Not Just in Retail

The many application areas of association analysis

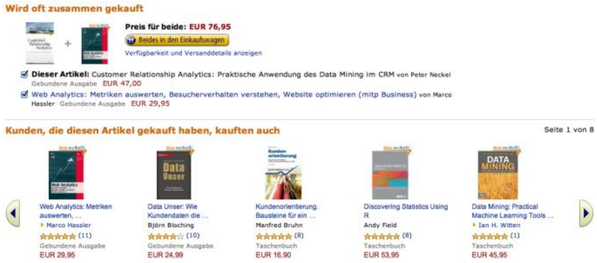
By Peter Gerngross

What is association analysis?

The earliest known example of association analysis – also known as market basket analysis – comes from the late 1980s and was ascribed to the U.S. retailer Walmart. Analysts had found out that diapers and beer were purchased together remarkably often on Friday evenings at Walmart stores. The Walmart managers exploited this phenomenon by placing beer and diapers next to each other in the supermarkets and thus succeeded in increasing sales even more.

We now know that this anecdote is in fact an urban legend[1] and didn't actually happen quite as claimed. Nevertheless, this simple example shows what association analysis is all about: It's about taking a number of grouped events (for example, event: purchase of a product; grouping: shopping cart in a supermarket) and finding frequently occurring combinations of events.

The procedure was first used in retail and also became well-known through its use here. Today, everyone knows the most important feature of the Amazon website (see Fig. 1): "Customers who bought this item also bought ..." For this function, Amazon analyzes large quantities of customer transaction data in near time and is therefore able to create additional buying incentives at the right time – when the customer is in the middle of making a purchase anyway.



Wird oft zusammen gekauft

Preis für beide: EUR 76,95

Beides in den Einkaufswagen

Verfügbarkeit und Versanddetails anzeigen

☑ Dieser Artikel: Customer Relationship Analytics: Praktische Anwendung des Data Mining im CRM von Peter Neckel Gebundene Ausgabe EUR 47,00

☑ Web Analytics: Metriken auswerten, Besucherverhalten verstehen, Website optimieren (mitp Business) von Marco Hassler Gebundene Ausgabe EUR 29,95

Kunden, die diesen Artikel gekauft haben, kauften auch

Seite 1 von 8

Web Analytics: Metriken auswerten, ... Marco Hassler (11) Gebundene Ausgabe EUR 29,95

Data Unser: Wie Kundendaten die ... Björn Bloching (10) Gebundene Ausgabe EUR 24,99

Kundenorientierung. Bausteine für ein ... Manfred Bruhn (8) Taschenbuch EUR 16,90

Discovering Statistics Using R Andy Field (8) Taschenbuch EUR 53,95

Data Mining: Practical Machine Learning Tools ... Ian H. Witten (1) Taschenbuch EUR 45,95

Results of association analyses on Amazon.de[2]

From an historical perspective, association analysis is seen as the first new data mining class[3]. At the same time, it is the most transparent – and the most underrated. For instance, its use is not restricted to retail, as could be assumed. It can also provide important insights into a wide range of issues in many industries, as this white paper will demonstrate. Very often, after all, the correlations found are used for operational purposes.

In the use of association methods, there are only a few parameters and key figures that can influence the result. You also don't need to be a statistician to understand the results: They can be understood almost intuitively and can be used immediately by both business users and customers, as the Amazon example shows.

Thanks to the development of efficient algorithms and also of ever more powerful computers, large data quantities can be processed with affordable standard hardware. In practice, multi-millions of rows are not uncommon.

The development of efficient algorithms was decisive here: Even with the most powerful computers, not all combinations could be calculated. For example, with just 100 products, the number of all possible product combinations (the quantity of all product subsets) is $2^{100} \approx 10^{30}$. In 1993, the first algorithm to solve the problem efficiently was published: the Apriori algorithm[4,5]. This was subsequently improved upon, and new algorithms were also found. As a result, datasets with seven-digit numbers of different products can be analyzed in minutes or even in seconds.

In Section 2 of this white paper, the most important types of association analysis will be presented in brief. Section 3 contains application scenarios from different industries, which illustrate the diverse ways in which the method can be used. Finally, in Section 4, the most important key figures and parameters of simple association analysis are described and explained using a simple example.

