Applying Market Basket Analysis to the Service Industry.

Make sense of it all.™ 1010data

Beyond the Grocery and Retail Store: Applying market basket analysis to the service industry.

A 1010Data White Paper, by Bart A. Lewin

Copyright © 2009, Bart A. Lewin, All rights reserved. Used here by permission.

What is Market Basket Analysis?

Market basket analysis (MBA) is a well-known data mining technique used primarily by retailers to understand which products consumers purchase together during a store, on-line or catalog transaction. They use this analysis to, among other things, discover cross and up-selling opportunities, develop promotions, determine where best to place products, and determine optimal product inventory levels. Consumers most commonly encounter a promotion derived from MBA when, during an on-line purchase, another product is suggested to them prefaced by a statement like, "Customers who bought this item also bought…." As another example, think of the Thanksgiving promotions offered by grocery stores. A grocer may offer turkey at deeply discounted prices as a loss leader, because they know that if they can get customers to come into the store for this item, these consumers may also purchase other profitable complementary items such as cranberry sauce, pumpkin pie, beverages, and/or yams.

A Simple Illustration

The Association Rule (AR) is the most common data mining method used to perform market basket analysis. The AR algorithm required is quite simple, leading to rapid performance on large data sets. [1]

Taking the Thanksgiving promotion example above, let's provide a simple illustration of how the Association Rule technique works.[2] Table 1 depicts eight imaginary grocery store transactions taken during the prior year's Thanksgiving promotion. The following assumptions make this analysis easier.

- Each transaction consists of one "standard package" of the same variety or brand of each item purchased. Another way to look at this is that all those who bought yams, bought exactly the same product (e.g., SKU).
- Each transaction was performed during a single visit.

Customer ID Number	Basket of Items Purchased
1	yams, cranberry sauce
2	yams, turkey
3	yams, cranberry sauce, turkey
4	yams, turkey

1 http://dms.irb.hr/tutorial/tut_assoc_rules.php as accessed on May 30, 2009.

2 This illustration adapted from Lewin, Bart A., A. K. Singh, Andrew Cardno; "Let's Talk Turkey: Applying retail market basket analysis to Gaming"; *Casino Enterprise Management Magazine*; December, 2008; pp 10-14.

5	cranberry sauce, turkey
6	yams, cranberry sauce, turkey
7	yams, turkey
8	cranberry sauce, turkey

Table 1 Example grocery store transactions.

Without applying a formal Association Rule algorithm, it is easy to see in this very simple example how useful MBA can be for determining product placement. Observe that most transactions (7 out of 8) included turkey (the focus of the promotion), 5 of the 7 turkey buyers also bought yams, and 4 bought cranberry sauce. Knowing this, the grocer may decide this year to place yams near the turkey, because 63% of all customers and 71% of turkey buying customers purchased these items together. Cranberry sauce should also be placed near the turkey because 57% of the turkey buyers also bought this item. Placing all three of these items near each other this year may encourage more customers to buy all three items. Since only 1 customer bought yams and cranberry sauce without buying turkey (13% of all customers who perhaps bought their turkey elsewhere) suggests that turkey does indeed drive the purchase of all three items.

Association Rules Algorithm

Another way to express the relationships between the Thanksgiving products is by using an "If/Then" construction. One of the "If / Then" Association Rules resulting from this analysis could be stated as "*If* a customer purchases turkey, *then* the customer is likely to also purchase yams". Association Rule engines typically provide their output in this way, along with the associated Support ("Supp") and Confidence ("Conf") factors. Table 2 shows the output for the Orange Association Rules engine.[3]

The row highlighted in yellow on Table 2 could be read, "If a customer bought turkey, they were also likely to buy yams." The usefulness of this rule is expressed as the AR Support and Confidence factors mentioned above. Support is the percentage of those who bought both turkey and yams divided by the total number of transactions (# of transactions including turkey and yams / total number of transactions), which in this case is 62.5%. Confidence is the percentage of those that bought turkey and also bought yams (# of transactions including turkey and yams / total number of transactions that included turkey), which in this case is 71.4%. These factors give us an indication of the accuracy and usefulness of the rules. Typically, analysts look for rules generating at least 50% support and 50% confidence levels.

