AUTO–ID LABS

The not so unique Global Trade Identification Number

Exploring inconsistencies in online product information sources

Stephan Karpischek, Florian Michahelles, Elgar Fleisch

Auto-ID Labs White Paper WP-BIZAPP-057

June 2011

Stephan Karpischek
Researcher and PhD Student
ETH Zürich

Dr. Florian Michahelles
Associate Director of the
Auto-ID Labs Zurich/St. Gallen
Manager of the labs of Prof. Fleisch
at D-MTEC
ETH Zürich

Prof. Dr. Elgar Fleisch
Professor of Information and
Technology Management at
ETH Zürich and
University of St. Gallen (HSG)

Contact:

Stephan Karpischek
ETH Zürich
Phone:+41 44 632 42 22
Fax: +41 44 632 17 40
skarpischek@ethz.ch
www.im.ethz.ch

Business Processes & Applications

Originally presented at the 7th European Workshop on Smart Objects: Systems, Technologies and Applications (RFID-SysTech 2011), 17 - 18 May 2011, Dresden, Germany.

Abstract

The Mobile shopping apps for consumers often identify retail products by Global Trade Identification Number (GTIN) using barcodes or RFID technology and rely on online data sources to provide basic product information. Several different sources exist, and reports show that the available product information in them is often inconsistent with implications for both retail businesses and consumers. In this paper we compare the product information provided by 10 different online sources for more than 82,000 retail products based on GTINs which were scanned by users of mobile barcode scanning apps. We find inconsistent information for nearly half of the products and report our results from analyzing these inconsistencies.

1. Introduction

Mobile shopping applications for smart phones provide product information for consumers and allow for a wide range of services, e.g. price comparison, nutritional information, or sharing of comments and opinions on products. Many of these applications use barcode scanning to identify the Global Trade Identification Number (GTIN) of retail products. Mobile product identification using RFID technology, in particular Near Field Communication (NFC), has been the subject of research for many years and is about to reach consumers with the newest generation of smart phones, e.g. the Google Nexus S.

Several online data sources for product information based on the GTIN as primary key exist. While the GTIN should be unique by definition, reports show that the same basic product information, e.g. the product's name, provided by different sources is often inconsistent. According to a GS1 UK report [3] product supply chain data is inconsistent in over 80% of instances resulting in huge costs for retailers and suppliers. A more recent study [4] indicates that the quality of product information in mobile shopping applications is even worse and a problem for consumers: The study compares product information returned from three different mobile shopping apps for 375 grocery products in the UK with information provided by the brand owner and finds that “only 9% of scans returned the correct product description”.

This paper contributes additional findings on the consistency of retail product information comparing product information from 10 different online sources for more than 82,000 retail products based on products scanned by users of mobile barcode scanning apps.

2. Our Approach

We have developed a server system which aggregates product information for a given GTIN from 10 different sources available online. Some sources provide a web service with an Application Programming Interface (API). Product information which is available for download is queried from a database.

The product information sources used for this study:

- Amazon - The Amazon eCommerce Web Service offers five different services for product information for the US, Canada, UK, Germany, and Japan.
- codecheck.info is a Swiss web service and provides user generated and editorial information about retail products and their ingredients for German speaking countries.
- upcdatabase.com is privately operated and provides user generated product information for over a million products in the U.S.
- openean - The Open EAN Database is a German web based database for product information.
- affili.net is one of Europe's leading affiliate marketing networks offering a product information web service.
- Bestbuy is a U.S. electronics retailer which provides a free web service for its product catalogue.

To ensure that we use valid GTINs of real products for our study, we use GTINs scanned by the users of mobile barcode scanning apps for iPhone and Android smart phones. One of the apps, my2cents, we have implemented ourselves for iPhone and Android. The my2cents app is described in more detail in [1, 2]. In addition we extract GTINs from the server logs of codecheck, another mobile barcode scanning app operated by codecheck.info, an independent Swiss product information provider. Together the users of the two apps scanned 218,722 different products.

