Affinity Analysis: Identifying Exploitable Product Synergies in Sales Data

Ed Ramsden
Principal, Sensorlytics LLC

Executive Overview:

Understanding the relations between one's products, as determined by customer buying patterns, can yield potentially useful insights into customer behavior and a needs. Such information can be useful and for discovering cross-selling opportunities and for determining how to more effectively promote one's products. In a B2B context, where a significant part of a sales organization's activities may be oriented to providing customers with detailed product information and application support, being able to introduce a customer to appropriate new products both creates new sales opportunities and can increase the effectiveness of the sales organization.

While product relations can sometimes be discovered through casual observation, anecdotal evidence, or information directly supplied by the customer, these non-systematic approaches may often miss subtle (but still significant and potentially exploitable) relations. When one has databases of customers and their purchases, or can readily assemble them, another option is to employ data mining techniques to identify potentially interesting product relations. This white paper describes a how one such technique, affinity analysis, works and how it can be applied to discover exploitable relations in product data.

Introduction

Understanding the relations between one's products, as determined by customer buying patterns, can yield potentially useful insights into customer behavior and a needs. One example is in being able to predict what a customer is likely to want based on items that they have already bought. In e-businesses, this is the basis for automated 'recommender' systems, such as those used by Amazon and Netflix to suggest books and movies to customers. For more traditional brick and mortar retail operations, such information can also be useful for discovering selling opportunities and for determining how one promotes products. For example, if the purchases of two products are closely correlated, it makes little sense to offer discounts on both simultaneously. Finally, in a B2B context, where a large part of a sales organization's activities may be oriented around providing customers with detailed product information and aid in product selection, being able to introduce a customer to appropriate products both creates new sales opportunities and increases the sales organization's effectiveness.

While product relations can sometimes be discovered through casual observation, anecdotal evidence, or information directly supplied by the customer, these non-systematic approaches may often miss subtle (but still significant and potentially exploitable) relations. When one has databases of customers and their purchases, or can readily assemble them, another option is to employ data mining techniques to identify potentially interesting product relations. These 'mined' relations then may be screened against additional data and practical business considerations to validate their actual significance and utility. This white paper describes a simple method known as 'affinity analysis' for performing such data mining operations.

Data mining is often viewed as an enterprise-level operation, and one that is only relevant when one has gigabytes of data to analyze. Affinity analysis is also applicable to analyzing modest amounts of data (from a few hundred to a few thousand customers), and are therefore a feasible undertaking for many smaller types of sales organizations.

Terminology:

Before proceeding with a into the subject of product relations, the definition of some basic concepts and terminology will clarify the subsequent discussion:

Product Range

The total set of products from which a customer may choose. For a retail concern, this would be the total of all product SKUs. A catalog distributor's product range would be the set of all catalog numbers for their product line. In cases where there are a very large number of products in relation to buyers it often makes sense to combine them into closely related groups, and consider a purchase of any item in a group to be a purchase of that 'group'

Basket

The set of products selected by a given customer from the product range. For purposes of the analysis techniques described in this white paper, the quantity of individual products selected by a customer is not considered, only the fact that the customer selected some quantity of that product. The basket of a customer who bought two gallons of milk and a loaf of bread would be considered identical to the basket of a different customer who bought one gallon of milk and two loaves of bread.

Relation

A set of one or more distinct products that occurs within a customer's basket. Each customer's basket may contain multiple simultaneous relations between products. In the case of a customer with a basket containing bread, cheese and milk, examples of two relations would be <BREAD-MILK> and <MILK-CHEESE>. Note that the order of items within a relation is not important because we are interested solely in association between items and not causation. For this reason, the relation <BREAD-MILK> can be considered identical to <MILK-BREAD>.

Relation Order

The number of distinct products that comprises a relation. Consider a customer with a basket containing BREAD, CHEESE and MILK. This basket would contain the following relations:

Order 1 (1 item): <BREAD>, <CHEESE>, <MILK>
Order 2 (2 items): <BREAD-CHEESE>, <BREAD-MILK>, <CHEESE-MILK>
Order 3 (3 items): <BREAD-CHEESE-MILK>

Order-1 relations consist of a single product, and don't provide much information beyond a product's average popularity. Higher order relations, however, yield information about how products relate to each other, and provide information that can be used to aid in more effective cross-selling.

