Scalable Service for Predictive Learning based on the Professional Social Networking Sites

Evgeny Nikulchev
Moscow Technological Institute
Moscow, Russia

Gregory Bubnov
Moscow Technological Institute
Moscow, Russia

Dmitry Ilin
Moscow Technological University
Moscow, Russia

Egor Mateshuk
Moscow Institute of Physics and Technology
Dolgoprudny, Russia

Abstract—**Professional social networking sites are widely used as a tool for obtaining specific information such as technology trends and professional skills demand. The article is aimed to consider the evolution of services for professional communities through integration of analysis of the patent activity, analysis of the academic research activity and analysis of the labour market trends. Authors have developed the prototype of a predictive learning software service which intended to fill the gap between professional social networking sites and e-learning systems, including massive open online course systems. It includes functionality for monitoring of professional skills demand on the labour market and analysis of patents for each corresponding technology. The software service will help to determine demand for professional skills, to actualise an applicant's skillset, to organise professional communities and to build individual learning programs for studying of skills and technologies which are predicted to grow in demand on the labour market.**

Keywords—Online social networks; Social networking sites; Technology life cycle; Predictive learning; Patent activity analysis; Professional skills; Topic detection; LinkedIn; ResearchGate

I. INTRODUCTION

There is a problem of choosing the trend for professional development, which is directly related to career growth. First of all, it concerns specialists in knowledge-intensive areas. Professionals have a number of needs, some of which have been solved to some extent by existing developments. These needs include:

- Self-promotion – creating a profile that represents one's professional skills in the best way.

- Improvement of professional skills – supplementing knowledge with the most advanced and sought-after skills from employers.

- Identification of trends – it is formed on the basis of the need for improvement of professional skills, namely the identification of skills demanded by employers.

While the first two items are well studied, the question of identifying the trends for professional development remains in the background. In addition, existing solutions do not combine the tools that fill all three needs simultaneously.

The problem of self-promotion is solved through social networking sites (SNS). They take a significant part in the life of professional (including scientific) communities [1]. While some SNSs such as Facebook, VK, Twitter are focused on self-presentation, others such as LinkedIn and ResearchGate are focused on self-promotion [2]. There are also less common variations of SNSs known as decentralised SNSs [3] [4]. Their characteristic is the qualitative difference in the audience [5]: users of professional social networking sites (PSNS) are mostly middle-aged people who are interested in building a network of professional relations.

More than 80% of large international companies search for candidates using SNSs, and the majority of them use PSNS, such as LinkedIn [6] [7]. Measurements of applicants' job search effectiveness (92% for professional contacts and 41% for SNSs) were obtained by recruitment company Antal Russia in scope of the research [8]. The research confirms the key role of PSNS. As the research shows [9], PSNS also reduces the amount of false information about professional skills of a person.

This influence increases the quality of information that is used in research as open data. However, it does not solve the difficulties of the natural language processing [10]. For this reason PSNS provides such a tool as definition of the skills and expertise [11] [12]. It is complemented by the confirmation function of the other members of the community, thus ensuring moderation. The same skills and corresponding keywords are often present in unstructured form in the online recruitment agencies (e.g. Indeed.com, HeadHunter). An example of the correspondence between skills and job description is shown in Figure 1.

Fig. 1.   Correspondence between skills and job description

E-Learning systems are used to improve the accessibility of education [13]. Such services of massive open online courses, like Coursera [14], largely solve the problem of self-education. In addition to these, there are learning management systems like Moodle [15]. They are aimed at the learning process itself and the delivery of knowledge, they have built-in elements of social networks, but they do not give an answer to the question: what should be taught to a particular professional?

The skills that a specialist should have include the ability to navigate in the trends in their field of knowledge. Different methods can be used to determine the actual trends:

- Survey of expert opinion.

- The use of trend assessment services (from simple ones, such as Djinni.co, to the most complex ones, for example, Owlin, Quid).

- Technology Life Cycle analysis (TLC) [16].

It is clear that the first option is the most common and least objective. Online services for the trend analysis do not specialise in the skills that appear in PSNSs and in vacancy texts. TLC is based on the analysis of patent activity and appears to be non-trivial for personal use, and the method is not skill-oriented, which limits the possibility of its application.

According to the mentioned problems, the article will consider a possible way of developing services oriented towards predictive learning for professional communities. The extension of ways to use tools such as "skills" with addition of information about innovations coming from scientific environment can open up new prospects in interaction of job seekers and employers.

