Scientific Visualization, 2022, volume 14, number 2, pages 62 - 76, DOI: 10.26583/sv.14.2.06

Trend Visualization of Academic Field: Proposed Method and Big Data Review

E.V. Antonov[1,A,B], A.A. Artamonov[2,A], A.V. Rudik[3,A], M.I. Malugin[4,A]

[A] National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russia
[B] Plekhanov Russian University of Economics, Moscow, Russia

[1] ORCID: 0000-0003-1498-9131, eantonov@kaf65.ru
[2] ORCID: 0000-0002-9140-5526, aartamonov@kaf65.ru
[3] ORCID: 0000-0002-1757-1681, avrudik@kaf65.ru
[4] ORCID: 0000-0003-1623-8518, mimalugin@kaf65.ru

Abstract

Research trends analysis is essential for scientists or governments to understand the present and predict the future of the field. Nowadays it is time-consuming to examine papers for following up on the latest trend in specific research interests. In our study, we present the distributed architecture of the system for automated data collection and analysis. Furthermore, we propose the data extraction workflow for collecting data from multiple sources. The interactive dashboard is implemented with a set of different visualizations for tracing research trends. As a practical implementation of the developed system, a research trend analysis of Big Data technologies is carried out. The set of 34062 articles was processed and collected on that topic from 25 selected internet sources. Finally, a review of Big Data technologies is presented using the developed dashboard, and cases of its use are considered. An analysis of the topic on a specific period, country, and the field is shown and discussed. In addition, the authors try to give a perspective of the future development of Big Data field and its association with related fields.

Keywords: trend analysis, visualization, data collection, data mining, big data.

1. Introduction

Recently, there has been a dramatic increase in academic publications number due to scientific community growth, research collaboration, university reward systems, and simplification of the publishing process [1]. Moreover, the information technology field is developing rapidly. Examining papers from a large number of specific journals is no longer satisfying to follow up the latest trends in specific research interest. As a result, systems for searching and obtaining relevant information, as well as recommendation systems, are developing, and new methods that can help scientists understand the present and predict the future of the field are appearing.

Many recent studies have focused on research trends analysis, which can help the government or researchers decide which topic is worth investing in or working on. The literature review shows that bibliometric analysis is a common method used to study characteristics of the academic fields as well as research trend analysis [2, 3, 4, 5]. In this method, scientists use various citation databases. For example, Xu M. et al. [6] analyzes composting using the Web of Science database, Herrera-Franco G. et al. [7] use the Scopus database to analyze research trends in geotourism. However, Singh V. K. et al. [8] compared three citation databases in their study (Web of Science, Scopus, and Dimensions) and showed that there are visible variations in search output, along with differential coverage of subject areas from those databases. Thus, we need to use data from different possible citation

databases to fully receive up-to-date information about the research field in question and its trends.

Collecting and processing data from various sources in manual mode are time-consuming, so there is a need to automate this process [9]. Only a few works in literature demonstrate system architecture that is aimed at automating the process of analyzing research trends. Park S. and Lee M. propose a system that automatically collects research papers and analyzes them in their study [10]. The system consists of three main models: data crawler, data analysis, web interface. As an example, they used the system to analyze research trend in the information security field. As a data source, authors collected research papers (4251 papers in total from 2009 to 2018) from selected information resources using web scraping technology. The data analysis module involves a morpheme analyzer for extracting keywords and NetworkX program package for drawing a network graph. The web interface allows users to examine the results of the analysis using the graph.

We propose a different architecture of the system for automated data collection and analysis. In our previous work [11], we used Apache Airflow platform for distributed data collection from various information sources and an integrated interactive dashboard with a set of different graphs for analysis. In this study, we present the improved system and propose the system as a tool for research trend analysis. Furthermore, we give a review of the Big Data field as an example of the result.

2. Methodology

2.1 System architecture

This study is an extension of the previous work. We have selected a set of software components that enables the implementation of information and analytical support system and have built the architecture. The detailed description is presented in the previous paper [11], and simplified architecture scheme is shown in Fig. 1.