In general, as one examines higher-order relations, the number of relations, both potential, and actually occurring in a data set increases rapidly. The following table shows the number of relations of each order that appear in a single customer's selection of only ten items:

Relation Order	# Relations among 10 items
1	10
2	45
3	120
4	210
5	252

Analyzing such potentially large numbers of relations is beyond the scope of manual calculation except in very small and limited cases. Even with the aid of suitable software it can be very time-consuming to effectively analyze sets of relations much beyond order-2 or order-3 for moderately large data sets.

Finding Product Synergies

The overall goal of affinity analysis is to identify relations which might be exploited in some way. This technique is also commonly referred to as 'market basket analysis'. Many qualitative factors of a business, social and regulatory nature affect one's ability to take advantage of a relationship between two or more products. A discussion of these qualitative factors, however, is outside of the scope of this white paper and will not be discussed here. We shall be restricting our discussion to the quantitative aspects of identifying product synergies.

What is of interest from a quantitative standpoint is how often a relationship is observed, and whether this occurrence rate is 'real' or merely an a result of random chance.

The first step in an analysis consists of enumerating all relationships (of a particular order) for each basket in the set of customers. To illustrate, consider a list of customer shopping baskets, the first four of which are shown here:

(1) Milk, Eggs, Juice
(2) Bread, Milk, Juice
(3) Bread, Cheese, Mustard
(4) Mustard, Milk
(5) ....

If one counts the order-1 relationships, one ends up with a frequency table of product selections. Such a table (for a store with a total of only six product types) may look like this:

Product Relation	Frequency
<Milk>	33
<Eggs>	11
<Juice>	22
<Bread>	24
<Cheese>	13
<Mustard>	21

The order-1 frequency tells us how popular an individual product is, and can be plotted out in the form of a histogram. Although it tells us that any given product may be more or less popular than another, it doesn't reveal any information about how the sale of a given product might be related to the sale of another.

It is also possible to build a count of the second-order relationships occurring in the data. With six item types to choose from, there are 15 potential order-2 relations that can exist between products. To continue with our example, one might obtain the following frequency table:

Product relation	Frequency	Product relation	Frequency
<Milk-Eggs>	10	<Eggs-Mustard>	3
<Milk-Juice>	10	<Juice-Bread>	5
<Milk-Bread>	5	<Juice-Cheese>	7
<Milk-Cheese>	4	<Juice-Mustard>	2
<Milk-Mustard>	5	<Bread-Cheese>	6
<Eggs-Juice>	3	<Bread-Mustard>	5
<Eggs-Bread>	4	<Cheese-Mustard>	6
<Eggs-Cheese>	3	–	–

Lift Factor

In the above table, <Milk-Eggs> seems to be a popular combination, as does <Milk-Juice>. One the other hand, very few people seem interested in the <Juice-Mustard> combination. The question is whether these occurrence frequencies indicate some kind of relationship, either positive (in the <Milk-Eggs> case) or negative (in the <Juice-Mustard> case).

One way to determine the significance of a relation is to compare the observed frequency against an estimate of 'expected frequency'. This estimate can be obtained based the frequencies of occurrence of the lower-order relations (in this case order-1) making up the relation of interest. The ratio of the observed frequency to the expected frequency is referred to as 'Lift'. A ratio greater than one indicates a positive correlation between the two products, where buying one of the products is positively associated with buying the other. Conversely, a lift ratio less than one indicates a negative correlation, where buying one of the products is associated with buying less of the other. Note that lift only points out correlations - it doesn't directly indicate causal relations between products.