The full association analysis of this example would include 12 rules (e.g., If turkey, then yams, if turkey, then cranberry sauce, if turkey and yams then cranberry sauce, etc.) with varying confidence and support factors. The equation for calculating the number of Association Rules is as follows:

$R = 3^d - 2^{d+1} + 1$

Where

R = total number of association rules, and

d = the number of distinct products.

Looking at this equation, the number of AR's increases exponentially with the number of available products, indicating that real-world analysis can get quite complex.

3. Orange is modeling software developed by the Artificial Intelligence Laboratory at Univerza v Ljubljani http://www.ailab.si/orange/

Rules	Supp	Conf
▼Cranberry Sauce=No		
Cranberry Sauce=No -> Turkey=Yes	0.375	1.000
Cranberry Sauce=No -> Yams=Yes	0.375	1.000
Cranberry Sauce=No -> Yams=Yes Turkey=Yes	0.375	1.000
Cranberry Sauce=No Turkey=Yes -> Yams=Yes	0.375	1.000
Cranberry Sauce=No Yams=Yes -> Turkey=Yes	0.375	1.000
▼Cranberry Sauce=Yes		
Cranberry Sauce=Yes -> Turkey=Yes	0.500	0.800
Cranberry Sauce=Yes -> Yams=Yes	0.375	0.600
▼Turkey=Yes		
Turkey=Yes -> Cranberry Sauce=No	0.375	0.429
Turkey=Yes -> Cranberry Sauce=Yes	0.500	0.571
Turkey=Yes -> Yams=Yes	0.625	0.714
Turkey=Yes -> Yams=Yes Cranberry Sauce=No	0.375	0.429
Turkey=Yes Cranberry Sauce=No -> Yams=Yes	0.375	1.000
Turkey=Yes Yams=Yes -> Cranberry Sauce=No	0.375	0.600
▼Yams=Yes		
Yams=Yes -> Cranberry Sauce=No	0.375	0.500
Yams=Yes -> Cranberry Sauce=No Turkey=Yes	0.375	0.500
Yams=Yes -> Cranberry Sauce=Yes	0.375	0.500
Yams=Yes -> Turkey=Yes	0.625	0.833
Yams=Yes Cranberry Sauce=No -> Turkey=Yes	0.375	1.000
Yams=Yes Turkey=Yes -> Cranberry Sauce=No	0.375	0.600

Table 2 Sample output from the Orange Association Rules engine.

Association Rules have some distinct advantages over many statistical models. The output consists of plain English rules that a typical business user can understand, validate and adjust in accordance with their experience. It also works with transactions of varying lengths (e.g., customer A bought turkey and yams, and customer B bought turkey, yams and cranberry sauce during their visit). Also, the goal is pure discovery. No predetermined target is required.[4] These are distinct advantages over black-box modeling such as regression analysis or neural networks, where the outcome is typically a probability that an event will occur, and it takes expertise in statistics to be able to understand how the result was derived and the model's quality.

4 In contrast to this, regression analysis, requires identifying a dependent variable (the target), and the outcome is of this analysis is the probability that a particular event will occur given a set of conditions as described by independent variables.

Taking Market Basket Analysis (MBA) Beyond Grocery and Retail

Given the characteristics of Association Rules, it seems that when a customer visits a service industry outlet, like a hotel, casino or even a dry cleaner, they are also buying a market basket of goods and services. For instance, during a hotel visit, a customer may request a top floor suite, order a particular group of items from room service, purchase in-room movies, use the business center, purchase a variety of items in the fine dining room, among many other things. Why, then, not apply MBA to this situation? In the case of the Lodging Industry, the hotel management system[5], the point of sale system, the in-room movie system, etc., commonly digitally capture purchase transactional data. Looking again at Table 1, it is possible to substitute any item from a hotel's products and services offerings at any level in the product hierarchy (see Figure 1 for an example of a small portion of a hotel products and services hierarchy) for the grocery items, and perform MBA.

