Big Data and Machine Learning in Government Projects: Expert Evaluation Case

Nikita Nikitinsky[1], Sergey Shashev[2], Polina Kachurina[3], Aleksander Bespalov[4]

[1]NAUMEN, Russia
[2]Tsentr Razrabotki, Russia
[3]DocSourcing, Russia
[4]Saint Petersburg Electrotechnical University "LETI", Russia

Abstract. In this paper, we present the Expert Hub System, which was designed to help governmental structures find the best experts in different areas of expertise for better reviewing of the incoming grant proposals. In order to define the areas of expertise with topic modeling and clustering, and then to relate experts to corresponding areas of expertise and rank them according to their proficiency in certain areas of expertise, the Expert Hub approach uses the data from the Directorate of Science and Technology Programmes. Furthermore, the paper discusses the use of Big Data and Machine Learning in the Russian government project.

Keywords: government project · Big Data · Machine Learning · expert evaluation · clustering

Introduction

Big Data projects for the government sector embody several prerequisites that expert believe are the hallmarks of fast analysis based on effective resources of information. Machine learning helps to build the hierarchy of importance of different parts of this information and gives a possibility to design semi-automated or completely automated services. World practices in this field are diverse. Currently there is a high degree of uncertainty as to which extent it is possible to use automated systems and where only human subjective evaluation works. However even conservative view on the issue allows using Dig Data and machine learning in prior analysis – it reduces the scope of the study area.

Information gathering and evaluation of heterogeneous distributed sources in experience and skills evaluation has previously been a manual process. At the same time the complexity of the operational environment increases due to the increase of labor mobility. Even in classical sociology the studies of social mobility were engaged in comparative inquiries. Nowadays retrospective questions and the use of cohort approach (comparing data with the early stage mobility studies) are not that useful: contemporary society created a new paradigm of existence. In these circumstances, further compara-

tive and longitudinal mobility studies have little point. However fast services for expertise evaluation, especially for collecting data on experience and expertise of professionals who evaluate technological projects seeking for state financing, are of great demand.

The framework outlined above lead to specific methods of research, used in this paper. First, it is a descriptive research on the worldwide experience. Secondly, in order to bring a cross-field study we try to analyze state-of-the-art technology and its use in a narrow sphere of e-government. Thirdly, we try to go in a very detail in description of the Russian Expert Hub system and as a conclusion – to compare it with the best world practices.

Current state-of-the-art technology and projects regarding information collection, fusion and analysis have a clear focus on Big Data and machine learning. The main goal of this article is to study the major international cases of government experience of the use of these technologies – which are a part of e-government, and then to depict the Russian case of expert evaluation. By comparing several clustering algorithms. The main method of our studies is the experimental method. Conclusions should derive from comparisons and be useful for further cases of Big Data and machine learning deployment for government projects. Also there might be a possibility to use this experience in other government projects. Russia is one of the leaders in software development and its market players are interested in fast development and potentially even in the export of technological products and solutions.

Machine Learning and Data Analysis for government projects in Russia

First thing we need to understand is that Data analysis field and e-governance in Russia are phenomena of completely different nature. They intersect during specific cases and amount of such cases is growing but both of them have their own features and specific history. In addition, both of this fields progress rapidly and information sources older than 10 years are almost outdated and have only historical interest.

Official history of Russian e-governance begins in 2000 with Okinawa charter of global information society [1,2,3] which was signed by Russia. The initial position of Russia in these matters was quite weak. In 2003, IT minister Reyman L. stated that only 1% of federal government workers use internet [4].

In 2002 governmental 2.57 billion dollars program “Electronic Russia” (E-Russia) began [2,3], [5,6]. Mostly it was covering the problem of delivering municipal services and information by internet. Results of this program were evaluated very diversely. In 2005 year Putin V. stated that IT market grew from 2% to 5.3% of total GDP, however 40 thousand of localities in Russia have no internet access [8]. Informational and service coverage showed growth from 2000 to 2005 but service coverage was only 6% in 2005 [2]. In 2012 year from 10 basic UN E-Gov objectives Russia targeted only 5 and had some success only in 4 of them [6]. Whole program was widely criticized for ineffectiveness [2,3], [5,7].

