Introducing a New Scalable Data-as-a-Service Cloud Platform for Enriching Traditional Text Mining Techniques by Integrating Ontology Modelling and Natural Language Processing

Alexey Cheptsov[1], Axel Tenschert[1], Paul Schmidt[2], Birte Glimm[3], Mauricio Matthesius[4], Thorsten Liebig[5]

[1] High-Performance Computing Center Stuttgart, Nobelstr. 19,
70569 Stuttgart, Germany
cheptsov@hlrs.de, tenschert@hlrs.de
[2] Institute of the Society for the Promotion of Applied Information Sciences at the Saarland University, Martin-Luther-Str. 14, 66111 Saarbrücken, Germany
Paul.Schmidt@iai-sb.de
[3] University of Ulm, Institute of Artificial Intelligence, 89069 Ulm, Germany
birte.glimm@uni-ulm.de
[4] Objectivity, Inc., 3099 North First Street, Suite 200 San Jose, CA 95134 USA,
mmatthesius@objectivity.com
[5] derivo GmbH, James-Franck-Ring, 89081 Ulm, Germany
liebig@derivo.de

Abstract. A good deal of digital data produced in academia, commerce and industry is made up of a raw, unstructured text, such as Word documents, Excel tables, emails, web pages, etc., which are also often represented in a natural language. An important analytical task in a number of scientific and technological domains is to retrieve information from text data, aiming to get a deeper insight into the content represented by the data in order to obtain some useful, often not explicitly stated knowledge and facts, related to a particular domain of interest. The major challenge is the size, structural complexity, and frequency of the analysed text sets' updates (i.e., the 'big data' aspect), which makes the use of traditional analysis techniques and tools impossible. We introduce an innovative approach to analyse unstructured text data. This allows for improving traditional data mining techniques by adopting algorithms from ontological domain modelling, natural language processing, and machine learning. The technique is inherently designed with parallelism in mind, which allows for high performance on large-scale Cloud computing infrastructures.

Keywords: Data-as-a-Service, Text Mining, Ontology Modelling, Cloud computing.

1 Introduction

The modern IT technologies are increasingly getting data-centric, fostered by the broad availability of data acquisition, collection and storing platforms. The concepts of linked and open data have enabled a principally new dimension of data analysis, which is no longer limited to internal document collections, i.e., "local data", but comprises a number of heterogeneous data sources, in particular from the Web, i.e., "global data". However, existing data processing and analysis technologies are still far from being able to scale to demands of global and, in case of large industrial corporations, even of local data, which makes up the core of the "big data" problem. With regard to this, the design of the current data analysis algorithms requires to be reconsidered in order to enable the scalability to big data demands. The problem has two major aspects: (1) the solid design of current algorithms makes the integration with other techniques that would help increase the analysis quality impossible, and (2) sequential design of the algorithms prevents porting them to parallel computing infrastructures and thus do not fulfil high performance and other QoS user requirements.

Text analytics is an important application area of data mining algorithms. At the same time, it is one of the applications impacted worst by the big data aspect: along on the Web there are several tens of billions web pages, which might be potentially related to the search context. Also on intranets of large and medium size companies there are a huge number of stored documents, e.g., as Word documents, Excel tables, emails, etc., which frequently need to be analyzed. The main purpose of such an analysis is to get a deeper insight into the content of textual information collected in documents in order to obtain some useful, often not explicitly stated knowledge and facts, related to a particular search query. However, automatic knowledge extraction from large text collections, also considering external data sources (e.g., Wikipedia), is a nontrivial and very challenging task. It requires the availability of high performance computing facilities as an "on demand" infrastructure as well as an experimental platform for implementing and later on for running interdisciplinary text analysis algorithms in the way that ensures fulfilment of both quality and efficiency requirements. Unfortunately, existing approaches seem to be neither efficient enough to ensure a proper quality of results (in terms of analysis automation, performance, etc.) nor scalable to catch up with the big data requirements.

