Electron Markets
DOI 10.1007/s12525-013-0136-4

GENERAL RESEARCH

Detecting incorrect product names in online sources for product master data

Stephan Karpischek · Florian Michahelles · Elgar Fleisch

Received: 4 December 2012 / Accepted: 20 June 2013

Abstract The global trade item number (GTIN) is traditionally used to identify trade items and look up corresponding information within industrial supply chains. Recently, consumers have also started using GTINs to access additional product information with mobile barcode scanning applications. Providers of these applications use different sources to provide product names for scanned GTINs. In this paper we analyze data from eight publicly available sources for a set of GTINs scanned by users of a mobile barcode scanning application. Our aim is to measure the correctness of product names in online sources and to quantify the problem of product data quality. We use a combination of string matching and supervised learning to estimate the number of incorrect product names. Our results show that approximately 2 % of all product names are incorrect. The applied method is useful for brand owners to monitor the data quality for their products and enables efficient data integration for application providers.

Keywords Correctness · Data quality · GTIN · Product master data · Product names · Quality assessment

JEL classification L15

Responsible editor: Hans-Dieter Zimmermann

S. Karpischek (✉) · F. Michahelles · E. Fleisch
ETH Zürich, WEV G 222.2, Weinbergstrasse 56/58, 8092 Zürich, Switzerland
e-mail: skarpischek@ethz.ch
URL: http://www.im.ethz.ch/

F. Michahelles
e-mail: fmichahelles@ethz.ch
URL: http://www.im.ethz.ch/

E. Fleisch
e-mail: efleisch@ethz.ch
URL: http://www.im.ethz.ch/

Introduction

Research and industry agree that data quality is a critical issue in organizations, and that insufficient data quality can have a substantial negative business impact (Wang and Strong 1996; Ballou et al. 2004; Haug et al. 2011). Recent research on the exchange and quality of product data in the consumer packaged goods industry (CPG) has focussed on intra- and inter-organizational supply chain scenarios (Legner and Schemm 2008; Hüner et al. 2011; Otto et al. 2011a).

One cornerstone of product data for the consumer packaged goods industry is the global trade item number (GTIN), which is specified and distributed by Global Standards One (GS1), a non-commercial organization with member organizations in countries worldwide. GTINs were formerly known as European Article Number (EAN) and Unique Product Code (UPC) and most often come in the form of one-dimensional barcodes printed on product packaging.

By definition GTINs are used to identify consumer goods items “at any point in any supply chain” (GS1 2012). Barcodes are typically scanned with laser scanners. The recognized GTIN is then mapped to predefined product master data such as a product name, description, price, and other attributes like packaging size, or weight. Since its introduction in the 1970s the use of GTINs has made supply chains and the exchange of product master data more efficient. Traditionally, its use has been restricted to company-internal use and industrial supply chains, and typically ended at the supermarket check-out.

Recently, the use of GTINs has become popular among consumers who scan product barcodes with mobile applications to access additional information about products of interest (Brody and Gottsman 1999; Ohbuchi et al. 2004; Adelmann et al. 2006; Reischach et al. 2010). As the use of GTINs is no more limited to company-internal use and supply chains, the quality of corresponding product master data affects brand image and consumer trust.

Published online: 04 August 2013

Previous research on matching product names from two e-Commerce websites found "erroneous UPC identifiers provided by the merchants on some of the product offers" when matching product names (Bilenko et al. 2005). A more recent study by GS1 claims that missing and incorrect product names in mobile B2C applications are a problem not only for users and the application providers but also for brand owners. Wrong or missing data affect the consumers' trust in the app and decrease the willingness to buy the product under consideration (Coussins et al. 2011). Our own experience with the development of a mobile barcode scanning application shows data quality problems with user-generated product data (blinded for review), e. g., incorrect product names for European groceries from Amazon (blinded for review).

There is currently no authoritative source of product master data for consumer goods today which is accessible, complete, and useful for B2C applications. Data pools in the Global Data Synchronization Network (GDSN) are targeted at supply chain requirements, cover only a subset of available products, and data access requires individual contracts with every company (Nakatani et al. 2006; Schemm et al. 2007; Schemm and Legner 2008). Providers of consumer-facing applications need to aggregate data from many sources, with different formats, schema, and quality. Several services offer public application programming interfaces to product data like, e. g., Google or Amazon. However, to our best knowledge no assessment of data quality in these sources has been done yet. Haug and Arlbjørn (2011) note the lack of academic research related to data quality and based on empirical evidence from many companies.

In this paper we assess the quality of product names in publicly available sources for a set of GTINs which were scanned during real-world usage of a mobile shopping application. Our goal is a better understanding of product master data quality online and to raise awareness of brand owners for emerging data quality problems. This is important because of the negative effects missing or incomplete data have on both the consumer experience and on the reputation of brand owners and retailers. To quantify the problem we focus on the product name as one of the most important attributes of a consumer product. Motivated by reports on missing and wrong product names (Coussins et al. 2011) we focus on correctness as the most relevant quality dimensions.

