Measuring and Mitigating Product Data Inaccuracy in Online Retailing

Runhua Xu[1] and Alexander Ilic[2]

[1]ETH Zurich, Zurich, Switzerland
[2]University of St.Gallen, St.Gallen, Switzerland
rxu@ethz.ch,alexander.ilic@unisg.ch

Abstract. Nowadays, consumers rely more on rich and accurate online product information to make purchasing decisions. As one of the first studies, we quantify how accurate online product data is today and evaluate existing approaches of mitigating inaccuracy. The result shows that the accuracy varies a lot across different sites and can be as low as 20%. However, when aggregating product information across different Web pages, the accuracy can be improved on average by 11.3%. Based on the analysis, we propose an attribute-based authentication approach based on Semantic Web to further mitigate online data inaccuracy.

Keywords: Data Quality, Data Accuracy, Product Data, Linked Data

1 Introduction

Product information on the Web is gaining more importance to help consumers choose the right products from an overwhelming number of offerings. Thanks to the widespread availability of Smartphones, this is not only a phenomenon for online retailing but also in the rapidly growing category of online-influenced offline sales. Consumers quickly want to understand the price and feature differences of similar products. In the EU, there is also a regulatory component that forces manufacturers and retailers to publish rich information for all food products sold online from 2014 onwards [5].

However, there is currently no standard for product identification on the Web and also the standards for describing product attributes in a machine-readable form such as schema.org are still in their infancy. Product data is mainly published by merchants in a human readable form and is often inaccurate [3], which leads to consumer confusion and lose of sales [11]. While data quality has been intensively studied within organizations and also some studies have provided approaches for analyzing and consolidating data online, there is little insight into the quantitative dimension of the online product data quality. With fast development of Semantic Web and linked data technologies, we see the potential to leverage such technologies to mitigate online data inaccuracy.

Our work comprises three main contributions. First, we quantify product data inaccuracy on the Web and describe the extent of the problem by analyzing data of e-Commerce websites. In addition, we review and compare different

mitigation approaches to improve the data accuarcy by aggregating product information across the Web. Furthermore, we propose a new approach based on Semantic Web, linked data and digital signature to mitigate the data inaccuracy.

2 Related Work

2.1 Data Quality Dimensions

Data quality (DQ) is commonly regarded as a multi-dimensional concept and its definition differs depends on the context. Wang and Strong [12] categorized DQ into 16 dimensions like accuracy, completeness, and security. However in a quality model developed by Bovee et al. [2], different essential attributes, namely accessibility, interpretability, relevance and integrity, were used to evaluate DQ. Although researchers evaluated DQ from different aspects, they used some common dimensions. Knight and Burn [9] and Barnes and Vidgen [1] compared different DQ dimensions and found that data accuracy was the most important and frequently used dimension. Thus, data accuracy was used to evaluate data quality in the study.

2.2 Data quality on the Web

Previous research on online data quality assessment mainly focuses on evaluating the quality of Websites. Eppler and Muenenmayer [4] presented a methodology to measure DQ on the Web by using Web mining tools and site analyser. Such tools helped to quantitatively evaluate quality criteria like accessibility, consistency and speed in a Website. But data accuracy can only be measured qualitatively based on user surveys. Barnes and Vidgen [1] developed a framework to assess quality of Websites. Based on 376 online questionnaires, they evaluated and compared the quality of three largest online bookshops from aspects like usability, information and trust. Different from survey-based DQ assessment, Frber and Hepp [6] developed a framework to evaluate DQ quantitatively in the Semantic Web context. They calculated data quality scores according to the four chosen dimensions: accuracy, completeness, timeliness and uniqueness. The framework was applied on 1.3 million semantic triples of BestBuy. However, more than 90% of the triples are related to geographic data. The remaining triples are email addresses, phone numbers and opening hours instead of product data.

2.3 Mitigation Strategies to Improve Data Quality

Consumers get confused when they find multiple values on the Web for one product attribute. In previous research, mitigation strategies have been generated by aggregating data from multiple sources. In a framework developed by Mendes et al. [10], different approaches were used to pick up the correct value for each attribute. These approaches were 1) taking average, maximum or minimum values, 2) taking first, last or random values, 3) taking most frequently used values, and

4) taking values from specific URLs. Based on these approaches, the framework decided which values to keep, discard or transform when multiple values exist for a single attribute.

