Consiglio Nazionale delle Ricerche
Istituto di Calcolo e Reti ad Alte Prestazioni

Large-Scale Extraction of Product Information on the Web

Ermelinda Oro, Massimo Ruffolo

ICAR-CNR-06-2015 **Novembre 2015**

Consiglio Nazionale delle Ricerche, Istituto di Calcolo e Reti ad Alte Prestazioni (ICAR)
– Sede di Cosenza, Via P. Bucci 41C, 87036 Rende, Italy, URL: *www.icar.cnr.it*
– Sezione di Napoli, Via P. Castellino 111, 80131 Napoli, URL: *www.na.icar.cnr.it*
– Sezione di Palermo, Viale delle Scienze, 90128 Palermo, URL: *www.pa.icar.cnr.it*

Large-Scale Extraction of Product Information on the Web

Ermelinda Oro and Massimo Ruffolo

Altilia srl
National Research Council (CNR)
Via P. Bucci 7/11C, 87036
Rende (CS), Italy
{linda.oro,massimo.ruffolo}@icar.cnr.it
{linda.oro,massimo.ruffolo}@altiliagroup.com

Abstract. E-Commerce is growing at an explosive rate but in a highly competitive, increasingly transparent environment. Thus, online retailers and product manufacturers need effective, affordable and reliable price and product intelligence solutions, capable to answer to needs like price monitoring, better assortment management, improved and timely product perception analysis. The highest competition in the e-commerce market, and the limitations in existing price and product solutions, pose challenges to solve pains of online retailers and product manufacturers. In this Industrial paper, the Altilia company, spin-off of the Italian National Research Council, presents its model and system named "Price and Product Intelligence Advisor" and describes lesson learned for bringing to the market an innovative and scalable solution that is able to offer price and product intelligence services to both online retailers and product manufacturers, even of small size.

Key words: E-Commerce, Price Intelligence, Product Intelligence, Big Data, Big Data Platform, Semantic Analysis, Sentiment Analysis, Opinion Analysis, Ontology,

1 Introduction

E-Commerce is growing at an explosive rate and todays manufactures and retailers work in a highly competitive and increasingly transparent environment. The growth in the e-commerce industry strongly depends on the ability of online retailers and product manufactures to offer better products at optimal conditions to their customers. Competitions has tremendously evolved in last few years due to the increased: (i) price and product features sensitivity of consumers, (ii) aggressiveness of competitors, (iii) price transparency and product showrooming made possible by the wide smartphone adoption, make price and product intelligence a fundamental decision support tool for companies seeking to optimize their own pricing and product strategies with respect to their competitors, (iv) availability of huge amount of buyers comments about product of all categories on online retailers web sites, and on social/online media, that makes almost impossible to manual curate this massive amounts of unstructured big data. For

these reasons price and product intelligence poses many issues, and it is still a complex task to perform due to performance problems and limitation of existing solutions.

Considering our experience and the deep analysis of existing tools we performed in last months, we discovered that only scanty IT solutions actually provide to Chief Marketing Officers and Category, Brand, Product Designer and Managers, limited real-time and actionable insights that can be acquired from big data available online. Existing price and product intelligence tools are still too complex, expensive and limited for target customers willing to pay. In addition, frequently potential customers have to spend a lot of resources in manually harnessing the output of current price intelligence software, by expert like data scientists, difficult to find and costly. Most of existing commercial solutions: (i) Monitor only specific product categories. (ii) Are unable to access web sites requiring login or form filling. (iii) Are unable to align in automatic way customers catalogues with their internal product taxonomy. (iv) Dont have a wide library of dynamic pricing algorithms for different products categories. (v) Have limited semantic capabilities to understand perceptions, sentiments, opinions, issues of consumers in buying and using products. Due to these limitations many online retailers and manufactures tackle to their pricing problems in three main different ways: (i) Ask advice services to specialized market research companies (e.g. GFK Eurisko, Nielsen). (ii) Buy raw data from price intelligence providers to analyze competitors product catalogues (e.g. Semantics3). (iii) Lease specialized applications from a price intelligence companies (e.g. Boomerang ecommerce). However, such approaches are time consuming and expensive. In particular, the first is very expensive and slow; the second approach demands for a well-staffed IT team, and the third has a high cost of ownership. Finally, many comparison shopping service websites (like for: insurances, hotels, airlines, car hire companies) offers to consumers a merely price comparison exploiting data received by sellers, that is not an effective solution for retailers and manufacturers that need a comparison with their selling policy.

