Analysis System of Scientific Publications Based on the Ontology Approach

Viacheslav Lanin, Svetlana Strinuk

National Research University Higher School of Economics, Perm, Russian Federation
lanin@perm.ru, strinuk@mail.ru

Abstract. The article describes an approach to scientific publications repository creation based on ontology approach and corpus linguistics methods, processing of unstructured data (scientific papers) using GATE. Implementation of discussed methods is intended to decrease significantly labor intensity of information search and analysis, provide operational use of information in research.

Keywords: ontology, GATE, scientific publication

1 Introduction

The number of academic publications has been growing day by day. It can be explained by the fact that the Internet has made a lot of publications and e-libraries available (RINC, Springer, ACM etc). Complexity of academic papers search on the particular subject is increasing in this regard. To improve search quality and speed information resources have to be systemized and well arranged, a user has to be given convenient navigating and query facilities.

To solve the task of text data processing statistics (latent semantic search), graph and ontological methods are implemented. Every method listed above has its drawbacks. Latent semantic analysis does not take into account semantics. Graph methods cannot be applied to academic search due to absence of evident links among documents. Ontologies have a limited application owing to lack of ontologies building; moreover building indexes and its support are time consuming. Semistructured character of information and heterogeneity of its sources involve implementing tools and methods of artificial intelligence to sort out the tasks of text data processing (text mining, Semantic Web technology and agent technology).

2 Document ontology

To search, analyze and classify, catalogue and store information efficiently consolidating knowledge about their content and structure is needed.

Information about the following aspects of electronic documents is critical: document size (format), document type, document layout (document structure).

While creating ontological resource notions about all three aspects of representing information are included in the document. Each element is described by ontology. Notions from different aspects should be interconnected therefore adjacent electronic document ontology is created. There are many projects of development document ontology (for example, Dublin core [4], project ontologies «docOnto» [3], Document ontology SHOE [5], Document Ontology of Research Centre Linked Data DERI [9], Muninn project document ontology [9]), but each existing document ontology has its advantages and disadvantages for solving our tasks. So, we create own ontology specialized on academic paper description.

3 Academic paper description

In this research most popular academic paper structure was analyzed. Rules describe the article in the academic journal as "original research, which should faithfully reflect the content and results of the research". Hypothesis should be put forward and evidence should be provided to prove it. The article normally provides clear accurate findings.

According to these recommendations most typical elements of academic papers were identified (see Table 1). Each section has a particular function; its tasks are formalized and described in guidelines and numerous article writing handbooks provide substantial information on writing each section. Understanding functions, which each section has, makes further identification of key article elements and search automation. Article title, authors' names, affiliation are not worth processing as they are unique elements of the structure. The key words set describes field of research in terminology terms, this set is relatively verified sample of the most frequent terms. Ontology of scientific publication were described on OWL language.

4 System implementation

The demands to these systems were augmentability, support of amount of languages, possibility to work with thesauruses and other ontological resources. After analyses of existing systems GATE [2] (General Architecture for Text Engineering) was chosen. GATE is a set of Java tools for natural languages processing. This open code system suits operations of processing texts of any size. It is necessary to note that in linguistic resources, which are used while working with GATE there are three types of data: documents, corpuses and annotations.

The following tasks are solved through means of GATE: organization of annotated storage of articles, implementation of mechanism of key words automated highlighting, realization of mechanism of automated structure analysis of publications in undirected formats introduction, identifying key structure elements of the article, identifying relationships between articles. Obviously, functional capabilities of GATE are limited; to solve all the mentioned tasks own solutions through implementation of GATE API are planned.

