Ecological Informatics 26 (2015) 203–216

Contents lists available at ScienceDirect

Ecological Informatics

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

Enrichment of the phenotypic and genotypic Data Warehouse analysis using Question Answering systems to facilitate the decision making process in cereal breeding programs

Jesús Peral [1], Antonio Ferrández [1], Elisa De Gregorio [1], Juan Trujillo [1], Alejandro Maté [1], Luis José Ferrández [1]

Department of Software and Computing Systems, University of Alicante, Carretera San Vicente S/N, Alicante 03080, Spain

ARTICLE INFO

Article history:
Received 30 November 2013
Received in revised form 16 April 2014
Accepted 7 May 2014
Available online 15 May 2014

Keywords:
Business Intelligence
Data Warehouse
Question Answering
Information Extraction
Information Retrieval
Genetic information

ABSTRACT

Currently there are an overwhelming number of scientific publications in Life Sciences, especially in Genetics and Biotechnology. This huge amount of information is structured in corporate Data Warehouses (DWs) or in Biological Databases (e.g. UniProt, RCSB Protein Data Bank, CEREALAB or GenBank), whose main drawback is its cost of updating that makes it obsolete easily. However, these Databases are the main tool for enterprises when they want to update their internal information, for example when a plant breeder enterprise needs to enrich its genetic information (internal structured Database) with recently discovered genes related to specific phenotypic traits (external unstructured data) in order to choose the desired parentals for breeding programs. In this paper, we propose to complement the internal information with external data from the Web using Question Answering (QA) techniques. We go a step further by providing a complete framework for integrating unstructured and structured information by combining traditional Databases and DW architectures with QA systems. The great advantage of our framework is that decision makers can compare instantaneously internal data with external data from competitors, thereby allowing taking quick strategic decisions based on richer data.

1. Introduction and motivation

According to the 2011 Gartner Group report (Gartner Group report, 2011), worldwide information volume is growing at a minimum rate of 59% annually. Thus, the available information for a company is progressively increasing. This information is accessible from any computer, and comes from both structured and unstructured sources of data. The structured data is predetermined, well defined, and usually managed by traditional Business Intelligence (BI) applications, based on a Data Warehouse (DW), which is a repository of historical data gathered from the heterogeneous operational databases of an organization (Inmon, 2005; Kimball and Ross, 2002).

The main benefit of a DW system is that it provides a common data model for all the company data of interest regardless of their source, in order to facilitate the report and analysis of the internal data of an organization. DW structures the data in terms of Facts and Dimensions. A fact is the center of the analysis, and typically represents a business activity. For example, gene effects on a trait could be considered a fact. In order to evaluate the performance of the activity, a fact includes fact attributes, also called measures, which are represented as cells in an OLAP cube. In our example, the influence degree of the gene could be a measure. Furthermore, a fact can be analyzed from different perspectives, which constitute dimensions that provide contextual information for the analysis, and are represented as axis in an OLAP cube. For example, we could analyze gene effects by looking at the trait associated or at the plant family whose traits are being studied. Moreover, each dimension may have its own structure, allowing us to analyze the fact at different levels of aggregation, and establishing relationships between levels. For example, the hierarchy for the species dimension could be species (lowest level), which can be aggregated into families, and families can be aggregated into classes.

However, there is a wide consensus in that the internal data of organizations to take right decisions is not enough, even more in current highly dynamic and changing markets where information from competitors and clients/users is extremely relevant for these decisions. Thus, the main disadvantage of traditional DW architectures is that they cannot deal with unstructured data (Rieger et al., 2000). Currently, these unstructured data are of a high relevance in order to be able to make more accurate decisions, since the BI applications would empower their functionality by considering both data from inside the company (e.g. the reports or emails from the staff stored in the company intranet) and outside (e.g. the Webs of the company competitors) (Trujillo and Maté, 2012).

For example, let us consider a scenario where a plant breeder enterprise needs to enrich its genetic information (internal structured DW)

E-mail addresses: jperal@dlsi.ua.es (J. Peral), antonio@dlsi.ua.es (A. Ferrández), edg12@alu.ua.es (E. De Gregorio), jtrujillo@dlsi.ua.es (J. Trujillo), amate@dlsi.ua.es (A. Maté), ljfp1@alu.ua.es (L.J. Ferrández).
[1] Tel.: +34 96 590 3400.

http://dx.doi.org/10.1016/j.ecoinf.2014.05.003
1574-9541/

with recently discovered genes related to specific phenotypic traits (external unstructured data obtained from the Web) in order to choose the desired parentals for breeding programs. The plant breeder enterprise will find that there are an overwhelming number of scientific publications in Life Sciences, specifically in Genetics and Biotechnology (Matos et al., 2010). According to the Medline database, about 2 scientific papers in Life Sciences are incorporated per minute, and there are more than 1000 journals in Biology currently published worldwide.[2] Moreover, increasing bioinformatics work has resulted in a large amount of information stored in Biological Databases (e.g. UniProt, RCSB Protein Data Bank, GenBank, and CEREALAB) that remains uninterpreted. For these reasons, the current rate of scientific publications requires search strategies that allow us to extract biological information easily and efficiently (Altman et al., 2008; Jensen et al., 2006).