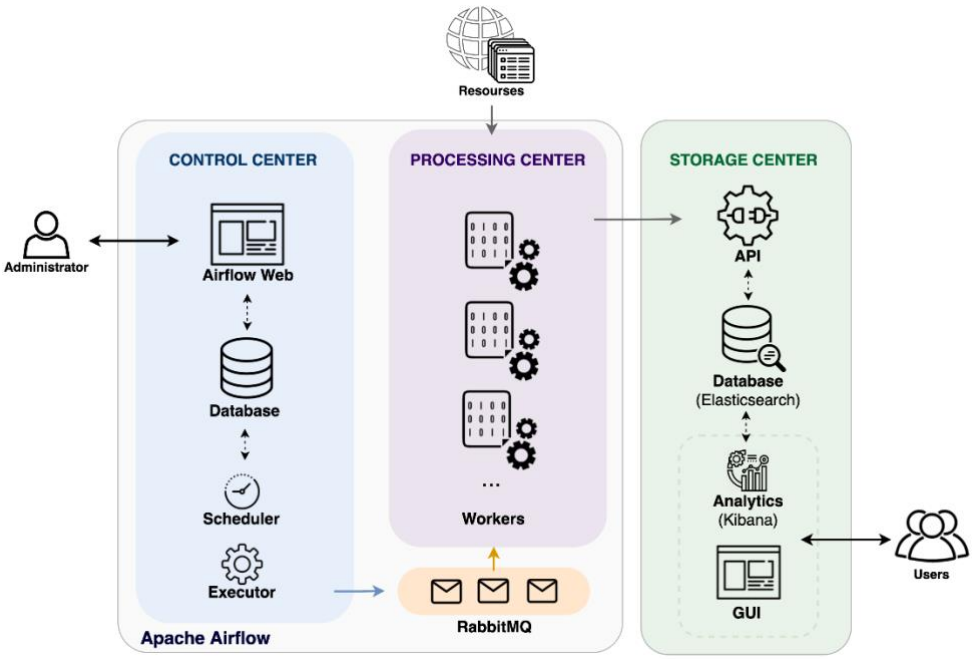


Fig. 1. Information and analysis support system architecture

The architecture is distributed and divided into three main blocks: Control center, Processing center, and Storage center. Control and Processing centers represent Apache Airflow platform. Apache Airflow is an open-source platform for development, planning, and monitoring workflows using the Python programming language. The Control center is applied

for monitoring and planning the whole data collection process, and the Processing center performs data extraction workflow from various web resources using a set of workers. The Control center uses RabbitMQ broker to inform workers to perform data collection tasks. The Storage center contains the Elasticsearch database, Kibana – an instrument for visualizing and analyzing data, developed application programing interface (API) and graphical user interface (GUI) that provide a result of the analysis.

2.2 Data extraction workflow

A workflow on Apache Airflow platform is a directed acyclic graph (DAG) [12]. It represents a Python file and consists of a set of tasks (functions) that are performed in a certain order and cannot be looped. The workflow execution starts with the scheduler that uses the management module to send tasks to the RabbitMQ broker queue, then a worker accepts a task and executes it. The task can be repeated in case of failure and returned to the queue or can be ended in case of success. In the study, we propose the article data extraction workflow from the web resource (see Fig. 2). The workflow can be divided into five main steps.

First, we need to identify new articles that have not been collected. If $\boldsymbol{U}$ is a set of all article URLs that we can locate on the web resource, then a set of new URLs $\boldsymbol{U'} = \{u_1', ..., u_n'\}$ is a proper subset of $\boldsymbol{U}$ and $\boldsymbol{U'} = \boldsymbol{U} / \{u_1, ..., u_k\}$, where $n \in \mathbb{N}_0$ – is a natural number of not collected article, $k \in \mathbb{N}_0$– is a natural number of the collected article, u_i' – is a URL of not collected article, $i \in [1; n]$, u_j – is a URL of the collected article, $j \in [1; k]$. In the first run of the workflow, we have to collect all the existing article URLs in the web resource, in further executions only the last 50-200 URLs can be considered (depending on the web resource).

