Semantic analysis implementation in engineering enterprise content management systems

A V Ivaschenko¹, A A Stolboa², D N Krupin³, A V Krivosheev¹, P V Sitnikov¹ and O Ja Kravets⁴

¹ Samara State Technical University, 244, Molodogvardeyskaya str., 443100, Samara, Russian Federation
² Samara National Research University, 34, Moscow highway, 443086, Samara, Russian Federation
³ SEC “Open code”, 55, Yarmarochnaya str., 443001, Samara, Russian Federation
⁴ Voronezh State Technical University, 14 Moscow ave., Voronezh, 394026, Russian Federation

E-mail: anton.ivashenko@gmail.com

Abstract. The paper introduces a new solution for semantic analysis implementation in modern enterprise content management (ECM) systems. The system of semantic analysis is intended for the intellectual analysis of enterprise official and technical documents based on machine learning, namely the extraction of the specified attributes from them for further use. In this paper it is proposed to implement semantic search using the extracted data configurator, which is responsible for creating and managing ontologies. From the configurator of the extracted data by the name of the document type, a graph is generated containing attributes to be extracted (official terms and sections, dates, etc.), regular expressions to search for sentences that probably contain the desired attribute, Yargy and regular rules for extracting attributes from the arrays of sentences. The proposed solution was successfully probated and tested on a dataset containing engineering enterprise contract agreements and protocols.

1. Introduction

Modern enterprises require new technologies for processing of project and engineering documentation capable of semantic analysis and search. According to the basic trends in industry new computer aided systems need to process digital twins. Despite the comprehensive facilities of present day project data management and project lifecycle systems, automation of documentation and archival repositories remains insufficient.

The main challenge is concerned with the complicated search and time consuming construction of new project documents with high volume of matching content. To solve these problems it is proposed to remove the workflow records system with enterprise content management (ECM) software that provides improved search facilities and a possibility of semantic analysis of textual and pictorial fragments of project documents.

In this paper there is proposed a new software solution. It is presented how the problem of semantic analysis of large volumes of textual data can be performed using the knowledge base powered by
Ontology. The examples of ECM system practical implementation prove the efficiency of Ontologies utilization for documentation processing systems at modern industrial enterprises.

2. Related works
The concept of Industry 4.0 [1, 2] describes a solution vision based on implementation of modern IT technologies to develop cyber-physical systems for smart factories. Based on existing and rich experience of manufacturing automation it picks out the best stack of technologies that can be applied in practice to improve the efficiency of the general supply chain. In this respect the opportunity to organize and support virtual communication between the decision makers in integrated information space becomes critical.

Modern level of digitalization allows implementing Internet based services for engineering enterprise content management [3, 4]. The process of interaction of users in integrated information space at modern production enterprises and supply chains generates a sequence of events of the exchange of documents, messages and other information objects. The number of the events is big (large physical data volume); they vary and require high-speed processing. In this regard, the task of managing the collection and processing of information data in the system of acquisition and processing system with a stratified architecture may be referred to the Big Data problem [5].

One of the solutions can be close to subject-oriented approach for business processes management (S-BPM), which conceives a process as a collaboration of multiple subjects organized via structured communication [6]. There can be proposed a model for the interaction of actors (subjects) in integrated information space, which can be implemented using the multi-agent software. The ideas of indirect and conditional project management generating soft influence over highly motivated autonomous actors are being successfully implemented in digital social and economic systems [7–9].

To coordinate decision makers on a semantic level the mechanism of knowledge description in the form of Ontologies can be used [10, 11]. Ontology represents the semantic network combining concepts, attributes, relations and rules that describe the situation from the predefined point of view. Ontologies of human actors differ, although for a certain problem domain they tend to agree on similar perception and understanding. Due to this fact problem domain Ontologies are treated as impersonal and objective. Ontologies of computer agents can be formally predefined and used to specify the definite and explicit logic of their behavior.

3. Materials and methods
Enterprise content management systems [12] refer to a new type of software, which is close in purpose to documentation workflow and product lifecycle systems, but does not duplicate them in general. According to the classical definition, ECM extends the concept of content management by adding a timeline for each content item and possibly enforcing processes for the creation, approval and distribution of them.

