


# Market and Technology Monitoring driven by Knowledge Graphs

Andreas Belger<sup>1</sup><sup>[0000-0001-7594-3677]</sup>, Renato Budinich<sup>1</sup>, Ralph Blum<sup>1</sup>, Martin Zablocki<sup>2</sup>, Roland Zimmermann<sup>3</sup><sup>[0000-0001-7380-5908]</sup>

<sup>1</sup> Fraunhofer SCS, Nordostpark 93, 90411 Nuernberg, Deutschland  
{andreas.belger, ralph.blum, renato.budinich}@iis.fraunhofer.de

<sup>2</sup> Trivadis AG, Elisabethenanlage 9, 4051 Basel, Schweiz  
martin.zablocki@trivadis.com

<sup>3</sup> Technische Hochschule Nuernberg Georg Simon Ohm, Kesslerplatz 12, 90489 Nuernberg  
roland.zimmermann@th-nuernberg.de

**Abstract.** In this paper, we describe an ongoing research project that aims to detect and trace trends for markets and technologies, hidden behind the vast amount of diverse information populated through the whole world. Our goal is to detect and follow upcoming and ongoing trends in a domain-agnostic and automatized fashion. In this paper we describe our experiences from the initial project steps and our approach using a continuously growing Knowledge Graph. We use a general model that allows us to capture identified mentions and relationships and resolve them into a number of entity and fact classes. Based on two business use cases we present first results where we already gained new insights into various technological developments without the intervention of human domain experts.

**Keywords:** Knowledge Graph, Semantic Web, Text Mining, Market and Technology Monitoring

## 1 Introduction

The idea of this research project is to work on tools, which reveal lines of timely developments by analyzing a “stream” of publicly available information, usually issued on a daily, weekly or monthly basis in public domains. More specifically the focus is on timely monitoring of technologies readiness (or maturity). Those technologies are propelled by a variety of stake holders (as e.g. universities, research institutes and tech companies) in certain market or branch surroundings. Chronologically, such information is first viewed in the form of patents, scientific publications, domain publications and, with some delay in general news relating technologies to market applications and distinctive use cases. In this paper, we describe an ongoing joint research project of Fraunhofer Supply Chain Services (SCS), Technische Hochschule Nueremberg (THN) and Trivadis AG to retrieve such information from different sources continuously whenever it discloses. The project considers continuously information starting from 2018, which report on the e-mobility domain and retrieve information from those sources to answer the following sample questions:

- A. *Which companies may constitute potential acquisition targets or sales leads in the e-mobility market?*
- B. *In what stage of development are the existing technologies and which are emerging in the e-mobility market?*

In the early stage of the project we tested Latent Dirichlet Allocation (LDA) [1] to group different documents into topics. Regardless of how well LDA works, there was still a significant amount of manual work required to interpret these results, e.g. by characterizing resulting topics. We further utilized the word2vec method for word embedding [2], leveraging the semantical properties of the resulting vector space to find other companies and technologies that are similar or related to few manually selected ones. Nevertheless, we were faced with the challenge of manually keeping track of the provenance and source text of each entity of interest, since word2vec is agnostic about these details.

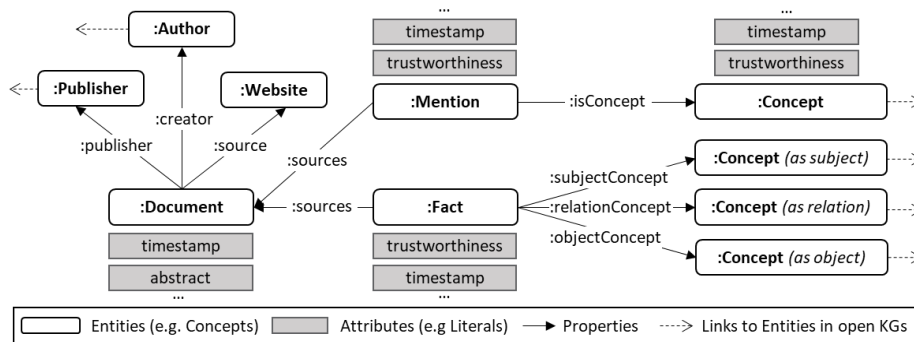
Reconsidering these experiences, we decided to use text analysis methods in combination with Semantic Web technologies. The representation of such information as a knowledge graph (KG), by means of the Resource Description Framework (RDF), allows not only to model complex networks of information, but also to infer latent structures [3]. However, constructing a KG from unstructured data, such as written text and providing a common interface for the business end-users is a challenging task.

