Journal of Universal Computer Science, vol. 21, no. 6 (2015), 757-776
submitted: 30/11/14, accepted: 20/2/15, appeared: 1/6/15 © J.UCS

Design and Implementation of an Extended Corporate CRM Database System with Big Data Analytical Functionalities

Ana I. Torre-Bastida, Esther Villar-Rodriguez,
Sergio Gil-Lopez and Javier Del Ser
(TECNALIA. OPTIMA Unit, E-48160 Derio, Spain
{isabel.torre, esther.villar, sergio.gil, javier.delser}@tecnalia.com)

Abstract: The amount of open information available on-line from heterogeneous sources and domains is growing at an extremely fast pace, and constitutes an important knowledge base for the consideration of industries and companies. In this context, two relevant data providers can be highlighted: the "Linked Open Data" (LOD) and "Social Media" (SM) paradigms. The fusion of these data sources – structured the former, and raw data the latter –, along with the information contained in structured corporate databases within the organizations themselves, may unveil significant business opportunities and competitive advantage to those who are able to understand and leverage their value. In this paper, we present two complementary use cases, illustrating the potential of using the open data in the business domain. The first represents the creation of an existing and potential customer knowledge base, exploiting social and linked open data based on which any given organization might infer valuable information as a support for decision making. The second focuses on the classification of organizations and enterprises aiming at detecting potential competitors and/or allies via the analysis of the conceptual similarity between their participated projects. To this end, a solution based on the synergy of Big Data and semantic technologies will be designed and developed. The first will be used to implement the tasks of collection, data fusion and classification supported by natural language processing (NLP) techniques, whereas the latter will deal with semantic aggregation, persistence, reasoning and information retrieval, as well as with the triggering of alerts based on the semantized information.

Key Words: Big Data, Social Media, Linked Open Data, business intelligence, information fusion, ontology management, information modeling
Category: H.3.3, I.5, E.1, J.0

1 Introduction and Motivation

Nowadays most organizations and industries collect huge amounts of valuable information towards monitoring, analyzing and improving the performance of their business operations, decision making policies, development plans and long-term strategies. This trend has given rise to the so-called business intelligence concept [Moss and Atre 1998] (BI), which refers to the set of procedures and key technologies aimed at inferring business-valued knowledge from the data generated by the company and the contextual framework around it (e.g. related external factors and information sources), with the ultimate objective of 1) optimizing daily operations (operational BI); 2) designing medium-term focused initiatives (tactical BI); or 3) outlining long-term business goals (strategic BI).

Unfortunately, the most often encountered problem by BI systems rests on the high heterogeneity and dimensionality of the available data, which unchains a severe computational inefficiency in subsequent knowledge extraction approaches. Such processing issues associated with the data volume and heterogeneity have been lately embraced under the Big Data paradigm, which refers to all such scenarios where the velocity, volume and variety of the collected data go beyond the scales managed by traditional database management and mining tools. Such a quantum leap on the characteristics of the data is enabled by the upsurge of new data sources and its progressively higher involvement and added value in Big Data scenarios, among which the Linked Open Data (LOD, [Bizer et al. 1998]) and Social Media are lately gaining momentum in the research community. On the one hand, Social Media is considered as a context-rich relevant information source not only from the social perspective itself, but also as a decisional driver for organizations whose operations and/or products are strongly influenced by social interactions, user-generated content and behavioral patterns. This is the reason why business executives find in Social Media valuable data that must be captured, exploited and incorporated in their decision-making procedures. On the other hand, LOD provides a global, open structured informational repository with high semantic value that permits not only freely extracting information related to the company, but also discovering semantically defined relationships among connected entities. This being said, this research work postulates the combination of Social Media (such as those found in Facebook and Twitter) and LOD as a semantically rich global information source with potentiality to generate a significant added value in business operations and decisional processes.

This hypothesis is firmly buttressed by intense research being currently conducted towards exploring such benefits in brand recognition [Hoffman and Fodor 2010], competitive intelligence [Vuori 2011] or bench marketing [Bingham and Conner 2010]. Nevertheless, scarce studies have been carried out regarding customer relationship management by identifying potential customers or improving and enriching the stored information about the client portfolio of the company at hand. Likewise, there are very few contributions to the literature addressing competition analysis based on public information; the existing ones (e.g. [He, Zha and Li 2013]) focus exclusively on data repositories of a certain class (e.g. Social Media), hence discarding its combination with related information sources of different characteristics. Furthermore, from a technical perspective the heterogeneity of the data coming from these sources comes along with non-standard, unreliable schemas that require a significant human effort to extract, format and assimilate knowledge. Indeed, the removal of noisy content (understood as the process of filtering out data due to their semantic irrelevance or lack of integrity) is mandatory before any knowledge inference stage. Another related issue inheres in how to merge these datasets with traditional business core systems such as relational database management servers or corporative repositories [Pham and Jung 2014].

This paper aims at stepping further beyond the issues identified above by proposing a novel Client Relationship Management (CRM) system with extended analytical functionalities (information aggregation/fusion and knowledge discovery) applied over a semantically aggregated information database. Technically speaking, our proposed setup follows a semantic aggregation approach that allows retrieving, combining and analyzing information from emerging datasets (in particular, Social Media and LOD) with other corporate databases. This embodies an integral, universal platform that implements diverse BI functionalities which, without loss of generality, will be exemplified within this manuscript by 1) the retrieval of extended information through the customer database; and 2) the analysis of competitors/allies based on the cosine similarity of published projects and initiatives participated by every client within the CRM database. This research work will show how semantic tools and Big Data technologies for information collection and aggregation can be hybridized so as to yield BI insights leveraging not only corporate datasets, but also the information contained in LOD and social platforms.

The rest of this paper is structured as follows: for the sake of completeness, Section 2 and subsections therein briefly survey the main concepts related to Social Media, LOD, Big Data technologies and the state of the art about their use for BI scenarios. The proposed scheme and its core processing steps are described in detail in Section 3, along with the analytical functionalities that will be put to practice. Next, Section 4 discusses performance metrics and the produced outcome of the designed system when applied to the aforementioned use case. Finally Section 5 draws some concluding remarks and sketches future research lines related to this work.

2 Background and Related Work

The system proposed in this paper builds upon the combination of Big Data and semantic technologies, which are surveyed within this section from a bottom-up perspective.

