ASSOCIATION AND SEQUENCING

KEYS TO SUCCESSFUL MARKET BASKET ANALYSIS AND WEB MINING

Until recently, association and sequencing were often overlooked. Many data mining products often omitted one or both of these techniques. The recent CRM emphasis on maximizing customer lifetime value through cross-sell and up-sell has refocused attention on associations and sequencing, as they are essential to analyzing and understanding buying patterns for both the traditional retail and e-tail business.

Once a business acquires a customer, selling as many different products as possible to that customer can help to maximize the customer's value to the business. One way to accomplish this goal is to understand what products or services customers tend to purchase at the same time, or later on as follow up purchases. Addressing this business need is a very common application of data mining, and association and sequencing techniques can perform this kind of analysis. Although originally devised for marketing purposes, these techniques also have important applications in medicine, finance and other data rich environments where separate events might be related to each other, and where knowing about such relationships can be valuable knowledge.

Association and sequencing tools analyze data to discover rules that identify patterns of behavior. An association tool will find rules like

> *When people buy diapers they also buy beer 50% of the time.*

It is highly unlikely that this rule is true. In fact, the oft-cited correlation between sales of beer and diapers is probably a myth. However, it is convenient to use it for illustrative purposes.

A sequencing tool is very similar to an association tool, but it adds time to the analysis and produces rules like

> *People who have purchased a VCR are 3 times more likely to purchase a Camcorder in the time period 2-4 months after the VCR was purchased.*

When an association or sequencing algorithm is used to find the kinds of rules we've just seen it is frequently called *market basket analysis.* In fact, such analysis has become the primary use of the association technique, resulting in its frequently being called a market basket technique.

Business managers or analysts can use a market basket analysis to plan:

- *Couponing and discounting* – it is probably not a good idea to offer simultaneous discounts on beer and diapers if they tend to be bought together. Instead, discount one to pull in sales of the other.
- *Product placement* – place products that have a strong purchasing relationship close together to take advantage of the natural correlation between the products. Alternatively, place such products far apart to increase traffic past other items.
- *Timing and cross-marketing.* For example, assume that a sequencing analysis has produced the VCR/Camcorder rule described above. Clearly this suggests that mailing a camcorder promotion to VCR purchasers is best done so that it will arrive in their mailbox approximately 2-3 months after the VCR purchase.

Although most commonly used for market basket analysis, association and sequencing tools have useful applications in many other industries besides retail. Association and sequencing tools find patterns in transaction data, and most organizations capture transactional data. Understanding the patterns of behavior or activity can provide valuable insight.

In **health care**, there are possible applications in care management, procedure interactions and pharmaceutical interactions. Consider, for example, the following statements that might result from an analysis. Their possible application should be immediately obvious.

> *Patients who are taking drugs A, B, and C are 2.5 times more likely to also be taking drug D.*

> *Patients receiving procedure X from Doctor Y are 3 times less likely to get infection Z.*

There are many applications of association and sequencing in the **financial** industry. As with the medical applications, an example is worth a thousand words (and to someone it may even be worth a million bucks!)

> *The prices of stocks in industry Q are 1.8 times more likely to close up one day after stocks in industry R closed down.*

Our final examples deal with **fraud detection** in **telecommunications** and **insurance**:

> *International credit card calls longer than 3 minutes originating in area code 555 between 1:00 AM and 3:00 AM are three times more likely to go uncollected.*

> *Accident claims involving soft tissue trauma where attorney P represents the claimant are twice as likely to be fraudulent.*

For clarity we have stated our example rules in English. Most association products produce rules in a more compact tabular form as is shown in Figure 1, although some, like Exclusive Ore's XAffinity, offer the ability to translate the rules to English (see Figure 6, below).

