Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp. 145-148 recibido 01-11-2016 revisado 12-05-2017 aceptado 18-05-2017

KBS4FIA: Leveraging advanced knowledge-based systems for financial information analysis

KBS4FIA: Sistema inteligente basado en conocimiento para análisis de información financiera

Francisco García-Sánchez
Mario Paredes-Valverde
Rafael Valencia-García
Universidad de Murcia
Facultad de Informática
Campus de Espinardo, 30100,
Murcia, España
{frgarcia, marioandres.paredes,
valencia}@um.es

Gema Alcaraz-Mármol
Departamento de Filología
Inglesa, Universidad de
Castilla-La Mancha
Avda. Carlos III, s/n,
45071, Toledo, España
gema.alcaraz@uclm.es

Ángela Almela
Centro Universitario de la
Defensa (Universidad
Politécnica de Cartagena)
Base Aérea de San Javier,
30720, Santiago de la Ribera,
Murcia, España
angela.almela@cud.upct.es

Abstract: Decision making takes place in an environment of uncertainty. Therefore, it is necessary to have information which is as accurate and complete as possible in order to minimize the risk that is inherent to the decision-making process. In the financial domain, the situation becomes even more critical due to the intrinsic complexity of the analytical tasks within this field. The main aim of the KBS4FIA project is to automate the processes associated with financial analysis by leveraging the technological advances in natural language processing, ontology learning and population, ontology evolution, opinion mining, the Semantic Web and Linked Data. This project is being developed by the TECNOMOD research group at the University of Murcia and has been funded by the Ministry of Economy, Industry and Competitiveness and the European Regional Development Fund (ERDF) through the Spanish National Plan for Scientific and Technical Research and Innovation Aimed at the Challenges of Society.
Keywords: Knowledge acquisition, ontologies, opinion mining, natural language processing, linked data

Resumen: La toma de decisiones tiene lugar en un ambiente de incertidumbre, por lo tanto es necesario disponer de información lo más exacta y completa posible para minimizar el riesgo inherente al proceso de toma de decisiones. En el dominio de las finanzas la situación se hace, si cabe, aún más crítica debido a la complejidad intrínseca de las tareas analíticas dentro de este campo. La finalidad del proyecto KBS4FIA es la automatización de los procesos ligados al análisis financiero, utilizando para ello tecnologías asociadas con el procesamiento del lenguaje natural, el aprendizaje, la instanciación y la evolución de ontologías, la minería de opiniones, la Web Semántica y el Linked Data. Este proyecto está siendo desarrollado por el grupo TECNOMOD de la Universidad de Murcia y ha sido financiado por el Ministerio de Economía y Competitividad y el Fondo Europeo de Desarrollo Regional (FEDER) a través del Programa Estatal de I+D+i Orientada a los Retos de la Sociedad.
Palabras clave: Adquisición de conocimiento, ontologías, minería de opiniones, procesamiento del lenguaje natural, linked data

1 Introduction and main goal

The need to manage financial data has been increasingly coming into sharp focus for some time. Years ago, these data sat in warehouses attached to specific applications in banks and financial companies. Then the Web came into the arena, generating the availability of diverse data sets across applications, departments and other financial entities. However, throughout

ISSN 1135-5948

these developments, a particular underlying problem has remained unsolved: data reside in thousands of incompatible formats and cannot be systematically managed, integrated, unified or easily processed.

In a larger context, the abovementioned problem may be multiplied by millions of data structures located in thousands of incompatible databases and message formats. This problem is getting worse as techniques for the processing of financial domain big data continue to gather more data, reengineer massive data processing methods, and integrate with more sources. Moreover, financial analysis and, specifically, stock market price prediction is regarded as one of the most challenging tasks of financial time series prediction. The difficulty of forecasting arises from the inherent non-linearity and non-stationarity of the stock market and financial time series (Kazem et al., 2013). In the last few years, different data mining technologies such as neuronal networks or support vector machines have been applied to solve this problem, but satisfactory results have not been achieved (Rodríguez-González et al., 2011).

The present project aims to develop new knowledge-empowered methods for financial analysis based on the Semantic Web, ontology learning, deep learning, and natural language processing technologies. Specifically, our project is centred on different research areas such as knowledge acquisition and representation from natural language documents, subjective natural language processing and deep learning technologies.