As people know the DQ problem on the Web is common and Internet data is never perfect [3], to the best of our knowledge, no research has quantitatively measured online product data accuracy and has shown how big the problem is. Furthermore, current mitigation approaches can help to choose the most possible correct value when data conflict happens. But no research has evaluated and compared their effectiveness on improving data accuracy. Consequently, our research tries to address these two research gaps.

3 Research Method

To analyze online product data accuracy, 15 products on major e-Commerce Websites in Switzerland were selected. Due to the lack of adoption of a Semantic Web standard for product data, all attributes have to be extracted manually. To reduce the number of mistakes, we verified the data by using two people to encode it. Based on discussions with industry experts, electronics and high value goods from category Smartphone, laptop, digital camera, coffee machine and printer were selected as target products because 1) they are of high interest to consumers, 2) they have high relevancy for rich product data, and 3) most of their official specs are available online to serve as ground truth values. In each category, the top three most popular products on a countrywide well-known product comparison Website were chosen. Product data that came from popular online retailers that sold the product was gathered. In the study, data was defined as accurate when an attribute value is exactly the same as the one in the official specs. It was defined as similar if an attribute value rounds the official one. Otherwise, it was regarded as wrong. For instance, if the screen resolution of a digital camera was 1.2 MP on its official specs, then 1.0 MP would be regarded as similar while 2.0 MP would be considered as wrong. Based on the above definitions, data accuracy was measured and the results were compared.

Furthermore, mitigation approaches introduced in Section 2.3 were applied on the dataset. However, not all of them are suitable in online retailing environment. First, the timeliness of product data on the Web is difficult to be identified. Second, based on our definition, taking average value actually moves data from accurate to similar. For non-numerical attributes, it is impossible to average out data with different values. Accordingly, four mitigation approaches were selected: taking maximum values (MAX), taking minimum values (MIN), taking most frequently used values (FREQ) and taking values from specific URL (SURL). For non-numerical attribute values, MAX and MIN algorithms would take values according to their string lengths or alphabetical orders depends on the context. One of the most well-known e-Commerce shops, digitec.ch, was selected as the specific Website in the SURL approach. Based on the result, we analyzed possible causes for online product data inaccuarcy and proposed an attribute-based authentication approach to mitigate the inaccuracy.

4 Result Analysis

4.1 Data Accuracy

We collected 3000 attribute values on 173 product Webpages from 66 distinct online retailers in November 2013. Attribute values on each Webpage were compared with the ground truth values and then labelled as accurate or inaccurate. In this stage, both similar and wrong data were regarded as inaccurate. The accuracy of each product Webpage was calculated by dividing the number of accurate values by the number of available attributes on that Webpage.

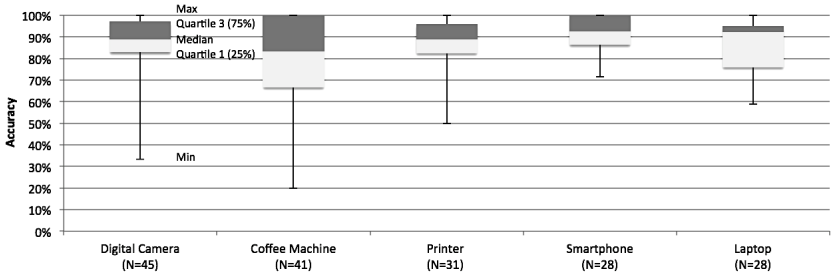

Fig. 1. Data accuracy of seleted products grouped by categories.

Fig. 1 shows the overall data accuracy for each category. Smartphone has the highest data accuracy with a median at 93%. Laptop has a similar median accuracy (92%) but its first quartile accuracy is only 75%. Digital camera and printer have almost the same accuracy except that the minimum accuracy of digital camera is only 33%. Coffee machine has the worst accuracy: It has the lowest minimum accuracy (20%), first quartile accuracy (67%) and median accuracy (83%). Overall, 88.9% of all the attribute values were accurate.

