Expert Systems With Applications 63 (2016) 20–34

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

An ontology-based data integration approach for web analytics in e-commerce

María del Mar Roldán García, José García-Nieto*, José F. Aldana-Montes[1]

Dept. de Lenguajes y Ciencias de la Computación, University of Málaga, ETSI Informática, Campus de Teatinos, Málaga - 29071, Spain

ARTICLE INFO

Article history:
Received 15 December 2015
Revised 15 June 2016
Accepted 16 June 2016
Available online 23 June 2016

Keywords:
Semantic model
Ontology
E-commerce
Web analytics

ABSTRACT

Web analytics has emerged as one of the most important activities in e-commerce, since it allows companies and e-merchants to track the behavior of customers when visiting their web sites. There exist a series of tools for web analytics that are used not only for tracking and measuring web traffic, but also for analyzing the commercial activity. However, most of these tools focus on low level web attributes and metrics, making other sophisticated functionalities and analyses only available for commercial (non-free) versions.

In this context, the SME-Ecompass European initiative aims at providing e-commerce SMEs with accessible tools for high level web analytics. These software facilities should use different sources of data coming from digital footprints allocated in e-shops, to fuse them together in a coherent way, and to make them available for advanced data mining procedures. This motivated us to propose in this work an ontology-based approach to collect, integrate and store web analytics data, from many sources of popular and commercial digital footprints. As article's main impact, we obtain enriched and semantically annotated data that is used to properly train an intelligent system, involving data mining procedures, for the analysis of customer behavior in real e-commerce sites. In concrete, for the validation of our semantic approach, we have captured and integrated data from Google Analytics and Piwik digital footprints allocated in 15 e-shops of different commercial sectors and countries (UK, Spain, Greece and Germany), throughout several months of activity. The obtained results show different perspectives in customer's behavior analysis that go one step beyond the most popular web analytics tools in the current market.

1. Introduction

In the last few years, web analytics has emerged as one of the most important activities in e-commerce, since it allows companies and e-merchants to track the behavior of customers when visiting their e-shop sites. Web analytic applications can also help companies to measure the results of traditional print or broadcast advertising campaigns. Web analytics procedure is based on measuring a visitor's behavior once on a given e-shop site, which includes its drivers and conversions (to actual customer). These data are typically compared against key performance indicators and used to improve a website or marketing campaign's audience response.

In the current market, there exist a series of tools for web analytics, such as: Google Analytics, Piwik, Clicky, and StatCounter; that are widely used not only for tracking and measuring web traffic, but also for analyzing the commercial activity, hence to improve the effectiveness of a website. However, these tools often focus on low level and limited sets of web metrics and attributes, without the possibility of providing specialized analyses. In most of cases, high level web metrics and sophisticated functionalities are available only for commercial (non-free) versions, which are rarely accessible by SMEs or individual e-merchants.

In this context, the SME-Ecompass European initiative[2] aims at providing e-commerce SMEs with accessible tools for high level web analytics. These software facilities use the different sources of data coming from different digital footprints allocated in e-shops. However, integrating data from multiple heterogeneous sources entails dealing with different data models, schema and query languages. Therefore, there is a clear demand of integrative proce-

* Corresponding author.
E-mail addresses: mmar@lcc.uma.es (M.d.M.R. García), jnieto@lcc.uma.es, jmgarcianieto@gmail.com (J. García-Nieto), jfam@lcc.uma.es (J.F. Aldana-Montes).

[1] This work is partially funded by FP7 EU project SME E-COMPASS under Grant No: 315637. It is also partially funded by Grants TIN2014-58304 (Spanish Ministry of Sciences and Innovation) and Regional projects P11-TIC-7529/P12-TIC-1519. Authors thanks to involved e-shops to kindly offer web tracking data for testing and validation.

[2] SME-Ecompass FP7 European initiative http://www.sme-ecompass.eu/

http://dx.doi.org/10.1016/j.eswa.2016.06.034
0957-4174/

dures for providing the advanced data mining algorithms with a uniform access to multiple heterogeneous web data sources.

The main hypothesis in this work is: **(H1) an ontology-based integration approach will help us to collect, fuse the data together in a coherent way, and store web analytics data, from many sources of popular and commercial digital footprints**. As a result, **(H2) we will obtain enriched and semantically annotated data that will be able to train data mining procedures for advanced analysis of customer behavior in real e-commerce sites**.

This motivated us to propose a semantic approach that uses an ontology as a mediated schema for the representation and consolidation in a knowledge base of the tracking data from web source's semantics. Semantic web ontologies become a key technology for intelligent knowledge processing, providing a framework for sharing conceptual models about a domain. Semantic mappings between the source schema and the ontology are then defined and used to transform the original data to RDF (Resource Description Framework) [3]. This way, data from heterogeneous sources are stored and integrated inside a single RDF repository, which can be now easily queried by high level algorithms. The goal is to properly feed artificial intelligence procedures capable of deciding how to perform marketing activities, such as: displaying a given advertisement targeted to certain category of clients, or decreasing the price of a product in a given region; then giving rise to sophisticated expert systems for e-commerce applications.

