Semantic Processing for the Conversion of Unstructured Documents into Structured Information in the Enterprise Context

Adam Bartusiak
Department of Computer Science
University of Applied Science Zittau/Görlitz
Görlitz, Germany
abartusiak@hszg.de

Jörg Lässig
Department of Computer Science
University of Applied Science Zittau/Görlitz
Görlitz, Germany
jlaessig@hszg.de

ABSTRACT

We present an on-going research project addressing the problem of massive amounts of unstructured data that is generated on a daily basis in most business organisations, regardless of size. Our motivation is to support in particular small and medium seized enterprises to gain a competitive advantage in the market. The goal is to improve their processes for extracting valuable business information from such disorganised data. To achieve this, we introduce a flexible and scalable data analysis framework capable of transforming various types of documents into semantically annotated structures. This includes emails, text files in various formats, slide presentations, blog entries, etc. Additionally, the solution provides a semantic search engine for structured retrieval of the analyzed information and a graphical layer to dynamically visualize the search results as an interactive graph. Throughout the paper, the architecture of two main engines that are responsible for data and text analysis and semantic search are described. We conclude that semantic processing of unstructured sources significantly improves data management and data integration within the enterprises.

Keywords

information retrieval, unstructured data, NLP, semantic annotations, semantic search

1. INTRODUCTION

Unstructured data offers a great potential for business organisations. Regardless whether it comes from external sources (websites, blogs, social media) or it is generated within regular business activities (organisation's email, MS Office documents, PDF files, presentation slides), in most cases it holds a lot of useful knowledge that can be utilized for different purposes such as analytics, business relationships, decision support, etc. and consequently for obtaining a sustainable competitive advantage on the market [3]. According to [11] the digital universe, which is mostly built from unstructured documents, is doubling in size every two years. This rapid expansion of unstructured data constantly drives the development of new technologies for information retrieval or knowledge extraction. Semantic Web is one of the main research contexts setting new initiatives and trends for finding, transforming, analyzing, integrating and visualizing knowledge within online resources. Research results achieved in the Web domain gain attention and are adopted by enterprises as feasible solutions for dealing with unstructured data within their organisations [8].

We follow the approach of utilizing latest Semantic Web solutions for unstructured data processing in the business field. The main focus is set on the analysis and integration of different sorts of unorganised data that is often distributed across the entire business organisation. Typically these resources are organised into hierarchies such as file systems, ontologies or relational databases, and dedicated tools are required to access and browse them. In our NXTM project (Networked XML Topic Maps) a new scalable and customizable architecture for information and knowledge management is introduced. It enables the user to keep control over such unstructured resources from a single endpoint and make them easily searchable and navigable.

This paper is a preliminary report on the NXTM system. In the next sections we discuss related work (Section 2), present the goals and objectives of our project (Section 3.1) and explain its architecture (Section 3.2). In the end we summarize our work and describe future work (Section 4).

2. RELATED WORK

As our project consists of two main modules responsible for semantic annotation of the resources and searching the analyzed data, we review existing work related to these two research fields.

2.1 Semantic annotation

There exist a number of web based *commercial* (AlchemyAPI[1], DandelionAPI[2], OpenCalais[3]), *open source* (Entityclassifier.eu[4], DBpedia Spotlight[5]) or *community-driven*

[1]http://www.alchemyapi.com
[2]https://dandelion.eu
[3]http://www.opencalais.com
[4]http://entityclassifier.eu/thd
[5]https://github.com/dbpedia-spotlight

SEMANTiCS 2016, September 12-15, 2016, Leipzig, Germany

DOI: http://dx.doi.org/10.1145/2993318.2993341

(FRED[6]) projects providing APIs for semantic analysis and knowledge extraction which annotate named entities, concepts, keywords, etc. from natural language input. Some of the solutions (AlchemyAPI, DBPedia Spotlight) contain algorithms solving the disambiguation problem among the discovered entities and automatically integrate them with the existing knowledge bases (e.g. DBPedia[7], Freebase[8]). A significant part of the platforms for semantic processing utilize one of the following solutions as underlying framework for natural language processing (NLP) and semantic tagging: i) *Apache UIMA*[9], ii) *NLTK*[10], iii) *GATE*[11], or iv) *LingPipe*[12]. Additionally there are a number of platforms for manual and semi-automatic text annotation available (e.g. Pundit [5]). An extensive overview on project for semantic annotation has been included in [12].