In this paper, we suggest and discuss a new technique to develop innovative applications for information retrieval from unstructured text corpora allowing for the determination of causal inter-contextual relations between the analyzed text entities (i.e., documents, web pages, etc.). The main innovation of the technique consists in enabling an interdisciplinary approach that allows traditional data mining algorithms to be enhanced towards incorporation of domain-specific ontology modelling methods as well as template-based, self learning natural language processing technologies in order to ensure a fully automated, reliable, and efficient information retrieval. The technique is inherently designed in a parallel fashion, which allows it to scale well on high performance computing infrastructures, such as Cloud computing. Apart from this, the use of a Cloud also offers a solution for the data privacy problem – whereas the critical (in terms of security and privacy restrictions) data will be processed on

Private Cloud infrastructures, the use of Public Cloud resources will be restricted to the processing of publically available and accessible data.

The paper is organized as follows. Section 2 gives an overview of state-of-the-art R&D activities in the related areas, namely, ontology modelling and semantic techniques, data mining and machine learning, natural human language understanding, and scalable data management. Section 3 introduces our approach to combine the above-mentioned techniques in a text analysis platform. Section 4 discusses the cloud-based architecture of the platform. Section 5 introduces two typical application scenarios that can take advantage of the proposed technique. Section 6 summarizes the main ideas of the paper and draws up a conclusion.

2 Related Work and Progress the Beyond State-of-the-Art

This section provides an overview of the technologies related to the topic of our research as well as an introduction of major enhancements and improvements targeted by our new analysis approach.

2.1 Parallelization and Scale-Up Technique for Data-Centric Processing

The big data problem represents several challenges to the efficient use of existing popular information retrieval platforms for text data such as SMILA [1] or GATE [2], mainly related to scaling up their algorithms to meet rapidly growing data demands. Due to a high complexity of the problem, current approaches to information retrieval focus on very restricted domains. Those challenges have been addressed by partially parallelizing computationally expensive parts of the text processing workflows. Being an obvious solution, this is not sufficient to meet the increasing big data demands. Instead of a top-down parallelization approach, which is currently followed by most of the analysis systems, whereby the parallelization begins at the application level and very rarely reaches the processing algorithms, the basic support of parallel processing should be provided in a bottom-up way. This means the parallelisation should already be included in the design of the underlying processing algorithms, in order to provide extensive support for developing highly parallel and thus efficient applications. In our research, we follow the second way – by reconsidering the basics of text processing algorithms, we aim to develop an innovative platform that will incorporate best practices of the currently available tools and techniques and also consider relatively recent developments in service-oriented data processing (e.g., SOA4ALL project [3]) and large-scale semantic web reasoning (e.g., LarKC project [4]).

In terms of the parallelization technology, Hadoop (a Java-based implementation of MapReduce of Yahoo [5]) is currently enjoying a prominent position in data-centric distributed and parallel computing. However, the bottom-up development approach, followed by us, puts very strict requirements on the platform in order to ensure a “near peak performance” utilization rate of the computing facilities by parallel programs, which Hadoop cannot meet due to the service-oriented and Java-based architecture design. With regard to this, a use of alternative approaches, such as Message-Passing Interface (MPI) or Partitioned Global Address Space (PGAS), is

becoming the latest trend. However, the major challenge for applying these parallelization techniques is constituted by a Java programs' execution environment (Java Virtual Machine, or JVM), which is extensively used for implementing data-centric applications nowadays. Whereas the abstraction to the underlying hardware architecture, offered by JVM, simplifies the application development for the users, the access to special hardware's features, such as cluster's high-bandwidth and low-latency network interconnect (e.g., Infiniband), is only possible through virtualized interfaces, which thus considerably limits the performance.

Nevertheless, the latest advances of such tools as ompiJava [6], which is an implementation of Java bindings for Open MPI [7], along with the promises of ongoing projects, such as JUNIPER [8], which aims to develop an efficient PGAS-based parallelization model for Java applications, offer a promising outlook on the perspectives of using both MPI and PGAS technologies complementary to Hadoop in data-centric parallel applications design.