Product relation	Observed Frequency	Expected Frequency	'Lift' Ratio
<Milk-Eggs>	10	5	2.0
<Milk-Juice>	10	9	1.1
<Juice-Mustard>	2	6	0.3

Computing the lift ratios for product pairs can yield surprising results not obvious from casual examination of the data. For example, although the <Milk-Eggs> and <Milk-Juice> relations occurred equally as often, the <Milk-Juice> relation did not occur any more often than would be expected from the individual occurrence rates of <Milk> and <Juice> considered separately. This suggests there is no special relation between the two. On the other hand, <Milk-Eggs> occurred twice as often as expected, suggesting a possible connection.

In contrast, <Juice-Mustard> occurred about a factor of three *less* often than expected. This suggests that there might be a special inhibiting relation between the two separate items. For example, one possible reason for this lack of a relation is that customers for the two individual products are part of different market segments. Another possible reason is that the two products are competitive substitutes for one another.

Confidence

Another useful metric that can be applied to the data is the *confidence* that if a typical customer buys one product 'A' they will buy another product 'B', or A=>B. The confidence of A=>B can be estimated as the frequency that someone will buy both A and B divided by the probability they will buy A (without regard to their buying B).

Product A	Product B	Frequency A	Frequency A & B	Confidence A=>B
<Milk>	<Eggs>	33	10	30%
<Milk>	<Juice>	33	10	30%
<Juice>	<Mustard>	22	2	9%

Where lift is useful in identifying interesting relationships, confidence is useful in evaluating their exploitable commercial potential.

Statistical Significance

Just because you can observe a relationship in your data doesn't mean that it is *significant*. From a statistical viewpoint, a relation becomes insignificant if it is likely to be a result of random chance, as opposed to a genuine effect. From a practical standpoint, a result becomes insignificant if it can't be exploited in some practical manner. Since practical significance is highly dependent on the overall context of the problem, this discussion shall be limited to the subject of statistical significance.

One way of determining statistical significance is to calculate a Z-score for a given relation. If the Z-score is less than a particular value at a given confidence level (e.g. > 1.65 for 10%), then it is typically assumed that a result is not significant. Conversely, if the score exceeds the threshold value, the relation is held to be significant. It is important to note that the score is not a guarantee of significance, but merely an indicator that the measurement can or can't be distinguished from the noise in the data.

When performing an affinity analysis, one typically considers and tests large number of relations. This creates a problem that numerous relations will show high significance levels, even if they really aren't significant. For example, consider the case where one has identified 100 relations. A Z-score test at the 5% level will typically indicate significance for five of these just on the basis of random chance.

One way of improving the screening process is to increase the confidence level. This unfortunately also has the effect of weeding out potentially significant relations that just aren't that pronounced. A more useful procedure is to use an evaluation process based on separate *exploratory* and *confirmatory* analyses.

The first step in this process is to break up the data into two separate sets, one to be used in each analysis. In the exploratory analysis one looks for all relations in the first data set that meet a reasonable significance level (e.g. 95%). In the the the confirmatory analysis, one again generates the relations, but using the second data set. If a relation appears significant in *both* phases, then it is more likely to be truly significant and not merely a statistical fluke. A key point in performing multiple analyses is that one must use different data for each. Performing multiple analyses on the same data sets will not yield meaningful confirmation.

While performing confirmatory analysis does not guarantee that all of the relations one discovers will be 'real', it does reduce the likelihood that one's final results will not be the result of random noise in the data.

