Lessons Learned and Research Agenda for Big Data Integration of Product Specifications (Discussion Paper)

Luciano Barbosa[1], Valter Crescenzi[2], Xin Luna Dong[3], Paolo Merialdo[2], Federico Piai[2], Disheng Qiu[4], Yanyan Shen[5], and Divesh Srivastava[6]

[1] Universidade Federal de Pernambuco luciano@cin.ufpe.br
[2] Roma Tre University {name.surname}@inf.uniroma3.it
[3] Amazon lunadong@amazon.com
[4] Wanderio disheng@wanderio.com
[5] Shanghai Jiao Tong University shenyy@sjtu.edu.cn
[6] AT&T Labs – Research divesh@research.att.com

Abstract. The product domain represents a challenging scenario for developing and evaluating big data integration solutions: the number of sources providing product specifications is very large, and ever increasing over time. The volume of available data is impressive, and these data keep changing very frequently. In this paper, we present ongoing efforts, challenges and our research agenda to address big data integration for product specifications.

1 Introduction

This paper describes our recent experiences dealing with the issue of extracting and integrating data from product specification pages on the Web, the challenges that we encountered and opportunities we leveraged. We provide an overview on our ongoing research activities on this problem, we discuss the lessons that we have learned from our experience, and illustrate open problems and future directions.

Integrating data from the product domain represents a challenging issue, presenting most of the difficulties associated with big data solutions at the web scale. Thousands of websites contain an impressive number of product pages. Each page provides information about a single product: usually general product specifications (typically technical and physical features), and some source-specific data, such as, price and customers' reviews.

Integrating product data might enable many valuable applications, such as data driven market analysis, question answering, and price comparison. However, data integration at the web-scale raises intriguing *challenges* due to the volume, variety, velocity, and veracity of data [3].

SEBD 2018, June 24-27, 2018, Castellaneta Marina, Italy.

– In this context, the *volume* of data refers not only to the large number of products but, even more importantly, to the number of sources. In order to achieve coverage and diversity, we need to process a very large number of sources across the entire web, not just a small number of pre-selected sources.
– The *variety* of data is directly influenced by the number of sources and arises at many different levels, affecting every task of our pipeline. At the product category level, websites, especially the head ones, organize products according to such a vast plethora of categorization strategies that it gets very difficult, if not impossible, to reconcile them into a unified taxonomy. At the product description level, heterogeneities are related to attributes and values, which are published according to different granularities (e.g., physical dimensions in one field vs three separate fields for width, length, height), formats (e.g., centimeters vs inches) and representations (e.g., the color of a product as a feature vs distinct products for different colors).
– The *Velocity* of data concerns the rate of appearance and disappearance of pages in sources as well as the rate of appearance and disappearance of web sources. Also, while some attributes (such as technical and physical features) are quite stable over time, the contents of the individual pages can change daily, for example for prices and reviews.
– The *Veracity* of data deals with honest mistakes that can occur in web pages, but also with deceits, that is, deliberate attempts to confuse or cheat (e.g., providing imprecise or erroneous product characteristics).

2 End-to-End Data Integration Pipeline

To address the above challenges, we are working on an end-to-end data integration solution for product specification pages. For simplicity of presentation, we describe our approach as a linear pipeline, as depicted in Figure 1, where tasks are performed in sequence and independent of one another. However, there might be feedback loops between the tasks, as intermediate results can indeed influence the performance and the behavior of the preceding tasks and of the end to end solution.

In our vision, the information need is expressed by an input set of sample pages. We observe that products are typically organized in *categories*, and hence we expect that the sample pages refer to products from the categories of interest. Our approach is inspired by the Open Information Extraction [4] paradigm: the schema for the target data is not specified in advance, and the categories of the target products do not refer to a predefined product taxonomy, but they are rather inferred from data in the product pages of the input sample and in the product pages that are gathered from the Web along a progressive and iterative process.

Every task produces output to feed the successive task in the pipeline, but intermediate results could find other compelling application scenarios as well. To this end, we advocate that the pipeline must include an empirical evaluation benchmark for every task.

