conTEXT – Lightweight Text Analytics using Linked Data

Ali Khalili, Sören Auer, and Axel-Cyrille Ngonga Ngomo

University of Leipzig, Institute of Computer Science, AKSW Group,
Augustusplatz 10, D-04009 Leipzig, Germany
{lastname}@informatik.uni-leipzig.de,
http://aksw.org

Abstract. The Web democratized publishing – everybody can easily publish information on a Website, Blog, in social networks or microblogging systems. The more the amount of published information grows, the more are important technologies for accessing, analysing, summarising and visualizing information. While substantial progress has been made in the last years in each of these areas individually, we argue, that only the intelligent combination of approaches will make this progress truly useful and leverage further synergies between techniques. In this paper we develop a text analytics architecture of participation, which allows ordinary people to use sophisticated NLP techniques for analysing and visualizing their content, be it a blog, twitter feed, website or article collection. The architecture comprises interfaces for information access, Natural Language Processing (currently mainly Named Entity Recognition) and visualization. Different exchangeable components can be plugged into this architecture.

1 Introduction

The Web democratized publishing – everybody can easily publish information on a Website, Blog, in social networks or microblogging systems. The more the amount of published information grows, the more are important technologies for accessing, analysing, summarising and visualizing information. While substantial progress has been made in the last years in each of these areas individually, we argue, that only the intelligent combination of approaches will make this progress truly useful and leverage further synergies between techniques. Natural Language Processing (NLP) technologies, for example, were developed for text analysis, but are often cumbersome and difficult to use for arbitrary people and it is even more difficult make sense of the results produced by these tools. Information visualization techniques, such as data-driven documents, on the other hand can provide intuitive visualizations of complex relationships.

We showcase *conTEXT* – a text analytics architecture of participation, which allows ordinary people to use sophisticated NLP techniques for analysing and visualizing their content, be it a blog, twitter feed, website or article collection. The architecture comprises interfaces for information access, Natural Language

Processing (currently mainly Named Entity Recognition) and visualization. Different exchangeable components can be plugged into this architecture. An online demo of the conTEXT is available at `http://rdface.aksw.org/nlp`.

Motivation Currently, there seems to be an imbalance on the Web, hundreds of millions of users are sharing stories about their life on social networking platforms such as Facebook, Twitter and Google Plus. However, the conclusions, which can be drawn from analysing the shared content are rarely shared back with the users of these platforms. The social networking platforms on the other hand exploit the results of analysing user generated content for targeted placement of advertisements, promotions, customer studies etc. One basic principle of data privacy is, that every person should be able to know what personal information is stored about herself in a database. We argue, that this principle does not suffice anymore. In addition, people should be able to find out, what patterns can be discovered and what conclusions can be drawn from the information they share.

When Judy updates her Facebook page regularly over years, she should be able to discover what the main topics were she shared with her friends, what places, products or organizations are related to her posts and how these things she wrote about are interrelated. Being able to understand what conclusions can be drawn by analysing her posts will give Judy at least some of the power back into her hands she lost during the last years to Web giants analysing big user data.

The text analytics architecture and implementation we present in this article helps to mitigate the analytical information imbalance. With almost no effort, users can analyse the information they share and obtain similar insights as social networking sites.

Approach conTEXT lowers the barrier to text analytics by providing the following key features:

- No installation and configuration required.
- Access content from a variety of sources
- Allow refinement of automatic annotations and take feedback into account
- Show instant benefits to users
- Provide a generic architecture where different modules for content acquisition, visualization and NER can be plugged together.

2 Workflow and interface design

Workflow As shown in Figure 1, the process of text analytics in conTEXT starts by collecting data from the social web. conTEXT utilizes standard data access points like RSS, ATOM feeds, SPARQL endpoints and REST APIs as well as customized web crawlers for WordPress, Blogger and Twitter to collect html/textual data.

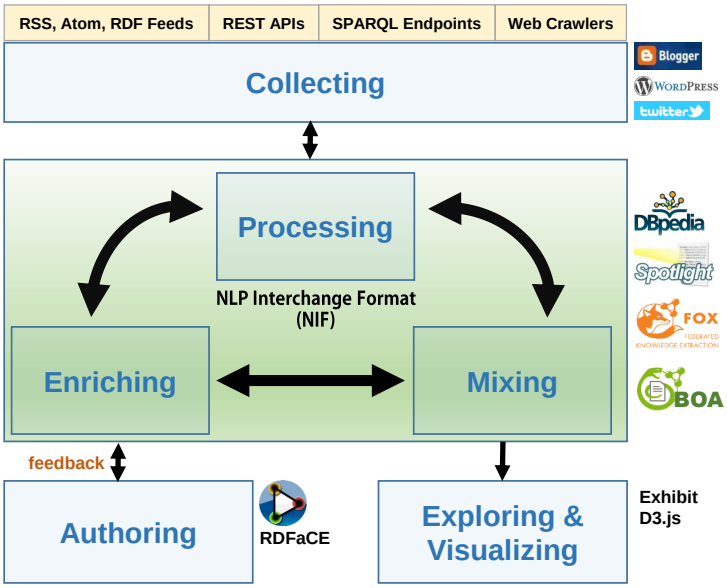

Fig. 1. Text analytics workflow in conTEXT.