2.2 Ontology Modelling and Semantics Analysis

Information Retrieval (IR) powered by Ontology Modelling Modelling (OM) has been investigated in several previous research works. For example, the approach introduced in [9] discusses an application from the Business Intelligence domain that allows for automatic extraction of facts by using domain-specific OM. Existing frameworks, such as the above-discussed SMILA, also allow for the integration of ontologies. To the best of our knowledge, the ontologies are, however, mainly used as data stores rather than for supporting the knowledge extraction by inferring implicit knowledge. An approach with a deeper integration of ontologies is proposed in [10], where the results of IR are purely based on OM; the approach does not consider learning methods, though.

Compared to existing alternatives, our approach takes the advantage of OM (along with decision making algorithms) to constitute the core of a model-based, intelligent (i.e., self-learning) knowledge extraction platform. The approach suggests that IR of domain knowledge should be conducted in a semi-automated way by using "seed patterns" (see Section 4 for more details) through machine learning instead of being manually created. Domain knowledge is used to identify antagonisms in the learned facts as well as to revise the ontology. The problem of the ontologically "clashing" and mismatching concepts is resolved by a decision-making system, which refines the ontology based on deductive reasoning algorithms, e.g., the one described in [11]. However, there is a clear need for such algorithms to be optimized and improved in order to address the big data scale, which involves, e.g., the computation of disjoint ontology classes or domains in an incremental and parallel fashion as described in [12].

2.3 Data Mining and Machine Learning

Data Mining (DM) is the technique of automatic information extraction from structured and unstructured machine-readable data sources, often accompanied by

natural language processing algorithms. A prominent example of the domain that has been revolutionized by DM is Business Intelligence, which allows companies to define their market strategy in a tight alignment with the current business goals and company's profile. In this relation, traditional statistical methods as well as manually created rules are extensively used. Unfortunately, the classical DM methods are only capable of identifying the statistical significance between the commonly collocated terms, e.g., *Barack Obama*, *Angela Merkel*, *state heads*. Therefore the meaning behind these collocated concepts remains uncovered and the relationship to the other concepts cannot be discovered.

One of the most promising concepts to deal with this problem is the use of Machine Learning (ML) in order to improve the extraction of new facts and generalization of the already extracted ones based on the predefined patterns and examples (also called "seeds"). NELL [13] is one of the pilot projects investigating the ability of ML algorithms to improve the IR quality on the web scale. NELL crawls over the sites on the Web, identifies newly appeared concepts (e.g., *persons*, *cities*) and retrieves the relationship between them (e.g., the *city* in which the *person was born*). The knowledge extraction is performed in a completely automated way. The major disadvantage of this approach is a poor ability to tackle with noisy and non-trustful data, which is often referred as a semantic drift [14], which leads to the propagation of the non-credible facts to the newly learned fact base. On a small scale of data, the negative impact of the semantic drift can be overcome by constant monitoring and controlling the quality of the information extraction process by a human, i.e., a specialist of the problem domain. However, this makes neither meeting ad-hoc solutions nor scaling to big data demands technically possible.

Our suggested approach to overcome this limitation is to use the domain-specific OM algorithm to evaluate and control the ML process. The first proof-of-concept prototype, discussed in [15], reveals a positive impact of such an integrated "ML-over-OM" approach. By optimally balancing the OM workflow on the parallel hardware resources, which should be achieved by means of the techniques discussed in Section 2.1, we expect to scale up the ML algorithm to the real complexity, i.e., web-scale, use cases.

2.4 Language Processing

Natural Language Processing (NLP) is an important technique for text-based analytics and has found a wide application in a number of scientific and technological domains. As an example, NLP has been extensively used and proved successful for a long time in language checking, language quality control, information retrieval (amongst others [16]), machine translation and many others. At the same time, NLP is also much affected by the big data problem. The major challenge is not only the size (the length of a single document as well as the size of the document collection), but also the heterogeneous nature of language (ranging from a free-style email communication to semi-structured Wikipedia articles). In order to sufficiently address these challenges, the NLP technology should be designed in a highly scalable manner. Modern hybrid methods that allow linguistic analysis techniques to be combined with the corpus-based processing [17][18] offer a promising vision on how the high

scalability can be achieved. However, these methods are pretty shallow in terms of the depth of semantics to be retrieved. Therefore a combination with the machine learning techniques discussed in Section 2.3 would be of an enormous advantage. Moreover, the integration with ontology modelling methods will allow further components of an IR workflow to use the NLP techniques, e.g., for named entity recognition (retrieval of names related to persons, organizations, places, vehicles, etc.), sentiment analysis (e.g., by means of special forms of text mining), extraction of relationships between the objects, etc.

