УДК 004.912
DOI: 10.37005/2071-9612-2020-4-147-154

BIG DATA APPROACH FOR THE STUDIES OF THE JOB MARKET AND RELATED AREAS

Belov Sergey[1], Javadzade Javad[2], Kadochnikov Ivan[3], Korenkov Vladimir[4], Zrelov Petr[5]

*[1]Lead developer;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
e-mail: belov@jinr.ru.*

*[2]Software developer;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
e-mail: jjavadzade@yandex.ru.*

*[3]Lead developer;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
e-mail: kadivas@jinr.ru.*

*[4]Doctor of technical Sciences, Laboratory director;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
e-mail: korenkov@jinr.ru.*

*[5]PhD, Department head;
Joint Institute for Nuclear Research,
Laboratory of Information Technologies;
141980, Moscow reg., Dubna, 6 Joliot-Curie st.;
e-mail: zrelov@jinr.ru.*

This paper discusses some approaches to intellectual text analysis in application to automated monitoring of the labour market. The construction of an analytical system based on Big Data technologies for the labour market is describedd. Were compared the combinations of methods of extracting semantic information about objects and connections between them (for example, from job advertisements) from specialized texts. A system for monitoring the Russian labour market has been created, and the work is underway to include other countries in the analysis. The considered approaches and methods can be widely used to extract knowledge from large amounts of texts.

Keywords: Big Data, labour market, machine learning.

For citation:
Belov S., Javadzade J., Kadochnikov I., Korenkov V., Zrelov P. Big Data approach for the studies of the job market and related areas. System Analysis in Science and Education, 2020;(4):147–154. Available from: http://sanse.ru/download/420. DOI: 10.37005/2071-9612-2020-4-147-154.

МЕТОДЫ БОЛЬШИХ ДАННЫХ ДЛЯ ИЗУЧЕНИЯ РЫНКА ТРУДА И СВЯЗАННЫХ С НИМ ОБЛАСТЕЙ

Белов Сергей Дмитриевич[1], Джавадзаде Джавад Намик оглы[2], Кадочников Иван Сергеевич[3], Кореньков Владимир Васильевич[4], Зрелов Петр Валентинович[5]

[1]*Ведущий программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
e-mail: belov@jinr.ru.*

[2]*Инженер-программист;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
e-mail: jjavadzade@yandex.ru.*

[3]*Инженер-программист 2 категории;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
e-mail: kadivas@jinr.ru.*

[4]*Доктор технических наук, директор лаборатории;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
e-mail: korenkov@jinr.ru.*

[5]*Кандидат физико-математических наук, начальник отдела;
Объединенный институт ядерных исследований,
Лаборатория информационных технологий;
141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6;
e-mail: zrelov@jinr.ru.*

В данной статье рассмотрены некоторые подходы к интеллектуальному анализу текста применительно к автоматизированному мониторингу рынка труда. Описано построение аналитической системы для исследования рынка труда на основе технологий Больших Данных. Проведено сравнение комбинаций методов извлечения семантической информации об объектах и связях между ними (например, из объявлений о вакансиях) из специализированных текстов. Создана система мониторинга российского рынка труда, ведется работа по анализу рынков труда других стран. Рассмотренные подходы и методы могут быть широко использованы для извлечения знаний из больших объемов текстов.

Ключевые слова: большие данные, рынок труда, машинное обучение.

Для цитирования:
Big Data approach for the studies of the job market and related areas = Методы Больших Данных для изучения рынка труда и связанных с ним областей / С. Д. Белов, Д. Н. Джавадзаде, И. С. Кадочников, В. В. Кореньков, П. В. Зрелов // Системный анализ в науке и образовании: сетевое научное издание. – 2020. – № 4. – С. 147–154. – На англ. языке. – URL: http://sanse.ru/download/420. – DOI : 10.37005/2071-9612-2020-4-147-154.

