Automating Semantic Analysis of Website Structures for Ontology-based Benchmarking

Conceptual Model and Implementation in Retail Banking

Nikola Vlahović, Mirjana Pejić Bach, Andrija Brljak
Faculty of Economics and Business
University of Zagreb
Zagreb, Croatia
e-mail: nvlahovic@efzg.hr

Abstract— **Companies use benchmarking to improve the efficiency of their business processes, organizational structures and response to changes in their business environment. Benchmarking incites additional effort and drain on company resources. In this paper, we present an approach that may offer new outlook in making benchmarking less costly and time consuming. The goal of this paper is to present a conceptual model of the system based on grounded theory that uses current information retrieval methods, natural language processing and available web resources to create semantic ontology of best practices in structuring web-based information content. The developed model that is the result of the described approach can be used as a benchmarking model and tool for various purposes that we will illustrate using a banking web sites case study.**

Keywords- semantic annotations; ontologies; automatic information retrieval; business applications; web mining; natural language processing.

I. INTRODUCTION

Organizations rely on business intelligence in order to better understand data they store in their information systems and data that is being generated outside the organization. Valuable information generated through deep analysis of data may be used to improve their operational efficiency, organizational structure, quality of decision making, competitive potential and overall performance. Innovative approaches in discovering new information from databases have been used over the past few decades, with substantial success. Still, new approaches are implemented in order to retain comparative advantage over their competitors. Semantic networks and ontologies have been rarely in the focus of business intelligence, even though for a long time the potential is recognized [2]. In this paper, we propose a concept of an ontology-based methodology that can be used to create semantic model of a particular problem domain or area of interest that can be used by the company to conduct comparative analysis and determine its advantage and disadvantages in relation to good practices learned and inferred from available data sources outside organization.

The goal of this paper is to present an approach to analyzing web site structure in order to create a structural semantic model for particular type of web sites that can capture current best practice in web information organization. Different organizations organize their information in different ways and discovering the most prevalent approach may indicate best practice in presentation of information. The inferred model is an ontology created through inductive process and it serves as a benchmarking tool. The model consists of nodes that are interlinked as in any semantic network. These nodes represent typical web pages that are encountered in a particular class of web pages. They are described by their topic using keywords. The relations between nodes represent the hierarchy of web pages within the web site. These structures and position of each node in the structure is determined by the inductive learning processes that take information from a collection of existing web pages as will be described in this paper. The evaluation of sample web site structures heavily relies on pattern matching, natural language processing and available English language corpuses. The main contribution of this work is the implementation of semantic web as a benchmarking tool in business intelligence of an organization while preserving acceptable level of cost, time requirements and engagement of other resources.

The rest of this paper is organized as follows. Section II describes the background on implementations of web and text mining, ontologies and semantics in managerial decision-making processes, specifically benchmarking as an important tool for managerial decision making. Section III describes the conceptual model of automated information retrieval, analysis, contextualization and creation of benchmarking model based on semantic content analysis. Section IV addresses the implementation based on revealing best practices in structuring banking web sites and describes the developed semantic model. Section V presents the discussion of presented work, points out main conclusions and presents further steps in the development of applications for the described model in various practical areas of economics and business.

II. BACKGROUND

In this Section we will explain the role of benchmarking in Business Intelligence as well as most recent developments regarding benchmarking analysis in current literature with special attention to semantic web and ontologies.

A. Benchmarking and Business Intelligence

Benchmarking is the process of analyzing business processes by comparing their performance and other properties with current best practice or industry standards for particular business domain. Companies have been applying this approach as an important part of decision support to make their processes more efficient [4]. Various quantitative and qualitative methods have been used to analyze available information about current best practices, such as knowledge management, knowledge-based systems, simulation modelling, datamining, etc. The main disadvantage of the currently used and proposed methodologies is that they are either time-consuming, costly or require overwhelming amount of resources. Therefore, the processes in decision support and business intelligence aim to automate various steps of implemented methodologies such as information retrieval, modelling, or analysis. Improvement in automation of these processes can be achieved with sufficient organization of information. This is one of the main reasons why semantic networks and ontologies have been identified as technologies that would greatly benefit business intelligence [2].

B. Related work

Over the past couple of decades there have been very few implementations of semantic web and ontologies in decision making [1]. Available implementations usually pertain to general web information and web services [6] published in traditional web sites with the purpose of adding a level of semantics in order to enhance search procedures [7]. Other implementations are concerned with the organization of specific expertise knowledge in various fields such as genetics [8], molecular biology [9], but also disaster management [1], e-governance and public data [10], project management [11] and social networking [12]. Special importance is given to the development of ontologies that are dedicated to syntactical information for various languages, such as lexical databases of English, German and other languages [13]. Lexical databases are often incorporated in other (semi)-automated systems for information retrieval as they can add the syntax layer to the retrieval procedure improving natural language processing significantly. Most of these ontologies are manually created to ensure correctness of concepts [14]. Manually created ontologies suffer from similar disadvantages as other methodologies implemented in business intelligence domain (high cost, time consuming, labor intensive, etc). On the other side there are informal ontologies that can be created in semi-automatic way either through volunteered information retrieval by virtual communities or by programmable information retrieval and analysis. Just in the last five years some implementations that are based on ontologies and semantic web have been proposed and developed [4] [5] that use semi-automated procedures.