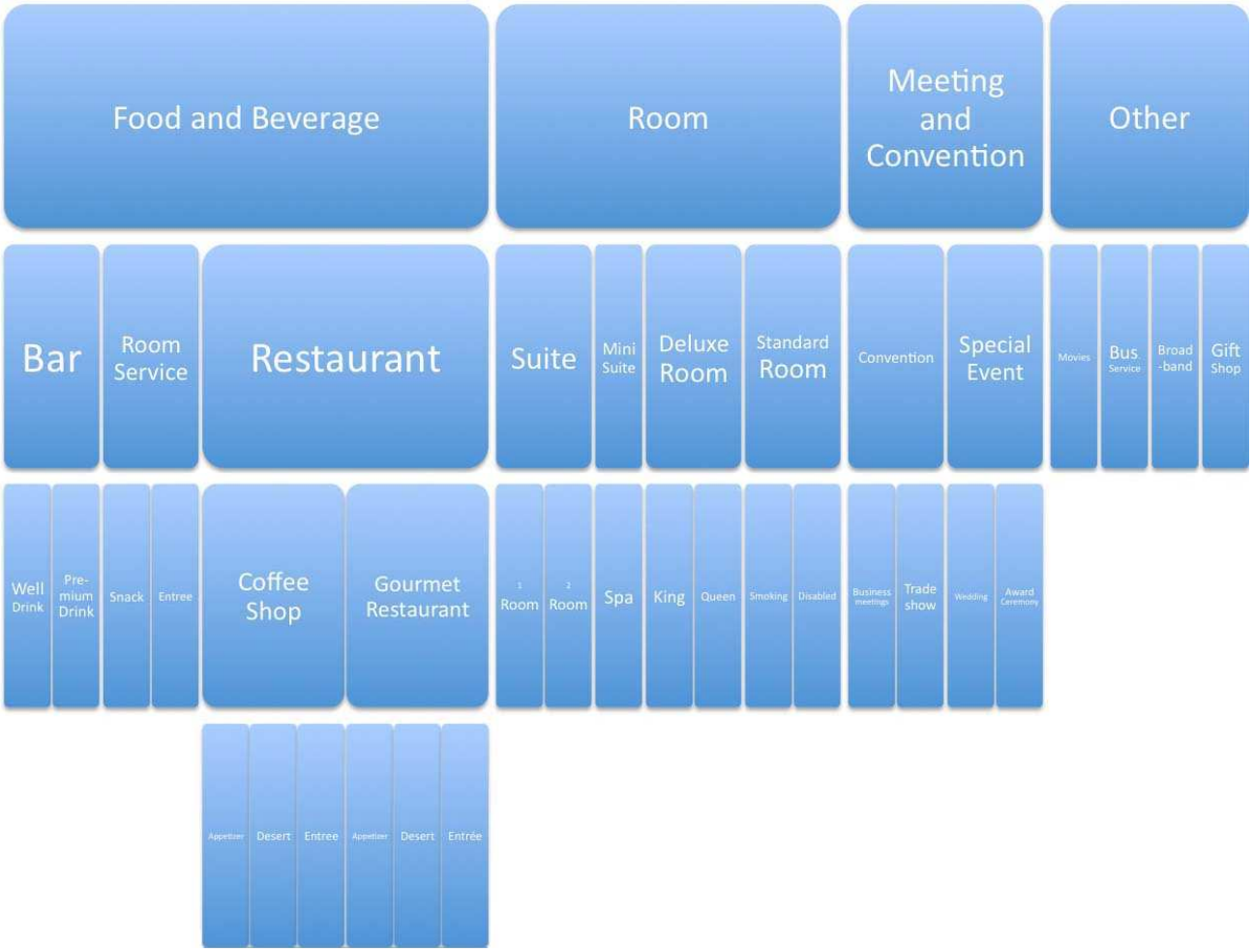

Figure 1 Example of a portion of a hotel product and service hierarchy

Add other factors such as the particular characteristics of a product (e.g., a desert in the gourmet restaurant serviced on a red plate and the same item served on a blue plate), or the location of where the

5 Hotel management systems are used to manage guest room reservations, room assignment, product the bill, etc.

product is offered (e.g., the gift shop located near the front desk as opposed to the one near the elevators), and one can see the possibilities and the rewards are endless. MBA for a hotel might be used to discover what a typical technology trade show attendee typically purchases, in order to construct an attractive package.