2.5 Scalable Data Management

The advances of modern large-scale infrastructures, such as Cloud computing, offer an extensive computing power to conduct challenging experiments quickly. However, the computation-centric orientation leads to a very high complexity of the data-centric application scenarios implementation. Data management services, such as Amazon S3, which are mature and well-established on the market, are very limited as far as efficient processing of stored data is concerned. This would be crucial for any integrated text analytics platform as the example in [19] indicates. Highly-optimized high performance computing systems are also characterized by extremely poor support offered to data-centric processing algorithms, which is the case when the major challenge of the application is constituted by the size of data rather than the complexity of the processing algorithm itself.

From the technology perspective, relational databases are typically used for storing textual data. This is also the case for the above-discussed SMILA and Gate frameworks. Additionally, XML databases may be used on top of relational databases for managing the ontology as well as the indexing system, which somewhat speeds up the rate of data access and, consequently, the overall performance. Despite being efficient and robust for small-scale data, this approach does not scale well on the big data range. Moreover, relational database technologies are not very suitable to be applied for statistical algorithms required for the text analysis, which is due to the need of processing sparse matrices.

The shortcomings of relational databases in their application to text analysis algorithms can be overcome by using graph data bases, such as InfiniteGraph [20] – a distributed graph database tailored to store and process structured and linked data, implemented in the Java language. Indeed, the use of graph data structures for representing textual and ontological concepts offers a lot of opportunities in terms of performance and scalability improvement as compared to traditional, relational database approaches. Moreover, the exceeding horizontal scalability features of graph databases offer a good basis for the deployment on parallel computing resources. However, the major challenge of integrating graph databases into the existing text analysis platforms is constituted by the solid design of those frameworks, which makes the seamless integration impossible and requires the underlying algorithms to be redesigned.

3 Approach to Enrich Data Mining by Enabling Ontology Modelling and Natural Language Processing

The state-of-the-art analysis, briefly summarized in Section 2, reveals that the traditional approaches to text mining stand to benefit from enrichment by innovative ontology modelling and natural language processing techniques. This would not only improve the quality of the analysis, thus increasing its practical value for a number of data science domains, but also enable the application to problem sizes on big data scale. The suggested workflow of the typical information retrieval process can be summarized as shown in Fig. 1.

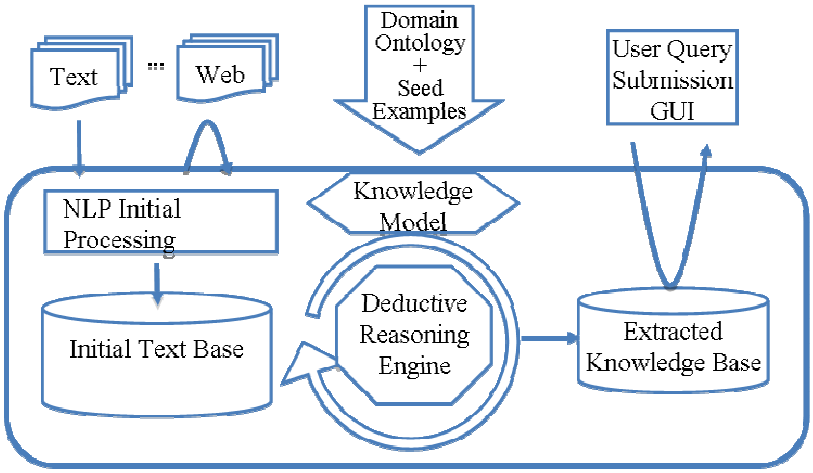

Figure 1: Schema of the workflow combining data mining, deductive reasoning, natural language processing, and ontology modelling methods to improve the information retrieval from text corpora.