We aim to measure the correctness of product names retrieved from publicly available sources to provide researchers and practitioners with unbiased and reliable numbers on the quality of product names. Our research questions are: How can we efficiently identify incorrect product names? And: How big is the problem of incorrect product names in publicly available sources for Swiss and German consumer packaged goods?

We use a combination of string matching and supervised learning to detect incorrect product names: First we measure the similarity of product names with authoritative names, then we train a classifier on these similarity measures. We measure the performance of the classifier on detecting incorrect product names in independent test sets and use the classifier to estimate the number of incorrect product names.

The rest of this paper is structured as follows: The next section provides an overview of related work. Then we describe the methodology used to answer our research questions. In the following section we present results and evaluate the proposed method to identify incorrect product names. The next section discusses results, limitations, and possible applications of our study. The paper ends with conclusions and an outlook on future work.

Related work

Data quality assessment

Data quality is a multi-dimensional concept and the definition of quality dimensions to conceptualize this has been a key issue in data quality research for many years. Lee et al. (2002) compared academic and practitioners' views on data quality dimensions and consolidate objective and subjective quality dimensions into one model. Batini and Scannapieco (2006) compared different approaches to defining quality dimensions, and Batini et al. (2009) gave an overview of the classifications of data quality dimensions found in two decades of data quality literature. They derived a common set of four dimensions: accuracy, completeness, consistency, and time-related dimensions (Batini et al. 2009).

In this paper the focus is on measuring the quality dimension accuracy. In the context of this paper and to avoid confusion with the notion of accuracy in machine learning, we prefer the term correctness over accuracy and use it synonymously to describe the extent to which data are correct or free from error (Wand and Wang 1996).

Product master data

This paper contributes to research on the quality of product master data. Master data describe features of a company's core entities such as customers, suppliers, or products (Otto et al. 2011b). According to Hüner et al. (2009) "[c]orporate master data is defined as data used in more than one division." According to another more popular definition "master data is typically created once and re-used many times, and does not change too frequently." (Knolmayer and Röthlin 2006).

Product master data is master data specific to a company's products such as product names, images, product descriptions, or ingredients. In data pools product master data are provided by CPG companies for use in a B2B context within industrial supply chains (Nakatani et al. 2006; Schemm and Legner 2008). When the same data are used by consumer-oriented services quality problems, e. g., missing or incorrect product names, emerge which have not been visible before, an effect which has been described by English (2005) as "deficient for downstream processes".

A pilot project conducted by GS1 aims to provide a trusted source of product master data for consumer applications (Anarkat et al. 2012). The project evaluates a technical implementation of such a service but does not address the problem of data integration or data quality.

Product name matching

Matching product names is a common example for research on entity matching and is done by measuring the similarity or distance between product names, i. e., strings. For identifying incorrect product names we are interested in finding non-matching instead of matching product names, however, this is basically the same problem from an opposite point of view, so we can use the same string-based metrics and algorithms which also provide good results for name matching.

Cohen et al. (2003a) compare the performance of different string distance metrics for name matching and implement an open-source software toolkit for matching names. They find a combination of the classical Levenshtein edit distance and the Jaro-Winkler method to perform best (Cohen et al. 2003b).

Bilenko et al. (2005) compare name matching algorithms for matching entity names for several data sets including product names taken from two e-Commerce websites. They note that UPC codes are "golden standard labels for evaluating linkage accuracy" (Bilenko et al. 2005) which at first seems to confirm the widespread assumption that GTINs are globally unique. However, from their experiments with matching product names they also report an unexpected "sharp drop in precision […] due to erroneous UPC identifiers provided by the merchants on some of the product offers" (Bilenko et al. 2005). They also find that different UPC codes for differently colored variants of the same product penalize observed precision values. To our knowledge this is the first time quality problems related to GTINs and product names are mentioned in academic research.

The authors do not further follow this issue or its consequences. With this paper we aim to follow up on their observations and further study the occurrence of "erroneous" GTINs in more detail and on a larger scale.

One major challenge for name matching in general is to find the optimal configuration parameters, e. g., similarity thresholds, to differentiate between matches and non-matches. Machine learning techniques can be used to automate this process, e. g., using supervised learning after training labeled examples for matches and non-matches (Cohen et al. 2003a; Bilenko et al. 2005; Köpcke et al. 2010).

Motivated by these previous research results we base our approach to identify incorrect product names in publicly available sources on a combination of string similarity measures and supervised learning to find the best configuration parameters and similarity thresholds.

Methodology

To identify incorrect product names we use the following process: First we define a set of GTINs and a set of publicly available sources we want to use. Then we collect product names for the selected GTINs from these sources. We also collect authoritative names for the same GTINs and measure the similarity of the collected product names with the authoritative names. A random sample of product names is labeled as correct or incorrect, and the similarity measures and labels are used to train a supervised learning classifier. Finally, the classifier is used to predict incorrect product names.

Data collection

We base our study on product barcode scans from users of a mobile shopping application for iPhone. Using GTINs scanned by consumers ensures that they are really used for real-world products. codecheck.info is an independent Swiss product information platform on which users collect information on products and their ingredients. The organization offers a mobile application which has been installed by more than a million users in Switzerland and Germany (Scandit 2011). With the mobile application users scan product barcodes in order to obtain information regarding product ingredients in food and cosmetics, in particular possibly unhealthy or ingredients and better alternatives.

