Tools for Internet Competitive Intelligence Based on Ontology

Viacheslav Lanin
Business Informatics Department
National Research University Higher School of Economics
Perm, Russian Federation
vlanin@hse.ru

***Abstract* — This paper proposes the approach to extract information from electronic documents on the Internet with the aid of ontologies. It applies to generic web page structure description developed on the ground of most widespread information placement and distribution mechanisms. Also, the proposed approach will allow end users to compose high-level resources (electronic documents) description almost by means of natural language. The most noticeable distinction on this approach is that the main emphasis is placed on the data placement structure, in its turn, enabling the overall search results improvement.**

***Index Terms* — website, web page structure ontology, ontology traversal, web page annotation.**

I. Introduction

Competitive intelligence is a systematic process initiated by organizations in order to gather and analyze information about competitors and the general socio-political and economic environment of the firm [1]. Nowadays a main source of information is the worldwide web. The most of information captured in various electronic documents, especially those found in the internet, kept in the semistructured or wholly unstructured form. Semi-structured data are the form of data not conforming with formal structure of relational data models or other table forms, yet containing specific markers to separate its semantic or logic parts [2]. Therefore, these data are often called to be selfdescribing. In addition, modern search engines exploits so-called "bag-of-words" document model and does not pay any attention to the word placement in the given document. That is why, the search engine end user has to formulate a very complicated query taking into account content particularities of those unstructured data. Eventually, this user-formulated search query will not guarantee highly relevant return result, as well as its processing may take a big amount of time.

This discrepancy is supposed to be resolved in the following way: the information search and extraction will be considered from the structure and placement view, rather than analyzing only its substantive aspects and characteristic features. This mechanism will let simplify the search query formulation process to a great extent, as well as its structure enabling querying specific document parts and sections.

In order to translate this possibility into action, it is necessary to develop a resource capturing document structure description provided that it is generic and suitable to develop a description for different kinds of electronic documents; with this aim in the view the most widespread structures and information placement templates were examined. This resource is supposed to be developed in the form of ontology keeping, in the context of web documents, HTML-tags identifying the web document section. Thus, to query a web a document described by means of this ontology, firstly, the specific document place is to be identified using traversal algorithm and then place-oriented content query is to be implemented.

II. Related works

Web scraping (web harvesting, web data extraction) is data scraping used for extracting data from websites [3]. Search bot [4] (web-crawler, scraper) is an application, often a part of standard search engine designed to web pages looking over in order to register them or update their description in a search engine database. Web-crawlers can download all visited pages for further processing allowing end users querying information of these pages faster. The required web-crawler settings including web pages visit procedure, visit frequency and information extraction criteria are determined by the information search mechanism being applied. It has to be mentioned that in case the web-crawler utilizes web document structure analysis, it will be able to identify valuable information and references faster narrowing search considerably.

In addition, there are several web-crawlers operation alternatives. Firstly, focused crawlers designed to extract web documents meeting specific prearranged criteria defined by either search query analysis or DOM examination. Secondly, "deep" web-crawlers focused to formulate a request to extract "hidden" pages. Furthermore, there exist crawlers exploiting specific heuristic algorithms, however, due to pure generic heuristics development difficulties, they have not become common use.

Product	Method	AddIn	AJAX	API	Price
Import.io	Automatic	+	+	+	>169 $

Product	Method	AddIn	AJAX	API	Price
Webhose.io	automatic	-	-//-	+	>50$
Diffbot	automatic	-	-//-	+	>899 $
Dexi.io	semi-automatic	+	+	+	>120 $
ParseHub	semi-automatic	+	+	+	>149 $
Spinn3r	semi-automatic	+	+	+	500$
Scraper	semi-automatic	-	+/-	-	0$
OutWit Hub	semi-automatic	-	+/-	-	0$
RapidMiner	semi-automatic	+	+/-	-	0$

A. Existing Ontologies for describing documents

There are several ontological resources with can be used for describing different aspects of documents. First of all, it is Dublin core (DC) standard [5]. DC is used to describe documents of various types. The metadata standard is separated to two levels: Simple and Qualified. First contains of 15 elements and having three additional elements. Second refines additional semantics of the elements. Each element in Dublin Core is optional and repeatable, it makes this standard widespread and flexible. But only documents tags having indirect correlation with the document content can be defined by Dublin core. It is not enough for describing structure and other essential aspects of the electronic document.

