Semantic Technology for Financial Awareness

Anna Satsiou[1], Artem Revenko[2], Ioannis Praggidis[3], Eirini Karapistoli[3], Georgios Panos[4], Christoforos Bouzanis[4], Ioannis Kompatsiaris[1]

[1]Information Technologies Institute, CERTH, 57001 Thessaloniki, Greece

[2]Semantic Web Company, Mariahilfer Straße 70/Neubaugasse 1, 1070 Vienna, Austria

[3]Dept. of Economics, Democritus University Thrace, 69100 Komotini, Greece

[4]Adam Smith Business School, University of Glasgow, Glasgow G12 8QQ

satsiou@iti.gr, a.revenko@semantic-web.at, gpragkid@ierd.duth.gr, ikarapis@duth.gr, Georgios.Panos@glasgow.ac.uk, Christoforos.Bouzanis@glasgow.ac.uk, ikom@iti.gr

ABSTRACT

This paper sheds light on how semantic technology can be a key driver for promoting financial awareness and capability, through the presentation of the ideas behind the PROFIT platform and its functionalities. The PROFIT platform is designed by the newly funded namesake EU project, as a solution catering to the exact need of action enhancement for greater financial awareness and capability that has been identified as a major target for improved social performance, personal and household finance protection and, ultimately, greater societal well-being. Towards this goal, the platform will provide the following functionalities that are based on semantic technology: a) a novel financial education toolkit available to the wider public (b) advanced crowd-sourcing tools to process financial data, extract and present collective knowledge, (c) financial forecasting models exploiting the market sentiment to identify market trends and threats, (d) novel personalized recommendation systems to support financial decisions according to the user's profile (financial literacy level, interests, demographic characteristics etc.) (e) a reputation-based incentive scheme to motivate good quality contributions to the platform.

CCS Concepts

• Information systems□ Information systems applications, World Wide Web, Information Retrieval • Applied Computing□ Education• Theory of Computation□ Semantics and Reasoning

Keywords

Linked open data; knowledge discovery; terminology, thesaurus & ontology management; financial education; financial forecasting; personalized recommendations; societal aspects.

1. INTRODUCTION

The recent financial crisis has generated interest in better understanding how to promote more responsible and prudent individual saving and borrowing behaviour [1]. The Eurozone debt crisis, with the resulting fiscal consequences throughout Europe, further stimulated the discussion on how to promote the efficient allocation of financial resources and greater financial stability [2]. The ability of citizens to make informed financial decisions is critical to developing sound personal finance, which can contribute to increased saving rates, more efficient allocation of financial resources, and greater financial stability. From the EU's viewpoint, it can also be conducive to the enhancement of the European identity, open democracy and active citizenship for informed political awareness.

SEMANTICS 2016: Posters and Demos Track
September 13-14, 2016, Leipzig, Germany

Technological developments have enabled and enhanced the availability of large volumes of information on themes relevant to financial decision making. Their potential benefits, though, are hindered by the cognitive limitations by individuals when it comes to the processing of large volumes of information, as well as the documented widespread financial illiteracy even within developed economies, including those of the European Union.

Acknowledging such needs, the newly-funded EU project PROFIT (Promoting Financial Awareness and Stability) [3] brings together researchers and professionals with expertise across accounting and finance, economics, information technology, computer/software engineering and education, along with a range of private, third-sector and institutional partners, to develop a financial awareness and capability platform for improved social performance, client protection and, ultimately, greater financial stability and societal well-being.

PROFIT platform will be built on Open Source components with the following functionalities: (a) novel financial education tools and toolkits available to the wider public (b) advanced crowd-sourcing tools to process financial data, extract and present collective knowledge, (c) financial forecasting models accounting for market sentiment to identify market trends and threats, (d) novel personalized recommendation systems to support financial decisions according to the user's profile (financial literacy level, interests, demographic characteristics etc.) (e) incentive mechanisms to encourage the active participation of citizens through many different channels like posting and rating of financial articles, Q&A, voting in relevant polls, etc.

Semantic technology is a pivotal part of the PROFIT platform architecture as it can be seen in Figure 1, supporting all the functionalities of the platform. A continuously evolving semantic knowledge graph is used to describe entities, relations and notions relevant to financial awareness, recommendation services, reputation mechanism and to enhance sentiment analysis, as will be described in the next sections.

