Ontology based Knowledge Extraction with Application to Finance

Özgür Bağlıoğlu and Mesut Çeviker
Department of Computer Engineering, Middle East Technical University, Cankaya, Ankara, Turkey

Keywords: Ontology, Knowledge Extraction, Semantic Search, Information Extraction, Information Systems, Knowledge Discovery.

Abstract: Public and private enterprise finance performance is reflected and affected by unorganized, unstructured data such as news, reports (IMF, OECD and other periodical reports) as well as structured statistical data extracted by Statistical Institutes and other organizations. The role of raw data in influencing performance and decision making is not negligible. In this context, this paper presents knowledge extraction methodology for precise and fast decision making in finance by using ontological tools. For this purpose, we firstly design finance ontology and collect datasets. The aim of this ontology is to support the knowledge management in the finance domain and to increase the productivity through evidence base, comprising raw finance data to be retrieved from various operational sources. We then propose to populate the ontology by using past project properties and project progress reports. After population of data, we plan to develop and use a semantic search engine to gather meaningful data i.e. knowledge. The semantic search engine will assist decision makers to make better decisions. The output of this work will be also used as an input for decision making and scenario based future prediction for finance as this study is a part of a larger project called “ontology based decision support system”.

1 STAGE OF THE RESEARCH

Projects, constituting the majority of the new initiatives, are the main activities for the continuity of the organizations. Organizations should select more plausible projects to accomplish their vision and strategies. However, there are many different factors and indicators to be considered while planning projects and making investment choices. It may be a challenge to decide useful activities amongst many initiative proposals for organizations having finite resources. Thus, financial management is of key importance for the organizations for proper management of funds.

In this work, finance domain is investigated from strategic management perspective, with different dimensions, namely planning, performance, budgeting, process, accounting, control, and risk. These aspects have different impact factors in classification and success of projects.

- Planning has a master vision (long term plans and programs) guiding what type of activities should be conducted.
- Process defines the structure of activities (projects, resources etc.).
- Budgeting is high level financing of activities.
- Accounting is generally about the realization of financial transactions.
- Control and risk are the environmental factors that affect the progress of activities from the finance perspective.

If all these aspects can be combined in a unified knowledge framework, the evaluation and monitoring of activities (projects) will become easier.

This paper presents common domain ontology for decision making. The aim of the ontology is to establish core factors affecting performance in decision making about projects (whether to invest, to continue, to maintain etc.). The domain ontology has been defined by domain experts using key phrases in finance domain.

This research is now on population of ontology stage using project dataset (consisting of project details and progress reports). Population helps ontology to evolve and knowledge base will be created on finance. We are using EU open dataset Cordis for

Bağlıoğlu Ö. and Çeviker M..
Ontology based Knowledge Extraction with Application to Finance.

population of ontology. Details are given in methodology phase.
After this stage, we plan to include semantic search capabilities to populated ontology. This helps to extract relevant knowledge for decision makers on approving new project proposals.

2 OUTLINE OF OBJECTIVES

The objectives can be divided into 3 groups as this study is conducted in 3 subfields:

1. Ontology creation and maintenance: Common Finance ontologies are defined from different perspectives (process, budget, planning risk etc.). This ontology will evolve as population and decision making continue. These ontologies provide explicit and formal information describing the concepts and their relationships for the target domain. Developed ontologies support the knowledge management in the finance domain and they can serve as a basis for the knowledge extraction.
2. Ontology population and enrichment: Automated crawling of projects, plans and other related documents is the first objective. In population stage ontology will evolve. We use graph database for ontology population stage. Population using a graph database and maintaining the updates between ontology model and graph database concurrently is another objective. The updates will be maintained in two ways: from model to graph, from graph to model.
3. Semantic search: We will evaluate the performance of ontology based semantic search with classical information retrieval model. Our objective is to gather more specific and related results with semantic search capabilities. Also the results will be used as an input for scenario based decision support system on finance which is another doctoral thesis of my college.

