

Using Ontology for Natural Sciences and Technologies for Vacancies Analysis

Natalia V. Loukachevitch^a, Andrey A. Komissarov^b, Boris V. Dobrov^a, and Sergey V. Shternov^a

^a *Lomonosov Moscow State University, Leninskie Gory, 1, Moscow 119899, Russia*

^b *University 2035, Nobel st., 1, Western Administrative District, Mozhaisky District, Skolkovo Innovation Center, Moscow*

Abstract

In the paper, the ontology-based approach for online vacancy posts is considered. The Ontology on Natural Sciences and Technologies is used as a basic resource for automatic processing. Vacancies are categorized using two large systems of categories. Besides, for a vacancy post, prerequisites can be inferred, which are implicitly contained in the post and required to be learned by a job seeker.

Keywords

Digital economy, text categorization, logical inference, vacancies posts, ontology

1. Introduction

Currently, the concept of digital economy is widely discussed. Digital economy comprises businesses based on digital computing technologies and Internet, including such technologies as artificial intelligence, big data, machine learning, web services, electronic commerce, etc. [1]. Companies actively search for employees, which are experiences in digital economy occupations. Many job-seekers search for the possibility to study digital economy disciplines to find better employment.

In this paper we consider an approach of automatic document processing of online job posting in the sphere of digital economy to help users to find an appropriate job and also to understand what educational prerequisites are supposed to be behind the vacancies posts in digital economy, what disciplines should be studied by job-seekers to find a job in the digital economy sphere.

In this paper we present the vacancy processing pipeline based on Ontology on natural sciences and technologies (OENT) [2–4], which presents thousand concepts and terms in multiple scientific and technological domains as a single semantic network. In previous projects, large volumes of concepts and terms in IT and cybersecurity domains were already introduced into the OENT ontology [5]. During the current study we checked and added missed digital economy terminology into OENT.

We implemented two systems of ontology-based automatic document categorization using OENT concepts. One category system contains subdomains and aspects of the digital economy domain. Another subject heading system is a general system of scientific categories. Using the OENT relations, we implemented the inference of prerequisites required for specific vacancies: for example, in a vacancy post, very specific skills (specific software products) can be required but it is possibly to infer what should be learned to obtain knowledge and skills for the current vacancy in more general terms.

DTTL-2021: International Workshop on Digital Technologies for Teaching and Learning, March 22-28, 2021, Kazan, Russia
EMAIL: louk_nat@mail.com (N. Loukachevitch); komissarov@2035.university (A. Komissarov); dobrov_bv@mail.ru (B. Dobrov); shternov@gmail.com (S. Shternov)
ORCID: 0000-0002-1883-4121 (N. Loukachevitch); 0000-0001-6925-5245 (A. Komissarov); 0000-0003-3110-0779 (B. Dobrov); 0000-0003-2901-5604 (S. Shternov)



© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

2. Related work

Online job services usually use information retrieval and text categorization techniques to provide better access to available job vacancies [6]. There exist international standards on classification of occupations: international standard ISCO [7] and European standard ESCO [8]. ISCO is a four-level hierarchically structured classification system classifying jobs into 10 major categories and 436 unit groups. ESCO data model includes the ISCO hierarchical structure as a whole, and extends it through (i) a further level of fine-grained occupation descriptions and (ii) a taxonomy of skills, competences and qualifications. It supports 24 languages of European Union.

Boselli et al. [8], consider the task of classification of job vacancies using several machine learning methods such as SVM (linear and RBF kernels), Random forest, and Neural networks. They use a collection of 57740 Italian vacancies classified by domain experts. The dataset was split to train, validation, and test sets. Vacancies are classified into 9 major classes. The best results are obtained by the SVM classifier (0.93% F-measure).

Many works are devoted to extraction of skills from vacancies postings [10-12]. Ontologies are used for description and analysis of jobs and related knowledge, skills and abilities [13-15].