In this paper, we introduce a method to overcome the limitations of competitors and take hold in the large and remunerative price and product intelligence target market. The proposed method is named Altilia Price and Product Intelligence Advisor (APPIA). APPIA is able to:

i) Timely and continuously collect product prices and customer comments/recommendations on products from online retailer web sites.
ii) Provide product price comparison to online retailers that may dynamically predict and apply the best product price on the base of competitors offers.
iii) Apply semantic analysis to shopper comments in order to provide manufactures with the perception of each product feature in comparison with competitive products.

2 Price and Product Intelligence System

The goal of product intelligence is to accelerate the rate of product innovation, thereby making the product and its owners more competitive. Essentially price and product intelligence is based on a big data gathering, management and analysis process involving the following steps:

i) **Big data extraction and acquisition** that imply to find, access and extract data from product pages available on disparate online retailers web sites, threads of forum and blogs, post of social media, etc.
ii) **Product identification and matching** aimed at determining through algorithms, whether or not the product matches exactly, or if it is a comparable product in order to refer all available data to the right product.
iii) **Big data analysis** that is the process of creating data collections related to prices, shipping cost, availability, buyers comments and the analysis of this data by the right algorithm capable to provide descriptive analysis, trends, predictions, semantic analysis of natural language texts.
iv) **Big data quality check** that imply regular checks for accuracy.
v) **Reporting and alerts** that is related to the ability to gain actionable insights from the big data that has been gathered.

An initial version of APPIA is developed on top of MANTRA Smart Data Platform[1], the Altilia's product, that is an innovative and advanced big data analytics technology having revolutionary web data collection, semantic and predictive, capabilities. In particular, MANTRA Smart Data Platform is an hybrid-cloud platform that exploits big data, cloud, and machine intelligence technologies to enable: (i) Big data acquisition from heterogeneous data sources by proprietary data extraction algorithms and data source connection tools. (ii) Big data harmonization by advanced and semantic data manipulation, transformation, wrangling, alignment and integration algorithms and tools. (iii) Semantic big data enrichment by natural language processing algorithms that performs entity resolution and extraction, sentiment and opinion analysis, concepts, facts, actions, relations and events identification and extraction. (iv) Big data aggregation and analysis by most innovative and powerful machine learning algorithms that exploit distributed and stream computing paradigms. (v) Big data exploration by advanced and powerful semi/un/structured data querying that make also use of NLP-based approaches. (vi) Big data visualization by advanced, dynamic, multi-dimensional, and heterogeneous data charts. The design and implementation of all algorithms, which the MANTRA Platform is based on, are founded on rigorous research activities of founders in the area of data capture, artificial intelligence, semantic technologies, machine learning, web scraping, big data management, natural language processing.

management architecture.

MANTRA is based on the contextual workflow approach that enables to develop applications as composition of MANTRA APPs. Each MANTRA APP is a program providing a specific set of functionalities (e.g. data capture, acquisition, analysis; semantic enrichment) that abstract and hide the complexity of performed activities to users. A key element of differentiation of the MANTRA platform is the ability to allow customers to address many application problems and business use cases by a single, unique, powerful, and highly scalable technology.

APPIA developed on top of the MANTRA Smart Data Platform provides a software stacks designed to optimize big data processing tasks applying a unique combination of innovative semantic analysis, machine learning, artificial intelligence algorithms over

[1] Altilia - We make data smart. http://www.altiliagroup.com/

extremely large numbers of high volume streams of possibly noisy and incomplete big data. Figure 1 shows how contextual workflows works to perform price intelligence.