Next big governmental attempt in these matters was State programme: "Information Society 2011-2020" which was issued by government in 2010 year [5], [8,9]. Significant growth e-governance services of was stated. Public opinion poll showed that 66% of internet users are ready to user e-gov and according to official statistics 10.6% of Russians have interactions with electronic services at least once[9].

Under governmental patronage large business accelerator, IIDF was founded in 2013[10]. One of its goals was to deliver high quality IT and e-gov services to the people. However, only 7 of 152 successful projects are somehow connected with e-governance [11]. Moreover, very few of IIDF successful projects exploited machine learning or data analysis. In the end of 2015 IIDF representative stated that 500 million of rubles will be invested in big data soon[12].

History of data analysis and machine learning is less dramatic. Yandex Company, which is 4-th largest search engine in the world, started big data analysis trend in the middle of 2000s [13]. In 2007 it opens the school of data analysis[14]. Trend was picked up by many educational institutions ITMO [15], HSE [16], MIPT [17] and so on. It was stated on governmental level, that data analysis and machine learning are development priorities for Russia and country can become competitive in this fields [18]. Data analysis market is quite small ($340 m. in 2014) but its growth rate is almost 40% per year[19]. Main buyers of analytical solutions are banks and telecom. Advertising firms use big data storages most intensively compared to other business directions but absence of world famous successful business stories in this field slows the growth down [20].

Finally, when we reached big data and machine learning in governmental activities and services we can see that government does not use them much by now. At most analytical firms directed to business and education. However, demand for these services is obvious and undeniable but somehow hidden from statistics.

Main problem is that there are three basic levels of governance in Russia:

- State level
- Regional level
- Municipal level

State level has several success stories in developing analytical solutions. All of them can be counted by one hand however, they are quite massive. Total revenue of the leading IT companies in the public sector of Russia in 2013 was $4321 m. This is 77% of their total revenue in Russia [21]. For instance – Federal Pension Fund created analytical services based on SAP HANA, Sberbank launched several complex solutions based on Teradata, Federal Tax service uses various instruments like Teradata, Oracle Exadata and SAP to create analytical layer and monitor tax payers' activities [21], [23]. Some of the state governmental projects are listed in table 1.

Regional level and municipal level are almost completely hidden from view. At analytical companies' sites there can be found proposals of analytical solutions for every level of governmental structure, but very few stats of success histories on regional or municipal levels can be found. It's obvious that some projects require analysis and it is done for them. For instance, in news there can be found that several students created algorithm for optimal car trafficking on municipal toll roads entrances [23]. Obviously,

it is a part of some municipal project, but this project was not tagged with “machine learning”. It is very hard to evaluate real volume of analytical demand in regions. However surely it was rising in recent years [21,22].

Table 1. Recent state ML projects.

Governmental client	Implemented solutions
Sberbank	Marketing and sales, risk management scoring, CRM, anti-fraud
Federal tax service	Establishment of the analytical layer for federal data warehouse
Pension Fund	Analytics and reporting
Federal Compulsory Medical Insurance Fund	Analytics and reporting
Federal Road Agency	Traffic jams forecasting system
Ministry of Finance	Security system, civil service positions classification system
Ministry of Education and Science	Expert-analytical prediction system, automated e-learning resources examination system, financial analytical system
Central Bank	Automated support system for IT departments, real estate analytical system
Supreme Court of Arbitration	E-governance integration, HR system,
Federal Treasury	"Electronic Russia budget" system, security
Roscosmos State Corporation	Computing networks integration and control systems
Federal Service for Hydrometeorology and Environmental Monitoring	Forcasting system update
Ministry of Natural Resources and Environment	Decision Support System
Federal Service for State Registratio Cadastre and Cartography	Automation of real estate registration service, analytics
Federal Financial Monitoring Servic	Automated classification and clustering system
Federal Drug Control Service of Russia	Data storage and analytics

The overall technological progress dictates a shift towards the use of the latest solutions in database management, data processing and automation of prior services. Despite the differences in systems of government administrative entities, the new generation of clerks brought a renewed vision on automation and the use of technology in government projects. This in its turn stimulates emergence of new specific projects and demands. One of such projects – the Expert Hub system, will be presented in the next part of the article.