The concept section describes the idea of the service and provides a list of the addressed issues. The design section provides an overview on the architectural approaches for scaling the development of the service software solution. Also, it provides an overview on the software components that implement the architectural approach. The prototype section provides examples of usage of the predictive learning service. The discussion section describes a list of problems to be solved for further evolution of the service. In addition, this section describes the revealed features of unstructured data in online recruitment agencies. The conclusion section summarises results and provides ideas for integration of the service.

II. THE CONCEPT

The labour market is focused on the practical skills in the context of the interaction between employee and employer. The SNSs have formed instrument specifying the skills and expectations of the parties. However, at this point the concept of "skills" does not involve additional sources of information. Community members often use analytical reports made by recruitment companies and technological reports for analysis of current skills.

It is proposed to develop a service that automatically generates analytical reports on demand for skills in the labour market and on the development of technology based on the analysis of patent activity.

It is noted [16] [17] that the increase in the patent activity leads to the development of technologies in knowledge-intensive areas and forms new professional skill. For example, the development of cloud technologies followed the increase in the number of patents. Currently the configuration of virtual machines in the cloud is a common skill for a system administrator.

Thereby it is reasonable to develop the predictive learning service for PSNS. The service can help to achieve following goals:

Identify the level of demand for a particular skill on the market. It is important for all participants of the market for short-term and strategic planning. It is necessary to take into account the number of vacancies indicating the skill, as well as the patent and research activity.

Refresh a person's skill set in line with the labour market. It is a known fact that knowledge and skills become irrelevant over time. A person interested in finding a new job often gets the task to fill skill gaps. In addition, professionals need continuing education.

Identify the least-filled segments of the labour market. It is obvious that a skilled person should pay some attention to the segments where competition is lower, as it increases the chances of successful employment. The result of this can become an equal distribution of specialists in a professional environment that will undoubtedly have a positive effect on the labour market as a whole.

Determine the market value of skills. Currently, the salary is formed based on expert evaluation of the labour market in most cases. There are cases when a single job offer has a list of requirements which cover multiple job offers. Requirement analysis tools could be introduced in scope of the service to solve the imbalance problem.

Organise professional communities. PSNS specialised in certain industries may organise communities based on the skills, thereby forming the subject of discussion. This will have a positive impact on the environment of professionals through the mutual exchange of experience, which will lead to continuing education.
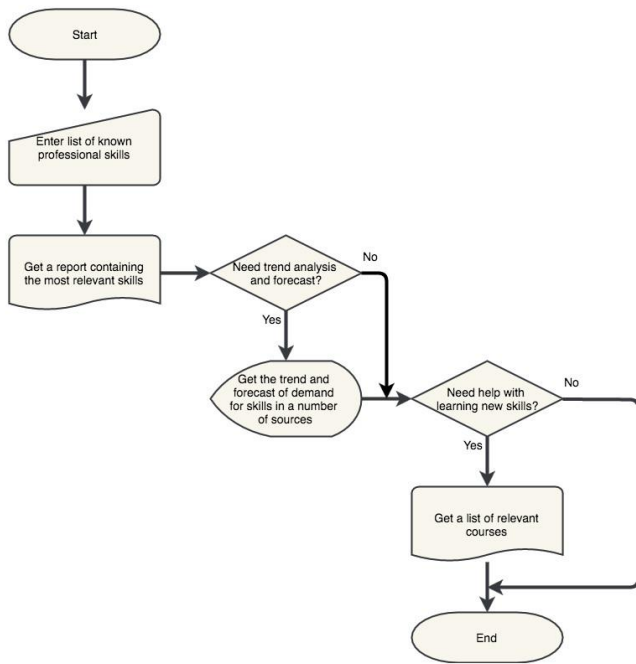
Fig. 2. Flowchart of the predictive learning service usage

Introduce professional standards for skills. This will require an active cooperation with industry leaders and research institutions. Additionally industries may introduce certification programs on key skills. This will improve the quality of training of specialists by giving them more precise boundaries of professional competence.

Consider the service usage algorithm for an applicant, which wants to improve one's skills (Figure 2). At the first stage user specifies a list of known skills with which user going to search for a job. The predictive learning service returns a report with suggested skills based on the content of vacancies. The second stage is optional. User can request charts with a forecast of demand for the suggested skills. If the user considers one or more skills appropriate for studying, user can fetch a selection of the most relevant training courses from PSNS by means of a built-in e-learning system.