Table 1. Essential article parts

Article part	Description
Abstract	Abstract contains important information about most important sections of the article. It does not provide references. Normally, objectives, methods, procedures and the main conclusions are described.
Introduction History Background	Introduction focuses on providing sufficient information about the field of research that is why this section usually has a lot of references. Introduction also contains objectives and tasks of the article. This section refers to general situation in the field of research. Introduction provides specification of the scale of research; rationale of choice of methods, preliminary results and conclusion.
Previous research Literature review	The main function of Previous research/Literature review – is to analyze published sources, which illustrate one way or another the problem the article refer to. Previous research/Literature review might be written as a general review or a review of literature for a particular period of time.
Present Approach Objectives Hypothesis Model Analysis Methodology	A key article sector, varying from article to article, Present Approach/Objectives/Hypothesis/Model/Analysis Methodology section describes uniqueness of the approach to problem solution and approach development. Hypothesis and interpretation methodology of data collected are presented in this section. It gives detailed method, methodology and procedure description.
Results Statistical Analysis	Results/Statistical Analysis section gives the summary of the results/data sometimes tables, diagrams and other visuals are added.
Theoretical Implications Summary Conclusion	Theoretical Implications/Summary/Conclusion section is usually a final section of the paper containing critical analysis and interpretation of results.
References	References section provides references to sources organized in accordance with editorial guidelines and instructions
Appendix	Important nonintegrated data are placed in Appendix Section.
Acknowledgements	Acknowledgements section is typically placed at the beginning or in the end of the article and is expression of gratitude to all who helped in research, writing the article etc.

The first stage of processing the corpus is creating aggregate document storage and filling it with articles. It is necessary to provide convenient and efficient support (storage and adding) of raw documents. Creating a separate catalogue to store documents with rubrics identified by experts is vital to implement GATE in future. Besides, Alfresco is implemented in Java language, which makes ontology processing

components integration with Semantic Web tools easier, as these tools are implemented on this language.

Linguistic markup is one of the key concepts of corpus linguistics. Linguistic markup identifies texts various parameters, allowing to achieve intelligent search in corpus. Text markup allows include metadata attribute to texts and their components. Basic markup is provided by GATE ready functions: tokenization and paragraphs, sentences and words markup on its basis; morpho-syntactic analysis (identifying the part of speech). More complicated markup (bibliography, the credits, etc.) may be realized by GATE tools improvement. The set of key words represents the paper in general and characterizes the work from the point of its relevance. Therefore characterizing the text via key words is critical for efficient academic search. Identifying key concepts cannot be executed through basic functions of GATE that is why additional module with application of GATE API is implemented in the system. To develop research prototype frequency approach is used due to its ease of use.

5 Conclusion

Implementation of discussed methods is intended to decrease significantly labour intensity of information search and analysis, provide operational use of information in research, and increase the amount of information from different sources available for processing. The basic mechanism of the system is knowledge oriented which allows providing integrated solutions to the tasks. Now it can be seen that the basis for creation of the intellectual system supporting research and providing efficient feedback is developed.

Acknowledgements. The reported study is supported by RFBR, research project №14-07-31273.

References

1. Bird S., Liberman. M. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, University of Pennsylvania, Philadelphia, PA, 1999.
2. Cunningham H., Maynard D., Bontcheva K. Text Processing with GATE. – Gateway Press CA, 2011.
3. CNXML/DocumentOntology http://mathweb.org/wiki/CNXML/DocumentOntology
4. Dublin Core Metadata Element Set, Version 1.1 http://dublincore.org/documents/dces/
5. Document Ontology (draft) http://www.cs.umd.edu/projects/plus/SHOE/onts/docmnt1.0.html
6. Grishman. R. TIPSTER Architecture Design Document Version 2.3. Technical report, DARPA, 1997. http://www.itl.nist.gov/div894/894.02/related_projects/tipster/.
7. Muninn Documents Ontology http://rdf.muninn-project.org/ontologies/documents.html
8. XML Languages http://cnx.org/help/authoring/xml
9. Varma P. Project Documents Ontology http://vocab.deri.ie/pdo.

Система анализа научных публикаций на основе онтологического подхода

Вячеслав В. Ланин, Светлана А. Стринюк
Национальный исследовательский университет «Высшая школа экономики»
vlanin@live.com,strinuk@mail.ru

Аннотация. В статье описывается подход к созданию хранилища научных публикаций с поддержкой семантического индексирования на основе онтологического подхода, методов компьютерной лингвистики и обработки неструктурированных данных. В качестве инструментальной среды для обработки текстов используется платформа GATE. Для анализа публикаций используются специально разработанные онтологические ресурсы, описывающие структуру публикаций и их формат. Также при обработке текстов используются словари ключевых слов и частотные характеристики текста. Реализация предлагаемого подхода позволит упростить поиск и анализ публикаций по заданной тематике, выявить связи между ними.

Ключевые слова. онтология, GATE, научная публикация.