3. Results

In the time between 28 February 2010 and 8 March 2011 we queried the information sources with 218,722 different GTINs. 81,782 of them (37.39%) could be mapped to a product name using at least one of the available product information sources. For 42,067 or 51.43% of all mapped GTINs the sources returned exactly one product name. In 2,134 cases exactly the same product name was returned by more than one source. For 39,715 products (48.56%) the sources returned more than one distinct value for the product name, i.e. at least two product information datasets were inconsistent.

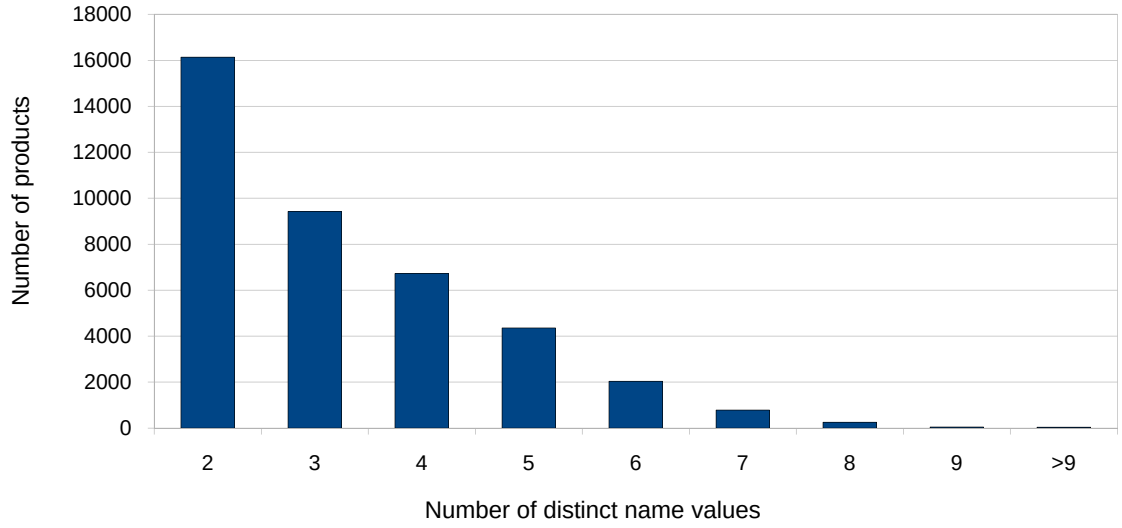

Fig. 1 Distribution of the number of distinct product name values.

Figure 1 shows the distribution of the number of distinct product name values for all products with more than one name. For 16,136 products two distinct name values were returned, for 36 products more than 9 different names were returned. The maximum number of product names returned for the same product was 243. Note that some information sources can return more than one result per GTIN.

Our initial analysis shows that some differences are due to translations of product names to local languages. Many product names differ only in minimal wording, e.g. uppercase versus lowercase letters, or punctuation. We also find problems with character encoding, manual encoding of umlauts, e.g. “ä” spelled “ae” and spelling mistakes.

Many differences are due to different naming conventions, e.g. for books some product names include the author while others don't. Also the order of words in the names is often inconsistent. Some product names include additional information from retailers, either informing about product specifities, e.g. “Osram Tropfenlampe CLAS P CL 60 E27 please note: German product but we supply a UK adapter if necessary”, including category information, e.g. “Kurzlehrbuch Embryologie (eBook)”, or plain advertisment, e.g. “Kaufen Sie keine Billigtinte! Kaufen Sie Qualitätstinte der Fa. TINTENTANKER – Vertrieb durch CMS + Ein Gutschein von 5,00 EUR.”

The products with the highest numbers of different names are mobile phones where the same device with the same GTIN is available in many different product bundles, e.g. the mobile phone bundled with a contract from a mobile network operator. Many online shops offer mobile phones and tend to put a lot of information about the bundle offer in the product name field, e.g. information about the mobile network operator contract details.