In recent years, several tools have also been developed for construction of ontological databases, semantic networks and taxonomies. Some of the examples include Ontolearn [15], OntoLT [16], SOAT [17] and TextOntoEx [18]. These tools can be used to create ontologies from natural language texts and serve as a link between linguistics and ontology engineering. SOAT is created to use Chinese language corpus while the rest of the tools use English corpus. Neither of these tools, though, include functionalities specific for benchmarking, i.e., created models do not allow for visualizations and comparative exploration tools that is typical for business intelligence tools. In order to increase the variety of possible applications and problem domains, and to redefine ontological models to serve as benchmarking tools for decision making in business organizations, we will propose a specific methodology. This methodology is based on ontologies and several other technologies that can be implemented in business environment and improve decision making and managerial planning tasks. Specifically, we will concentrate on automated information retrieval from web sources and analysis based on the iterative creation of benchmarking model that can provide insight in current state of the organization of web information in particular industry or business domain.

III. AUTOMATED INFORMATION RETRIEVAL, ANALYSIS AND ONTOLOGY MODELLING

In this Section we will describe conceptual model of automated information retrieval, analysis, contextualization and creation of benchmarking model based on semantic content analysis. Key aspects of using existing lexical databases will be described as well and the algorithm of evaluating and improving the structure of the semantic model during the learning phase of the model creation process.

A. Best Practice Ontological Model

The ontological model is based on inductive synthesis approach. This is an approach originally implemented in automatic software application development based on second order logic [3]. In the case of ontology induction, the set of software specifications is replaced by a set of individual structures, so the goal of the algorithm is to create a generic ontological model that can describe all of the most typical, prevalent and semantically justified properties contained within each instance of the set. For the purpose of this paper, the structures used refer to individual web sites and the organization and structuring of their individual web pages based on the information content published. In order to automate this task, a complex algorithm was developed that goes through several different phases (Figure 1).

 ISBN: 978-1-61208-654-5

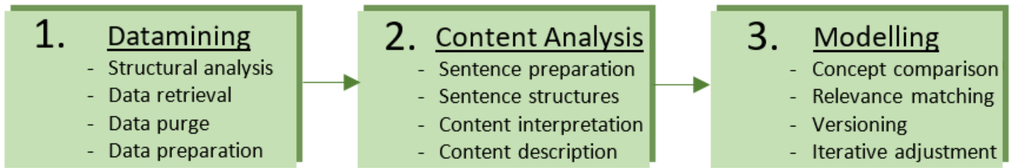

Figure 1. Overview of phases and steps in automated informationretrieval, analysis and ontology modeling

First is the stage of datamining activities Initially, information retrieval is conducted for each resource. Based on the initial structure hierarchy of constituent elements is analyzed. In case of web sites, retrieved navigation is analyzed so that each subpage within the web site structure can be accessed. Then, the rest of the content for each structure can be retrieved. Depending on content retrieved, irrelevant information or code is removed during data purge. In this case most of the HTML code that does not pertain to content itself is removed. Finally, data is prepared for the next stage of the analysis that involves Natural Language Processing (NLP) methods and procedures.

While it is necessary to perform first stage of the process online, second stage can be performed offline since all of the prepared data is stored in local database. In this stage the system still has to access online resources that include language lexicons in order to perform language analysis of the content. Firstly, content of each page is divided into sentences. The structure of each sentence is determined during syntax analysis (described in next section). Depending on the structure and syntactical role of words, a set of potential keywords is determined – token words. For each token word meaning and role is determined using lexical databases, in order to assess the interpretation of content. Here, it is important to consult language thesaurus, determine the definite meaning of potential keywords and create set of keywords that best describe the content of each page. Meanings that are most common to the set of token words are used to resolve disambiguation of any particular token word that has more than one acceptable meaning.

Finally, in the third stage of the process each page is represented by its content description that is then compared to existing concepts in the ontological model. Depending on the result of this analysis further steps in changing the ontological model are determined depending on the relevance of the page to the current model. Before any change is committed to the model versioning of the model is stored in order to provide insight in the evolution of the model or to manually select the most appropriate model for benchmarking. If there are new concepts introduced into the model additional tuning of the model may be performed at the last step of this activity. This process is repeated for each analyzed structure that will shape ontological model,

Based on this process a set of tools and appropriate interfaces were developed. The architecture of the completed system is given in Figure 2.