As a preparatory step, the analyzed document corpus is uploaded to a document base, which can be done either by users manually or also performed automatically by the analysis platform. For the latter purpose, the platform offers a special service that crawls the internet pages that are in scope of the research, downloads them, recognises the text format and stores the extracted text in the base. A set of smart converters for the most wide-spread document formats, e.g., doc, pdf, html, etc., are provided by the platform as well. The analysis begins with the identification of sentences as well as principal lexical forms in them, supported by computer linguistic and natural language processing (NLP) methods, as discussed in Section 2.4. At the next phase, a bootstrapping analysis mechanism is used to identify new terms (e.g., Persons) as well as relations (e.g., role in the company) that constitute the basic knowledge model of the whole text corpus. The process is then iteratively repeated and at each step the knowledge model is enhanced and extended with the new set of terms (such as other persons), definitions, and relations. The key role during this process is taken by a deductive reasoning algorithm, used for checking the semantic

consistence of the retrieved statements. The latter is performed based on a domain-specific ontology schema. The schema is produced from the domain ontology, enhanced by the analysis of seed examples, as described in Section 2.3. The major advantage of the ontology-based reasoning algorithm is that it allows for a certain level of automation of the knowledge retrieval processes and does not require any human inspector as in case of the traditional knowledge discovery methods. The growing algorithmic complexity, caused by the need to make reasoning over the initial data corpus at each iteration of the analysis, is supposed not to have any considerable impact on the computational performance of the analysis due to enabled inherent parallel implementation of the reasoning algorithm, leveraging the high performance computing resources, see Section 4 for details on the platform's system organization.

4 Analytics Platform's Architecture Design

The practical realisation of the analysis approach discussed in Section 3 requires an elastic (in terms of offered resources) and rich (in terms of ensured data services quality) infrastructure, where the data should be collected and analysed in a centralised way, as well as a software platform that leverages the infrastructure's storage and computing facilities to solving on-demand (in terms of the needed scale and performance) data analytics tasks. Text analysis is a typical task that can be offered as a service – the users are interested in the information contained in the data rather than the data themselves. Therefore, the suggested text analysis platform's architecture is conceptualized according to the "Data-as-a-Service" (DaaS) system organization (see Fig. 2). It is worth mentioning, that our notion of DaaS-based cloud model is a little bit broader than the one given by Wikipedia [21] and assumes a consolidated service stack built on top of the infrastructure (IaaS), data analysis platform (PaaS) and user-centric application (SaaS) services. Following subsections discuss some distinctive features of the suggested data-centric cloud architecture design with regard to the analysis platform, infrastructure, and user-centric services.

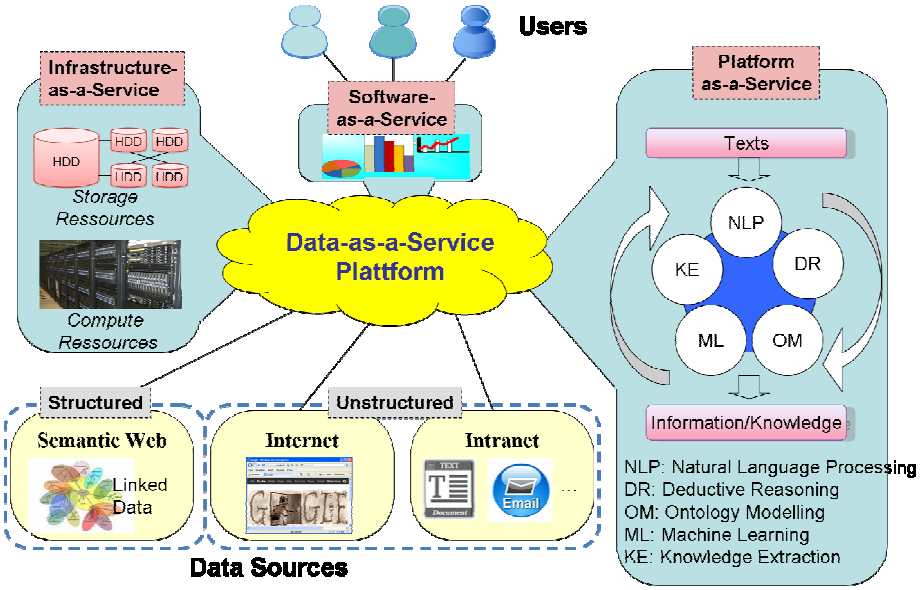

Figure 2: Data-centric cloud model for the enhanced text mining algorithm.