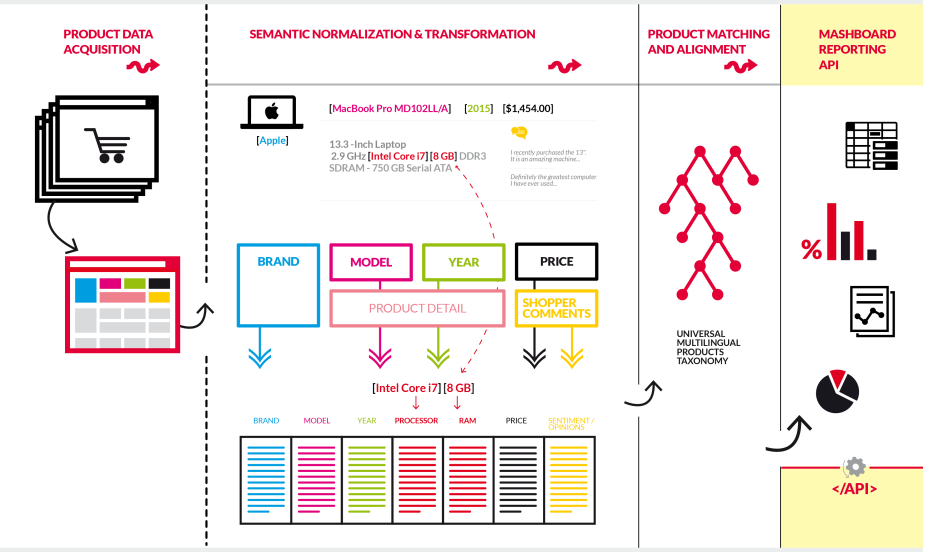

Fig. 1. How contextual workflows works to perform price intelligence

3 Competive Landscape and Related Work

The competitive landscape has been analyzed by considering the following segmentation variables: (i) geography, (ii) product features, and (iii) revenues level. Regarding the geography, competitors are mainly based in USA and UK. A less number of competitors is based in Canada, other European countries, and India. As key benchmarking, we identify the features described in the following Table where in italics we describe why APPIA Solution overcome competition.

In analyzing the competitive landscape we considered the set of features described in Table 2 made available to customers by each competitor, the offer innovativeness, the market penetration and the revenue as scoring criteria. On the base of above mentioned segmentation variables, competitors can be roughly classified in the 3 categories shown in the following table.

One closest target competitor, that recently raised 8.5 million dollars in a Series A round of financing from US based VCs is boomerangcommerce.com. This is a US based company providing innovative price intelligence and dynamic price optimization product with advanced price intelligence features based on machine learning techniques. However, it do not have the ability to semantically process product information, so far they do not offer customers the catalogue alignment and cover only a limited set of