Our study builds upon the server logs with requests from the codecheck iPhone application. The logs represent the first month of usage after the launch on the iTunes app store. From March to April 2010 2,028,778 products were scanned. For every scan the logs show a request with the corresponding GTIN as a query parameter.

The collected GTINs were used to request product master data from several sources. We selected services which can be accessed for free or at little cost and preferred sources which covered a wide range of products. The first three digits of the GTIN are the country prefix and denote the GS1 country organization with which a GTIN was registered with some special ranges for restricted distribution, coupons, or the publishing industry. Based on the country distribution of the scanned GTINs we focused on services for German speaking markets.

- **Amazon.com** is the world's largest e-commerce retailer and provides a Product Advertising API as part of the

Amazon e-Commerce Web Services. We used SOAP requests and the ItemLookup method with 'ItemType' set to 'EAN', providing the GTIN as value for 'ItemId' and 'ResponseGroup' set to 'Medium'. For every GTIN we queried the API four times with different country parameters 'DE', 'UK', 'FR' and 'US'. From the first returned result we used the attribute 'title' as product name or, if no 'title'-attribute was present, the attribute 'label'.

- **Google** provides a RESTful Search API for Shopping for querying product offers that have been uploaded to Google by merchants. We used the country parameter 'DE' and provided the GTIN as value for the parameters 'q' and 'restrictBy'. From the query results we used the attribute 'title' as product name. Google limits the number of requests to 2,500 per day, the number of results per GTIN is limited to 25 by default.
- **codecheck.info** provides a RESTful API for their product database. We used the GTIN as value for the parameter 'EAN'. The API always returns exactly one product. From the returned result we used the attribute 'name' as product name.
- **affili.net** is one of Europe's leading affiliate marketing networks. They provide a platform to establish affiliate marketing partnerships with a large number of online shops and a web service to search for products of online shops with which partnerships are established. We used SOAP requests and the GTIN as value for the attribute 'query' of the method 'SearchProducts'. From the returned results we used the attribute 'title' as product name. We limited the number of results to the default value of 10.
- **openean.kaufkauf.net** provides a RESTful API for accessing their open EAN/GTIN database of mostly German products. A small fee is needed to register a queryid. We used the GTIN as value for the attribute 'ean'. From the returned result we used the attribute 'name' as product name or, if no 'name'-attribute was present, the attribute 'detailname'.

All queries were done from a server application which was implemented in Ruby on Rails. The GTINs collected from the codecheck logs were processed by the server one per minute over several weeks. This allowed us to keep the number of requests within the limitations of the online services.

We requested company information for all GTINs from GEPIR, a service provided by GS1, with which a GTIN can be mapped to the owning company, i. e., the company which registered the GTIN. We used the company names provided by GEPIR to group scans and GTINs by company and identify popular companies.

SA2 WorldSync operates data pools for master data exchange in consumer goods supply chains and provided us with a static list of product names for the set of GTINs. Product names in the SA2 data pool are typically provided by the brand owner, i. e., the company which registered the GTIN and markets the product. The product name can thus be considered to be authoritative, correct, and "ground truth". To assess the correctness of the product names from the online information sources we compare them with authoritative product names from the SA2 data pool. Figure 1 provides an overview of the data collection process.

Dataset description

The server logs contain 2,028,778 barcode scans, in total we can extract 262,794 different and valid 13-digit GTINs for our dataset. As some sources restrict the number of requests per time we need to further restrict the set of GTINs to be able to finish the study within reasonable time. We choose to ignore 133,145 GTINs which appear only once and obtain a set of 129,649 valid 13-digit GTINs which are the basis for further analysis. GEPIR returned a company name for 91,940 GTINs (70.91 %), the other GTINs could not be resolved. In most cases the owner refused to make company information available. The dataset contains GTINs from 11,450 companies in 109 different countries.

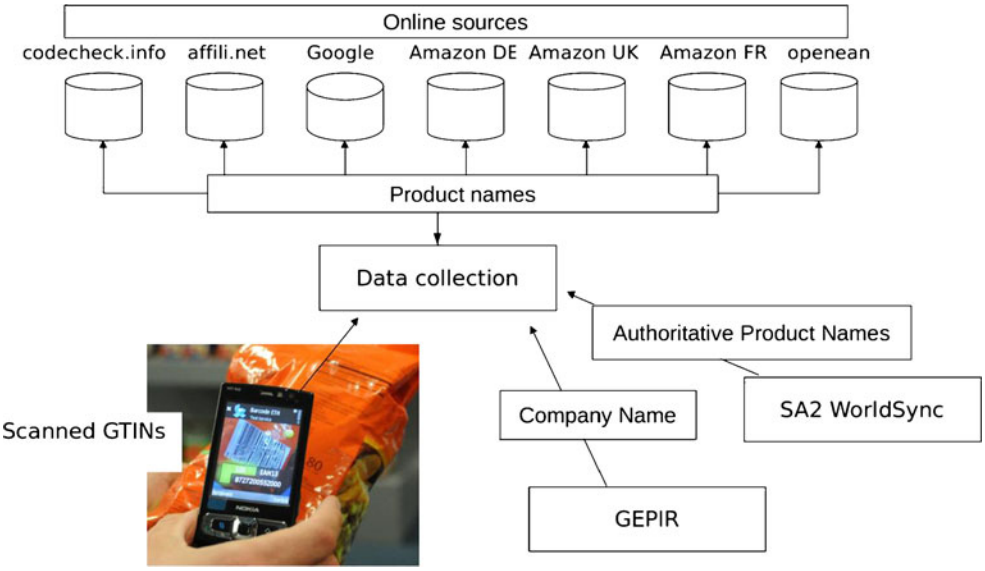
We queried product information from 8 publicly available online sources (codecheck, Google, Amazon DE, UK, US, and FR, openEAN, and affili.net) for the set of 129,649 GTINs. For 13,702 GTINs we received both an authoritative name from the SA2 Worldsync data pool and at least one product name from a publicly available source. In total we collected 140,195 product names for these 13,702 GTINs from publicly available sources which can be compared with authoritative product names.

