Experimental Integration of University Knowledge Bases Founded on Semantic Web Technologies

Victor Telnov[1][a], Konstantin Odintsov[2][b]
[1]National Research Nuclear University MEPhI, Obninsk, Russia
[2]Lomonosov Moscow State University, Moscow, Russia
telnov@bk.ru, mr.odintsov2@gmail.com

Keywords: Semantic web, Knowledge base, Machine learning, Text classification, Cloud computing.

Abstract: The paper deals with the issues of finding and researching optimum algorithms for classification of textual network content in Russian and English in the interests of its integration with existing knowledge bases. Testing of the algorithms is carried out by the method of cross–-validation. The novelty of the presented research is due to the application of the Pareto's optimality principle for multi–-criteria evaluation and ranking of the studied machine learning algorithms, provided that there is no a priori information about the comparative importance of the criteria. The features of the software implementation of efficient classification and semantic annotation algorithms as part of a scalable semantic web portal hosted on a cloud platform are discussed. The proposed software solutions are based on cloud computing using DBaaS and PaaS service models to ensure the scalability of data warehouses and network services.

1 INTRODUCTION

Classification and integration of weakly structured knowledge based on ontologies is a non-trivial task of informatics. It is usually easy for a human to understand whether two or more entities are related based on cognitive associations, while a computer is not always able to do this. Empowering machines with common sense and domain understanding has been and continues to be an important challenge in the field of artificial intelligence.

As of 2023, educational web portals of universities and open corporate knowledge management systems are isolated from each other, do not use the capabilities of the semantic web and machine learning methods. For distance learning, the educational platforms Coursera, edX, Moodle, etc. are used, which are based on taxonomies, thesauri and are not very compatible with each other in terms of data formats. The project considered here (Knowledge graphs on computer science) is a step towards the virtual integration of educational resources, initially for university knowledge bases in computer science and programming. The long-term goal of the entire project is a public semantic educational web portal of a new generation, which is based on machine learning methods (Geron, 2019) and semantic web technologies (W3C Semantic Web). Potential beneficiaries of information solutions that are designated as project results (target audience): students, teachers, managers, experts, researchers and specialists in the field of computer science and programming.

The novelty of the implemented approach is as follows. In order to implement virtual integration (federalization) of external weakly structured data with existing knowledge bases, it is necessary to somehow intelligently embed external data into existing ontologies. A number of machine learning algorithms are known that do this with varying degrees of accuracy. Using the Pareto's optimality principle (Vilfredo Pareto), each time we choose the best virtual knowledge integration algorithm by performing multi-criteria optimization of the algorithms, provided that there is no a priori information about the comparative importance of the criteria.

Research groups from Stanford University, the Massachusetts Institute of Technology, the University of Bari, the University of Leipzig, and

[a] https://orcid.org/0000-0003-0176-5016
[b] https://orcid.org/0000-0000-0000-0000

the University of Manchester are focusing on the development of the Semantic Web, related issues of machine learning and natural language processing (NLP). In Russia, specialized research is carried out at the Competence Center of the NTI MIPT, ITMO University, at the Faculty of the CMC of Moscow State University, at the ISP RAS named after. V.P. Ivannikov.

Within the framework of the project (Knowledge graphs on computer science), the tasks of creating and connecting two new components of the web portal are solved: 1) the "Semantic Classification" agent, which implements classification algorithms for textual network content in Russian and English for automated filling and virtual integration (federalization) of knowledge bases on computer science and programming; 2) agent "Ontology Editor WebProtege", which provides control of remote access to knowledge bases by professors and knowledge engineers and provides tools for collaborative editing of ontologies. Expected result: algorithms for seamless virtual integration of external data sources with knowledge bases in Russian and English. Seamless integration means the ability to navigate integrated external data sources using an RDF browser (Telnov, Korovin, 2020).

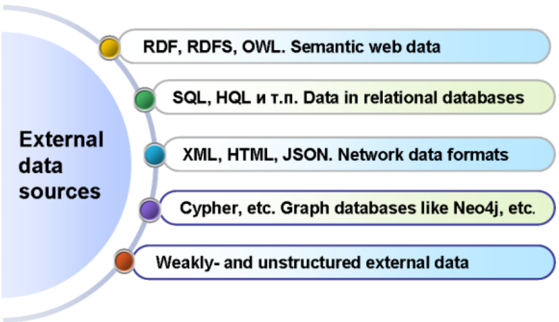

Figure 1: External source data formats for virtual integration with knowledge bases.

