НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Key skills extraction in labour market for IT sphere

Andrei A. Ternikov, PhD Student

National Research University Higher School of Economics

May 12, 2021

Labor Market Trends

- Digitalization and automation on labor market
- Growing number of vacancies (new professions)
- Skill-sets broadening (combination of skills)

Scientific Side Contribution

- New methodology and algorithms provision for skills analysis
- Model calibration on Russian data

Practical Side Contribution

- Request for renewal of professional standards
- Request for forecasting methods from State labor authorities

Motivation

Research Problem

Rethinking of the educational policy regarding the formation of skills cannot be separated from the demand on the labor market, which needs effective tools for identifying combinations of professional skills that are required by employers.

Research Focus

Current work is concentrated on the creation of the algorithm of key skill-sets, determining the particular occupational group, extraction in the IT sphere.

Research Question

Which skills are needed by companies from different occupational groups of IT sector?

Related Work and Novelty

Theoretical significance

- Introducing the generalized skills extraction algorithm for unstructured database of job advertisements
- Processing and use of large portion of multilingual data in time prospective
- Introducing the approach for skill-sets (combinations) extraction

Practical significance

- Current work provides broader understanding of the demand side of labor market and more evidence for the educational system in order to maintain and renew educational standards
- Algorithm allows to identify and standardize key skills which are applicable for creation of the system of Russian classifiers for occupations and skills

Data Preparation

Initial Data

- HeadHunter IT-sphere vacancies (351,623 observations)
- 13,347 unique frequent skills

Processed Data

- 343,669 vacancies
- 1,730 skills
  - Similar terms (synonyms)
  - Generalized terms (common term for subset of skills)
  - Multi-lingual terms (auto-translated)
- 2-tuple $(id, skills)$ data structure, where id — ID of vacancy and $skills$ — subset of standardized skills

Vacancies' Clustering Algorithm

Supporting Matrices

J (filled with $1 - Jaccard(A, B)$), **C1** (filled with $1 - conf(A \to B)$) and **C2** (filled with $1 - conf(B \to A)$), where A, B — sets of IDs for corresponding $skills$

Algorithm

1. Input: $\mathbf{J}, \mathbf{C1}, \mathbf{C2}$, where i-th row and column attributes to the same skill; threshold t describes the reasonable number of elements in a particular cluster.
2. Hierarchical agglomerative clustering (HAC) procedure with 2-cluster separation is run till the number of elements in the bigger cluster is greater than t.
3. Pairwise aggregation of small clusters with Girvan-Newman algorithm.
4. Iterated reduction of the number of clusters.
5. Matching of non-matched terms. Then, the previous step is repeated.
6. The result of the algorithm is a disjoint set of clusters including all input terms. ■

Main Results and Future Work

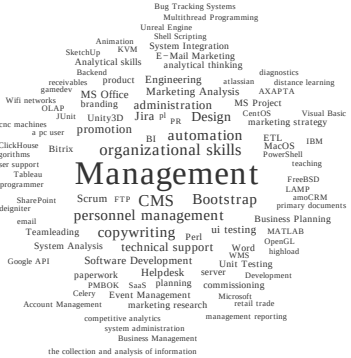
Main Results

- 13 job occupations clusters are obtained
- 16.9% of skills are supposed to be mismatched (negative Silhouette scores)
- Algorithm prevents unnecessary split of clusters into smaller ones and at the same time avoids too much clusters' aggregation
- Detection of relatively same-sized clusters, which could be described by the human
- Novelty: skill-based approach and the use of categorical data structure with unweighted links between categories

Future Work

- Investigation of spheres of Economics & Management; Healthcare
- Use of resumes: finding matching effect

(a) Top-100 well-matched terms.



(b) Top-100 mismatched terms.

Figure: Word clouds of terms and its frequencies separated by Silhouette scores.

#	Name	Size	Top-15 representatives
1	Marketing	110	SEO, Advertising, google, technical website audit, contextual advertising, Internet Marketing, optimization, web analytics, Yandex Direct, yandex, social network, site, SMM, search, project
2	Hardware	184	Linux, Windows, equipment, server administration, configuring dns, network equipment, setting up the pc, software setting, IP, TCP, pc repair, server configuration, Active Directory, Windows Server, maintenance
3	Big Data	57	Python, C++, BigData, Data Analysis, Machine Learning, SCALA, Hadoop, ElasticSearch, Mathematical Statistics, Data Mining, Spark, Mathematical Modeling, Data Science, Kafka, Cassandra
4	Software	467	HTML, JavaScript, CSS, PHP, SQL, Git, MySQL, Java, OOP, jQuery, C#, PostgreSQL, MSSQL, 1C-Bitrix, Framework
5	Administration	96	SAP, C, Unix, Unit Testing, Qt, STL, System Integration, Teamleading, Boost, ABAP, Unreal Engine, ARM, Embedded, teaching, sed
6	Security	52	audit, Information Security, Cisco, security, competitive analytics, antivirus protection network, means of cryptographic protection of information, technical means of information protection, implementation of information systems, domains, Juniper, audit information systems, SIEM, DLP, Check Point
7	Web Design	104	Adobe, Testing, organizational skills, copywriting, Bootstrap, Web Design, writing skills, content, Functional testing, UI, Graphic Design, UX, layout, video, writing
8	Engineering	70	Design, Project Documentation, repair, documentation, Engineering, control, AutoCAD, Visio, gost, automation of processes, process control system, circuitry, programming, circuit design, normative-technical documentation
9	Analytics	83	Excel, technical support, processing, database, powerpoint, ERP, Reporting, paperwork, Financial Analysis, VBA, financial statements, SAP ERP, work with current customer base, analytical studies, primary documents
10	Soft Skills	140	Communication skills, responsibility, result orientation, stress resistance, diligence, Customer Support, care, analytical thinking, Event Management, Bitrix, dedication, punctuality, Game Development, E-Mail Marketing, initiative
11	Management	76	Management, personnel management, administration, ui testing, Recruitment, dbms, BI, Delphi, Business Planning, Web Application Development, Xcode, ExtJS, Strategic Planning, mobile app, personnel evaluation
12	Testing	80	QA, Business Analysis, Selenium, modeling of processes, UML, BPMN, ITIL, Redmine, ITSM, Blockchain, banking software, IDEF, Test Automation, Quality Control, manual testing
13	ERP	211	sales, Project Manager, teamwork, Negotiation skills, English, literacy, Business communication, 1C, customer, Presentation skills, business correspondence, accounting, 1c programming, B2B, the company

Q&A