The Expert Hub System

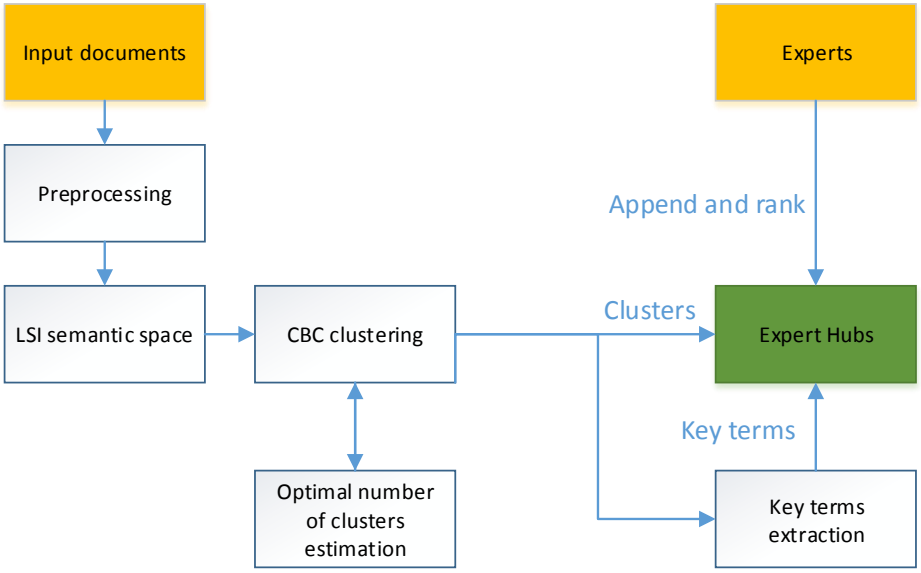

Fig. 1. Schema of the system

Concept. In this part, we will make a general description of the system and then – go into a more detail describing the algorithms that were used, focusing with a special attention on experiments with the algorithms and the way in which optimal variants were chosen.

To increase the implementation speed of innovational solutions in government the Xpir project was created [25]. Its main goal is to provide information support for Russian scientific and technical society. This platform contains science news, conferences information, data on Russian and international funds and organizations. The Expert Hub system prototype was originally created as a module for the Xpir project.

The idea of the Expert Hub approach is to use the documents from Examination System in order to define areas of expertise with semantic space construction and clustering, then to relate experts to the corresponding areas of expertise and rank them according to their proficiency in certain areas of expertise. The Examination System is an internal system in the Directorate of Science and Technology Programmes for evaluating research project proposals. In this system, invited or employed experts are reviewing incoming grant proposals and deciding whether a given research project should be or should not be awarded with a grant or other kind of benefit.

Data. The Directorate of Science and Technology Programmes provided us with 30 000 documents created by 13545 experts for the study.

Data preparation and preprocessing. First, we extract all the so-called metadata from the documents – author names, document titles etc. This data is later used for reference purposes. Then, we conduct the tokenization of the contents, and remove all the punctuation marks as well as the stop-words. We consider words having almost no

meaning, such as prepositions and conjunctions, stop-words. As the final step of data preprocessing, we lemmatize the contents in order to reduce the number of unique terms as different forms for one word by converting them into one conventional form.

Semantic space construction. After preprocessing the input documents, we create LSA term-document semantic space of all documents from data we have, where each row denotes document and each column is a word.

LSA (Latent Semantic Analysis) is a technique for Natural Language Processing, which is widely used for solving various tasks in information retrieval. The underlying idea of LSA is that words with similar sense tend to occur in similar contexts. Thus, this technique can deal with homonymy. We employ LSA as it is faster and can work with larger data sets compared to other approaches [26].

