Development of computer service for analysis of demanded skills in the professional environment

Dmitry Ilin[1,*], *Denis* Strunitsyn[1], *Mikhail* Fedorov[2], *Evgeniy* Nikulchev[3,4] and *Gregory* Bubnov [2,3]

[1]Moscow Technological University, 107996, Moscow, Russia
[2]Moscow Institute of Physics and Technology, 141700, Dolgoprudny, Moscow Region, Russia
[3]Moscow Technological Institute, 119334, Moscow, Russia
[4]Moscow State University Technology and Manadgment, 109004, Moscow, Russia

Abstract. At the present time the technology market is constantly growing, new technologies emerge every day so it makes it hard to track all of them and evaluate their potential. As the retraining process takes significant time, the acute need to change the staff training approach from reactive to proactive arises. There are services for monitoring and modeling purposes developed by Quid, Owlin, Djinni. Another approach is the expert analysis. But every of the defined information sources has its drawbacks, which leads to a decision to develop a software and mathematical solution. The article suggests methods of skills demand monitoring for the IT sector, based on comprehensive monitoring of the Web, including employers' demands, company history and publications.

1 Introduction

Besides of expert analysis of technology market demands, which is subjective, there is a number of services for tracking and systematization of data from the Web. They can be useful for determination of technology trends. Examples of such services are:
- Web-platform by Quid;
- Technology Trends Index by Owlin and KPMG;
- Service of salary analysis by Djinni;
- Google Trends.

Consider each of the services to determine the market shares.

The service developed by Quid company indexes and analyzes millions of documents. As the result it offers an ability to visualise reports on topics which usually took months of research. The target audience of Quid are teams of the world largest companies, as well as marketing teams. Thus, the service is focused on the premium market and offers the most advanced functionality.

Technology Trends Index also works with millions of documents, but its scope is limited to 8 subjects.

The service provided by Djinni analyzes average salaries and offers a subscription for a number of preselected majors. Its scope is limited to a short list of IT job positions and data can be retrieved only as an email subscription. It also does not offer forecasting functionality.

Google Trends analyzes search-terms and offers forecasting for some of them. Yet, usage of a single criteria cannot be considered as an objective approach.

Thus, it is expedient to develop a system in order to perform data collection from public sources and to analyze all available factors to forecast expertise relevance rise and fall. Suggested term for that kind of systems is data mining system of relevant skills analysis (DMSRSA).

The described kind of systems can be classified based on its target audience:
- End-user systems, for job applicants;
- Systems for commercial organizations;
- Systems for educational organizations.

These systems can also be classified based on its service delivery method:
- Intranet systems;
- SaaS.

A combination of public sources representing global trends and internal source (person's or company's data) should improve forecasting accuracy and applicability. Such public sources are: the number of scientific paper publications per year (Google Scholar), the number of patent filings per year (Google Patents), data based on search requests (Google Trends), and also the number of available job positions based on data from internet-recruitment services. End-user systems can use an approach which includes their skills as the starting point to evaluate all related skills. It will improve the value of the recommendations because one can study related skills with less efforts.

* Corresponding author: i@dmitryilin.com

The usage of such a solution could provide a number of competitive advantages for organizations, such as:

- Determination of the fastest growing technologies on the market;
- Timely retraining of company's highly qualified personnel;
- Precise choice of the most relevant classes for university studies;
- Improvement of educational programs based on market demands.

End-user systems can provide competitive advantage to job applicants by timely giving them directions to study the most demanded skills in their area of competence.

2 Materials and methods

The hypothesis is that having enough statistical data about a company's staff workload on specific expertise, and also based on the law of large numbers regarding public sources, one can forecast the market demand for a specific category of professionals with high precision.

The algorithmic base of the software solution should be built based on the intellectual time series analysis, which is a subsection of data mining.

The software solution has to have a number of features such as: data collection, data storage, and statistics processing. It is beneficial to implement data collection using 2 different approaches: real time data collection and scheduled data collection on daily basis.

Data storage is a trivial task and does not apply any restrictions. It can be implemented using DBMS MySQL.