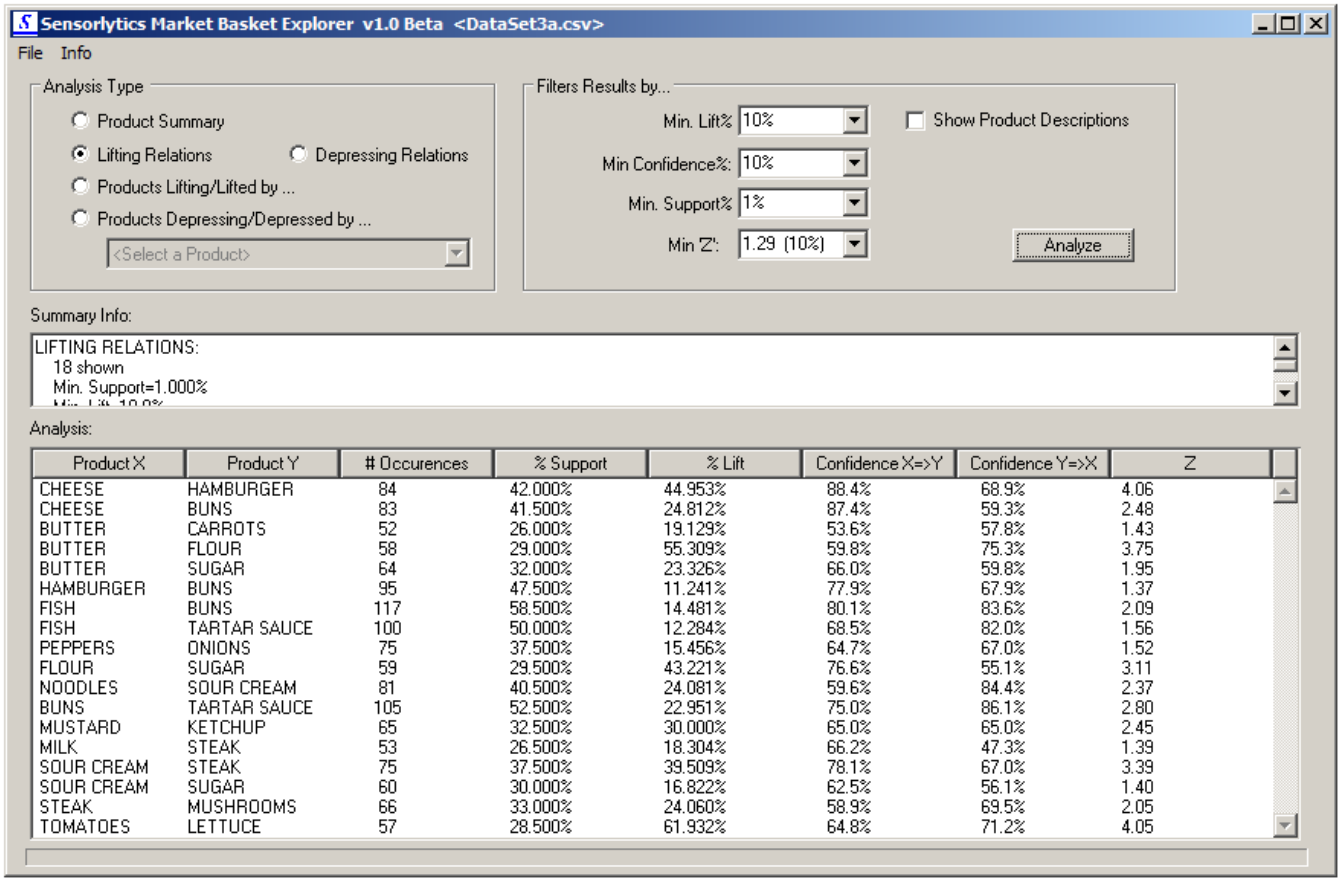
Once one has discovered statistically significant results, one needs to determine if they are significant in a practical sense. When examining large data sets, even very small effects can be tagged as statistically significant, but the magnitude of the effect may be so small that it is impractical to exploit in a commercial sense. One common example is when one sees a a relation that occurs two or three times, but was predicted to occur at a very much smaller rate (e.g 0.001). This type of relation will often show a very high level of statistical significance. From a practical standpoint, however, can you exploit a pattern only observed in one or two customers? It is important to keep in mind that statistical significance does not automatically imply real-world significance!

An important effect we have noted in certain data sets is that this technique can identify structural relations that have been built into the product line by the vendor. For example, many observed relations in a set of web site click data had startlingly high statistical significance ($Z > 50$), which we believe was a result of having pages being located very closely on the web site. For example, two links located on the same page will have a much higher likelihood of being clicked than two links on far-removed pages. By placing the links on the same page, the site-owner had created a strong predisposition for a significant relationship. In the case of retail or B2B purchases, product bundling and cross-selling strategies are also likely to introduce these kinds of biases. For example, if one offers bananas at half price when one buys ice cream, one is likely to see a strong relationship between bananas and ice cream that may not naturally exist except when the promotion is in effect.