Analysis platform

The interdisciplinary nature of the integrated text mining approach, as discussed in Section 3, fosters the collaboration between independent groups of data scientists, each working on a corresponding part of the text mining workflow, whether it is natural language processing, deductive reasoning, ontology modelling, machine learning, or knowledge extraction. Thus, the main task of the platform is to provide means for developing services that implement the certain component's functionality by ensuring the needed level of interoperability between the loosely coupled services in order to be integrated in a common application, i.e., a workflow. Furthermore, the services can incorporate previously developed services, which have been designed in a way that allows their easy adoption by the upper-level services. In particular, such an approach has been investigated and successfully implemented in the LarKC project [3], whereby applications are developed as a workflow that is constructed of several services (or plug-ins, in the LarKC terminology) connected through a common data- and workflow management system. A "Data-as-a-Service" concept, as suggested by us, extends the basic principals of LarKC by providing such an analysis platform as a service, which means that all the essential components of the platform's software stack, such as execution runtime, database, web services host, workflow manager, etc., are permanently deployed on the cloud. The platform's advantage from the service-oriented design is manifold: (1) it does not introduce any additional requirements to the configuration of the user's underlying software or hardware layers – the client part of the cloud platform is constituted by light-weight web services, accessible via a plain schema (typically XML-based) from any internet- (in case of a public cloud) or intranet (in case of a private cloud) connected device; (2) it can be

relatively easily migrated to another cloud-based infrastructure in order to help meeting the concrete application's requirements (including Service Level Agreement, security, costs, etc.) or even spawn over multiple cloud providers, i.e., be offered through a federated cloud environment; (3) it matches automatically the application demands to the available infrastructure resources, so that the users do not need to make the allocation themselves; (4) it exploits the inherent parallelism of the analysis algorithms to fully utilize the infrastructure capacities.

Infrastructure

An on-demand infrastructure provisioning is one of the key features of the suggested cloud-based platform design. Depending on the concrete application scenario, the underlying hardware can virtually scale to meet the application demands, e.g., the size of data to be stored and analyzed, as well as to satisfy the predefined functional and Quality of Service requirements, such as the maximal query answering time, the number of parallel processed of user queries, etc. The infrastructure consists of three basic types of resources: compute cluster, storage disc farm, and local area network (LAN). Compute clusters offer hardware for performing computation, i.e., CPUs and RAM, loosely connected to each other by means of a (e.g., TCP-IP based) LAN, or, in case of supercomputers, through a high-speed network like Infiniband or Gigabit Ethernet. Some cluster nodes might be equipped with a local disk, however a more common case is when all nodes are getting access to a shared storage, available through a networked file system, such as Lustre. The main task of the cloudware is thus to manage all those hardware resources constituting the cloud infrastructure among the user experiments, or jobs, which is done by means of a cloud resource manager. Resource manager is an interactive service, having a role of the infrastructure frontend with regard to the platform's system software.

User-centric software

The use of the cloud platform by the users is facilitated by user-centric software, which includes interfaces to submit and control experiments, analyse results, etc. For this purpose, a set of intuitive general-purpose web services has been designed. Depending on the use case scenario, these services can be extended to meet user and application specific requirements.

5 Use Case Scenarios

This section presents two exemplary applications of the proposed text mining platform. The aim is to demonstrate the advantages of the platform with regard to typical analysis scenarios from science and industry. None of the scenarios could have been implemented with the existing tools so far.

5.1 Open Media Data Analytics on the Web

The everyday work of journalists and news agencies is largely influenced by the availability of free media data, published on the internet. The market success of a media platform is largely dependent on the timeframe between the news' arrival and their releasing in both internet and traditional (printed) media. The crucial point here is that the newly arrived information has to be validated with respect to its truthfulness and actuality before being published. Having a high performance contextualization platform, which would enable information extraction from all those large-sized, heterogeneous, and unstructured data sources, would be of a great advantage for journalists.