3 RESEARCH PROBLEM

This study is about knowledge extraction for precise and fast decision making in finance by using ontological tools. The first stage is the design finance ontology. The aim of this ontology is to support the knowledge management in the finance domain and to increase the productivity through evidence base, comprising raw finance data to be retrieved from various operational sources. We then propose to populate the ontology by using project attributes and project progress reports previously executed. The third phase is development of a semantic search engine to gather knowledge and input parameters for assisting decision makers. The output of this work will be also used as an input for decision making and scenario based future prediction for finance.

4 STATE OF THE ART

As depicted previously, the studies in this field are mainly conducted in three fields: ontology development, ontology population, semantic search. This section provides literature review about these areas.

4.1 Ontology Development

Ontology is a common linked vocabulary which is developed to share common understanding between researchers working in same domains. Ontologies cover machine interpretable definitions of domain concepts and relations among them (Noy, 2001). Ontology consists of vocabulary model, used for extracting knowledge, which includes domain specific concepts and notions and the relations between them. Ontologies are helpful in domain specific queries, and allow for gathering more relevant results compared to classic information retrieval methodologies. Some of the studies can be given as examples such as OntoSeek and Inquizit (Baclawski 2000).

The initiative comprehensive research done in ontology development is conducted by Jones et al. (1998). In this work, TOVE, Enterprise model approach, Methontology, KBSI IDF5 and many other ontology extraction methodogies are explained and compared in detail. Although there are substantial differences between these approaches, fundamentally, the common points are similar to research conducted by Noy et al. (2001). The ontology development methodology used within this research is similar to Noy's work; we also present practical application of the development methodology.

In another research conducted 2003, it is indicated that ontology technology is mature enough and there exist several tools for extracting ontology. This paper argues that there should be common standard for ontology development and this may increase the transfer of common information between common domain researchers. This study investigates several ontology development tools, explains some ontology definition languages and the detailed comparisons are made (Corcho, 2003).

In addition to the above-mentioned studies, there exists simple, applicable and more understandable ontology development work, preferred because of these factors, conducted by Noy and McGuinness. This work explains both fundamental ontology science and ontology creation steps. According to this work, although there are many tools for developing ontologies; the fundamental methodology includes the following steps:

- Determining the domain and scope of an ontology
- Enumerating important terms in domain
- Defining the entities and concepts in the ontology
- Determining the taxonomic relations between entities (such as class hierarchy)
- Determining properties and types of entities
- Determining the relationships between concepts (Noy, 2001)

This research is another application of ontology development with practical points in mind. The methodology in that paper is simple and applicable to other areas besides finance.

4.2 Ontology Population

Ontologies provide standardized means of modelling, querying, and reasoning over knowledge bases. After developing ontology, the instances of entities and relations should be identified. This process is a knowledge acquisition activity that relies on semi-automatic methods to generate instance data from unstructured and structured data sources (Topic Ontology Population, 2014). Through this process, the ontology will become capable of representing large amount of information using small number of axioms (entities and relationships).

Ontology population task can be examined from different perspectives. The preliminary studies include term extraction methods in which much research is conducted and this area becomes mature enough. Term extraction (named entity recognition) includes linguistic processes or rule based methods to extract noun phrases that express terms and special entities such as location, special name, and time phrases (Cimiano, 2006). However as ontology have concepts instead of named entities; the extraction of instances becomes more challenging.

Hahn and Schattinger (1998) introduce a methodology for automating the maintenance of domain specific taxonomies based on NLP methods. The ontology is updated as acquisition and extraction of texts. The extraction depends on the linguistic and conceptual quality of various forms of evidence underlying the generation and refinement of concept hypothesis. This approach is based on qualification calculus where several hypotheses about instance and concept are created and ranked. For example, a hypothesis “the printer <A>” would create an instance <A> from concept printer. The quality of hypothesis is generated according to grammatical constructions of lexical items (i.e. parse trees) and terminological knowledge base which derives concept hypothesis. The dataset is information technology magazines of 101 texts. The knowledge base contains 325 concepts and 447 conceptual relations. The results indicate high recalls but low precisions (about %30). The drawback of this linguistic approach is that it needs a set of accurate tools for linguistic processing i.e. domain specific POS-taggers, parsers.