3. Ontology on natural sciences and technologies

Ontology on natural sciences and technologies (OENT) is intended for automatic document processing of scientific and technological documents and papers [3, 4]. It related to RuThes-family of resources [2] having the same structure. OENT comprises large volumes of concepts and terminology from several scientific disciplines and technological domains presented as a single semantic network of concepts with corresponding text entries and relations between concepts. The ontology started from creating text collections of text materials (specialized web-sites, school and university text books) in mathematics, physics, geology, biology, chemistry. These collections were used for term candidate extraction. Then terminologists worked with the extracted terms, added corresponding concepts, synonyms and variants and relations into the OENT ontology using available terminological dictionaries in specific domains.

Further, the ontology was expanded on the basis of automatic document processing of scientific papers, also concepts in technological domains (oil and gas industry, hydro power, IT domain and computer security, psychology) were added based on specific projects, in which the ontology was used. OENT comprises as a subpart also Socio-political thesaurus, which allows processing scientific documents containing various socio-political or financial problems. Concepts in OENT have one or more domain labels. Table 1 describes the numbers of concepts according to the domain labels.

Table 1

Subdomains in OENT Ontology

Domain in OENT	Number of concepts
Mathematics	4053
Physics	10737
Chemistry	9166
Geology	4465
Biology	10630
Medicine	6269
Geographical objects	8105
Informatics, IT	5420
Technology (Oil and gas industry, power generation, electronics)	16121
Socio-political concepts	26662
Others	--
Total	106697

4. Structure of OENT ontology

In structure, OENT is a semantic network of concepts. A concept has a unique, unambiguous name. If an unambiguous and clear name in form of an existing word or a phrase cannot be found, than an ambiguous word can be used for naming and supplied with a “relator” (a brief note in parentheses). The concepts have list of text entries that can convey these concepts in texts. The list of text entries for a specific concept can comprise single words of different parts of speech, including ambiguous ones, and phrases that can be either idiomatic or compositional groups. Large rows of synonyms and term variants are collected to provide better recognition of concepts in texts. For each text entry, lemmatized form of words is also stored to provide possibility to match text entries with document after text lemmatization. Ambiguity of a word is described via linking of a word with several concepts.

For example, concept *Машинное обучение с учителем* (supervised learning) has four text entries: *машинное обучение с учителем*, *обучаться с учителем*, *обучение с учителем* (training with supervisor), *обучать с учителем* (to train with supervisor).

. The OENT relation set includes *class-subclass*, *part-whole*, *ontological dependence*, and *symmetric association* (related) relations.

One of the main relations in RuThes-like resources including OENT is the relation of conceptual ontological dependence, which shows the dependence of the existence of one concept on another. An example of such an attitude is the relationship between the concepts *Tree – Forest*, where *Forest* is a dependent concept requiring the existence of the *Tree* concept. The relation of the conceptual dependence is denoted as directed association $asc_1 – asc_2$. Symmetric associations (*asc*) are also possible in only restricted number of cases.

For the relations, properties of transitivity and inheritance are defined, which have the possibility to infer some relations absent in texts. The following properties are defined:

- transitivity of the class-subclass relation:

$$class(c_1, c_2) \wedge class(c_2, c_3) \rightarrow class(c_1, c_3);$$

- transitivity of the part-whole relations:

$$whole(c_1, c_2) \wedge whole(c_2, c_3) \rightarrow whole(c_1, c_3);$$

- inheritance of the whole relationship to subclasses:

$$class(c_1, c_2) \wedge whole(c_2, c_3) \rightarrow whole(c_1, c_3);$$

- inheritance of dependence association relations and symmetric association relations on types and parts:

$$class(c_1, c_2) \wedge asc_1(c_2, c_3) \rightarrow asc_1(c_1, c_3);$$

$$class(c_1, c_2) \wedge asc(c_2, c_3) \rightarrow asc(c_1, c_3);$$

$$whole(c_1, c_2) \wedge asc_1(c_2, c_3) \rightarrow asc_1(c_1, c_3);$$

$$whole(c_1, c_2) \wedge asc(c_2, c_3) \rightarrow asc(c_1, c_3)$$

Considering all possible relation paths existing between two thesaurus concepts C_1 and C_2 , it was supposed that those paths that can be reduced to a single relation with the application of the above-mentioned rules of transitivity and inheritance indicate semantic relatedness between concepts C_1 and C_2 , so called semantic paths.