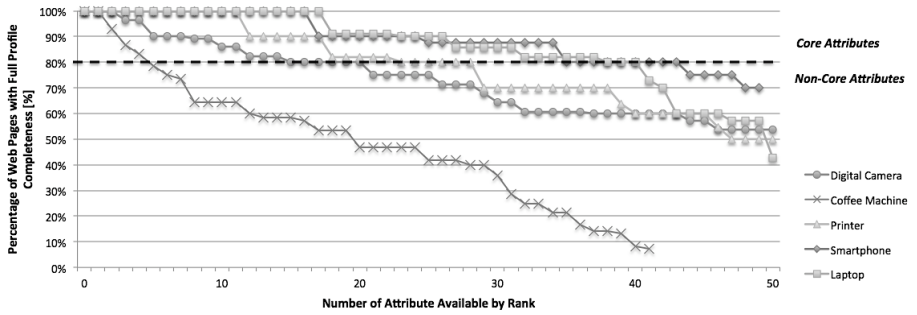

Fig. 2. Ranking attributes based on the percentage of Webpages that publish them.

Data inaccuracy existed in different attributes has various levels of impact on manufacturers, retailers and consumers. Therefore, attributes were further divided into core and non-core attributes based on popularity. A more sophisticated approach of categorization is required in further research.

As shown in Fig. 2, attributes of all products in each category were ranked according to the percentage of Webpages that publish them. Taking coffee machines for example, the first attribute has 100% on the y-axis, which means all the Webpages involved in the study that sell the product publish values for this attribute. As most Webpages publish values for the attribute, we assume it to be a core attribute. For the last attribute, conversely, only 7% of all the Websites that sell the product publish values for it, which makes it to be a non-core attribute. According to Pareto principle, the boundary between core and non-core attributes is set to 80% in the study.

Table 1. Data accuracy of core attribute values gathered from 173 Webpages.

Product Category	*Core Values*	*Accurate Values*		*Similar Values*		*Wrong Values*	
	M(SD)	*M(SD)*	*Pct.*	*M(SD)*	*Pct.*	*M(SD)*	*Pct.*
Digital Camera	102.3 (93.2)	85.7 (74.0)	83.7%	11.7(11.5)	11.4%	5.0 (8.7)	4.9%
Coffee Machine	21.0 (6.6)	18.7 (5.8)	88.9%	0.7 (0.6)	3.2%	1.7 (2.1)	7.9%
Printer	91.0 (20.0)	75.3 (8.3)	82.7%	9.4 (6.8)	10.3%	6.3 (5.5)	7.0%
Smartphone	125.3 (43.0)	116.0 (43.3)	92.6%	8.7 (3.1)	6.9%	0.6 (0.6)	0.5%
Laptop	119.3 (40.0)	100.3 (24.9)	84.1%	7.3 (3.2)	6.1%	11.7 (15.0)	9.8%
All Products	91.8 (57.4)	79.2 (44.4)	86.3%	7.5 (6.6)	8.2%	5.1 (8.0)	5.5%

After identifying core attributes, data accuracy was measured on core attributes as shown in Table 1. Take the first row for example, each digital camera selected in the study has on average 102.3 attribute values gathered from different Webpages, and 85.7 or 83.7% of these values are accurate, 11.4% are similar and 4.9% are wrong. Smartphone has the highest accuracy with 92.6% while printer has the lowest accuracy with 82.7%. Laptop has the highest percentage of wrong values (9.8%). Overall, 86.3%, 8.2% and 5.5% of all the core attribute values are accurate, similar and wrong, respectively.

4.2 Data Inaccuracy Mitigation Approaches

One strategy to mitigate data inaccuracy is to aggregate attribute values of a product from multiple sources. As described in Section 3, four mitigation approaches were applied on the dataset. Table 2 shows the result. Different from the previous analysis, the accuracy measured here is on attributes instead of attribute values. A product attribute has inaccurate attribute values on some Webpages may still be judged as accurate if a mitigation approach finally selects an accurate value for the attribute. The second column in the table shows the worst possible accuracy. For instance, Smartphone has a 75.7% worst possible accuracy, which means 75.7% of all the Smartphone attributes have consistent and accurate values; only the remaining 24.3% attributes have multiple values on different Websites. In the worst case, a mitigation approach chooses inaccurate values for all the attributes with multiple values, which makes the data accuracy to be 75.7%. However, if MAX is used, the accuracy will be improved to 82.4%. Compared to the worst case, the four mitigation approaches can on

average improve the accuracy by 11.3%. FREQ performs best but it still leaves around 9% inaccurate attributes.