III. DESIGN

A. Architecture of the predictive learning service

Consider the main functional components of the service for a PSNS. The service consists of following subsystems:

- Data collection
- Data storage
- Analysis and forecast
- User interface

Data collection subsystem might use, in addition to PSNS data, external data sources. It will improve the reliability of the analysis due to comprehensive monitoring of the Web. The open data sources are the most useful because of the lowest costs of both hardware and human resources.

External sources can be classified according to the provided data:

- Patent activity (e.g. Google Patents, Thomson Reuters)
- Research activity (e.g. Web of Science, Scopus)
- User search activity (e.g. Google Trends)
- Online recruitment agencies (e.g. Indeed, HeadHunter)

Due to the high variability of the data sources there is an issue of data homogenisation. From the perspective of the post-processing, time series are the most suitable for the task of analysing the demand for skills. So, time series should be taken as the basic structure of the stored data. A search of complementary skills requires a unique identifier for each information entry. In the case of online recruitment agencies it could be the URL that uniquely identifies the vacancy. The result is the templates of database entities structure for the first case (Figure 3) and for the second (Figure 4).
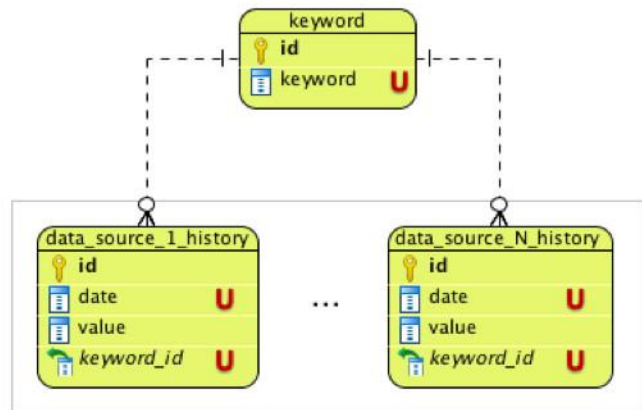

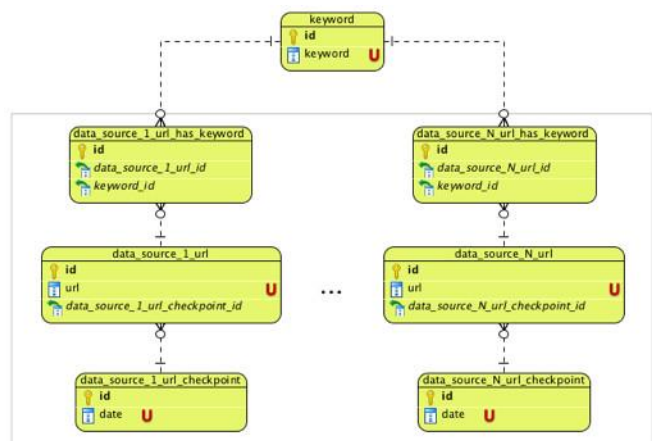Fig. 3. Template of database entities structure for time series

Fig. 4. Template of database entities structure for the search of complementary skills

As stated in the study [18], implementation of an adequate approach to the problem of extension of functionality can improve the efficiency and reliability of the development process. The service needs to be designed taking into account its separation into loosely coupled modules. The components of data collection from external sources need to be built with

the ability to operate independently from the rest of the service modules; data collection can be a resource-intensive process, as the sources can contain data in poorly suitable for processing format, or contain varying amounts of noise. This separation is possible in strict compliance with the principle of single responsibility, not only for the models, but also for the modules.

Inversion of control could be involved to decouple software modules. Among the possible ways of implementation (factory, service locator, dependency injection) dependency injection should be considered as the most applicable to the problem. There are difficulties with the code testing in the case of the factory pattern. The factory methods need to be modified to support unit testing frameworks [19]. Usage of service locator is possible, but implies that all classes should be dependent on the locator. It also negatively affects the code testability. The dependency injection approach has been criticised because of the complexity of the software solution foundation. However, since the foundation is rarely subject to change, in the scope of the service development the problem is considered overrated.