One of the initial researches about automatic population is Artequakt, developed by Southampton University (Alani, 2003). Artequakt automatically extracts knowledge about artists from the web, make ontology population and generate biographies about artists. Generation of biographies consists of three steps:

- Knowledge extraction: responsible from gathering instances along sentences from web documents.
- Information management: stores and consolidates the information so that biography generation phase can query the knowledge base using inference engine.
- Narrative Generation: This phase has some story templates about biography, and according to it and using knowledge base generated in information management phase, generates automatically biographies about artists.

Artequakt uses GATE and WordNet to extract instances of ontology entities and relations. In extraction phase, the output is in xml format and almost all information is generated in structured format, so population becomes a very easy task. The success of this research comes from the dataset which is almost structured and easy to extract. Besides, this dataset is very narrow and has specific vocabulary set because of narrow domain.

The most comprehensive survey performed in ontology population by Wimalasuriya and Dejing (2010) review the details of different ontology based information extraction systems (OBIE). In this study, common layout and architecture of OBIE systems is defined and the studies are compared according to extraction methods, ontology generation methodology and types of resources.

In this survey, according to extraction methodology, most of the researchers use linguistic rules-specifying regular expressions that capture certain types of information. This method is combined with NLP methods such as part of speech taggers and noun phrase chunkers that enables wide range of rules. Another extraction technique is using gazetteer lists which recognize individual words or phrases instead of patterns. This technique is widely used in named entity recognition task (for example organization names). Next extraction technique is classification technique such as support vector machines, maximum entropy models and decision trees methods. For this technique, classifiers are trained to identify instances and relations of ontology. Different linguistic features (POS tags, capitalizations, individual words) are used as input for classification. Another extraction method analyzes HTML or XML tags to generate extraction rules. This method generally depends on dataset and the pages or documents must be structured. Last method named web based search queries for the instance whether it is relevant to entity or relationship in ontology. It uses Hearst patterns like "<CONCEPTS> such as <INSTANCE>". This pattern is queried and if the search results have reasonable number of outputs then instance candidate is labeled as instance.

In the survey, ontology generation methodologies of OBIE systems are also examined. One approach considers the generated ontology as an input to the system. The ontology is constructed manually with this approach. In another approach, ontology is automatically or semi-automatically constructed by using information extraction. Initially ontology may exist and updated with extracted information, or ontology is constructed from scratch automatically. The main aim is to update ontology with new information.

This paper mentions about the types of components of ontology that are extracted. These components are classes, data type properties, object properties, instances, property values of instances and constraints. OBIE systems using Ontology generation process generally extract information related to classes only. Other OBIE systems generally extract only instance names. Some systems also extract property values of instances.

The survey also compares OBIE studies according to datasets used as input. Many OBIE systems process the data from quite different sources. These sources are Wikipedia pages, documents or html files from domain, web pages from a particular site. PANKOW and OntoSyphon which uses web based search as information extraction have no restriction of the source. This fact implies that these methods can be applied to any domain including other languages. The survey also indicates that there are no de-facto standard text corpora for performance evaluation of different OBIE systems.

Lastly, in the survey the performance evaluation of OBIE systems is performed using standard information retrieval metrics: precision recall and F-measure.

4.3 Knowledge Base and Semantic Search

A knowledge-based system consists of a knowledge-base that represents facts about the world and an inference engine that can reason about those facts and use rules and other forms of logic to deduce new facts or highlight inconsistencies. The ideal representation for a knowledge-base is an object model (often called an ontology in AI literature) with classes, subclasses, and instances.

After population of ontology, knowledge base will be ready for querying. Semantic search is an approach to gather more relevant search results by understanding user's intent and contextual meaning of terms.

Current trend in semantic search is to rely on generally the World Wide Web, so semantic search engine independent of domain is hot research area. Hakia is a meaning based search engine that presents search results based on meaning matching rather than popularity of search terms (Alan, 2008). Because of the research scope, we will deal only with ontology based search engines instead of semantic search engines.