Table 2. Improvements of algorithms on mitigating data inaccuracy.

Product Category	Worst Possible Data Accuracy	Improved Data Accuracy with Mitigation Approaches			
		MAX	MIN	FREQ	SURL
Digital Camera	88.2%	93.3%	88.9%	97.0%	89.1%
Coffee Machine	61.9%	73.8%	66.7%	80.9%	77.8%
Printer	80.5%	92.2%	84.4%	87.0%	91.4%
Smartphone	75.7%	82.4%	86.5%	93.2%	91.3%
Laptop	71.9%	82.5%	87.7%	85.9%	89.1%
Total Products	79.0%	87.3%	84.9%	90.9%	88.6%

4.3 Potential Causes for Data Inaccuracy

The result provides insights into potential causes for data inaccuracy, thus, we suspect four main types of errors. 1) Data input errors: retailers publish product data imprecisely on the Web due to manual mistakes. 2) Data lack errors: When official information is not available, some retailers use a default value for all the products in the same category. 3) Data update errors: In some cases, manufacturers update product information to correct previous errors. But retailers do not replace old values with the updated ones. 4) Data copy errors: Instead of publishing product data themselves, some retailers just copy the HTML source file from other retailers that sell the same product and publish it on their Websites. Consequently, if an error exists on a Webpage, it will be amplified by other retailers who copy the page.

5 New Approach to Mitigate Data Inaccuracy

We developed an attribute-based authentication approach to mitigate data inaccuracy through automatically differentiating manufacturer data and open data on the Web. Based on the Public-Key Cryptography, two Web tools were developed: The first one was used by manufacturers to generate a digital signature for each authentic attribute value. The second one was used by anyone on the Web to verify these signatures. Together with attribute values, online retailers can publish the corresponding signatures they get from manufacturers as meta-data in products' HTML files, as shown in the example code below.

```
<div>Width:
 <meta itemprop="width" href="http://schema.org/QuantitativeValue"
 content="13.55cm | 311b09f8ad88aaa83d549043e037a02a">13.55cm
</div>
```

When an application crawls Webpages to generate a consolidated view for a product, it can retrieve both product attributes and signatures (the string in

the content tag in the above code) from a Webpage and call the second Web tool API to verify whether the attribute values are authentic. Authentic values will always overwrite inauthentic ones when multiple values exist for a single attribute. When manufacturers update product information, the first Web tool will make sure that old signatures become invalid and new ones come into force. Thus, only the latest attribute values will be verified as authentic by the second tool. Furthermore, retailers who copy online product data will also copy the signatures since the signatures are integrated in HTML files. Therefore, the new approach can contribute to reduce the data input, update and copy errors.

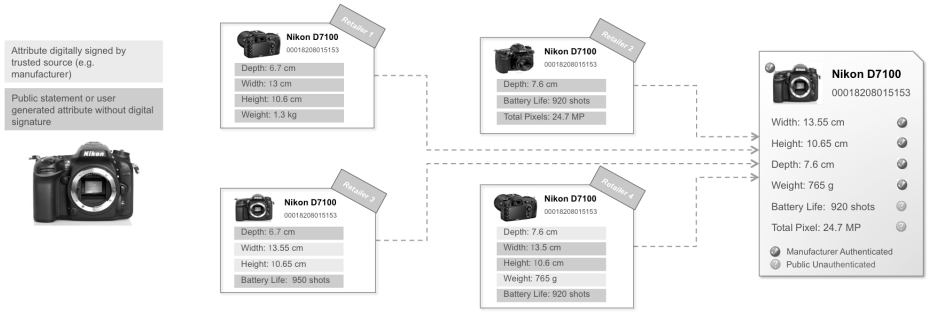

Fig. 3. An attribute-based authentication approach to mitigate data inaccuracy.