[1] See, for example, [Fi]

[2] Source: http://www.amazon.de

[3] Other methods that are now seen as part of data mining, such as decision or clustering methods, developed from conventional statistics.

[4] Origina article on the Apriori algorithm: [AgImSw1993]

[5] The algorithms will not be discussed in further detail in this white paper. For good descriptions of the most important algorithms, see, for example, [TaStKu2005].

Types of association analyses

The term "association analysis" encompasses a whole range of different questions that have one thing in common: They investigate the frequent common occurrence of elements or events. There are different methods depending on the type of relationship between the events. We distinguish between simple relationships, associations, and sequences.

Relationships

A relationship is a common occurrence of elements that have the same reference value. The reference value that links elements is also called a "transaction" in this context. For example, a shopping cart ("transaction") links the products butter and bread ("elements") that are in it. The relationship is that both items are bought by a customer together.

In the analysis of relationships, we look for combinations that occur relatively frequently. For example, a result of relationship analysis could be that butter is purchased together with bread in 40% of the shopping carts examined. One use of such relationship patterns in marketing is to identify interesting product bundles.

In the figure "Results of association analyses on Amazon.de[2]" the upper part ("frequently bought together") is the result of a relationship analysis. The lower part ("customers who bought this item also bought"), on the other hand, is typical of an association analysis.

Associations

Unlike relationships, associations do not only look at the common occurrence of elements, but also at their dependencies. This results in statements such as: "Two thirds of customers who buy butter, also buy bread at the same time" – so the "direction" of the relationship is also examined.

And the frequency of the relationship is interesting here, too: Associations must be relevant in terms of quantity, that is, the common occurrence of the elements examined should display a certain minimum frequency.

The analysis of associations leads to what we refer to as association rules. In the above example, the rules found would be expressed as follows: "If butter, then bread," or in its short form: "butter -> bread." The objective of an association analysis is therefore to find interesting or relevant rules that are as informative as possible.

Receipt no.	Product
4711	Butter
4711	Cheese
4711	Bread
4712	Cheese
4712	Bread
4712	Muesli
4712	Milk
4713	Butter
4713	...
...	...

Table 1: Input data record for association analysis

However, association rules do not have to comprise pairs of elements. Both the left and the right side of the rule can contain more than one element. A further permissible rule would be, for example: "bread, cheese -> wine, butter."

To be able to analyze relationships and associations, just two data characteristics are required[6]: first, the element (for example, the product purchased) and second, the reference value assigned to the element (for example, the receipt number). Table 1 shows a typical input data record for a relationship and association analysis.

Sequences

In the analysis of sequences, there is a third aspect: the chronological order of the elements (here known more appropriately as "events"). Not only is the common occurrence of events examined, but also the chronological sequence of these events.

The generated rules have, for example, the form "purchase of product A followed by purchase of product B leads to purchase of product C" (short form:

[6] In addition, a taxonomy could be examined. A product taxonomy is the hierarchical summarization of products in groups. Taking the example of a supermarket, a taxonomy chain could be as follows: "Kerrygold butter 250g" -> "butter" -> "dairy products" -> "chilled products" -> "food."

"A, B -> C"). Unlike with association analysis, the rules "A, B -> C" and "B, A -> C" are different with association analysis.

In addition to the event and the reference value, a third data characteristic is required with sequence analysis. When applied to retail, you would, for example, select the following three characteristics:

- Event: Purchase of a product
- Reference value: Customer
- Time: Purchase date

In the case of a mail-order company selling sports equipment, the input data record could be as follows:

Customer no.	Product	Purchase date
1234	Dumbbell set	5/3/2013
1234	Gym mat	5/23/2013
1234	Barbell	5/25/2013
1235	Running shoes	6/1/2013
1235	Running pants	6/1/2013
1235	Running jacket	9/23/2013
1235	Running shoes	10/24/2013
1236	Indoor sneakers	7/9/2013
1236	...	...
...	...	...

Table 2: Input data record for sequence analysis

An additional result of sequence analysis is statements about time periods between the events. For example, in the above example, the average amount of time between the purchase of the dumbbell set and the gym mat as well as between the gym mat and the barbell can be calculated for the rule "dumbbell set, gym mat -> barbell." This information can be used in direct marketing, for example, to control the times of advertising stimuli.