Then the article data needs to be collected, and it can be done using the web scraping method [13, 14, 15]. In our case, we also use Selenium Webdriver for simulating human Internet behavior. As a result, we extract unstructured information from the web page by the URL u_i' and get structured article data a_i.

The next step is to identify characteristics c_i for each collected article a_i. In the study, we propose a set of five text keywords as a characteristic. Text keywords can be extracted using an unsupervised method based on the statistical features provided by the YAKE software package [16]. Moreover, the library computing a score that determines the level of relevance (the lower the score, the more relevant the keyword is). Therefore, $c_i = \{(k_1, s_1), ..., (k_5, s_5)\}$, where k_l is a keyword from the text, s_l is a score of relevant (the lower the score, the more relevant the keyword is), $l \in [1; 5]$ is the number of keywords sufficient to define the topic, the value can be changed, it is obtained by testing the proposed method.

Next, we need to select articles that are eligible, which means that each article matches the topic. In our work, we analyze research trends in one field (Big Data) as an example and identify eligible articles by the calculated score s_l in a previous step. If the "big data" term is in the set c_i (is in the top five by relevance) then the article a_i is eligible. Consequently, a set of eligible articles $\boldsymbol{D} = \{d_1, ..., d_m\}$ is selected, where $d_t = \{a_t, c_t\}$ – is an article on the topic, and $t \in [1; m]$, where $m \in \mathbb{N}_0 <= i$. However, at this step, any classification or clustering method can be used for assigning articles to one or another topic. For example, Grin D. et al. [17] describe clustering methods in their work and propose the framework that can be used for this step. Shtanko A. and Kulik S. [18] presented intelligent module for data processing and the algorithm that can be considered for the task in their study. Kadhim A. I. [19] describes, compares, and gives examples of different text classification techniques and their implementation that can be applied in this workflow as well.

Finally, articles $\boldsymbol{D}$ are unified, enriched by organizations' geolocation using Yandex Map API (https://developer.tech.yandex.ru/), and are uploaded to a database. In our study, all data is stored in the NoSQL database Elasticsearch.