One of the ontology search engines is developed by Maedche et al. In this study, an infrastructure for searching, reusing and evolving ontologies in a distributed environment is discussed. This work presents an approach for evolution of distributed ontologies, which is based on keeping change information in the form of evolution logs. The

generated tool named KAON uses inference engine for answering conjunctive queries in the form of SPARQL. The search is based on WordNet and lexical matching of element names providing more intuitive searching (Mädche, 2008).
Another study is SEWISE, which is an ontology based web information search engine. This study is experimented in the financial domain where ontology is developed for web finance news. Statistical text mining techniques (text indexing, text categorization, text summarization, keyword extraction) are used to refine and extract HTML pages to XML files which are semantically tagged. These XML files include semantic knowledge of the finance news which enriches textual information. XML repository can be queried using rich Xqueries to gather relevant information (Gardarin, 2003).

Another research conducted by Buitelaar et al. focuses on ontology based information extraction from soccer web pages. In that study, a tool named SmartWeb Ontology Based Annotation (SOBA) component, which automatically populates a knowledge base by information extraction from soccer match reports found on the web, has been developed. They extract information from heterogeneous sources such as tabular structures, text and image captions in a semantically integrated way by using SProUT which multilingual NLP platform. It implements a novel paradigm in which information extraction, knowledge base updates and reasoning are tightly interleaved (Buitelaar, 2006).

Another study is conducted by Alan et al. describes a video annotation and querying system which is capable of semi-automatic annotation of videos from text. The domain is soccer domain and video segments are enriched with metadata (textual content). This approach provides semantic search in videos. This study helps user to query videos according to important parts of games (e.g. all goals by Hakan or Nihat) (Alan, 2008).

Similar research conducted by Kara et al.(2012) is about ontology based information extraction and retrieval. This study is conducted in soccer domain from UEFA and SporX. The extraction process is achieved by using domain specific ontology. The retrieval system is enhanced using semantic indexing where the domain specific information extraction is used. The domain specific queries (e.g. fouls committed by Daniel) give better results than traditional keyword based information retrieval. Also this system can answer the queries without need of SPARQL. This study is a base of our research with finance domain.

5 METHODOLOGY

In this section, steps of proposed ontology development procedure will be explained. After ontology development step, ontology population and semantic search methodologies will be explained.

5.1 Ontology Building

The ontology extraction methodology includes many migrations from words to the ontology. Firstly, domain knowledge is gathered by discussing key terms that are used in budgeting and accounting. Then these terms (more than 1500) are simplified to key terms (about 120) that are generally used in this domain. The key terms are defined and all these terms are discussed in detail through regular meetings with domain experts. This process also includes grouping the key terms into manageable categories so that ontology can be easily divided into small and easily definable sub ontologies.

Afterwards, we develop audit, budgeting, control, accounting, performance, planning, process, risk ontologies according to the clustering of key terms and relations between these terms. We discuss the key terms and relations within ontologies, in case of necessity we simplify ontologies by further consolidation of terms or relations. After determining all groups of ontologies, we merge relations and obtain a small set of relations. Lastly, we combine all groups of ontologies into one consolidated ontology and define new relations between these sub-ontologies.

While extracting budgeting and accounting ontologies, some groups seem to pose less importance at first sight. For example “risk”, “audit” or “control” has more general terms. The reason behind this is to focus on budget and account groups. Moreover, ontology extraction is an evolving process, new terms can always be added to these groups while decision making and performance management models are developed. Similarly, the relation between performance and budgeting can be extended. The main aim of this initial ontology is to capture domain knowledge and to identify the core relations in finance.

While developing the ontology, a number of domains exist such as budgeting, accounting, etc. The main problem here is the way of construction of a combined ontology including all the related domains. As a solution, instead of one big and unreadable ontology, related ontologies are developed for each domain separately.

Another problem to be solved is the identification of too many different relations for the ontologies. In order to address this problem, the relations having similar meanings are combined under a general relation or transferred to a mostly used associated relation. Therefore, a limited set of relations are specified after many iterations.

The ontology is developed in Protégé, a knowledge modelling tool and OWL is used as knowledge representation language. For finance domain we generated budget, performance, planning, process, accounting, control, risk and audit ontologies. We are still in the process of developing the properties of entities so far. The reason is while populating ontology new properties of entities are generated. The ontology maintenance (update) phase will survive till the end of research as it evolves with enrichment of ontology and upcoming new datasets.