The main contributions of this study are summarized as follows:

- We have developed a semantic approach for the data integration and consolidation of multiple web analytics data sources. These data are daily accumulated from many heterogeneous digital footprints allocated on actual e-shops.
- We have designed and implemented for the first time an OWL (Web Ontology Language) ontology (Dean & Schreiber, 2004) for web analytics. This ontology considers a large and complemented set of attributes and metrics, which have been token from several representative web analytics tools in the market.
- To test hypothesis H1, we have captured and integrated data from Google Analytics and Piwik digital footprints allocated in 15 e-shops of different commercial sectors (retail, tourism, electronics, pharmacy, etc.) and countries (UK, Spain, Greece and Germany), throughout several months of activity. The data are integrated following the same (standard) format and stored in a common RDF repository.
- To test hypothesis H2, obtained "semantized" data are used to train advanced data mining algorithms to perform customer's profile analyses. In particular, these algorithms are tested with success in two cases of study to classify the visitor's behavior and product preference.

The remaining of this article is organized as follows. In Section 2, background and literature overview are presented. Section 3 presents the current state and practices in web analytics for e-commerce. In Section 4, the semantic approach is described, giving details of the service architecture and the OWL ontology. After this, the validation procedure is reported in Section 5. Finally, main conclusions and future work are given in Section 6.

2. Background and related work

This section describes the main background concepts. A review of current related works in the specialized literature is carried out to point out their main differences with regards to our approach.

[3] RDF in W3C https://www.w3.org/RDF/

2.1. Background concepts

- *Ontology.* Ontologies provide a formal representation of the real world, shared by a sufficient amount of users, by defining concepts and relationships between them (Gruber, 1993). In computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. These primitives are typically concepts (classes), attributes (properties), class members (class instances) and relationships (property instances). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.

Ontologies are part of the W3C standards stack for the semantic web, in which they are used to specify standard conceptual vocabularies in which to exchange data between systems, provide services for answering queries, publish reusable knowledge bases, and offer services to facilitate interoperability across multiple, heterogeneous systems and databases.

- *RDF.* Resource Description Framework is a basic ontology language used for representing information about resources on the web (Staab & Studer, 2009). Resources are described in terms of properties and property values using RDF statements. Statements are represented as triples, consisting of a subject, predicate and object. RDF Schema (Staab & Studer, 2009) (RDFS) "semantically extends" RDF to enable us to talk about classes of resources, and the properties that will be used with them. It does this by giving particular meanings to certain RDF properties and resources. RDFS provides the means to describe application specific RDF vocabularies. RDF and RDFS provide basic capabilities for describing vocabularies that describe resources, metadata and ontologies.

- *SPARQL.* It is an RDF query language for ontology models and databases, capable of extracting and manipulating information stored in RDF format. Essentially, SPARQL is a graph-matching query language that can be used to extract knowledge from the model such as the one proposed in this article. Given a data source D, a query consists of a pattern, which is matched against D. The combinations of values resulting from this matching constitute the result of the query (Pérez, Arenas, & Gutierrez, 2009). SPARQL has strong support for querying semi-structured and tagged data, e.g. data with an unpredictable and unreliable structure. SPARQL supports queries to networked, web data sources identified by URIs. In fact, it is a W3C recommendation for RDF data.

- *OWL.* In 2004, the W3C ontology working group (Dean & Schreiber, 2004) proposed OWL as a semantic markup language for publishing and sharing ontologies on the World Wide Web. From a formal point of view, OWL is equivalent to a very expressive description logic where an ontology corresponds to a Tbox (Gruber, 1993). This equivalence allows the language to exploit description logic researcher results. OWL extends RDF and RDFS. When compared to RDF models, OWL adds more vocabulary for describing properties and classes: relations between classes (e.g. disjointedness), cardinality (e.g. "exactly one"), equality, richer typing of properties, characteristics of properties (e.g. symmetry), and enumerated classes (McGuinness & Harmelen, 2004).

- *OWL-DL.* Syntactic variant of the SHOIN (D) description logic (Haase & Stojanovic, 2005) with a different terminology to OWL, which is based on RDFS, hence the support for data values, data types and data type properties. OWL-DL restricts OWL into two distinct ways (Horrocks & Patel-Schneider, 2003): first, some syntactic constructs like recursive descriptions in them are not allowed; second, classes, individuals and properties (respectively concepts, individuals and roles in description logics) must all be disjoint. In this work, we use OWL-DL syntax to formalize the proposed ontology here for our semantic model. A summarized description of basic OWL-DL semantics syntax is shown in Table 1, where an informal logic syntax is represented (left