

SmartCOPI

Smart Consolidation of Product Information

Maurice van Keulen (University of Twente), m.vankeulen@utwente.nl
Dolf Trieschnigg (Mydatafactory), dtrieschnigg@mydatafactory.com

Abstract

Maintaining the quality of detailed product data, ranging from data about required raw materials to detailed specifications of tools and spare parts, is of vital importance in many industries. Ordering or using wrong spare parts (based on wrong or incomplete information) may result in significant production loss or even impact health and safety. The web provides a wealth of information on products provided in various formats, detail levels, targeted at a variety of audiences. Semi-automatically locating, extracting and consolidating this information would be a “killer app” for enriching and improving product data quality with a significant impact on production cost and quality. The new to COMMIT/ industry partner Mydatafactory is interested in both the web harvesting and data cleansing technologies developed in COMMIT/-projects P1/Infiniti and P19/TimeTrails for this potential and for improving Mydatafactory’s data cleansing services. The ICT science questions behind data cleansing and web harvesting are how noise can be detected and reduced in discrete structured data, and how human cognitive skills in information navigation and extraction can be mimicked. Research results on these questions may benefit a wide range of applications from various domains such as fraud detection and forensics, creating a common operational picture, and safety in food and pharmaceuticals.

Three partners and their status

The **new to COMMIT/ industry** partner Mydatafactory provides product data cleansing solutions to global and regional companies. Mydatafactory focuses on one thing: fixing product data quality problems by delivering customers the best software to cleanse, update, structure, standardize and enrich their product data. In an industrial setting product data is used throughout the production chain: it ranges from data about required raw materials, and specifications of tools and spare parts needed in the production process, to detailed information about the delivered products. The number of items involved in the process may easily grow to several hundreds of thousands. The Mydatafactory software is intelligent; it improves and adapts itself depending on the usage. Where others work with rule-based functions, Mydatafactory works with machine learning, language processing and dictionary-driven technologies. In this way, product data cleansing processes can be shortened from weeks/months to hours/days, costs reduced to 20%, and errors reduced from over 50% to less than 1% of the products. People involved in this Zwaluw project will be Arjan Herskamp (director), Dolf Trieschnigg (data scientist, previously a PostDoc researcher at the University of Twente, not involved in COMMIT/), as well as software engineers from Mydatafactory and sister-company Calago. The **Old to COMMIT/ academic** are project leader Maurice van Keulen and Software Engineer Jan Flokstra from the University of Twente both involved in COMMIT/-projects P1/Infiniti and P19/TimeTrails. Data enrichment and cleansing are important drivers in smart industry. Mydatafactory has the ambition to be the best in what they do by staying on the forefront of technology, and is as such an attractive partner for the upcoming COMMIT2DATA program. The **New to COMMIT/ academic** is Brend Wanders, currently finishing his PhD at the University of Twente. Brend worked on probabilistic data integration and cleansing technology for bioinformatics.

His cooperation with the FedSS project (ITEA2) has led to a promising approach for data cleansing based on his open source probabilistic datalog engine¹.

Mydatafactory is interested in both the web harvesting and data cleansing technologies developed in COMMIT/-projects P1/Infiniti and P19/TimeTrails and by Brend himself, for their potential for enriching product data from web sources and for improving Mydatafactory's data cleansing capabilities.

Project description

The philosophy of COMMIT/ is use-inspired research, so we first present the use. We then highlight the ICT science questions that were inspired by this (intended) use. Finally, we discuss how solving these more generic ICT science questions can have other applications in other top sectors.

Use: consolidating complete and high quality product information

Product information is of vital importance in many industries. Maintaining accurate and detailed information is needed in many parts of the production chain: not only do industries have to provide precise information about the produced products, they also have to manage product information involved in the production process, ranging from information about required raw materials to detailed specifications of spare parts used in the production line. Incomplete product information may result for instance in ordering the wrong spare part. This could result in a disrupted production line and production loss, but could also impact health and safety. Using inappropriate substances in a food- or pharmaceutical context may result in non-complying end-user products. Gartner² estimates that "data quality affects overall labor productivity by as much as a 20%". For the specific case of product data, these effects might even be larger.

Completing and verifying product information is hard and time-consuming: contradictory evidence may be present in a variety of enterprise resources (such as ERP and PIM systems). This information could be extended with vendor information from the web, which provides a wealth of information of product information, provided in various formats and targeting at a variety of audiences (end-users, domain specialists). A complicating factor is the varying quality of the available information. The people responsible for the product descriptions on the corporate webpage cannot be expected to be as knowledgeable as the domain specialists in the factory. An error in the product information in a published catalogue or website can therefore remain undetected. Finally, a lot of implicit information is available in the minds of the people involved in the production process.

The goal of Mydatafactory is to aid companies in dealing with product data quality issues in a manner that uses human effort and expertise as efficiently as possible by employing advanced machine learning and information retrieval technology. Semi-automatically harvesting, extracting and consolidating these three types of knowledge sources (enterprise, web and people) would be a "killer app" to improve product data quality with a significant impact on production cost and quality.

Inspiration: the ICT science questions

The application emphasizes the need for advances in two related areas: *data cleansing* and *web harvesting*. Both have their own specific underlying ICT science question.

- **How can noise be detected and reduced in discrete structured data?**

In analogy with electronics, errors in source data, conflicts between sources, mistakes due to ambiguity in data processing (e.g., classification, natural language processing and entity resolution), can be seen as *data noise*. The analogy becomes even stronger when a probabilistic data approach is used as it more faithfully represents uncertainty and

¹ Habib, M.B. and Wanders, B. and Flokstra, Jan and van Keulen, M. (2015) *Data uncertainty handling using evidence combination: a case study on maritime data reasoning*. In: Proceedings of DEXA workshop ISSASIM 2015, 1-4 Sep 2015, Valencia, Spain. IEEE Computer Society.