Fig. 3 demonstrates how the approach works. In the demo, Retailer3 and Retailer4 publish accurate product data with digital signatures on their product Webpages, whereas Retailer1 and Retailer2 publish data without signatures or with invalid signatures. When the demo application crawls the four Webpages, it verifies each attribute signature and decides which value to trust when conflict occurs. The consolidated result is shown in the rightmost picture, where manufacturer authentic data with digital signature is always selected when data conflict happens, thereby improving data accuracy. For attribute values without signatures, FREQ is used to select a proper value.

6 Conclusion, Limitation and Further Work

The study revealed a significant problem of product data inaccuracy on the Web. On average over 11% all attribute values and 13% core attribute values were inaccurate. We showed that the data accuracy can be increased by crawling the web and consolidating the data with several approaches. We investigated four popular approaches and showed that they were able to increase the accuracy on average by 11.3%. However, the most frequently appeared attribute value is not always accurate. Thus, we suspect that the dynamics of copying information from one page to another amplifies in several cases the problems. Based on the findings, we proposed an attribute-based authentication approach to mitigate online product data inaccuracy.

Our study is not without limitation and provides several opportunities for further research. First, product data was extracted manually in the study, which results in the relative small number of samples. It might be interesting to leverage Semantic Web and linked data technologies to automatically extract and compare online data of a larger number of products from different categories. Second, we categorized core and non-core attributes only based on their popularity. In future study, we plan to conduct a survey study to better understand what attributes are more important to consumers. Third, we suspected four potential causes for online data inaccuracy. We acknowledge that small sample size might lead to biased understanding, therefore, 20 expert interviews will be conducted on manufacturers, distributors and retailers to validate the suspected causes. Furthermore, we developed a prototype to demonstrate our attribute-based authentication approach. In the next step, we will collaborate with online retailers and manufacturers to implement the approach in a field trial and then evaluate to what extent it can improve data accuracy.

References

1. Barnes, S., Vidgen, R.: An Integrative Approach to the Assessment of E-commerce Quality. Journal of Electronic Commerce Research. 3(3) (2002)
2. Bovee, M., Srivastava, R.P., Mak, B.: A Conceptual Framework and Belief-function Approach to Assessing Overall Information Quality. Int. J. Intell. Syst, 18, 51–74 (2003)
3. Cho, V.: Data Quality on the Internet. Services and Business Computing Solutions with XML: Application for Quality Management and Best Processes. 171–176. Information Science Reference, New York (2009)
4. Eppler, M., Muenzenmayer, P.: Measuring Information Quality in the Web Context: A Survey of State-of-the-Art Instruments and an Application Methodology. In: 7th International Conference on Information Quality, PP. 187–196. Cambridge (2002)
5. European Union.: Regulation (EU) No 1169/2011 of the European parliament and of the council of 25 October 2011. Official Journal of the European Union (2011)
6. Frber, C., Hepp, M.: A Semantic Web Information Quality Assessment Framework. In: 25th European Conference on Information System, pp.76, Helsinki (2011)
7. Klein, B.D. (2001). User Perceptions of Data Quality: Internet and Traditional Text Sources. The Journal of Computer Information System, 41, 9–18 (2001)
8. Knap, T., Michelfeit, J., Necasky, M.: Linked Open Data Aggregation: Conflict Resolution and Aggregate Quality. Computer Software and Applications Conference Workshops (2012)
9. Knight, S., Burn J.: Developing a Framework for Assessing Information Quality on the World Wide Web. Informing Science Journal, 8, 159–172 (2005)
10. Mendes, P.N., Mhleisen, H., Bizer, C.: Sieve: Linked Data Quality Assessment and Fusion. In: Joint EDBT/ICDT Workshops, 116-123, ACM New York, New York (2012)
11. Redman, T.C.: The Impact of Poor Data Quality on the Typical Enterprise. Magazine Communications of the ACM, 41, 79–82 (1998)
12. Wang, R., Strong, D.: Beyond Accuracy: What Data Quality Means to Data Consumer. Journal Management Information System, 12, 5–33 (1996)