2.2 Semantic search engines

In contrast to full text search machines, semantic search engines additionally provide an interface for building precise queries which can directly refer to the underlying semantic schema. In order to efficiently retrieve information from semantic structures stored for example as RDF triples (built from: objects, its attributes and instances), a semantic SQL-like query language (called SPARQL) has been designed. However, this approach requires an accurate and complex knowledge about the underlying ontologies used in the search system what can be a serious obstacle for users who are not familiar with the given semantic environment. As a result, multiple approaches simplifying the process of semantic query building have been proposed. In [7] semantic search engines have been categorised into four groups characterized by different user interface design: i) *form-based* search engines, where special forms are provided suggesting the users existing ontologies, classes, properties and instances (e.g. Broccoli [1]); ii) *RDF-based querying languages fronted search engines*, which provide sophisticated querying languages to support semantic search; iii) *semantic-based keyword search engines*, which enhance the performance of the traditional keyword search engines techniques by making use of available semantic data (SIREn [4]); and iv) *question answering tools* that use existing semantic annotations to answer questions in natural language format (e.g. IBM Watson). A comparison of other existing semantic search engines is included in the state-of-the-art section in [2] and in [9] presenting a more abstract view on this topic.

2.3 Information retrieval and knowledge management systems

There exist research projects which integrate an entire set of solutions for semantic data extraction, storing, indexing and retrieving and thus build complex platforms for Knowledge Management (KM). One of the first famous publications in this context is a project called KIM[6] which extensively describes different issues concerning semantic annotation, indexing and retrieval of structured data. It also implements exemplary infrastructure for information extraction (based on GATE), storage and query, custom ontology and knowledge base, and entity retrieval. In [10] an extension of the Onto-DOM platform for document annotation and retrieval is proposed. The authors present a new conceptual architecture of a KM system for enterprises and provide an improved strategy for meaning interpretation by capturing not only nouns but also other modifiers of the annotated concepts (e.g. adjectives).

[6]http://wit.istc.cnr.it/stlab-tools/fred
[7]http://wiki.dbpedia.org
[8]https://developers.google.com/freebase
[9]https://uima.apache.org
[10]http://www.nltk.org
[11]https://gate.ac.uk
[12]http://alias-i.com/lingpipe

3. THE NXTM PROJECT

In this section we present an overview of the on-going research project NXTM, which is a document management platform aiming to create, integrate and manage knowledge in the business environment. The framework consists of multiple solutions for semantic processing of unstructured data, storage, retrieving and browsing sets of the analyzed documents.

3.1 Goals and Objectives

The main goal of our system is to enable enterprises to convert massive amounts of unstructured documents that are being generated within daily business activities into valuable knowledge resources. In this way, this unorganised and unused data can be utilized for trend analytics, decision support, problem solving, discovering new facts and dependencies, etc. In our opinion, such a system needs to meet the following requirements: i) syntactic and semantic processing of documents in various formats for metadata and entity recognition; ii) dynamic linking of documents - analysis of semantic relevance and similarity between documents; iii) periodic data re-analysis for tracking content and structure changes in the documents and thus improving the search reliability; iv) possibility of manual curation of search results by the user for better data integration and relevance improvement; v) management of security policies for controlling document access rights while browsing the search results; vi) ease of system integration - the platform should be available as a standalone web solution or as a plug-in for existing content and document management systems; and vii) a flexible and intuitive graphical user interface enabling an easy access and navigation over the analyzed data.

The entire architecture needs to be modular and flexible so that implementation, maintenance and further improvement of the system can be performed directly by the users. Also the underlying solutions for data modelling and search should be characterized by a low degree of complexity.

3.2 Architecture and Design

While designing our system we have been trying to meet all the requirements defined in the previous section. The NXTM project consists of two main modules that work as independent web applications: data analysis engine and search engine.

3.2.1 Data analysis engine

The core functionality of the system is based on an analysis engine that converts heterogeneous information sources into structured data. As the underlying framework for data analysis we use Unstructured Information Management Architecture (UIMA). This platform enables the definition of single processing blocks and putting them together into a pipeline. Each block acts as a separate analysis engine re-

sponsible for a particular NLP operation (e.g. tokenization, segmentation, POS tagging, etc.) and stores its results into a CAS (Common Analysis System) object in form of Standoff annotations. Beside analysis engines, UIMA provides two other system components: i) *collection readers* for converting provided documents into raw CAS objects and supplying them into the analysis pipeline; and ii) *CAS consumers* for digesting and serializing the CAS objects at the end of the pipeline, according to users needs.

In the NXTM project we implement custom collection readers for importing different types of text documents and XML files representing dynamic web content entries (e.g. blog entries). Additionally we utilize existing open source libraries written for UIMA that perform various NLP tasks and are provided by research communities such as DKPro[13], Apache OpenNLP[14] or StanfordNLP[15]. In particular, the analysis phase is characterized by the following steps (Figure 1):

1. **Loading and pre-processing**: documents to be analyzed, updated or removed from the system are firstly uploaded into a database throughout an independent web service (so called Connector interface). Then, using UIMA collection readers, the documents are loaded from the database into the analysis pipeline one after another.