LSA is based on the well-known singular value decomposition technique (SVD):

$$M = U\Sigma V^* \tag{1}$$

where M is m × n matrix whose entries come from some field K, U is m×m matrix, Σ is m×n diagonal matrix with non-negative real numbers on the diagonal and V* is an n×n unitary matrix over K.

We apply Log Entropy weighting function for LSA as this function works well in many practical studies [27].

Particularly, each cell aij of a term-document matrix A is computed as follows:

$$p_{ij} = \frac{tf_{ij}}{gf_i}, \quad g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}, \; a_{ij} = g_i + \log(tf_{ij} + 1) \tag{2}$$

where n is total number of documents, g_i is the global weight, tf_{ij} is the number of occurrences of term i in document j, and gf_i is the total number of times the term i occurs in the corpus.

Semantic space clustering. We cluster the LSA semantic space (currently we use DBSCAN + CBC hybrid clustering algorithm for this task, see Experiments section to know why we used it). As the result of clustering, we obtain centroids and document vectors belonging to those centroids. The optimal number of clusters is computed with Silhouette score using Grid search hyperparameter optimization approach. We then call each cluster an Expert Hub. Since we know, which document belongs to which expert, we may estimate areas of expertise for every expert based on their documents - and the documents are already distributed to clusters.

CBC (Clustering by committee) is a centroid-based clustering algorithm, which was designed with motivation to cluster texts written in natural languages. The algorithm consists of three phases. In Phase I, each element's top-k similar elements are computed for some small value of k. In Phase II, a collection of tight clusters is constructed, using the top-k similar elements from Phase I, where the elements of each cluster form a so-called "committee". The algorithm tries to form as many committees as possible on the condition that each newly formed committee should not be equal or much similar to any already existing committee. All the committees violating this condition are simply discarded from further computing. In the final phase of the algorithm, each element e

is assigned to its most similar cluster or clusters if we apply soft clustering approach [28].

DBSCAN (Density-based spatial clustering of applications with noise) is a density-based clustering algorithm. Given a set of points in some space, it groups together points that are close to each other, marking as outliers points that lie alone in low-density regions. [29].

Silhouette score is an internal clustering validation measure, which is the measure that does not employ any external knowledge about the data e.g. known class labels. It just evaluates the quality of clustering based on the data used for clustering and the result of clustering. Silhouette coefficient compares the average distance from element to element within a cluster with the average distance to elements in other clusters, assigning highest scores to the algorithm producing dense clusters (with high similarity within the cluster) located far from each other (low similarity between clusters). [30].

Grid search hyperparameter optimization approach is a simple approach for selecting the best hyperparameters (e.g. parameters set by a researcher and not learned by algorithm itself) by generating candidate hyperparameters from a grid of possible hyperparameter values specified by a researcher [31].

Key terms extraction. We extract key terms (including n-grams) from LSA semantic space for each Expert Hub based on documents belonging to that Hub. The extracted key terms represent the Expert Hub making it possible for user to name the Expert Hub. In addition, we extract keywords for every area of expertise for every expert for the same purpose.

We experimented with two methods of key term extraction:

1. Computing research area vector for an expert as average vector of his or her documents belonging to the research area. Then, we select top-20 lemma vectors from the whole semantic space, which are similar to the research area vector. These top-20 lemma vectors are selected to represent the research area for the expert. This approach has a feature that among words representing research area for the expert there may occur words not presented in documents of the expert. Caveat of this approach is that we retrieve only unigrams as the semantic space consists of lemma vectors representing single words (bag-of-words approach)
2. We compute research area vector for an expert as average vector of his or her documents belonging to the research area as in first approach. Then we take top-20 lemma vectors, which are most similar to the research area average vector, only from documents of the expert. For the most similar words, we look for n-grams in documents based on rules, which we created. We estimate the LogEntropy weights of a bigram as maximum weight of unigram constituents of a bigram. This approach can retrieve n-grams up to trigrams and consider terms occurring only in this expert' documents.