Introduction

Last years, the prospects for the digital transformation of economic processes were actively discussed. It is quite a complex problem having no solution with traditional methods. Opportunities for the qualitative development of the transformation are illustrated by using Big Data analytics, particularly text analysis, to assess the needs of regional labour markets in the workforce. The problem is solved using the authors' developed automated information system of monitoring matching employers' staffing needs with the training level. The system presented uses the information gathering from open data sources and provides additional opportunities to identify a qualitative and quantitative interrelation between the education and the labour market. The system is targeted at a wide range of users: authorities and management of regions and municipalities; universities' management, companies, recruitment agencies; graduates and prospective students.

Links between the labour market and the educational system is rather complex, with many parties involved (government, universities, employers, individuals, etc.). Both horizontal and vertical mismatch between skills and qualifications from one side and the market's requirements from another is still widely observed in both developing and developed countries. To discover both qualitative and quantitative correlations between the education system and the labour market reasonably fast, we proposed an intelligent system to monitor the employers' demands and match them with the educational standards and programs. The analysis is based on stringing together job requirements and single competencies from the educational standards, the lowest levels of the labor market models, and the education system correspondingly. We used machine learning technologies for semantic parsing and the vector representation of words and short sentences to automate the processing. Big Data approaches and technologies are in use for collecting and processing the data. The system allows to estimate a need for specific professions for regions, consider matching professional standards with real market jobs, and plan the number of funded places in colleges and universities. Having historical data, it is also possible to make some further predictions.

1. Labour market and professional education system

The interaction of the labour market and educational system is a complex process, with many parties involved (government, universities, employers, individuals, etc.). In the ideal world, this interaction would be coherent and perfectly balanced. Mostly it affects youth employment, so-called school-to-work transition. Since the Great Recession unfolding in 2008, youth unemployment has been at the forefront of political and academic debates. In most countries, young employees have suffered more in recession than have older ones [1]. The high unemployment rate, especially among young in a country or region [2] could cause the growth of social tensions or even be a breeding ground for extremism.

Many researchers are giving attention to the volatile labour market and youths' complications with entering it. There are plenty of entities to influence the area, e.g., contract policies for new employees, state programs, etc. Governments still invest a lot in education, so do individuals. However, both horizontal and vertical mismatch between skills and qualifications from one side and the market's requirements from another is still widely observed in both developing and developed countries [3, 4]. This may hinder youths from entering the labour market, causing the fall of education-related expectations, or making people inactive (out of employment, education, and not looking for a job).

From the employer's perspective, a successful worker's qualifications and skills should be on the level required for the job. For potential employees, education quality means competitive advantages. Most of the approaches to discover the real needs of the market use per-area employer and worker surveys. Conducting such polls takes certain time and resources and could not ensure complete labour market coverage.

To discover both qualitative and quantitative correlations between education and labour market in a reasonable time, we proposed an intellectual system to monitor the demands of employers and match them with the standards and programs of higher education [5] As a source of the real-life market needs, it was decided to use job advertisements from job search resources on the Internet (job hunting sites, state and city employment offices, etc.). For the education, texts of the state educational standards along with universities’ educational programs are involved.

The analysis is based on stringing together job requirements and single competencies from the standards. Generally, this cannot be easily defined, seeing as variant vocabulary is used from both sides, source information is quite unstructured, and so forth. Then, to automate the processing as much as possible, we used machine learning technologies for semantic parsing. Since the wording in both requirements and competencies details usually consists of about 10 words, the calculation of the semantic distance between short sentences lies at the core of the method.

The creation of semantic models is one of the well-known key problems of natural language processing. For our task, the way proposed was manipulating vector representation of words and short sentences. Among contemporary word embedding approaches, predictive models based on neural networks are on the top. One of the principal tools for vector representation is word2vec [6] hailed from Google's research.

Having a semantic model, elaborately trained with the text corpus traditional to the subject concerned (for example, IT industry, pharmacology or academic sector, etc.) allows to reveal semantically closes words and phrases, and then give a quantitate assessment how the competencies match the job requirements. Starting from this point, it is possible to rise to higher-level concepts, using the connection through competence-requirement links. For instance, to score contents of the educational profile against the real work field.

