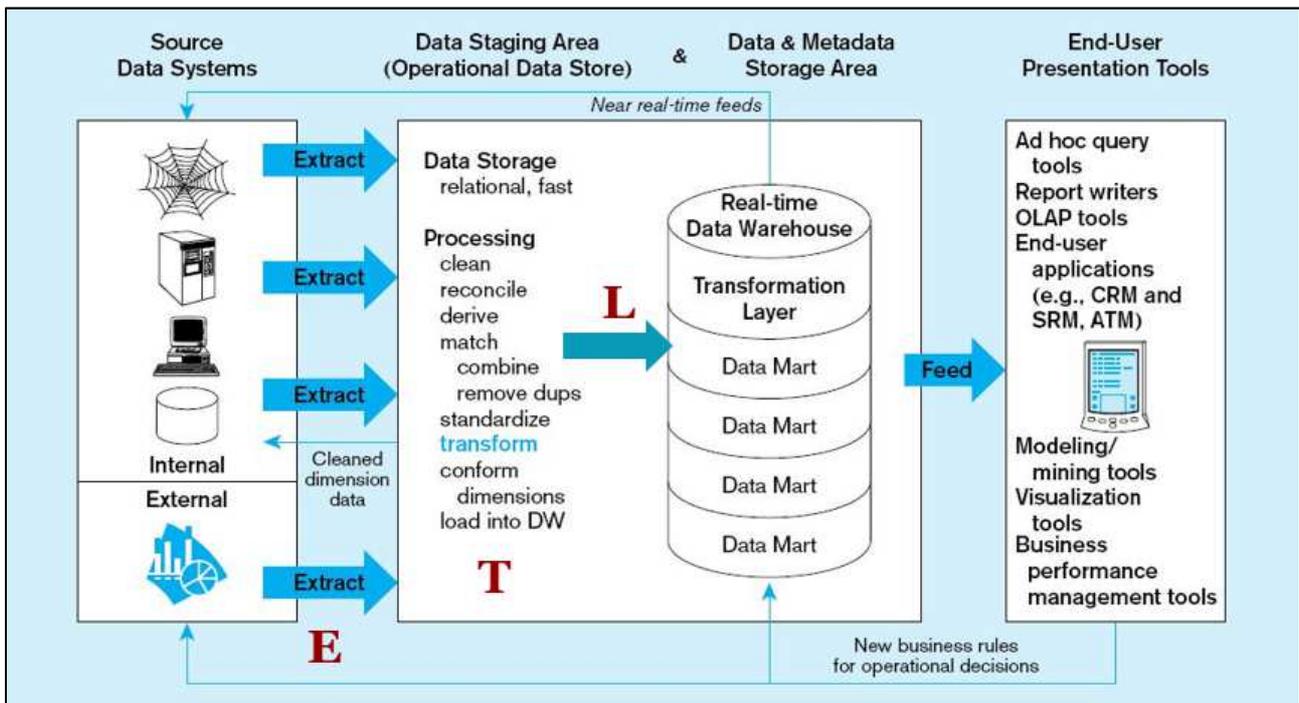


Figure 1. Logical data mart and real-time data warehouse architecture [5].

A logical data mart and real-time data warehouse architecture is practical and essential for developing moderate-sized data warehouses with three processes of E (extract), T (transformation) and L (load). A framework shown in Figure 1 [5] is a popular one with the following unique characteristics: 1) logical data marts are not physically separate databases but rather different relational views of one physical, slightly denormalized relational data warehouse, 2) data are moved into the data warehouse rather than to a separate staging area to utilize the high-performance computing power of the warehouse technology to perform the cleansing and transformation steps, 3) new data marts can be created quickly because no physical database or database technology needs to be created or acquired and no loading routines need to be written, and 4) data marts are always up-to-date because data in a view are created when the view is referenced; views can be materialized if a user has a series of queries and analysis that need to work off the same instantiation of the data mart. Moreover, the architecture should be scalable as data marts may not be small and the enterprise data warehouse continuously grow.

4.2. Elements of a BI System

Tools and frameworks for the generic BI system is vital for businesses competing in the rapidly changing industry. A BI system is designed to retrieve, analyze, transform and report data for business intelligence. The applications generally read data that have been previously stored, often, though not necessarily, in a data warehouse or data mart.

As shown in Figure 2, BI servers provide two major functions: management and delivery [6]. The management function maintains BI application and metadata about the authorized delivery of BI results to users. BI application includes three layers: 1) data source (e. g., operational data, data warehouse, data mart and human interviews), 2) techniques of analysis such as RFM (recency, frequency, monetary), OLAP (OnLine Analytical Process), market basket and, data mining, 3) the results produced from one of the application techniques. Today, practically the BI results can be delivered to “any” devices using computers, smartphones and tablets.