Application scenarios

Association analysis is classically used in the (retail) trade, as has already been mentioned. The methods were developed for this industry and have been used operationally for a long time. Application scenarios are discussed in detail in the relevant literature[7].

Nevertheless, there are also application scenarios in many other industries. Some of them are described in this section.

Example 1: Optimization of the supply of replacement parts for large medical devices

Initial situation

A leading producer of medical technology has concluded maintenance contracts with doctors and hospitals for the devices it manufactures. These are valuable large devices such as X-ray units or computer or magnetic resonance imaging systems, the prices of which are typically in the one- to two-digit million euro range.

If a device malfunctions, a technician travels to the site to maintain the equipment. Replacement parts are usually required, which are often very valuable and have high transport costs.

During normal operation, the devices permanently send technical data about their condition and status to the manufacturer. This sensor data is stored in a data warehouse and is available for evaluation purposes.

Requirements

If a device malfunctions, the technician takes with him a few standard replacement parts that are often required. But in many cases, he discovers on-site that other components are needed. This means high transport costs for special parts and a delay that leads to a long downtime for the device.

In the project, analyses are performed with the aim of forecasting in advance – that is, after the machine has broken down and before the technician has left for the site – which replacement parts are necessary for the maintenance work.

Solution

On the basis of the log data stored in the data warehouse, rules are generated with the help of association analysis that accurately determine the set of replacement parts needed when a machine breaks down.

[7] See, for example, [NeKn2005] p. 222 ff

The log data is stored in the form of event data records that contain the two key pieces of information, namely event ID and the device ID. In addition, the event's timestamp is stored.
To generate the rules, device breakdowns from the past are examined. From these breakdowns, we can see (in retrospect) which replacement parts were needed. Every replacement part has an ID. This information is used to generate "artificial" event data records for defective devices. These contain the component ID and the device ID.
Using association analysis, frequently occurring combinations of the form
event 1 + event 2 + … + event N -> component A + … + component X
are sought, that is, conclusions can be made about the replacement parts needed for the maintenance work on the basis of a combination of log events.

Results

By implementing the procedure described, it was possible to increase considerably the rate of correct replacement parts taken to the customer. The consequence was that drastic savings could be made in maintenance costs and downtimes were much shorter.

Example 2: Cross-selling in multichannel sales of financial services

Initial situation

A financial services provider has a number of subholdings that use different distribution channels for customer acquisition and for product distribution to the installed base. The sales channels are in some cases entirely separate in organizational terms. Product ownership among customers differs greatly between the distribution channels, although cross-selling between the subholdings has been pushed during the past years.

Requirements

The financial service provider's central marketing department is interested in intensifying the use of cross-selling. This requires more in-depth information about entry-level products, additional purchases, and product purchase chains.

The goal is to automatically generate customer-specific product recommendations for inbound sales, the website, and the branches.

Solution

A central dataset was set up that comprises the products purchased by the customers of all the subholdings. It contains the customer number, the product code including taxonomy, and the product's date of purchase.
A sequence analysis was performed based on this dataset. First, this enabled typical entry-level products to be determined. Second – and more importantly – frequently occurring product purchase chains could be determined. An example would be

1. time deposit + 2. checking account
leads to
3. securities account

Such chains can be integrated very well into CRM software solutions or operational call center systems and thus give sales employees important information about the best-possible product offering for the customer at the current time.

Results

Telephone inbound sales and branch sales rose significantly. Furthermore, it was possible to increase the banner click rates on the website. Ultimately, the procedure makes an important contribution to reaching the goal of continuously increasing the groupwide cross-selling rate.

Further project steps

More "shopping cart objects" are to be included that, in a strict sense, are not actually products – such as product terminations and competitive information. This should, on the one hand, increase the forecast quality and, on the other hand, enable other results such as the termination of customer relationships to be modeled.
Another possibility would be "reverse modeling." For example, it could be interesting to find out in which cases checking account customers do not conclude a credit card contract.

Example 3: Website optimization for an insurance company

Initial situation

An insurance company operates a website that, on the one hand, informs existing and potential customers about products and services and, on the other hand, enables customers to conclude insurance contracts online. This means a large amount of information is provided on the various subpages. At the same time, however, concluding contracts on the website should be made as simple as possible.