2. RELATED WORK

There is wide range of websites that aspires to the term financial awareness platform ranging from yahoo personal finance to classical economic media (i.e., the Wall Street Journal, the CNBC, Forbes, etc.), general media (i.e., the New York Times, CNN, etc.), stock exchange and securities websites (i.e., The New York Stock Exchange, Börse Franfurkt, etc.) as well as encyclopedic websites (i.e., investopedia.com, mymoney.gov, about.com, etc.). More specialized financial literacy education toolkits are still at early stages of conception and availability with existing primary examples involving the Workplace Financial Fitness Toolkit (WFFT), funded by NYSE and the Euronext Foundation and

relevant programme by banks and financial institutions, e.g. the Danske Bank. Moreover, most of the existing financial education training courses have limited access rights and target audiences, such as the young, university students etc. Thus, to a certain extent, their novelty and impact so far has yet to be established.

Departing from these observations, PROFIT will provide a holistic approach to financial awareness establishing a user-centred financial awareness platform that is based on semantic technology obtaining finance-related crowdsourced data from the web and its users in order to create new knowledge towards offering financial information, education and advanced forecasting tools to help users understand financial data and trends and empower them in decision making. The platform will be freely available based on Open software components and licensed under suitable Open license to be available for the general use and re-usage. To the best of our knowledge, such a multi-functional platform for the user is not available in the European environment so far. Moreover, the **PROFIT platform will be among the first Open Source paradigms to illustrate and analyse the possibilities of exploiting the semantic technology in the financial awareness domain showcasing the benefits therein as well as the problems and obstacles to overcome in the future.**

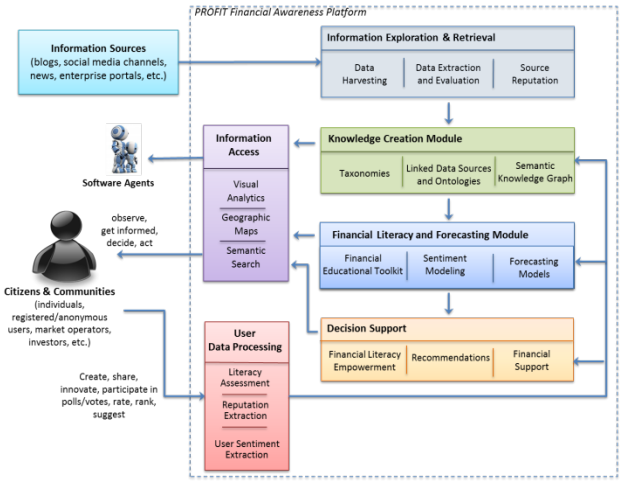

Figure 1. High-level Architecture of PROFIT platform.

3. SEMANTIC FRAMEWORK FOR PROFIT PLATFORM

The semantic technologies enable users to process information, and make inferences based on conceptualizations and observations from data [4]. In the case of financial and educational data, the semantic connection between a term in a resource (e.g., a book) and the same term in a world-wide context is vital for understanding the context and, therefore, the real meaning of the term [5,6]. In PROFIT the data model consists of several ontologies and thesauri, each serving a special purpose. Following the best practices of the semantic web, the prepared controlled vocabularies reuse the well-known existing ones. In order to enable the full power of semantic technologies PoolParty (www.poolparty.biz) is used for managing controlled vocabularies, linked data, and processing texts. The ontologies are available at http://profit.poolparty.biz/PROFIT-User.html and http://profit.poolparty.biz/PROFIT-Finance.html, respectively.

A number of use cases and scenarios demonstrating how the PROFIT platform and the respective semantic technologies can be used and exploited are listed in [23].

3.1 User Ontology

The goal of the user ontology is to acquire, process and assign the user information that could be relevant for further analysis in the platform: understanding user interests; finding out user skills and knowledge, eliciting user attitudes. For this purpose, an ontology is designed and created for the PROFIT platform that contains users' actions, relations, creative works, contributions in the platform, and attitudes.

Users are represented by the "Person" class taken from Schema.org [5] ontology. User actions fall into "Action" and "InteractAction" classes taken from Schema.org as well. These actions are connected to different subclasses of CreativeWork: "EducationalMaterial", "InteractiveCreativeWork", and "InteractiveEducationalMaterial". Further on, users may express "Opinion" (taken from the MARL ontology [6]).