2.1 Social Media and Linked Open Data

To begin with, two different classes of datasets are considered: Social Media and Open Data/Linked Open Data, the latter being semantically structured as opposed to the unstructured nature of the former. Indeed, Social Media platforms nowadays store unprecedented amounts of raw yet valuable data due to the fact that the user role has shifted sharply from being a mere information consumer to a data provider. Interestingly, Social Media is defined in [Jung 2012] as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content". Thus, it can be considered as a contextually rich source of knowledge with business-wise relevance in sectors such as retail, commerce, bank and health, among many others [Lo 2008].

On the other hand, Open Data stands up for the idea that certain data should be freely available to be used and republished at will, without restrictions from copyright, patents or other control mechanisms [Auer et al. 2007]. A special instance of this concept is the Linked Data paradigm, which refers to a set of best practices for publishing and interlinking structured data on the Web. With this definition, Bizer et al. [Bizer, Heath and Berners-Lee 2009] defined the linked data paradigm and provided a mechanism to build the Web of Data, founding on the basis of semantic Web technologies and being considered as a simplified version of the Semantic Web. The data model for representing interlinked data is RDF (Resource Description Framework, [Hoffman and Fodor 2010]), where data is represented as node-and-edge-labeled directed graphs. Some published Linked Data datasets contain billions of triples, whose cardinality is steadily increasing to yield the so-called Linked Data Cloud, i.e. a group of accessible data sets on the Web containing links pointing at other Linked Data sets. The Linked Data principles are enumerated as follows: 1) Linked Data uses URIs (Uniform Resource Identifiers) as names for things; 2) Linked Data uses HTTP URIs so that people can look up those names; 3) when a user looks up an URI, Linked Data provides useful information using the standards RDF and SPARQL (SPARQL Protocol and RDF Query Language); 4) Linked Data includes links to other URIs, so that they can discover more things. This being said, LOD refers to the combination of Open Data and Linked Data, i.e. semantically defined repositories of open data.

2.2 Big Data Technologies

In the last couple of years a research trend has crystallized within the computer science community towards the development of new data storage, retrieval and processing technologies that allow efficiently analyzing very large and diverse amounts of structured and unstructured data. In this context, one can distinguish three classifications of technologies, depending on the task that they accomplished: Real-time analysis, batch analysis and storage. As will be later shown in the paper, elements from this threefold classification will be included in the design of the proposed system.

On the one hand, data streams are monitored and processed in real time for detecting patterns by virtue of Complex Event Processing (CEP) techniques. A CEP approach can be understood as a *backwards* database. In other words, in a common database the data is stored and queries are subsequently launched once the storage has been fully committed; however, in CEP setups queries are first implemented and collected before the data is released. The flow of information is non-persistent and is stored in memory during a time window defined a priori within the queries [Gonzalez and Ortiz 2014]. There are multiple alternatives to implement CEP functionalities such as Esper [Esper 2014], WSO2, Aleri, Software AG and Yarn, among many others [Carvalho et al. 2013]. In general, these systems provide specific methods to define and incorporate data flows, and define specific language to fully describe complex events. In regards to the proposed CRM platform, the functionalities provided by CEP engines are strongly matched to the processing requirements associated to the inclusion of Social Media, which essentially comprise a set of preliminary filters and classifications for the data streams derived from data providers such as Facebook and Twitter.

On the other hand, in what relates to batch analysis new models of parallelization have emerged in the last decade. This is the case of the Map-Reduce framework [Dean and Ghemawat 2004], which is a programming model and an associated parallel data processing framework aimed at analyzing large volumes of data on large clusters based on the *divide-and-conquer* principle. A Map-Reduce program is called a job, and is composed of Map and Reduce tasks. Summarizing, a job takes a set of key/value pairs as an input and produces a set of key/value pairs as an output. A Map-Reduce program conceives the computation as two distributable functions:

1. `Map`, which converts the input from the sources in key/value pairs by filtering and sorting the input data based on a certain property of the data themselves.

2. `Reduce`, which implements an aggregation or summarization of the outputs from the set of `Map` tasks. In this task the input key/value pairs are sorted and clustered by the key.

In this work the Map-Reduce programming model is adopted to alleviate the computational cost of certain processing stages, such as the disambiguation of entities or the semantic inference over the aggregated database.

Regarding Big Data storage technologies, NOSQL databases have become the most widely used solution in practical setups. The persistence mechanism that will be utilized in the scope of this article relies on a Cassandra cluster that has been reported to feature good distribution and scalability properties [Lakshman and Malik 2009]. Other repositories such as Relational Database Management Systems (RDBMS) or triple stores [Rohloff et al. 2007] are not suitable for the envisaged application of the CRM system due to their low scalability when handling data of high volume and heterogeneity.

2.3 Related Work

In business intelligence – especially in the area of competitive information – data gathering process may involve a large number of technologies and design strategies, which have unchained an intense research activity in the related literature. Among them, the so called Social Big Data – conceived as the extraction of knowledge from Social Media – has been broadly adopted in data analysis systems; in this matter, studies such as [Rappaport 2011] highlight the essential role Big Data can take when exploiting Social Media in the field of business intelligence, which is further argued by presenting practical cases where Social Media data is shown to yield measurable business advantages. In this context, the work in [Dey et al 2011] presents a preliminary study focused on using text mining techniques for collaborative intelligent information gathering. The main difference with the approach here presented lies on the used techniques (our work resorts to semantic fusion and Big Data technologies) and the application domain, which in our case gravitates on the construction of an intelligent knowledge base of customers.

Another work related to our scheme is the one presented in [Shroff 2011], which describes a framework for the fusion of business intelligence in various industrial sectors such as manufacturing, retail or insurance. Once again and unlike our proposal, this contribution hinges on its global and general case-based implementation without concentrating on a specific problem. Furthermore, techniques used in this reference reduce to the so-called blackboard architecture and locality sensitive hashing, which are far from the semantic fusion approach considered in this paper.

Another interesting and more specific contribution is [Agichtein et al. 2008], which elaborates on a high-quality Social Media information gathering scheme, but only managing data posted in the *Yahoo!Answers* social platform. Other investigations also discuss the advantages of data fusion on information collected from Social Media as in [Cui et al. 2010], where multiple features in the Social Media environment (textual, visual and user information) are fused for its subsequent use on a retrieval algorithm for large Social Media data (*flickr*). Likewise, in [Lovett et al. 2010] a use case of a shared on-line calendar is presented and enhanced with events generated by user social networks and location data using fusion techniques. Furthermore, we highlight the proposal in [Jung 2013] since it is, to the best of our knowledge, the only reference found in the literature utilizing semantic fusion techniques. However, its purpose is certainly out of the scope of business intelligence and fails to provide enough technical details on its methodology and assessment for a fair comparison with our proposed system. Finally, a survey about technologies, applications and challenges of linked data mash-ups has been reported in [Hanh et al. 2014]. In this reference a use case close to our approach employs semantic web pipes for integration, as opposed to our approach which utilizes ontology mapping/alignment techniques to the same end. Furthermore, this reference only considers freebase datasets and does not hence exploit any Social Media.

