Business Analytics on Knowledge Graphs for Market Trend Analysis

Jens Albrecht[1], Andreas Belger[2], Ralph Blum[2], Roland Zimmermann[1]

[1] Technische Hochschule Nürnberg Georg Simon Ohm, Kesslerplatz 12, 90489 Nürnberg
{jens.albrecht, roland.zimmermann}@th-nuernberg.de
[2] Fraunhofer SCS, Nordostpark 93, 90411 Nürnberg, Deutschland
{andreas.belger, ralph.blum}@scs.fraunhofer.de

Abstract. We describe an ongoing research project that aims at automating information retrieval for technology and innovation management. It is built around a knowledge graph which is created automatically from selected news sources. Based on the knowledge graph, quantitative measurements of mentions on trend-relevant entities as well as changes in the knowledge graph over time are combined to offer insights into market trends for business users.

Keywords: Knowledge Graph, Semantic Web, Text Mining, Trend Analysis

1 Need for automated Technology and Innovation Management

Keeping track of technological developments, monitoring market changes and detecting innovations are essential activities for every enterprise to strategically define its value-creation processes for future success. However, generation of such knowledge is time-consuming due to the diversity of data and the complexity of relationships between technologies, applications and market participants which evolve over time. Automation of such information retrieval for technology and innovation management (TIM) is the goal of an ongoing research project at TH Nuremberg and Fraunhofer SCS. Currently, this approach is under development with a domain focus on e-mobility and smart energy topics. For maximum control of programmability and minimized operational costs, the research project aims at leveraging open source or free-to-use software wherever possible.

The use of text analysis methods in combination with semantic web technology offers the chance to automatically capture information, structure relationships and dynamically analyze changes. Information about entities of interest and their relations is stored in a knowledge graph [1]. The representation as a graph allows not only to model complex networks of information, but also to infer latent structures, e.g. subnetworks around influential players. The graph is a slowly changing structure, and its dynamic

development over time provides the basis for trend exploration. Thus, analytic queries on graphs can detect upcoming topics, influential players and new technologies.

2 Knowledge-Graph-Centric Process

The core process to support TIM in business aims at automating data acquisition and knowledge graph development to a large degree, while at the same time allowing for intuitive assessment by business analysts during trend exploration. Figure 1 shows an overview, centering around the creation of a knowledge graph termed "Trend Graph". The process is divided into three main stages with corresponding research questions:

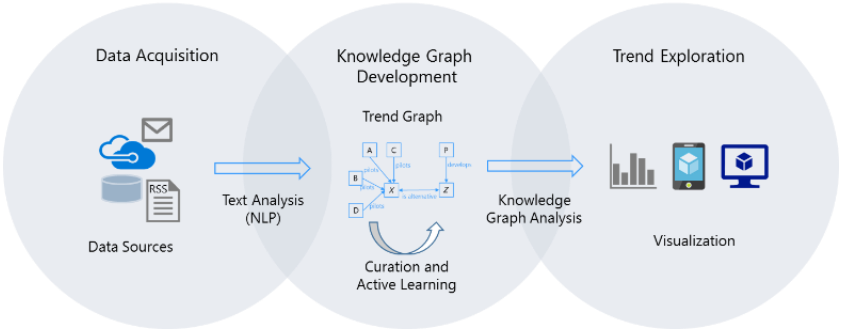

Fig. 1. Knowledge-Graph-centric process for Technology and Innovation Management (TIM)

1. *Data Acquisition:* How can reliable and representative data sources be selected to narrow in on relevant technology and market information thus reducing noise in gathered information while maintaining relevance of data for business users?
2. *Knowledge Graph Development:* How can market-specific entities (enterprises, technologies, products, events, etc.) be recognized, unambiguously identified and inserted with relevant relations between multiple entities into a knowledge graph? How is the historic development of such entities documented within a knowledge graph to allow analysis of technology and market developments over time?
3. *Trend Exploration:* What options are available to extract signals for market-relevant trends from a complex knowledge graph while at the same time hiding this complexity from business users? How can such analysis be automated, and results be visualized to offer access to the relevant factors and relationships between entities?

Data acquisition is currently based on manually selected RSS feeds (>500 are regularly monitored) which deliver news items for selected domains. The current sample consists of over 260,000 items in the domain of "e-mobility". Additional channels will be incorporated (e.g. Twitter, Blogs, patent databases) to enhance representativeness of facts and opinions. The focus of this paper, however, lies on *Knowledge Graph Development* and *Trend Exploration*.

3 Knowledge Graph Development

The key challenges for knowledge graph development are coverage of relevant information, correctness as well as consistency of the extracted information, and freshness, i.e. up-to-date information [1], [2]. To limit the number of concepts and relation types in the graph and therefore the effort for manual curation, it is helpful to define a domain-specific ontology [3], [4].