Considering the key role of semantic search in ECM concept, the corresponding content analysis should be based on modern technologies. The system of semantic analysis of documents is intended for the intellectual analysis of enterprise official and technical documents based on machine learning [13], namely the extraction of the specified attributes from them for further use. In this paper it is proposed to implement semantic search using the extracted data configurator, which is responsible for creating and managing ontologies. Some technologies useful for its implementation are listed below.

Documents acting as input data of the implemented subsystem can be presented in various formats, including in the form of scans of documents. So, one of the tasks is image recognition in order to obtain text. Tesseract [14, 15] is the highest quality and most popular open source library for optical recognition, and is one of the freeware software. The advantages of this library include easy learning in the recognition of various fonts, which improves the quality of recognition.

One of the serious reasons why the quality of recognition is deteriorating is the rotation of the original image: a rotation of one degree gives a loss of recognition accuracy of 20%. In order to eliminate such errors, it is necessary to solve the problem of preprocessing the original image, namely,
performing the rotation of the original image. To implement this type of pre-processing, the Leptonic library was selected, which, according to world trends, is the most popular open source library for image processing and analysis [16].

Another reason for recognition errors is that the image size is too small, because Tesseract perceives characters that are less than 10 pixels in size for interference, and discards them, which can lead to reduced accuracy and loss of significant information. To resize the image, interpolation is used, the essence of which is to determine the unknown values of the added pixels based on the available data. One of the interpolation methods is the nearest neighbor method, which is the least expensive [17]. As a result of its application, each pixel simply becomes larger, but the image quality is low.

Bilinear interpolation takes into account the square (2x2) of the nearest known pixels, and the weighted average value of these pixels is used as the interpolated value, however, with this type of interpolation, the graininess of the image remains [18]. Bicubic interpolation takes into account 4x4 known pixels and is the most optimal in quality and speed, which is proved by its use in various editors (for example, Adobe Photoshop). In this regard, it is proposed to use Bicubic interpolation to increase the original images [19, 20].

Information extraction is the most critical feature for semantic analysis and search. One of the easiest and most effective ways to extract information is using regular expressions [21, 22]. In addition, regular expressions can be used to preprocess text. This technology is based on a special system for recording samples for search. The sample that defines the search rule in Russian is also sometimes called the "template" or "mask". Most programming languages support regular expressions for working with text strings. The use of this technology does not require additional costs in the form of the use of paid closed solutions, for example, Stanford NLP, spaCy, NLTK, etc. as it would be when using the technology of extracting entities or the use of neural network solutions.

To extract texts in Russian, the Yargy library is used [23]. In it, the rules are described in the format of context-free grammars and dictionaries of keywords. A bank of ready-made rules for names, dates, addresses and other entities is available in the Natasha repository. Another text extraction tool is the use of tomita-parser developed by Yandex. These solutions have several differences:

- yandex/tomita-parser has a console interface, while Yargy is a Python library;
- yandex/tomita-parser uses its own language and Protobuf files to describe the rules; in Yargy, grammars and dictionaries are described in Python;
- along with Yargy, ready-made rules are published in the Natasha repository for retrieving names, dates, addresses and other entities. Yandex does not disclose its rules for yandex/tomita-parser.

4. Solution architecture
We propose to implement the semantic analysis module as an ECM subsystem in the form of an independent web service. The main modules of the developed subsystem are:

- data extraction module, consisting of a service for processing contracts and a service for processing documents.
- optical recognition module (OCR);
- service for the classification of documents;
- configurator of extracted data.

The architecture is presented in figure 1. Documents from external software are transferred as input data to the subsystem of semantic analysis on the OCR module. After the processes of optical recognition and classification of the document received in the subsystem, the document type is sent to the configurator of the extracted data. From the configurator of the extracted data by the name of the document type, a graph is generated containing:
- attributes to be extracted (official terms and sections, dates, etc.);
- regular expressions to search for sentences that probably contain the desired attribute;
- Yargy rules for extracting an attribute from an array of sentences found;
- regular rules for extracting attributes (different from regular expressions for searching sentences by predicate name).

![Image](image.png)

**Figure 1.** Solution architecture.

The result of the extracted data configurator is an associative array, where the keys are the names of the attributes (AttrName), and the values are the rule sets to apply (RegExpRule, YargyRule, YargyRegExpFilter). This associative array together with the text of the document is sent to the information extraction module.