² https://www.data.com/export/sites/data/common/assets/pdf/DS_Gartner.pdf

untrustworthiness in the data.³ Data cleansing becomes a form of noise reduction while protecting 'the signal' (correct data). In an electronic signal, the *global* concept of frequency plays an important role in noise reduction; with discrete structured data, a data item's context formed by surrounding and related data can be used to reason away or dampen the presence of noisy erroneous data while protecting or even boosting correct 'signal' data. Relevant sub-questions are related to determining trustworthiness of (combined) data, granularity of trust, reasoning with probabilistic data and how to make best use of the human-in-the-loop.

- **How to mimic human cognitive skills in information navigation and extraction?**

Web sites and natural language texts contain valuable information represented in a form befitting *human* information gathering. Unconsciously, humans use their cognitive skills to determine which link to click for navigation and which string of characters represents a value of a certain attribute for a certain entity. More automation in web harvesting, especially in enriching and consolidating data with information from the internet, requires computers to obtain more autonomy and robustness in information navigation and extraction in this human-shaped information space. Two central human capabilities are vital: being able to doubt and to reason with context information. We believe that a probabilistic approach combined with machine learning, information retrieval and data cleansing techniques is the road towards computers being able to autonomously and robustly handle the uncertainty, ambiguity and reliance on context that a human-shaped information space brings with it.

Vision: other applications

Although data cleansing and web harvesting have applications in almost any domain imaginable, we would like to highlight a few concrete ones outside the smart industry domain (the domain of this proposal) namely "Big data for security" and "Big data for life".

- **The role of the internet in fraud risk analysis and forensics**

Governmental organizations responsible for keeping certain types of fraud under control, often use data-driven methods for both immediate detection of fraud, or for fraud risk analysis aimed at more effectively targeting inspections of persons and organizations. A blind spot in such methods, is that government data may represent a 'paper reality'. Fraudsters will attempt to disguise themselves with disinformation painting a world in which they do nothing wrong. This blind spot can be counteracted by enriching their data with traces and indicators from more 'real-world' sources such as social media and internet. This allows for two new strong kinds of indicators: (a) discrepancies between government data and real-world traces, and (b) other involved persons and bystanders leaving behind observations and opinions about the subjects, i.e., the human as a sensor. There is an established contact at the Inspection of the Ministry of SZW with whom we already have conducted a pilot with the web harvesting technology developed in both COMMIT/-projects.

- **Creating a Common Operational Picture (COP)**

In projects TEC4SE and FedSS, we collaborate with Thales on consolidating and cleansing data from multiple heterogeneous sources for the purpose of creating a COP. We have targeted so far applications for the coast guard (FedSS) and for police and others aimed at crowd surveillance at massive events (TEC4SE).

- **Safety in food and pharmaceuticals**

Product data quality problems form a major safety risk in food and pharmaceuticals. Wrong application of drugs or food additives due to incorrect information may incur severe health risks. The product data cleansing technology of this project is immediately applicable to this domain. Furthermore, observations and opinions left behind by humans on the internet are a valuable source for indications of problems, for example, reports of sickness attributed to the

³ van Keulen, M. (2012) Managing Uncertainty: The Road Towards Better Data Interoperability. IT - Information Technology, 54 (3). pp. 138-146. ISSN 1611-2776

use of some product or suspicions that the contents of a package do not precisely coincide with what is said on the label.

How does the research contribute to overall goals?

This project is mainly based on the following research results

- The web harvesting software developed in P1/WP7 and P19/WP6 including the pilots we have conducted with it with COMMIT/ partners as well as other organizations (some of which mentioned in the above).
- Probabilistic database and probabilistic data integration technology developed over the years in MultimediaN/AmbientDB, as well as specific for natural language processing and information extraction in the local project Neogeography and regional project TEC4SE.
- Data cleansing technology based on Probabilistic Datalog as developed in the local project PayDIBI (bioinformatics domain) and the ITEA2 FedSS project (maritime COP), which are in part based on the ambiguity handling approach developed in the Neogeography project.

As mentioned under “Vision” above, harvesting and information extraction for enrichment and consolidation of data have applications in many domains of the top-sectors COMMIT2DATA aims at. The need for such technology is heard ever more loudly: empower (digital) investigators in fraud detection and forensics, using the man-in-the-street as sensor, creating a Common Operational Picture, etc. Such signals call for more support by, hence autonomy and robustness of, computers in information gathering, combination, and cleansing.

Besides these needs from societal sources, a similar need in e-science can be heard. Combining and analyzing data in novel ways has become a main method for research, tackling research questions that could not be answered before. Extraction, transformation, cleansing, and integration of existing data sources has become a primary activity of an e-scientist, often consuming more than half of the time of a PhD project. A significant reduction in an e-scientist’s ‘data fiddling’ effort has the potential of significantly improving their productivity and as a consequence scientific progress in their disciplines. Moreover, mistakes in data understanding and cleansing may significantly invalidate results or their interpretation in pursuit of meaning and causality.

In both society and science, keywords like data quality, unstructured data, heterogeneity, data interoperability, (fine-grained) trust, disinformation, human-in-the-loop, etc. occur ever more often. The use-inspired research of this proposal has been shaped with the aim of bundling and improving previous and on-going efforts in both COMMIT/ as well as other research projects, giving the research results a new domain context, and lining up research and partners for cooperation in COMMIT2DATA.