When turning the scope of this literature survey to the specific functionalities assessed in this paper, there are very few articles dealing with the unsupervised classification of enterprises. The most recent work is [He, Zha and Li 2013], where text mining is applied to analyze unstructured text content posted on the Facebook and Twitter profiles of three large pizza chains. The ultimate aim in this reference is to monitor and analyze not only the customer-generated content on the own social site of every company, but also the textual information posted on the sites of their competing counterparts. Competition is analyzed among a predefined set of companies under study based on raw data extracted from Social Media. In contrast, our research work is based on discovering similarity patterns on semantized open data via text mining and unsupervised learning techniques, to which Social Media can be incorporated in a straightforward manner. To this end we will contextualize this functionality as a technical means to discover potential allied/competing organizations based on their participation in public funding programs, which is a critical BI requirement for research institutions and institutes. This target contextual application is related to [Lozano. Duch and Arenas 2006], where community detection algorithms are applied to a graph representations of the interactions among institutions and companies as a function of their participation in the 6th Framework Program of the European Commission. However, this reference defines the similarity relationship among organizations as the number of projects in which they coincide. Our work takes a step further by semantically defining such a similarity metric based on the description of the projects where each organization participates.

3 System Overview and Architecture

The architectural diagram of the proposed system and its compounding modules are depicted in Figure 1. Each of such modules is responsible for performing all functionalities and tasks required to implement the two use cases in study: 1) the initialization of the CRM database and the retrieval of extended about each of its entries from the LOD and Social Media; and 2) the discovery of close competing/allied organizations in terms of the cosine similarity of their participated projects.

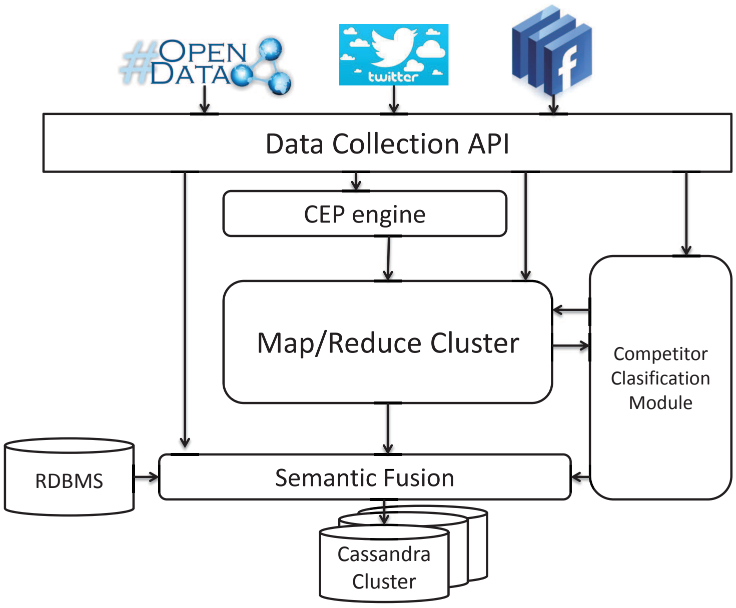

Figure 1: Overview of the proposed system architecture.

First of all, the module labeled as *Data Collection API* (Application Programming Interface) is in charge of collecting external information such as tweets related to the company at hand, customer feedback and comments or business-related open data, independently of the use case under study. Subsequently, the CEP engine extracts meaningful information from the collected unstructured data based on a set of filters, whose output is set as the input to a Map-Reduce cluster that allows efficiently refining and analyzing the captured data (e.g. the unsupervised discovery of competitors or allies that will be later analyzed as an exemplifying functionality of the proposed platform). This refined information is fused and merged with structured data (e.g. semantized information extracted from LOD) and corporate data coming from existing RDBMS or external sources. This is accomplished by means of a semantic fusion module which stores the semantically enriched aggregated information in a Cassandra cluster by following the RDF embedding procedure presented in [Illarramendi et al. 2011]. Triples that conform our semantic data model are distributed over the Cassandra cluster nodes by arranging two structures organized in columns composed by different fields ("composite-columns"). As shown in Figure 2, these structures represent triples `(subject,predicate,object)` where all resources can be a variable. Over this semantic model heterogeneous data retrieval, inference and analysis actions can be performed.

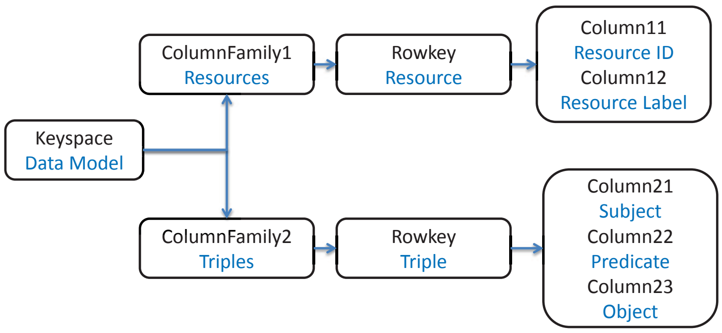

Figure 2: Schematic comparison of the Cassandra data model (black) versus semantic data model (gray) when used to store RDF data.

Let us delve into the different modules of the proposed system. Information is collected from two different sources: Social Media and Open Data/Linked Open Data. Specifically for the former, Facebook posts and comments from specific user IDs are captured along with Twitter feeds containing certain keywords or being generated by specific user IDs. Such keywords are extracted via Term Frequency-Inverse Document Frequency (TF/IDF) from corporate documents and website of the company being analyzed. To this end, streaming APIs supplied by such social networks have been utilized. As for Open Data, the proposed architecture accommodates any source of open information, but for the previously introduced specific application open data from the European Union Open Data Portal [EU Open Data Portal 2014] will be considered. This portal is a single point of access to a growing range of data produced by the institutions and other bodies of the European Union. Furthermore, Linked Open Data will be also integrated as another information repository for the system. In this regard, there are several datasets related to the business domain – such as DBpedia, CrunchBase or Freebase – which can be queried by the SPARQL query language or web services.

From these datasets, structured information about customers is obtained, which is latter mapped to the semantic model of our system.