External data integration algorithms are under construction for the following common formats, see Figure 1:

1. Semantic data, RDF, RDFS, OWL formats. SPARQL Query Language.
2. Data in relational database tables. Query languages SQL, HQL, etc.
3. Popular network data formats XML, HTML, JSON.
4. Data in graph databases like Neo4j, etc. Cypher query language.
5. Weakly structured (CSV) and unstructured external data.

The following seven knowledge graphs on computer science and programming act as training sets:

1. Programming technologies.
2. Object-oriented programming.
3. Front-end web programming.
4. Paradigms and programming patterns.
5. Cloud services and technologies.
6. Semantic Web.
7. Integrated knowledge graph.

To solve the problems of classifying textual network content and integrating it into knowledge bases, the following machine learning methods are studied:

1. Softmax classifiers (maximum entropy model).
2. Classifiers based on the Support Vector Machine.
3. Naive Bayes classifiers.
4. Classifiers based on Terminological Decision Trees.
5. Classifiers using the Nearest Neighbors method.

In the course of solving the above tasks, public software (Apache Jena, Stanford NLP, Scikit-learn, Weka) and original program code are used.

2 RESEARCH METHODOLOGY

The subject of research in the project (Knowledge graphs on computer science) is algorithms for the virtual integration of external semantic and non-semantic data with knowledge bases in the field of computer science and programming. Research methods include design, software implementation, tuning and testing of virtual data integration algorithms. Cross-validation is used to test the effectiveness of natural language text classification algorithms. Algorithm performance indicators are known machine learning metrics *Precision*, *Recall*, *F1-score*. The *Precision* metric characterizes the ability of the algorithm to distinguish classes from each other. The *Recall* metric measures the algorithm's ability to detect a particular class at all. The *F1-score* metric is informative in cases where the values of the first two metrics differ significantly from each other.

Having received the results of testing five classification algorithms on seven knowledge graphs in Russian and English, then a set of Pareto-optimal algorithms is determined, which are the best in the aggregate of all computational experiments performed. The optimization problem is formulated as follows. It is required to choose the best classification method, taking into account all the calculated quality indicators, without making any a priori assumptions about the relative importance of these indicators. For this, in the class of transitive antireflexive binary relations, the Pareto's relation in the Euclidean space is considered. This relationship is also called the dominance relationship. The essence of this relationship is as follows. It is said that some element x from some set dominates another element y from the same set if x is not worse than y in all aspects (criteria) and at least one aspect is superior to y. The set of all non-dominated elements is called the Pareto's set. The binary Pareto relation provides a universal mathematical model of multi-criteria context-independent choice in the Euclidean space. Based on the Pareto's relation, a choice function is obviously constructed, which generates the set of elements that are the best, taking into account all the calculated metrics, without any a priori assumptions about the comparative importance of these metrics. The corresponding mathematical calculations are presented in (Telnov, Korovin, 2022; Telnov, Korovin, Odintsov, 2023).

From a practical point of view, the project (Knowledge graphs on computer science) is implemented on the Jelastic cloud platform in the J2EE and Python runtimes according to the PaaS service model. Semantic repositories are hosted on a cloud platform according to the DBaaS service model. This provides scaling of data storage and network services, guaranteeing the software against braking with a sharp increase in the intensity of network requests. The development is based on the standards and technologies of the semantic web: RDFS, OWL, SPARQL, as well as the description logics ALC and SROIQ(D) (Telnov, Korovin, 2020). Full-text educational objects and media content are hosted in arbitrary remote data storage and video hostings. The choice of a specific remote storage is not critical, arbitrary repositories equipped with content display tools (Google Drive, Yandex.Disk, YouTube, etc.) can be used.

3 RESEARCH RESULTS

During the implementation of the project (Knowledge graphs on computer science), on seven corpora of specialized texts on computer science and programming, the effectiveness of relatively simple, intuitive machine learning methods for solving the problem of automated filling from the WWW and integrating knowledge bases without direct human participation was studied. Moderate amounts of initial data were used to test machine learning methods. Each of the seven involved knowledge graphs contained no more than one thousand objects and no more than one hundred classes.

Each training set was randomly divided three times into three samples of approximately the same size. Each of the three samples was in turn declared a control sample, the remaining two samples were combined into a training sample. The text classification algorithm was adjusted according to the training sample, and then classified the objects of the control sample. The described procedure was repeated three times for each text classification algorithm and for each knowledge graph. To assess the quality of classification algorithms, macro-averages of machine learning metrics *Precision*, *Recall*, *F1-score* were used, when specific metric values are averaged over all classes, regardless of the number of objects in these classes.