The Lodging Industry in particular is interesting because of their variety of products, the fact that they digitally collect information for them (e.g., as room charges), and they directly associate purchase transactions with customer demographic information as a part of doing their business (we typically provide our name, e-mail and a physical address to the hotelier). But, one can now readily see how the analysis would apply to many service industry situations.

We Have the Data

It is almost cliché to say that the world is undergoing a data explosion. According to Cisco®'s 2008 Visual Networking Index Forecast for 2007-2012, IP traffic will double every two years during this period, resulting in an annual bandwidth demand for the world's IP networks to reach 522 exabytes[2] (more than half a zettabyte--10^{21}).[6]

Driving some of this network traffic are new ways to collect information. Through the use of affinity programs, retail and grocery outlets, airlines, hotels and resorts are all able to digitally associate purchase transactions to individual consumers. To us, this means that turning data into useful information will become more difficult, but if accomplished, can allow for marketing, operational, and financial analysis to reach heights never before imagined.

We Have the Tools

Fortunately, new and improved Database Management Systems (DBMS), network components, digital processors, data visualization tools, and data mining techniques are being introduced with full capabilities to meet the data explosion challenge.[7]

Extract Load Transform (ELT) versus Extract Transform Load (ETL)

The availability of mounds of data, however, is not enough, because there are data quality issues to consider. The data required to perform MBA typically comes from numerous sources (e.g., cashiers, outlets, kiosks, call centers, websites), in numerous formats. This data must be cleaned and standardized in order for useful analysis to be performed.

The trend for doing this today is to load the atomic data (raw data at the lowest level of detail) from all sources, and use database views and / or procedures to perform the necessary transformations. This technique has distinct advantages over the traditional practice of performing any data hygiene and standardization before it is loaded into a carefully designed data structure inside the database. For instance, any required changes to the transformation algorithms may be accomplished by changing the

---

6 "Cisco Visual Networking Index Projects Global IP Traffic to Reach Over Half a Zettabyte in Next Four Years" http://newsroom.cisco.com/dlls/2008/prod_061608b.html; as accessed April 5, 2009.

7 Feinberg, Donald, Mark A. Beyer; "Magic Quadrant for Data Warehouse Database Management Systems; 23 December 2008; Gartner RAS Core Research Note G00163473.

Tan, Pang-Ning, Steinbach, Michael, and Kumar, Vipin. (2006). Introduction to Data Mining. Addison-Wesley.

metadata inside the database, without having to perform a time consuming and risky data conversion process. Even if the raw data structure changes, it is possible to create database views and procedures to accommodate it, without having to change the structure of the legacy data (both structures may easily coexist).

This is only possible, because today's lightning fast Database Management Systems, are fully up to the task of performing these manipulations in real-time *as the data is being retrieved for analysis.* For MBA, this could mean performing these manipulations on billions of rows of data.

Associating Rows Within a Database Table

Typically, when purchase transaction data is captured by the retail or service sectors, the detail sales information has a separate row for each line item purchased. Each line item is linked to a transaction (a market basket), and is accompanied by the revenue and location of the purchase along with other pertinent information. Table 3 is a screen print from one such table.