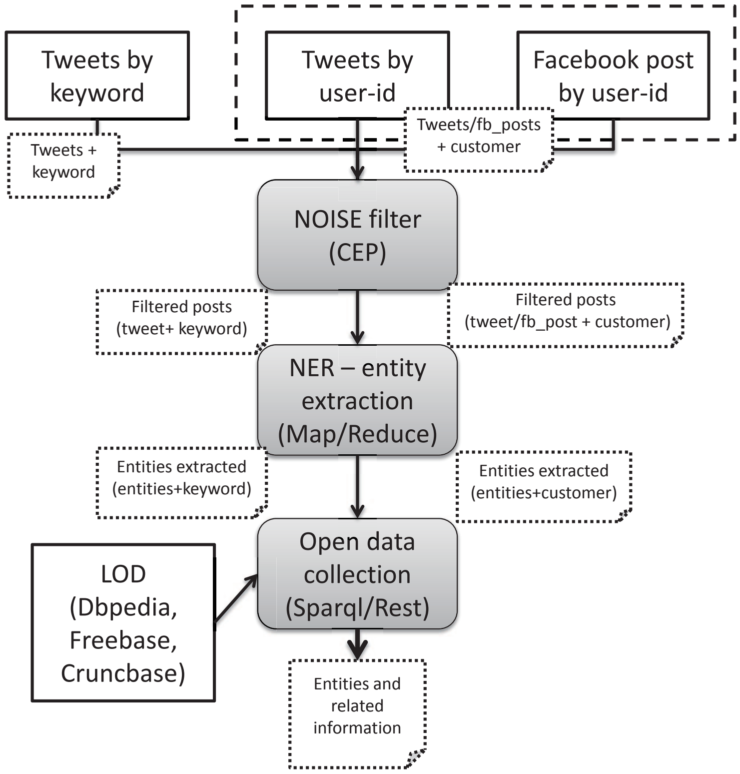

Figure 3: Data collection process flow.

The data collection is technically detailed in Figure 3. This task is composed by three sub-processes: data collection and noise reduction, extraction of disambiguated entities and harvest of related entities' information available as open data. In a first step, the different Social Media data streams are captured by using the aforementioned APIs. Next, posts from Twitter or Facebook are preprocessed by using the Freeling API to carry out the language analysis, calculating their corresponding synsets (i.e. a group of data elements that are considered as semantically equivalent, represented by an identifier). The collection of pairs formed by each post and its synsets are the input events to a set of rules that allow deducing whether the post (tweet/comment) can be considered of relevance for the business domain. This stage is what has been coined as *noise* filter. This filter, which brings down to a set of rules, is implemented by means of a CEP engine fed with rules built upon a set of synsets that represent a descriptive context of the business domain, e.g. `[concept:business; synsets:`

`08056231,08058098]`. If any of the synsets belonging to ongoing posts can be matched to any one of the synsets that compose the context, the rule is activated and the post is filtered as relevant for the business domain. Otherwise, the post is discarded since it is assumed that its content is *noise* in regards to the sector at hand.

Once posts have gone through the noise filter, the result will be deemed valuable since it is likely to provide meaningful information about the domain, which is then fed to the sub-process in charge of entity extraction. Named Entity Recognition (NER) refers to the module or function used for detecting any kind of entities such as cities, organizations and people, and is mostly utilized by text and language processing as a contributor for semantic information. In our case, filtered posts may contain any named entity corresponding to an already existing customer, a potential client or even a competitor working in the same market sectors. On this purpose, the Daedalus Topic Extraction API has been used and integrated it on a Map-Reduce framework to parallelize the algorithm responsible for extracting entities. The output obtained from the Map-Reduce job is a set of entities grouped by post.

Finally, for each of the previously extracted entities, we will collect the information available in the Linked Open Data sets (Freebase, DBpedia) and other open data repositories such as CrunchBase. This information will be merged and aggregated to the existing data from corporation relational databases, with the final aim of feeding the semantic model.

3.1 Semantic Fusion: Aggregation, Model and Interlinking

The semantic aggregation process has two main goals: to improve the existing information for customers of the organization and to discover new potential customers. The entire process is detailed in Figure 4. First of all, a classification process is applied to each post to determine whether its contents relate to any entity existing in the semantic data model. Depending on the result of this classification the system follows two different alternative flows. In the positive case, the semantic model is updated with the new information about customer and its partnerships/relationships. Otherwise, the data gathered from the Linked Open Data Cloud is mapped into a new instance within the semantic model. These processes are supported by a set of previously computed semantic links between our model and the LOD datasets vocabularies, which are calculated following the ontology alignment process proposed in [Torre-Bastida et al. 2014].

With regard to the definition of our model schema, well-known semantic vocabularies will be reused, to promote interoperability with other RDF repositories or datasets. Our ontology model is based on the combination of the `schema.org` ontology along with that used in DBpedia and vocabularies such as SKOS [Miles et al. 2005] to specify semantic relationships and links. New classes or properties are also modeled in the case that existing vocabularies do not provide their definition.

Finally, the new instances of the semantic data model are stored in the selected Cassandra NOSQL cluster database. For this task, we have implemented an specific

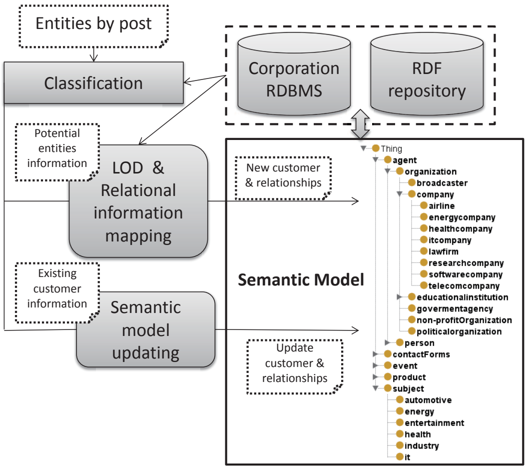

Figure 4: Semantic data aggregation.

RDF storage API over the Cassandra Client APIs and relying on CummulusRDF work [Ladwig and Harth 2011]. Given the size and growth rate of the data to be handled, we selected this type of storage rather than other RDF repositories due to their high scalability and fault tolerance.