The collected data as text corpus is then sent to Natural Language Processing (NLP) services for processing. conTEXT employs *DBpedia Spotlight*[5] and *FOX* (Federated knOwledge eXtraction Framework)[1] for annotating named entities in the text. The processed data is then enriched by two mechanisms: 1) dereferencing the DBpedia URIs of the entities to add more specific information to the named entities (e.g. adding longitude and latitudes for places) and 2) matching the entity co-occurences with pre-defined natural language patterns provided by *BOA* (BOotstrapping linked datA)[1] in order to extract the possible relationships between the entities.

The processed data can also be mixed with the other existing corpora as a data Mashup to provide more context for the results. For example, a user's Wordpress blog corpus can be integrated with a user's Twitter and Facebook corpus. In order to provide interoperability when mixing and enriching the results of different NLP services, conTEXT employs the *NIF* (NLP Interchange Format)[2]. NIF also enables the fast and easy integration of other existing NLP services into conTEXT.

The processed, enriched and mixed results are presented to users using different views for exploration and visualization of the data. *Exhibit*[3] for structured data publishing and *D3.js*[2] for data-driven documents are employed for dynamic exploration and visualization. Additionally, conTEXT provides an au-

[1] `http://aksw.org/Projects/FOX`
[2] `http://d3js.org/`

Parameter	Description
text	annotated text.
entityUri	the identifier of the annotated entity.
surfaceForm	the name of the annotated entity.
offset	position of the first letter of the entity.
feedback	indicates whether the annotation is correct or incorrect.
context	indicates the context of the annotated corpus.
isManual	indicates whether the feedback is generated by user or by other NLP services.
senderIDs	identifier(s) of the feedback sender.

Table 1. NLP Feedback parameters.

thoring user interface based on *RDFaCE* (RDFa Content Editor)[4] to enable users to refine the annotated results. User-refined annotations are sent back to the NLP services as constructive feedback for the purpose of learning in the system.

Progress indicator interfaces The process of annotating a large text corpus can be long and time-consuming. Users might need to wait a long time until the whole text corpus is collected and annotated. Therefore it is very important to provide users with some immediate results and to advise them about the progress of annotation task. For this purpose, we have designed special UIs to keep users awake until the complete results are available. The first indicator interface is an animated progress bar which reflects the percentage of the collected/annotated results. The second indicator interface is a real-time tag cloud which is updated while the annotation is in progress.

Authoring interfaces Lightweight text analytics proposed by conTEXT provides a beneficial incentive for users to adopt semantic text annotation. Users will get more precise results as they refine the annotations. On the other hand, NLP services will also take advantage of these manually-polished annotations to learn the right annotations. conTEXT employs RDFaCE on top of the faceted browsing view and thus enables users to edit existing annotations while browsing the data. The manual annotations will be collected and sent to the corresponding NLP service. The feedback encompasses the parameters specified in Table 1;

Exploration and visualization interfaces The dynamic exploration of content indexed by the annotated entities will result in faster and easier comprehension of the targeted content for text analytics. conTEXT creates a novel entity-based search and browsing interface for end-users to review and explore their content. On the other hand, conTEXT provides different visualization UIs which present, transform, and convert semantically enriched data into a visual representation, so that, users can read and query them efficiently. Visualization UIs are supported by noise-removal algorithms which will tune the results for better representation and will highlight the picks and trends in the visualizations.

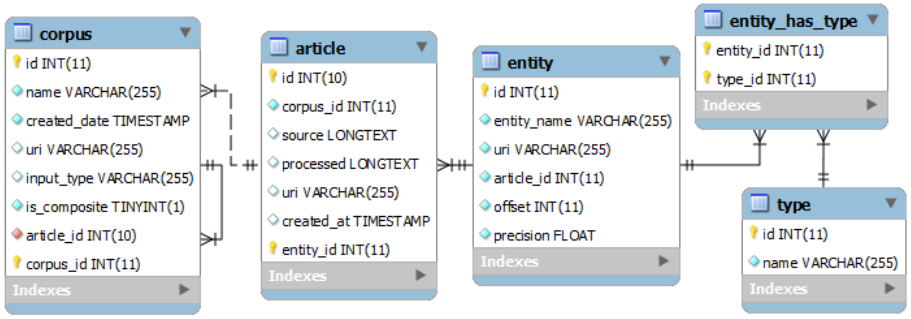

Fig. 2. conTEXT underlying data model.