First, we describe the approach of a general model to integrate entities and relations in a KG and how we extract these entities and relations from a continuous data flow by applying state-of-the-art Natural Language Processing (NLP). Secondly, we analyse the automatically built KG including 3,9 million entities and 54 million relations by applying the sample questions. Finally, we close with open questions targeted to the community.

## 2 **Methods**

For market and technology monitoring, we define a temporal development through the three following stages: (I) research: as description of functionality, (II) prototype: demonstration of functionality and (III) market solution: deployment on the commercial market. Different actors and events can describe each of these stages. For example, stage (I) is dominated by actors such as universities and research institutions. The relevant events are described by verbs like study, develop, observe. The focus is not on interpreting each text correctly, but rather on drawing conclusions from the entire stream of data. We use an easy-to-understand model, which is expressive enough to capture the described aspects and reduce complexity to being able to interact with the KG. This approach ensures that on the one hand we are able to disambiguate relationships from different sources, which actually represent the same thing, and merge them by means of a domain-specific ontology. On the other hand, information is made unambiguous without losing the provenance of the information. In addition, temporal changes should be mapped so that trends can be derived.

Figure 1 illustrates our simplified RDF model representing three main components: (1) input documents, (2) identified mentions, (3) retrieved facts. Each component includes metadata such as timestamps and trustworthiness regarding the method that was used to derive the RDF Triples from unstructured text. *Mentions* represent the particular appearances of an entity as a substring of the text, while *Concepts* represent a general disambiguated version of them. The relation between these concepts describes how they relate to each other. In order to merge similar relations and reduce their number, we clustered all relations based on the ones we need as *Facts*. We differentiate between mentions and concepts in order to be able to use a Named Entity Recognition (NER) tool, which can find new instances of specific types of entities in the texts without relying on Named Entity Linking (NEL) or the databases in the LOD it refers to.



**Fig. 1.** A simplified RDF model for market and technology monitoring

We collected a list of 1,082 potentially relevant RSS feeds in the field of e-mobility. From these we incrementally gather the new abstracts and integrate them into our KG. While processing the abstracts we are storing the metadata of the documents (such as source and title) in the KG. We then pass the texts through the publicly available Spotlight API [4] which links any recognized mention  $m$  to its DBpedia [5] concept  $c_m$  and store them in the KG. See also Fig. 1.

For the fact extraction we are currently employing a rule-based approach: we manually choose a set  $V = \{v_1, \dots, v_n\}$  of verbs of interest (e.g. buy, sell, produce,...) and look at their neighbors when considering the graph of synonyms built from WordNet [6]. This way we build  $C_1, \dots, C_k$  classes of verbs with similar meaning to ones representing events meaningful for market and technology monitoring. To detect the three stages of technologies' lifecycles described in the opening of this section we used  $k = 3$ , with  $C_1$  being a similarity class for the verb "develop",  $C_2$  for "test" and  $C_3$  for "order". We then use NLTK's part-of-speech tagger [7] to identify in the text corpus triples of the form  $s, v_j, o$ , where  $s$  and  $o$  are mentions that have the grammatical function of subject and object in the sentence while  $v_j$  is an element of  $V$ . We can finally create the corresponding *Fact* of the form  $c_s, C(v_j), c_o$ , where  $C(v_j)$  is the class of verb  $v_j$ , and reinsert this into the graph, for example in the form depicted

in Fig. 1. This allows us to query for facts between relevant entities somewhat independently of the particular formulation used to describe them in the original text. For interacting and monitoring the temporal changes, we define Sparql queries which are made available via a REST API of the KG database to a standard Business Intelligence Frontend. Due to the general model, the interacting remains small in its output triple size for monitoring at larger scales.