3.2 Information Retrieval, Inference and Alert Generation

Once the information has been converted to RDF format following our semantic model and stored in the NOSQL database, several added-value operations can be implemented over the stored data:

- Information retrieval: In our case SPARQL – the current W3C recommendation for querying RDF data – is selected to allow the user to perform selective queries. In particular, the JENA API and the FUSEKI SPARQL server [Grobe 2009] have been chosen for implementing this module. Figure 6 in Section 4 illustrates an example of this type of queries.
- Inference: Based on the information stored in the semantic repository RDFS/OWL inference can be performed aimed at discovering new hidden relationships among different organizations. This task is accomplished using the technique presented in [Urbani et al. 2009], where RDFS/OWL inference is implemented on a Map-Reduce parallelized framework. However, it is important to emphasize that in our proposed system the connection between the inference mechanism and the persistence module does not require the use of any intermediate files or processes, thus the inferred data is persisted over the NoSQL database. Additionally, the system

also enables the definition of specific business-related semantic rules using the Semantic Web Rule (SWRL), OWL and RuleML languages.

– Alert generation: Finally, an alert generation module is included for monitoring data and triggering events that indicate that a number of conditions specified in the alert have been fulfilled. For its implementation a listener is utilized during the loading and inference process, which allows detecting whether alert conditions have been met.

3.3 Discovery of Semantic Relationships among CRM entities

Research companies usually undergo strong competition in public funding programs as a means to economically support their research lines. In this sector the goal pursued by the board of governors, CTO's and decision makers alike is to find new research fields in which disruptive technologies can be applied with significant added value and a measurable business impact. To this end, funds are usually captured within public programs fostered by governmental and public institutions such as the European Union (EU), who periodically arranges competitive calls for project proposals as a major supporting instrument and catalyst for research and innovation.

Focusing on this envisaged context, a use case involving companies and institutions having participated in projects and initiatives in the past 7th EU Framework Program (FP7) has been designed towards validating one of the advanced analytical functionalities implemented in the proposed BI system. The purpose is to discover similarities among such companies based on the description of the projects where they have been involved, in such a way that a unsupervised learning model subsequently unveils semantically close organizations that may correspond to potential competitors or collaborators.

To this end, information has been collected from [EU Open Data Portal 2014] corresponding to the projects funded by the European Union under the FP7 program from 2007 to 2013. For each granted project references, acronyms, dates, funding, programs, participant countries, subjects, abstracts and objectives are provided in the dataset. The dataset is in CSV format, which is parsed so as to be readable in our proposed platform. As summarized in Algorithm 1, the definition of a cosine similarity among organizations starts by computing the TD/IDF measure for the abstracts of all granted projects which permits representing them as multidimensional vectors with as many dimensions as words found in the abstracts. Bigrams and trigrams are also computed and included in the final vector space in order to avoid any lose of meaningfulness when breaking complex terms or multiwords. Once every project within the dataset is represented by its contents in a vectorial fashion, a representation of every organization participating in the project is created. The Bag of Words (BOW) of every organization once transformed into a vector must satisfy the uniqueness condition for every word. Due to the non-zero probability of encountering identical terms across different projects (which

Algorithm 1 Pseudo-code of the cosine similarity matrix construction process

Require: `URL` of the dataset of FP7 projects granted by the European Commission.
Ensure: A similarity matrix $\mathbf{D}$, whose entry $\mathbf{D}[i,j]$ denotes the cosine distance between partner i and j in the aforementioned database.

Collect the data: `data_dump = COLLECT(URL)`
Parse the data: `lst_projects = PARSE(data_dump)`
Let $N = |$`lst_projects`$|$, i.e. the number of projects in the dataset.
Let P denote the total number of different partners in the dataset.
Let $\mathbf{T}$ represent the dictionary of all tokens found when processing the abstracts of the projects in the dataset. This variable is set empty at the beginning of the loop and filled within the algorithm loop.
Let $P \times |\mathbf{T}|$ matrix $\mathbf{M}$ contain the Bag of Words (BOW) representation of all organizations in the dataset, whose p-th row is composed by the TF/IDF value for every token in $\mathbf{T}$ for organization $p \in \{1,\ldots,P\}$. This variable is initially empty, and will be progressively filled within the algorithm loop.
Let variable $\mathbf{C}$ denote a counter of the number of non-zero TF/IDF values corresponding to a certain token and partner.
for `project` **in** $\{1,\ldots,N\}$ **do**
    Register the partners participating in the project at hand: `lst_partners[project]=PARTNERS(lst_projects[project])`
    Extract the tokens from the project abstract (also considering bigrams, trigrams and multiwords). This step removes common morphological and inflexional endings from words in the abstract via the Porter stemming algorithm [Porter 1980]: `tokens = EXTRACT_TOKENS(lst_projects[project])`
    for `partner` **in** `lst_partners[project]` **do**
        for `token` **in** `tokens` **do**
            Compute the TF/IDF metric of the extracted token from the project with respect to the whole corpus: `tf_idf = TF-IDF(lst_projects[project],ztoken, lst_projects)`
            Add the TF/IDF metric to the BOW entry of $\mathbf{M}$ indexed by `partner` and `token`: $\mathbf{M}$ [`partner,token`] $+=$ `tf_idf`
            Update counter $\mathbf{C}$ if `tf_idf` > 0: $\mathbf{C}$ [`partner,token`] $+= 1$
            Update the dictionary of tokens $\mathbf{T}$ with `token` if `token` $\notin \mathbf{T}$: `ADD_TOKEN(`$\mathbf{T}$`,token)`
        end for
    end for
end for
for `partner` in $\{1,\ldots,P\}$ **and** `token` in $\mathbf{T}$ **do**
    $\mathbf{M}$ [`partner,token`] = $\mathbf{M}$ [`partner,token`]/$\mathbf{C}$ [`partner,token`]
end for
for $\texttt{partner}_1$ in $\{1,\ldots,P\}$ **and** $\texttt{partner}_2$ in $\{\texttt{partner}_1 + 1,\ldots,P\}$ **do**
    Compute cosine similarity metric between $\texttt{partner}_1$ and $\texttt{partner}_2$: `sim = COMPUTE_SIMILARITY(`$\texttt{partner}_1$`,`$\texttt{partner}_2$`,`$\mathbf{M}$`)`
    $\mathbf{D}[\texttt{partner}_1,\texttt{partner}_2] =$ `sim`
end for

becomes higher when handling stemmed tokens), the TF/IDF average is computed so as to guarantee a coherent and solid representation of the processed texts. Next, a cosine similarity matrix is built by considering each pair of companies in the database, which is finally represented in a visually understandable fashion by means of Multidimensional Scaling (MDS [Kruskal and Wish 1978]). This statistical technique allows displaying a distance matrix by placing each object (i.e. organization) in a low-dimensional space such that the between-object distances are preserved as much as possible.