Information extraction module performs the following stages:

- splitting the text into several fragments of n sentences in each (to distribute the extraction task by core);
- compiling the list of sentences related to the required attributes (RegExpFilter) using regular expressions extracted from the Ontology;
- processing the main extraction, consisting of several cycles, on the received list of offers.

In the main cycle of the third stage of the information extraction module, the selected rows are processed. Each line is processed by one of two internal loops: in one, the element is checked for compliance with Yargy rules (YargyRule), in the other, for compliance with regular expressions from the rule list (RegExpRule).

Ontology configurator is responsible for creating and managing the retrieved data. The configurator provides the following functions:

- creation of subject-oriented ontologies through the description and management of real-world entities and their attributes;
- data storage and providing access;
- visualization, editing, adding, deleting ontologies;
- management of optical text recognition settings;
- settings management for integration with corporate systems.

Attributes of domain-specific ontologies are matched with data represented by sets of regular expressions and Yargy rules. Data is stored in a specialized storage (Triplestore) or in the form of xml-files in rdf / xml format. The configurator provides the ability to access data at the time of the request from the information extraction module or administration module.

The creation and description of a subject-oriented ontology is carried out using a specialized ontology editor. The editor is designed to build conceptual schemes. It provides the ability to display concepts, attributes, and relationships using text descriptions or graphic elements.

Classification service provides the following functions:
• Compilation of a data set:
  • collection of documents of various types;
  • markup of documents;
  • transformation of the document to a form suitable for machine processing in several
    variations to assess the effectiveness of each approach to the text vectorization;
  • Training of machine learning models, namely the selection, training and stacking of models
    for the classification of contracts entering the system.

Documents processing service is also introduced to analyze various types of documents and
provides:

• implementation of methods for extracting the required units of information through regular
  expressions and rules for the Yargy library;
• correlation of regular expressions and rules with ontological entities;
• preservation of regular expressions and rules in ontologies in the form of literals and separate
  entities with many literals, for use in treaty processing.

5. Implementation results
The results of implementation are illustrated in figure 2. The proposed solution was probated and
tested on a dataset containing enterprise contract agreements and protocols. The documents refer to
official legal workflow. From one side, they are structured and contain typical parts (like dates, parties
and formal clauses). From the other side, they differ in format, structure and filling, which makes
semantic analysis a non-trivial task.

![Figure 2. Implementation.](image)

The subsystem accepts a POST request with a document (file), an indication of the task number
(task_id) and document type (doc_type). The resulting file is stored in a database for temporary
storage of processed files with the identification number of the task (task_id) to which it belongs.
Next, the task in the form of its task_id number and document type doc_type is sent to the task queue
(celery). Celery worker restores a document from the database by task_id task number.

Then the doc_type check is performed: if the document type refers to the specific type, the
use_column_processor marker is set, which will be transferred to the optical text recognition unit to
use the “solution for extracting the parties”. When transmitting an implicit document type (“protocol”
or “contract”), the Subsystem calls a classification function that determines the specific type of
document from the headers (for example, “protocol_sd” or “service_contract”).

The subsystem sends a document with the use_column_processor flag to the OCR server Arcadia.
The Subsystem receives a response from the Arcadia server containing text containing markers at the
places where the source document contains tables and tables in the form of two-dimensional arrays. Next, the function of searching and retrieving attributes is called, into which the following data is transmitted: document text (plain_text), table information in the form of an associative array “table id” - “two-dimensional array” (dictionary) and the exact type of document (classified_type).

An XML structure (as a string) comes from the attribute extraction function, which contains the nodes corresponding to the nodes of the graph that describes the attributes of a particular type of document. A request is generated, which includes task_id, an XML structure with the extracted information, recognized text with table markers, and a table in the form of an associative array “table id” - “two-dimensional array”. The subsystem sends the generated request to the external software [24, 25].

User interface is presented in figures 3, 4. Ontology is represented for the user in the form of a semantic network. The user can edit the concepts and attributes, add new rules and manage the document types.