2. **Document structure and content analysis**: the diverse analysis is performed by single analysis engines in a chained manner. Initially, language identification, metadata and mime-type discovery is performed using Apache TIKA[16]. In the next step, the NLP-engines responsible for segmentation, morphology, syntax and semantic analysis are executed. Depending on document's type and language, appropriate models are then utilized. They adopt the functionality of particular engines in order to achieve optimal performance. At the end of this stage, named entities are discovered and stored as annotations in the CAS object.

3. **Similarity calculation and document clustering**: during the analysis of the document a term frequency vector is created which reflects the scope and topic of the document. In the vector space model, documents with a higher TF-IDF similarity measure, form vectors which point in the same direction. During the clustering process, similar vectors are hierarchically aggregated. According to this approach, documents that are thematically related to the searched vector (query) can be quickly found. Additionally, the similarity measure is used to calculate the length of the edges between nodes in the result graph, which reflects the relevance between particular documents.

4. **Storing the documents with extracted data in a database and updating the index of the search engine**: new information gathered and discovered during the analysis process, such as entities, meta-data or keywords, consists of attribute-value annotations and must be stored in a database for further processing. The hierarchical nature of these data structures suits better noSQL persistence

[13]https://dkpro.github.io/dkpro-core
[14]http://opennlp.apache.org
[15]http://nlp.stanford.edu
[16]http://tika.apache.org/

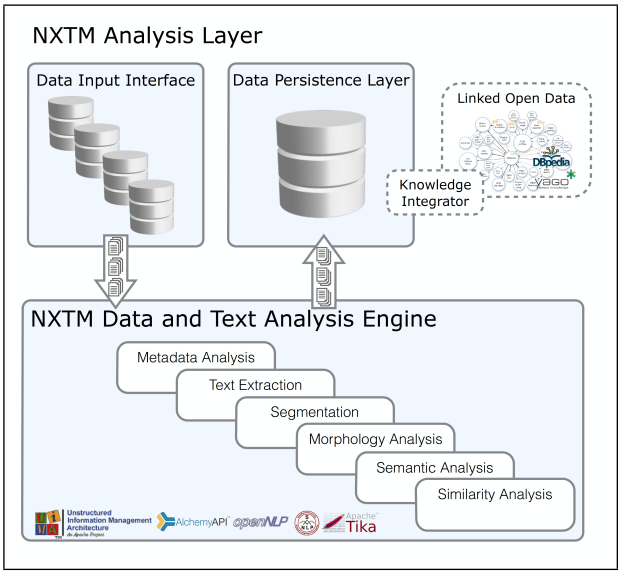

Figure 1: Analysis layer of the NXTM platform

solutions (i.e key-value stores, JSON databases) rather than relational databases at first glance. However, the RDBMS systems are still a standard technology in the business world. Typically SQL-based databases and their components are already existing in most of the business organisations. Introducing a new persistence technology may complicate the process of implementation and integration of our platform with user's existing infrastructure. As a result, the NXTM project uses an RDBMS (PostgreSQL[17]) as a persistence layer.

3.2.2 *Semantic search engine*

Building direct queries for retrieving the analyzed data might be quite inefficient and problematic for an average user. In most cases the user of the system is unaware of the ontology used and is not experienced with query languages for semantic data like SPARQL. Keyword search as the most common approach for document retrieval is often characterized by very low precision when applied to structured resources, generating a high rate of false positive results. For instance a simple keyword query consisting of two phrases "author" and "Smith" will retrieve all documents where both the words appear. By using a semantic search engine, it is possible to treat these phrases as attribute-value pairs. This assumption can significantly increase the precision of search since documents containing these strict key-value dependencies will be retrieved only.

In the project we have decided to use a hybrid approach that combines keyword search and semantic search. After doing research on existing solutions, Semantic Information Retrieval Engine (SIREn [4]) has been chosen as underlying search machine. This project adopts a JSON tree data structure and orderings of tree nodes to model datasets, entities and their descriptions. It includes its own indexing model and query operators. By using SIREn it is possible to perform a keyword search in the particular nodes of a JSON tree, preserving the hierarchical relations of the nodes at the same time. In this way, referring to the example shown

[17]http://www.postgresql.org

before, a query searching the attribute node "author" and the value node "Smith" can be built easily.

3.2.3 *Representation layer*

A very important part of the project is the graphical interface visualising the analyzed data in the system. Our intention is to provide the user with most accurate search results in form of an interactive graph. Its nodes and edges represent a network of related documents, entities, metadata and their interrelations. Dynamic rendering enables a real time navigation over the graph by clicking and expanding other nodes. In this manner, a user can deeply explore a given dataset and easily discover new entities, facts and resources that might be relevant to the searched topic.