So far, many attempts to integrate a corporate DW of structured data with unstructured data have been reported (Badia, 2006; Henrich and Morgenroth, 2003; McCabe et al., 2000; Pérez-Martínez, 2007; Pérez-Martínez et al., 2008a,b, 2009; Priebe and Pernul, 2003a,b; Qu et al., 2007; Rieger et al., 2000). They are mainly based on systems that use Natural Language Processing (NLP) techniques to access the unstructured data in order to extract the relevant information of them but they do not reach a full integration of structured and unstructured data as our proposal manages.

In this paper, we present a framework which combines traditional DW architectures with Question Answering (QA) systems. QA systems represent the potential future of Web search engines because QA returns specific answers as well as documents. It supposes the combination of Information Retrieval (IR) and Information Extraction (IE) techniques. IR is the activity of obtaining information resources relevant to an information need from a collection of information resources. This activity is currently quite popularized by the Web search engines as Google. On the other hand, IE is the task of automatically extracting specific structured information from unstructured and/or semi-structured machine-readable documents. A typical application of IE is to scan a set of documents written in a natural language and populate a database with the information extracted (e.g. the name of products and their prices).

We start with a question or query in Natural Language (NL) posed by the decision maker, who also identifies the sources where to search the required information. We distinguish between *queries* and *questions* in order to highlight that a query refers to a request of data to the DW system, whereas a question requests data to the QA system. The former is likely to be much more rich and complex than simple questions, which may force to divide the query into several questions. The questions are analyzed by the Distributor/Integrator service of the framework and are passed to the corresponding node (e.g. the QA node to access external data or the DW node to access internal data). Then, each node processes the question in an autonomous way on its corresponding sources. Once the system receives all the results from the nodes, like internal DW, Web services or API's, it is capable of integrating and showing a dashboard to the user that allows him/her to take the right decision. Finally, let us add that we also take advantage of our unique well-checked hybrid method for building data warehouses. Our method starts by analyzing user requirements by means of interviews. Then, each requirement is checked against the data sources to ensure that the necessary data exists. Afterwards, the data warehouse is built in order to support queries from the presented approach. Therefore, we can ensure that the query posed on the DW node will return the correct data required by the decision maker (Mazón and Trujillo, 2008; Mazón et al., 2007).

The paper is structured as follows. In Section 2, we summarize the most relevant related work regarding combining traditional DWs with unstructured data. In Section 3, we introduce our framework for analyzing and integrating different data sources into a common dashboard. In Section 4, and in order to clarify our proposal, we introduce the case study that will be evaluated in Section 5, where we provide detail on the evaluation of the application of our proposal. We conclude the paper with the summary of our main contributions and our directions for future works.

2. Related work

Several attempts to integrate search of structured and unstructured data have arisen, in which the structured data is handled by a DW or a DB system, and the unstructured data is handled by an IR, IE or QA system. This integration should meet certain requirements in order to adequately provide integrated information for the users. These requirements include the detection of matching points between the structured and unstructured data, the integration of the results obtained by each system, and the preservation of high quality sources of information, i.e. the DW. In other words, the extraction of structured data from unstructured data is required in order to provide links with similar structured data. In this way, the user can represent and integrate the unstructured data in all the possible dimensions and hierarchies that a DW cube can contain. As a result the information returned by both systems could be perfectly integrated and analyzed together. However, these data cannot be mixed, as that would result in potential decrease of the accuracy of the data stored.

Regarding the connection between a DW system and an IR system, the work presented in Rieger et al. (2000) and Henrich and Morgenroth (2003) can be cited. However, as it is claimed in the work presented in McCabe et al. (2000), those efforts do not take advantage of the hierarchical nature of structured data nor of classification hierarchies in the text, so they implement an IR system based on a multidimensional database. Specifically, they focus on the use of OLAP techniques as an approach to multidimensional IR, where the document collection is categorized by location and time. In this way, they can handle more complex queries, like retrieving the documents with the term "financial crisis" published during the first quarter of 1998 in New York, and then drilling down to obtain those documents published in July 1998.

In Priebe and Pernul (2003a,b), the authors propose an architecture that introduces a communication bus where both systems publish their output. Each system picks up this output and uses it to show related information. For example, the query context of a DW access is used by an IR system in order to provide the user with related documents found in the organization's document management system. In order to solve the problem of the heterogeneity of both systems, they propose to use ontological concept mapping (e.g. the DW system uses "owner" for what is called "author" within the document metadata). They use an ontology for the integration, but it is only oriented to communicate both applications in enterprise knowledge portals. In this way, they handle queries like "sales of certain audio electronics products within the four quarters of 1998".

In LaBrie and St. Louis (2005), an alternative mechanism for IR ("dynamic hierarchies" based upon a recognition paradigm) that overcomes many of the limitations inherent in traditional keyword searching is proposed. This IR approach was used in BI applications but no integration between both applications was made.

In Pérez-Martínez (2007) and Pérez-Martínez et al. (2008a), the authors provide a framework for the integration of a corporate warehouse of structured data with a warehouse of text-rich XML documents, resulting in what authors call a contextualized warehouse. These works are based on applying IR techniques to select the context of analysis from the document warehouses. In Pérez-Martínez et al. (2009), the authors formalize a multidimensional model containing a new dimension for the returned documents. To the best of our knowledge, these papers are the most complete ones in combining and considering structured and unstructured data in a common DW architecture.

[2] http://www.e-journals.org/botany/ (visited on 24th of March, 2013).