The ontology is accompanied by a thesaurus of particular notions, i.e. different possible instances of the ontology classes. For example, the particular instances of creative works are news articles, blog posts, polls, etc. The possible user's actions are create, rate, download, comment, etc. The data model is designed to be expandable with as little effort as possible: if a new action or a new type of creative work becomes available there is no need to change the data model since the new instance can be added to the thesaurus in the respective concept scheme.

3.2 Financial Ontology

Generating a financial ontology is no easy task. The goal of the financial ontology is to describe and capture the differences between various textual resources on finance, and on personal finance in particular. The ontology contains main classes of financial objects, such as economic agents, financial systems, economic variables and factors. The relations capture the notions of being dependent on something, owning something, being a member or having a representative.

The classes constitute a high level structure for the description of financial objects' content. Some helper classes and relations are taken from other ontologies: Dublin Core Metadata [7], FIBO Foundations [8], and Schema.org [5]. The ontology is well suited to capture the main features without further details, therefore offering a broad range of usages. In the future steps of the project, it is planned to check how well the created ontology is able to capture the differences and the similarities of different textual financial resources and possibly to extend the model.

The thesaurus of particular financial instances is expected to be much more complex than the one of the user data; it should contain a vast amount of financial concepts to facilitate the annotation and classification of textual resources. In order to prepare such a thesaurus, a novel procedure is being implemented, that is described in the following subsection.

3.2.1 Financial Thesaurus

Corpora are widely used to facilitate the construction of thesauri [9, 10]. In order to facilitate the preparation of PROFIT financial thesaurus, corpora are analysed. Several topics of interest are chosen, including the GDP of the Eurozone, the Eurostoxx50, the oil price and the Eurodollar exchange rate.

The corpora are extracted from an immense archive at the website "www.investing.com". The archive contains news articles from the sources: "Investing.com", "Reuters", "FinanceMagnates", "International Business Times", "ecPulse", "LFB Forex". Three

corpora were fetched: the corpus related to oil prices containing 14209 articles, the corpus related to Eurodollar exchange rate containing 19119 articles, and the corpus related to Eurostoxx50 containing 5834 articles. The news articles date back to 2010.

As next step, the STW Economics is chosen as the initial thesaurus by the experts in the field. The thesaurus contains 6521 concepts [11]. The fetched news articles are annotated against the thesaurus. As the result of annotation, the following important data are obtained:

1. For each concept in thesaurus: a) number of times extracted, b) depth from bottom, c) number of times children nodes are extracted.
2. For each document from corpora: a) terms found in this document that are top ranked in PoolParty, b) number of concepts extracted and distances between these concepts. Distance is a function of the shortest path and depth.
3. Overall statistics: a) average and standard deviation of the number of times each concept extracted, b) average and standard deviation of the number of extracted concepts in documents.

Based on these data, suggestions about concepts, and branches of concepts that are rarely used (a concept together with its children) are made. Based on these suggestions, an expert obtains a list of concepts that could be deleted or extended with new labels so that they could be extracted more often. The expert also receives a preselected collection of texts that are poorly annotated. Each text is supplied with a list of free terms that are significant for the whole corpus and are present in the text; therefore, he is enabled to improve the annotation of the text via adding and converting the suggested free terms into new concepts.

4. PROFIT PLATFORM FUNCTIONALITIES

4.1 Financial Education

One of the basic functionalities of the PROFIT platform is to provide a financial educational toolkit aimed at providing individuals with financial knowledge, skills and capacity to use resources and tools (including financial products and services) in order to make informed financial decisions. This toolkit will be a novel collection of primary financial education material. Links will also be provided to relevant practical assessments and information and other tools, such as videos, narratives and games, that the users will be able to utilise needed based on their requirements, goals and performance.

The toolkit will employ a wide array of personal finance material and information-acquisition methods, ranging from taxonomies of standard concepts and definitions (as described in section 3), to customized collections of content for broad specific user cases. The latter will target user audiences, such as entrepreneurs, latent entrepreneurs, the elderly, the young, the immigrants, etc. Further customisations in the content will be informed based on the updated literature, suggesting that e.g. females, trainees, and active citizens are likely to have different financial-literacy enhancement requirements [2].