While very high values of Z may not be especially useful from a traditional statistical standpoint, we have noticed that they can be used to help discover potentially important relations. In data sets we have examined where certain relations are to expected (and where it would be truly surprising if they were not discovered by a data mining algorithm), a ranking by Z has sometimes been a more effective means of 'discovering' these relations than ranking by other metrics such as lift or confidence.

Software

Sensorlytics Market Basket Explorer (MBE10B.EXE) is a software tool developed by Sensorlytics LLC for use with Microsoft Windows™ that allows a user to interactively analyze small sets of customer basket information, with an emphasis on finding both positive and negative relations between products.



Product X	Product Y	# Occurences	% Support	% Lift	Confidence X=>Y	Confidence Y=>X	Z
CHEESE	HAMBURGER	84	42.000%	44.953%	88.4%	68.9%	4.06
CHEESE	BUNS	83	41.500%	24.812%	87.4%	59.3%	2.48
BUTTER	CARROTS	52	26.000%	19.129%	53.6%	57.8%	1.43
BUTTER	FLOUR	58	29.000%	55.309%	59.8%	75.3%	3.75
BUTTER	SUGAR	64	32.000%	23.326%	66.0%	59.8%	1.95
HAMBURGER	BUNS	95	47.500%	11.241%	77.9%	67.9%	1.37
FISH	BUNS	117	58.500%	14.481%	80.1%	83.6%	2.09
FISH	TARTAR SAUCE	100	50.000%	12.284%	68.5%	82.0%	1.56
PEPPERS	ONIONS	75	37.500%	15.456%	64.7%	67.0%	1.52
FLOUR	SUGAR	59	29.500%	43.221%	76.6%	55.1%	3.11
NOODLES	SOUR CREAM	81	40.500%	24.081%	59.6%	84.4%	2.37
BUNS	TARTAR SAUCE	105	52.500%	22.951%	75.0%	86.1%	2.80
MUSTARD	KETCHUP	65	32.500%	30.000%	65.0%	65.0%	2.45
MILK	STEAK	53	26.500%	18.304%	66.2%	47.3%	1.39
SOUR CREAM	STEAK	75	37.500%	39.509%	78.1%	67.0%	3.39
SOUR CREAM	SUGAR	60	30.000%	16.822%	62.5%	56.1%	1.40
STEAK	MUSHROOMS	66	33.000%	24.060%	58.9%	69.5%	2.05
TOMATOES	LETTUCE	57	28.500%	61.932%	64.8%	71.2%	4.05

Data sets are accepted in comma-separated text format (CSV) so that they may be prepared using a variety of popular spreadsheet and database programs. MBE can provide the following information:

- Independent popularity of given items
- Mutual lift or suppression between items
- Confidence of one purchase driving another
- Support rate of relations
- Statistical Significance of relations

The program provides both interactive output, the ability to 'filter' results by focusing on a particular products and by statistical criteria, as well as the ability to sort results based on output values. In addition, the program can output results in the form of both fixed-width column text files and comma-delimited CSV files for further analysis with other programs. A fully-functional, but limited-capacity demo version of MBE (2000 products maximum) may be obtained free of charge by contacting Sensorlytics.

Further Reading

Devore, J.L., *Probability and Statistics for Engineering and the Sciences, 7th ed.,* Thomson/Duxbury, Belmont CA, 2008.

Krippendorf, K., *Information Theory: Structural Models for Qualitative Data*, Sage Publications, T

About Sensorlytics

Marketing, Communication and Analytic Services for the Sensor and Technology Industries

Sensorlytics LLC is a consultancy that provides services to help sensor and technology companies achieve their marketing goals. Our primary service offerings include:

Market Research & Analysis
Technical Communications
Product Definition
Analytic Services

For more information on what we do, the services and resources we provide, and how we might be able to assist you, please visit our website at:

www.sensorlytics.com