Table 1. Key features of the proposed solution

FEATURE	**DESCRIPTION**
Product Data Sources	It identifies the set of product information sources that the specific player is able to process. Possible source are: standard e-commerce web sites (e.g. mediamarkt.de), private sales online stores (e.g. saldiprivati.com), and web sites that requires form filling (e.g. http://online.linear.it/WebSite/DatiAssicurativi). *No competitors are able to automatically access private sales online stores and web sites that requires form filling, this drops away large share of the potential market.*
Product Data Acquisition Methods	It refers features related to method used for collecting product information. Players mainly use web scraping, web sites monitoring, API access, manual data entry. *Only few competitors use automatic methods to acquire product information from online private stores. So, only few competitors are able to provide real-time price intelligence capabilities. The APPIA Solution will benefit from advanced innovative patent pending web data extraction algorithms available in the MANTRA Smart Data Platform™.*
Product Categories	It defines how many product categories a player provides to its customers. There are very few players that are general purposes and address a wide range of products in different categories, more players are specialized in specific product categories. Many players have a limited internal product taxonomy and coding. *There are no existing solutions that already have a Universal Multilingual Product Taxonomy so, typically players work for a specific geography or for specific product categories.*
Product and Brands Intelligence Facilities	It is a set of features related to the capability of a player to manage products, catalogues and assortment information and the ability to semantically process product information contained in product recommendations available on e-commerce web sites, social and online media contents. Such features are: ability to search and match equal product over many competitors, the ability to search by similar products in order to compare assortments and catalogues, the ability to semantically analyze product recommendations and consumers comments for providing insights on products (sentiments, opinions, issues), the ability to automatically align competitors catalogues with the customer catalogue for saving time and money and making a more complete and effective users experience. By these features it is possible for brands manufactures to compare products prices on all ecommerce players for checking price violations, frauds and optimizing price strategies and assortments. *No players offer these features. The APPIA Solution can benefit from advanced and innovative semantic algorithms available in MANTRA Smart Data Platform and that we intent to develop by this project. This set of features enables to aggregate, analyze and compare product information coming from different ecommerce players.*
User Interaction	It refers the set of features that enable users to access and use product information and analysis results. *Some players provide only API for accessing product information, other have mono or multi-devise form of reporting and/or price changes alerting. Some have integration with e-commerce platform. The APPIA Solution will provide a complete set of SaaS-based reporting, Product search and retrieval functionalities, Price alerting, and a complete API to provide raw data to customers that would like to use their internal BI environment saving already done investments.*

Table 2. Key features of the proposed solution

Strong featuring	**Good market visibility**	**Limited competition**
Boomerangcommerce, Pricesquid, Priceweave, Profitero, Wiser	360pi, 24ispriceintelligence, Insitetrack, Price2spy, Workit, Market-hub	Cartbound, Competitormonitor, Indix, Mindrest, Semantics3, Marketyze, PriceZombie, Price Tracker

product categories. The Figure 2 shows the positioning of APPIA with respect to competitors.

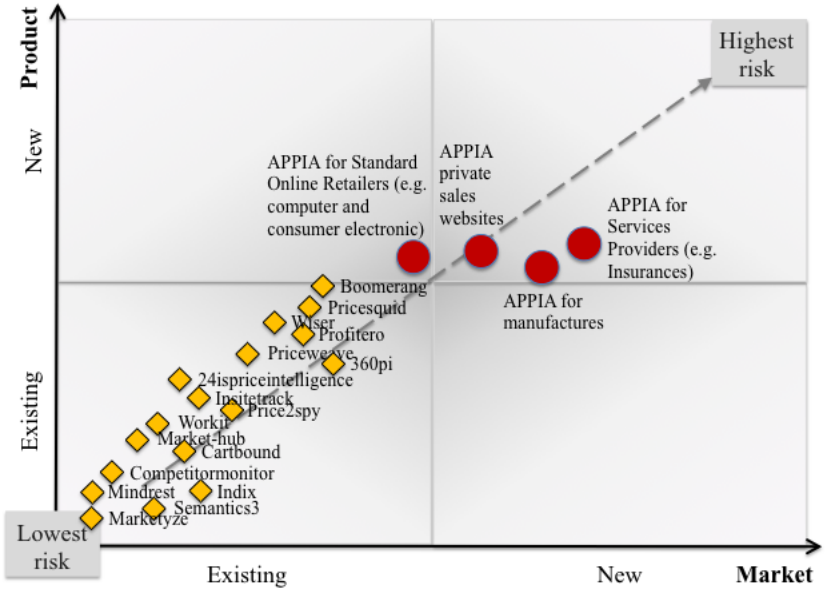

Fig. 2. The positioning of APPIA vs main competitors

APPIA combines innovative algorithms coming from many different scientific and technological areas. In particular, in addition to overcome limitation of existing commercial products that provide product price comparison, APPIA has patented and patentable ideas that enable to:

i) Timely and continuously collect product prices and customer comments/recommendations on products from online retailer web sites. In literature exists different algorithms that performs web data extraction, but heterogeneous big data processing massively involving semi/unstructured data is still an open research field. In [1] presents an overview of the literature in the field of Web Data Extraction. The

ecent paper [5] introduce a tool that produces structured interoperable data from product features, i.e.,attribute namevalue pairs, on the web.