Model levels (shown in details in Figure 1):

- Education

Competence → Type of activity → Educational profile → Field of study.

- Labour market

Requirement → Profession → Focus area → Career field.

Continuous collection and analysis of job ads from major sources offer the possibility of monitoring the job market with its trends, focusing on industries or professions, and having statistics for cities and regions.

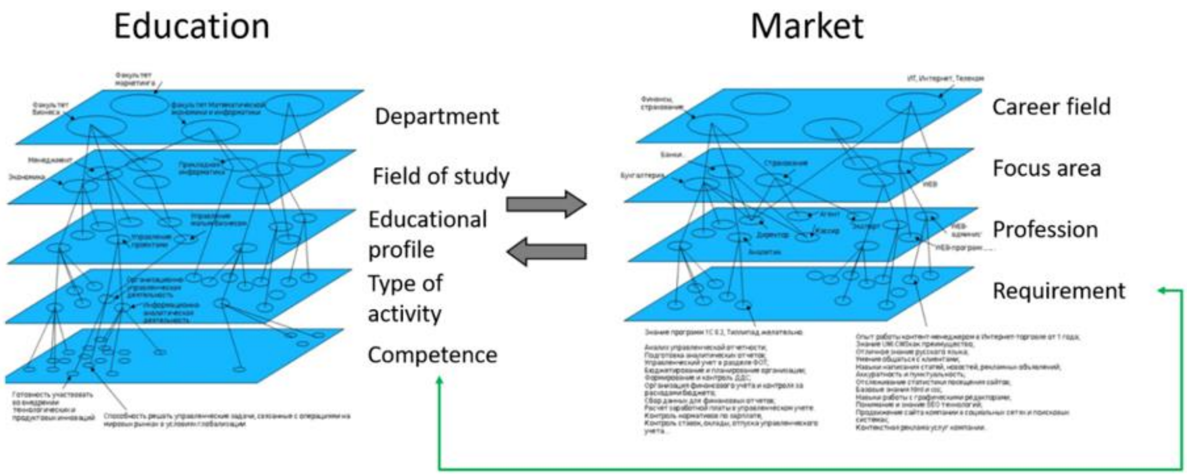
Education
Market
Факультет маркетинга
Факультет бизнеса
Факультет математической экономики и информатики
Менеджмент
Экономика
Прикладная информатика
Управление проектами
Организационно-управленческая деятельность
Информационно-аналитическая деятельность
Готовность участвовать во внедрении технологических и продуктовых инноваций
Способность решать управленческие задачи, связанные с операциями на мировых рынках в условиях глобализации
Department
Field of study
Educational profile
Type of activity
Competence
ИТ, Интернет, Телеком
Финансы, страхование
Банки
Бухгалтерия
Страхование
WEB
Агент
Эксперт
Директор
Кассир
Аналитик
WEB-админист
WEB-програм...
Знание программ 1С 8.2, Таплипад желательно
Анализ управленческой отчетности;
Подготовка аналитических отчетов;
Управленческий учет в разделе ФОТ;
Бюджетирование и планирование организации;
Формирование и контроль ДДС;
Организация финансового учета и контроля за расходами бюджета;
Сбор данных для финансовых отчетов;
Расчет заработной платы в управленческом учете;
Контроль норматива по зарплате;
Контроль ставок, окладах, отпуска управленческого учета...
Опыт работы контент-менеджером в Интернет-торговле от 1 года;
Знание UMI CMS как преимущество;
Отличное знание русского языка;
Умение общаться с клиентами;
Навыки написания статей, новостей, рекламных объявлений;
Аккуратность и пунктуальность;
Отслеживание статистики посещения сайтов;
Базовые знания html и css;
Навыки работы с графическими редакторами;
Понимание и знание SEO технологий;
Продвижение сайта компании в социальных сетях и поисковых системах;
Контекстная реклама услуг компании.
Career field
Focus area
Profession
Requirement

Fig. 1. Mutual mapping between models of the education system and the labor market at different levels of hierarchies

Taking advantage of machine learning and Big Data technologies, the developed system could make near-real-time reporting. Having historical data makes it possible to determine the labour market's current expectations from the education system and make some further predictions. It could be used for tailoring educational programs and a number of future professionals to the needs of the region.