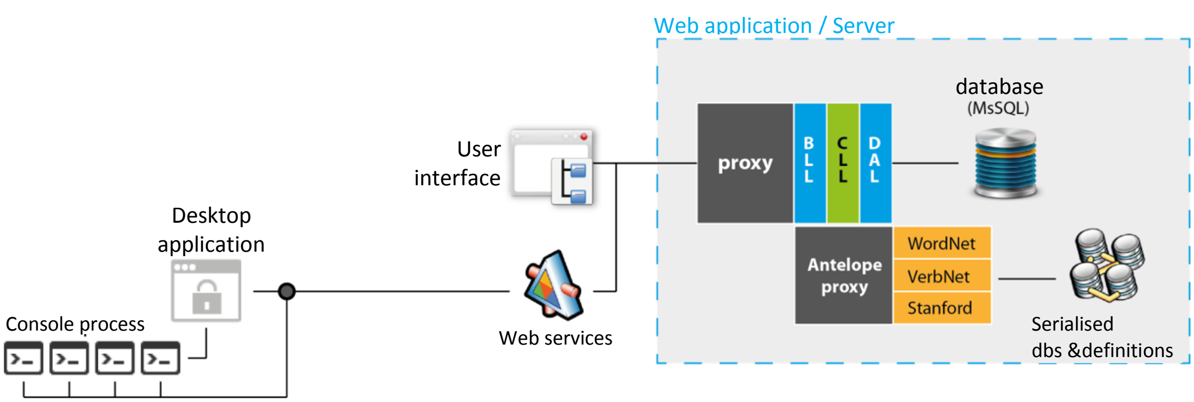

Figure 2. Overview of system components required for ontology modeling

The system is composed of multiple tiers. It consists of several components that include web information retrieval agent (as console process), parser and analyzer (as part of desktop application). Key elements of the system include the use of lexical databases and basic NLP procedures, as well as heuristics specifically developed to effectively determine keywords for content description. These features will be explained in more detailed in the rest of this Section.

B. Natural Language Processing and role of Lexical Databases

Natural Language Processing (NLP) methods that are used in this example deal with determining the meaning of text given in various sources. Meaning itself will be represented by a set of keywords that are best suited for particular content, i.e., web page. Keywords will determine position of each web page within the ontological model instead of using only title for each page. In order to adequately estimate meaning of the page content, it is important to consider the language of the text presented to the NLP procedures. There are particular properties of text specific to each language. For this purpose, only web sites and web pages written in English language are taken into account. NLP procedures can take advantage of existing language ontologies and lexicons that are available and accessible online.

There are several lexical databases available. WordNet is one of the first and most comprehensive lexicons for English language. Subsequently, other languages also developed their lexicons using WordNet organization model. Nouns, verbs, adjectives, adverbs are all grouped into cognitive synonym groups called SynSets. Each SynSet represents a clearly defined concept that fosters conceptual-semantic relations or lexical relations to other SynSets. There are 117000 SynSets, each with its own short description and example of usage in sentences.

In order to use lexical database, it is important to determine sentence structure and role of each word in each example. Stanford parser is a probabilistic parser used to determine grammatical structure of a sentence and group words into phrases and determine their function. It plays an important role in segmentation of sentences as it presents a sentence as Stanford Dependency (SD). SDs are hierarchical graphical representations of relations between words in sentence each described as a triplet: relation name, governing term and the dependent term.

Finally, modular solution that serves both as a language lexicon and parser is Proxem Antelope project and currently commercially available software tool Proxem Studio. This tool connects previously described lexicons and parsers and serves as an interface to procedures required for the analysis given in Figure 3.

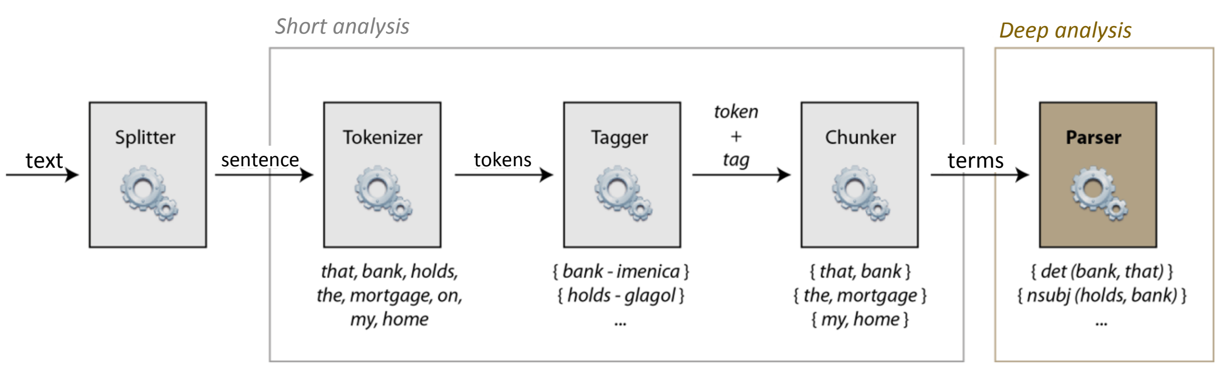

Figure 3. Overview of activities performed during syntax analysis

Content analysis goes through three steps: splitting text into sentences, performing short analysis and then performing deep analysis. During short analysis each sentence is disassembled into words or tokens. Each token is then tagged with the appropriate role in the sentence. For each tagged token, meaning is provided from the language lexicon and a set of relevant terms is created. These terms are then subjected to deep analysis by the parser that uses SDs to extract most probable keywords for the initial text.