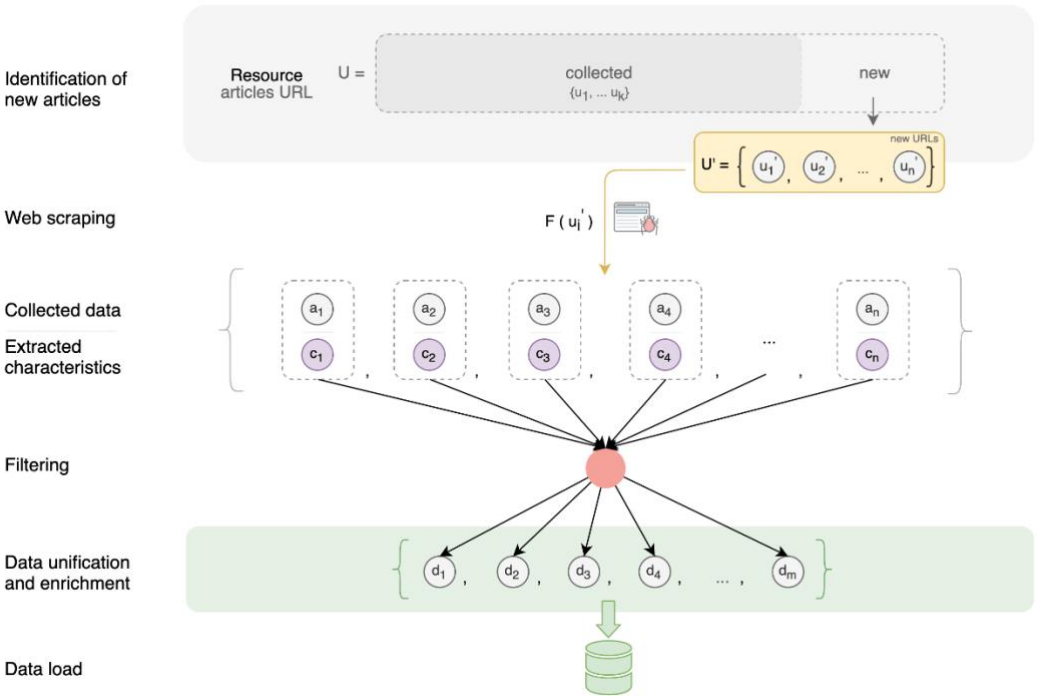


Fig. 2. Data extraction workflow scheme

Data extraction DAG is developed based on the data extraction workflow proposed in the article. The graph view of the workflow in Apache Airflow platform is shown in Fig. 3, it consists of nine separate tasks. The tasks are performed sequentially or in parallel. For example, the task "load_to_the database" depends on two tasks “unify_data” and “get_geolocation” that are executed in parallel, and the "load_to_the database" does not perform until the two previous are completed with success.

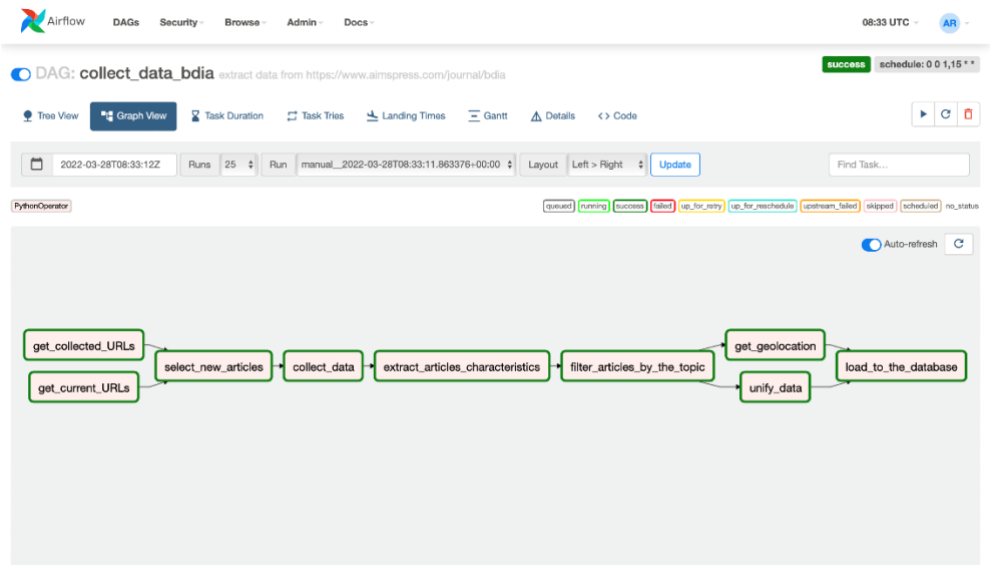


Fig. 3. Data extraction DAG

DAG can be developed to extract data from one or several resources. We develop data extraction DAG for each resource separately. Mainly because we can monitor and service the process of data extraction from each selected web resource and set up a schedule for execution. In our method, only two tasks of nine should be redeveloped for each data extraction DAG, the rest are the same. These two tasks ("get_current_URL" and "collect data") are responsible for web scraping process and depend on web page markup or resource API. DAG schedule is set using cron schedule expression. For example, the DAG in Fig. 3 has

"0 0 1,15 * *" expression and is performed on the 15th and 30th of each month. Thus, you can configure workflows to distribute the execution over time.

2.3 Data analysis tool

In the research, the NoSQL Elasticsearch database is chosen for storing data. We use Kibana which works hand in glove with Elasticsearch and provides data visualization, the possibility of generating dashboards, performing a basic search, and filtration. Kibana is a well-known instrument that is used for various tasks as an analytics tool [20, 21, 22]. A dashboard in Kibana is a collection of metrics, charts, graphs, and maps that are gained on one page. All elements are dynamic, interactive, and adaptable. Kibana is empowered with methods to perform a search on the whole dashboard.