Data processing is essential for the software solution. Forecasting should be performed at 2 levels: forecasting of each public data source trend and forecasting of internal company data using public sources forecast results as overlay data.

The result of the algorithm execution will be a chart representing trends based on the achieved forecast. Future releases of the program will include an expert system.

It has been decided to use a company's statistical data as the main data source for forecasting. The forecast will be based on the dynamics of the quantitative measure of expertise involved in a company's projects.

However, usage of a single data source could not be considered as fully representational. It was decided to use public sources as well.

Usage of global market information, as well as local market information, should lead to more precise forecasting results.

It is reasonable to assume that the number of references to a specified competence in scientific publications for a selected time span should be in direct proportion with this competence's relevance.

The choice is: Google Patents, Google Scholar, Google Trends, HeadHunter, and Indeed.

Most of these services do not offer suitable API. Retrievable data are not ready for direct processing because responses are provided as semi-structured data. Thus, every service requires preprocessing of different data structures. The most solid way is to cache all retrieved data in a relational database prior to any kind of processing.

Consider each of the services in details.

Google Trends is a public web-service, based on Google search, and it shows popularity index of words in search requests. Trend dynamics can be requested for a specified time span and it includes extra geolocation data. Retrievable data consist of JavaScript code elements, which can be parsed and converted to JSON-object. After that it can be converted to the internal application representation for further processing.

Google Patents is a search engine indexing patents and patent filings from varying sources. Optical character recognition has been applied to the oldest patents. All foreign patents are translated to English using Google Translate to be indexed and available in search results. These facts describe the service as reliable. The information from this service can be retrieved in JSON format.

Google Scholar is a scientific publication search engine, which works with all available categories of publications and disciplines. It has been online since November 2004. It indexes most peer-reviewed online journals of Europe and America's largest scholarly publishers. Google Scholar indexes public scientific articles as well as papers published in commercial journals. As the service does not provide API, the only possible way to retrieve data is HTML-page analysis with parsing of specific DOM-tree nodes.

HeadHunter – a Russian internet-recruitment company. Its website contains relevant vacancies for numerous occupations. Indeed Inc. – an American internet-recruitment company. Both services offer APIs, which receive GET-requests and response with a semi-structured data set in XML format.

Linear regression and SMO regression from Weka library (implemented within the forecasting plugin) are being used in the software solution as forecasting methods, as well as the autoregressive model method. The library is licensed under GNU GPL [1]. Another method which is being used is the original method described by [2].

3 Experimental research

Solution market analysis has shown that there is a wide number of software solutions with comparable functionality to collect, analyze and visualize forecast results. The closest system class in the applied sense is the marketing information system class (MkIS). Its main aim is to identify, measure and forecast marketing [3–5]. This aim is close to trend forecasting in the IT sector. Yet, most of the software solutions available on the market are a combination of CRM and PRM systems with an addition of campaign success evaluation, so the market does not have any application software for the defined problem. Thus, it can be considered as an additional confirmation of the expediency of implementation.

Google Trends service offers information in a convenient format so it can be easily aggregated. It has its own data visualization. It offers information from 2007 to the current year, but statistics for the current year are incomplete, so the data for this year should not be included into calculations.

Google Patents offers information on a number of patents found using a specified keyword. Patent filings are considered as the most relevant value and it follows aggregation of search result count. As an advantage it offers a wider date range than Google Trends but it also has a number of disadvantages. Firstly, the data collection process takes significantly more time because each value can only be obtained by using a specific request. Secondly, the number of search results for 2014 and 2015 years together does not match the sum of results for 2014 and 2015 years retrieved separately. Thirdly, the charts built based on collected data on various competences are, in most cases, identical and differ only by amplitude as it is shown in fig. 1.