5.2 Ontology Population

This phase consists of many challenging steps: enhancing ontology using properties and features of datasets, dataset identification and purification, information extraction step by using concepts and relations in ontology.

One of the important issues in determining the entity properties is the dataset specifications. The dataset should consist of different types of information. Some of these are:

- Organization types and names
- High level plans, programs
- Project specifications and progress reports

As the size of dataset can be very large, we do not prefer using Protégé's repository for instances of ontology entities. There exist some alternatives for this big data management issue. One of them is assuming ontology as a graph and using graph database. The other one is saving all instances to a database and using mapping between database items and ontology entities or relations. The last option is using ontology editor's repository which is in XML structure, but it might not be feasible for retrieval or search due to performance reasons.

We prefer to use graph database for population of instances. Graph databases are often faster for related and hierarchical data sets. Ontology consists of various triple stores. Triple stores (subject-predicate-object) consume a lot of disk space in relational databases. The retrieval and search operations also slow for very large datasets. Neo4j stores whole graphs as opposed to "just" triples. Also it is fast for querying and scales very well to handle larger datasets (Vicknair,2010) (Neo4j, 2014).

The mapping between Protégé and Neo4j OWL models will be maintained during all phases of search.

The dataset used in populating ontology is taken from EU Research Projects Portal (Cordis). CORDIS is the primary repository for EU-funded RandD projects covering a myriad of science, technology, and research-related fields and topics. Dating from before 1990 to the present, they relate not only to the Seventh Framework programme (FP7), but also to previous framework programmes. Project details are published on CORDIS after they are made available to CORDIS by the Commission service responsible. This service uses the breadth of the CORDIS repository as a base to bring together a wide variety of information related to individual projects, including: project details, descriptions, funding, programmes; project results, documents, reports, summaries; project participants; links; publications; multimedia. This dataset is open and has about 50000 project results with 3 framework programs and many subprograms. The analysis of programs and subprograms are also public. This dataset will be used for initial population of ontology. This dataset is in English and population with free text will be easy because of rich framework set exist in English language (GATE, WordNet etc.).

We plan to include different kinds of datasets for enrichment of ontology. One of the fundamental sources will be structural database which is used by the Ministry of Finance (MoF). This database is used for budgetary operations of projects conducted by this Ministry. We propose to enhance the raw data using semantic methodologies and generate ontology instances from this data. Additionally, by using other textual reports defined at the beginning of this section, we enhance the knowledge base. We will use Turkish language for this dataset which is among the most commonly used 20 languages in the world. In Turkish there is no generic NLP tool like GATE as far as we know. The Wordnet like studies in Turkish is not mature enough also. These points make the extraction process more complicated and time consuming for second dataset.

Also there may be cross lingual maintenance issues for ontology enrichment which are open to research.

5.3 Semantic Search

After population of ontology, the ontology will evolve and knowledge base will be ready for semantic search. For semantic search we base our implementation on traditional IR model: open source search engine called Lucene. In this phase we will compare the performance of our semantic knowledge with classic keyword based retrieval. The search result will be an input for rule based decision making system which queries about the performance of projects.

Decision making and future prediction phases will be conducted by our colleagues in the Artificial Intelligence group.

The ontology and semantic search capabilities facilitate decision support systems, and project assessments will be based on more concrete parameters for performance, this is a step forward in an innovative evidence based decision making. The general structure of this project is depicted in Figure 1. As seen from the figure this study consists of two phases. One of them is ontology based knowledge extraction which we mention in details. The second phase is decision making by using semantic search capabilities which is a future work of this search.

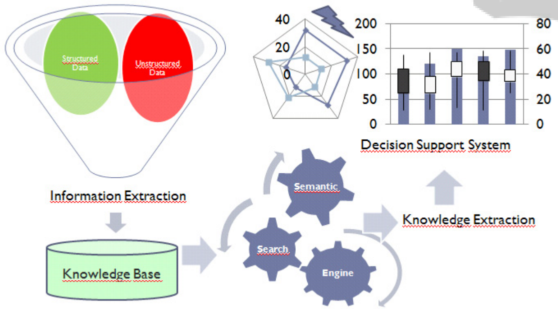

Figure 1: General Architecture of Ontology Based Decision Support System.