Dashboard for research trend analysis is constructed (see Fig. 4) and made of nine elements: time-series line graph, map, two pie charts, three data tables, and two tag clouds.

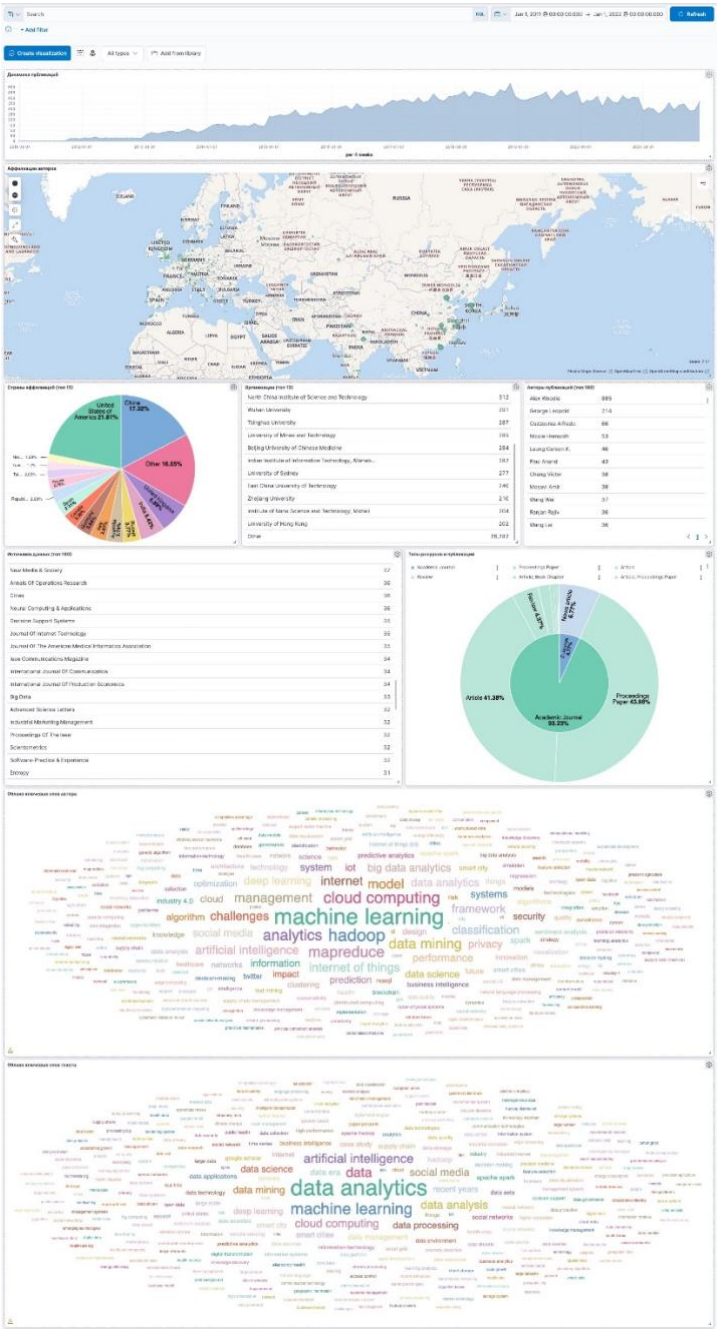


Fig. 4. Dashboard for research trend analysis on the period from 2011 to 2021

The time-series line graph shows the number of publications per 4 weeks. The map represents organizations that are interested in research on the topic, the bigger the point the more articles they publish. One pie chart shows the percentage and the number of countries participating in the research, and the other shows the proportion of different published paper types. The tables represent the number of publications by affiliations, by authors, and by information sources. And tag clouds indicate used words or expressions, the bigger a word/expression on the cloud the more it is used. The first tag cloud shows the keywords that authors emphasized in their study, and the second shows extracted keywords from the article text.

