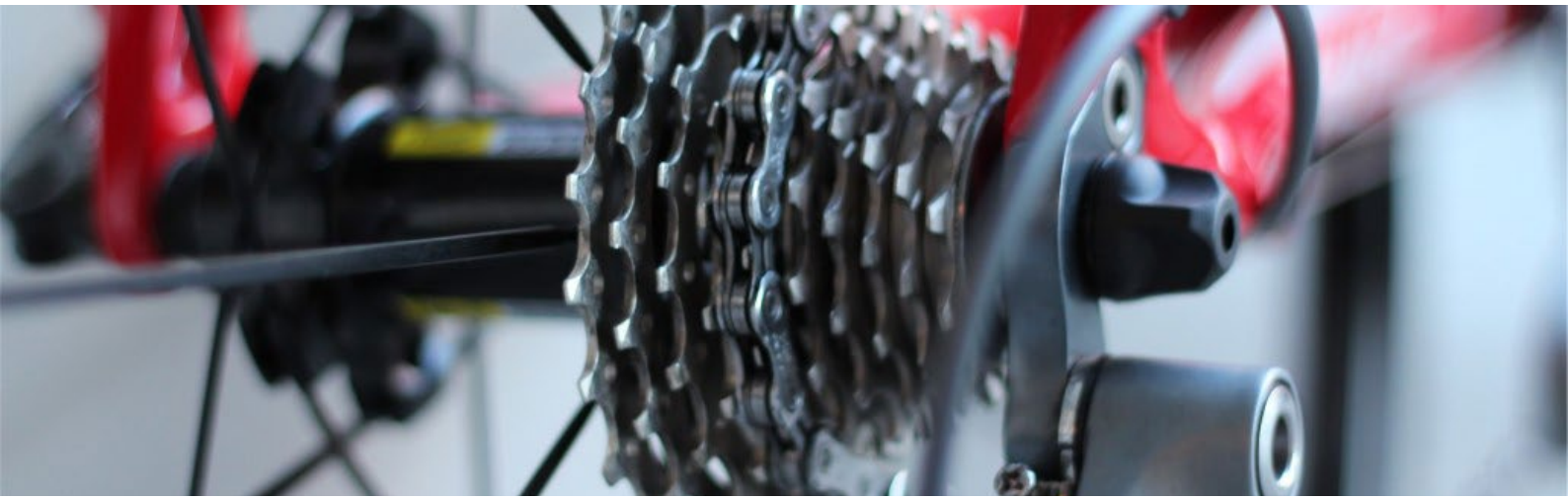
Product matching excellence is the key to successful market intelligence in the age of big data



Summary

Proper identification and matching of product items lie at the heart of web-based market intelligence. While manual research is futile in the face of big data processing requirements, many automated systems produce results of limited value. The key to achieving high market transparency is a software's ability to adapt its product matching logic to the specifics and idiosyncrasies of any given industry. Because all products were not created equal.

Businesses need more precise market information at faster speeds

The disruptive effects of e-commerce on industries and businesses are felt worldwide. They have led to an increasing need for market information at ever increasing rates. Not only e-commerce businesses but enterprises of various industries struggle to keep up with market trends, with existing and future competitors - and with their customers, who can find almost any information online. Comparison shopping becomes the norm. It forces businesses to gather and analyze market information permanently. Because competitiveness always relies on information advantages.

The product is the heart of the matter

For most goals pursued by market and business intelligence, products and their features are the main focus. Businesses need to know how their products, prices and conditions compare on the internet in order to make sound decisions and stay ahead of the competition.

A product is a product is a product - or is it?

In theory, you should encounter no substantial obstacles when trying to find and compare identical products on the internet. Numbering systems such as EAN were introduced to make this an easy task. But the reality of online shops, shopping portals as well as in-house databases is far from such an ideal state of transparency, for various reasons:

- In many cases, EAN or other unique identifiers (e.g., isbn, pzn) are lacking
- In many other cases, EAN or other unique identifiers are incorrectly assigned (sometimes deliberately, to mislead competitors)
- Product descriptions use different wordings: Converse All Star Ox Plimsolls vs. Converse CHUCK TAYLOR ALL STAR - Trainers – navy

Product descriptions give different pieces of information. Consider these widely differing descriptions which two merchants give for one and the same power supply:

Product data	
Type	TEX 120-124
Weight	1 kg
Weight	Synchronised
Height	56 mm
Width	174 mm
Length	93 mm
Min. input voltage	85 Vac
Connection	Plug
Max. output current	5 A
Max. input voltage	264 Vac
Power	120 W
Output volate (max.)	28 Vdc
Category	AC7DC PSU module
No. of outputs	1x

Fig.1 Varying of product data I

Data	
Input Voltage AC Max:	264V
Input Voltage AC Min:	85V
Output Current Max:	5A
Output Power Max:	120V
Output Voltage Nom.:	24V
Power Supply Output Type:	Fixed
SVHC:	To Be Advised

Fig. 2 Varying of product data II

- Attributes carry varying degrees of relevancy in different product categories
  - In the above example the following attributes are important for matching: min / max input voltage, max output current, power
  - On the other hand the attributes height, width, length have in this case no relevance for the matching. But these attributes are for example highly relevant when matching furniture articles

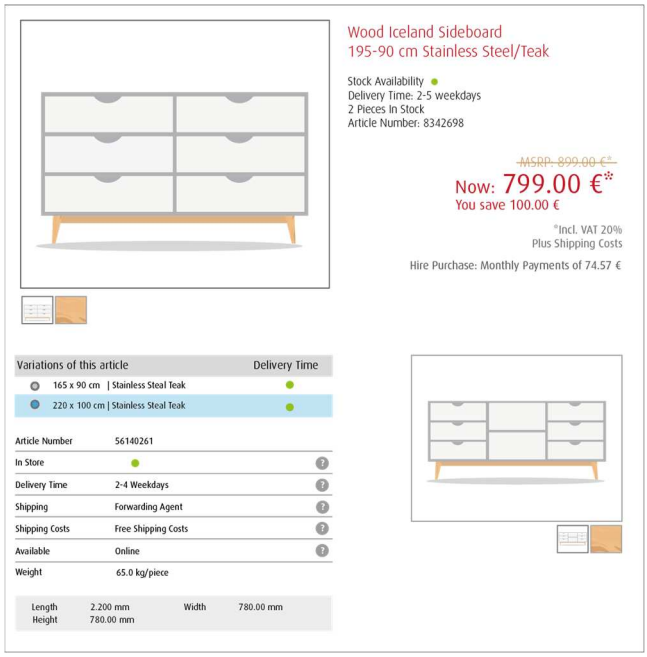


Fig. 3 Varying of product data III

- In the example width and length are relevant to correctly match different variants
- Markets are multi-lingual
- What exactly constitutes "a product", can vary

A product can be defined on different levels of detail and depends on how variants are treated: model level, model + color level, model + color + size level.

Fig. 4 Varying of product data III

Product equivalence therefore is contingent on many factors. Many automated systems lack the ability to properly take these into account. They have severe difficulties with identifying the correct products.

Recall and precision: the hard currency in product matching

When matching products, the ultimate goal for any system is to find the perfect balance between coverage and accuracy, or in technical terms: *recall* and *precision*. In an ideal world, both would be 100%:

- Recall: you want to identify every existing equivalent product in the sources you analyze.
- Precision: you also want to make sure that only equivalent products find their way into the matching algorithm

Under real world conditions, though, there is a trade-off between recall and precision.

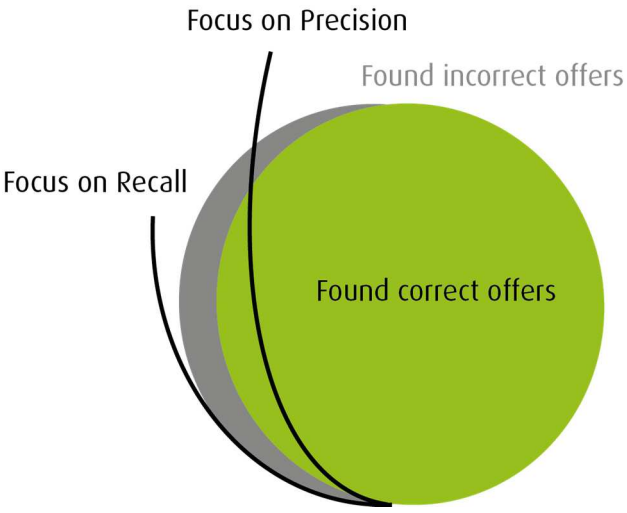


Fig. 5 Recall vs. Precision

A very narrow definition of what qualifies as an equivalent product will leave a large blind spot on the market analysis. The matching process will ignore many cases where equivalent products are presented and/or described differently. Data quality in terms of *precision* will be strong, but *recall* will be weak.

