1 Using Machine Teaching in Text Analysis

Case Study on Using Machine Teaching with Knowledge Graphs

Thomas Cook
Cambridge Semantics
USA
Thomas.Cook@CambridgeSemantics.com

Rajib Saha
Parabole.ai
USA
rajib@mindparabole.com

Aditya Narayanamoorthy
Parabole.ai
USA
aditya.narayanamoorthy@mindparabole.com

Sandip Bhaumik
Parabole.ai
USA
sandip@mindparabole.com

ABSTRACT

Parabole.ai with its TRAIN™ [1] machine teaching platform and Cambridge Semantics with its AnzoGraph® DB [2] offering have been working together to improve the analytical value from diverse data sources and ease the process of unstructured data knowledge discovery and text analysis. In this paper we outline some of the learnings from implementing NLP (Natural Language Processing) and ML (Machine Learning) with Knowledge Graphs.

BACKGROUND

Parabole.ai is a machine teaching company that automates knowledge extraction from unstructured information. Cambridge Semantics Inc. is a modern data management and enterprise analytics software company. AnzoGraph® DB is a fast scalable massively parallel processing (MPP) graph database offering by Cambridge Semantics. AnzoGraph DB is based on open standards and supports data integration and rich analysis with graph algorithms, inferencing, data warehouse-style analytics, geospatial and more. Parabole.ai has been using its machine teaching and knowledge graph analytics platform with AnzoGraph® DB to jointly address unstructured data & text analysis for customers.
Typical joint customers have been large corporations with diversified offerings.

KEYWORDS

Graph database, natural language processing, machine learning, NLP, data integration, knowledge graphs

2 CUSTOMER NEEDS AND PROBLEM DEFINITION

The proliferation of the variety of data sources in the corporate world continues as new sources from inside and outside enterprises get more and more varied. Enterprises are looking to extract knowledge from unstructured data and combine this extracted knowledge with data within their databases to execute business processes or gain insights.

Figure 1: Human Defined Structure and Definitions

Category	Context	Definition
Environment	Air Quality	The category addresses management of air quality impacts resulting from stationary (e.g., factories, power plants) and mobile sources (e.g., trucks, delivery vehicles, planes) as well as industrial emissions. Relevant airborne pollutants include, but are not limited to, oxides of nitrogen (NOx), oxides of sulfur (SOx), volatile organic compounds (VOCs), heavy metals, particulate matter, and chlorofluorocarbons. The category does not include GHG emissions, which are addressed in a separate category.
Social Capital	Human Rights & Community Relations	The category addresses management of the relationship between businesses and the communities in which they operate, including, but not limited to, management of direct and indirect impacts on core human rights and the treatment of indigenous peoples. More specifically, such management may cover socio-economic community impacts, community engagement, environmental justice, cultivation of local workforces, impact on local businesses, license to operate, and environmental/social impact assessments. The category does not include environmental impacts such as air pollution or waste which, although they may impact the health and safety of members of local communities, are addressed in separate categories.
		The category addresses a company's ability to create and maintain a safe and healthy workplace environment that is free of injuries, fatalities, and illness (both chronic and acute). It is traditionally accomplished through implementing safety management plans, developing training requirements for

For broad adoption, data extraction must be automated, as must be the discovery and linking of key business concepts as well as the process for managing and upkeep of data and information models so that all enterprise data, whether inside or outside of the organization can be searched, discovered and analyzed.

Historical dictionary-based approaches are static, labor-intensive to maintain, and poorly identify linked concepts. These systems are increasingly being replaced by automated model discovery and business solutions built on a Knowledge Graph.

Knowledge Graphs link entities and provide understanding of N-dimensional relationships and concepts for contextual understanding of the data. Simplifying how these Knowledge Graph data models get built and evolve as the business needs change, continues to be a challenge. Building and maintaining Knowledge Graph data models with unstructured data are even more complex and, if not maintained, can quickly get outdated.

There are various knowledge graphs that are used in existing applications, such as the Google Knowledge Graph [3] or DBPedia [4]. However, these are usually manually curated, and it is not often clear how these can be created automatically from varied data sources.

How do you easily create a Knowledge Graph? How do you further automate the process so that new concepts or terms can easily be

identified from the corpus of data and linked to the existing corpus of knowledge? How do you allow Knowledge Graphs to evolve as the organization learnings evolve?

Parabole.ai and Cambridge Semantics look to address these and other emerging needs as companies look to automate their unstructured data and text analysis processes. In this paper, we outline what we have done and some of the lessons we have learned.