Correctness

We use the term correctness synonymously to accuracy for the data quality dimension we want to measure (Wand and Wang 1996), i. e., we differentiate between correct and incorrect product names for a given GTIN. We define a product name as correct when the name clearly describes the same product as the authoritative name from the SA2 data pool for the same GTIN. Reversely, a product name is incorrect when it describes a different product.

The need for authoritative data to distinguish correct from incorrect product names limits the data set for further analysis to all GTINs for which both an authoritative name and at least one product name from online sources could be retrieved. In order to identify incorrect product names efficiently, we measure the string similarity of a product name with the authoritative product name from SA2 WorldSync.

Fig. 1 Data collection process



String similarity measures

We use the following string similarity measures (Cohen et al. 2003a; Elmagarmid et al. 2007):

- Equality (Eq): Two strings are equal when they have the same length and every single character is the same. This measure returns a binary value, i. e., either 0 when the names are not the same or 1 when they are.
- Levenshtein distance (Lvsht), also known as edit distance, measures the number of edit operations to change one string to the other (Cohen et al. 2003a; Elmagarmid et al. 2007).
- Jaro-Winkler defines two strings as similar when the beginnings of the strings are similar. We use two ratios 0.25 (JW25) and 0.5 (JW50), i. e., the first 25 % respectively 50 % of the name are compared (Elmagarmid et al. 2007).
- Word coefficient (WCo): The number of words two strings share, divided by the average number of words.
- q-grams: The string is split into chunks of the length q. The number of shared chunks is then divided by the average number of chunks. With padding the beginning of the string is filled with q-1 padding characters. We use 2-grams and 3-grams with and without padding (2gr, 3gr, and 2grp, 3grp) (Elmagarmid et al. 2007).

All measures return a value in the range from 0 to 1, where 1 means perfect match, i. e., the names are considered identical, and 0 means no similarity at all. While character-based string similarity measures (Equality, Levenshtein, Jaro-Winkler) take the order of the words into account, the order does not matter for token-based measures. Before comparing the product names punctuation characters are replaced with space and all upper case letters are changed to lower case.

When there is more than one authoritative name, e. g., because of different languages, the values for all authoritative names are computed and the maximum result of every measure is used. Table 1 shows an example of measures for three different product names compared with the authoritative name in the first row.

Supervised learning

A random sample of 500 GTINs is chosen from the 129,649 GTINs of the dataset. The corresponding 5,248 product names are manually compared with the authoritative product names and labeled as either correct (0) or incorrect (1). 94 of the sample's product names (1.79 %) are labeled as incorrect.

We want to train a supervised learning algorithm on the string similarity measures of the random sample to classify product names as correct or incorrect. As the number of incorrect product names is much smaller than the number of correct product names, the dataset is imbalanced. Incorrect product names are a rare class.

Based on machine learning text books (Bishop 2009; Mitchell 1997), the online machine learning class by Andrew Ng (Ng 2011), and previous comparisons of classification models in machine learning literature (Michie et al. 1994; Caruana and Niculescu-Mizil 2006) we compare the performance of the following 11 different classification models on the given dataset using implementations in Matlab (R2011b):

- Linear discriminant analysis (LDA), the default algorithm for the classify function in Matlab.
- Linear discriminant analysis with empirical prior (LDA Emp. Prior) taking into account the imbalanced distribution of classes.

Table 1 String similarity measures example

Product name	Eq	Lvsht	JW25	JW50	WCo	2gr	2grp	3gr	3grp
Gelierzucker 2PLUS1 500 G									
Einmachzucker und Geliermittel: Südzucker 1×500 g Südzucker Gelierzucker 2+1 - Zucker	0	0.37	0.46	0.46	0.35	0.36	0.34	0.30	0.28
Gelierzucker	0	0.65	1.00	1.00	0.40	0.63	0.62	0.61	0.59
Stereo Mikroskop Objektive Objektivpaar 3×	0	0.24	0.49	0.49	0.00	0.09	0.09	0.00	0.00

- Naive Bayes using Matlab NaiveBayes.fit with default Gaussian distribution.
- Naive Bayes Kernel, with kernel smoothing density estimate.
- Classification tree (Tree), using Matlab Classification Tree.fit with default settings.
- Classification tree (Tree Pruned), pruned to optimal depth using the ClassificationTree.prune method.
- Support vector machine (SVM linear), using Matlab svmStruct with default linear SVM.
- Support vector machine using Matlab svmStruct with a radial basis function (RBF) kernel (SVM RBF).
- Support vector machine using libsvm (LIBSVM RBF) with default RBF kernel.
- Support vector machine using libsvm with linear SVM (−t 0) (LIBSVM linear).
- Logistic regression (LOGREG), based on the logistic regression model presented in Andrew Ng's online machine learning class (Ng 2011).