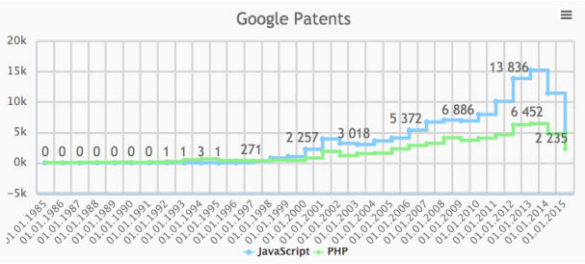

Fig. 1. Example of Google Patents statistics

Google Scholar has mostly the same advantages and disadvantages as Google Patents. Unlike Google Patents, it does not offer search date ranges shorter than a year. As well as this, there is a crawler bot protection, which triggers after several requests. It requires the user to enter a CAPTCHA or to pass a similar test to prove that the requester is not a crawling bot.

The main goal of using HeadHunter is to collect number of available job positions found by a keyword. This service has following disadvantages:

- No functionality to request historical data by a specified date range.
- The data collection mechanism must provide data storage for request results. It requires the user to have a list of keywords before the collecting process.

However, this service provides open access to API and gives precise results. Indeed this service is similar to HeadHunter in most cases. Yet it has a couple of differences:

- API requires the user to be registered.
- Searching using keywords such as LESS could give unnecessary information because it intersects with a commonly used word.

In general, the number of found vacancies in comparison to HeadHunter is ten times larger. Yet, some competences have a local market influence, as shown in fig. 2.

These services can be used as the main data source for an applicant's request to determine which skills are the most relevant to learn. It is implemented using intersection of vacancies for different keywords. For example, for a person with skillset of PHP and MySQL the most relevant skills are HTML, XHTML, CSS, JavaScript and JScript, as it is shown in fig 3.

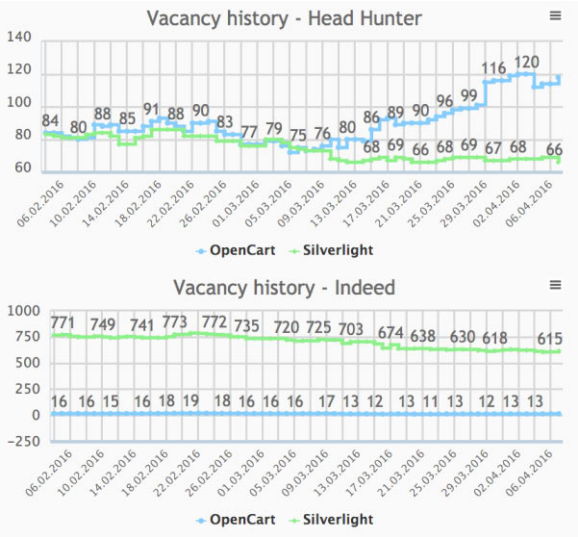

Fig. 2. Comparison of HeadHunter and Indeed statistics

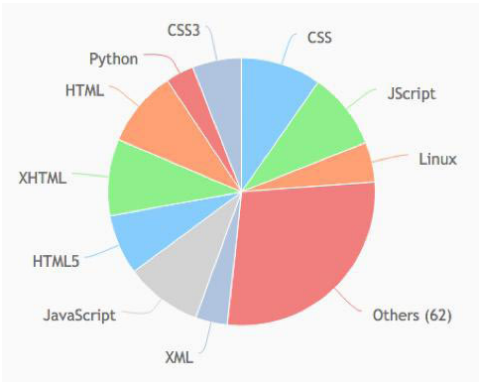

Fig. 3. Relevant skills based on data from HeadHunter

The initial data fetched by keyword "Linux" as a professional competence (fig. 4.) has been processed using the aforementioned methods to simulate a known time span of 5 years. Data for 2014 and 2015 years were excluded from the experiment, because the values were unreliable. The SMO regression method was not applicable to the initial data set, so it was excluded from further comparison. The most important results of the simulation are the differences between actual values and simulated ones. The resulting absolute difference values can be viewed in fig. 5.