Transaction data from the website visitors is already gathered in an anonymized form and stored in the data warehouse. This data contains a session ID (a key that uniquely identifies a website visitor from the time of entering to the time of exiting the website), the subpages visited during the session, and the point in time in each case.

Requirements

The website requires optimization. As part of this optimization, access to the product purchase pages should be made as easy as possible.

For this purpose, analyses should be performed to determine whether there are frequently used and lengthy click paths that lead to a product purchase page and that can be simplified.

Solution

The web data stored in the data warehouse was investigated using a sequence analysis. All the data required for this was available. In addition, it was necessary to define which subpages are categorized as "product purchase pages."

The sequence analysis delivers rules like

Page 1
followed by page 2
followed by ...
followed by page M
leads to
page N

Accordingly, rules where "page N" is a product purchase page are interesting. Furthermore, the rules had to have a certain minimum length (that is, "lengthy" click paths were sought) and they should occur relatively frequently.

It was possible to find rules that meet these requirements automatically using sequence analysis. But some human intervention is still necessary: The rules (that is, click paths) had to be subsequently checked in terms of content, in order to determine whether and, if applicable, how the path to a product purchase can be simplified.

Results

In many cases, the path to an online purchase was made considerably easier for prospects with the intention of purchasing an insurance product. After the website optimization, a new analysis of the click paths revealed that the rate of "dropouts" who did not make it to the purchase page due to the unnecessary and complicated paths decreased considerably. The effectiveness of the website as a sales channel was therefore increased significantly.

Interesting associations

Central key indicators: support and confidence

How can we tell whether an association rule is relevant? First of all, there are a number of central key indicators that we can use to find potentially interesting rules. Put differently: From the huge number of possible rules, a tiny minority is filtered out. These can be examined in more detail, while the vast majority of rules can be ignored with confidence.

The first key indicator – support – specifies the relative frequency of the element combinations contained in the rule.

Let's assume that the rule "A -> B" should be investigated. To calculate the rule's support, we count the number of transactions that contain both A and B. This figure is divided by the total of the shopping carts N.

$$\text{Support}\ (A \rightarrow B) = \frac{\#(A \text{ and } B)}{N}$$

With support, we can assess the quantitative relevance of a rule. In most cases, a rule is not relevant if it only occurs in a few transactions. Confidence is the second central key indicator in association analysis. It specifies the reliability of an association rule.

Let's look at the rule "A -> B." As with support, the number of transactions in which both elements A and B occur is counted first of all. Then we count the number of transactions in which element A occurs. Confidence is the quotient of both figures:

$$\text{Confidence}(\text{A} \rightarrow \text{B}) = \frac{\#(\text{A and B})}{\#(\text{A})}$$

Confidence can be interpreted as a conditional probability: If we know that element A is in a shopping cart, how probable is it then that B is also in the cart? It specifies the reliability of the rule: How sure can I be that the rule applies?

In most data mining tools, support and confidence are the only parameters that must be set in an association analysis. In each case, you enter a minimum value and, as a result, receive only rules in which support or confidence surpasses the minimum value. Setting these restrictions is essential for efficiently determining rules. The algorithms implemented in the tools use the restrictions for excluding large amounts of the possible quantity of rules.

Many data mining tools specify further indicators for the rules found, which can be used to restrict the quantity of good rules still further. An example here is lift, which specifies the factor by which the confidence of a rule exceeds the simple (unconditional) frequency[8].

$$\text{Lift (A}\rightarrow\text{B)} = \frac{\text{Confidence (A}\rightarrow\text{B)}}{\text{P(B)}}$$

Examples of relevant and irrelevant associations

Not all associations that are determined algorithmically are also relevant. First, relevant rules must be reliable and quantitatively relevant, which is automatically ensured through filtering using the parameters of confidence and support.

A further criterion for a good association rule is that it is not obvious. In our supermarket example, "beer -> mineral water" would be an obvious rule. You don't need to perform a data mining analysis to establish this rule. All you need is common sense. In such cases, the analysis would at best be suitable for confirming or disproving the obvious rule.