The sample is partitioned into training set, validation set and test set, to fit the model on the training set, select the best parameters for the model using the validation set, and estimate the performance of the best model using the test set.

We partition the labeled dataset (N=5,248 with 94 members of the rare class) into 10 training and independent test sets using 10-fold stratified cross-validation. Stratified sampling in the training and test set generation ensures that every product name is used for training, validation and testing and that every subset has some members of the rare class. In our case the rare class corresponds to incorrect product names.

On each fold of the training set we perform a grid search (Hsu et al. 2010) to find the best parameters for this particular dataset. The best combination of parameters is used again to train a classifier on the training set and evaluate its performance on the independent test set. This process is repeated for each of the ten folds.

With an imbalanced data set having one rare class, in our case incorrect product names, accuracy (Acc) alone is not a reliable measure for the performance of the classifier (Joshi 2002; He and Garcia 2009). In addition, we use recall and precision and a combined F-score. These measures are based on the number of true positives (TP), false negatives (FN), and false positives (FP) with respect to the rare class and defined as follows (Joshi 2002; He and Garcia 2009):

$$\text{Precision},\quad \text{Pr} := \text{TP}/(\text{TP}+\text{FP}) \tag{1}$$

$$\text{Recall},\quad \text{Rec} := \text{TP}/(\text{TP}+\text{FN}) \tag{2}$$

True positives (TP) in our case are product names that are in fact incorrect and classified as incorrect by the classifier. Accordingly, false positives (FP) are product names that are in fact correct, i.e., they describe the same product as the authoritative product name, but they are classified as incorrect. True negatives (TN) are product names that are correct and also classified as correct, while false negatives (FN) are product names that are incorrect and classified as correct.

Optimizing a classifier is a trade-off between recall and precision. We choose to optimize for recall first as we prefer to find as many incorrect product names as possible, i. e., we want to avoid false negatives. False positives, i. e., to falsely classify product names as incorrect which are in fact correct, are not as big a problem in practice.

As a measure combining precision and recall we report the F2-score, which puts more emphasis on recall than on precision compared to the more balanced F1-score:

$$\begin{aligned}\text{F2}-\text{score},\quad \text{F2} := {} & (5 * \text{Precision} * \text{Recall}) \\ & /(4 * \text{Precision} + \text{Recall})\end{aligned} \tag{3}$$

Following Forman and Scholz (2010) we do not only report average scores over the cross-validation folds, which could be misleading, but sum up the true positives, false positives and false negatives from all folds and calculate the scores on these sums.

In order to evaluate the applicability of this supervised learning approach for consumer goods companies and to illustrate the extent of the problem of incorrect product names for brand owners, three companies are selected based on their popularity, i. e., the sum of products scanned, and the amount of authoritative data, i. e., the number of scans for which authoritative product names are available. We then select all corresponding product names for the GTINs belonging to these three companies from the dataset and again label these product names manually as either correct or incorrect.

Results

Classification model selection

We compare the performance of 11 classification models on the labeled dataset. Table 2 shows the results for 11 different classification models.

Due to the imbalanced dataset the accuracy (Acc) is high for most classifiers. The combined F2 measure is best for logistic regression. We choose logistic regression for further analysis.

Logistic regression performance

The previous subsection compared different classifiers and showed that logistic regression performs best. In order to better estimate the number of incorrect product names, we provide more details on the performance of the classification model in this section. Table 3 shows the results of 10-fold cross-validation for logistic regression in more detail. Each row shows the validation and test measures for one particular fold, using a tenth of the labeled sample set (N=525 for the first eight folds, N=524 for the last two). The validation recall is always a perfect 1 while the recall on the test sets is nearly always perfect with only one false negative occurring in fold number 2.

Precision on the test sets varies from 0.41 to 0.63 with a mean of 0.52, standard deviation 0.07. The 95 % confidence-interval of the mean precision using a *t*-test is 0.47 to 0.57. When we calculate the overall test measures from the sums of true and false positives, and false negatives over all 10 folds recall is 0.99, precision is 0.51, and the F2-score is 0.83. These numbers do not reflect the performance of a single classifier but the performance which can be expected from such a classifier trained and validated on this dataset.

Figure 2 shows learning curves for the classifier, i. e., the F2-scores for training and validation sets plotted for different training set sizes. For training sets larger than 4,000 the training and validation F2-scores converge at around 0.83. This means that adding more training data is unlikely to increase the performance of the classifier.