Sales Item Detail

Columns 1-14 of 16, Rows 1-30 of 737,538,011

Transaction ID	Store	Date	Time	Item SKU	Units	Sales	Cost	Extended PLU Price	Subcat	Cat	Dept	Selling Unit	Account
286021531	1005	1/01/08	9:54:51	78394366	-1	-5.00	-1.84	-5.00	3423	14	1	78394366	502037957
286021532	1005	1/01/08	9:56:10	499343B7	1	1.10	0.56	1.10	1153	22	2	499343B7	506260478
286021532	1005	1/01/08	9:56:10	4943598A	1	0.50	0.25	0.50	1103	22	2	4943598A	506260478
286021534	1005	1/01/08	15:12:25	69833A96	1	3.00	1.45	3.00	2202	32	3	69833A96	504999738
286021534	1005	1/01/08	15:12:25	69833A96	1	3.00	1.45	3.00	2202	32	3	69833A96	504999738
286021534	1005	1/01/08	15:12:25	6833565B	1	2.25	1.35	2.25	2209	32	3	6833565B	504999738
286021535	1005	1/01/08	15:14:11	GV435969	1	1.10	1.00	1.10	1153	22	2	GV435969	500006709
286021535	1005	1/01/08	15:14:11	GV43596A	1	1.10	1.00	1.10	1153	22	2	GV43596A	500006709
286021535	1005	1/01/08	15:14:11	46336CB7	1	1.65	1.10	1.65	1305	29	2	46336CB7	500006709
286021536	1005	1/01/08	15:16:00	678743A4	1	0.34	0.13	0.34	1405	20	2	678743A4	500991748
286021536	1005	1/01/08	15:16:00	678743A4	1	0.34	0.13	0.34	1405	20	2	678743A4	500991748

Table 3 Sample purchase transaction table screen shot taken from 1010Data's Database Management System (http://www.1010data.com).

In order to perform market basket analysis, however, the individual rows containing the line item information must be grouped together by transaction, or market basket as depicted in Table 1 above. For many relational Database Management Systems, creating a row containing a variable number of columns is quite cumbersome, but this is exactly what is required for most Association Rules. Facilitating MBA calls for a Database Management System with preferably built in functions that allow rows within a table to be associated with one another based on key columns. This is very different than the typical on-to-many relationship view where the detailed records are simply listed below the master records.

However, once again, modern high performance DBMSs have embedded row-grouping functions in their engines, and many even incorporate statistical modeling functions as well. It is well understood that these functions perform faster when embedded in the tool, and have the added convenience of not requiring the analyst to export the data from the database and re-import it into a separate analytical tool.

Business Value of Using MBA in the Services Industries—Conclusion

Market Basket Analysis is an effective marketing and operational tool for not just industries where only real products are sold, but also in service industries where by treating services as products for analytical purposes similar results can be achieved. In fact, Market Basket Analysis may be even more effective when applied to service industries because of their ability to collect a plethora of data and directly associate purchase information to known individuals.

The following three Use Cases illustrate the business value of applying MBA to service industry marketing.

- Lodging Industry - Once a customer market segment has been established, which for a hotel this might be business travelers that have visited the hotel once in connection with a convention, MBA might be used to discover the types of services they enjoyed during the group visit. This information would be used to develop an offer to entice them to stay any time they visit the location on business travel or for a weekend getaway with their family. It may also be used to entice the meeting planner to have another meeting at the hotel by offering a package that includes the products and services enjoyed by the participants at the last meeting.

- Dry Cleaning Service - A dry-cleaning chain, might study the types of garments (e.g., suits, shirt laundering), household items (e.g., comforters, drapery), or other services (e.g., alterations, delivery) particular households have serviced over a period of time. Rather than simply providing overall discounts, more attractive offers might be developed by packaging common services purchased over a year by a particular market segment that both encourages the customer to utilize the store and purchase other services (e.g., have household items cleaned) the customer may not have thought of.

- Operational Analysis - MBA might also be used for operational analysis as well. For instance, if particular combinations of services are purchased during holidays, the inventory levels of materials for providing those services and for staffing may be adjusted to reduce waste and eliminate shortages.

In order to affectively apply MBA, one must employ a high performance Database Management System for the purpose of cleaning and formulating raw data for ready usage by Association Rules engines. Fortunately, there excellent and mature tools are commercially available today, and the analytical is well understood.

Companies that use and deploy these new tools around MBA and who have the ability to use analytics for their benefit, will most likely be the winners in this new explosive data rich environment we find ourselves competing in today.