As relevant consumers, we consider authorities, universities, big employers, recruiting agencies. Young people could also reap a benefit from knowing the tendencies while planning their studies and career path.

2. Linking market's requirements with educational competencies

Modeling the semantics (meaning) of the word is one of the critical problems related to natural language processing. The semantic analysis results are used in search engines [7], automatic translation systems, and other fields related to natural language text processing [8].

At the moment, in the approaches of vector representations of words (word embedding), the top place is taken by the so-called predictive model based on the use of neural networks [9]. One of the principal tools for vector representation of words is word2vec [6].

The basic principle of word2vec is to find relations between contexts of words according to the assumption that words that appear in similar contexts tend to indicate similar things (being semantically close). The problem solved by word2vec could be formalized as follows: to minimize the distance between the vectors of words that appear next to each other and maximize the distance between the vectors of words that appear quite far. "Near" in this case means "in similar contexts". For example, the words "analysis" and "research" are often found in similar contexts, word2vec analyzes these contexts and concludes that these words are close in their meaning. Analysis of contexts is performed on large corpora of text. In our task, we used the corpus of the Russian Wikipedia and a national corpus of the Russian language and models of distributional semantics RusVectōrēs [10].

There are attempts to create a predictive model for the translation of a document to a vector space [11]. However, the task of comparing short sentences on the similarity of meaning has specific characteristics, and the use of existing models for translating words or documents to a vector space without modifications gives an unsatisfactory result.

Considering that the text of the language of educational competences, as well as the wording of the requirements in the vacancy announcements, contain an average of about 10 words, the task of evaluation of the semantic closeness of two short sentences is in the basis of the analytical part of the system. Authors have developed the algorithm [12] of sentence translation to vector space based on word2vec.

Thus, each word is mapped to a vector of dimension n (this parameter affects the accuracy of the model). Metric space of mappings of words is used to be called semantic. Projections of the vectors of the words close by meaning are close together as well and form some semantic clusters.

Vector representation allows calculating the "similarity" of words based on the calculation of cosine distance.

So, for two words w_1 and w_2, represented in the form of vectors 1) and 2), the formula to calculate the semantic closeness:

$$cos\left(\vec{V}(w_1), \vec{V}(w_2)\right) = \frac{\vec{V}(w_1) \times \vec{V}(w_2)}{|\vec{V}(w_1)| \cdot |\vec{V}(w_2)|}. \tag{1}$$

By analogy with the calculation of the similarity of words, it is to calculate the semantic proximity of the competencies and requirements, which are short statements with 10 words on average. The calculation of the vector of described sentences (s), where s = {w_1, w_2, …, w_k}, is defined as a weighted average of vectors of the constituent words:

$$\vec{v}(s) = \frac{\sum_{i=1}^{k} p_i * \vec{v}(w_i)}{\sum_{i=1}^{k} p_i}, \tag{2}$$

where p_i is the weight of a word, which is calculated as the ratio of the frequency of use of the word to the dimension of the lexicon of the selected level of the hierarchy on the side of the education system or labour market, k is the number of words in a sentence.

Then it is calculated the semantic proximity of the sentences using the formula given above. It is worth noting that words that have no particular meaning (conjunctions, particles, prepositions, pronouns, and so on) do not participate in the formation of the vector for the sentence.

3. Monitoring system for the labour market

The aim of implementing the information systems for monitoring and forecasting the situation on the labour market and analysis of staffing requirements is to provide additional opportunities to identify qualitative and quantitative relationships between education and the labour market.

The system is developed for a wide range of users and is intended primarily for heads of regions, universities, companies, recruitment agencies. It is expected that the project will provide a tighter link between the educational system in the country and the labour market, will give the opportunity to adjust curricula, to open new educational programs or to adjust the existing ones in accordance with the economic objectives of the regions, to implement efficient recruitment and training. After that, it is assumed that the system will become a useful tool for young professionals, starting seeking for a job in their chosen profession, and also the persons trying to choose their professionalization.