C. *Created Heuristics in Analysing Web Page Context and Creating Ontological Model*

Now that each element of the hierarchy (in this case Web page) has a list of relevant key words describing the content, this structure is used to advance the ontological model (Figure 4). Comparative analysis between the current ontological scheme and prepared web page is performed. Based on the quantitative heuristics web page elements will be transformed into nodes of the ontological scheme. There are several options available: node will be added to the ontological model, removed from the structure or repositioned.

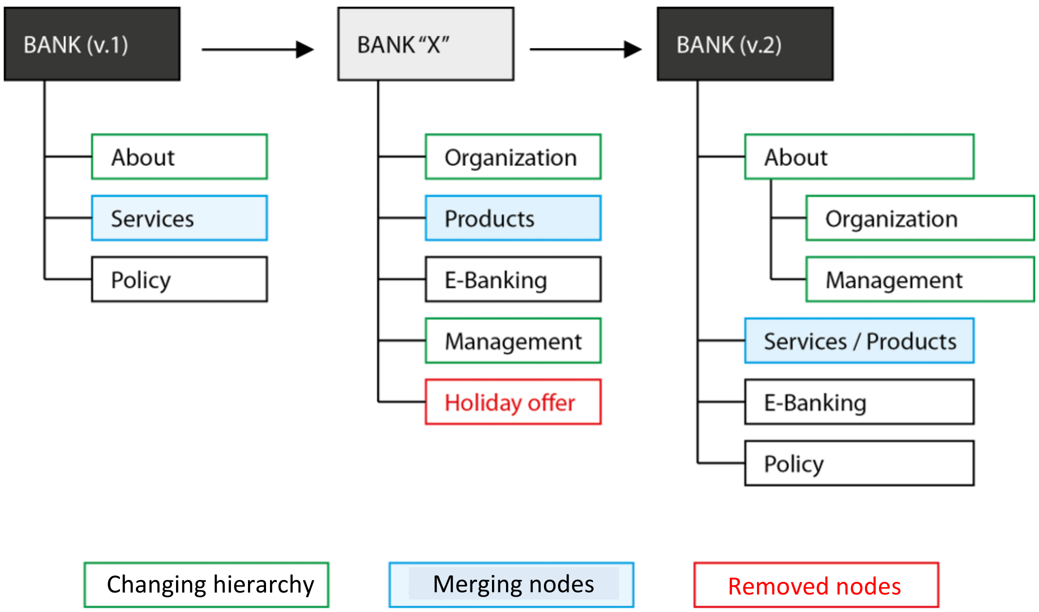

Figure 4. Example of operations over the ontology scheme during creation phase

Comparative analysis is performed for each node of the new web page (Bank X) and current version of ontological model (Bank v.1). Here, the heuristic list of keywords is chosen for each node and it is compared to each keyword set for nodes of the current ontological model. Distances between keywords are calculated in three steps.

In the first step, a set of potential elements, SE, is determined as (1)

$$SE = \underset{x \le 10}{\arg\max} f(x) = \{K_S \cap K_E\}, \quad |K_S \cap K_E| \ge m, x \in N, m > 0 \tag{1}$$

where maximum number of potential elements that are used for comparison x is ten. If K_S is a set of keywords describing the web page and K_E is a set of keywords describing elements of the ontological scheme, m is a heuristics parameter describing the minimum number of matching keywords both in K_S and K_E. Next weights for each matched keyword in SE is calculated as (2)

$$w_{SE} = \sum |K_{SE} \cap K_E|, \; K_{SE} \subseteq SE \tag{2}$$

w_{SE} is calculated as a sum of repetition of keywords contained in SE. Finally, distance between two sets K_S and K_E are calculated as (3)

$$M = \underset{x=1}{\arg\max} f(x) = \left\{\sum \frac{SE}{w_{SE}}\right\}, |SE| > 0, w_{SE} > 0 \tag{3}$$

where M is the sum of weighted keywords in SE. Since the goal is to find the smallest distance between the two sets, i.e., most similar pair of nodes from potential web page keywords and ontological model inverse of w_{SE} is used.

After ontological model is adjusted with information from each subsequent web page, additional *ex post* tuning using the same procedure is performed since new information changes the content of the set of potential elements SE, enabling additional corrections between nodes of the ontological model.

 ISBN: 978-1-61208-654-5

As we can see in Figure 4 depending on the comparison analysis, each node from new web site instance will influence the ontological model with three possible outcomes: (1) if the matching threshold is not passed the node will not be included in the model, (2) if there is matching between two or more nodes of the web page with one node in the model, these nodes will become sub-nodes of the ontological model increasing the hierarchical depth of the model and (3) if there is only one node of the web site that matches with only one of the nodes in the model these nodes will merge.

In initiating phases of the development of ontological model new nodes can be added to the model manually to help determine the most important features of the problem domain.