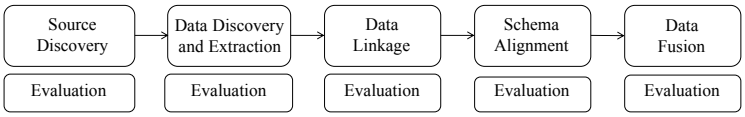

Fig. 1. Our end-to-end big data integration pipeline for product specifications.

Source discovery aims at efficiently finding and crawling product websites in order to gather pages that refer to the products of interest. One might believe that discovering product websites is a minor task, as a relatively small number of *head sources* can offer enough data for most of the products. For example, amazon.com already provides data about an impressive number of products. However, valuable information is actually published by an enormous number of *tail sources* [1, 2], i.e., sources that each provide a small number of product entities. These tail sources are important because they improve coverage. They can often offer *tail entities*, i.e., products that are present in a small number of sources, as well as *tail attributes*, i.e., product properties that are present in a small number of entities. Also, tail sources ofter refer to *tail categories*, i.e., small niche categories of products. Finally, *tail sources* contribute to information diversity, as they provide values that depend on the local source, such as, product reviews and price.

Data discovery and extraction has the objective of processing the pages harvested in the previous task in order to locate and extract product attribute names and their values. As we mentioned above, we do not rely on a predefined schema, but rather extract attributes bottom-up, with the goal of discovering not just *head attributes*, but also *tail attributes* that cannot always be described in advance.

Data linkage seeks to cluster pages from different sources that refer to the same products. It is worth observing that in the traditional data integration pipeline, schema alignment is performed before record linkage [3]. Unfortunately, with a very large number of sources, such a traditional approach becomes infeasible because of the huge variety and heterogeneity among attributes. We propose then to perform data linkage before schema alignment as we can take advantage of the opportunity that products are named entities, and hence a product specification page usually publishes the product identifier.

Schema alignment addresses the challenge of semantic ambiguity and aims to reconcile the attributes offered by different sources, that is, to understand which attributes have the same meaning and which ones do not, as well as identify value transformations to normalize different representations of the same attribute values. Since we do not rely on a global schema given in advance, correspondences among attributes are established bottom-up leveraging the results of the previous data extraction and data linkage phases.

Data fusion tackles the issue of reconciling conflicting values that may occur for attributes from different sources. Data fusion aims at evaluating the trustworthiness of data, deciding the true value for each data item, and the accuracy of the sources. To address these challenges, data fusion techniques rely on data redundancy, which further motivates the need to process many sources.

2.1 Our approach: redundancy as a friend

Our approach to develop the above pipeline aims at taking advantage of the *opportunity* that products are named entities, and hence a product specification page usually publishes the product identifier. Web sources that deliver product specification pages publish product identifiers mainly for economic reasons: websites need to expose the product identifiers to let them be indexed by shopping agents and available to customers who search products for comparing prices or consulting specifications. Large e-commerce marketplaces strongly encourage sellers and retailers to publish product identifiers, as they improve efficiency both for the internal management of data and for the exchange of data with search engines like Google and Bing.

The presence of identifiers allows us to drive the pipeline from source discovery to data integration by leveraging the opportunity of *redundancy of information at the global level*, and the *homogeneity of information at the local level*.

- At the global level, we observe that head (popular) products are present in several head (large) sources as well as in many tail (small) sources. Therefore, we expect that identifiers of head products are spread across many sources. Further, many head products in a category will often co-occur in multiple sources.
- At the local level, we observe that the structure and the semantics of information, within each source, tend to be regular. Hence, we expect the product specification and product identifiers presented in a given page are published according to the same structure for every page in the same source.

Figure 2 illustrates the key intuitions underlying our approach from source discovery to data integration to meet the goal of *effectively* and *efficiently* dealing with head and tail sources, hence including all pages of head and tail entities. Starting from known head entities in head sources, we take advantage of homogeneity of information at the local level to extract product specifications and identifiers for tail entities in head sources (even head sources offer many tail entities). Then, we exploit the presence of head entities across sources: searching head identifiers, we discover tail sources (even tail sources offer a few head entities). Again, we exploit homogeneity of information at the local level to extract identifiers and specifications for tail entities in tail sources.

eBay, Amazon, Google Shop explicitly require sellers to publish the id for many product categories. For example, see eBay's rules:
http://for-business.ebay.com/product-identifiers-what-they-are-and-why-they-are-important.