Predicting correctness

The previous subsection estimated the performance of the logistic regression classifier. In this section we apply the resulting classification model to the full dataset in order to estimate the number of incorrect product names. When run on the full dataset, i. e., all product names for which we had authoritative names and thus similarity measures (N=140,195), the classifier predicts 5,527 positives. Based on the performance measurement in the previous subsections, we estimate that between 46.8 % and 56.9 %, i. e., between 2,588 and 3,145 product names, are incorrect, which corresponds to 1.8–2.2 % of all product names.

In order to further evaluate the relevance of our study and the applicability of the proposed supervised learning approach for CPG companies, additional sets of product names corresponding to all GTINs for three selected companies are used. The companies are selected based on the number of scans and available authoritative product names in the dataset. Company 1 is a global manufacturer of soft drinks, company 2 is a global manufacturer of cosmetics, and company 3 is a global manufacturer of cigarettes. Table 4 shows the datasets and results for these companies.

Discussion

Results

Our first research question was how to efficiently identify incorrect product names.

We proposed a classification model based on a combination of string matching and supervised learning. We

Table 2 Performance comparison of classification models

	TP	TN	FP	FN	Acc	Pr	Rec	F2
LDA	94	4,758	396	0	0.92	0.19	1.00	0.54
LDA Emp. Prior	60	5,114	40	34	0.99	0.60	0.64	0.63
Naive Bayes	87	5,072	82	7	0.98	0.51	0.93	0.80
Naive Bayes Kernel	90	5,018	136	4	0.97	0.40	0.96	0.75
Tree	58	5,123	31	36	0.99	0.65	0.62	0.62
Tree Pruned	63	5,115	39	31	0.99	0.62	0.67	0.66
SVM	77	5,075	79	17	0.98	0.49	0.82	0.72
SVM RBF kernel	70	5,086	68	24	0.98	0.51	0.74	0.68
LIBSVM	72	5,118	36	22	0.99	0.67	0.77	0.74
LIBSVM linear	72	5,106	48	22	0.99	0.60	0.77	0.73
LOGREG	93	5,066	88	1	0.98	0.51	0.99	**0.83**

Table 3 Performance of the logistic regression classifier for 10 folds

Fold	Validation			Test							
	Pr	Rec	F2	TP	TN	FP	FN	Acc	Pr	Rec	F2
1	0.50	1	0.83	9	510	6	0	0.99	0.60	1	0.88
2	0.54	1	0.86	8	509	7	1	0.98	0.53	0.89	0.78
3	0.51	1	0.84	9	505	11	0	0.98	0.45	1	0.80
4	0.51	1	0.84	9	509	7	0	0.99	0.56	1	0.87
5	0.52	1	0.84	9	505	11	0	0.98	0.45	1	0.80
6	0.50	1	0.83	10	506	9	0	0.98	0.53	1	0.85
7	0.51	1	0.84	10	507	8	0	0.98	0.56	1	0.86
8	0.50	1	0.83	10	509	6	0	0.99	0.63	1	0.89
9	0.52	1	0.85	10	503	11	0	0.98	0.48	1	0.82
10	0.53	1	0.85	9	502	13	0	0.98	0.41	1	0.78

measured the performance of the proposed classifier to identify incorrect product names. The high recall scores of 0.99 respectively 1.0 with a precision acceptable for practical purposes show the usefulness of the proposed approach for detecting incorrect product names given a set of GTINs and corresponding product names which are known to be correct.

Our second research question was: “How big is the problem of incorrect product names in publicly available sources for Swiss and German consumer packaged goods?”

Our results show that for about half of the scanned GTINs at least one product name could be found in publicly available sources and that approximately 2 % of the found product names are incorrect. The percentage of incorrect product names varies heavily across companies. For two companies with very popular products we found much higher numbers of incorrect product names, in the case of the cigarette manufacturer nearly two third of the product names are wrong.

Limitations

The results of this study are limited in several aspects: The product scans which determine the set of GTINs are mostly from German and Swiss users for products in the corresponding regional markets. The available authoritative data further limits the results of the similarity measures to mostly German products. The applicability of the results to other regions is subject to further analysis.

Scans are taken only from one app, so the selection of products might not be representative. In addition, scans