### 3 Insights and future research

In this section, we report our first results produced out of 452,549 abstracts about e-mobility from April to September 2019.

To answer question *A*, we analyzed data such as type and size of a company, which was provided by the disambiguated DBpedia concepts. The Semantic Web structure allowed us to analyze along multiple meaningful dimensions, e.g. find all companies in the same sector as any given (already recognized) company. Regarding question *B*, as mentioned in section 2 we defined three classes of verbs corresponding to developing, testing and ordering product technologies. As a preliminary analysis of the efficiency of the method, we manually checked a small number of the produced facts and found false-positive rates of 15%, 39% and 14% for the three classes respectively. In order to do a deeper analysis where we could compute also false negatives, and due to the lack (to our knowledge) of a domain relevant dataset, we are in the process of manually annotating a random sample of our text corpus. The labels identify whether a certain text contains a fact from one of the above defined classes involving a company. Regarding the second part of question *B*, we are unfortunately not able to give a fully satisfactory answer yet.

Our current plans for future research aim at extending into further market domains and on the technical side to enrich the structure of the KG. We are currently working on including new sources (such as social media), separating NEL and NER steps using tools such as Flair [8], SpaCy [9] and Agdistis [10] to detect also entities that have no current entry in DBpedia. To extract relations between entities in a more automated way, we consider investigating FRED [11] and PIKES [12]. FRED is a service that extracts semantic representations from natural language text offering a REST API and Python library for querying. PIKES is a Java-based suite for Knowledge Extraction that automatically extracts entities of interest and facts about them from text. Regarding the analysis, aside from the data contained in the KG, we want to start leveraging the structure of the KG itself: which methods from social networks analysis could be adapted in order to detect different types of node neighborhoods that could signal relevant features? How can we further integrate and exploit temporal aspects and dynamical changes? How can we define a semantic model which captures something like a "trend" as part of the graph and enables us to detect new and emerging ones?

We provide further information about the presented research project on the website: [www.th-nuernberg.de/future-engineering](http://www.th-nuernberg.de/future-engineering).

## References

1. Blei, D. M., et al.: Latent Dirichlet Allocation. In *Journal of Machine Learning Research* 3(4–5), 993–1022 (2003).
2. Mikolov, T. et al.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013).
3. Kertkeidkachorn, N., Ichise, R.: T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. In *The AAAI-17 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning*, pp. 743–749 (2017).
4. Daiber, J., et al.: Improving Efficiency and Accuracy in Multilingual Entity Extraction. In *Proceedings of the 9th International Conference on Semantic Systems (I-Semantics)*. Graz, Austria, (2013).
5. DBpedia Homepage, <https://wiki.dbpedia.org/>, last accessed: 2020/03/10
6. Miller, G. A.: WordNet: a lexical database for English. *Communications of the ACM* 38(11), 39–41 (1995).
7. Natural Language Toolkit Homepage, <https://www.nltk.org/book/ch05.html>, last accessed: 2020/03/10.
8. Akbik, A., et al.: Contextual String Embeddings for Sequence Labeling. In *27th International Conference on Computational Linguistics*, pp. 1638–1649 (2018).
9. SpaCy Homepage, <https://spacy.io/models>, last accessed: 2020/01/15.
10. Usbeck, R., et al.: AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Da-ta. In *ECAI 2014*, pp. 1113–1114 (2014).
11. Gangemi, A. et al.: Semantic Web Machine Reading with FRED. *Semantic Web Journal* 8(6), 873-893 (2017).
12. Corcoglioniti, F. et al.: Frame-based Ontology Population with PIKES. *IEEE Transactions on Knowledge and Data Engineering* 28(12), 3261-3275 (2016).