The structure of service modules should be based on the principle of convention over configuration [20]. Thus it is possible to avoid large amounts of duplicated code related to the interaction of system components. Magento 1.9 e-commerce platform is a known example of a system in which such approach could significantly reduce extra efforts. In practice module definitions have identical configurations in most cases; cases of non-standard module configurations are often considered to be examples of the lack of understanding of the principles of Magento platform. This is confirmed by the fact that third-party developers have implemented a plugin for PHPStorm IDE [21] [22], which allow developers to generate default configuration files.

B. Software solutions for the predictive learning service

Applicability of existing software solutions was analysed regarding the problem. Open source solutions were considered. As the service should be embedded into PSNS, user interface should be based on the HTML technology. It also involves the client-server approach for interaction with users.

Two main alternatives of relational database management systems (RDBMS) were reviewed for data storage: MySQL and PostgreSQL. An important prerequisite for the service is the ability to be horizontally scalable. It requires usage of replication. It is worth noting that there is no need for sharding, as the number of external data sources is limited. Comparison MySQL 5.5.31 and PostgreSQL 9.1 demonstrates that the CRUD operation performance with usage of replication significantly higher in PostgreSQL in most of the experiments [23]. In this regard, it was decided to use this particular RDBMS.

The following technologies were considered as the basis for server-side development:

- PHP

- Java (Vaadin framework)

- Python (Django framework)

- JavaScript (Node.JS platform, Express.js framework)

PHP language, although it is the most common tool for the development of server-side components of a web application, is not suitable for the development of components for PSNS. The language does not provide convenient tools for implementation of the daemon services. In addition, performance indicators are relatively low [24], the language is not suitable for building scalable systems.

Java is a suitable tool for the development of scalable services [24], but it has several drawbacks:

- Project compilation takes significant time

- It is verbose, which results in a lower developer's performance

Consideration of Vaadin framework has shown that it is much better suited for internal company systems. Its usage is not advisable for SNSs, due to the lack of full control over the generated web application client-side code.

Python does not have these disadvantages of Java. Consider the framework Django, which is the de facto standard. It supports RDBMS, but does not support NoSQL-storages. This is not an issue at this stage, but involves additional risk to the project. Generation of CRUD interface can be an advantage for developers, but it is a small advantage for the production environment. In addition, the need to use two different programming languages for the client-side and server-side has a negative impact on the developer efficiency.

JavaScript is actively used for client-side development, but with the advent of Node.JS platform it is used for server-side system components as well. It has lower performance compared to Java, yet it is acceptable within the problem scope. Also, there are no disadvantages of Python. Express.js framework is the most common; it does not imply significant limitations to the architecture of software solutions. Thus, combination of JavaScript, Node.JS, and Express.js considered suitable for server-side development of the service.

BottleJS and AngularJS 1.5.x were selected as an addition to the Express.js, which provided support of dependency injection on server- and client-side, respectively. It is important to note the similarity of these frameworks, which positively affects the uniformity of the service code base.

IV. PROTOTYPE

The prototype of predictive learning service for PSNS was developed in scope of the experiment, the collection. It performs homogenisation and analysis of data from 5 external sources. Some of the sources provide data for more than one indicator that reflects the labour market state. Consider the current features of the service.
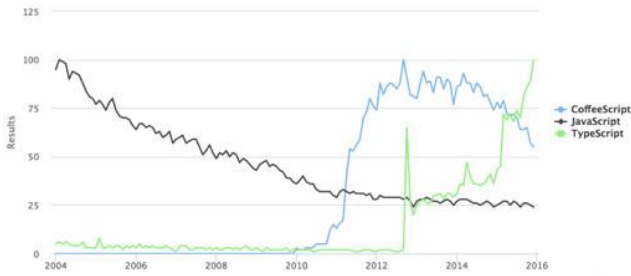
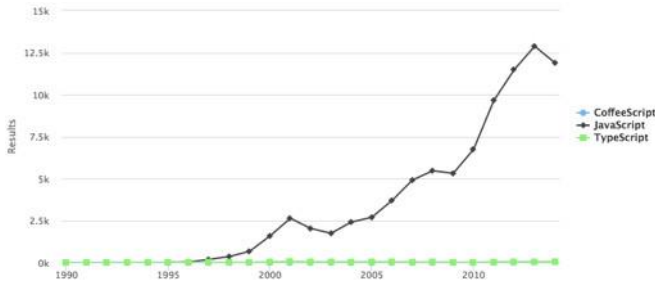Fig. 5.   Relative amount of search requests (Google Trends)