Our approach uses semantic web technologies for the implementation of a business knowledge graph, because standards like Resource Description Framework (RDF) and SparQL provide easy access to and integration of external knowledge from global open data sources like DBPedia [5] or YAGO [6]. The graph consists of strongly typed nodes and relationships defined in a domain-specific ontology. Contextual metadata like temporal validity or trustworthiness are included to support data curation and analysis [7].

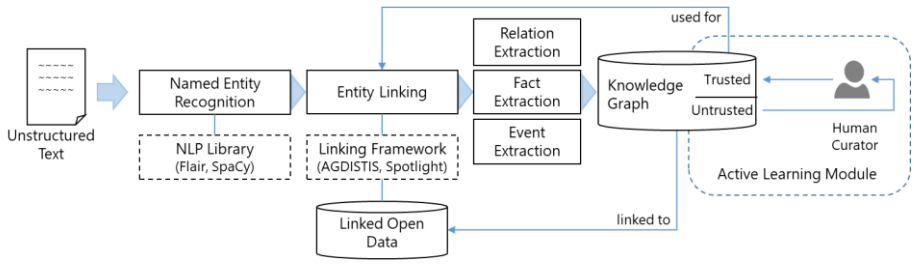

Fig. 2. Subprocess for Information Extraction and Knowledge Graph Development

During named entity recognition, *mentions*, e.g. potential hits for named entities like organizations, persons, products and date/time values are identified. The named entity recognition (NER) modules of Flair [8] and SpaCy [9] are used as ensemble to increase the accuracy of this step. Both provide state-of-the-art deep neural language models with pretrained word embeddings. The detected mentions need to be disambiguated and linked to unique entities (URIs) in the knowledge base. Open frameworks like AGDISTIS [10] can be utilized to link entities to public ontologies like DBpedia, which allow to infer further information like the type, size and location of a company thus ensuring basic meaningfulness of the knowledge graph. For each detected entity the link to the originating document and the date of publication are added to the knowledge graph as lineage information.

Furthermore, the confidence (trustworthiness) of each detection step is evaluated and stored in the knowledge graph. All information below a certain confidence threshold is marked “untrusted” and per default excluded from analysis. Entities included in the knowledge graph, which are initially given low trustworthiness (“untrusted”), need to be disambiguated by human curators as part of an active learning loop (see figure 2). Unknown entities such as new organizations are checked manually once and from thereon used automatically to match entity candidates in newly arriving texts.

The next step extracts relations, facts about entities and events using open information extraction algorithms. Events, i.e. expressions related to time, are particularly

interesting for trend analysis. The relations must be mapped to or newly integrated into the knowledge graph in a similar process as the entities.

The knowledge graph is developed as an RDF data model on the specifications of the W3C standards. All information is modeled as triples consisting of nodes and relationships stored in an RDF graph database. The current (June 2019) knowledge graph consists of 17,791,689 RDF triples.

4 Trend Exploration

Analyzing the Knowledge Graph is based on a descriptive analysis of selected mentions and related concepts. Initial questions involve for example the number of announced initial purchases by industrial or public buyers or the geographical distribution of mentions as well as key words within selected mentions and their variation over time. Figure 3 shows an example created with Microsoft Power BI based on the current graph for e-mobility where the sub-domain of electric busses has been selected. Close to 400 relevant mentions are identified. Key words in these mentions are shown in a word cloud (Fig. 3, right) disclosing the context around the selected mentions.

With cross-apply-filtering it is possible to select key words of interest and then characterize those by their geographical distribution to identify e.g. hot-spots for the initial installation and use of electric busses (Fig. 3, left). Mentions of commercial e-vehicle manufacturers are identified and counted (Fig. 3, middle), allowing to infer market relevance. Thus, end-users analyze the knowledge graph and infer knowledge about technologies and markets with a business intelligence (BI) tool.

Fig. 3. Trend Graph exploration example (BI tool: Microsoft Power BI)

Basis for the visualization is a group of different SparQL queries resulting in tables for mentions, enterprises, geography and data sources (e.g. RSS feeds) that are linked to each other within the BI tool to form a common star schema. In a generalized BI perspective, the knowledge graph resembles a core data warehouse while the frontend utilizes BI self-service capabilities to realize a data mart, which is optimized towards a specific group of end-users.

As text sources are stored additionally as full-text in the knowledge graph, a direct reference is permanently granted. SparQL queries are predefined and can be parametrized to some degree (e.g. restrict the selection to certain concepts) by end-users via parameter tables. The queries are available via a Rest-API of the triple store and can then be accessed by the BI frontend.