IV. BEST PRACTICE SEMANTICS: BANKING WEB SITES

In order to analyze the presented conceptual model a prototype system was developed as proof of concept. The goal of the prototype was to determine best practices in structuring web information on banking web sites in East European countries. Architecture of the developed system is given in Figure 5.

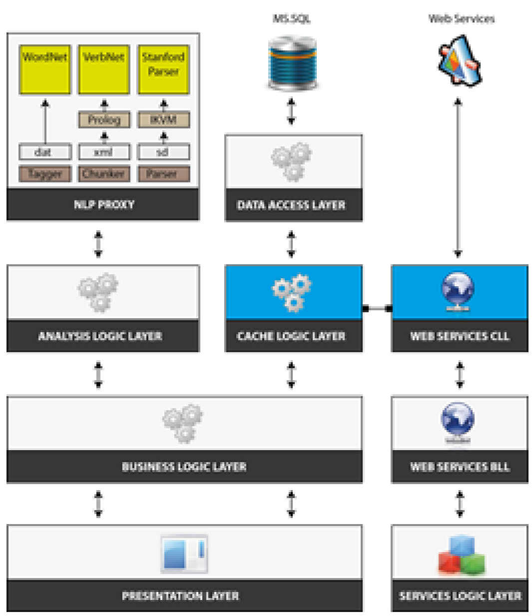

Figure 5. System architecutre for the developed prototype of the system

The system was given a database of URL addresses for a set of 42 banking web sites that have published information in English language. Total number of webpages included in the research was 2600. Retrieval of website structure and webpage documents was automated using a dedicated software agent. Software agent accessed each web site and parsed the content to search for web map or web navigation. Once located it was able to retrieve hyperlinks to each page in web map, access these pages and retrieve their content. All of the content was further prepared and stored into a local database that was used to initiate content analysis and iterative generation of the ontological benchmarking model of banking web sites. The induction process of the ontological model was conducted iteratively as described in earlier Section. Basic Statistics of the developed ontological scheme model are given in Table 1.

TABLE I. BASIS STATISTICS FOR DEVELOPED ONTOLOGY SCHEME

No. of iterations	No. of ontology elements	No. of matched pages
30	142	813
Total no. of used keywords	**Average no. of keywords per element**	**No. of matched pages per element**
2.393	16 (σ = 9,1777)	5 (σ = 7,9491)

Total number of iterations of the heuristic algorithm that calculated distances of each new node and its position in the model was 30. Final structure of the model included 143 content elements or pages in hierarchy. Each category generated a list of keywords from a total pool of generated keywords for the model 3993. Finally, Average number of keywords associated with each element is 16 with rather high variance of over 9. Average number of pages associated with each element of the ontological model is 5.

Part of the developed model can be seen in Figure 6.