Rules that occur systematically should also be excluded. Example: An online shop encloses a brochure with each order. This is saved in the receipt data. If these items are not eliminated from the data before the analysis, some – or, depending on the support restriction, many – rules of the form "product A -> brochure" will occur that all have a confidence of 100%.

Sometimes, only rules with a certain form are relevant. In the above example about sensor data, we only want rules with exclusively replacement parts on the right. Rules of the form "event 1 -> event 2" are irrelevant here.

In general, we can say that good rules should display the following characteristics:

- Reliable
- Quantitatively relevant
- Previously unknown
- Not systematically occurring
- Relevant in terms of content

Sample calculation for a supermarket

The purchases of five customers are noted in a supermarket:

Customer	Shopping cart
Anton	Butter, cheese, bread
Berta	Cheese, beer, muesli, milk
Carl	Butter, bread, beer
Dora	Butter, cheese, milk
Eric	Cola, muesli, milk

Table 3: Sample calculation

In total, the customers bought six different products. What rules can we now generate from this? Three of

[8] For indicators and the most commonly known algorithms, see, for example, [TaStKu2005] p. 327 ff

the five customers have butter in their shopping cart (Anton, Carl, and Dora). Furthermore, Anton and Carl have also bought bread.
The following rule should now be investigated: "If a customer buys butter, then also bread." The central key indicators are then calculated:

Support

$$\text{Support (butter}\rightarrow\text{bread)}$$

$$= \frac{\text{Number of carts with butter and bread}}{\text{Total number of carts}}$$

$$= 40\%$$

Confidence

$$\text{Confidence (butter}\rightarrow\text{bread)}$$

$$= \frac{\text{Number of carts with butter and bread}}{\text{Number of carts with butter}}$$

$$= \frac{2}{3} \approx 66.7\%$$

We can interpret this confidence as the (conditional) probability of the purchase of bread, if it is known that a customer buys butter.

Lift

The simple probability of the purchase of bread is

$$\text{P(bread)} = \frac{\text{Number of carts with bread}}{\text{Total number of carts}} = \frac{2}{5} = 40\%$$

From which we can derive lift:

$$\text{Lift (butter}\rightarrow\text{bread)}$$

$$= \frac{\text{Confidence (butter}\rightarrow\text{bread)}}{\text{P(bread)}} = \frac{\frac{2}{3}}{\frac{2}{5}} \approx 1.67$$

Bibliography

[Fi] Donald Fisk: Beer and Nappies – A Data Mining Urban Legend. Source: http://web.onetel.net.uk/~hibou/Beer%20and%20Nappies.html

[DMS2010] mayato Data Mining Study 2010 - Field Test and Benchmarking. Source: http://www.mayato.com/de_DE/kompetenzen/fokusthemen/data-mining-studie.html

[AgImSw1993] R. Agrawal, T. Imieliński, A. Swami: Mining association rules between sets of items in large databases. Proceedings of the 1993 ACM SIGMOD international conference on Management of data - SIGMOD '93. p. 207

[TaStKu2005] P. Tan, M. Steinbach, V. Kumar: Introduction to Data Mining. Addison Wesley 2005

[NeKn2005] Peter Neckel, Bernd Knobloch: Customer Relationship Analytics: Praktische Anwendung des Data Mining im CRM (The Practical Application of Data Mining in CRM). Dpunkt Verlag 2005

The author

Peter Gerngross is an expert in data mining, business intelligence, and analytical CRM with the analyst and consulting company mayato. He focuses on the qualitative optimization and processing of customer data, as well as its use in sales.

Peter Gerngross | mayato GmbH
peter.gerngross@mayato.com

About mayato

mayato is an analyst and consulting company that specializes in business intelligence and business analytics. A team of experienced IT and BI architects, statisticians, analysts, and experts on special subjects such as fraud detection, data mining, and analytical CRM operates from its offices in Berlin, Bielefeld, and Heidelberg. Among mayato's customers are renowned companies from a range of industries. As a partner of several software providers, mayato is committed to remaining neutral and – first and foremost – to delivering its own high-quality services. For more information, visit: www.mayato.com.

mayato GmbH

Am Borsigturm 9
D-13507 Berlin
Germany

Tel. +49 30.4174.8657
info@mayato.com

www.mayato.com