
# Why is product matching hard?

**4th Aug, 2015**

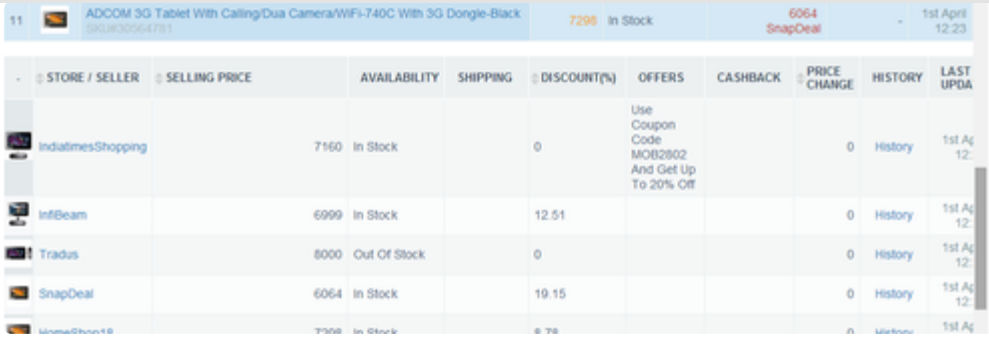
BY DATAWEAVE

Product Matching is a combination of algorithmic and manual techniques to recognize and match identical products from different sources. Product matching is at the core of competitive intelligence for retail. A competitive intelligence product is most useful when it can accurately match products of a wide range of categories in a timely manner, and at scale.

Shown below is PriceWeave's Products Tracking Interface, one of the features where product matching is in action. The Products Tracking Interface lets a brand or a retailer track their products and monitor prices, availability offers, discounts, variants, and SLAs on a daily (or a more frequent) basis.

| # | . | PRODUCT | YOUR PRICE | AVAILABILITY | LOWEST PRICE / STORE | MAX PRICE CHANGE(%) / STORE | LAST UPDATE | FAVORITE |
|---|---|---|---|---|---|---|---|---|
| 1 | | 65w Original AC Adaptor / Charger For Various Dell Laptops SKU#303306144 | 850 | In Stock | 419 Amazon | InfiBeam | 1st April 12:23 | |
| 2 | | 8 Bottle Pyramid Wine Rack SKU#30851937 | 1590 | In Stock | 1427 IndiatimesShopping | - | 1st April 12:23 | |
| 3 | | Buy 98 Degree North Men's Pullover-Dark Green SKU#31156983 | 895 | Out Of Stock | 0 | - | 1st April 12:23 | |
| 4 | | Buy 98 Degree North Men's Pullover-Purple Mil SKU#31156969 | 1495 | Out Of Stock | 0 | - | 1st April 12:23 | |
| 5 | | 98 Degrees North 98 Degree North Men's Pullover-Red SKU#31156955 | 1495 | In Stock | 1495 HomeShop18 | - | 1st April 12:23 | |
| 6 | | Aapno Rajasthan Cotton Double Bed Sheet Set - BS12734 SKU#30673817 | 995 | In Stock | 995 HomeShop18 | - | 1st April 12:23 | |
| 7 | | Aapno Rajasthan Cotton Double Bed Sheet Set - BS12739 SKU#30673827 | 995 | In Stock | 995 HomeShop18 | - | 1st April 12:23 | |

**A snapshot of products tracked for a large online mass merchant**




**Expanded view for a product shows the prices related data points from competing stores**

Product Matching helps a retailer or a brand in several ways:

- Tracking competitor prices and stock availability
- Organizing seller listings on a marketplace platform
- Discovering gaps in product catalog
- Filling the missing attributes in product catalog information
- Comparing product life cycles across competitors

Given its criticality, every competitive intelligence product strives hard to make its product matching accurate and comprehensive. It is a hard problem, and one that cannot be complete addressed in an automated fashion. In the rest of this post, we will talk about why product matching is hard.