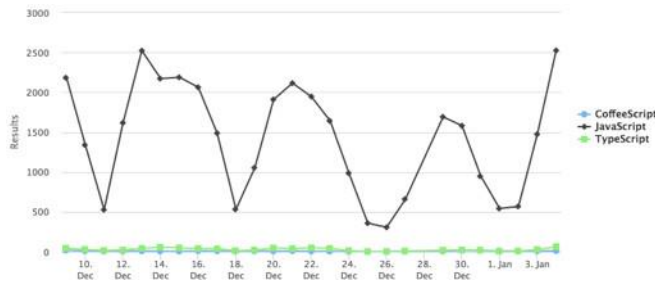Fig. 6.   Annual patent activity (Google Patents)

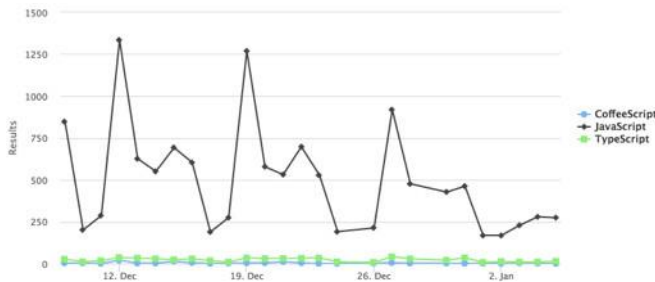
Fig. 7.   Vacancies in the United States (Indeed)

Fig. 8.   Vacancies in Russia (HeadHunter)

Despite the growing interest in new programming languages like CoffeeScript and typescript, they are rarely used in commercial segment. Figure 5 shows the change in popularity of Google searches on each of the languages. Each of the series in the chart is independent of the other. Thus, we cannot assume that users search the information about CoffeeScript more often than about JavaScript. A number of sources provide evidence on low demand for these languages in the commercial sector: patent activity (Figure 6), vacancies in the United States and Russia (Figures 7 and 8 respectively).

The result matches the expectations: since CoffeeScript and TypeScript are compiled to JavaScript, it requires developers to know all these technologies, thereby rising requirements for employees. It is not beneficial in the segment of commercial

software development, since it involves additional risks. However, these languages are in demand in the Open Source community. Atom source code editor, which is developed using CoffeeScript, is a good example.

The service allows tracking the current stage of technology life cycle on the basis of patent activity. Data on several well-known RDBMS can be examined as an example (Figure 9). Also, one can note that the XSLT, a known XML-document processing language, becomes obsolete (Figure 10).
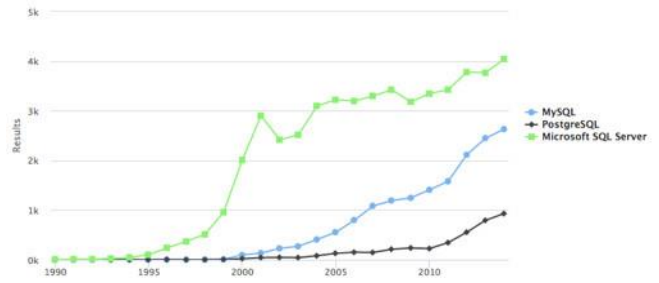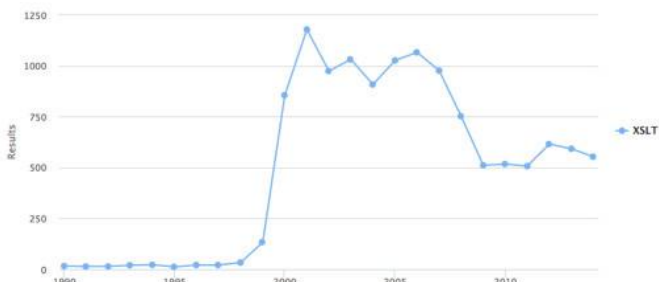

Fig. 9.   Patent activity (RDBMS)