3 COMBINING ML & KNOWLEDGE GRAPHS

Parabole.ai's TRAIN™ software blends subject matter expertise with ML learning techniques and Knowledge Graphs.

3.1 Human-Defined Concept Definition:

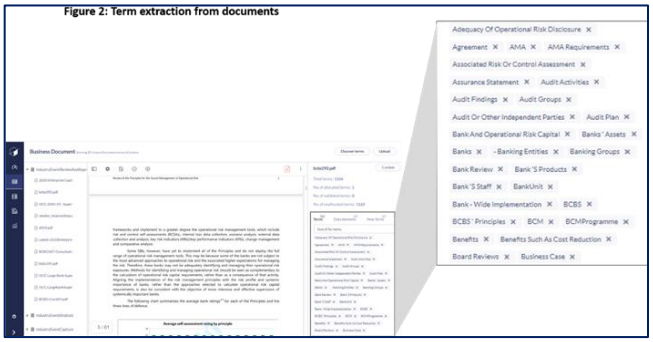
In Parabole.ai's approach, subject matter experts (SMEs) outline the initial knowledge discovery category, context and definition in a simple interface or an Excel spreadsheet. For example, in Figure 1, for a social impact investing use case, financial services firms may consider the "Environment" as a key category to collect data for knowledge discovery and analysis. In that "Environment" category, "Air Quality" may be an important element for evaluation. SMEs would define in natural language what "Air Quality" is, what factors are important in "Air Quality" and the relationship or value to the "Environment" category.

This SME description of structure and definitions provided in the spreadsheet or word interface is then used for creating the Knowledge Graph for socially responsible investing. This is done by automatically extracting the linguistic and semantic properties of words in the descriptions, as well as capturing relationships among the words and across definitions.

3.2 Knowledge Extraction:

Parabole.ai's TRAIN™ software extracts important key phrases and concepts (all called "terms") from target documents stored in various data sources (e.g. SharePoint, File Systems, ElasticSearch), on the scale of thousands of documents. NLP and ML techniques are used to extract these terms and then map them to other related terms, structures and concepts in the Knowledge Graph which was automatically created from SME definitions in the Concept Definition phase. Machine Learning identifies new terms that were not previously considered by SMEs in their Concept Definition phase. These terms can be considered and accepted by SMEs to further enhance the Knowledge Graph (See Figure 2).

Figure 2: Term extraction from documents

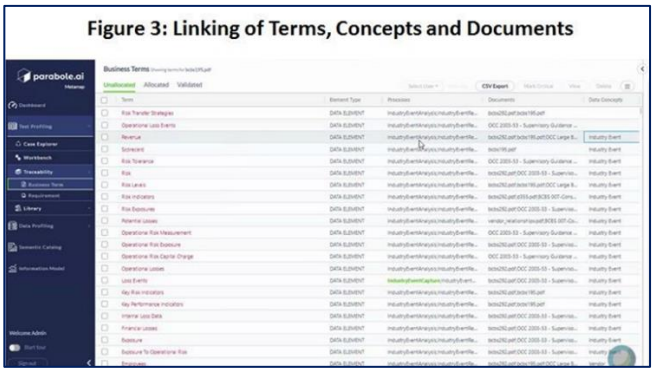


3.3 Linking of Terms, Concepts and Documents:

Linking of terms, concepts and documents is automatically done by the software. This linking provides the data lineage that can be used to materialize and showcase sections of the document from where the fact was extracted and is available when the data is being used by a user for search, discovery or analysis (See Figure 3).

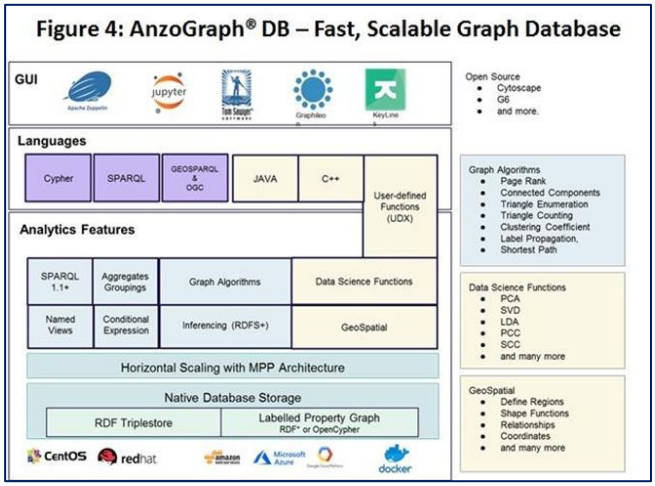
Carrying on from the previous example of "Environment" and "Air Quality" as important categories and elements in a domain, at this stage, there may be other terms and concepts linked to "Air Quality" that have been extracted from the various input data sources for the knowledge. These terms may be present in the original definition of the element, such as "Industrial Emissions", but they may also be linked through the Knowledge Graph using linguistic similarity, such as "Industrial Sector Emissions", or semantic similarity, such as "Factory Pollutants".