3 Implementation

conTEXT is implemented using *PHP* and *JavaScript*. The application relies heavily on *AJAX* and *JSON* format as input for dynamic client-side visualization and exploration.

As shown in Figure 2, conTEXT incorporates *Corpus*, *Article*, *Entity* and *Entity_Type* classes to represent and persist the data for text analytics. A corpus is composed of a set of articles or a set of other corpora (in case of mixed corpus). Each article includes a set of entities defined by URIs and annotation score. Entity_type stores the related type(s) for each entity.

4 Views

conTEXT provides the following views for exploring and visualizing the annotated corpora:

- *Faceted browsing* allows users to quickly and efficiently explore the corpus along multiple dimensions (i.e. articles, entity types, temporal data).
- *Places map* shows the locations as well as their corresponding articles in the corpus.
- *People timeline* shows the temporal relations between people mentioned in the corpus.
- *Tag cloud* helps to quickly identify the most prominent entities in the corpora.
- *Chordal graph view* shows the relationships among the different entities in a corpus. The relationships are extracted based on the co-occurrence of the entities and their matching to a set of predefined natural language patterns.
- *Matrix view* shows the entity co-occurrence matrix. Each cell in the matrix reflects the entity co-occurrence by entity types (color of the cell) and by the frequency of co-occurrence (color intensity).
- *Trend view* shows the frequency of entities over the times.

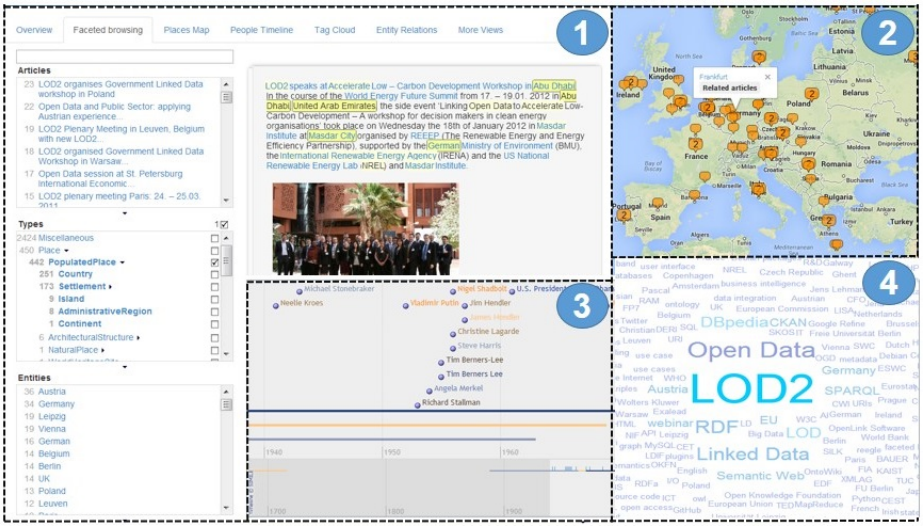

Fig. 3. Different views for search and exploring an analyzed corpus: 1) faceted browser, 2)map view, 3) timeline, 4) tag cloud

– *Image view* shows a picture collage created from the entities Wikipedia images. This is an alternative for tag cloud which reflects the frequent entities in the corpora by using different image sizes.

5 Conclusion

With conTEXT, we showcased an innovative text analytics application for end-users, which integrates a number of previously disconnected technologies. In this way, conTEXT is democratizing and downsizing NLP technologies, so they can be easily and beneficially used by arbitrary end users. conTEXT provides instant benefits for annotation and empowers users to gain novel insights and complete tasks, which previously required substantial development. Such tasks include in particular:

– Finding all articles or posts related to a specific person, location or organization.
– Identifying the most frequently mentioned terms, concepts, people, locations or organizations in a corpus.
– Showing the temporal relations between people mentioned in the corpus.
– Discovering typical relationships between entities.
– Identifying trending concepts or entities over time.
– Find posts where certain entities or concepts co-occur.

