Matching of product data

How you can create a clean database using adaptive match-algorithms

Introduction

Whether an online shop or brand-name manufacturer, companies in numerous sectors are facing the challenge of managing an increasing volume of product data. One of the problems is multiple-entry product data (“doublets”, “duplicates”). If, for example, a product is obtained from various suppliers, then that item can appear repeatedly in the product databank under various titles and descriptions. In the worst case scenario, both the consumer and product manager will be advised that the item is not available at all. The customer will then purchase from another online shop and the product manager create a repeat order, even though sufficient stock already exists

The process of automated identification of similar or equivalent datasets is termed “matching”. The domain and the application scenario determine here when two datasets are considered similar or equivalent, and thus a “match”.

When matching is tasked with identifying duplicates in the product master-data of an online shop, then two datasets become a match when they represent the same product. By contrast, however, in the use of matching for competitor monitoring by brand-name manufacturers, two datasets become a match when relating to comparable products.

The advantages of matching arise from the following particular cases:

_ Improvement of data quality Matching helps to identify and remove duplicates, thus creating a clean database. In online shops, for example, this improves usability and provides for contented customers who will gladly purchase (again) here.

_ Automated price monitoring: Matching facilitates a comparison of your own prices with competitor offerings and as a result permits sound analyses of competitor pricing strategies.

This whitepaper should convey to you an overview of the new possibilities for the matching of product data. You will discover how to create a clean and reliable database using these new automated methods.

Status quo

Even though the majority of globally-active companies have long since recognized the great importance of a clean and reliable database to the efficiency of their business processes, only a few of these have already drawn the necessary consequences. This is the key finding of the study "Strategic Management of Master-Data Quality", in which strategy and organization advisors Camelot Management Consultants surveyed 56 decision-makers from globally-operating companies spanning all sectors and sizes about their approach to quality assurance for their company data. Over half of those companies surveyed stated that inadequate master-data quality still massively affects the processes along the entire value-creation chain in a negative way. Around 60% of those polled see an enormous backlog demand in the measurability and control of data quality, as well as in the introduction of automated tools.[1]

An independent user survey by TDWI Europe also came to a similar conclusion. Only 36% of those surveyed were satisfied with the data quality at their company. The majority polled envisage positive effects with an improvement in data quality, above all in the areas of reporting, customer retention and process optimization. Of particular interest is that some participants assume the annual costs to their company due to lacking data quality amounts to over 20 TEUR.[2]

Up until now, in-house employees have often been engaged in manually correcting product data or recording competitor prices. In the case of small product data volumes, this is possible with relatively few problems. With increases in data volumes, however, the effort required then rises disproportionately.

The overwhelming majority of automated matching tools in existence have so far offered no specialist solutions for matching product data. The specific challenges to be overcome for product data will be clarified below.

Challenges

The biggest obstacles to the matching of product data lie in the handling of heterogeneous product descriptions and a shortfall in consistent numbering for clear product identification (EAN/GTIN numbers, for example). Varying product descriptions arise among other things from the use of:

_ Domain-specific terms and abbreviations: In the fashion sector, for instance, the term "lg. A." is an abbreviation for "long arm". While over in the electronics segment, LG is a brand-name manufacturer.

_ Heterogeneous designations for sizes and quantities information: 168 units vs. 3x56 units.

_ Heterogeneous notation of model designations: DHI655FX vs. DHI 655 FX vs. DHI-655 FX.

_ Synonyms: In the fashion world, for example, a hoody, hoodie and hooded pullover all describe a pullover with an integral hood.

One general challenge to matching is its effectiveness: To evaluate this, two measures are normally applied: Precision and recall.

[1]https://www.hs-heilbronn.de/6691397/Trendstudie_SDQ_2013_Management_Summary.pdf

[2]http://www.emagixx.de/images/Unternehmen/Studie_Datenqualitaet_in_Unternehmen.pdf

Precision measures the proportion of actual matches against recognized ones and is thus a gauge of accuracy. A high precision implies that recognized matches can be accepted as such without due concern.

Recall, on the other hand, measures the proportion of recognized matches against all actual ones, and is thus a gauge of matching completeness. A high recall therefore attests that many of the actual existing matches were found.

Depending upon usage case and quality of the initial data, it is difficult to maximize both the precision and also the recall. Normally, it is therefore necessary to weight both targets on an application-specific basis. When the focus rests upon exactness (precision), it must be accepted that some potentially correct matches will not be recognized. Focus, on the other hand, resting upon completeness (recall) potentially results in some false associations.