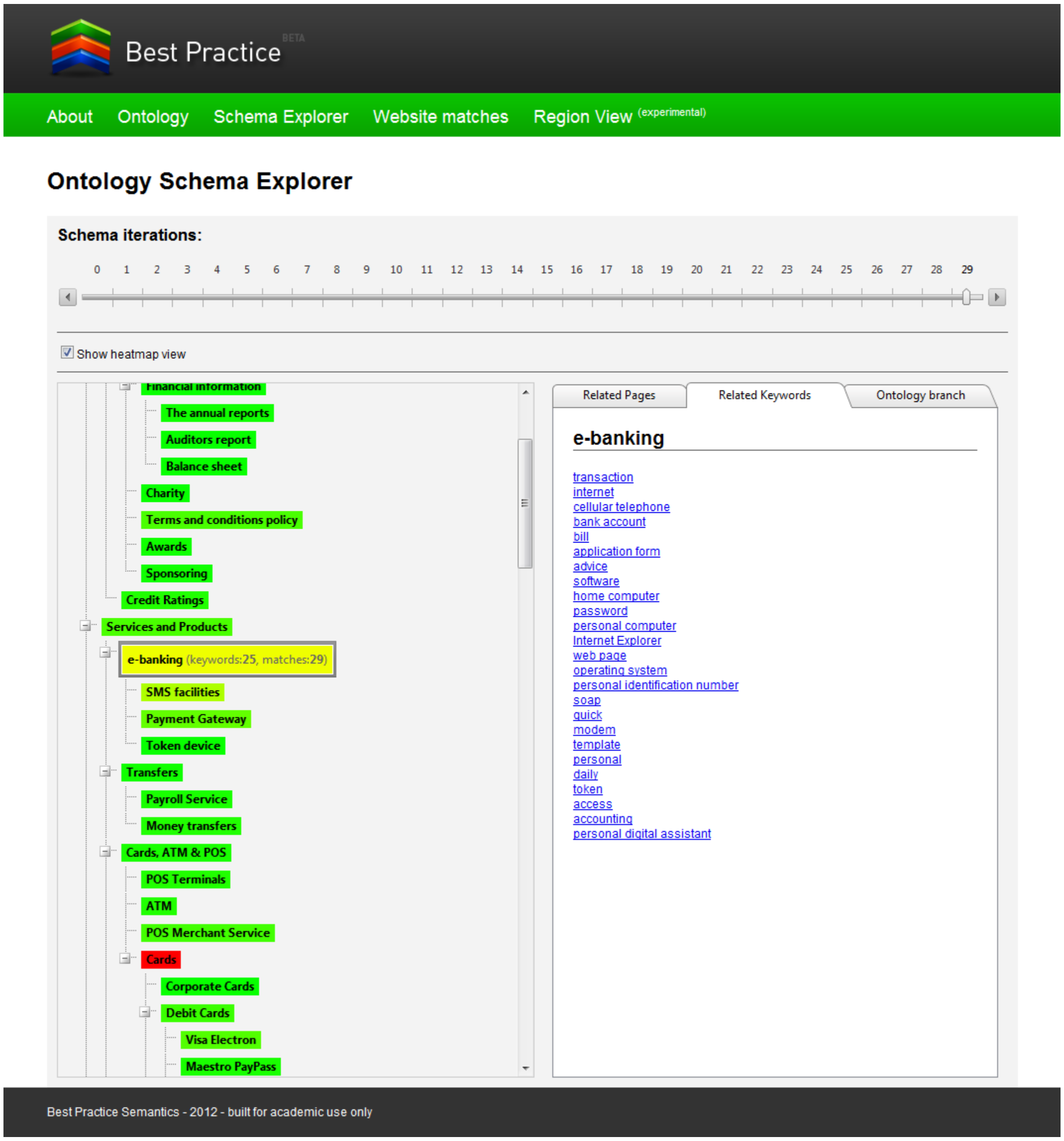

Figure 6. User interface of the developed prototype with portion of ontology model

The Ontology scheme Explorer shown in Figure 6 can also be used to view different versions of the scheme for each iteration, but it can also provide the list of keywords associated with each node presented in the model.

Some of the most common elements in bank website structure are given in Table 2 with respective list of keywords.

Keywords associated with particular node/web page category can also provide additional insight into the content of each page. Here we can see some web sites provide multiple pages with similar information, where the ontological model grouped these pages as same node. List of keywords can be used for further analysis and visualization. For example, presentation of the model benefits from color coded information, where green color is used to show categories that are present in most web sites, while less common categories are represented in yellow or gray color.

Additionally, Ontology Schema Explorer can provide comparison tool and show parallelly the scheme with best practices and structure of chosen web site with additional information about the similarities and differences between a web site and benchmarking structure (Figure 7).

 ISBN: 978-1-61208-654-5

TABLE II. MOST FREQUENT WEBSITE CATHEGORIES AND ASSOCIATED KEYWORDS

#	Name of Node	Number of pages	Most common keywords associated with the element
1.	MasterCard Standard	62	account, bank card, card game, credit card, customer, debit card, hotel, interest rate, internet, issue, logo, merchant, opportunity, personal identification number, swipe, transaction...
2.	Cards	35	application form, bank account, bank card, business card, calling card, credit card, debit card, internet, kind, larceny, merchant, overdraft, phone number, regular payment, seller...
3.	Depository Services	32	bank account, booklet, credit, depositor, depository, entrepreneur, interesting, investor, issuer, labour contract, mediator, quarterly, redemption, time deposit...
4.	International payments	31	bill, call mark, check, documentary, duty, European Union, foreign country, foreign exchange, franchise, futures contract, giro account, International, rate of exchange, savings, transaction...
5.	e-banking	29	access, accounting, application form, bank account, bill, cellular telephone, daily, home computer, internet, Internet Explorer, modem, operating system, password, personal digital assistant, personal identification number, quick, software, template, token, transaction, web page...
6.	Current Account	21	account, application form, bank account, check book, debit, foreign exchange, guardianship, income, interest rate, legal status, national, old-age pension, overdraft, pension, savings, Social Security, standing order, transaction, wage...
7.	SMS facilities	21	bank account, cellular telephone, customer, daily, debit, foreign exchange, password, phone number, stand-in, transaction, transfer payment...
8.	SME Financing	20	business activity, capital, cash flow, collateral, distribution channel, documentary, employee, entity, equipment, expertness, export, factorization, financing, guarantor, investment, letter of credit, mortgage, overdraft, postponement, production cost, real property, savings, subcontractor, suiting, supplier, wage...
9.	Individual customers	19	advisory, applicant, check, customer, employee, entity, financing, free, giro account, income, interest rate, memo, memorabilia, net income, pension, poor people, precious metal, private, safe-deposit, savings, Social Security, valuable, wage...
10.	Frequently Asked Questions (FAQ)	17	advice, alias, anti-virus program, at home, call, card, consent, database, encoding, income, letter of credit, overdraft, password, personal computer, personal identification number, phone number, prerequisite, procurement, purchase price, rate of exchange, small letter, smart card, software...