In some cases we encounter very different products, with completely different names, images, and product categories, i.e. at least two information sources provide conflicting information for the same GTIN. The inconsistencies between product names are analyzed using different string similarity algorithms to quantify the differences between product names. To exclude differences in punctuation, uppercase vs. lowercase or character encoding problems, the case of all characters is changed to lower case and all characters except lower case letters, numbers and space are stripped, i.e. matching “a-z0-9 “. Also from the 12,299 responses from the Japanese Amazon Web Service 1,084 product names containing Japanese translations with mostly Japanese characters were removed, leaving 38,631 products to analyze.

Three common algorithms for string similarity are used: The Levenshtein ratio is a metric between 0 and 1 based on the edit distance, i.e. how many edit steps are necessary to get from one string to the other, with 1 for similar strings. The Jaro-Winkler ratio gives more weight to a common prefix, i.e. if the first 25% of a string are similar, the ratio is 1. The Levenshtein setratio attempts the best match between any of the words in the first and the second product name. These algorithms were implemented in python using the pylevenshtein code library [5]. A good overview over string similarity metrics is [6].

We also implement a simple matching coefficient algorithm, which takes the nature of the product names to compare into account with the goal of finding different products being described with the same GTIN. From two product names given the algorithm first selects the product name with less words. If both product names have the same wordcount, the shortest string is chosen. Then the matching coefficient for the longer name is calculated, i.e. the ratio of words in common. A ratio of 1 means that all words from the shorter product name also appear in the longer product name, which is a clear indicator that both product names describe the same product. On the other hand a ratio of 0 is a clear indicator that both product names describe different products.

For products with two names four ratios using the described algorithms are calculated. For products with more than two names the means of the ratios of all possible combinations are calculated, together with standard deviations, minimum and maximum values. From the four ratios a combined average ratio is calculated using the mean. Figure 2 shows box plots of the different ratio distributions. The mean of the combined ratio over all 38,631 products is 0.708 with a standard deviation of 0.189. The median is 0.751.