There are projects oriented to formal structure document description. For example, ontology «docOnto» [6] and CNXML document ontology (Connexions Markup Language). Also Document ontology SHOE, Document Ontology of Research Centre Linked Data DERI and Muninn WW1 [7] should be mentioned. But all existing document ontology is not suitable for describing webpage structure, so it is necessary to develop special ontology.

B. Ontology-Based Information Extraction

In the article [8] problems of structured web-forum information extraction were addressed. Authors identified two mechanisms: template-dependent and template-independent.

Template-dependent algorithms are based on soi-disant wrapper exploitation. Wrapper is a regular expression or a tree structure and can be developed manually or automatically. Nevertheless, wrappers suitable for diverse web page templates maintenance will be worth lump sums of money, making them almost impossible for practical application. That is why, in order to have more generic solutions to the stated problem of structured information extraction template-independent methods were proposed. They consider information extraction to be a segmentation problem, hence probabilistic language models are being applied. Therefore, these mechanisms depend on the page or document template considerably weakly. As a matter of actual practice, template-independent mechanisms analyze each page features individually.

In the work [9] authors offered an approach to enhance collaborative writing process, as long as state-of-the-art collaborative writing systems do not implement the semantic integrity checking while multiple authors are working on shared documents.

To provide the semantic coherence, authors developed three ontologies including document structure ontology, rhetorical ontology and annotation ontology. The document structure ontology captures relationships among main document parts like sections, subsections, paragraphs and so on. It is the rhetorical ontology that allows to perform semantic integrity checking keeping document rhetorical elements and their interrelationships. As for the annotation ontology, it captures document metadata, in other words, annotations, also letting to classify documents pursuant to the general subject matter, author or other features. The Figure 2 shows how these ontologies were organized, and as seen, document structure ontology occupies the middle level and has connections to both rhetorical and annotation ontologies.

Authors' primary focus was on the rhetorical ontology developed by means of OWL and descriptive logic because it is the ontology that takes the main part in sematic coherence checking process. Generally, this approach exploits semantic annotations in the other way compared to their ordinary applications. As a result, it also improves the coordination among several authors working on the same document concurrently.

In the article [10] authors proposed the ontology-based mechanism designed to extract information from construction tendering documents.

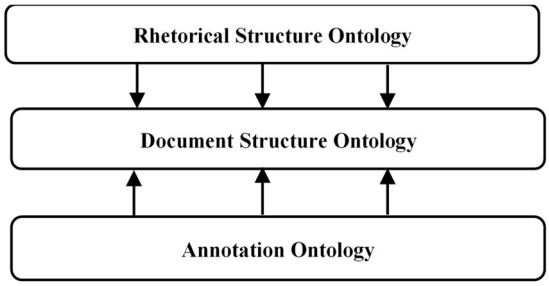

Fig. 1. Rhetorical Structure Theory Ontology Layers

The first step of the offered mechanism is tendering documents collection in the repository, then they are processed by the information extraction tool and transformed into machine-readable format and added to the knowledge base. At the finishing steps of this mechanism, tendering documents are being ranked by the special automatic ranking tool exploiting the knowledge base produced at the previous stage. The result of this algorithm is ranked tendering documents list allowing specialists to decide on tender leaders more effectively. However, it has to be mentioned that the authors did not exploited document structure analysis, by the way here the structure obeys the strict template, they only used domain

ontology designed as a result of domain experts interviews. helping to populate the knowledge base. The Figure 3 shows the generic architecture of this approach.

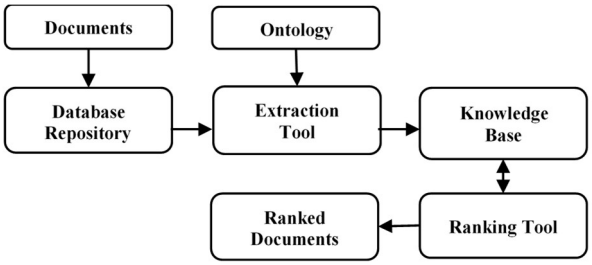

Fig. 2. Tendering Documents Ranking Tool Architecture

III. The Proposed Approach

This section contains the detailed description of the structure-centered approach to the information extraction problem solution.

A. Web Document Structure Ontology Description

In suggested approach the information search method is based on an ontology which describes document structure analysis and ontologies implementation is offered. The developed ontology is separated to two levels: website and web page. First level describes hole website structure and contains such concepts as main page, contact page, news page, product description page and other commonly used page types. Second level describes information blocks with in particular page. Figure 4 and Figure 5 show an example such two-level separation.