V. DISCUSSION AND CONCLUSIONS

As we can see from the developed prototype the presented concept can be implemented using currently available technologies in order to provide additional information about current status of published web information in various problem areas. It is important to stress that currently language and lexical databases are main constraining factor in further implementations of this approach with regards to assessing information in languages other than English. Actual implementation of the system may prove to require additional costs for the development in comparison to more standardized business intelligence tools used in benchmarking, but this approach does offer a new and fresh look at the available data that is rarely covered by current business intelligence or decision support tools.

If we take a closer look at the ontological model for banking web sites for the South Eastern Europe, we see that there is a high difference in web site quality between banks both in structuring web site and content. Many of the banks are part

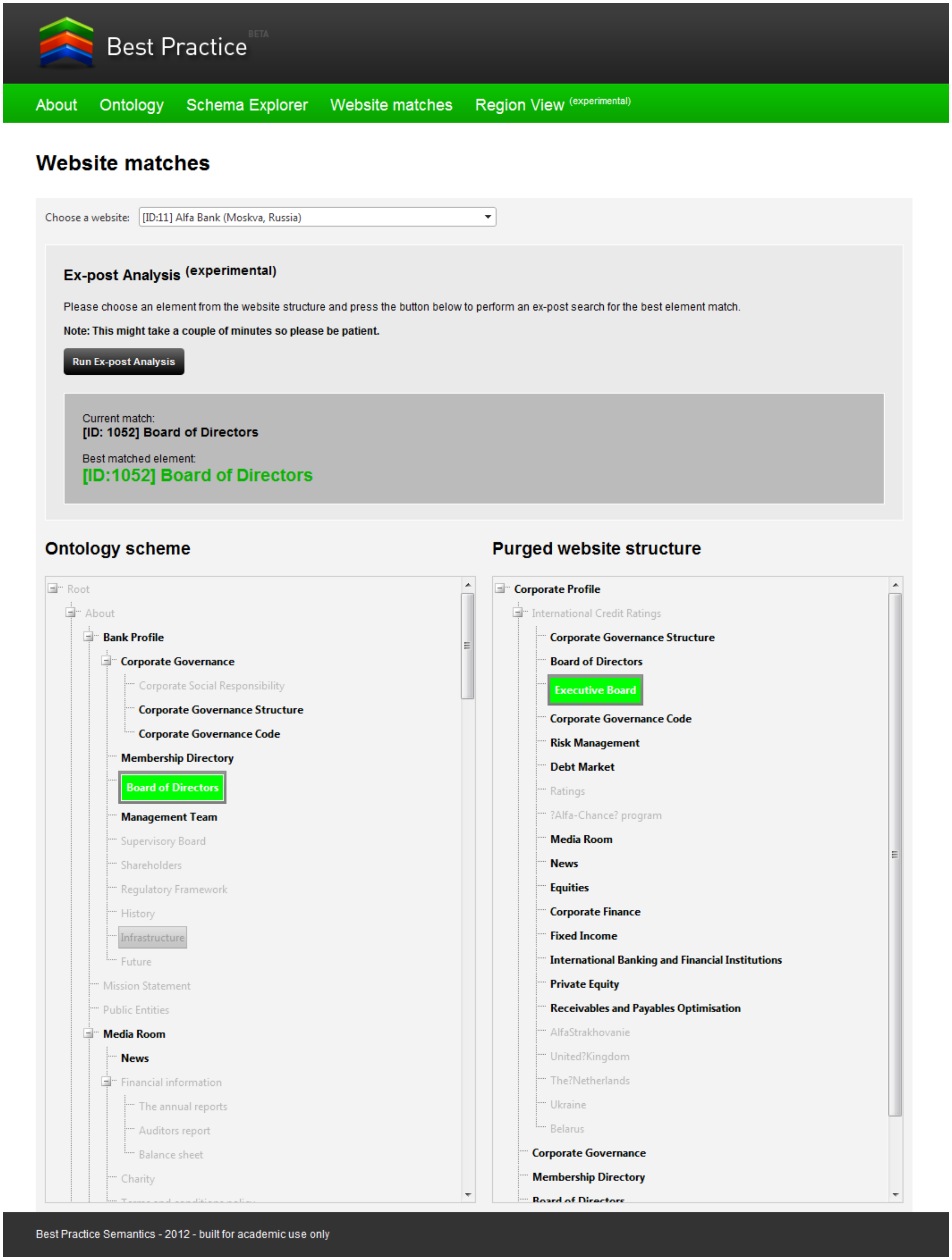

Figure 7. Comparative analysis of web site structure with (a) ontology scheme on the left side and (b) specific website structure on the right

of larger group of international banks that implement web content management solutions of the parent bank, so the specific local customization is not implemented consistently or adequately. The ontological model recognized this by grouping several pages of some of the same banks into the same category of the ontological model, while excluding several retrieved pages that did not pass the relevant threshold of the model. This is very important information that may be used to improve shortcomings of web sites of commercial banks in this region.

In conclusion, we see that there is still adequate potential in developing new approaches to automated and semi-automated tools that can help with Web information retrieval and generation of decision assistive models for decision making. Ontologies implemented in the area of benchmarking analysis in business can provide new means of analyzing publicly available data about markets and competition allowing companies to improve their processes and strategies.