5. General design of automatic document processing

Typical technical decision for automatic document processing based on OENT (and other RuThes-like resources) includes the following main parts [16]:

- Linguistic resources in form of linguistic ontologies made in the format of RuThes linguistic ontology including an ontology, several subject category schemes for automatic document categorization,
- ALOT instrument for natural language processing of documents, which includes tokenization, morphological analysis, terminological analysis, thematic analysis, text categorization and summarization,
- Visualization of A LOT results including visualization of the thematic structure of a single document and also visualization of large text collections.

5.1. Stages of text processing

The main stages of ontology-based ALOT document processing include:

- Tokenization and lemmatization, that is, the transfer of word forms to dictionary forms (lemmas);
- Matching with the ontology based on the lemma representation of the document. Multiword terms from an ontology are matched with the text using lemma sequences;
- Disambiguation of ambiguous text entries, which means selection of a correct concept for an ambiguous word in context;
- Grouping semantically related concepts into so-called thematic nodes. This provides better determination of concept weights, which are calculated on the basis of the concept frequency in the given document and the significance of the corresponding thematic node;
- Forming the conceptual index of the document. Conceptual index of a document consists of concepts found in the document and their assigned weights. The weight of a concept accounts for the significance of the corresponding thematic node and the frequency of the concept in the document;
- Calculation of category weights in dependence of concepts included into the rules of the inference for this category;
- The results of document processing, including the word index, the conceptual index, the calculated categories, etc. are loaded into an information-analytical system.

5.2. Text categorization

The mainstream technology of automatic document categorization is the machine-learning approach. This approach assumes that there is a sufficient training collection for learning the algorithms. However, many organizations have a need in automatic text categorization, when even a category system of subject headings may be absent and should be created from scratch or with the use of existing similar categorial systems.

In such conditions, machine-learning approaches cannot be applied, and knowledge-based methods of text categorization, i.e. exploiting manual rules for describing categories, are more acceptable. When one creates a hierarchical system of categories and rules for the text categorization in a broad subject domain, it is convenient to use the ontology support, because the ontology allows working not with separate words and expressions, but with concepts and substructures of the ontology [16].

When creating a linguistic profile of a category, each category R is described by a disjunction of alternatives, each clause is a conjunction:

$$R = \bigcup_i D_i ; D_i = \bigcap_j K_{ij}$$

Conjuncts, in turn, are described by experts using the so-called supporting concepts of an ontology. For each supporting concept, a rule for its extension $f(\bullet)$ is set, which determines how, together with the supporting concept, to take into account the concepts subordinate to it in the hierarchy:

- without extension (denoted by the symbol “N”),
- full expansion along the OENT hierarchy tree (symbol “E”) , expansion only by subclass relations (symbol “L”),
- expansion by all types of relations by one level of the hierarchy (symbol “W”),
- expansion by one level of the hierarchy, not including subclass relations (symbol “V”).

The supporting concept can be either "positive", that is, add lower-lying concepts to the description of a conjunct, or “negative”, that is, cut out its subordinate concepts from the rubric description. The result of the application of the extension of the basic concepts is a set of ontology concepts and their text entries, which fully describes the conjunct:

The weight of a conjunct depends on the maximum weight of the concept of the ontology included in it. The weight of a clause is intended to take into account not only the sum of the weights of its constituent conjuncts, but also the measure of the proximity of the conjuncts in the text. The category weight is the maximum weights of the alternatives included in the category description.

The screenshot displays the OENT interface with the following components:

- Концепты (Concepts):** A list of concepts including "ЯЗЫК ПРОГРАММИРОВАНИЯ PYTHON" (highlighted), "ЯЗЫК ПРОГРАММИРОВАНИЯ PYTHON ВЕРСИЯ...", and "ЯЗЫК ПРОГРАММИРОВАНИЯ PYTHON ВЕРСИЯ...". A search filter "ФИЛЬТР КОНЦЕПТОВ" is active.
- Отношения (Relations):** A table showing relationships between concepts. The table has columns: "Отношение", "Асп.", and "Наименование концепта".