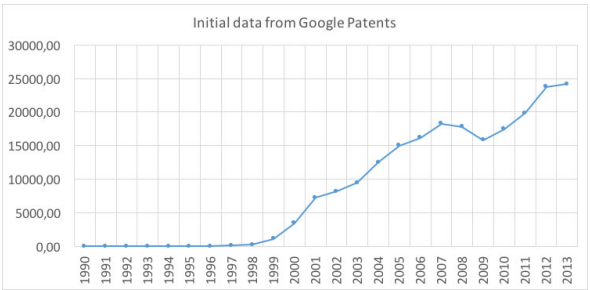

Fig. 4. Initial data

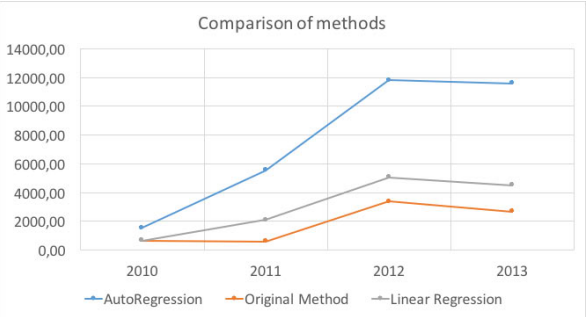

Fig. 5. Simulation results – absolute error values

Calculations of MAPE for the simulation methods are shown in table 1. It's worth mentioning that the original method has the smallest MAPE.

Table 1. Precision validation of suggested methods.

Method	MAPE
Linear Regression	13.57%
AutoRegression	33.61%
Original method	8.06%

However, the results were varying during the experiments with different data sets. Thus, it is better to evaluate an algorithm using specific dataset prior to choosing it for making a forecast.

Based on the results, a method was developed for the monitoring of relevant expertise in the IT sector. It consists of the following steps:

1. Data collection and identification of current knowledge and competences used in a company or a university.
2. Data collection from public sources based on a keyword list.
3. Analysis of the dynamics of trends in the Web.
4. Selection of the most suitable forecasting methods based on model simulation results.
5. Forecasting of local trends, taking into account the dynamics of the Web.
6. If demand for the expertise increases, then there is a need to expand the number of employees working in forecasted competence.
7. If a recession is forecasted, then there is a need for staff cutbacks or retraining.

4 Discussion

It should be assumed that not all organizations collect statistics on their used competencies, which imposes a number of restrictions on the integration of the developed system into business processes.

Precision of different forecasting methods is varying for different time series. According to this fact it is reasonable to consider usage of adaptive selection of the most applicable forecasting method based on automated evaluation of simulation models for each data set.

There is a plan to expand the software to the level of an expert system to reduce the qualification requirements for the user in future [6–7].

The end-user systems might require a relevance index for evaluation of skills. It should include following information for each of skills:

- Data on intersection of the skill with current user skills
- Data on the market share of the demanded skill
- Data on the trend of the demanded skill

5 Conclusions

Various approaches and personal skills are required for effective learning in modern society: for example, ability to make relations and see the sense between spheres of knowledge, concepts and ideas is one of the major skills ensuring effective activity in the modern world. Timely renewal of knowledge is a necessary feature of modern education. Moreover, this is a process of decision-making (ability to choose, analyze, organize, classify, evaluate incoming information), which presupposes high level of development of informative and cognitive independence of students' personality.

References

1. X. Chen, Y. Ye, G. Williams, X. Xu, LNCS, **4819**, 3–14 (2007)
2. V.N .Petrushin, S.A. Drozdov, G.O. Rytikov, Cloud of science, **2**, 247–264 (2015)
3. S. M. S. Freihat, IJRRAS, **11.2**, 326, (2012)
4. M. S. Ezekiel, J. F. Eze, J. A. Anyadighibe, AJTR, **2.2**, 154, (2013)
5. S. Titov, E. Nikulchev, G. Bubnov, Learning Practices as a Tool for Quality Costs Reduction in Construction Projects, Quality-Access to Success. **16**, 68–70 (2015)
6. V.N. Petrushin, E.V. Nikulchev, D.A. Korolev, Histogram Arithmetic under Uncertainty of Probability Density Function, Applied Mathematical Sciences, **9**, 7043–7052 (2015)
7. N.N. Astakhova, L.A. Demidova, E.V. Nikulchev, Forecasting Method for Grouped Time Series with the Use of k-Means Algorithm //Applied Mathematical Sciences, **9**, 4813-4830, (2015)