The results obtained in the course of computational experiments allow us to draw the following conclusion. Among the five tested methods for classifying texts in natural languages, the "Support Vector Machine" method is the leader. It is the only one included in the set of Pareto-optimal algorithms. The "Nearest Neighbors Classifier" method is not much inferior to it. The "Maxent Classifier (Softmax)" method looks like an outsider compared to other methods. It should be noted that the "Support Vector Machine" method is binary, that is, it allows you to distribute elements in just two classes. This technical limitation is overcome by multiple one-against-all and one-against-one classifications.

The educational product under construction is a semantic web portal on computer science and programming (Knowledge graphs on computer science) esigned in accordance with the architectural pattern MVC (Model-View-Controller) (The Model View Controller Pattern). Algorithms for processing and presenting data are separated from each other and from the actual data (learning objects). Learning objects are combined through ontologies into knowledge graphs, which are placed in semantic repositories on a cloud platform.

Figure 2 below shows an enlarged diagram of the components of a semantic educational web portal on computer science and programming

(Knowledge graphs on computer science). New components are highlighted in color. In more detail, the structure of the software product is as follows:

1. Semantic repository as a storage of knowledge graphs, equipped with the Apache Jena engine.
2. Search widgets to quickly dive into knowledge graphs. Their interface is similar to how the search query string works in popular search engines.
3. Intelligent RDF browser for visual navigation of knowledge graphs. The intuitive visual way of navigation is reminiscent of a computer adventure game, does not require special skills and is accessible to an inexperienced user. Students learn how to interact with the RDF browser in minutes.
4. Component for context-sensitive search of network content in the WWW and specialized resources. Uses the capabilities of standard search engines.
5. A component for semantic annotation of textual network content in the interests of filling and updating knowledge graphs.
6. New software component "Semantic Classification" for automated classification of textual network content in the interests of filling, updating knowledge graphs and integrating knowledge bases.
7. Plug-in software component "Ontology editor WebProtege" for managing remote access to knowledge graphs (ontologies) by teachers, knowledge engineers and for joint editing of ontologies.
8. Public access points to international knowledge bases DBpedia, Wikidata

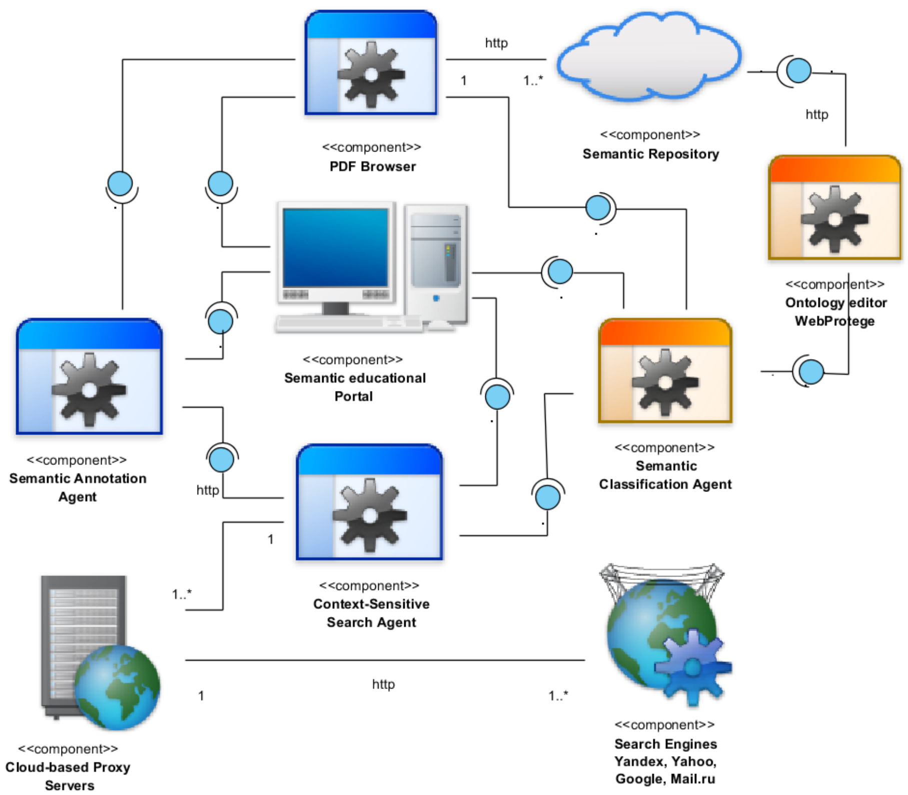

Figure 2: Component diagram of the scalable semantic web portal. New components are highlighted in orange.