LeftSide		RightSide	Confidence	Lift	Support
pet food		pet supplies	1.0000	1.92	0.001665
canned veg, cereal		dairy	0.7659	1.0058	0.001370
cereal, produce		dairy	0.6683	1.35	0.001171
cereal		dairy	0.6680	1.97	0.009983
beer, cereal		dairy	0.3427	2.12	0.000971

Figure 1.Common tabular format fore association rules, sorted by confidence

Some important elements of association tools can be seen in the sample rules and illustrations we've seen so far. Did you notice that our first rule about diapers and beer stated an explicit percentage (*50% of the time*) while the second rule used the comparative phrase *3 times more likely*? In the second case the probability is compared to the probability of the event occurring independently. For example, if people normally buy beer 5% of the time, then the first rule could have said "10 times more likely." The ratio in this kind of comparison is called *lift*, and a key goal of an association or sequencing data mining exercise is to find rules that have the desired lift.

Figure 1 shows five typical association rules. For each rule, a confidence, a lift, and the support computed, is reported. The precise meanings of confidence and support are explained in the next section. We will also return later to some implications that might be drawn from the particular rules in Figure 1.

TERMINOLOGY: UNDERSTANDING THE RULES

Each rule has a *left-hand side* (when people buy diapers) and a *right-hand side* (they also buy beer). Sometimes the left-hand side is called the *antecedent* and the right-hand side is called the *consequent*. In general, both the left-hand side and the right-hand side can contain multiple items, but for simplicity we'll stick with single items for now.

A rule has two measures, called *confidence* and *support*. To see what these terms mean and how they are computed, consider the following beer and diaper example:

500,000 transactions
20,000 transactions contain diapers
30,000 transactions contain beer
10,000 transactions contain both diapers and beer

Support measures how often items occur together, as a percentage of the total transactions. In this example, beer and diapers occur together 2% of the time (10,000 / 500,000). Confidence measures how much a particular item is dependent on another. Since 20,000 transactions contain diapers and 10,000 also contain beer, when people buy diapers, they also buy beer ½ of the time (a rule that matches our introductory example). The confidence for this rule is ½ or 50%.

The *inverse rule*, which would be stated as

When people buy beer they also buy diapers 1/3 of the time

has a confidence of 33 1/3% (computed as 10,000 / 30,000).

Note that these two rules have the same support (2% as computed above). Support is not dependent on the direction (or implication) of the rule; it is only dependent on the set of items in the rule.

In the absence of any knowledge about what else was bought, we can also make the following assertions from the available data:

People buy diapers 4% of the time.
People buy beer 6% of the time.

These numbers -- 4% and 6% -- are called the *expected confidence* of buying diapers or beer, respectively, regardless of what else is purchased. *Lift* measures the difference between the confidence of a rule and the expected confidence. You can measure this difference by taking a ratio.

Lift is one measure of the strength of an effect. If we found a rule that said that people who bought diapers also bought beer 8% of the time, then the effect of buying diapers on beer buying behavior is fairly small if the expected confidence is 6%. If, on the other hand, the confidence is 50%, and lift is more than 8 times (when measured as a ratio), then the interactions between diapers and beer is very strong.

The above rules were all based on item sets with two items. Now let's consider item sets with three items, adding the following to our example:

10,000 transactions contain wipes
8,000 transactions contain wipes and diapers
220 transactions contain wipes and beer
200 transactions contain wipes, diapers and beer

The complete set of 12 rules is presented in Figure 2 along with their confidence, support and lift. There are 6 possible rules for item sets of three items (rules 7-12 in Figure 2). In addition, there are four additional 2-item rules (rules 3-6 in Figure 2).

	Left hand side	→	**Right hand side**	**Expected Conf**	**Confi-dence**	**Lift Ratio**	**Support**
1	Diapers	→	Beer	6.00%	50.00%	8.33	2.00%
2	Beer	→	Diapers	4.00%	33.33%	8.33	2.00%
3	Diapers	→	Wipes	2.00%	40.00%	20.00	1.60%
4	Wipes	→	Diapers	4.00%	80.00%	20.00	1.60%
5	Wipes	→	Beer	6.00%	2.20%	0.37	0.04%
6	Beer	→	Wipes	2.00%	0.73%	0.37	0.04%
7	Diapers and Wipes	→	Beer	6.00%	2.50%	0.42	0.04%
8	Diapers and Beer	→	Wipes	2.00%	2.00%	1.00	0.04%
9	Wipes and Beer	→	Diapers	4.00%	90.91%	22.73	0.04%
10	Diapers	→	Wipes and Beer	0.04%	1.00%	22.73	0.04%
11	Wipes	→	Diapers and Beer	2.00%	2.00%	1.00	0.04%
12	Beer	→	Diapers and Wipes	1.60%	0.67%	0.42	0.04%

Figure 2. A complete set of 2-item and 3-item rules for the beer and diapers example.