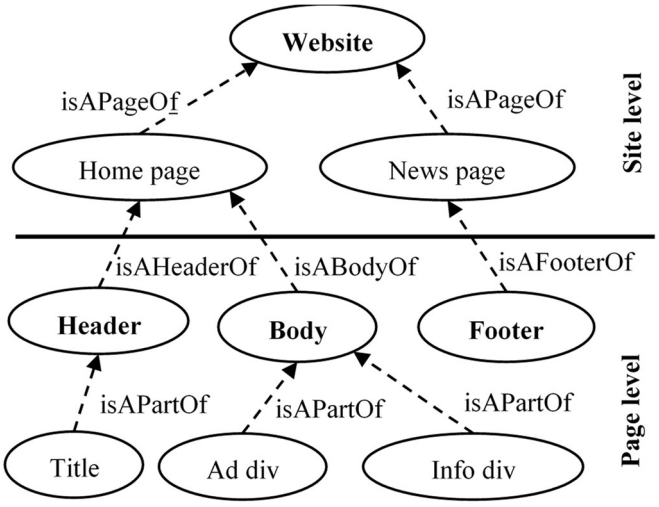

Fig. 3. Fragment of Two-level Web Document Ontology

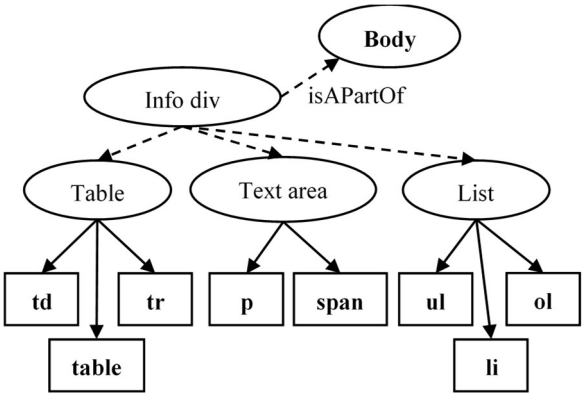

Fig. 4. Fragment of Data associated with Page Level

Figure 6 shows main architect components needed for suggested approach implementation.

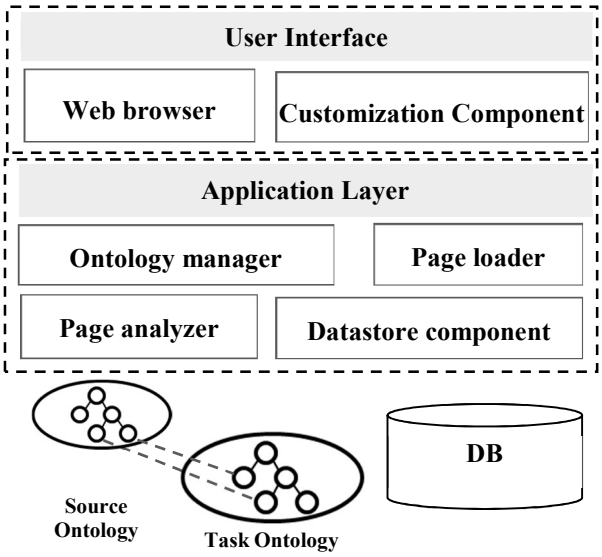

Fig. 5. Main components of the ontology-driven system for monitoring global processes on the basis of Internet news

This example query will return the address of a home page navigation block keeping it in a addressBlock property; PREFIX construct allows to abbreviate ontology URI usually including the whole address of a web page created to capture the developed ontology:

```
PREFIX ontostr: <ontology URI>
SELECT ?addressBlock,
      ?block,
      ?page WHERE {
 ?page a ontostr:HomePage.  ?block a
ontostr:NavigationBlock;
ontostr:isANavBlockOfAHomePage ?page.
 ?block ontostr:Address ?addressBlock.
 <other constraints>
}
```

Figure 6 shows user interface of developed tool implemented suggested approach.