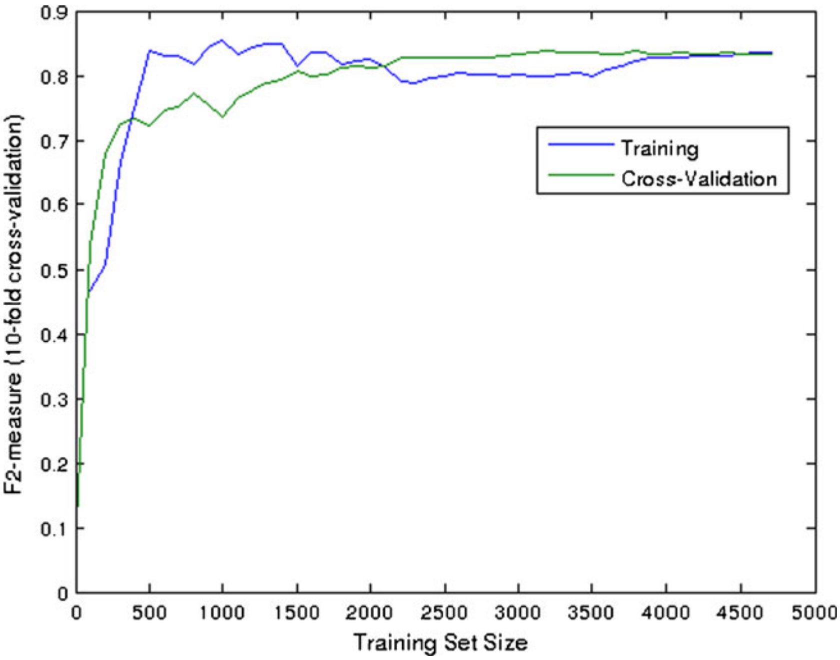


Fig. 2 Learning curves, F2-measures of training and cross-validation sets for different training set sizes

Table 4 Datasets and results for selected companies

Company (Country)	Product names	Incorrect	Recall	Precision	F2-score
Coca Cola (BE)	153	33 (21.57 %)	1	0.97	0.99
Beiersdorf (DE)	4980	11 (0.22 %)	1	0.58	0.87
Philip Morris (DE)	80	53 (66.25 %)	1	0.85	0.97

result from the first month of operation after the launch of the mobile app, so the product selection might be biased towards users trying out the app by scanning the next available barcode versus a representative use of the app for informational needs.

Available product data in publicly available sources is subject to continuous change. Due to the request limitations of the used services the data collection had to happen over a long period of time. In addition, data for the same GTIN has not always been collected at the same time, e. g., for most products the Google API has been queried after the other services. This makes a direct comparison of the performance of the used sources less meaningful.

Checking product names for correctness with the presented approach is limited to available authoritative data. Only products from companies with an existing master data management process have authoritative data in the SA2 data pool. On the one hand these products might have better data quality online, on the other hand popular products might be subject to GTINs being reused by unauthorized parties.

Possible applications

The method we applied can be used in several practical applications:

- Application and service providers who aggregate product master data from different sources can easily detect incorrect product names and implement quality checks on aggregated data to flag or delete incorrect data.
- Service providers who allow user input of product names can check the user input for correctness and warn the user if the entered information is likely to be wrong.
- Consumer goods companies and brand owners can monitor product master data for their GTINs in publicly available sources. This can be done continuously, so unauthorized and potentially harmful use of GTINs can be detected early.
- GS1 can benefit from monitoring public sources for incorrect product master data and adherence of brand owners to GTIN allocation rules.

We think that even results which are considered false positives in the context of our study could be interesting for practical applications as the product names—if not incorrect—are at least inconsistent with the authoritative names and as such still an indicator of consistency problems. We found a large number of false positives to indicate authoritative product names with little or confusing information while product names from publicly available sources for the same GTIN seemed to represent the products correctly.

More authoritative data exist in data pools but are not easily accessible. In particular we had no access to product master data from GDSN data pools. The presented approach to detect incorrect product names depends on the availability of authoritative product master data but only for about a fifth of the scanned GTINs authoritative product names are available. We believe that more authoritative product master data should be available and accessible online. This will not only help application and service providers to deliver better product information: When companies make their product master data easily accessible they will benefit from higher product master data quality in public sources and consumer-facing services and applications.

Conclusions

In this paper we applied a method using supervised learning to identify incorrect product names and evaluated its performance and applicability for consumer goods companies. We measured the correctness of product names from publicly available sources for a set of 13,702 GTINs and found that approximately 2 % of the product names are incorrect when compared with authoritative data.

The presented performance estimations provide a baseline for future improvements of the supervised learning approach, e. g., the use of other classifiers like decision trees or support vector machines might result in higher precision scores. Future work could eliminate the need of authoritative data by comparing, e. g., vector representations of (non-authoritative) product names and detecting outliers. Using semi-supervised or unsupervised learning techniques could further reduce or eliminate the need for labeling. Outliers could then be detected by comparing new product names with a model of brand names and functional names for a given company prefix or company.

Future research could measure the consistency of product names and develop a summarizing measure of consistency for a set of products, e. g., for a brand or company. This would allow to compare the performance of brands and companies regarding product master data quality.

References

Adelmann, R., Langheinrich, M., & Flörkemeier, C. (2006). Toolkit for bar code recognition and resolving on camera phones—Jump starting the internet of things. *Proceedings of Workshop Mobile and Embedded Interactive Systems (MEIS06) at Informatik*. Dresden, Germany.

Anarkat, D., Horwood, J., Green, C., & Bowden, M. (2012). *GS1 trusted source of data pilot report*. Retrieved February 21, 2012, from http://www.gs1.org/docs/b2c/GS1_TSD_Pilot_Report.pdf

Ballou, D., Madnick, S., & Wang, R. (2004). Special section: assuring information quality. *Journal of Management Information Systems, 20*(3), 9–11.

Batini, C., & Scannapieco, M. (2006). *Data quality: Concepts, methodologies and techniques*. Springer.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys, 41*(3), 1–52.