As shown in Figure 2, our solution uses D3JS[18] project for visualising the graph. It is an open-source JavaScript library providing a set of data visualisation components such as graphs, tables, plots, maps and several other useful features.

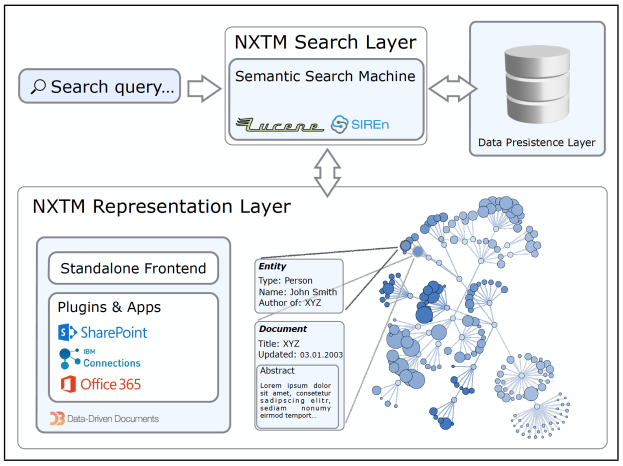

Figure 2: Search and representation layer of the NXTM platform

Depending on the search results, each node of the graph can represent either a document, an entity or meta-information. The edges of the graphs constitute different relations between the nodes. Their lengths express additional dependency information between resources such as the similarity between both resources or the confidence score for newly extracted metadata and predicates.

4. CONCLUSION AND FUTURE WORK

In our work we have introduced a new platform for annotation, search and representation of semantic information gathered from various unstructured resources. The main goal of the project is to improve the process of Knowledge Management within enterprises which directly influences the stability and the competitiveness of business organisations on dynamic business markets. We consider our approach as an efficient way for representing semantic data which enables exploration of new facts, entities and resources that are interconnected with the initial search results. Due to the fact that the NXTM system is still an on-going project and its prototype is being currently implemented, the paper does not yet include the performance results of this platform. The performance measures of the NXTM system will be expressed by the calculation and evaluation of precision, recall and F-measure figures that measure the correctness of semantic annotation (entity recognition) or semantic search.

Future work will focus on i) improving the performance of semantic annotation by including third party annotation engines providing REST APIs for their annotation services into the analysis chain; ii) enhancing local semantic structures through the integration with existing online knowledge bases, according to Linked Open Data principles.

[18]https://d3js.org

5. REFERENCES

[1] H. Bast, F. Bäurle, B. Buchhold, and E. Haussmann. Broccoli: Semantic full-text search at your fingertips. *arXiv preprint arXiv:1207.2615*, 2012.

[2] R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. *Hybrid search: Effectively combining keywords and semantic searches.* Springer, 2008.

[3] E. Blomqvist. The use of semantic web technologies for decision support –a survey. *Semant. web*, 5(3):177–201, July 2014.

[4] R. Delbru. Siren: Entity retrieval system for the web of data. In *Proceedings of the 3rd Symposium on Future Directions in Information Access (FDIA)*, 2009.

[5] M. Grassi, C. Morbidoni, M. Nucci, S. Fonda, and F. Di Donato. Pundit: Creating, exploring and consuming semantic annotations. In *SDA*, pages 65–72. Citeseer, 2013.

[6] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, 2(1):49–79, 2004.

[7] Y. Lei, V. Uren, and E. Motta. Semsearch: A search engine for the semantic web. In *Managing Knowledge in a World of Networks*, pages 238–245. Springer, 2006.

[8] M. J. Murphy, M. Dick, and T. Fischer. Towards the semantic grid: A state of the art survey of semantic web services and their applicability to collaborative design, engineering, and procurement. *Communications of the IIMA*, 8(3), 2008.

[9] U. Spree, N. Feißt, A. Lühr, B. Peisztal, N. Schroeder, and P. Wollschläger. *Semantic Search–State-of-the-Art-Überblick zu semantischen Suchlösungen im WWW.* na, 2011.

[10] C. M. Toledo, M. A. Ale, O. Chiotti, and M. R. Galli. An ontology-driven document retrieval strategy for organizational knowledge management systems. *Electron. Notes Theor. Comput. Sci.*, 281:21–34, Dec. 2011.

[11] V. Turner, J. F. Gantz, D. Reinsel, and S. Minton. The digital universe of opportunities: Rich data and the increasing value of the internet of things. White paper, IDC, EMC, April 2014.

[12] V. Uren, P. Cimiano, J. Iria, S. Handschuh, M. Vargas-Vera, E. Motta, and F. Ciravegna. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, 4(1):14–28, 2006.