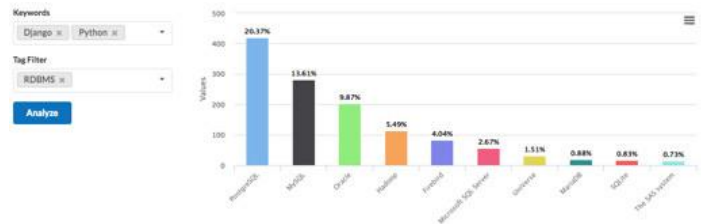Fig. 10. Patent activity (XSLT language)

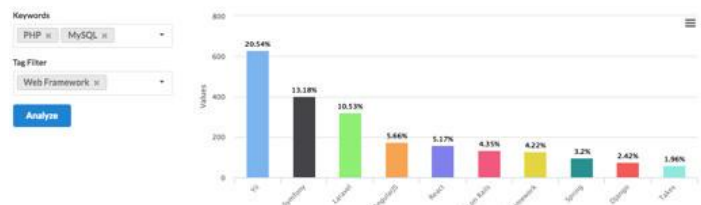Fig. 11. Relevant RDBMS for Python and Django

Fig. 12. Relevant web-frameworks for PHP and MySQL

The predictive learning service solves the problem of selection of the most relevant skills for the studying. Let's assume that the developer knows two main skills: Python and related web-framework Django. In order to increase the developer's own value on the labour market, it is relevant to study one or more RDBMS, with which the developer will have to interact. The outcome of labour market analysis

(Figure 11) shows the relevant results: the most demanded are the RDBMS PostgreSQL and MySQL

A similar study can be performed for PHP and MySQL. In this case, it is reasonable to determine the most relevant Web-frameworks. As it is shown in Figure 12, the most demanded frameworks are Yii, Symfony and Laravel. The presence of frameworks for other languages in the list is due to the fact that companies are looking for developers for a project with a small team (1-2 developers).

V. DISCUSSION

At this stage, the main problem is the lack of a common approach to the naming of professional skills:

- Members of PSNS name their skills with varying degrees of detail

- Skill may have more than a single name (e.g., it may have synonyms)

- The list of skills is not standardised, there may be spelling errors

- The skills are provided in unstructured form, which complicates the processing and reduces the accuracy of the results

An access to the PSNS application programming interface (API) is needed for more detailed analysis of the problem. However, LinkedIn and ResearchGate do not provide an access to the skills API.

An open data feature has been revealed during the experiments. The percentage of noise vacancies is varying for monolingual information retrieval (US recruitment agency – Indeed.com) and reaches unacceptable values. But while using a foreign language professional terms (Russian recruitment agency – HeadHunter) noise levels remain within acceptable limits and have small deviations. At the same time, noise vacancies are defined as vacancies, the description of which do not contain the desired word or contains it in a meaning that does not imply the chosen skill. Accordingly, the relevant vacancies are understood to be vacancies containing a skill in its immediate meaning.

Illustration of this feature is shown in Tables 1 and 2 and in Figures 13 and 14, respectively. The skills were selected based on the following distribution:

- 2 common programming languages (Java, PHP)

- 2 uncommon programming languages (Haskell, Boo)

- 1 obsolescent programming language (Objective-C)

- 2 obsolete programming language (Pascal, Clarion)

- 3 common DBMS of different application areas (MySQL, Microsoft Access, SQLite)

Possible error in the table values – 5 vacancies which does not affect the result. Haskell, Boo and Clarion skills were excluded from Russian online recruitment agency data due to the fact that the number of results was lower than the possible error.

TABLE. I. VACANCY TO NOISE RATIO (US ONLINE RECRUITMENT AGENCY – INDEED.COM)[a]

Skill	Total vacancies	Real vacancies	Noise vacancies
Java	1025	945	80
PHP	1025	897	128
Haskell	373	171	202
Boo	88	2	86
Objective-C	1025	379	646
Pascal	67	37	30
Clarion	412	24	388
MySQL	1025	784	241
Microsoft Access	1025	433	592
SQLite	471	320	151

[a.] Retrieved on October 15, 2016

TABLE. II. VACANCY TO NOISE RATIO (RUSSIAN ONLINE RECRUITMENT AGENCY – HEADHUNTER)[b]