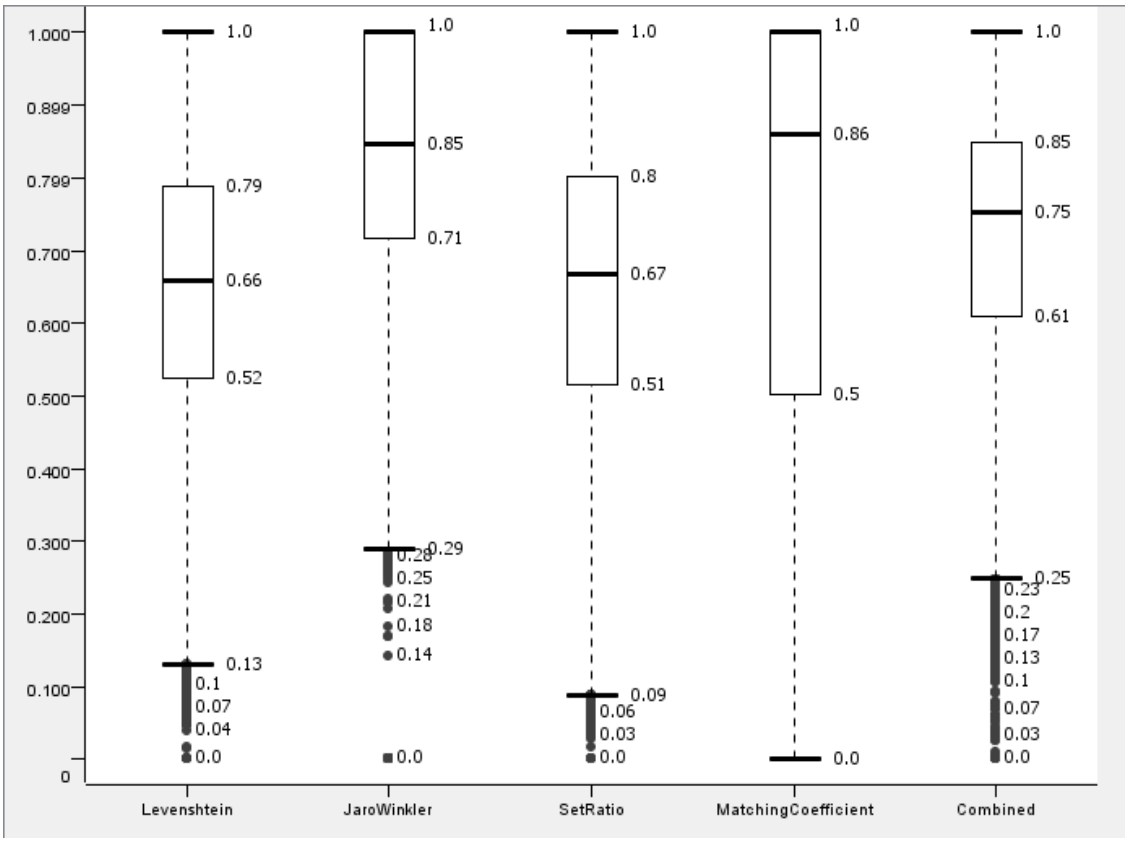

Fig. 2 Box Plots of the similarity and combined ratio distributions.

Of the 38,631 analyzed GTINs, 621 GTINs with a combined ratio of 1.0 have all identical product names except for punctuation or upper case vs. lower case letters. 18,760 GTINs (48.56%) with a combined ratio over 0.75 and below 1.0 have all similar product names with only minor typos, some additional product information in the product name, or smaller parts of the product name translated. For 13,854 products (35.86%) with a combined ratio between 0.5 and 0.75 the product names describe the same product but show some more difference in the product names. For 4,132 GTINs (10.7%) with a combined ratio between 0.25 and 0.5 the product names differ substantially and describe either different products or the same product with a very different product name. For 1,264 GTINs (3.2%) with a combined ratio below 0.25 the returned product names describe different products. Possible reasons for different product names are user generated content and translation of product names. Also, the unauthorized use of GTINs could be a reason.

For retail businesses and consumers trust in product information services is very likely to be low if the product is not correctly recognized. For providers and developers of product information services, the absence of an authoritative product information source is a problem. The integration of a feedback cycle will be necessary to improve the quality of product information and to establish trust with users.

4. Conclusion

We compared the product names returned by different product information sources for a set of GTINs taken from the usage of mobile barcode scanning applications. We found that the returned product names differ in nearly half of the cases. Analyzing the differences in more detail with similarity metrics we found only minor differences for half of the products with differences. Between 3.2% and 13.9% of the GTINs have product names describing different products. A more exact analysis needs improvement of algorithms or could be done by comparing product names manually.

References

[1] Karpischek, S., Michahelles, M., Fleisch, E. (2010): my2cents - a Twitter for products, The 2010 International Workshop on Smartphone Applications and Services (Smartphone 2010), Gwangju, Korea, 9 – 11 December 2010.

[2] Karpischek, S., Michahelles, M. (2010): my2cents – Digitizing consumer opinions and comments about retail products, Internet of Things 2010 Conference (IoT2010), Tokyo, Japan, 29 November - 1 December 2010.

[3] GS1 UK (2009): Data Crunch Report, London, UK, October 2009. Online: http://www.gs1uk.org/resources/help_support/WhitePapers/GS1_UK_Data_Crunch_Report_2009.pdf [11 March 2011].

[4] Coussins, O., Beston, T., Adnan-Ariffin, S., Griffiths, R. and Ross, S. (2011): Mobile-savvy shopper report. GS1 UK, London, UK, January 2011. Online: http://www.gs1uk.org/resources/help_support/WhitePapers/GS1_UK_Mobile-savvy_Shopper_Report_2011.pdf [11 March 2011].

[5] pylevenshtein - A fast implementation of Levenshtein Distance (and others) for Python. http://code.google.com/p/pylevenshtein/ [18 April 2011].

[6] Chapman, S (2011): Similarity Metrics for Information Integration. http://staffwww.dcs.shef.ac.uk/people/S.Chapman/stringmetrics.html [18 April 2011].