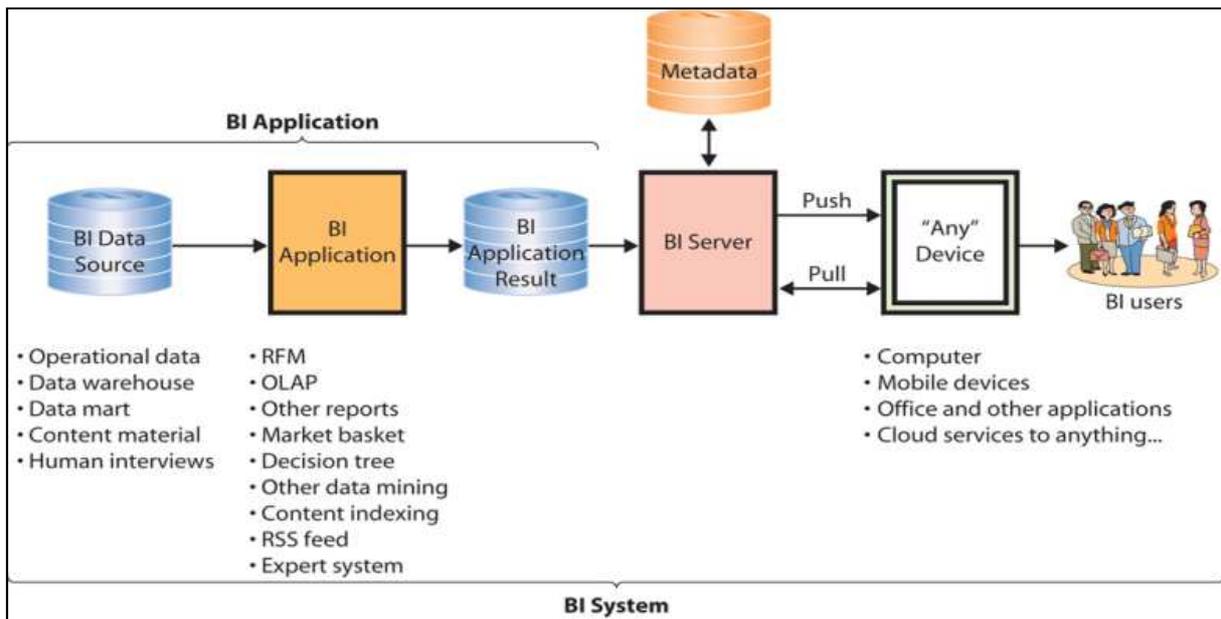


Figure 2. Elements of a BI system [6].

5. Market Basket Analysis – a Business Example

5.1. Case Description

Mountain High is a small startup recently opened by Mike and John, two MBA graduates, at the base of Mt. Spokane, who are attempting to put their newly learned entrepreneurial skills to practice. Mountain High specializes in the sale of ski equipment, and seeing that winter is the season for skiing, Mike and John are excited to get their first sale of high quality equipment. While Mike had specialized in Marketing while he pursued his MBA, John had taken a liking to the MIS side of the field. Mike

told John on a daily basis that the most important part of a business was knowing what to sell to whom, and that a computer could not do a better job than a person at recognizing these trends. While John understood the importance of having a feel for the customer’s wants, he knew that he could apply his knowledge from his Business Intelligence class to the data that they were obtaining, and that way provide Mike with even more relevant data that he might have missed.

5.2. Objective

The objective of the project that John was setting up, was to determine the items that are most often bought together, which could then be used for upselling the customers into

buying other products, and if possible, develop a set of association rules between product combinations that could say, with a certain degree of confidence, what items are expected to be purchased when another item is bought. He hoped that by showing Mike the positive impacts that Business Intelligence could have on their business, he would be more open to try different alternatives that John suggested.

5.3. Organization of the Data

John organized the data into an Excel spreadsheet, separating it into three columns. The first column contained the ID of the transaction, repeated for every item bought in said transaction. The second column containing the ID of the item bought, represented by its name rather than a number; this is for easier readability and identification of item pairs. The third column contained how many of each item was bought. The data is shown in Figure 3.

	A	B	C
1	customerId	itemId	itemCount
2		1 Mask	2
3		1 Skis	2
4		1 Poles	2
5		2 Mask	1
6		2 Skins	1
7		3 Poles	3
8		3 Skis	2
9		3 Skins	2
10		4 Shovel	1
11		5 Shovel	2
12		5 Skins	2
13		5 Mask	3
14		5 Skis	2

Figure 3. Data file.