![Figure 3. View of Ontology semantic network node details.](image1)

![Figure 4. View of Ontology literal details.](image2)

### 6. Conclusion
Implementing ontologies can significantly improve the quality and usability of semantic analysis in modern ECM systems. New opportunities for the users include the developed search facilities, and considerably new services of documents comparing and automated generation. This is especially required in ECM systems used for typical engineering projects management. Therefore intelligent ECM systems based on ontologies become an essential part of Industry 4.0.
Acknowledgments
The reported system usability research was supported by RFBR according to the research project № 20-08-00797.

References
[1] Kagermann H, Wahlster W and Helbig J 2013 Recommendations for implementing the strategic initiative Industrie 4.0 Final report of the Industrie 4.0 Working Group 82
[2] Lasi H, Kemper H-G, Fettke P, Feld T and Hoffmann M 2014 Industry 4.0. Business & Information Systems Engineering 4(6) 239-42
[3] Digital Russia 2017 New Reality Digital McKinsey 7 133
[4] https://www.cigionline.org/activity/global-commission-internet-governance
[5] Baesens B 2014 Analytics in a Big Data world: The essential guide to data science and its applications (Wiley) 232
[6] Fleischmann A, Schmidt W and Stary C 2015 S-BPM in the wild (Springer) 282
[7] Ivashenko A V, Barabanov V F, Podvalny E S 2015 Conditional management technology for multimedia interaction Automation and Remote Control 76(6) 1081-7
[8] Surnin O L, Sitnikov P V, Khorina A A, Ivashchenko A V, Stolbova A A and Ilyasova N Yu 2019 Industrial application of big data services in digital economy CEUR Workshop Proceedings 2416 409-16
[9] Ivashenko A, Korchivoy S and Spodobaev M 2020 Infrastructural models of intermediary service providers in digital economy Advances in Intelligent Systems and Computing 1038 594-605
[10] Pouchard L, Ivezic N and Schlenoff C 2000 Ontology engineering for distributed collaboration in manufacturing Proceedings of the AIS2000 Conference 1012-7
[11] De Nicola A, Missikoff M and Navigli R 2009 A software engineering approach to Ontology building Information Systems 34(2) 258-75
[12] Kampffmeyer U 2006 ECM enterprise content management (Verlag / Herausgeber Project Consult) 84
[13] Hey T, Butler K, Jackson S and Thiyyagalingam J 2020 Machine learning and big scientific data Philosophical Transactions of the Royal Society A 378(2166) 20190054
[14] Sun P, Xie Q, Wu Z, Feng X, Cai J and Jiang Y 2019 Characters recognition based on tesseract-OCR 2019 IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference 102-6
[15] Gupta S and Meena A K 2018 A practical implementation of automatic document analysis and verification using Tesseract 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems 401-6
[16] Vasileiadis M, Kaklanis N, Votis K and Tzovaras D 2017 Extraction of tabular data from document images Proceedings of the 14th Web for All Conference on The Future of Accessible Work 1-2
[17] Savchenko A V 2017 Maximum-likelihood approximate nearest neighbor method in real-time image recognition Pattern Recognition 61 459-69
[18] Li J, Su J, Zeng X 2019 A solution method for image distortion correction model based on bilinear interpolation Computer Optics 43(1) 99-104
[19] Li Y, Qi F and Wan Y 2019 Improvements on bicubic image interpolation 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference 1 1316-20
[20] Aràndiga F 2016 A nonlinear algorithm for monotone piecewise bicubic interpolation Applied Mathematics and Computation 272 100-13
[21] Elton D C, Turakhia D, Reddy N, Boukouvalas Z, Fuge M D, Doherty R M and Chung P W 2019 Using natural language processing techniques to extract information on the properties and functionalities of energetic materials from large text corpora arXiv:1903.00415
[22] Miyazaki Y and Shinshi K 2017 Searching mathematical expressions with regular expressions tool and its application to extract mathematical concepts Society for Information Technology & Teacher Education International Conference 1974-8

[23] Radishevskii V L, Kulnevich A D, Chugunov R A and Shevchuk A A 2018 Distributed GLR-parser for natural language processing CEUR Workshop Proceedings 2267 374-7

[24] Mihaylov D 2019 A Way to Accelerate the Process of Gathering Information for Decision-making International Journal on Information Technologies and Security 4(11) 39-50

[25] Ivanova D and Elenkov A 2019 Big Data Analytics for Air Quality Monitoring Assessment Based on IoT Platform International Journal on Information Technologies and Security 2 (11) 43-50