6 EXPECTED OUTCOME

One of the outputs of this work is -ontology, can be used in many areas of finance. Ontology will be used in “Ontology Based Decision Support System with Application to Public Finance”. This is a project carried out by the MoF using traditional decision making tools such as cost benefit analysis.

The aim of all these studies is to increase the performance of finance and to efficiently use the resources. The proposed product will be used as a part of a decision and support system with business intelligence capabilities in public finance. So that, the managers will give better decisions based on reliable data and assess the impact of their decisions in intelligent and simulated environment before applying it.

After the completion of these studies, decision makers in finance will have chance to see more objective metrics and decision of project proposal (whether to accept or reject), decision of ongoing project (whether to continue or stop- whether to allocate more money or not) will be easier. Another decision can be made according to metrics if there is some need for some type of investments can be seen easily.

ACKNOWLEDGEMENTS

This work is partially supported by Ministry of Science, Industry and Technology in the program of with grant number 0265.STZ.2013-2 and Stratek (Strategic Technologies Research and Development) company.

REFERENCES

Alan, O., Akpinar, S., Sabuncu, O., Cicekli, N., and Alpaslan, F. (2008, October). Ontological video annotation and querying system for soccer games. In*Computer and Information Sciences, 2008. ISCIS'08. 23rd International Symposium on* (pp. 1-6). IEEE.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., and Shadbolt, N. R. (2003). Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE, 18* (1), 14-21.

Baclawski, K., Cigna, J., Kokar, M. M., Mager, P., and Indurkhya, B. (2000). Knowledge representation and indexing using the unified medical language system. In *Pac Symp Biocomput* (pp. 493-504).

Buitelaar, P., Cimiano, P., Racioppa, S., and Siegel, M. (2006). Ontology-based information extraction with soba. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC).*

Cimiano, P. (2006). *Ontology learning from text* (pp. 19-34). Springer US.

Corcho, O., Fernández-López, M., and Gómez-Pérez, A. (2003). Methodologies, tools and languages for building ontologies. Where is their meeting point?. *Data and knowledge engineering, 46* (1), 41-64.

Gardarin, G., Kou, H., Zeitouni, K., Meng, X., and Wang, H. (2003, June). SEWISE: An Ontology-based Web Information Search Engine. In *NLDB* (Vol. 2003, pp. 106-119).

Hahn, U. and Schattinger, K. (1998). Towards Text Knowledge Engineering. In Proceedings of the 15th National Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, pp. 524-531.

Jones, D., Bench-Capon, T., and Visser, P. (1998). *Methodologies for ontology development*. In Proc. ITandKNOWS Conference of the 15th IFIP World Computer Congress. 1998. pp. 20-35.

Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N. K., and Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing.*Information Systems*, *37* (4), 294-305.

Kozaki, K., Kitamura, Y., Ikeda, M., and Mizoguchi, R. (2002). Hozo: an environment for building/using ontologies based on a fundamental consideration of “Role” and “Relationship”. In *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web* (pp. 213-218). Springer Berlin Heidelberg.

Madhu, G., Govardhan, D. A., and Rajinikanth, D. T. (2011). Intelligent Semantic Web Search Engines: A Brief Survey. *arXiv preprint arXiv:1102.0831*.

Mädche, A., Motik, B., Stojanovic, L., Studer, R., and Volz, R. (2003, May). An infrastructure for searching, reusing and evolving distributed ontologies. In *Proceedings of the 12th international conference on World Wide Web* (pp. 439-448). ACM.

Noy, N. F., and McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology.

Topic Ontology Population, http://semanticweb.org/wiki/ Category: Topic_Ontology_population. 8 March 2014.

Vicknair, C., Macias, M., Zhao, Z., Nan, X., Chen, Y., and Wilkins, D. (2010, April). A comparison of a graph database and a relational database: a data provenance perspective. In *Proceedings of the 48th annual Southeast regional conference* (p. 42). ACM.

Wimalasuriya, D. C., and Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*. “And Now for Something Completely Different: Using OWL with Neo4j” http://neo4j.com/blog/and-now-for-something-completely-different-using-owl-with-neo4j/