ii) Apply semantic analysis to shopper comments in order to provide manufactures with the perception of each product feature in comparison with competitive products. In [2] shown that a recommendation system (RS) can also be effectively applied to e-commerce retailers to promote their products and services. Amazon published its patent in 2001 acting as real and effective usecase for RSs applyed in the e-commerce domain [4, 7, 6]. Also recently new RSs approaches are presented [3, 8] proposed a framework. We include in APPIA semantic analysis features (such as accurate sentiment/opinion analysis) that are at the base of recommendation systems.

4 Conclusions

The APPIA project has been launched under the pressure of existing Altilia customers, and prospects in the sales pipeline, that constantly express the nailing needs for an innovative, flexible, scalable, easy-to-adopt and -use price and product intelligence solution. The APPIA value proposition is based on the following uncommon features that drive technical and commercial advantages over competitive solutions.

Technical and usability advantages descend from the SaaS nature of APPIA that enables customers to: (i) Easily and dynamically access in real-time prices, perceptions, and description of a huge number of products sold by worldwide online retailers. (ii) Get data and insights of interest (e.g. catalog configuration of a give online retailer) in different ways, as web report, in tabular forms, as JSON/CSV files, and so on, directly available for humans or useful to feed other systems and applications. (iii) Connect the APPIA API in order to integrate APPIA into their digital environment.

Commercial advantages are a consequence of the technical uniqueness that enable customer to: (i) Provide on-line retailers with a complete apperception into competitors product pricing, brand assortment, promotions and stocks, giving the possibility to make automatic, smarter, more informed and profitable pricing decisions. (ii) Help manufactures in improving on sale products positioning, getting insights about buyers perception of product on the base of all comments available on the web. (iii) Reduce costs and augment margins by shrinking time to insights to minutes and by automating and optimizing pricing, promotions and assortments decisions that currently requires complex time consuming manual work. (iv) Address data scientists shortage and the high costs related to big data analytics projects by an easy to learn and use application that speed-up big data driven solutions adoption. Such advantages over competitive solutions enable to save costs in exploiting big data for market/product and price intelligence applications because the TCO of APPIA is at least 30% less than competitive solutions and improve revenues and margins because insights are based on an comprehensive view of worldwide markets. These aspects are very disruptive for the entire big data driven market analysis panorama in Europe and in the World.

References

[1] Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner. Web data extraction, applications and techniques: a survey. *Knowledge-based systems*, 70:301–323, 2014.

[2] Yung-Ming Li, Chun-Te Wu, and Cheng-Yang Lai. A social recommender mechanism for e-commerce: Combining similarity, trust, and relationship. *Decision Support Systems*, 55(3):740–752, 2013.

[3] Ting-Peng Liang, Xin Li, Chin-Tsung Yang, and Mengyue Wang. What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach. *International Journal of Electronic Commerce*, 20(2):236–260, 2015.

[4] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[5] Tuğba Özacar. A tool for producing structured interoperable data from product features on the web. *Information Systems*, 56:36–54, 2016.

[6] Badrul M Sarwar, George Karypis, Joseph Konstan, and John Riedl. Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering. In *Proceedings of the fifth international conference on computer and information technology*, volume 1. Citeseer, 2002.

[7] J Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166. ACM, 1999.

[8] Wayne Xin Zhao, Jinpeng Wang, Yulan He, Ji-Rong Wen, Edward Y Chang, and Xiaoming Li. Mining product adopter information from online reviews for improving product recommendation. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 10(3):29, 2016.