Отношение	Асп.	Наименование концепта
АССОЦ	2	БИБЛИОТЕКА JUPYTERHUB
АССОЦ	2	БИБЛИОТЕКА ДЛЯ PYTHON
АССОЦ	2	ВЕБ-СЕРВЕР TORNAДО
АССОЦ	2	ДИСТРИБУТИВ PYTHON ANACONDA
АССОЦ	2	МЕНЕДЖЕР ПАКЕТОВ PIP
АССОЦ	2	ПРОГРАММИРОВАНИЕ НА PYTHON
АССОЦ	2	ПРОГРАММИСТ НА PYTHON
- Синонимы (Synonyms):** A list of synonyms for "PYTHON", including "PYTHON-BИБЛИОТЕКА", "БИБЛИОТЕКА PYTHON", "БИБЛИОТЕКА ДЛЯ PYTHON", "БИБЛИОТЕКА НА PYTHON", and "ПИТОНОВСКАЯ БИБЛИОТЕКА".
- Search and Navigation:** A search bar with "PYTHON PROGRAMMI" and a count of 64. Navigation buttons for page 2129 of 2134 are visible.

Figure 1: OENT interface, where description of Python programming language is presented including text entries and relations

6. Processing of job vacancies using OENT ontology

For processing job vacancies posts, a large collection of vacancies texts in the domain of information technologies was lemmatized, frequency lists of lemmas and phrases were extracted. The most frequent IT concepts and concrete programming tools or devices absent in OENT were added to the OENT with description of variant text entries and relations between OENT units.

Figure 1 shows the working interface of the OENT ontology maintenance. In the left upper corner the list of the OENT units (concepts and named entities) is located. The cursor is set on the Python programming language unit. In the lower left corner, various Python synonyms and variants in

Russian and English are presented. In the upper right corner, relations of the Python language are described. In the current case concepts and named entities, which are dependent from Python existences can be seen, for example Python library. Below variants for mentioning Python library are given.

The Figure 2 shows the coverage of requirement from a machine learning vacancies post with OENT concepts of the following text.

Experience in implementing projects in the field of data analysis and machine learning;
 Understanding of the main methods and algorithms of machine learning, experience of their application in practice;
 Strong knowledge of Python, as well as basic data analysis and machine learning libraries (pandas, numpy, sklearn, scipy, matplotlib, lightgbm, tensorflow, pytorch, etc.);
 Good knowledge of SQL and experience with databases;
 Good knowledge of probability theory and mathematical statistics.

Red colour denotes unambiguous terms or names found in OENT, brown and blue colour indicates ambiguous words, ambiguity of which was resolved automatically. It can be seen that good coverage of the vacancy is provided.

Опыт реализации проектов в области анализа данных и машинного обучения;
 Понимание основных методов и алгоритмов машинного обучения, опыт их применения на практике;
 Уверенное владение Python, а также основными библиотеками анализа данных и машинного обучения (pandas, numpy, sklearn, scipy, matplotlib, lightgbm, tensorflow, pytorch и др.);
 Хорошее знание SQL и опыт работы с базами данных;
 Хорошее знание теории вероятностей и математической статистики.

Figure 2: Example of a vacancy post and its coverage with the OENT concepts.

All extracted vacancies are automatically categorized using the above-described technology.

With this aim, the specialized system of subject headings of the Digital Economy domain was created for classification of vacancies text according to new computer technologies professions. The subject heading system comprises 158 categories. The top level of the subject heading system includes the following subdivisions:

- Digital economy competence,
- Digital economy instruments (software, hardware, Internet),
- Digital economy objects (transport, industrial equipment, power equipment),
- Influence of digital economy (law regulation of Internet),
- Digital economy positions (programmers, analytics, administration, etc.)

Also for vacancies categorization according to different domain, the subject heading system Sciences was used. The subject heading system comprises 268 categories. The main subdivisions of Sciences categories are as follows^

- Mathematics and Informatics
- Physics, mechanics, and astronomy,
- Chemistry
- Biomedicine
- Earth sciences
- Sciences about human and society
- Information technologies
- Engineering technologies

The categories of Digital economy subject headings automatically obtained for the example text (Figure 2) are as follows: Data warehouse, Databases, Artificial intelligence, Libraries of programs, Programming languages. Sciences categories for the given text include: Computer technologies, Informational technologies, Mathematics, Machine learning, Databases, Artificial Intelligence.