The use of semantic technologies emphasizing on personal finance and personal financial planning, along with responsible banking and finance, can be seen as advantages of the PROFIT platform. So can be the potential emphasis on the public communication of economics and finance, i.e. on issues on public finance, for the formation of public attitudes e.g. [22].

4.2 Financial Forecasting

The forecasting component of PROFIT aspires to improve to some extent the financial decision making process of the platform users by providing information about future movements of core financial and economic indexes. This includes newly generated information using semantics [14] along with structured numerical data as well as existing information from forecasting projections of well-established organisations (such as ECB, IMF, and the EC).

The forecasted series to be analysed include the GDP of the Eurozone, the Eurostoxx 50, which is Europe's leading Blue Chip Index for the Eurozone, the oil price and the Eurodollar exchange rate. The selection of these series is based on the interconnectivity of financial markets and economic activity [15]. According to the seminal paper of Tetlock [14], both the market sentiment and the uncertainty about future developments have gained significant ground as predictors for upcoming recessions [16]. This gives us the intuition to extract valuable information about these terms by exploiting unstructured information through semantic-based text classification and analysis from selected sources on the web. Within PROFIT, we will rely on some form of sentiment analysis to turn the text into data for analysis via application of natural language processing (NLP) and analytical methods [17].

Although there is a proliferation with regard to the proposed methods and datasets for forecasting, results could be characterized rather poor or they cannot provide accurate estimations for a long time. This may be due to the fact that many models employ specifications that impose restrictions not letting the system to move freely. Utilising the recent literature as a baseline and relying on the strengths of semantic technology, we will attempt to overcome this problem by utilizing various methods. Among these methods, a new dataset containing semantically enhanced information about the market sentiment and the uncertainty, as described in the previous paragraph, will be used in order to further enhance the forecasting power of our semantically aware prediction models.

4.3 Personalised Recommendations

Another functionality of the PROFIT platform is the provision of personalized recommendations to the users according to their interests and needs. Recommendations will be mainly about content/articles/posts and educational material of interest, and about users with similar preferences. Semantic technology will support this functionality by storing the users' IDs and different resources (educational resources, news articles, etc.) as linked data, and mapping them in a graph. This graph contains the resource itself and different pairs of predicates and objects that are relevant with the resource. For example, a graph describing a user may contain the pair (has name, <name of the user>), (knowledgeable in, <field of finances>), (has positively evaluated, <news article>), etc. With these representations we can compare the graphs of the different resources in order to find out similarities. The graph comparison is based on similarities of predicates (labeled edges) and objects (nodes). The predicates are split into disjoint groups, e.g., one group can represent the degree of awareness of a user about a certain field: is knowledgeable in, read about, is expert in, etc. In the group, the similarity between different predicates is predefined. The similarities of objects is computed based on the thesaurus structure and, possibly, on the frequency of occurrences in the corpora. The similarities between different graphs is assessed using the SoftCosine measure [18]

where the coordinates are pairs (predicate, object). Similarity propagation among users will also be explored [19].

Normally, in order to use aggregated data and make a decision based on a group of users, common patterns are precomputed. Those are frequent subgraphs of the user graphs. Associative rules [20] are used to draw a conclusion about the entity to be recommended to a user matching a common pattern.

4.4 Reputation-based Incentive Mechanism

PROFIT platform will also consist of a reputation-based incentive mechanism [21] to encourage users participation and promote good quality contributions (posts, articles, answers to questions, etc). The reputation metric for a given user will be calculated based on the ratings and comments her contributions received by other users in the system, as an indication of their quality. However, when there are no ratings (or very few) for particular contributions, the reputation metric will also be judged by a semantics-based quality assessment of the contributions. In particular, the relevancy of the contribution to the PROFIT platform will be assessed. The relevancy score is based on the number of concepts from the PROFIT thesaurus that is used in the user post and the relation between the concepts; e.g., if the post text is long, however only a few general concepts (e.g. "economics", "finance") are used, then the contribution is assumed to be of an arguable relevance. Besides that, the readability of the contribution will be assessed .

In general, users of high reputation will be awarded by the platform with social status recognized in the platform through gamification elements, like special avatars and badges, moderation rights to the platform, as well as particular extrinsic awards provided by the FEBEA member banking organisations, like discounts to their products, merchandising material, etc.