The greatest amount of lift, is found in the ninth and tenth rules, which both have a lift greater than 22, computed as 90.9/4 and 1/0.04. Looking at the ninth rule, the lift of 22 means that people who purchase wipes and beer are 22 times more likely to also purchase diapers than people who do not purchase wipes and beer. Also, note the *negative lift* (or a lift ratio of less than 1) in the fifth, sixth, seventh and last rules. The latter two rules both have a lift ratio of approximately 0.42. We can interpret the negative lift on the seventh rule to mean that people who buy diapers and wipes are less likely to buy beer than one would ordinarily expect (i.e., in the absence of diapers and wipes). It is not coincidental, but might seem somewhat counterintuitive, that the lift of a rule and its inverse rule are the same. One can easily prove that they will always be the same from the formula that computes lift, an exercise we leave for the interested reader.

While one would be most interested in rules with very high or very low confidence, sometimes such rules model an anomaly. Let's take another look at the rules in Figure 1. The first rule, with a confidence of 1 (100%), says that whenever people bought pet food they also bought pet supplies. A confidence of 100% seems highly unlikely, as it means that no one bought pet food without also buying pet supplies, but further investigation might show that the data was for one day only, and that there was a special giveaway on that day so that anyone who bought pet food was given a free pet toy. The other rules, because they all have dairy on the right hand side, may also warrant another look. Because milk or eggs are so commonly purchased, "dairy" is quite likely to show up in many rules. However, because it is such a common ingredient in a market basket, rules involving dairy might not be practical. For this reason, the ability to exclude specific items can be very useful in an association tool.

In general, analysts are looking for rules that have a very high or very low lift, that have support that exceeds a threshold, and that do not involve items that appear on most transactions. Rules with a low value for support might simply be due to a statistical anomaly. Even if they are accurate rules, acting on rules with little support might not lead to much gain since a low level of support suggests that the event occurs infrequently. Increasing your profit by $100 on something that occurs three times a year is probably not worth doing.

PREPARING DATA

As with all data, the data used by an association algorithm is made up of entities and attributes. (The terms "entity" and "attribute" here do not mean exactly the same thing as when used in relational database modeling.) The entity might be a market basket – and the attributes are all of the items purchased at one time. Or the entity might be a web session – where the attributes are the IP address, date and time, browser, pages visited, and purchase behavior associated with the individual who visited the site.

The data that is used by an association algorithm needs to be in one of two formats that we'll call *horizontal* and *vertical*, respectively. In the horizontal format (see Figure 3) there is one row for each entity, and there are columns for each attribute. For example, analyzing the effectiveness of a medical procedure would require one row for each patient, with separate columns for each drug, doctor and outcome recorded. For market basket analysis with the horizontal format, there is one row for each market basket, with columns for each (type of) product.

A significant problem for the horizontal format is that the number of columns can become quite large. For market basket analysis, where the number of products might exceed 100,000, similar products need to be grouped together to reduce the number of columns to a reasonable quantity. Another problem with the horizontal format is that the schema is data dependent. When a new product is added to the market basket analysis, or when products are categorized in a different way, then the schema needs to be changed to add or reorganize columns. An additional problem arises in trying to use the horizontal format for sequence analysis of web logs, as it is very difficult to represent either the time element or repeated page visits.

ID	diapers	beer	wipes	dairy	cereal
132	Y			Y	
428		Y	Y	Y	

Figure 3. Data in horizontal format.