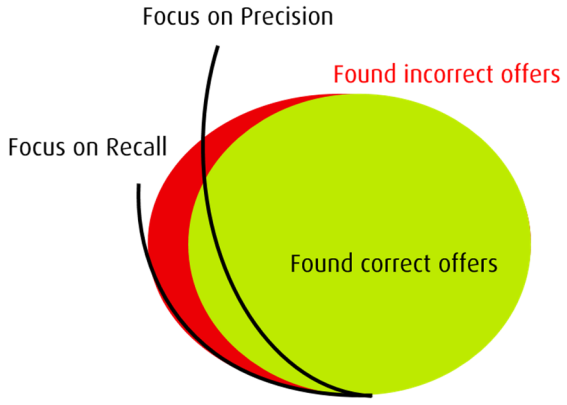

Figure 1: Recall versus Precision

Further challenges to the matching of product data arise, depending upon specific application, from the following factors:

_ Data volume: Depending upon the situation, a huge volume of data has to be handled, whose scale directly influences the runtime whilst matching.

_ Frequency: The more frequently the matching of a (large) database needs to be repeated, the more extensive the process becomes.

_ Alteration of the database: The more often the database alters from which matching is performed, the more often new (extensive) matching processes must take place.

Benefits

Companies that use automated software solutions for the matching of product data profit in several respects. The internal resources for manually checking and associating of the data is avoided. Particularly in areas of higher dynamics, the efforts put in to matching under altering conditions can be drastically reduced so that fewer internal resources remain tied up.

Furthermore, depending upon the particular application of the matching, an additional individual use appears.

Examples:

After optimizing product data, visitors to an online shop find the desired product information significantly more easily, since the data is now clearly structured and also delivers better results in the search. Well maintained and consistent product information creates more trust in the customer. The benefits to the online shop reveal themselves in higher customer satisfaction and rising sales figures.

Consistent product data also provides for increased transparency during the purchase. Shoppers receive a quick overview of available stocks and can easily compare the purchase prices from various suppliers despite varying designations. This creates the potential for higher margins and prevents duplicate stocks from being held.

Besides the internal matching of the own products of various suppliers, shop operators have the option of comparing their own product range directly with the offerings of various competitors. Of particular interest here is the monitoring of prices for corresponding products held in the product lines of competitors. With the help of precise and reliable market data, retailers can then tailor their price-setting strategy to the competitor situation and thus increase their competitiveness. This procedure is described in more detail in the whitepaper “Price monitoring on the Internet - how you can optimize margins using structured competitor data”.

Matching Process

The matching process can be roughly subdivided into the four phases of data extraction, preprocessing, matching and reporting. These four phases are detailed below using examples from the blackbee software.

1. Data Extraction

Firstly, the data is provided from which a matching is to be performed. The data to be matched can exist within various source systems, such as different databases or text data of varying formats. blackbee supports diverse source systems and data formats. Furthermore, the integration of product feeds and the querying of diverse web sources are possible, such as web services, price comparison portals and online shops.

2. Preprocessing

The data is typically preprocessed first of all, in order to easier facilitate the actual matching procedure. Included in the preprocessing stage, among others, are the standardisation of attribute values and the supplement of missing attribute values, as well as the further addition of other attributes.

Standardization delivers attribute values in a consistent format, which makes heterogeneous manufacturer designations comparable, such as “HP” and “Hewlett-Packard”. The supplementing of missing attribute values can be performed with the help of a reference list. Such a list, for example, will be assembled from available data for the attribute of “brand”, so that the brands for products with a missing “brand” attribute in either their title or description can then be supplemented. By adding further attributes, product data can be enriched with clearly identifiable characteristics. The product code can thus be extracted from the title or description of a product as an additional attribute. A product code is a unique character sequence selected by the manufacturer for the purpose of product identification. For a product with the title “Bosch DHI 655FX grey-metallic extractor hood fixture”, the product code would be “DHI 655FX”.