3. Results and discussion

The practical implementation of the developed methodology is carried out in the topic of Big Data technologies. The term "Big Data" has been widely used since the 2010s, though it starts many years before the current buzz [23].

We can use different types of Internet sources within the proposed method.I Information resources such as online academic journals about modern technology are selected in the study. Data are collected by the data extraction workflows from 23 online sources (for example, Datanami, PLoS ONE, Wiley Online Library) and 2 citations databases (Web of Science, Scopus).

In the study, the dashboard is implemented, and the full view is shown in Fig. 4. All the graphs presented below are screenshots and are made using the developed dashboard and its filter.

3.1 Conceptual overview

In total, 74559 articles data were processed, and 34062 of them were selected after the filtering step, according to the results of the data extraction workflows. Besides, 10446 organizations implementing research in the Big Data field were identified, and geolocations were obtained for each organization of collected data. Although the first article in the collected set dates to 2008, just 8 papers are from 2008 to 2011, and the rest (35054) are from 2011 to the present. The number of publications per day (533 articles) reached its highest point on November 15, 2018 (see Fig. 5). Therefore, the analysis is performed mostly from 2011 due to the growth in the number of articles on the topic.

Fig. 5. Time-series line graph of the number of publications per 4 weeks by publications from 2011 to 2021

The global map represents organizations doing research in the Big Data field (see Fig. 6). There are 15 big green points on the map demonstrating organizations that are more involved in research on the topic from the United States of America (2), China (8), India (2), South Korea (1), Australia (1), and Ghana (1). The organizations are presented in Fig. 7b.

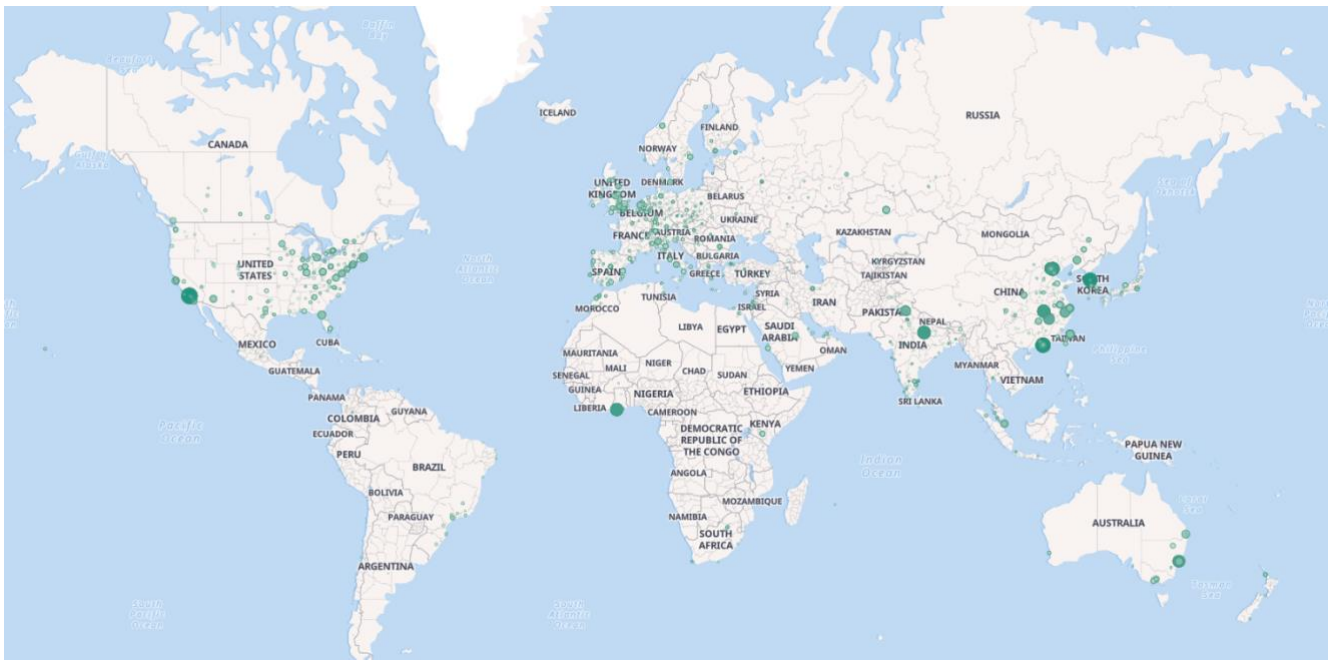
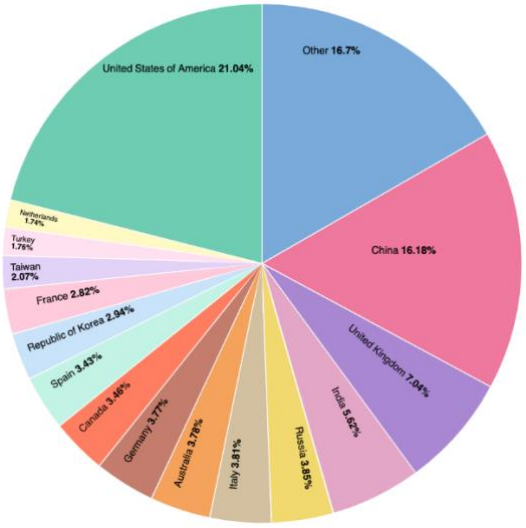