2 Project status

Thus far, a comprehensive analysis of the state of the art in the different topics involved in this project has been carried out. The key technologies that we have identified for the project comprise: (1) ontology models and linked data from the Semantic Web area for the financial domain; (2) knowledge acquisition from natural language texts; and (3) subjective language analysis.

2.1 Ontology models and linked data for the financial domain

Several ontologies in the financial context have been generated in the last few years, such as, for example, the BORO (Business Object Reference Ontology) ontology (Partridge, Partridge, and Stefanova, 2001), and the ones developed by the XBRL ontology Specification Group (XBRL International, n.d.).

Regarding information data sources, the emergence of the Open Data movement has contributed to the distribution without restrictions of relevant financial, economic, and business data across the Web, which can be consumed from software agents and applications, as well as by the users behind them. The Open Data approach has been adopted throughout the world, and governments in over 40 countries have established data-publishing sites aiming to ensure transparency in government activities, encourage secondary use of public data and create new markets.

In view of the aforementioned facts, integrating financial information, as well as performing a faster and more accurate analysis across these disparate financial information sources, a fundamental challenge remains. In order to address this challenge, it is necessary an approach that enables to connect and consume large quantities of data sources in a faster way as well as to perform a more accurate data analysis. In this sense, Semantic Web technologies are deemed as a promising mechanism for sharing large quantities of data via the Web, due to the fact that it provides Web information with a well-defined meaning and make it understandable not only by humans but also by computers (Shadbolt, Berners-Lee, and Hall, 2006), thus allowing these machines to automate, integrate and reuse high-quality information across several applications.

2.2 Knowledge acquisition

The Semantic Web arose with the aim of adding meaning to the data published on the Web. Ontologies constitute the technological key that allows for the representation of static knowledge in order to be shared and reutilized. The manual construction of ontologies is a hard and costly process which requires time and resources. In order to avoid this process, in the last years many studies about automatic construction and updating of ontologies have been carried out (Gil Herrera and Martín-Bautista, 2015). We can distinguish three main categories: ontology learning, ontology population, and ontology evolution.

On the other hand, nowadays there are a large number of public knowledge bases promoted by the best practices in order to publish and connect structured data on the Web.

This is known as Linked Data (http://linkeddata.org/). A serious problem found in this field points to the need of tools to access and consume the huge amount of knowledge that is available. In order to extract information from those knowledge bases, users need to know: (1) ontology languages, (2) some formal query language (e.g. SPARQL), and (3) the structure of the ontology vocabulary. That is why there have appeared different natural language interfaces (NLIs) or question-answer systems aiming to make access to ontology knowledge easier by hiding their formality and their language of search (Lopez et al., 2013).

2.3 Subjective language analysis

Sentiment analysis has become a popular topic towards the understanding of public opinion from unstructured Web data. In this sense, sentiment analysis is devoted to extracting users' opinions from textual data. The capture of public opinion is gaining momentum, particularly in terms of product preferences, marketing campaigns, political movements, financial aspects and company strategies. The focus of opinion mining is not on the topic of a text, but rather on what opinion that text expresses (Esuli and Sebastiani, 2005). It determines whether the comments in online forums, blogs or the like related to a particular topic (product, book, movie, company, etc.) are positive, negative or neutral. Opinions are very important when someone wishes to hear others' views before making a decision.

Recent studies attempting to create an automated system that performs an effective sentiment analysis have based their works on two main approaches: Semantic Orientation and Machine Learning. Moreover, several studies have been conducted in recent years in order to improve sentiment classification. These approaches work at different levels: document-level, sentence-level, and feature-level. The major issue with using these techniques is that a model that works well for opinion mining in one domain might not provide satisfactory results in others. To overcome such an issue, deep learning technologies are currently being successfully applied (Glorot, Bordes, and Bengio, 2011). At the same time, most of the studies on opinion mining deal exclusively with English documents, perhaps due to the lack of resources in other languages (Martín-Valdivia et al., 2013). An important aspect on which subjectivity and sentiment analysis require further efforts is the analysis of multilingual texts.

One of the main problems concerning the financial and, specifically, stock-market analysis is that current approaches have not been designed to consider external factors that are extremely relevant in order to quantify the impact of these factors. Some studies about the relationship between public sentiment and stock prices have been published in the last few years (Li et al., 2014). Besides, financial language is inherently complex, since financial terms refer to an underlying social, economic and legal context (Milne and Chisholm, 2013). Consequently, not many sentiment analysis approaches have been validated in the financial domain and the results obtained are unpromising (Salas-Zárate et al., 2017).