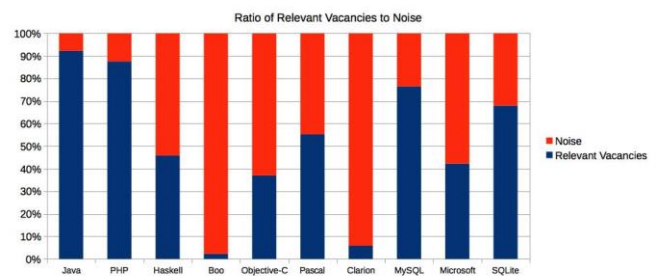| Skill	Total vacancies	Real vacancies	Noise vacancies
Java	484	448	36
PHP	416	386	30
Objective-C	53	50	3
Pascal	8	8	0
MySQL	346	298	48
Microsoft Access	20	18	2
SQLite	14	14	0

[b.] Retrieved on December 6, 2016

Fig. 13. Vacancy to noise ratio (US online recruitment agency – Indeed.com)
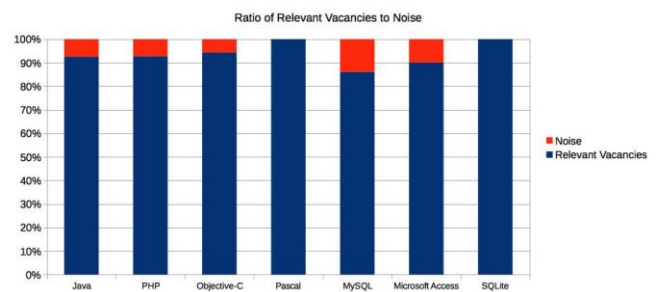

Fig. 14. Vacancy to noise ratio (Russian online recruitment agency – HeadHunter)

This feature can be used when working with multilingual resources to improve the accuracy of results. Data collection from monolingual external sources requires a filter that would reduce the percentage of noise to acceptable values.

Forecasting of the market requires development of a model focused on a given subject area for both short-term and strategic planning. At the moment, a study was conducted to short-term forecasting, in which the original method showed the most accurate result [25].

VI. Conclusion and Future Work

The developed service can be integrated into commercial PSNS. Due to the scalable architecture it can be modified to interact with a larger number of data sources, thereby increasing the value of information to users. In addition, specialised PSNS focused on specific areas of expertise (e.g., IT) can be built based on the idea of this service. Based on report from Bureau of Labour Statistics (US) [26], the average duration of unemployment is slightly higher than half a year. The unemployment period can be used to improve the skills of applicants and the service will allow them to choose a direction for professional growth more effectively.

In addition to integration into PSNS, the concept of predictive learning service can be integrated into educational institutions to improve the quality of the academic plan or into e-learning systems for organisation of user communities encouraging them to share their experience [27]. Furthermore, integration with commercial systems can be monetised not via the premium services, as it does not always increase the conversion [28], but via high-quality targeting of advertising campaigns.

Future work will be devoted to improving the analytical component of the service. Methods of forecasting were previously analysed, but it is also necessary to increase the level of reliability of the collected data. In addition, when integrating with professional social networks, software will be required to support the process of supplementing the skills base, since (as noted in [11]) it can be extremely time-consuming. Since the proposed concept does not imply the free introduction of skills names by users, it will be necessary to investigate alternative solutions for filling the list of professional skills.

References

[1] R. Van Noorden, "Online collaboration: Scientists and the social network.," Nature, vol. 512, no. 7513, pp. 126-129, 2014.

[2] J. van Dijck, "'You have one identity': performing the self on Facebook and LinkedIn," Media, Culture & Society, vol. 35, no. 2, pp. 199-215, 2013.

[3] A. Datta, S. Buchegger, L.-H. Vu, T. Strufe and K. Rzadca, "Decentralized Online Social Networks," in Handbook of Social Network Technologies and Applications, Springer Science & Business Media, 2010, pp. 349-378.

[4] Y. Wang, L. Wei, A. V. Vasilakos and Q. Jin, "Device-to-Device based mobile social networking in proximity (MSNP) on smartphones: Framework, challenges and prototype," Future Generation Computer Systems, 2015.

[5] M. M. Skeels and J. Grudin, "When Social Networks Cross Boundaries: A Case Study of Workplace Use of Facebook and LinkedIn," Group´09, vol. 10, no. 3, pp. 95-103, 2009.

[6] D. Aguado, R. Rico, V. J. Rubio and L. Fernández, "Applicant reactions to social network web use in personnel selection and assessment," Revista de Psicología del Trabajo y de las Organizaciones, vol. 32, no. 3, pp. 183-190, 2016.

