Computational Modelling for Bankruptcy Prediction: Semantic data Analysis Integrating Graph Database and Financial Ontology

Natalia Yerashenia[1] Alexander Bolotov[2]

[1] University of Westminster, London, United Kingdom, w1578366@my.westminster.ac.uk
[2] University of Westminster, London, United Kingdom, A.Bolotov@westminster.ac.uk

Abstract: In this paper, we propose a novel intelligent methodology to construct a Bankruptcy Prediction Computation Model, which is aimed to execute a company's financial status analysis accurately. Based on the semantic data analysis and management, our methodology considers Semantic Database System as the core of the system. It comprises three layers: an Ontology of Bankruptcy Prediction, Semantic Search Engine, and a Semantic Analysis Graph Database.

1 Introduction

We propose a concept of an intelligent, analytical system to perform the prediction of the companies' bankruptcy. The system processes financial information of a company and undertakes a comprehensive investigation of companies' financial activities during a particular designated time period. We aim at creating a *Bankruptcy Prediction Computational Model (BPCM)* which is capable of the automated construction of an expert analytical report, where various data and information are presented reliably and objectively.

The main feature of the proposed system is the consolidation of the information management with the decision-making process to serve the prediction. This involves modern methods of searching, processing and storing potentially large amount of heterogeneous data together with advanced machine learning methods. In this paper, we define the process of the *Semantic Database System* construction, a novel development, which comprises *an Ontology of Bankruptcy Prediction (OBP)*, the *a Semantic Search Engine (SDS)* and *a Semantic Analysis Graph Database (SAGRADA)*.

The remaining of the paper is organised as follows. In Section 2 we describe the problem set-up of the SDS. Section 3 describes the architecture of the Ontology and Graph Database and illustrates their functionality. Finally, in Section 4 we summarise the contributions of the research provided, discuss future work, and draw conclusions.

2 Problem Set-up

We argue that the financial dataset to be analysed for our purposes, figuratively speaking, can be characterised by four 'V' and 'R'. It shares most (four out of five big 'V') of the qualities of *Big Data* – Variety, Velocity, Veracity and Value [9] being not dependent on the Volume. However, we underline the fifth, 'R', feature of these financial data – an extremely high level of Relationships. Indeed, similar to big data, in our case, we have heterogeneous data, coming from different sources. These components of a company's financial system can be (and usually this is the most common practice) described in the form of *relational tables* (traditional database), e.g. it is easy to present a balance sheet or income statement in such a way. However, to show the interconnections between all elements of these tables, it is necessary to create a number of tables of a different structure containing thousands of objects. In this case, the efficiency of database management and search are substantially affected. For example, it becomes problematic to formulate a general query to several databases, because of the difference in objects and attributes of the domain or changes in objects over time. When the data are inserted, updated or deleted, the integrity constraints for the database with changing objects should be checked and assured that the data will be consistent after all modifications [7]. Also, there is a problem of the integration of new nodes into the system. When adding a new node, it is essential to check the data and the data schema for consistency with the information already available in the system [3].

Although traditional relational databases still dominate among data storage facilities, these systems would not be suitable for the purposes of our financial analysis being unable to tackle the requirements of the *'Big Four V + R'*. There are *NoSQL systems* that extend the capabilities of traditional databases by allowing to deal with the four 'V'.

3 Components of Semantic Database System

Developing a Financial Ontology. The ontology presentation format defines the mechanisms to store concepts and their relationships in the library; it is a method of transmitting ontological descriptions to other consumers and a method of processing its concepts. Specific ontology presentation languages have been developed as ontological description formats (OWL, RDF, KIF) [6].

Ontologies are used as data sources for many software applications such as information retrieval, text analysis, knowledge extraction, and other information technologies, allowing more efficient processing of complex and diverse information. This way of representing knowledge enables applications to recognise those semantic differences that are obvious to people but not known to the computer [2].

The main and most crucial component of the financial risk management of a company is the knowledge base.

Our approach to building an ontology describes the basic concepts of financial analysis, as well as the objects that serve as sources of knowledge for predicting a company's bankruptcy. It also contains the concepts and relationships required for the formation of a hierarchy of knowledge fields and the subsequent use of this hierarchy by various applications. In addition, expert rules and regulations can be described in terms of ontology, which significantly increases their level of succinctness and transparency for the users.

The structure and the content of the OBP are based on the experience of analysts specialising in the theory and practice of bankruptcy prediction [1]. This hierarchy reflects a number of the most popular indicators used to conduct a financial analysis of a company, as well as their origin (documents and concepts to which they relate) and the relationship of these indicators to each other. Financial analytic factors form the penultimate row of the hierarchy, while the principal generalising object is the concept of Company's Financial Records. The last row in the hierarchy contains linguistic variables that will be later involved in the development of machine learning computational modules.