Fig. 6. Organizations doing research in the Big Data field
by publications from 2011 to 2021

However, if we consider the top five countries by percentage of the number of publications (see Fig. 7a), then the United Kingdom (2659 articles or around 7%) and the Russian Federation (1406 articles or around 4%) are also included there. That is, probably, since there is no single research center on the topic in the countries and research is carried out throughout the country by multiple organizations.



a

Аффилиация	Кол-во статей
University of California, Los Angeles	359
Korea University	315
North China Institute of Science and Technology	288
Indian Institute of Information Technology, Allahabad	282
University of Mines and Technology	271
Wuhan University	271
Tsinghua University	263
University of Sydney	263
Beijing University of Chinese Medicine	252
East China University of Technology	224
Institute of Nano Science and Technology, Mohali	202
Zhejiang University	193
University of Hong Kong	185
China University of Technology	173
Massachusetts Institute of Technology	170
Other	26,464

b

Fig. 7. Pie chart of top countries (a) and table of top organizations (b)
by publications from 2011 to 2021

Using implemented tag clouds, we can compare the keywords stated by the authors and the extracted keywords from the article text. On this basis, according to the period from 2011 to 2021 (see Fig. 8), we can conclude that the most used term by the author is "machine learning", and as for the articles' text the term is "data analytics".

a

b

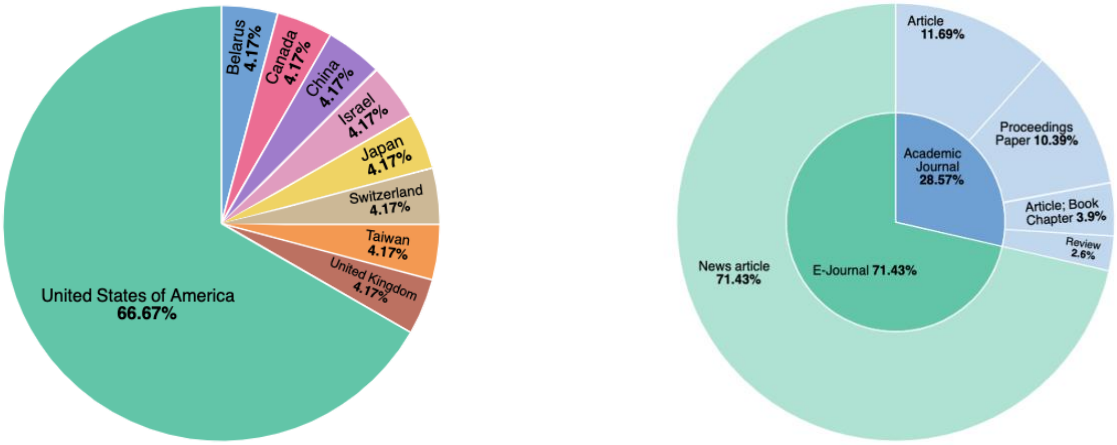
Fig. 8. Tag clouds of keywords state by the author (a) and of extracted keywords from text (b) by publications from 2011 to 2021