As bigrams represent areas of expertise better than unigrams, we employ the second approach in our system.

Expert assigning and ranking. We then append experts to the corresponding Expert Hubs and rank them based on their impact weight to the Hub.

Impact weight is computed based on multiple factors:

- scientific background of an expert (from the expert's profile in the Examination System)
- information about the previous expert assessments of an expert
- similarity of the documents of the expert belonging to the certain Hub (the higher similarity to the Hub documents have, the more impact weight the expert obtains).

Since every expert may have multiple different areas of expertise, we apply soft clustering method allowing the experts be related to several Hubs with various impact weight.

Experiments

In this section, we conduct experiments in order to define the best approach to clustering the term-document semantic space built from the documents of the experts. We try three different types of clustering algorithms (i.e. centroid-based, agglomerative hierarchical and density-based) and their combinations.

The aim of the experiments is to select the optimal clustering algorithm or combination of clustering algorithms providing the best clustering results. To measure the quality of clustering, we use Silhouette score. With optimal clustering results, the Expert Hub System should maintain the optimal quality of experts' allocation to the hubs.

In every experiment, for every clustering algorithm, we iteratively select certain number of clusters and on every cluster number we measure Silhouette score. The number of clusters with the highest measured Silhouette score we consider optimal for the clustering algorithm. As we expected the number of Expert Hubs to be from 40 to 120 depending on the possible degree of fragmentation of scientific fields, we conducted iterative clustering on this possible distribution of the clusters. For experimental purposes, we cluster the LSA space constructed from all the 30 000 documents.

Experiment 1. In the first experiment, we clustered the LSA space with just CBC algorithm. The optimal Silhouette score value (0.131) was obtained on 45 clusters (table 2). The keywords extracted from the clusters contained much common lexis and words irrelevant to clusters, which indirectly indicated bad quality of clustering.

Experiment 2. In the second experiment, we clustered the LSA space with just Agglomerative Hierarchical Clustering algorithm (also known as AGNES and AHC).

AGNES (AGglomerative NESting) or AHC is a standard agglomerative hierarchical clustering algorithm, consisting of two phases. Phase I initially starts with n clusters each containing a different element, Phase II embraces the merge of two most similar clusters (repeated n – 1 times) [28].

The results of clustering were better than in previous experiment, however, not much: the highest Silhouette score value was 0.1457 for 43 clusters (table 2). Key terms extracted from the clusters contained common lexis. This makes us to conclude that Agglomerative Hierarchical Clustering is also irrelevant method for clustering in our case.

As we may see from the above experiments, agglomerative hierarchical and centroid algorithms worked not very well on the data. We supposed, that the cause of such results was that the distribution of data points in the LSA semantic space contained much

outliers and the shape of the resulting clusters could be arbitrary. Thus, we decided to try a density-based clustering algorithm DBSCAN as its advantages included robustness to outliers and ability to locate arbitrary-shaped clusters

Experiment 3. In the experiment number three, we first clustered the LSA space with DBSCAN algorithm (with the following hyperparameters: epsilon = 0.5, minPts = 10 and cosine distance function). As this algorithm discovers the appropriate number of clusters by itself and this number of clusters may not fit our predefined possible cluster distribution, we applied CBC to the resulting average vectors of DBSCAN clusters and iteratively measured Silhouette score. The highest Silhouette score value (0.697) was achieved on 42 clusters (table 2). The key terms, which we extracted, contained a lot of special lexis and almost no common lexis. This indirectly indicates a good clustering, i.e., documents with a large number of common lexis were assigned to separate clusters.

As the final trial, we also experimented with applying AHC to the average DBSCAN vectors, but ended up with lower Silhouette score of 0.579 on 46 clusters (table 2).

Conclusion. We consider hybrid clustering algorithm, consisting of DBSCAN and CBC, the most appropriate algorithm to cluster the LSA space constructed from the data. For the data under study, we found out that based on experimental study the optimal number of clusters was between 40 and 46 clusters with the most probable number of 42 clusters.

