УДК 519.25;519.248;004.6;005.6 DOI: http://doi.org/10.25728/pu.2019.3.6

# ПОИСК АНОМАЛИЙ В ЗАДАЧЕ ПОВЫШЕНИЯ КАЧЕСТВА ОТКРЫТЫХ ДАННЫХ

М. Ю. Чесноков

**Аннотация.** Отмечено, что в настоящее время растет число проектов открытых данных (ОД) – публикации в свободном доступе информации государственных органов или частных компаний для последующего использования. Один из барьеров в получении выгод от ОД состоит в наличии проблем качества публикуемых данных. Проанализирована указанная проблема, причины ее появления, рассмотрены метрики и стратегии повышения качества ОД, предложена общая стратегия и ее имплементация для случаев наличия временного и категориального контекстов, предполагающие применение методов поиска аномалий.

**Ключевые слова:** открытые данные, качество данных, поиск аномалий.

## ЛИТЕРАТУРА

1. *Janssen, M., Charalabidis, Y., Zuiderwijk, A.* Benefits, adoption barriers and myths of open data and open government // Information systems management. – 2012. – Vol. 29, No. 4. – P. 258–268.
2. *Manyika, J., Chui, M., Farrell, D.* Open data: Unlocking innovation and performance with liquid information. – McKinsey Global Institute. – October 2013. – Vol. 21. – 116 p.
3. *Batini, C., Cappielo, C., Francalanci, C., Maurino, A.* Methodologies for data quality assessment and improvement // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – P. 16.
4. *Chandola, V., Banerjee, A., Kumar, V.* Anomaly detection: A survey // ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – Article No 15.
5. *Волков А. И., Рейнгольд Л. А.* Открытые данные: проблемы и решения // Прикладная информатика. – 2014. – № 3 (51). – С. 5–12. [*Volkov, A., Reingold, L.* Open data: problems and solutions // Journal of Applied Informatics. – 2014. – No. 3 (51). – P. 5–12. (In Russian)]
6. *Umbrich, J., Neumaier, S., Polleres, A.* Quality assessment and evolution of open data portals. – URL: https://aic.ai.wu.ac.at/~polleres/publications/umbr-etal-2015OBD.pdf.
7. *Neumaier, S., Umbrich, J., Polleres, A.* Automated quality assessment of metadata across open data portals // Journal of Data and Information Quality. – 2016. – Vol. 8, No. 1. – Article *No. 2*.
8. *Kučera, J., Chlapek, D., Nečaský, M.* Open government data catalogs: Current approaches and quality perspective // Inter. Conf. on Electronic Government and the Information Systems Perspective EGOVIS/EDEM – Springer-Verlag, Berlin, Heidelberg, 2013. – P. 152–166.
9. *Vetrò, A., Canova, L., Torchiano, M.,* et al. Open data quality measurement framework: Definition and application to Open Government Data // Government Information Quarterly. – 2016. – Vol. 33, No. 2. – P. 325–337.
10. *Wienand, D., Paulheim, H.* Detecting incorrect numerical data in DBpedia // European Semantic Web Conference ESWC 2014. Springer, Cham, 2014. – P. 504–518.
11. *URL:* http://wiki.dbpedia.org/about.
12. *Debattista, J., Lange, C., Auer, S.* A Preliminary Investigation Towards Improving Linked Data Quality Using Distance-Based Outlier Detection // Joint Inter. Semantic Technology Conf. JIST 2016, Springer, Cham, 2016. – P. 116–124.
13. *Chu, X., Ilyas, I. F., Krishnan, S., Wang, J.* Data cleaning: Overview and emerging challenges // Proc. of the 2016 Int. Conf. on Management of Data. – P. 2201–2206.
14. *Бусыгин В. П., Желободько Е. В., Цыплаков А. А.* Микроэкономика – третий уровень: учеб. пособие. – Новосибирск: Изд-во СО РАН. – 2005. – 704 с. [*Busygin, V. P., Zhelobodko, E. V., Tsyplakov, A. A.* Microeconomics. – Novosibirsk: Publishing house SB RAS. – 2005. – 704 p. (In Russian)]
15. *Baumol, W. J.* Welfare Economics and the Theory of the State / The encyclopedia of public choice. – Boston, MA: Springer, 2004. – P. 937–940.

16. *Гринберг Р. С., Рубинштейн А. Я.* Индивидуум & Государство: экономическая дилемма. – М.: Весь Мир, 2014. – 480 с. [*Grinberg, R. S., Rubinshtein, A. Ya*. Individuum & Gosudarstvo: ekonomicheskaya dilemma. – Moscow: Ves' Mir, 2014. – 480 p. (In Russian)]
17. *English, L.* Total information quality management: A complete methodology for IQ management // Dm Review. – 2003. – Vol. 9, No. 3.
18. *Batini, C., Cappielo, C., Francalanci, C., Maurino, A.* A comprehensive data quality methodology for web and structured data // Int. J. of Innovative Computing and Applications. – 2008. – Vol. 1, No. 3. – P. 205–218.
19. *Aggarwal, C. C., Sathe, S.* Outlier Ensembles: An Introduction. – Springer, 2017. – 276 p.
20. *Чесноков М. Ю.* Поиск аномалий во временных рядах на основе ансамблей алгоритмов DBSCAN // Искусственный интеллект и принятие решений. – 2018. – № 1. – С. 99–107. [*Chesnokov, M. Yu.* Poisk anomalii vo vremennykh ryadakh na osnove ansamblei algoritmov DBSCAN // Iskusstvennyi intellekt i prinyatie reshenii. – 2018. – No. 1. – P. 99–107. (In Russian)]
21. *Leys, C., Ley, C., Klein, O.*, et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median // J. of E. Social Psychology. – 2013. – Vol. 49, No. 4. – P. 764–766.
22. *URL:* https://data.mos.ru/.
23. *URL:* https://data.mos.ru/opendata/7704786030-narusheniya-pdd-vyyavlyaemye-s-ispolzovaniem-avtomaticheskoy-sistemy-fotovideofiksatsii-narusheniy-pdd (version 3.30 от 07.05.2018).
24. *URL:* https://data.mos.ru/opendata/7710474791-dannye-vyzovov-pojarnoy-slujby-po-ao-goroda-moskvy (version 1.34 от 8.05.2018).

*Статья представлена к публикации членом редколлегии В. М. Вишневским.*

**Чесноков Михаил Юрьевич** – Московский физико-технический институт (государственный университет), г. Долгопрудный, ✉ mikhail.chesnokov@phystech.edu.

*Поступила в редакцию 17.07.2018, после доработки 24.12.2018.*
*Принята к публикации 6.02.2019.*

## ANOMALY DETECTION FOR OPEN DATA QUALITY IMPROVEMENT

M. Yu. Chesnokov

Moscow Institute of Physics and Technology, Russia, ✉ mikhail.chesnokov@phystech.edu

**Abstract.** It is noted that an increasing number of Open Data (OD) projects are making governmental and corporate data available to public with free access and reuse. One of the barriers of getting benefits from OD is the quality of published data. This problem and its causes are analyzed, metrics and strategies of improvement of the quality of OD are considered, the general strategy using anomaly detection techniques and its' implementation for cases of time and categorical contexts are proposed.

**Keywords**: open data, data quality, anomaly detection.