In this paper we presented ontology-based benchmarking tool that may provide support to managers while providing semi-automated assistive decision models. This tool can be further improved in several different directions. Firstly, implementing additional analysis of created ontological model would show metrics for each node, explanation of the hierarchical position of a node within the model, etc. Secondly, adding the analysis of recovered data and visualizations in terms of color coding various content indicators, such as significance of content topics based on number of web sites including these topics, optional position of particular nodes in the structure, etc. This information may be used to guide decision makers during planning of their web site structure, ultimately improving their customers' experience. Finally, third possible improvement of the proposed model is further sophistication and automation of the ontology construction procedure so that it requires less manual intervention and corrections.

Future work includes developing approaches to automate additional steps during the creation of the benchmarking semantic model as it currently requires expert input, especially in the initial stages of the creation process. Another improvement of the model is the possibility to use other languages that have well defined language corpuses available.

REFERENCES

[1] C.-H. Chou, F. M. Zahedi, and H. Zhao, "Ontology for Developing Web Sites for Natural Disaster Management: Methodology and Implementation", IEEE Transactions on Systems, Man, and Cybernetics—part a: systems and humans, vol. 41, no. 1, January 2011, pp. 50-62.

[2] S. Drew, "From knowledge to action: the impact of benchmarking on organizational performance", Long Range Planning, Vol. 30 No. 3, 1997, pp. 427-441.

[3] S. Itzhaky, S. Gulwani, N. Immerman, and M. Sagiv, "A Simple Inductive Synthesis Methodology and its Applications," Proceedings of the ACM international conference on Object oriented programming systems languages and applications, Reno/Tahoe, Nevada, USA — October 17 - 21, 2010, ACM, 2010, pp. 36–46.

[4] F. Teuteberg, M. Kluth, F. Ahlemann, and S. Smolnik, "Semantic process benchmarking to improve process performance", Benchmarking: An International Journal Vol. 20 No. 4, Emerald Group Publishing Limited, 2013, pp. 484-511.

[5] K. N. Vavliakis, A. L. Symeonidis, G. T. Karagiannis, and P. A. Mitkas, "An integrated framework for enhancing the semantic transformation, editing and querying of relational databases", Expert Systems with Applications, Vol. 38, Issue 4, April 2011, pp 3844-3856.

[6] Auhood Alfaries, David Bell, and Mark Lycett, "Motivating service re - use with a web service ontology learning", International Journal of Web Information Systems, Vol. 9 Issue: 3, 2013, pp.219-241.

[7] M. Calaresu, and Ali Shiri, "Understanding Semantic Web: a conceptual model", Library Review, Vol. 64 Issue: 1/2, 2015, pp.82-100.

[8] M. Singleton, "Combining Phenotype and Genotype For Discovery and Diagnosis of Genetic Disease", Doctoral Dissertation, August 2015.

[9] C. Stanley Funk, "Recognition and Normalization of Terminology from Large Biomedical Ontologies and Their Application For Pharmacogene and Protein Function Prediction", Doctoral Dissertation, 2015.

[10] Bhaskar Sinha, Somnath Chandra, and Megha Garg, "Development of ontology from Indian agricultural e-governance data using IndoWordNet: a semantic web approach", Journal of Knowledge Management, Vol. 19 Issue: 1, 2015, pp.25-44.

[11] D. Ruikar, C.J. Anumba, A. Duke, P.M. Carrillo, and N.M. Bouchlaghem, "Using the semantic web for project information management", Facilities, Vol. 25 Issue: 13/14, 2007, pp.507-524.

[12] George Macgregor, "Knowledge Representation in the Social Semantic Web", Library Review, Vol. 60 Issue: 8, 2011, pp.723-735.

[13] A. A. Krizhanovsky and A. V. Smirnov, "An Approach to Automated Construction of a General□Purpose Lexical Ontology Based on Wiktionary", Journal of Computer and Systems Sciences International, Vol. 52, No. 2, 2013, pp. 215–225.

[14] N. A. Astrakhantsev and D. Yu. Turdakov, "Automatic Construction and Enrichment of Informal Ontologies: A Survey", Programming and Computer Software, Vol. 39, No. 1, 2013, pp. 34–42.

[15] R. Navigli, P., Velardi, and A. Gangemi, " Ontology learning and its application to automated terminology translation". IEEE Intelligent Systems, 18(1), 2003.

[16] M. Sintek, P. Buitelaar, and D. Olejnik. " A protege plug-in for ontology extraction from text based on linguistic analysis". In: Proceedings of the 1st European semantic web symposium (ESWS), 2004.

[17] S. H. Wu, and W. L. Hsu. " SOAT: a semi-automatic domain ontology acquisition tool from Chinese Corpus". In: 19th international conference on computational linguistics Howard international house and Academia Sinica, Taipei, Taiwan, 2002.

[18] M. Y. Dahab, H. A. Hassan, and A. Rafea. "TextOntoEx: Automatic ontology construction from natural English text", Expert Systems with Applications 34, 2008, pp. 1474–1480.