Figure 3: Linking of Terms, Concepts and Documents



3.4 Storing and Leveraging the Power of Horizontally Scalable Graph Database built for Analytics

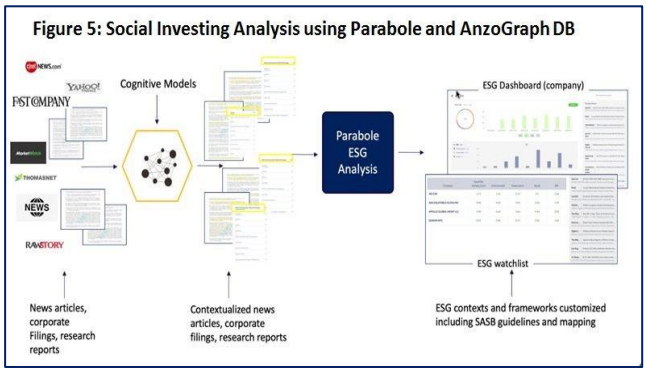
Core to storing, analyzing and using the NLP and ML data from Parabole.ai's TRAIN™ software is the use of a highly scalable, fast and flexible graph database uniquely positioned for handling data integration and analytics at scale.



The data extracted by TRAIN™ is stored and managed in AnzoGraph DB as a Knowledge Graph. It is important to have a fast database that easily scales as data needs grow. This allows customers to start with a single server machine when data volumes are small and add servers as data volumes or computational needs increase.

Graph databases naturally store entities and relationships and are built for Knowledge Graph types of data structures (i.e. graph data structures). AnzoGraph® DB is an ACID (atomicity, consistency, isolation, durability)-compliant MPP graph database that shards data across cores in a server or across servers when deployed as a cluster with each core individual processing the data. The database can be deployed in the cloud or behind the firewall. Inferencing capability offered by AnzoGraph® DB is important to find similar terms from the data compiled by Parabole.ai's TRAIN™ software. Other analytics functions such as aggregates, graph algorithms and geospatial are important to find insights from the data.

The end result is the creation of an end-to-end solution that allows organizations to access, link and analyze data from diverse sources and allow users to search, discover and conduct additional analysis (See Figure 5). This solution thus provides improvements over manual Knowledge Graph creation, and scales up the amount of input that can be handled, as well as the amount of data that can be subjected to analysis, thanks to TRAIN™ and AnzoGraph DB.



Currently, there are two categories of companies using TRAIN™ for sustainability research. Two private equity firms are using the solution to augment third-party research papers on portfolio companies. Many times, third-party research is not sufficient for private equity analysts to make informed decisions and create a due diligence file on companies within their coverage. These third-party reports are also static and are generally updated on a monthly basis. TRAIN™ provides the ability to research both portfolio and watchlist companies in near-real time, allowing analysts to "do their own work" and build solid accurate files for these companies ahead of their competition. Corporate clients are using TRAIN™ to verify their own sustainability efforts and verify the accuracy of how their efforts are perceived and being reported in the marketplace. It is also allowing these clients to measure sustainability based on their internal risk controls to identify potential exposure to ESG risk factors in a fast-changing environment in real time.

4 LESSONS LEARNED

It is not sufficient to use NLP and ML alone. Terms or facts extracted from documents without context creates a bag of words or phrases that cannot be understood or used easily.

Knowledge Graphs link these terms to provide context and meaning; however, manually creating and managing Knowledge Graphs is painful, labor-intensive and not sustainable as documents and unstructured data sources grow.

Providing an easy interface, approach and tools for SMEs to build the Knowledge Graph and maintain the Knowledge Graph allows better knowledge creation and management.

ML and reasoning are important tools for identifying related terms previously not considered and to provide new insights.

5 REFERENCES

1. https://www.parabole.ai/
2. https://www.cambridgesemantics.com/anzograph/
3. https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html
4. https://wiki.dbpedia.org/