7. Describing prerequisites for job vacancies in the digital domain

For describing prerequisites in the OENT ontology, additional relations were added: prerequisite relations and its inverse relation the potential relations. The prerequisite relation is treated as follows; Concept A is treated as a prerequisite of concept B, if to be able to work with B it is required (or desirable) to know or to be able to work with A.

The prerequisite relations can be added manually. For example, for machine learning concept, the linear algebra concept can be introduced as a prerequisite relation. But also it was important to use the properties and inference abilities of existing OENT relations for automatic generation of prerequisite relations on the basis available relations.

- 1) First, the dependence relations, which denote conceptual dependence of one concept from another concept, seems can be directly added as a prerequisite-potential relation. Fig, shows that Python programming language has the potential relations with Python-based libraries and instruments because it should be learned before using those instruments.

If A --> asc₁ (conceptually dependent on) --> C
Then A --> (has) prerequisite --> C

- 2) Second, lower level concepts has as prerequisites their class concepts within the same domain. For example, orthogonal matrix has the matrix concept as a prerequisite because matrices should be learned before to learn their specific subclasses.

If A --> (has) class --> C
Then A – has prerequisite C

- 3) The third rule presupposes inheriting prerequisites to subclasses and instances. For example, for the Python Library, the prerequisite is Python programming language, NumPy is a Python library, Python is also a prerequisite for it.

If A has prerequisite C and D is subclass of A
Then D has prerequisite C

- 4) The last rule expands the prerequisites from the established on to upper levels using subclass-class relations with tag V – (vozmozhno, possible in Russian). For example, the Python language is a prerequisite for the NumPy library, Python is a scripting language. We can infer that the scripting language concept is a possible prerequisite for the NumPy library. All the inference proceeds within the domain of a source concept.

If A --> (has) prerequisite --> C and C – (has) class D
Then A --> has prerequisite (V) --> D
Where V – vozmozhno (possible) aspect of the relations

All the automatically created prerequisite relations have a special label indicating their automatic origin. The relations can be corrected or removed, if necessary. In 2)-4) rules, not only directed class-subclass relations are used, but also class-subclass relations, which can be inferred, using the transitivity property of the class-subclass relations.

Figure 3 presents prerequisites calculated for NumPy library. The prerequisites set includes Python library, Python programming language, Library of programs, High-level programming language, etc.

Altogether, 8260 concepts from IT domain and mathematics have currently prerequisite relations, among which 647 are manually set and 93 K relations are inferred automatically (10 prerequisite relations for a concept on average).

Отношение	Асп.	Наименование концепта
ПРРКЗТ		БИБЛИОТЕКА ДЛЯ PYTHON
ПРРКЗТ		БИБЛИОТЕКА ДЛЯ АНАЛИЗА ДАННЫХ
ПРРКЗТ		БИБЛИОТЕКА ПРИКЛАДНЫХ ПРОГРАММ
ПРРКЗТ		БИБЛИОТЕКА ПРОГРАММ
ПРРКЗТ	В	ВЫСОКОУРОВНЕВЫЙ ЯЗЫК ПРОГРАММИРОВАНИЯ
ПРРКЗТ	В	ДИНАМИЧЕСКИ ТИПИЗИРОВАННЫЙ ЯЗЫК
ПРРКЗТ	В	ЗАЩИТА ИНФОРМАЦИИ
ПРРКЗТ	В	ИНФОРМАЦИЯ
ПРРКЗТ		КОМПЬЮТЕР

Синонимы
PYTHON-БИБЛИОТЕКА
БИБЛИОТЕКА PYTHON
БИБЛИОТЕКА ДЛЯ PYTHON
БИБЛИОТЕКА НА PYTHON

Figure 3: Prerequisites calculated for NumPy library

8. Conclusion

In the paper, the ontology-based approach for online vacancy posts is considered. The Ontology on Natural Sciences and Technologies is used as a basic resource for automatic processing. Terminology of the IT and digital economy domains was added to the ontology. Vacancies are categorized using two large systems of categories. Besides, for a vacancy post, prerequisites can be inferred, which are implicitly contained in the post and required to be learned by a job seeker.