4 DISCUSSION

The results presented in the previous section were compared with data obtained by independent researchers on other text corpora, including using advanced methods of deep machine learning. In a recent review (Minaee, Kalchbrenner, Cambria, Nikzad, Chenaghlu, Gao, 2022), Table 1 on page 27 shows the results of testing a number of machine learning algorithms for solving text classification problems. In particular, the Naive Bayes classifier in our study showed an average accuracy of 86%, while the same classifier on the much larger SST-2 text corpus gave an accuracy of 81.80%. Deep machine learning algorithms on the SST-2 corpus of texts give an average accuracy of 91%, which is only 5% better than the average accuracy of 86% that the Naive Bayes classifier provides on the modest datasets in the present study. From this it is possible to conclude that the support vector machine, the nearest neighbors method and the naive Bayes classifier provide sufficient competence of semantic knowledge bases as an artificial intelligence systems.

The initial filling and subsequent updating of knowledge bases, their integration with third-party educational resources, is the prerogative of university professors and knowledge engineers. To do this, the project (Knowledge graphs on computer science) provides for the connection of the program component "Ontology Editor WebProtege", which is designed to control remote access to knowledge bases (ontologies). An educational product can be replicated without restrictions, adapting the content of knowledge bases to the new types of educational programs or levels of education. In the course of further software development, it is planned to replenish knowledge bases with educational objects from at least the following external sources (National Open University of Information Technologies; MIT OpenCourseWare). The educational product in question is available from anywhere in the world where the WWW is available. It can be used in traditional and distance learning as the main or additional repository of lecture material, study literature, practical tasks, knowledge control tools, etc.

5 CONCLUSION

The created prototype of the knowledge base in the field of computer science and programming is currently used as part of the educational process at National Research Nuclear University "MEPhI" as part of master's programs. As a result of the introduction of such integrated knowledge bases into the educational practice of universities, one can expect an expansion of the professional horizons of students and an increase in the level of competence of professors due to the emergence of a new unified channel for accessing educational materials from third-party universities. The new educational product will provide university professors with an authoring tool to help create new courses in computer science and programming. The intelligent RDF browser will provide students, professors and all interested persons with the possibility of seamless interactive navigation through the knowledge bases of many universities in the world. This is expected to increase the attractiveness of university education in the eyes of applicants and third parties, as well as create conditions for increasing the level of competence of knowledge bases as an artificial intelligence systems.

FUNDING

The study was supported by the Russian Science Foundation grant No. 22-21-00182.

REFERENCES

Knowledge graphs on computer science. Intelligent search agents, http://vt.obninsk.ru/s/.

Geron, A., 2019. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc. Sebastopol, 2nd edition.

W3C Semantic Web, http://www.w3.org/standards/semanticweb.

Vilfredo Pareto, https://www.newworldencyclopedia.org/entry/Vilfredo_Pareto.

Telnov, V., Korovin, Y., 2020. Semantic Web and Interactive Knowledge Graphs as Educational Technology. In Cloud Computing Security, Ed. Dinesh G. Harkut. IntechOpen, London.

Telnov, V., Korovin, Y., 2022. Application of Machine Learning Methods for Filling and Updating Nuclear Knowledge Databases. *In Izvestiya vuzov. Yadernaya Energetika*, 4. 122–133.

Telnov, V., Korovin, Y., Odintsov, K., 2023. On the Issue of Optimum Machine Learning Methods for Filling and Updating Nuclear Knowledge Graphs. *In Lobachevskii J. Math,* 44 (1), 227–236.

Telnov, V., Korovin, Y., 2020. Semantic web and knowledge graphs as an educational technology of personnel training for nuclear power engineering. *In Nuclear Energy and Technology*, 5(3), 273–280.

The Model View Controller Pattern, http://griffon-framework.org/tutorials/5_mvc_patterns.html.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., Gao, J., 2022. Deep Learning Based Text Classification. A Comprehensive Review. *In ACM Computing Surveys*, 54(3), 1–40.

National Open University of Information Technologies, https://intuit.ru/.

MIT OpenCourseWare, http://ocw.mit.edu/search/?s=department_course_numbers.sort_coursenum.