5.4. Results

The first ten results of the data as outputted by RapidMiner (a data mining tool) were as follows:

Premise	Conclusion	Support	Confidence
sum(itemCount)_Skis	sum(itemCount)_Skins	0.4	0.667
sum(itemCount)_Skins	sum(itemCount)_Skis	0.4	0.667
sum(itemCount)_Skis	sum(itemCount)_Mask	0.4	0.667
sum(itemCount)_Mask	sum(itemCount)_Skis	0.4	0.667
sum(itemCount)_Skis	sum(itemCount)_Poles	0.4	0.667
sum(itemCount)_Skins	sum(itemCount)_Mask	0.4	0.667
sum(itemCount)_Mask	sum(itemCount)_Skins	0.4	0.667
sum(itemCount)_Poles	sum(itemCount)_Skis	0.4	1.000
sum(itemCount)_Skis	sum(itemCount)_Shovel	0.2	0.333
sum(itemCount)_Skins	sum(itemCount)_Shovel	0.2	0.333

Figure 4. Results produced from Rapidminer.

Here we can see the relationship between certain item combinations that are being sold by Mountain High. Interpretation of the result is illustrated as follows. The first Association Rule created by RapidMiner explains that the purchase of Skis implies the purchase of Skins. These two items are bought together in 40% of the transactions (two out of five in this case), and 66.7% probability of the times that Skis is bought Skins is bought as well. Furthermore, if we look at row number 10 (last row), we see that the purchase of Skins implies the purchase of Shovel. This combination happens in 20% of the transactions, and we see that every time Skins is bought, Shovel is bought with 33.3% of confidence.

These calculations can prove very helpful to Mountain High, since Mike can use this data to know what items to sell together or to suggest to customers who are shopping in their store. It can also help them with knowing which items can be bundled or sold at discounted prices if bought together. Mike can be in charge of speaking to the customers and suggesting what to buy for what occasions, while John can focus on creating bundles or incentives to convince people to buy items together. A safety factor can be used here, selling skis together with shovels and avalanche probes, which can help

in the possible occurrence of an avalanche.

6. Conclusion

Business Intelligence is an important part of everyday operations of many companies around the world. There are consulting companies who focus solely on the collection of information and the application of it to help other businesses as their business model. The more relevant information a consulting has, the more of a competitive advantage it is bound to have if it can use this information in a productive way to learn and grow. According to StatSoft.com [9] that argonauten 360°, the German consulting company, helped a call-by-call telecommunications provider rise to the top of its industry and cut costs, after argonauten managed to implement a well-rounded system for their employees and analysts to use. A market basket analysis can be run to discover relevant relationships existing between data points, and this information can then be used for many purposes in a business, including but not limited to finding better pricing information and bundling of items. Text mining [7] can be used to compare texts to see their similarities, which can be used for plagiarism issues, or making your program learn a

model in which to sort out documents to their respective category. The possibilities for data mining keep growing and there is no end in sight. There is nothing stopping someone from taking many disclosures of information that companies release, building a model, and comparing their disclosures with their performance. This way when a company announces new information, it can be applied to the model built to attempt to determine the company's future performance. The world of Business Intelligence is a very immersive and vast one, and there is always something new to learn for he or she who hungers for that knowledge. Anything is possible; you just have to have the right tools and information.

References

- [1] Berbel R. L. T. and Gonzalez S. SM (2015). "How to help end users to get better decisions" Personalising OLAP aggregation queries through semantic recommendation of text documents, *International Journal of Business Intelligence and Data mining*, Vol. 10, No. 1.
- [2] Biere, Mike (2010). *The New Era of Enterprise Business Intelligence*, IBM Press.
- [3] Chaudhuri, Surajit; Dayal, Umeshwar; Narasayya, Vivek (2011). "An Overview of Business Intelligence Technology", *Communications of the ACM* 54.8 pp. 88-98.
- [4] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer-Verlag.
- [5] Hoffer, J. A., Topi, H. and Ramesh, V. (2014). *Essentials of Data Management*, Pearson.
- [6] Kroenke, D. M. and Boyle, R. J. (2017). *Using MIS*, Pearson.
- [7] North, Matthew (2012). *Data Mining for the Masses*, A Global Text Project Book.
- [8] Rahman M. M., Maksud, U. A. and Rahman, S. M. M. (2015). "An open multi-tier architecture for high-performance data mining using SOA," *International Journal of Data Mining, Modelling and Management*, Vol. 7, No. 1.
- [9] StatSoft.com (2004), available at https://www.statsoft.com/Portals/0/Customers/Success_Stories/argonauten360.pdf.
- [10] Stock, Tom. (2011). "Using a Data Warehouse to Solve Risk, Performance, Reporting and Compliance-Related Issues," *Journal of Securities Operations & Custody* 3.4, pp. 305-315.
- [11] Turban, E., Sharda R. and Delen D. (2014). "Business Intelligence: A Managerial Approach", Prentice-Hall, 3rd Edition.
- [12] Victor, N. and Lopez, D. (2016). "Privacy models for big data: a survey", *International Journal of Big Data Intelligence*, Vol. 3, No. 1.
- [13] Wikipedia.a (2017), available at: https://en.wikipedia.org/wiki/Ken_Olsen.
- [14] Wikipedia.b (2017), available at: https://en.wikipedia.org/wiki/Hartsfield%E2%80%93Atlanta_International_Airport.