Based on our *Redundancy as a friend* approach, we developed a focused crawler, Dexter [6], to discover and crawl product web sites offering product pages for the input categories. Dexter starts from a seed set of product pages, and iteratively discovers and crawls new sources, from which extracts products specifications. Then, we developed a big data linkage approach specifically tailored for product pages that takes advantage of the redundancy of information at a global level, and homogeneity of structure and semantic at the individual source level [5].

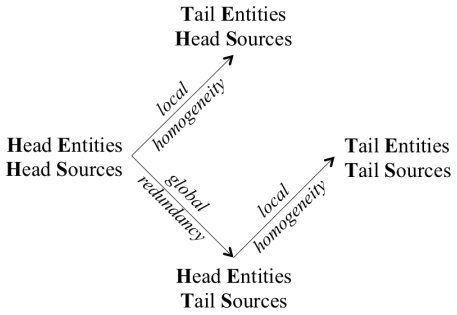

Fig. 2. Our approach.

3 Lessons Learned and Research Agenda

In an experimental evaluation performed between Sept 2014 and Feb 2015, we have trained the focused Dexter crawler [6] to gather product pages from 10 coarse categories: camera, cutlery, headphone, monitor, notebook, shoes, software, sunglasses, toilet accessories, televisions. The crawler discovered 3.5k websites, for a total of 1.9M pages. Each website contributed to provide pages for the different categories, and pages were grouped into $7,145$ clusters, corresponding to the local categories exposed by the websites (on average every websites has 2 local categories). The dataset is publicly available on-line.

Building a Benchmark Product Dataset – We compared the contents of our dataset with pages in Common Crawl, an open repository of web crawl data. About 68% of the sources discovered by our approach were not present in Common Crawl. Only 20% of our sources contained fewer pages than the same sources in Common Crawl, and a very small fraction of the pages in these sources were product pages: on a sample set of 12 websites where Common Crawl presented

Note that the on-line version (https://github.com/disheng/DEXTER) is an extension of the dataset presented in [6].
http://commoncrawl.org/

more pages than in our dataset, we evaluated that only 0.8% of the pages were product pages.

These results suggest the critical need for the community to build a suitable benchmark product dataset to conduct big data research. Our dataset can be considered a first step in this direction. Another important step would be that of maintaining the dataset over time, as discussed next.

Maintaining the Benchmark Product Dataset: Addressing the Velocity Challenge – In March 2018, we have checked all the URLs of the pages of the dataset. We have observed that just 30% of the original pages and 37% of the original sources are still valid (we consider a source valid if it contains at least one working URL). We also performed an extraction of the product specifications. We obtained complete specification from just 20% of the pages.

These numbers clearly indicate that the velocity dimension affects all the tasks of the pipeline. Developing solutions to collect snapshots over regular time intervals and perform data integration over time can open intriguing research directions. While some activities, such as checking the appearance/disappearance of sources can be done on monthly basis, others, such as crawling web sites to check appearance/disappearance of pages and changes in the pages should be performed more frequently. To this end, the development of efficient incremental solutions for source discovery and web crawling represent interesting research directions.

As our experiments emphasize, data extraction rules are brittle over time. The development of wrappers resilient to changes in the pages has always been a primary goal in data extraction research. A dataset with multiple snapshots over a long interval of time, as the one that we have advocated above, could serve as a benchmark for data extraction solutions.

Harnessing Velocity – We observe that while velocity represents a challenge, it could also become an opportunity. We observe that analyzing changes in the pages could help improve our data extraction and data linkage techniques.

For example, one of the main challenges of the product identifier extraction step is to eliminate identifiers that refer to suggested and related products, instead of the main product of the page. Examining the same product page over time may help us to more easily separate out the former part, since those may change faster over time than the description of the product in the page.