As a data source on vacancies, the following internet resources are used: portal "Work in Russia" (information website of the Russian labour agency), portals of staffing companies HeadHunter and SuperJob. As the governing documents, the registry of approved professional standards and Federal state educational standards of higher education are used [13]. The subject of a separate study is evaluating how complete the job ads represent the market's real demands.

The automated information system's implemented prototype is a web-oriented application with an intuitive user interface, ensuring reliable data storage.

The system is built on a modular principle. First, it includes the module collecting textual data (operating in automatic mode with the use of open sources, which are Internet portals and recruitment agencies).

Second, the load module and data storage, consisting of a distributed data store (provides replication and archiving).

Third, the automatic processing module performing the preparation of information for analysis, automatic linking requirements and competencies, and machine learning.

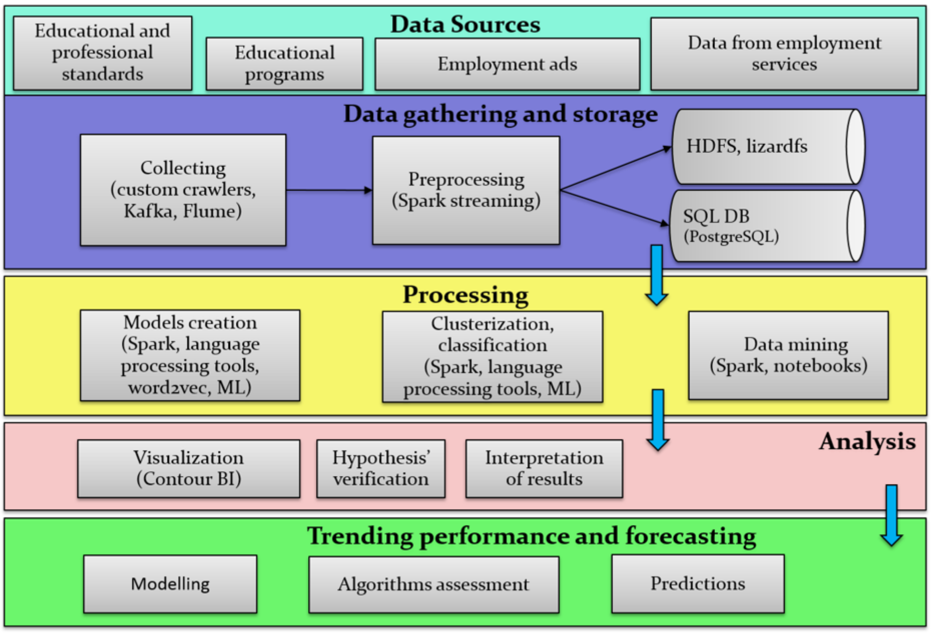
Data Sources
Educational and professional standards
Educational programs
Employment ads
Data from employment services
Data gathering and storage
Collecting (custom crawlers, Kafka, Flume)
Preprocessing (Spark streaming)
HDFS, lizardfs
SQL DB (PostgreSQL)
Processing
Models creation (Spark, language processing tools, word2vec, ML)
Clusterization, classification (Spark, language processing tools, ML)
Data mining (Spark, notebooks)
Analysis
Visualization (Contour BI)
Hypothesis' verification
Interpretation of results
Trending performance and forecasting
Modelling
Algorithms assessment
Predictions

Fig. 2. Information workflow in the labour market monitoring system

Fourth, user interfaces to generate and display reports based on business intelligence technologies.

The general scheme of data processing is shown in Figure 2.

Most approaches to identify the real needs of the market primarily used surveys among employers and employees. Conducting such surveys requires a certain amount of time and resources and cannot provide full coverage of the labour market. To identify both qualitative and quantitative correlations between education and labor market within a reasonable time of the proposed intelligent system of monitoring of the needs of employers and the analysis of their compliance with existing professional and educational standards. Results of this analysis may be recommendations for changes in educational programs.