The vertical format (see Figure 4), which is more commonly used by the data mining products, eliminates these problems by using multiple rows to store an entity, using one row for each attribute. The rows for a particular entity (i.e., a market basket or web session) are tied together with a common ID. This kind of representation is more normalized in the relational sense, and works much better when an entity can have great variability in terms of the number of attributes. For example, some people check out of the supermarket with only two or three items when others fill two carts with hundreds of items. For sequencing, the time attribute is simply another column, e.g., the time at which a particular page was visited.

ID	product
132	diapers
132	dairy
428	beer
428	wipes
428	dairy

Figure 4. Data in vertical format.

Association algorithms can only operate on categorical data. If you want to use non-categorical attributes, like income of the purchaser for example, the non-categorical data needs to be binned into ranges (e.g., 0-20,000; 20,001 – 40,000, 40,001 – 70,000, 70,001 and over), turning each range into an attribute.

Another common characteristic of association rule generators is an *item hierarchy*. A product hierarchy can be used to group similar items together. In our example we were using generic terms: diapers, wipes and beer. In reality, stores sell specific items: Coors beer in a 12oz six-pack, Bud Lite 16oz cans in a case, etc. While sometimes one might be interested in rules that deal in specific items, rules that are more general are frequently desired, like the rules we have been using as examples. Or, we might be interested in rules that differentiate between diapers sold in boxes vs. diapers sold in bulk.

ALGORITHMS AT WORK

At its heart, an association or sequencing algorithm is simply a counting algorithm, with the final probabilities computed by taking ratios between various counts. If item hierarchies are in use, then some translation (or lookup) is needed. In either case you must carefully control the sizes of the item sets under consideration because the combinatorial explosion problem comes into play.

Large grocery stores stock upwards of 100,000 different items. This means that there are approximately 0.5×10^{10} (5 billion!) possible item pairs, and 0.17×10^{15} sets of three items. You should use an item hierarchy to reduce this number to a manageable size. In a grocery store application there is unlikely to be a specific relationship between Pampers in the 30 count box, say, and Pabst Blue Ribbon in 12 ounce cans. If there is such a relationship, it is probably subsumed by the more general relationship between diapers and beer.

Using an item hierarchy reduces the number of combinations, and also helps to find more general, and probably more useful, higher level relationships such as those between any kind of diapers and any kind of beer. Unfortunately, even if you use an item hierarchy to group items together so that the average group size is 50 (reducing 100,000 items to 2,000 item groups), the combinatorial explosion will still raise its ugly head. With 2,000 item groups there are still almost 2 million paired item sets. An algorithm might therefore need to have up to 2 million counting registers. And, there are 1.3 billion 3-item item sets! Six-item item sets, should you want to consider them, will potentially require 9×10^{17} counters.

What we have been talking about is the *potential* number of combinations. In most cases, particularly in a supermarket, many of the combinations will never occur. This means that an algorithm will not need nearly as many counters as was implied by the analysis above. Nevertheless, some sort of dynamic memory or counter allocation and addressing scheme will be needed.

As an aside, how many combinations would there be in a typical transaction? Knowing this number will give us an idea of how many counters need to be updated for each transaction. Lets assume that

the typical market basket contains 30 items, and that some of these are in the same item group (for example, two kinds of cheeses), reducing the number of item groups to 25. Now we use the formula that tells us how many combinations there are of “n things taken m at a time.” Just for old time's sake, here is the formula:

$$\frac{n!}{m!(n-m)!}$$

The table below (Figure 5) gives the number of combinations for 25 things taken 2, 3, etc., at a time.

2	300
3	2,300
4	12,650
5	53,130
6	177,100
7	480,700
8	1,081,575

Figure 5. The number of combinations of 25 things taken 2, 3, etc., at a time.

So, if we are interested in item sets of up to six, there will be 245,480 (the sum of the first 5 entries in the right-hand column, above) combinations to analyze i**n each transaction**. If we’re looking at a million or so transactions, you can see that some care is needed in the algorithm to ensure reasonable performance.

SEQUENCING

Sequencing builds on the basic constructs in association by adding the concept of time. Lets see what is different in a sequencing algorithm.