9. Acknowledgements

This work was partially implemented during the project with ANO “University of National Technology Initiative 2035” (UNIVERSITY 2035), agreement № U-19/118 from 22.10.2019.

10. References

- [1] M. Tobji, R. Jallouli, Y. Koubaa, A. Nijholt (Eds), Digital Economy. Emerging Technologies and Business Innovation: Third International Conference, ICDEc 2018 Proceedings, volume 325, 2018.
- [2] N. Loukachevitch, B. Dobrov, Ontologies for Natural Language Processing: Case of Russian, in: Third International Conference Computational Linguistics in Bulgaria, 2018, pp. 93–103.
- [3] N. Loukachevitch, B. Dobrov, Development of Linguistic Ontology on Natural Sciences and Technology, in: Proceedings LREC-2006, 2006, pp. 1077–1082.
- [4] M. Tikhomirov, N. Loukachevitch, B. Dobrov, Assessing theme adherence in student thesis, in: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue, 2019, pp. 649–661.
- [5] N. Loukachevitch, B. Dobrov, Ontological resources for representing security domain in information-analytical system, in: Proceedings of OSTIS conference 2(8), 2018, pp. 185–191.

- [6] International Labour Organization. International Standard Classification of Occupations, 2012. URL: <http://www.ilo.org/public/english/bureau/stat/isco/index.htm>
- [7] European skills/competences, qualifications and occupations, 2020. URL: <https://ec.europa.eu/esco/portal/home>
- [8] S. Chala, S. Harrison, M. Fathi, Knowledge extraction from online vacancies for effective job matching, in: 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), IEEE, 2017, pp. 1–4.
- [9] R. Boselli, M. Cesarini, F. Mercorio, M. Mezzanzanica, Classifying online job advertisements through machine learning, *Future Generation Computer Systems* (86), (2018) 319–328. doi: 10.1016/j.future.2018.03.035
- [10] I. Wowczko, Skills and vacancy analysis with data mining techniques, *Informatics. – Multidisciplinary Digital Publishing Institute*, V 2, №. 4, 2015, pp. 31–49. doi: 10.3390/informatics2040031.
- [11] I. Kivimäki, A. Panchenko, A. Dessy, D. Verdegem, P. Francq, H. Bersini, M. Saerens, A graph-based approach to skill extraction from text, in: *Proceedings of TextGraphs-8 graph-based methods for natural language processing*, 2013, pp. 79–87.
- [12] A. Gugnani, H. Misra, Implicit Skills Extraction Using Document Embedding and Its Use in Job Recommendation, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, No. 08, 2020, pp. 13286–13293. doi:10.1609/aaai.v34i08.7038.
- [13] E. M. Sibarani, S. Scerri, C. Morales, S. Auer, D. Collarana, Ontology-guided job market demand analysis: a cross-sectional study for the data science field, in: *Proceedings of the 13th International Conference on Semantic Systems*, 2017, pp. 25–32. doi:10.1145/3132218.3132228.
- [14] M. Khobreh, F. Ansari, M. Fathi, R. Vas, S. T. Mol, H. A. Berkers, K. Varga, An ontology-based approach for the semantic representation of job knowledge. *IEEE Transactions on Emerging Topics in Computing*, 4(3), 2015, pp. 462–473. doi:10.1109/TETC.2015.2449662.
- [15] D. Çelik, A. Karakas, G. Bal, C. Gültunca, A. Elçi, B. Buluz, M. C. Alevli, Towards an information extraction system based on ontology to match resumes and jobs, in: *2013 IEEE 37th annual computer software and applications conference workshops*, IEEE, 2013, pp. 333–338. doi:10.1109/COMPSACW.2013.60.
- [16] N. Loukachevitch, B. Dobrov, The sociopolitical thesaurus as a resource for automatic document processing in Russian. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2), (2015), pp. 237–262. doi:10.1075/term.21.2.05lou.