The framework of the project created a prototype of an automated information system for monitoring and analyzing of employment needs of regions and identifying the relationship to market demand with educational and professional standards. The system included the software and technological solutions to the Situation centre for Russia's socio-economic development and its federal subjects.

With this system the analysis of the constantly updating large amounts of data, it is possible to determine how the training programs of higher education correspond to current market expectations, to anticipate changes in those expectations, and automatically provide recommendations on adjustment of training programs to the most exact conformity to these expectations. Development and adaptation of the system can be carried out according to the customer's requirements depending on the specifics of the task – characteristics of the region, university, etc. We believe that the created system, and the algorithms and principles on which it is based, can be used to solve a wider class of topical challenges. For this, the system can be reconfigured depending on the peculiarities of the task statement and the nature of input data.

Conclusion

Most approaches to identify the real needs of the market primarily used surveys among employers and employees. Conducting such surveys requires a certain amount of time and resources and cannot provide full coverage of the labour market. To identify both qualitative and quantitative correlations between education and labor market within a reasonable time of the proposed intelligent system of monitoring of the needs of employers and the analysis of their compliance with existing professional and educational standards. Results of this analysis may be recommendations for changes in educational programs.

In the framework of the project, it was created a prototype of an automated information system for monitoring and analysis of employment needs of regions and identifying the relationship to market demand with educational and professional standards. The system included the software and technological solutions to the Situation centre for the socio-economic development of Russia and the Federation's subjects.

With this system the analysis of the constantly updating large amounts of data, it is possible to determine how the training programs of higher education correspond to current market expectations, to anticipate changes in those expectations, and automatically provide recommendations on adjustment of training programs to the most exact conformity to these expectations. Development and adaptation of the system can be carried out according to the customer's requirements depending on the specifics of the task – characteristics of the region, university, etc. We believe that the created system, and the algorithms and principles on which it is based on, can be used to solve a wider class of topical challenges. For this, the system can be reconfigured depending on the peculiarities of the task statement and the nature of input data.

Acknowledgment

The work was supported by the Russian Foundation for Basic Research (RFBR), grant 18-07-01359 "Development of the information-analytical system of monitoring and analysis of labour market's needs for graduates of Universities based on Big Data Analytics".

References

1. Dolado J. No Country For Young People? Youth Labour Market Problems in Europe. London: Centre for Economic Policy Research, 2015.
2. Labour Market and Wage Developments in Europe. Annual Review European Commission, 2016. https://doi.org/10.2767/232054.
3. From University to Employment: Higher Education Provision and Labour Market Needs In the Western Balkans. Synthesis Report. European Commission, 2016. https://doi.org/10.2766/48413.
4. Wolf A. Review of Vocational Education: The Wolf Report. UK Department for Education, 2011. Ref: DFE-00031-2011.
5. Zrelov P. Automated system of monitoring and analysis of staffing needs for the nomenclature of specialties of the university. *Federalizm*, 2016;4(84):63-76 (in Russ).
6. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, 2013.
7. Efrati A. Google Gives Search a Refresh. *The Wall Street Journal*. Retrieved July 13, 2012.

8. Garcia E. M., España-Bonet C., Màrquez L. Document-Level Machine Translation with Word Vector Models. Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT), 2015:59-66.

9. Barkan O. Bayesian Neural Word Embedding. arXiv:1603.06571, 2015.

10. Kutuzov A., Kuzmenko E. WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, Springer, Cham, 2016; 661.

11. Le Q., Mikolov T. Distributed Representations of Sentences and Documents. arXiv:1405.4053, 2014.

12. Zrelov P., Petrosyan A., Semenov R., Filozova I., Korenkov V. Monitoring of the labour market needs for university graduates based on data-intensive analytics. Proceedings of the XVIII International Conference DAMID/RCDL'2016, October 11-14, 2016, Ershovo, Moscow Region, Russia.

13. Professional standards in Russia. http://profstandart.rosmintrud.ru.