On the other hand, a wider definition allowing more variations will enable you to get a more complete picture of the market while at the same time increasing the risk for comparing apples and oranges. This time, *recall* will be strong, but *precision* will be weak.

Systems need training to balance recall and precision

The main challenge for finding the perfect balance between recall and precision arises from the fact that you can easily validate *precision,* but not *recall. Recall* can only be determined for reference data sets. The real size of the blind spot simply cannot be known - you do not see what you cannot see. A carefully composed reference set does serve as a good approximation, though. Ideally, it reflects the way a company and its industry structure information about products: their specifications and classifications, their attributes used and how they are expressed, and the priorities as to what signifies equivalence to what extent.

It is the data set to test and train your system, so it understands how products are presented and described in your industry. The matching process needs to be refined and adapted until it produces results in accordance with the product matches which the reference data set has marked to be correct.

Here's the catch: This training data set needs to be created. And: the work is never truly finished.

Positive training data	Negative training data
Max 2016 Style: 127561C	Max Blue/White/Black
Max 2016 - Men Shoes Navy Blue-White-Black	Max 2016 - Men Shoes Cobalt Blue-White
Max High Top 2016 - Style: 127561C - blue/white	Max Top 2016 - Women 80674 - Women Shoes
Max 2016, Navy Blue/ white-black, 7 M US	

Fig. 6 Training data example

Teach your system well

In order to achieve excellent product matching results, a system needs the ability to go on a steep learning curve - not only when generating and processing training data, but also during on-going projects so you can integrate new learnings and keep track of changes in the market place. Therefore, advanced market intelligence system provide these key features:

- Rules, synonyms and dictionaries can be defined, added and managed at any time
  - Example rule: Different sizes (XS,S,M,L,XL,XXL) do not match
  - Synonym: Hoody = Kapuzenpulli
- A validator supports manual validation
- Manually validated matches can be seamlessly integrated into the workflow
- Enhance and link product data with additional information from internal and external sources

How complex the learning process turns out to be is largely dependent on the industry you are looking at.

The acid test: Can you do fashion?

The fashion industry is notorious for a seemingly unfathomable product landscape. Its complexities pose some of the biggest challenges to market intelligence software. Fashion products are exceptionally tough to match for several reasons:

- The same EAN is used for a whole set of variations on a piece of clothing
- The naming of attributes can be decidedly idiosyncratic (Adidas uses "electricity" for the color yellow)
- Attributes used to describe equivalent products can vary extremely
- Brand is not necessarily an important differentiator
- Product descriptions use a multitude of languages (a German website may use yellow or Gelb or jaune etc.)

The absence of brand names as a means to determine product equivalence is especially unfortunate from a market intelligence perspective. It means that product titles are rendered meaningless for product matching, and the process has to focus on attributes whose priorities are much more difficult to establish.

It's the process that makes the difference

It goes without saying that performance and scalability are important aspects. The algorithm used needs to be state-of-the-art. Smart data partitioning and a powerful infrastructure are required to keep calculating time from spiralling out of proportion in big data projects.

Usually, these requirements are met by serious market intelligence software services. The real difference lies in the process. The software will generate high quality results if it is able to integrate contextual knowledge into the matching process. It needs to reflect an industry's knowledge about its product landscape as well as the individual preferences of the company using it.

Implications

Calculating product equivalence produces excellent results if the software is able to seamlessly and efficiently integrate context-specific user knowledge into the process. The same holds true for any set of entities existing in different contexts which are subject to permanent change.

About Webdata Solutions GmbH

E-commerce service provider Webdata Solutions GmbH, Leipzig, was founded in 2012 as a spin-off from a research project at Leipzig University, and has quickly established itself as one of the world-wide market leaders for **online market analysis**. The company's innovative **blackbee** platform technology produces valuable and exclusive market insights for leading online retailers and manufacturers, substantially increasing their competitiveness and profitability. Every day blackbee crawls and matches millions of product data on the the internet and reduces the complexity arising from the plethora of product and product-related data, converting it into highly relevant business information. Webdata Solutions leverages the **true business potential** of data on the internet.

Dr. Hanna Köpcke
CTO

Contact

Telephone +49 (0) 341 – 351 361 – 70

E-Mail info@webdata-solutions.com

Internet www.webdata-solutions.com

Author: Dr. Hanna Köpcke
Graphics: Swantje Jung

Jacobstraße 5
04105 Leipzig