Table 2. Highest results for clustering algorithms

Clustering algorithm	Optimal number of clusters	Highest Silhouette Score
CBC	45	0.131
AHC	43	0.1457
DBSCAN + AHC	46	0.579
DBSCAN + CBC	42	0.697

Evaluation

To evaluate the Expert Hub System, we employed the expert analysis approach. First, we named the expert hubs with the appropriate names according to the key terms of those hubs (for example, Physical Chemistry, Biology etc.). Then, we selected the top-5 most highly ranked (i.e. relevant to the hub) persons from each of the 42 hubs ending up with the total number of 210 persons. After that, we asked our experts to check some bibliographic databases (i.e. RSCI, Scopus and Web of Science) to make sure that persons indeed should have been related to certain expert hubs. The criterion of the person relevance to the certain expert hub was the following: a person should have had more publications relevant to the topic of the hub he or she was assigned to than to every other topics.

The evaluation showed us, that 83.34% of experts (175 persons) met the criterion. Thus, we can suppose that the Expert Hub System prototype maintains relatively high accuracy in assigning experts to the corresponding hubs.

Results and discussion

In this study, we presented our attempt to create a system to automatically detect and rank experts in certain areas of expertise in order to provide governmental structures with the most highly qualified experts for reviewing incoming grant proposals and research projects. The Expert Hub System prototype operates well – the accuracy of assigning experts to the corresponding expert hubs is above 80%. The clustering algorithm with the best performance on the data we had for the study appeared to be the hybrid DBSCAN + CBC algorithm.

Furthermore, we described prior experiences of technologisation of government services and projects. The basic goals of the previous stage of government services development focused much on automation and storage of information, while nowadays it is possible to work with well-structured data, to shift from database management towards Data Mining, to use Big Data and Machine Learning for sophisticated projects.

Certainly, our study has many limitations. For instance, the evaluation of the system was not strictly formal and we evaluated only some aspects of the system. Moreover, the data we had for this study was relatively small in order to be applied to real processes in the Directorate of Science and Technology Programmes and other governmental structures.

Conclusion and future work

To conclude, we would like to say that the Expert Hub System prototype shows promising results and demonstrated decent performance during the evaluation. Product and market opportunities make the project scalable for other tasks, i.e. for HR solutions or for automated studies of competitors (especially for SME).

For a future study, we suggest:

- Conducting a more thorough and formalized evaluation of the system
- Applying other methods for creating the semantic space from documents in order to obtain better results. Currently, we consider word2vec and similar tools suitable for this.
- Conducting a research in order to better detect the optimal number of clusters, currently we think about applying semi-supervised approach to cluster analysis to handle the task.
- Carrying out usability studies of the system to discover its applicability to other tasks.

Acknowledgements Special thanks goes to the Directorate of Science and Technology Programmes for providing us with the data for this study. Ministry of Education and Science of the Russian Federation supported the research reported in this publication. Unique id of the research project is RFMEFI57914X0091.