Sequencing requires a primary ID, like a customer ID, which relates transactions that occurred at different times. By taking pair-wise combinations of all transactions that have the same primary ID, and computing the time difference between each pair, the algorithm identifies all before-and-after item pairs.

Because the time difference is not categorical, it has to be binned, or in some way turned into a categorical value based on the time-related objectives set by the user. Examples of time related objectives are "at any later time", "within six months" or "next visit", or a set of mutually exclusive ranges (e.g., next day, next week , next month , and next year). Some implementations support only "at any later time", others, like XAffinity support a wide range of time related objectives.

Note that taking pair-wise combinations again brings the categorical explosion into play. If a particular primary ID (e.g., a customer) has five transactions, then there will be 10 paired combinations (5 things taken 2 at a time).

As an aside, consider a regular customer at a supermarket. If he or she makes weekly visits to the market, and we want to do a sequencing analysis on a year’s worth of data (say 50 transactions with two weeks off for vacation), we’d better proceed cautiously. For just one such regular customer there will be 1,225 paired combinations! It would improve things a lot if we only considered “next visit” pairings, as that would reduce the number of combinations to 49. (The number of “next visit”

pairings is always one less than the number of transactions.) For a supermarket, anyway, the latter is also an analysis that makes more sense. We're much more likely to find interesting patterns from one week to the next than we are to find patterns relating January's market basket to July's.

In any case, our point is that an appropriate time-related objective can (and should) be used to reduce the number of combinations. This is important, because once the transactions have been combined, the sequencing algorithm is identical to the association algorithm with some additional restrictions. And, as we saw above, the association algorithm is already laced with combinatorial issues; the time related pairing just adds another layer of combinations!

We won't go into further detail here, except to point out to those who have a deeper interest the following restrictions that when applied to the association algorithm turns it into a sequencing algorithm.

- An item set must contain the time difference attribute.
- An item set must contain at least one "before" item.
- An item set must contain at least one "after" item.

INTERPRETING THE OUTPUT

Most association and sequencing products present results as rules, as shown in Figure 1, above, and Figure 6, below. In addition they often isolate specific rules of interest for examination. Some permit rules to be sorted by lift, confidence, support, left-hand side, or right-hand side. Others enable selection of a sub-set of the rules based on specified item, item group or thresholds for confidence or support. In some products, users specify selection criteria before executing the algorithm. In others all rules are generated and parameters are then used to select a subset for presentation. For problems that have a large amount of data, the first method might be more useful if you wish to reduce the amount of time needed to run the algorithm. On the other hand, if the first method is in use and you later decide to change the constraints, you'll have to wait while the algorithm finds new rules for you. If all rules are computed at once, interactive exploration of the rule set will be much easier to do.

Rules in Natural Language

Rules

Whenever someone obtains beer they are likely to also obtain diapers with a probability of 33.33%, which is 8.33 times more likely than the usual probability of 4.00%. This combination occurred in 2.00% of all baskets (100).

Whenever someone obtains beer they are likely to also obtain diapers AND wipes with a probability of 0.67%, which is 0.42 times more likely than the usual probability of 1.60%. This combination occurred in 0.04% of all baskets (2).

Whenever someone obtains beer they are likely to also obtain wipes with a probability of 1.00%, which is 0.50 times more likely than the usual probability of 2.00%. This combination occurred in 0.06% of all baskets (3).

Figure 6. Association rules presented in English by XAffinity.

ALL TOGETHER NOW

Association and sequencing techniques can help companies identify groups of products or services that customers have already demonstrated a tendency to acquire together, or in subsequent purchases. The infamous but probably mythical association between beer and diaper purchases is a good example of a retail application of association and sequencing.

In today's e-commerce environment, these techniques can be used not only in the traditional way but also to analyze user's click-streams to look for positive and negative relationships between web pages and subsequent behavior (e.g., purchasing a product or abruptly leaving a web site). This kind of analysis can help a company differentiate between "good" and "bad" pages on their web site.

Remember - detecting and assessing relevant patterns can benefit almost any business that accumulates large volumes of transactions.