4 Experimental Validation: Use Case

A prototype of the proposed BI system is implemented and deployed over a combined Map-Reduce Cassandra cluster. Tests are programmed in Java 1.6 and executed in a cluster of nodes with Linux Ubuntu 11.10. The cluster is composed by 6 nodes, each with the following features: two processors with 8 Xeon 5645 cores at 2.4 GHz, 8 GB RAM and 250 GB hard disk. Esper 5.1 is the CEP engine used in our testbed.

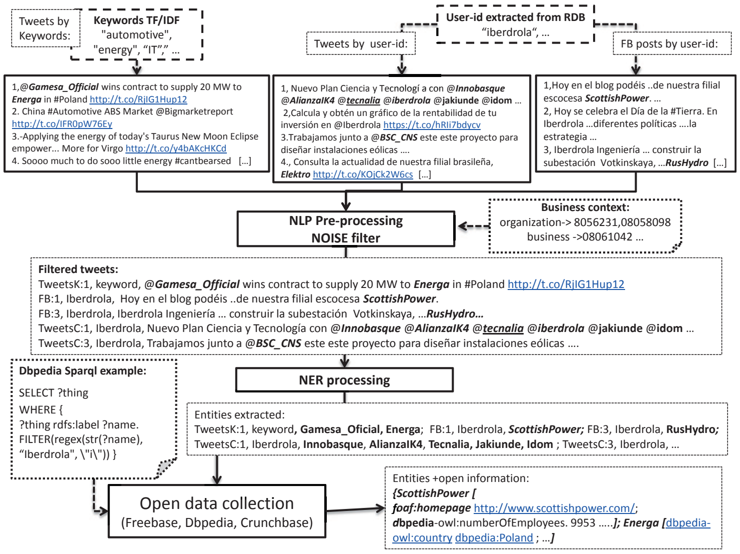

Figure 5: Data collection example.

To begin with, this section describes in detail an illustrative example of the process followed by our system from the data collection to the information retrieval by means of a SPARQL query. The data collection process is shown in Figure 5. The first input is the *real* data retrieved from Twitter and Facebook. Tweets and posts are preprocessed to transform them in synsets as explained in Section 3. These synsets are filtered (a noise filter for irrelevant data) using a macro that consists of a set of concepts representing the business domain (business context in the Figure). These filtered tweets and posts are subject to a named entity recognition procedure aimed at extracting the entities so as to collect from them the information available on the LOD.

The data model and instances generated by the semantic aggregation process and an example of information retrieval using a SPARQL sentence are depicted in Figure 6. As shown in the picture, the query returns a list of all organizations and its related subjects. It is important to notice that although `ScottishPower` is annotated as `energycompany`, this entity is also returned in the query, because in the ontological model (see figure 4) an `energycompany` is categorized as a subclass of `organization`. This unveils one of the advantages of using a semantic model for information retrieval.

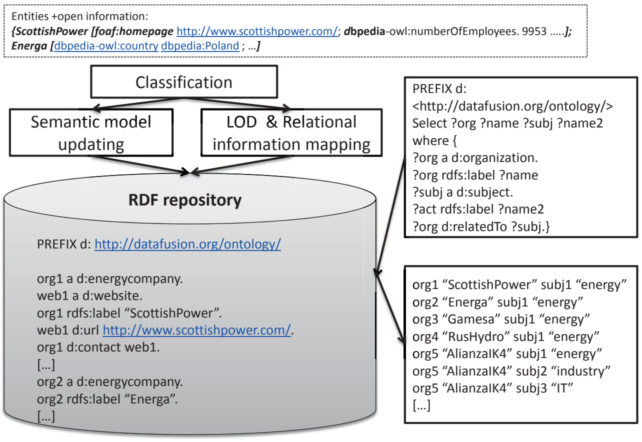

Figure 6: Generated semantic data model and example of an SPARQL query.

The process of discovering similarities among companies is illustrated in Figure 7. We center the scope of the use case on the discovery of companies and organizations having participated in the most similar FP7 projects to those where the Industry and Transport Division of TECNALIA RESEARCH & INNOVATION has been involved. To this end, once the data has been collected, the process previously explained in Algorithm 1 is applied over the abstracts of all projects in the dataset ($N = 25,432$ projects in total), yielding an overall dictionary of more than $200,000$ unique tokens. This BOW consists of `<token:tf-idf>` pairs associated to each partner in the project. It is important to denote that not only individual tokens have been extracted, but also bigrams and trigrams. In the flow diagram we can observe the set of `<token:tf-idf>` values for the project `215007` and associated to the participants `["fundacion tecnalia research & innovation telecom", "teknologian tutkimuskeskus vtt knowledge intensive services", "university of surrey research administration services",...,"alcatel-lucent bell labs france"]`. A list of $P = 15,017$ participants is then constructed upon the set of all tokens and its computed TF/IDF value, which gives rise to the partner vector representation introduced in Section 3.3. Finally we compute the cosine similarity between all pairs of partners.

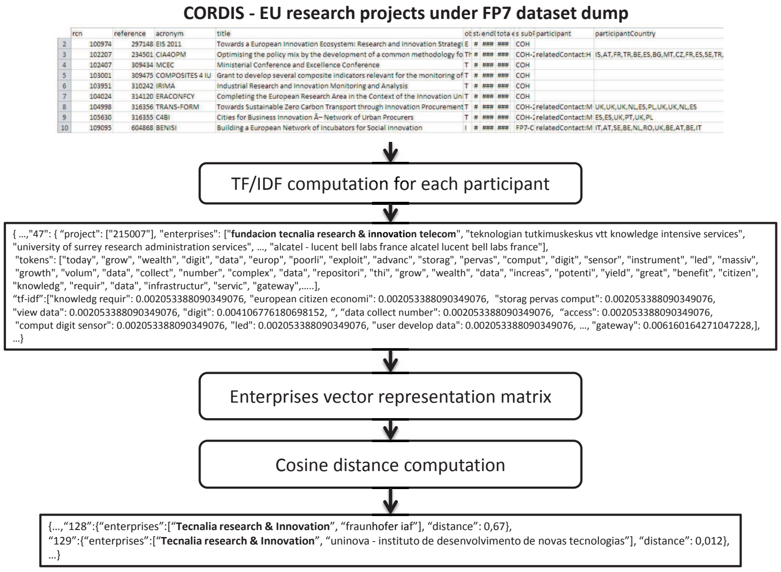

Figure 7: Flow diagram of the similarity computation among companies having participated in the European FP7 funding program.