**Product Matching Guidelines**

Amazon provides a guideline to sellers about how they should write product catalog information in order to achieve a good product matching with respect to their seller listings. These guidelines apply to any retail store or marketplace platform. The trouble is, more often than not these guidelines are not followed, or cannot by retailers because they don't have access to all the product related information. Some of the challenges are:

- Products either don't have a UPC code or it is not available. There are also non-standard products, unbranded products, and private label products.
- There are products with slights variations in technical specifications, but the complete specs are not available.
- Retailers manage a huge catalog of accessories, for instance Electronics Accessories (screen guards, flip covers, fancy USB drives, etc.).
- Apparels and Lifestyle products often have very little by way of unique identifiers. There is no standard nomenclature for colors, material and style.
- Products are often bundled with accessories or other related products. There are no standard ways of doing product bundling.

In the absence of standard ways of representing products, every retailer uses their own internal product IDs, product descriptions, and attribute names.

**Algorithmic Product Matching using "Document Clustering"**

Algorithmic product matching is done using some Machine Learning, typically techniques from Document Clustering. A document is a text document or a web page, or a set of terms that usually occur within a "context". Document clustering is the process of bringing together (forming clusters of) similar documents, and separating our dissimilar ones. There are many ways of defining similarity of documents that we will not delve into in this post. Documents have "features" that act as "identifiers" that help an algorithm cluster them.

A document in our case is a product description — essentially a set of data points or attributes we have extracted from a product page. These attributes include: title, brand, category, price, and other specs. Therefore, these are the attributes that help us cluster together similar products and match products. The quality of clustering — that is how accurate and how complete the clusters are — depends on how good the features are. In our case, most of the times the features are not good, and that is what makes clustering, and in turn product matching, a hard problem.

**Noisy Small Factually Weak (NSFW) Documents**

The documents that we deal with, the product descriptions, are not well formed and so not readily usable for product matching. We at PriceWeave characterize them endearignly as Noisy Weak and Factually Weak (NSFW) documents. Let us see some examples to understand these terms.

**Noisy**

- Spelling errors, non-standard and/or incomplete representations of product features.
- Brands written as "UCB" and "WD" instead of "United Colors of Benetton" and "Western Digital".
- Model no.s might or might not be present. A camera's model number written as one of the following variants: DSC-WX650 vs DSCWX650 vs DSC WX 650 vs WX 650.
- Noisy/meaningless terms might be present ("brand new", "manufacturer's warranty", "with purchase receipt")

**Small**

- Not much description. A product simply written as "Apple iPhone" without any mention of its generation, or other features.
- Not many distinguishable features. Example, "Samsung Galaxy Note vs Samsung Galaxy Note 2", "Apple ipad 3 16 GB wifi+cellular vs Apple ipad mini 16 GB wifi-cellular"

**Factually Weak**

- Products represented with generic and subjective descriptions.
- Colours and their combinations might be represented differently. Examples, “Puma Red Striped Bag”, “Adidas Black/Red/Blue Polo Tshirt”.

In the absence of clean, sufficient, and specific product information, the quality of algorithmic matching suffers. Product matching include many knobs and switches to adjust the weights given to different product attributes. For example, we might include a rule that says, “if two products are identical, then they fall in the same price range.” While such rules work well generally, they vary widely from category to category and across geographies. Further, adding more and more specific rules will start throwing off the algorithms in unexpected ways rendering them less effective.

In this post, we discussed the challenges posed by product matching that make it a hard problem to crack. In the next post, we will discuss how we address these challenges to make PriceWeave’s product matching robust.

*PriceWeave is an all-around Competitive Intelligence product for retailers, brands, and manufacturers. We’re built on top of huge amounts of products data to provide real-time actionable insights. PriceWeave’s offerings include: pricing intelligence, assortment intelligence, gaps in catalogs, and promotion analysis. Please visit PriceWeave to view all our offerings. If you’d like to try us out request for a demo.*

*Originally published at **blog.priceweave.com**.*

***- DataWeave Marketing***

***4th Aug, 2015***

ARTIFICIAL INTELLIGENCE | DATA ENGINEERING | E COMMERCE