3 Future work

Current technologies present several limitations and challenges, which we aim to deal with and solve in this project. Our overall aim is to overcome those drawbacks by exploring, developing and validating knowledge-based technologies and natural language processing technologies for financial analysis.

The goal is to develop a knowledge-based empowered financial monitoring and management platform to provide relevant sentiment data associated to economic structures. In a nutshell, the system will be designed to gather financial structured and unstructured data from various distinct sources such as social media, to enrich them semantically with annotations and to store them in a repository. Information gathering will be carried out by extracting data from both public information sources such as the Internet (forums, blogs, news), from corporate private information sources (i.e. from corporate sites), and from linked and open data sources.

Furthermore, the project aims to provide an innovative technique for analysing financial information through sentiment analysis of natural language texts such as financial news, blogs or tweets in different languages. The unstructured information will be extracted from online resources such as users' opinions, and will detect whether the analysed text is related to the financial domain or not.

Then, opinion mining techniques will be applied over this filtered information to obtain

polarity and reputation analysis. It will also be able to obtain relations between companies, sectors and geographical areas. With all this information stored, the system will be able to classify all of it, depending on the reputation of the source, or on the quantity of coincidences in an opinion. These factors will determine the weight of acquired data. The opinion of the ECB, the Federal Reserve, the IMF chairmen or Nobel Prizes in Economics carries more weight than a couple of journalists in local media or some anonymous comment in a forum. Furthermore, the system will attach more importance to an estimate if it is supported by a large number of different sources.

Finally, users will then be provided with different services for data access. These services will take advantage of the machine-readable semantic annotations of the financial information to provide more sophisticated high-quality functionality to the system's users. In particular, the results generated by the platform could be shown to end users in a number of ways: reports, business diagrams, dashboards, and personalized recommendations.

Acknowledgments

This project has been funded by the Spanish National Research Agency (AEI) and the European Regional Development Fund (FEDER / ERDF) through project KBS4FIA (TIN2016-76323-R).

References

Esuli, A. and F. Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05* (pages 617–624). New York, New York, USA: ACM Press.

Gil Herrera, R. J. and M. J. Martín-Bautista. 2015. A novel process-based KMS success framework empowered by ontology learning technology. *Engineering Applications of Artificial Intelligence*, 45: 295–312.

Glorot, X., A. Bordes and Y. Bengio. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. In L. Getoor and T. Scheffer (Eds.), *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pages 513–520). New York, NY, USA: ACM.

Kazem, A., E. Sharifi, F. K. Hussain, M. Saberi and O. K. Hussain. 2013. Support vector regression with chaos-based firefly algorithm for stock market price forecasting. *Applied Soft Computing*, 13(2): 947–958.

Li, X., H. Xie, L. Chen, J. Wang and X. Deng. 2014. News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69: 14–23.

Lopez, V., C. Unger, P. Cimiano and E. Motta. 2013. Evaluating question answering over linked data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 21: 3–13.

Martín-Valdivia, M. T., E. Martínez-Cámara, J. M. Perea-Ortega and L. A. Ureña-López. 2013. Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches. *Expert Systems with Applications*, 40(10): 3934–3942.

Milne, A. and M. Chisholm. 2013. *The Prospects for Common Financial Language in Wholesale Financial Services*. *SWIFT Institute Working Paper No. 2012-005*.

Partridge, C. and M. Stefanova. 2001. A Synthesis of State of the Art Enterprise Ontologies. In *LESSONS LEARNED. 2001, THE BORO PROGRAM, LADSEB CNR*.

Rodríguez-González, A., A. García-Crespo, R. Colomo-Palacios, F. Guldrís Iglesias and J. M. Gómez-Berbís. 2011. CAST: Using neural networks to improve trading systems based on technical analysis by means of the RSI financial indicator. *Expert Systems with Applications*, 38(9): 11489–11500.

Salas-Zárate, M. del P., R. Valencia-García, A. Ruiz-Martínez and R. Colomo-Palacios. 2017. Feature-based opinion mining in financial news: An ontology-driven approach. *Journal of Information Science*, 43(4): 458-479.

Shadbolt, N., T. Berners-Lee and W. Hall. 2006. The Semantic Web Revisited. *IEEE Intelligent Systems*, 21(3): 96–101.

XBRL International. (n.d.). XBRL: eXtensible Business Reporting Language. Retrieved May 22, 2017, from https://www.xbrl.org/