In Figure 8 a reduced yet insightful set of the 23 FP7 participants closest to TECNALIA's Industry and Transport Division after MDS processing of the similarity matrix **D** is displayed as an example of how any given institution could use this information to discover potential allies or detect competitors. Organizations close to the target in this downscaled space feature a high cosine similarity between their participated projects.

Finally, the performance metrics of each of the processes involved in our system (namely, data collection, semantic aggregation and information retrieval) are presented, along with the processing times taken by queries sent to Cassandra and Map-Reduce inference tasks. Identical metrics are also provided with respect to Algorithm 1, i.e. the discovery of similar organizations in terms of cosine similarity between their participated projects. Table 1 summarizes the obtained average duration for each of the processes carried out by our system, along with complexity indicators (in this case, managed data volume) associated to each one.

As for data collection and filtering (the latter implemented by the CEP engine), the processing is steady and continuous. For this rationale the metrics to determine their performance are based on the number of events or volume of data that can be processed per second. In our case, we obtain a peak practical performance of 106000 events per

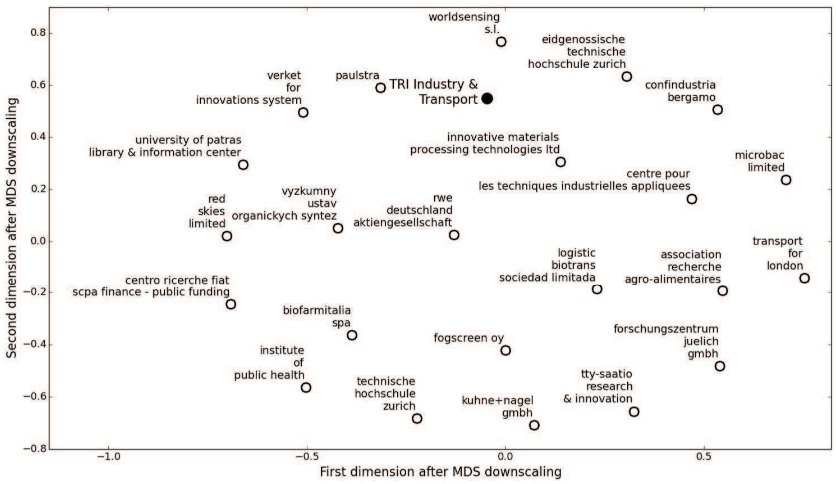

Figure 8: MDS representation of the closest companies to the Industry and Transport Division of TECNALIA in terms of cosine similarity among participated projects.

Process	**Time (hours)**	**Data Size**
Entity Extraction	0.82 hours	1 TB
Semantic fusion	0.64 hours	1 TB
Partner similarity	0.28 hours	1 TB

Table 1: Average execution times of the main processes of the proposed system.

second for the filtering tasks. On the other hand, the results for the query task over our semantic business model persisted into the Cassandra cluster are shown in Table 2. Finally, in regards to the semantic inference task we can state that for a dataset with a size of about 500 million triples (corresponding to about 250 GB of data), the average processing time registered in our practical experiments was 10 hours and 20 minutes, producing an output of roughly double the size of the input dataset (ca. 1 billion of inferred triples).

In summary, two main conclusions can be drawn from the obtained timing scores:

– Competitive and rational execution times: none of our processes takes longer than an hour to run for a fairly high amount of data (i.e. 1 TB). This supports the hypothesis that our system can be utilized not only for tactical and strategic BI, but also on operational BI in companies and organizations not subject to critical near-to-real-time operational decision making.

– High scalability: all technologies used in our approach are known to be highly

Query Type	Execution Time	Query Type	Execution Time
(s p o)	14,003	*(s p ?)*	25,612
(s ? ?)	9,127	*(? p o)*	13,535
(? p ?)	178,294	*(s ? o)*	94,471
(? ? o)	35,345	*(s p ?)* $\bowtie$ *(? p o)*	4,445,588

Table 2: Average execution times for different triple query patterns (microseconds) over the Cassandra cluster.

scalable and allow the system to adjust itself to data growth without significantly jeopardizing their performance times.

5 Concluding Remarks and Future Research Lines

This manuscript has gravitated on the problem of automatically creating and managing a customer database from a novel perspective: semantic aggregation. Input data comes from new sources such as Social Media and Linked Open Data. Furthermore, different modules have been implemented leveraging Big Data (Map-Reduce, Complex Event Processing) and semantic web (RDF repository, reasoner, SWRL) technology stacks. A use case exemplifies the multiple possibilities and potentiality offered to a corporation by our approach, ranging from the discovery of new customers to the knowledge base expansion of traditional clients. This springs profitable advantages in the business domain, where the decision making is a critical process and the collection of customer information is a key factor. The practical utility of our approach is validated by addressing a common BI problem in the research domain: the detection of allies and competitors based on the semantic similarity of their participated projects, which are public exponents of their research activity and interests. To address this task we rely on a similarity analysis between organizations participating in the European FP7 program, whose information is available as Open Data.

Future work will be devoted towards the study of new applications for the proposed BI architecture, as well as towards enlarging the technical scope of the semantic aggregation so as to e.g. include projects referencing entities, business concepts or places and properties that can be matched to relationships within the semantic model. Multilingual processing features will be also considered for their inclusion in the platform.

Acknowledgments

This work has been possible thanks to the support of the computing infrastructure of the i2BASQUE academic network.

References

[Agichtein et al. 2008] Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: "Finding High-quality Content in Social Media"; *International ACM Conference on Web Search and Data Mining* (2008) 183-194.

[Auer et al. 2007] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: "DBpedia: A Nucleus for a Web of Open Data"; *Lecture Notes in Computer Science*, 4825 (2007), 722-735.

[Bernhardt and Vasseur 2007] Bernhardt, T., Vasseur, A.: "Esper: Event Stream Processing and Correlation"; O'Reilly Technical Note (2007).

[Bingham and Conner 2010] Bingham, T., Conner, M.: "The New Social Learning: A Guide to Transforming Organizations through Social Media"; Berrett-Koehler Publishers (2010).

[Bizer, Heath and Berners-Lee 2009] Bizer, C., Heath, T., Berners-Lee, T.: "Linked Data - The Story so far"; *International Journal on Semantic Web and Information Systems*, 5, 3 (2009) 1-22.

[Bizer et al. 1998] Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: "Linked data on the web (LDOW2008)"; *Proceedings of the 17th International Conference on World Wide Web* (2008) 1265-1266.

[Carvalho et al. 2013] Carvalho, O. M. d., Roloff, E., Navaux P. O. A.: "A Survey of the State-of-the-art in Event Processing"; *XI Workshop de Processamento Paralelo e Distribuido* (2013).