Bilenko, M., Basu, S., & Sahami, M. (2005). Adaptive product normalization: Using online learning for record linkage in comparison shopping. *Fifth IEEE International Conference on Data Mining (ICDM'05),* 58–65.

Bishop, C. M. (2009). *Pattern recognition and machine learning*. Springer.

Brody, A. B., & Gottsman, E. J. (1999). Pocket bargain finder: A handheld device for augmented commerce. *HUC'99 Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing* (pp. 44–51).

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *ICML'06 Proceedings of the 23rd international conference on Machine learning,* (pp. 161–168).

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003a). A comparison of string distance metrics for name-matching tasks. In S. Kambhampati & C. A. Knoblock (Eds.), *Proceedings of the IJCAI2003 Workshop on Information Integration on the Web IIWeb03* (pp. 73–78).

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003b). A comparison of string metrics for matching names and records. *Proceedings of the workshop on Data Cleaning and Object Consolidation at the International Conference on Knowledge Discovery and Data Mining (KDD)* (Vol. 3, pp. 73–78).

Coussins, O., Beston, T., Adnan-Ariffin, S., Griffiths, R., & Rossi, S. (2011). *Mobile-savvy shopper report*. Retrieved February 21, 2012, from http://www.gs1uk.org/resources/help_support/WhitePapers/GS1_UK_Mobile-Savvy_Shopper_Report_2011.pdf

Elmagarmid, A., Ipeirotis, P., & Verykios, V. (2007). Duplicate record detection: a survey. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 1–16.

English, L. P. (2005). To a High IQ! Defining information quality: More than meets the eye. Retrieved February 21, 2012, from http://iaidq.org/publications/doc2/english-2005-04.shtml

Forman, G., & Scholz, M. (2010). Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement. *ACM SIGKDD Explorations Newsletter, 12*(1), 49.

GS1. (2012). GS1 general specifications, version 12. GS1.

Haug, A., & Arlbjørn, J. S. (2011). Barriers to master data quality. *Journal of Enterprise Information Management, 24*(3), 288–303.

Haug, A., Zachariassen, F., & van Liempd, D. (2011). The costs of poor data quality. *Journal of Industrial Engineering and Management, 4*(2), 168–193.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Hsu, C.-w., Chang, C.-c., & Lin, C.-j. (2010). *A practical guide to support vector classification* (pp. 1–16).

Hüner, K. M., Ofner, M., & Otto, B. (2009). Towards a maturity model for corporate data quality management. *Proceedings of the 2009 ACM symposium on Applied Computing SAC 09.*

Hüner, K. M., Schierning, A., Otto, B., & Österle, H. (2011). Product data quality in supply chains: the case of Beiersdorf. *Electronic Markets, 21*(2), 141–154.

Joshi, M. V. (2002). On evaluating performance of classifiers for rare classes. *IEEE International Conference on Data Mining* (pp. 641–644).

Knolmayer, G. F., & Röthlin, M. (2006). Quality of material master data and its effect on the usefulness of distributed ERP systems. *Advances in Conceptual Modeling-Theory and Practice, 362–371.*

Köpcke, H., Thor, A., & Rahm, E. (2010). Learning-based approaches for matching web data entities. *IEEE Internet Computing, 14*, 23–31.

Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. *Information Management, 40*(2), 133–146.

Legner, C., & Schemm, J. W. (2008). Toward the inter-organizational product information supply chain—evidence from the retail and consumer goods industries. *Journal of the Association for Information Systems, 9*(4), 119–150.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). Machine learning, neural and statistical classification. Ellis Horwood.

Mitchell, T. M. (1997). Machine learning. Mcgraw-Hill International Editions.

Nakatani, K., Chuang, T.-T., & Zhou, D. (2006). Data synchronization technology: standards, business values and implications. *Communications of the Association for Information Systems, 17*(1), 2–60.

Ng, A. (2011). Machine learning class. Retrieved 15 December, 2011, from http://ml-class.org

Ohbuchi, E., Hanaizumi, H., & Hock, L. A. (2004). Barcode readers using the camera device in mobile phones. *2004 International Conference on Cyberworlds*, 260–265.

Otto, B., Lee, Y. W., & Caballero, I. (2011a). Information and data quality in business networking: a key concept for enterprises in its early stages of development. *Electronic Markets, 21*(2), 83–97.

Otto, B., Hüner, K. M., & Österle, H. (2011b). Toward a functional reference model for master data quality management. *Information Systems and e-Business Management*, 1–31.

Reischach, F., Karpischek, S., Adelmann, R., & Michahelles, F. (2010). Evaluation of 1D barcode scanning on mobile phones. *Internet of Things 2010 Conference (IoT2010).*

Scandit. (2011). New Codecheck.info Android app now powered by Scandit. Retrieved February 21, 2012, from http://www.scandit.com/2011/10/07/new-codecheck-info-android-app-now-powered-by-scandit/

Schemm, J. W., & Legner, C. (2008). The role and emerging landscape of data pools in the retail and consumer goods industries. *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)* (pp. 320–320).

Schemm, J. W., Legner, C., & Otto, B. (2007). Global data synchronization—Current status and future trends. Institute of Information Management, University of St. Gallen.

Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM, 39*(11), 86–95.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: what data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.