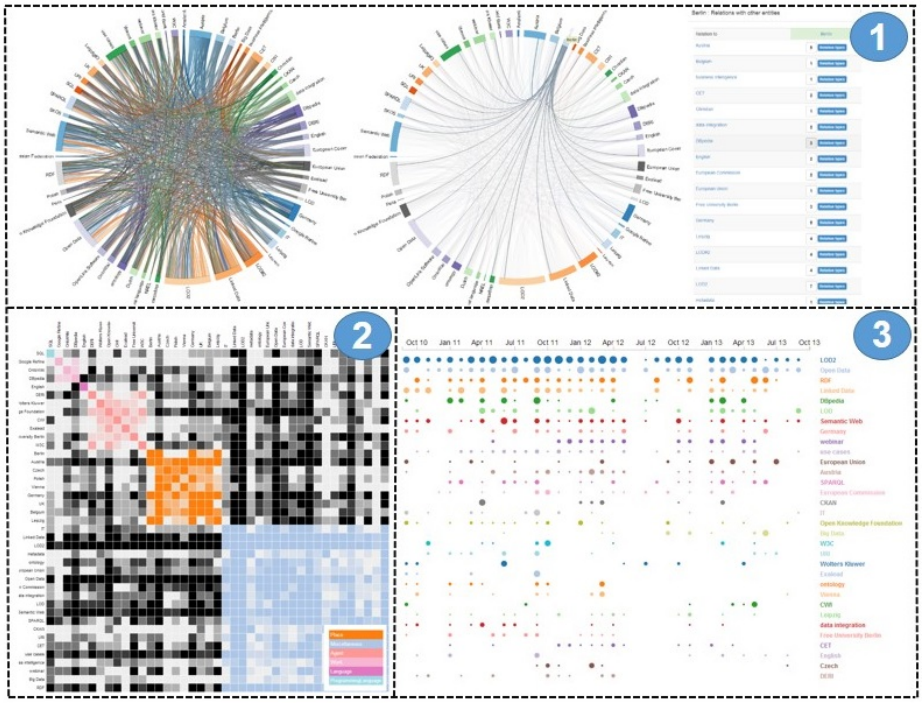

Fig. 4. Different views for visualizing an analyzed corpus: 1) chordal graph view, 2) matrix view, 3) trend view.

Appendix

5.1 Minimal Requirements

1. *conTEXT is an end-user application.* conTEXT can be easily used by any user and provides attractive exploration and visualization interfaces.
2. *Information sources are under diverse ownership or control.* conTEXT can process arbitrary textual information sources ranging from blogs over social network feeds to arbitrary websites.
3. *Information sources are heterogeneous.* conTEXT can deal with a number of syntactic information representations (feeds, APIs, plain text). conTEXT processes various forms of unstructured content. It maps extracted information to heterogeneous semantic structures.
4. *Information sources contain substantial quantities of real world data.* conTEXT can be used with arbitrary textual information sources. The text analytics and visualization functions are sufficiently scalable to process large quantities of real-world data.
5. *Meaning is represented using Semantic Web technologies.* conTEXT encodes and represents semantics in texts using RDFa.

6. *Data is processed and manipulated in interesting ways.* conTEXT empowers ordinary users to gain new insights from analysing their textual content with just a few klicks.
7. *conTEXT achieves things that alternative technologies cannot do.* Without semantic technologies and representations formalisms of the text analytics functions would hardly be possible. In particular, the faceted-browsing or timeline exploration interfaces require semantic background knowledge.

5.2 Additional Desirable Features

conTEXT fulfills most of the additional desired features. conTEXT provides an attractive and functional Web interface comprising a number of innovative visualizations of the extracted semantics. conTEXT is scalable in terms of the amount of data used since a variety of arbitrary information sources can be processed in real-time. Due to its NIF (NLP Interchange Format) interface a number of distributed components (e.g. Named Entity recognition) are working together. conTEXT provides novelty, in applying semantic technology to text analytics tasks that have not been considered before. With its semantic search and a large number of visualization techniques, conTEXT's functionality goes beyond pure information retrieval. The application has clear commercial potential and potentially large user base, since all social media users are potential conTEXT users. There is a use of dynamic data (i.e. social media feeds) in combination with static information (from the Data Web). The results are as accurate as possible, since user feedback is taken into account for subsequently improving the natural language processing.

References

1. D. Gerber and A.-C. Ngonga Ngomo. Bootstrapping the linked data web. In *1st Workshop on Web Scale Knowledge Extraction @ ISWC 2011*, 2011.
2. S. Hellmann, J. Lehmann, S. Auer, and M. Brümmer. Integrating nlp using linked data. In *12th International Semantic Web Conference, 21-25 October 2013, Sydney, Australia*, 2013.
3. D. F. Huynh, D. R. Karger, and R. C. Miller. Exhibit: lightweight structured data publishing. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 737–746, New York, NY, USA, 2007. ACM.
4. A. Khalili, S. Auer, and D. Hladky. The rdfa content editor. In *IEEE COMPSAC 2012*.
5. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY, USA, 2011. ACM.