The working version of the OBP is an informal conceptual representation model, which is an initial step of the proposed approach.[1]

Developing a Graph Database. *Graph DB* (for instance, Neo4J) are an example of *NoSQL databases* aimed at representing semantical data [4]. Graph databases are used for storing, processing and automated visualisation of standard structural elements. A typical Graph DB usually contains some reference information regarding objects [10]. Therefore, the user/designer does not have to spend time searching fo this information in the DB directories. It also reduces the number of possible human factor related errors. Graph DB enables to create standard elements automatically, which significantly reduces the design time [8].

Neo4j[2] is an open source Graph Database management system implemented in Java. This Graph DB environment stores data in a proprietary format specifically adapted for the presentation of graph information; this approach, in comparison with the modelling of a graph database, using a relational *Databases Management Systems (DBMS)*, allows for additional optimisation in the case of data with a more complex structure. Neo4j uses its own query language, Cypher[3], though the queries can be done in other ways, for example, directly through the Java API. Cypher is not only a query language but also a data manipulation language, as it provides CRUD functions for graph storage.

We emphasise that the OBP structure is an excellent basis for the Semantic Analysis Graph Database which is used as a repository of the financial data for BPCM model. So, we intend to apply an existing solution of creating and managing Graph Databases and integrate it into our novel approach.

We have implemented a prototype Graph DB, SAGRADA, in Neo4j. The basic concepts in a Graph DB are nodes (an object of the database), relations (graph edges) and their properties. In our case, the nodes of the graph are financial ratios, financial indicators, and the documents containing them. Our graphical repository has 29 nodes divided into three categories – Ratio, Criteria (financial indicator), Statement, and 52 relationships between them (of two types – direct and inverse).

[1] At the moment, our work concerns with supplementing the structure of this ontology, as well as with the development of its formal physical representation model utilising the OWL/RDF environment.

[2] `https://neo4j.com/product/`

[3] `https://neo4j.com/developer/cypher/`

4 Conclusions

Based on the analysis of various modern approaches to the processing and storage of the heterogeneous data related to the financial analysis, we proposed a novel intelligent methodology to construct a Bankruptcy Prediction Computational Model. Our methodology is based upon the utilisation and integration of the semantic data management methods. Following this methodology, we have introduced a novel layered architecture for this Computational Model, which integrates the Semantic Database System and a set of modern machine learning algorithms. We have implemented the principles of the new Ontology of Bankruptcy Prediction and the Semantic Analysis Graph Database on the example of a company financial record.

Further, we will improve the structure of the OBP Ontology creating its formal conceptual representation through OWL / RDF languages. We will also work on further enhancement of the SAGRADA itself. We will also tackle a problem of the data exchange between the structural parts of the SDS finding a way to transfer data in various directions automatically.

References

[1] Jodi L Bellovary, Don E Giacomino, and Michael D Akers. A review of bankruptcy prediction studies: 1930 to present. *Journal of Financial education*, pages 1–42, 2007.

[2] Helena Dudycz and Jerzy Korczak. Conceptual design of financial ontology. In *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1505–1511. IEEE, 2015.

[3] Michel Gagnon. Ontology-based integration of data sources. In *2007 10th International Conference on Information Fusion*, pages 1–8. IEEE, 2007.

[4] Venkat N Gudivada, Dhana Rao, and Vijay V Raghavan. Nosql systems for big data management. In *2014 IEEE World congress on services*, pages 190–197. IEEE, 2014.

[5] Xiufeng Liu, Nadeem Iftikhar, and Xike Xie. Survey of real-time processing systems for big data. In

Proceedings of the 18th International Database Engineering & Applications Symposium, pages 356–361. ACM, 2014.

[6] Xiufeng Liu, Nadeem Iftikhar, and Xike Xie. Survey of real-time processing systems for big data. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, pages 356–361. ACM, 2014.

[7] Ghonia Pooja Mansukhlal and C Malathy. Semantic integration with ontology based approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 2257–2260. IEEE, 2016.

[8] Jaroslav Pokornỳ. Graph databases: their power and limitations. In *IFIP International Conference on Computer Information Systems and Industrial Management*, pages 58–69. Springer, 2015.

[9] Xizhao Wang and Yulin He. Learning from uncertainty for big data: Future analytical challenges and strategies. *IEEE Systems, Man, and Cybernetics Magazine*, 2(2):26–31, 2016.

[10] Zuopeng Justin Zhang. Graph databases for knowledge management. *IT Professional*, 19(6):26–32, 2017.