References

1. Okinawa charter of global information society, http://www.iis.ru/library/okinawa/charter.en.html
2. McHenry W. and Borisov A. (2006) "E-Government and Democracy in Russia," Communications of the Association for Information Systems: Vol. 17, Article 48.
3. Pardo, Theresa, "Digital Government Implementation: A Comparative Study in USA and Russia" (2010). AMCIS 2010 Proceedings. Paper 330.
4. Reyman, L. (2003) "Information Technologies in the Work of Federal Governmental Agencies" (in Russian), Vestnik Svyazi International, 9, pp. 1-8
5. Vinogradovaa N., Moiseevaa A.,Open Government and "EGovernment" in Russia , Sociology Study, January 2015, Vol. 5, No. 1, 29-38 doi: 10.17265/2159-5526/2015.01.004
6. Bershadskaya, L., Chugunov, A., Trutnev, D. 2012. EGovernment in Russia: Is or Seems? In: Proceedings of the 6th International Conference on Theory and Practice of Electronic Governance (ICEGOV2012, Albany, New York, United States, 22-25 October 2012). Ed. by J.Ramon GilGarcia, Natalie Helbig, Abegboyega Ojo. N.Y.: ACM Press, 2012. 79-82. DOI: 10.1145/2463728.2463747
7. Highlights on meeting of the Presidium of the State Council on the development of information and communication technologies in the Russian Federation, http://www.vestnik-sviazy.ru/news/16-fevralja-2006-goda-v-g-nizhnijj-novgorod-pod-predsedatelstvom-prezidenta-rf-vv-putina-sostojalos-zasedanie-prezidiuma-gossoveta-rf/ (in Russian)
8. State programme: Information Society 2011-2020, http://government.ru/en/docs/3369/
9. Vidiasova L., Chugunov A., Mikhaylova E., E-Governance in Russia: Toward New Models of Democracy, 2015. Proceedings of the 2015 2nd International Conference on Electronic Governance and Open Society: Challenges in Eurasia. ACM, New York, NY, USA, pp. 44-49
10. IIDF homepage, http://www.iidf.ru/ (in Russian)
11. IIDF projects analysis, https://megamozg.ru/post/20382/ (in Russian)
12. IIDF will invest 500 million in Big Data, http://www.vedomosti.ru/technology/articles/2015/11/24/618040-frii-nuzhno-bolshe-dannih (in Russian)
13. From startup to IPO: How Yandex became Russia's search giant, http://www.ewdn.com/2011/05/17/from-startup-to-ipo-how-yandex-became-russias-search-giant/,
14. The Yandex School of Data Analysis , https://yandexdataschool.com/about
15. ITMO extreme computing programme, http://en.ifmo.ru/en/viewjep/2/5/Big_Data_and_Extreme_Computing.htm
16. HSE Machine learning course overview, https://www.hse.ru/data/2015/09/18/1082472447/1Applied%20Machine%20Learning%20-%202015-2016.pdf
17. MIPT github Machine learning course sides, https://github.com/vkantor/MIPT_Data_mining_in_action_2015/tree/master/Slides

18. The order of the Russian Federation Government from November 1, 2013 N 2036-p Moscow, ,http://rg.ru/2013/11/08/texnologii-site-dok.html
19. Structure of the big data market in Russia , http://rusbase.com/howto/big-data-in-russia/ (in Russian)
20. Russian big data in early stages, http://www.computerweekly.com/news/4500259726/Russian-big-data-in-early-stages
21. Kuraeva A., Kazantsev N., Survey on big data analytics in public sector of russian federation, Information Technology and Quantitative Management (ITQM 2015), Procedia Computer Science 55 (2015) 905 – 911
22. Largest Big data projects in Russia, http://www.cnews.ru/tables/a9249186ccefd9e546774ec36da1970ba20ca212/ (in Russian)
23. Big data for governmental sector, http://www.cnews.ru/reviews/ikt_v_gossektore_2014/articles/bolshie_dannye_novye_vozmozhnosti_dlya_gossektora/ (in Russian)
24. Toll roads traffic optimization algorithm, https://www.mos.ru/news/item/7968073 (in Russian)
25. Xpir - platform for communication and cooperation between scientists and entrepreneurs. www.xpir.ru
26. Deerwester S., Dumais S. T., Furnas G. W., Landauer T.K., Harshman R.., "Indexing by Latent Semantic Analysis", Journal of the American Society for Information Science, 41 (6), 1990, pp. 391-407.
27. T.Landauer, D.S. McNamara, S.Dennis and W. Kintsch., Handbook of Latent Semantic Analysis. Lawrence Erlbaum Associates, 2007
28. Pantel, Patrick A. Clustering by committee. Diss. University of Alberta, 2003.
29. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; Fayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231
30. Rousseeuw, Peter J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
31. Bergstra, James, and Yoshua Bengio. "Random search for hyper-parameter optimization." The Journal of Machine learning Research 13.1 (2012): 281-305