[Cui et al. 2010] Cui, B., Tung, A. K., Zhang, C., Zhao, Z.: "Multiple Feature Fusion for Social Media Applications"; *ACM SIGMOD International Conference on Management of Data* (2010) 435-446.

[Dean and Ghemawat 2004] Dean, J., Ghemawat, S.: "MapReduce: Simplified Data Processing on Large Clusters"; *Sixth Symposium on Operating System Design and Implementation* (2004) 137-150.

[Dey et al 2011] Dey, L., Haque, S. M., Khurdiya, A., Shroff, G.: "Acquiring Competitive Intelligence from Social Media"; *Proceedings of the 2011 ACM Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data* (2011) 3.

[Esper 2014] Esper CEP: `http://esper.codehaus.org/`. Accessed in December 2014.

[EU Open Data Portal 2014] European Union Open Data Portal: `https://open-data.europa.eu/`. Accessed in December 2014.

[Gonzalez and Ortiz 2014] Gonzalez, L., Ortiz, G.: "An Event-Driven Integration Platform for Context-Aware Web Services"; *J. of Universal Computer Science*, 20, 8 (2014), 1071-1088.

[Grobe 2009] Grobe, M.: "RDF, Jena, SparQL and the 'Semantic Web"', *37th Annual ACM SIGUCCS Fall Conference* (2009) 131-138.

[Hanh et al. 2014] Hanh, H. H., Cung, T. N. P., Truong, D. K., Hwang, D., Jung, J. J.: "Semantic Information Integration with Linked Data Mashups Approaches"; *International Journal of Distributed Sensor Networks* (2014) Article ID 813875.

[He, Zha and Li 2013] He, W., Zha, S., Li, L.: "Social Media Competitive Analysis and Text Mining: A Case Study in the Pizza Industry"; *International Journal of Information Management* 33, 3 (2013) 464-472.

[Hoffman and Fodor 2010] Hoffman, D. L., Fodor, M.: "Can You Measure the ROI of Your Social Media Marketing?"; *MIT Sloan Management Review* 52, 1 (2010), 41-49.

[Illarramendi et al. 2011] Illarramendi, A., Bermudez, J., Gonzalez, M., Torre, A. I.: "Diseño de un Repositorio RDF basado en Tecnologias NOSQL" (in Spanish); *XVI Jornadas de Ingeniería del Software y Bases de Datos* (2011).

[Jung 2012] Jung, J. J.: "Online Named Entity Recognition Method for Microtexts in Social Networking Services: a Case Study of Twitter"; *Expert Systems with Applications*, 39, 9 (2012) 8066-8070.

[Jung 2013] Jung, J. J.: "Cross-lingual Query Expansion in Multilingual Folksonomies: a Case Study on Flickr,"; *Knowledge-Based Systems*, Vol. 42, pp. 60-67 (2013).

[Jung 2014] Jung, J. J.: "Measuring Trustworthiness of Information Diffusion by Risk Discovery Process in Social Networking Services"; *Quality & Quantity*, 48, 3 (2014) 1325-1336.

[Kruskal and Wish 1978] Kruskal, J. B., Wish, M.: "Multidimensional Scaling"; *Quantitative Applications in the Social Sciences*, 11, Sage Publications (1978).

[Ladwig and Harth 2011] Ladwig, G., Harth, A.: "CumulusRDF: Linked Data Management on Nested Key-Value Stores"; *7th International Workshop on Scalable Semantic Web Knowledge Base Systems* (2011) 30.

[Lakshman and Malik 2009] Lakshman, A., Malik, P.: "Cassandra - A Decentralized Structured Storage System"; *Workshop on Large-Scale Distributed Systems and Middleware* (2009).

[Lo 2008] Lo, B.: "Social Media Analytics in Business Intelligence Applications"; M.Eng. Thesis, Massachusetts Institute of Technology (2008).

[Lovett et al. 2010] Lovett, T., O'Neill, E., Irwin, J., Pollington, D.: "The Calendar as a Sensor: Analysis and Improvement using Data Fusion with Social Networks and Location"; *12th ACM International Conference on Ubiquitous Computing* (2010) 3-12.

[Lozano. Duch and Arenas 2006] Lozano, S., Duch, J., Arenas, A.: "Community Detection in a Large Social Dataset of European Projects"; *SIAM Workshop on Link Analysis, Counter Terrorism and Security* (2006).

[Miles et al. 2005] Miles, A., Matthews, B., Wilson, M., Brickley, D.: "SKOS Core: Simple Knowledge Organisation for the Web"; *International Conference on Dublin Core* (2005).

[Moss and Atre 1998] Moss, L. T., Atre, S.: "Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications"; Addison-Wesley (2003).

[Pham and Jung 2014] Pham, X. H., Jung, J. J.: "Recommendation System Based on Multilingual Entity Matching on Linked Open Data"; *Journal of Intelligent & Fuzzy Systems*, 27, 2 (2014) 589-599.

[Porter 1980] Porter, M. F.: "An Algorithm for Suffix Stripping"; *Program*, 14, 3 (1980) 130-137.

[Rappaport 2011] Rappaport, S. D.: "Listen First!: Turning Social Media Conversations into Business Advantage"; John Wiley and Sons (2011).

[Rohloff et al. 2007] Rohloff, K., Dean, M., Emmons, I., Ryder, D., Sumner, J.: "An Evaluation of Triple-Store Technologies for Large Data Stores"; *OTM Confederated international conference on On the move to meaningful internet systems*, Volume Part II (2007) 1105-1114.

[Shroff 2011] Shroff, G., Agarwal, P., Dey, L.: "Enterprise Information Fusion for Real-time Business Intelligence"; *Proceedings of the 14th International IEEE Conference on Information Fusion* (2011), 1-8.

[Torre-Bastida et al. 2014] Torre-Bastida A. I., Villar-Rodriguez E., Del Ser J., Camacho D., Gonzalez-Rodriguez M.: "On Interlinking Linked Data Sources by using Ontology Matching Techniques and the Map-Reduce Framework"; *Lecture Notes in Computer Science* 8669 (2014) 53-60.

[Urbani et al. 2009] Urbani, J., Kotoulas, S., Oren, E., Van Harmelen, F.: "Scalable Distributed Reasoning using Mapreduce"; *Lecture Notes in Computer Science* 5823 (2009) 634-649.

[Vuori 2011] Vuori, V.: "Social Media Changing the Competitive Intelligence Process: Elicitation of Employees Competitive Knowledge"; Julkaisu-Tampere University of Technology (2011) 1001.