As an example, let's suppose a journalist has to take a decision about publishing a news item about some recent political event. The news might be outdated, come from an unreliable source, or might contain knowingly false statements, such as the wrong political allegiance or affiliation of the news's protagonist. Therefore, the information contained in this news must be validated according to other related news, e.g., the previously published ones, as well as publically available information from the internet. The analysis platform would automatically identify the information from the sources related to this news, categorize it, extract the main statements as well as the context they are used in (e.g., sentiments), etc. This helps greatly and serves as a recommendation for the publication. Moreover, the platform will allow the journalist to essentially extend the scope of news by considering other related events and articles. Considering the size of the analyzed data as well as the availability of near real time requirements for the discussed application scenario, this is a typical big data task that obviously requires such a platform as suggested in this paper. The research has already attracted attention of key players on the media market, such as Deutsche Welle. The potential end users of this technology are newspapers, radios, TV channels, and journals.

5.2 Strategic Business Decision Making in Enterprises

Intelligent taking of strategic business decisions is an important task for many enterprises that are planning or/and optimizing their research and prototyping activities towards achieving the identified innovation and technology management aims. For example, before reorienting its main production line to producing electrical vehicles instead of petrol or gas ones, car manufactures like BMW would need to analyze the results of a huge number of reports, studying the impact of this strategic decision from very different perspectives, including the technological, economical, and social ones. Successfully conducting this task often involves an extensive analysis of the internal document collections, such as marketing studies, technical reports, instruction manuals, technical email discussions, forums, trackers, and other textually captured and stored documents, whose aggregated volume can easily reach the size of several Petabytes. The role of the platform in the decision making process is not only to identify the documents related to a particular query as well as to retrieve the useful information contained in those document, but also to enrich this formation with other

knowledge coming from external, usually unstructured, data sources, e.g., from the Internet, in order to build a knowledge base to be used for the internal planning of the enterprise's strategy.

Unfortunately, a trivial syntactic look-up (e.g., via Google Search Appliance) has proved ineffective for such kind of search, since it only returns co-occurrences and does not retrieve the relationship between the searched terms and events. The suggested platform improves this kind of analysis by integrating enterprise-related domain knowledge regarding the search directly into the knowledge extraction process in the form of ontologies. The use of learning and reasoning during the knowledge extraction process further improves the amount and quality of the gained knowledge. Hence, the platform will retrieve knowledge for the concrete search from unstructured, distributed and heterogeneous big data in order to provide a highly-structured, comprehensive knowledge base, aligned with the current situation at the enterprise, in order to support meeting ad-hoc operative decisions.

6 Conclusion and Outlook

The paper introduced the results of an interdisciplinary research (spanning over information retrieval and analysis, semantic technologies, computational linguistic, data management, and high performance computing domains) aimed to elaborate a new technique to information and knowledge extraction from text corpora that fall within the scope of the big data aspect, i.e., the analysis of heterogeneous, unstructured, and large-sized data. The major motivation for the research lies in an increasing inability of the traditional techniques to address the big data challenges in textual data processing, as reported by numerous scientific and industrial research communities. The elaborated approach suggests enriching the traditional text mining techniques by integrating ontology modelling and natural language processing algorithms. The major benefits of the integrated approach can be summarized in the following points: (1) full automation of the knowledge extraction process thanks to adopting a domain ontology based deductive reasoning algorithm; (2) very high quality of the retrieved information thanks to the iterative, self-learning analysis method; (3) applicability to a wide range of data sources thanks to adaptive NLP methods; (4) short time-to-market and high cost effectiveness during adoption by the new application domains.

The approach should be implemented and deployed in a data-centric cloud environment, conceptualized in a way conforming to the "data-as-a-service" paradigm. The cloud services, implementing the elaborated algorithms, will be designed considering the most promising parallelization techniques, such as MapReduce, MPI, PGAS, etc., in order to ensure a high efficiency of the developed software when running in high performance computing and cloud environments.