The next development steps in the Trend Exploration area focus on defining maturity level measurements for technologies and identifying structural changes in selected areas of the knowledge graph. The following example illustrates the idea of structural change calculations regarding actors, technologies and application projects:

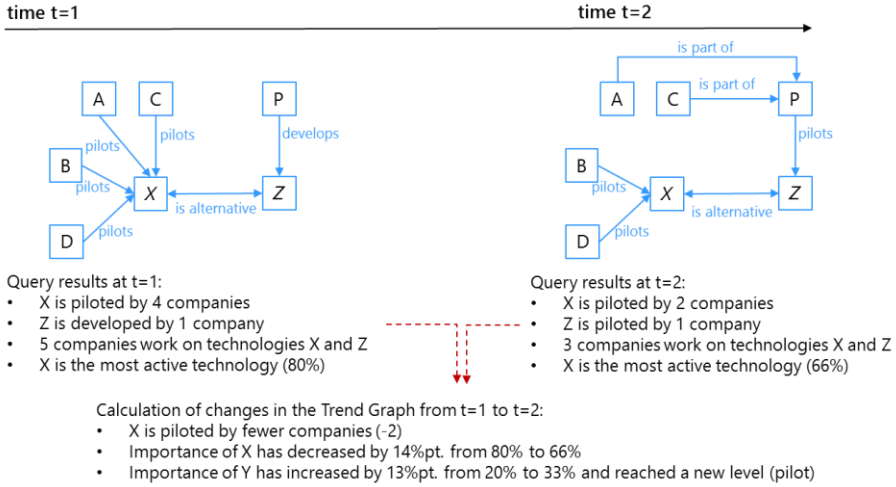

Fig. 4. Calculation of structural changes

The transition from time t=1 to t=2 involves a structural change in the knowledge graph with respect to the competing technologies X and Z. From a social network analysis perspective, the centrality of node Z has increased. One aim of future development is to test the applicability of centrality algorithms to identify structural changes (e.g. changed relevance of concepts) and create quantified indicators for trend exploration.

5 Insights from Pilot and Future Work

We presented the concept of a knowledge graph for trend analysis as an innovative approach for business intelligence. One of the key research challenges is the evolution of the information. Companies split and merge, products enter and leave the market. This kind of events introduces to areas for research, novelty detection and staleness detection. Regarding novelties, it would be helpful to generate signals immediately when interesting new information is integrated into the graph. We examined a sample from the knowledge graph (e-mobility/ grid topic) to determine whether the knowledge claims in the graph are interesting new information in comparison to *energate messenger*, a leading paid-content publisher for German energy market news [11]. The sample

includes 26 distinct articles published from January to April 2019. During this period 19 articles were published on *energate*. Results show that 10 of the knowledge graph articles were not available on *energate* while 16 were identical. Three experts performed the comparison independently. This indicates that the graph contains new and relevant information for the sample topic and even goes beyond the benchmark (paid-content provider). It is part of our further research to define how the significance of signals can be determined based on the content of the graph. But most information in the graph can become stale or invalid. However, staleness cannot be fully determined unless further evidence is found in the data sources. A model for data aging dependent on the kind of information would be helpful to generate some kind of staleness score influencing the trustworthiness of analyses. To determine the performance of these two aspects we are working on extended evaluation processes.

References

1. Noy N, Gao Y, Jain A, Narayanan A, Patterson A, Tylor J (2019) Industry-scale Knowledge Graphs. Lessons and Challenges. ACM Queue 17(2):1–28.
2. Kertkeidkachorn N, Ichise R (2017) T2KG: An End-to-End System for Creating Knowledge Graph from Unstructured Text. The AAAI-17 Workshop on Knowledge-Based Techniques for Problem Solving and Reasoning:743–749
3. Kim Y, Ju Y, Hong S, Jeong SR (2017) Practical Text Mining for Trend Analysis: Ontology to visualization in Aerospace Technology. KSII Transactions on Internet and Information Systems (TIIS) 11(8).
4. Wimalasuriya DC, Dou D (2010) Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science 36(3):306–323.
5. DBpedia Homepage. https://wiki.dbpedia.org/. last accessed: 2019/06/24
6. YAGO Homepage. https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/. last accessed: 2019/06/24
7. Krötzsch M (2017) Ontologies for Knowledge Graphs? 30th International Workshop on Description Logics, Bd 2017
8. Akbik A, Blythe D, Vollgraf R (2018) Contextual String Embeddings for Sequence Labeling. 27th International Conference on Computational Linguistics:1638–1649
9. Spacy Homepage. https://spacy.io/models. last accessed: 2019/08/15
10. Usbeck R, Ngomo A-CN, Roder M, Gerber D, Coelho SA, Auer S, Both A (2014) AGDISTIS - Agnostic Disambiguation of Named Entities Using Linked Open Data. ECAI 2014:1113–1114.
11. Energate Homepage. https://www.energate.de/medien/energate-messenger.html. last accessed: 2019/08/15