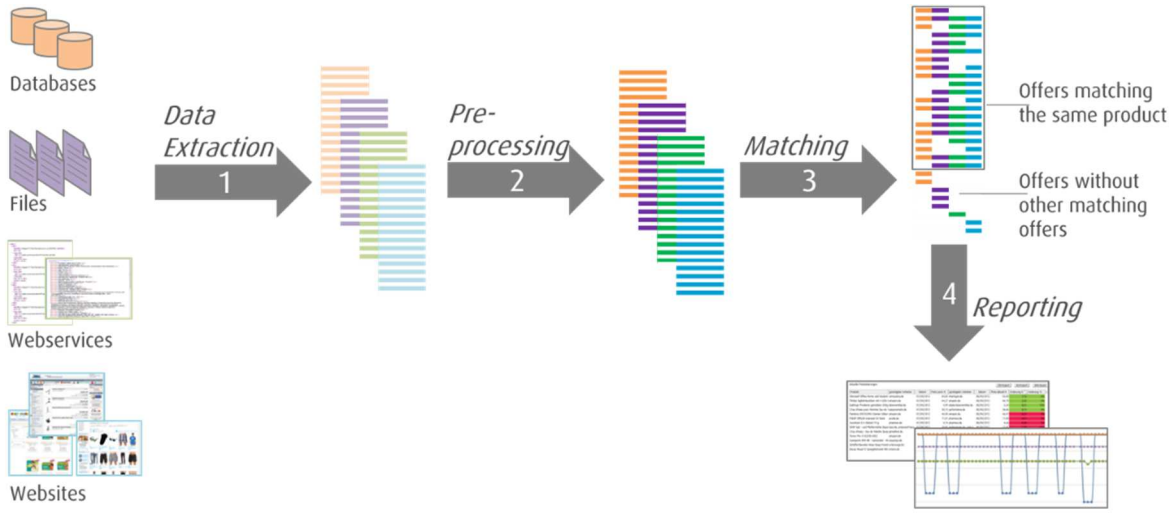

Figure 2: Matching Process blackbee

3. Matching

n this step, the individual datasets are compared against one another. It is essential with large data volumes to firstly limit the search space, thus avoiding the need to compare each dataset against every other one. In addition, there are also so-called blocking procedures to apply, whereby the product data is divided into blocks. This can take place in various ways. One possibility is the generation or use of a block key. All product data having the same block key (such as the product brand) are pooled into one individual block. Product data contained in the same block (namely products of the same brand) will now only be compared with one another.

The comparison of two products generally takes place by a comparison of their attributes. A variety of different similarity measurements exist for this comparison of attributes. Common measurements for similarity include edit distance, Jaccard, tf-idf and trigram. It is essential to the quality of matching that appropriate similarity measures are selected. In the majority of cases, one single measurement does not lead to an optimal result. A more promising tactic is to combine various measures for the different attribute values (such as the "article designation" and "brand" attributes). Determining an effective matching strategy, due to the multitude of existing procedures, is often challenging even to an expert. In this area, adaptive learning methods, such as those provided by blackbee, deliver crucial advantages. Machine learning techniques reduce the manual tuning efforts required by determining an optimised matching strategy semi-automatically ("learning"). These techniques require positive and negative "training data" in the form of examples for matches and non-matches. blackbee supports the generation and maintenance of such training data, as well as the development of a feedback mechanism. Association errors can be corrected by taking this feedback into consideration. The system learns from these corrections and thus improves in accuracy from one run to the next.

4. Reporting

The result of matching is a large bulk of correspondences. A correspondence here is a pair from two sets of product data identified by the system as matches and a similarity value indicating to what probability an actual match is likely. Depending upon specific application, these results can be processed for further analyses and used for a wide variety of reports. Regarding price monitoring, the report function allows, for example, an overview of the Top 5 vendors to be generated for each product. blackbee supports the export of results and generated reports in various formats (such as Excel, CSV).

Summary

Companies see a major backlog of demand in the area of data quality. Whereas the majority of providers for data quality solutions offer no matching functionalities or remain specialized in customer data, blackbee also provides a service for the matching and correction of product data. The introduction of adaptive matching algorithms for the quality improvement of product data offers companies a strategic advantage in dealing with large data volumes. This is particularly true for companies whose data frequently changes. When using the blackbee software solution, companies are in a position to process large volumes of product data in a short time and repeatedly correct it to remove duplicates. Data can also be brought together for a price comparison of the offerings from various vendors. The reduction in manual resources and avoidance of poor decisions based on faulty data represent two decisive advantages arising from the introduction of the blackbee software.

About Webdata Solutions GmbH

The e-commerce service provider came into being in 2012 as a spin-off from a research project at Leipzig University and today counts among the market leaders worldwide in the field of **online market analysis**. The solutions based on the innovative **blackbee** platform technology have been successfully put into practice by leading online retailers and manufacturers. Webdata Solutions reduces the complexity arising from the plethora of product and product-related data on the Internet, and also generates **information** focused **upon the benefits to business**. The company has set the target of broadly raising the potential of web data using blackbee and, by applying the **correct core information at the right time**, contributing towards e-commerce becoming a **transparent market**.

Hanna Köpcke
CTO

Contact

Phone +49 (0) 341 – 351 361 – 70

E-Mail info@webdata-solutions.com

Internet www.webdata-solutions.com

Authors: Hanna Köpcke
Carina Röllig

Jacobstraße 5
04105 Leipzig