The future research will concentrate on the following actions: (1) further elaboration of the interdisciplinary analysis approach in cooperation with providers of the identified use case; (2) more precise assessment of the performance and scalability promises, enabled by the new approach; (3) implementation of a platform's prototype; (4) spreading out the data-as-a-service innovations for implementing data-centric

algorithms in diverse science and technology communities; (5) validation of the software implementing the elaborated algorithms in a cloud computing environment.

References

1. SMILA Framework Website: http://www.eclipse.org/smila/
2. GATE Projekt Website: http://gate.ac.uk/
3. A. Cheptsov: Semantic Web Reasoning on the Internet Scale with Large Knowledge Collider. International Journal of Computer Science and Applications, Technomathematics Research Foundation, Vol. 8, No. 2, pp. 102 – 117, 2011.
4. C. Pedrinaci, D. Lambert, M. Maleshkova, D. Liu, J. Domingue, R. Krummenacher: Adaptive Service Binding with Lightweight Semantic Web Services. In Service Engineering. European Research Results (S. Dustdar and F. Li eds.), Springer, 2010.
5. J. Dean and S. Ghemawat: MapReduce - simplified data processing on large clusters. In Proc. OSDI04: 6th Symposium on Operating Systems Design and Implementation, 2004.
6. A. Cheptsov and B. Koller: Message-Passing Interface for Java Applications: Practical Aspects of Leveraging High Performance Computing to Speed and Scale Up the Semantic Web. International Journal on Advances in Software, vol 6 no 1 & 2, year 2013, pp. 45-55.
7. E. Gabriel et al.: Open MPI: Goals, concept, and design of a next generation MPI implementation. In Proc., 11th European PVM/MPI Users' Group Meeting, Budapest, Hungary, September 2004, pp. 97–104.
8. A. Cheptsov and B. Koller: JUNIPER takes aim at Big Data. inSiDE - Innovatives Supercomputing in Deutschland, vol. 11, No. 1, Spring 2011, pp. 68-69.
9. H. F. Saggion: Ontology-based Information Extraction for Business Intelligence. Proceedings of the 6th International Semantic Web and 2nd Asian Semantic Web Conference, pp. 843-856, Springer, 2007.
10. B. Yildiz: Ontology-Driven Information Extraction. (T. U. Wien, Hrsg.), 2007.
11. J. Langbein: Concept and implementation of a self-learning, ontology-based retrieval of entities and relations in natural language texts (in German), 2012.
12. B. H. Glimm: A Novel Approach to Ontology Classification. Journal of Web Semantics. Science, Services and Agents on the World Wide Web , 14 (84-101), 2012.
13. D. Movshovitz-Attias and W. W. Cohen: Bootstrapping Biomedical Ontologies for Scientific Text using NELL. BioNLP-2012, 2012.
14. A. Carlson, J. Betteridge, E. Hruschka, and T. Mitchell: Coupling Semi-Supervised Learning of Categories and Relations. Proceedings of the NAACL HLT Workshop on Semi-supervised Learning for Natural Language Processing, 2009.

15. Carnegie Mellon University: Read the Web. http://rtw.ml.cmu.edu/rtw/, 2012.
16. U. Deriu, J. Lehmann, and P. Schmidt. Creation of a technique ontology based on filtered language technologies (in German). In Proceedings Knowtech, Bad Homburg, 2009.
17. D. Jurafsky and J. Martin: Speech and Language Processing, Upper Saddle River, New Jersey, 2009.
18. C. Manning and H. Schütze: Foundations of Statistical Natural Language Processing. MIT Press, 2004.
19. E.L. Goodman, D. Mizell: Scalable In-memory RDFS Closure on Billions of Triples. In Proceedings of the 4th InternationalWorkshop on Scalable SemanticWeb Knowledge Base Systems. Shanghai, China, 2010.
20. InfiniteGraph Website: http://www.objectivity.com/infinitegraph
21. Data-as-a-Service Wikipedia Entry: http://en.wikipedia.org/wiki/Data_as_a_service