Schema Alignment – We are currently working on the development of techniques to perform schema alignment for our product domain. The main difficulties that we have to face are due to the heterogeneity at the schema and at the instance level, due to the large number of independent sources.

To give a concrete example of the heterogeneity at the schema level, consider the dataset collected using the Dexter crawler, described earlier in this section. The specifications extracted from these sources contain more than $86k$ distinct attribute names (after normalization by lowercasing and removal of non alphanumeric characters). Most of the attribute names (about $85k$) are present in

less than 3% of the sources, while only 80 attribute names occur in 10% of the sources, with the most popular attribute name occurring in just 38% sources.

The solution that we are investigating exploits data linkage to cluster attributes that share the same values. Since heterogeneity occurs also at the instance level, we aim at progressively finding correspondences between the cluster resolving increasingly complex matches between values with different representations.

Addressing the Veracity Challenge and Data Fusion – Analyzing the sources of our dataset we noticed that some clusters contain product pages from different categories. In some cases, the errors are in the sources: some websites adopt unexpected (or even completely wrong) criteria as, for example, classifying monitors under a laptop category or vice-versa; other websites aggregate products and related accessories (which represent another category of products). In other cases, the errors are due to wrong classifications by the system.

To overcome these issues we want to investigate different solutions. First, we believe that introducing feedbacks in our pipeline could significantly improve data quality, especially for precision. For example, alternating source discovery and data linkage we could select better identifiers to feed to Search for source discovery, thus increasing the precision, with respect to the target category, of the sources. Similarly, results from schema alignment could help improve the precision of linkage, as pages whose attributes do not align are unlikely to represent the same product. Another promising direction to improve precision without penalizing recall is to study solutions to exploit humans in the loop. In particular, we aim at developing and evaluating techniques based on active learning and crowdsourcing to continuously train the classifiers with effective and updated training sets.

We have observed many inconsistencies between values of a product attribute across sources. Existing data fusion techniques [2, 3] can help to resolve these inconsistencies when they are due to honest mistakes, possibly in combination with extraction errors. However, the product domain also exhibits inconsistencies due to deceit, where sources may deliberately provide imprecise or erroneous product characteristics. Identifying and effectively addressing such inconsistencies at web scale is an important direction of future work.

Beyond Source Homogeneity – Our approach to data extraction is based on the assumption that pages are structurally homogeneous at the local source level. This is a valid assumption for a vast majority of websites, but there are exceptions. For example, some websites that publish used products have a weak template and leave the seller the freedom to publish the product specification without any structural imposition. For some application scenarios, one can simply drop these sources. If on the contrary they are important, data extraction should be performed by solutions that do not rely on the template.

Beyond Product Specifications – So far we have considered the extraction of the identifiers and of the specifications. However important data that complete the

product description are price and reviews. Challenging issues for the extraction of price are to distinguish the price of the principal product in the page from the prices of other products, such as suggested products, similar products, and the actual price from discounts or list price. Reviews represent important information in many applications. An interesting problem is how to combine structured data from the specification with the unstructured data of the reviews.

References

1. Dalvi, N., Machanavajjhala, A., Pang, B.: An analysis of structured data on the web. Proceedings of the VLDB Endowment **5**(7), 680–691 (2012)
2. Dong, X.L.: How far are we from collecting the knowledge in the world? In: Keynote at 19th International Workshop on Web and Databases. ACM (2016)
3. Dong, X.L., Srivastava, D.: Big data integration, vol. 7. Morgan & Claypool Publishers (2015)
4. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12), 68–74 (2008)
5. Qiu, D., Barbosa, L., Crescenzi, V., Merialdo, P., Srivastava, D.: Big data linkage for product specification pages. In: Proceedings of the 2018 ACM SIGMOD International Conference on Management of data. ACM (2018)
6. Qiu, D., Barbosa, L., Dong, X.L., Shen, Y., Srivastava, D.: Dexter: large-scale discovery and extraction of product specifications on the web. Proceedings of the VLDB Endowment **8**(13), 2194–2205 (2015)