5. CONCLUSIONS

In this paper, we described how semantic technology can be exploited in the financial awareness and capability domain, enabling the various functionalities of the PROFIT financial awareness platform. PROFIT platform is still under development and will be pilot-tested via the collaboration with the members of the European Federation of Ethical and Alternative Banks (FEBEA), an institution committed to the responsible banking and finance agenda. The outcomes of the project are expected to enable inferences and specific appropriate practices that can be made available to the wide public in the EU.

6. ACKNOWLEDGMENTS

This work has been supported by the EU HORIZON 2020 project PROFIT (contract no: 687895)

7. REFERENCES

[1] Klapper Leora, Lusardi Annamaria, and Georgios A. Panos. Financial literacy and its consequences: Evidence from Russia during the financial crisis. Journal of Banking and Finance, 37(10), pp. 3904-392, 2013

[2] Lusardi, Annamaria, and Mitchell S. Olivia. "The Economic Importance of Financial Literacy: Theory and Evidence", Journal of Economic Literature. 52(1), pp. 5-44, 2014.

[3] EU HORIZON 2020 PROFIT project "Promoting Financial Awareness and Stability" : http://projectprofit.eu/, 2016-2018

[4] J. Hendler, "Web 3.0 Emerging," in Computer, vol. 42, no. 1, pp. 111-113, Jan. 2009.

[5] Vafopoulos, M. N., Vafeiadis, G., Razis, G., Anagnostopoulos, I., Negkas, D., & Galanos, L. (2016). Linked Open Economy: Take Full Advantage of Economic Data. Available at SSRN 2732218.

[6] T. Tiropanis, H. Davis, D. Millard and M. Weal, "Semantic Technologies for Learning and Teaching in the Web 2.0 Era," in IEEE Intelligent Systems, vol. 24, no. 6, pp. 49-53, Nov.-Dec. 2009.

[7] Ronallo, J., HTML5 Microdata and Schema. org. Code4Lib Journal,16, 2012.

[8] Westerski, A., & Sánchez-Rada, J. F., Marl Ontology Specification, V1. 0 May 2013.

[9] Weibel, S., Kunze, J., Lagoze, C., & Wolf, M. Dublin core metadata for resource discovery (No. RFC 2413), 1998.

[10] Bennett, M. The financial industry business ontology: Best practice for big data. Journal of Banking Regulation, 14(3), 255-268, 2013.

[11] Cimiano, P., Hotho, A., & Staab, S. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. J. Artif. Intell. Res.(JAIR),24, 305-339, 2005.

[12] Ahmad, K., Tariq, M., Vrusias, B., & Handy, C. Corpus-based thesaurus construction for image retrieval in specialist domains (pp. 502-510). Springer Berlin Heidelberg, 2003.

[13] Neubert, J. Bringing the" Thesaurus for Economics" on to the Web of Linked Data. LDOW, 25964, 2009.

[14] Tetlock, P. C. Giving content to investor sentiment: The role of media in the stock market. The Journal of Finance, 62(3), 1139-1168, 2007.

[15] Stock, J. H., & Watson, M. W. Forecasting output and inflation: the role of asset prices. National Bureau of Economic Research, 2001.

[16] Ludvigson, S. C., Ma, S., & Ng, S. Uncertainty and Business Cycles: Exogenous Impulse or Endogenous Response? National Bureau of Economic Research, 2015.

[17] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," Foundations and Trends in Information Retrieval, vol. 2, nos. 1–2, pp. 1–135, 2008.

[18] Sidorov, Grigori; Gelbukh, Alexander; Gómez-Adorno, Helena; Pinto, David. "Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model". Computación y Sistemas 18 (3): 491–504. 2014.

[19] Satsiou, A., Tassiulas, L., Propagating Users' Similarity towards improving recommender systems, IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technologies (1) : 221-228, 2014

[20] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami.. Mining association rules between sets of items in large databases. SIGMOD Rec. 22, 2 (June 1993), 207-216, 1993.

[21] Katmada, A., Satsiou, A., Kompatsiaris, I. Incentive Mechanisms for Crowdsourcing Platforms, accepted for publication in Proceedings of the 3rd International Conference on Internet Science, 2016.

[22] Klapper L., Lusardi A., and G. Panos Financial literacy and trust in financial institutions. Working paper, 2016.

[23] D1.2: PROFIT Use Cases and User Scenarios, http://projectprofit.eu/material/, 2016.