[7] J. K. H. Chiang and H. Y. Suen, "Self-presentation and hiring recommendations in online communities: Lessons from LinkedIn," Computers in Human Behavior, vol. 48, pp. 516-524, 2015.

[8] "Issledovanie rynka truda i obsor zarabotnykh plat. Rossia [Research of the labor market and salary review. Russia]," Antal Russia, 2016.

[9] J. Guillory and J. T. Hancock, "The Effect of Linkedin on Deception in Resumes," Cyberpsychology, Behavior, and Social Networking, vol. 15, no. 3, pp. 135-140, 2012.

[10] A. Sapountzi and K. E. Psannis, "Social networking data analysis tools and challenges," Future Generation Computer Systems, 2016.

[11] M. Bastian, M. Hayes, W. Vaughan, S. Shah, P. Skomoroch, H. Kim, S. Uryasev and C. Lloyd, "LinkedIn Skills: Large-Scale Topic Extraction and Inference Mathieu," in Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14, 2014.

[12] D. Nicholas, D. Clark and E. Herman, "ResearchGate: Reputation uncovered," Learned Publishing, vol. 29, no. 3, pp. 173-182, 2016.

[13] P. C. Sun, R. J. Tsai, G. Finger, Y. Y. Chen and D. Yeh, "What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction," Computers and Education, vol. 50, no. 4, pp. 1183-1202, 2008.

[14] J. Knox, "Digital culture clash: "Massive" education in the e-learning and digital cultures MOOC," Distance Education, vol. 35, no. 2, p. 164–177, 2014.

[15] T. Martín-Blas and A. Serrano-Fernández, "The role of new technologies in the learning process: Moodle as a teaching tool in Physics," Computers and Education, vol. 52, no. 1, pp. 35-44, 2009.

[16] H. M. Järvenpää, S. J. Mäkinen and M. Seppänen, "Patent and publishing activity sequence over a technology's life cycle," Technological Forecasting and Social Change, vol. 78, no. 2, pp. 283-293, 2011.

[17] L. Gao, A. L. Porter, J. Wang, S. Fang, X. Zhang, T. Ma, W. Wang and L. Huang, "Technology life cycle analysis method based on patent documents," Technological Forecasting and Social Change, vol. 80, no. 3, pp. 398-407, 2013.

[18] B. Klatt and K. Krogmann, "Software Extension Mechanisms," Fakultt fr Informatik, Karlsruhe, Germany, Interner Bericht, vol. 8, 2008.

[19] M. A. Brown and E. Tapolcsanyi, "Mock Object Patterns," in The 10th Conference on Pattern Languages of Programs, 2003.

[20] M. Bächle and P. Kirchberg, "Ruby on rails," IEEE Software, vol. 24, no. 6, pp. 105-108, 2007.

[21] J. A. Moreno, Aplicación de comercio electrónico para pequeñas y medianas empresas a través de las tecnologías Open Source, Madrid, 2016.

[22] E. Piatti, "Magicento, PhpStorm (PHP IDE) plugin for Magento," [Online]. Available: http://magicento.com/. [Accessed 30 12 2016].

[23] C. O. Truica, F. Radulescu, A. Boicea and I. Bucur, "Performance evaluation for CRUD operations in asynchronously replicated document oriented database," in Proceedings - 2015 20th International Conference on Control Systems and Computer Science, CSCS 2015, 2015.

[24] G. Kotsis and L. Taferner, "Performance Comparison of Web - based Database Access," in Proceedings of DCABES, 2002.

[25] D. Ilin, D. Strunitsyn, M. Fedorov, E. Nikulchev and G. Bubnov, "Development of computer service for analysis of demanded skills in the professional environment," SHS Web of Conferences, vol. 29, no. 02017, 2016.

[26] The employment situation - October 2016, Bureau of Labor Statistics; U.S. Department of Labor, 2016.

[27] I. V. Osipov, A. A. Volinsky, E. V. Nikulchev and A. Y. Prasikova, "Online e-learning application for practicing foreign language skills with native speakers," Technology, Innovation and Education, vol. 2, no. 3, pp. 1-8, 2016.

[28] I. Osipov, E. Nikulchev, A. Volinsky and A. Y. Prasikova, "Study of Gamification Effectiveness in Online e-Learning Systems," International Journal of Advanced Computer Science and Applications, vol. 6, no. 2, pp. 71-77, 2015.