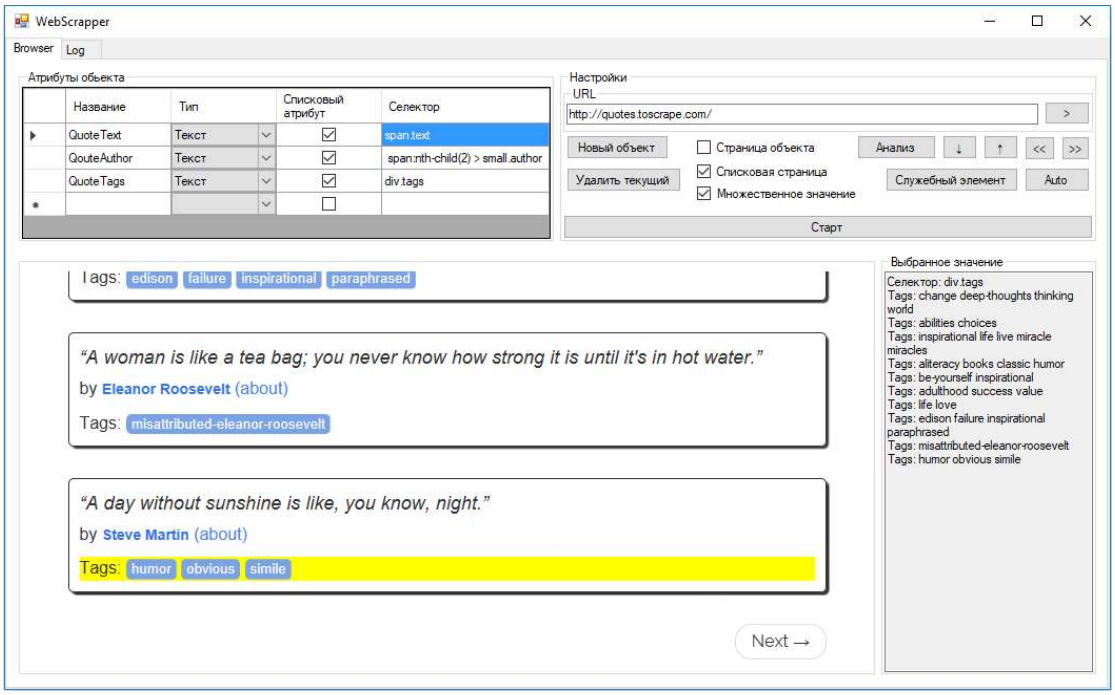

Fig. 6. User interface of prototype implementation

IV. Conclusion

This paper aims at offering an approach to refine information extraction process by shifting the emphasis of information search from the content analysis to the data structure and their placement in documents. Special attention during the work was given to the web documents because they are the most widespread ways to capture valuable information nowadays.

The web document structure ontology being developed is to be integrated with generic multidimensional ontology of electronic documents [11].

References

[1] T. Colakoglu "The Problematic Of Competitive Intelligence: How To Evaluate & Develop Competitive Intelligence," Procedia Social and Behavioral Sciences: 7th International Strategic Management Conference 24, 2011 pp. 1615-1623.

[2] P. Bunneman "Semistructured Data," in Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, May 11-15, 1997, Tucson, Arizona, United States. pp.117-121

[3] G. Boeing and P. Waddell "New Insights into Rental Housing Markets across the United States: Web Scraping and Analyzing Craigslist Rental Listings," Journal of Planning Education and Research.

[4] R. Cai, J.-M. Yang, W. Lai et al. "iRobot: An Intelligent Crawler for Web Forums," in Proceeding of the 17th international conference on World Wide Web (WWW 2008), pp. 448-456, April 21-25, 2008. Beijing, China.

[5] Dublin Core Metadata Element Set, Version 1.1 [Online]. Available: http://dublincore.org/documents/dces/.

[6] CNXML/DocumentOntology [Online]. Available: http://mathweb.org/wiki/CNXML/DocumentOntology.

[7] Muninn Documents Ontology [Online]. Available: http://rdf.muninnproject.org/ontologies/documents.html.

[8] J.-M. Yang, R. Cai, C. Wang et al. "Incorporating Site-level Knowledge for Incremental Crawling of Web Forums: a List-wise Strategy," in Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2009), pp. 1375-1383, June 28, 2009.

[9] C. Rahhal, N. Desliva and H. Naja-Jazzar, "OntoReST: A RST-based Ontology for Maintaining Semantic Consistency in Collaborative Writing," [Research Report] 2007, p.18.

[10] R. Mohemad, A.R. Hamdan, Z.A. Othman et al., "Ontological-Based Information Extraction of Construction Tender Documents," in Proceedings of the 7th Atlantic Web Intelligence Conference, AWIC 2011, Fribourg, Switzerland, January, 2011. pp. 153-162,

[11] I. Shalyaeva, L. Lyadova and V. Lanin "Events analysis based on Internet information retrieval and process mining tools," in: Proceedings of 10th International Conference on Application of Information and Communication Technologies (AICT2016). Baku: Publishing Department of Qafqaz University, 2016. pp. 168-172.