Furthermore, we can see that authors set the keywords that are generally about the article major topics, and in extracted keywords from text we can discover where or what tools/methods can be applied.

3.2 Case analysis

A researcher can interactively perform operations on the automated collected data, including adjusting time intervals, setting filtering by authors, organizations, keywords, etc., and applying multiple filters.

The academic field can be analyzed in a specific period. For example, let's analyze the period from 2008 to 2011. There are 77 published papers and 51 of them (around 67%) were published by the United States of America researchers, the studies (55 articles or around 71%) were mostly published in online sources, according to the graphs in Fig. 9. However, with time, more and more studies on the topic have been published in academic journals (see Fig. 4).



a b

Fig. 9. Pie charts of top countries (a) and of resource type (b) by publications from 2008 to 2011

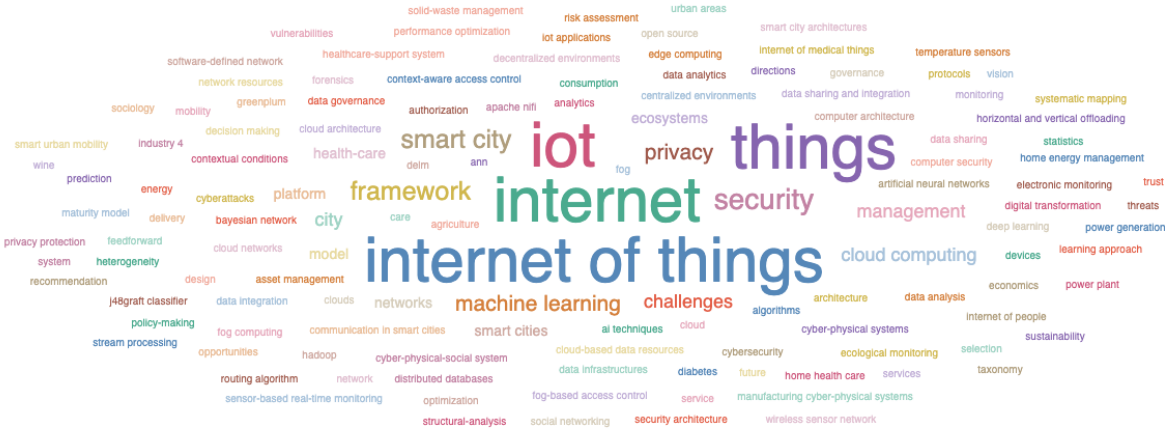
Tags cloud analysis shows that most of the research in this period focused on hardware issues and on companies that granted hardware. For example, we can see keywords like “data intensive” or “high performance computing”, and companies like IBM, Cloudera or Oracle that provide solutions to the issues. All keywords used in publications about Big Data from 2008 to 2011 can be seen in Fig. 10.

a

b

Fig. 10. Tag clouds of keywords state by the author (a) and of extracted keywords from articles’ text (b) by publications from 2011 to 2021

The academic field can be analyzed on a specific topic. For instance, consider the term internet of things (IoT) in tag clouds using the filter (see Fig. 11). There are certain discrepancies associated with the keyword "smart city" in the extracted sample and the author keywords - the term is a major according to keywords from the articles' text. It can mean that authors emphasize many different keywords in the works associated with the IoT, but they are considered a “smart city” discrepancy. Referring to the time-series line graph, it can be revealed that research in the field began in 2013-2014 (see Fig. 12) and there are 2374 articles on the topic in total. Besides, the theme is still relevant for today.



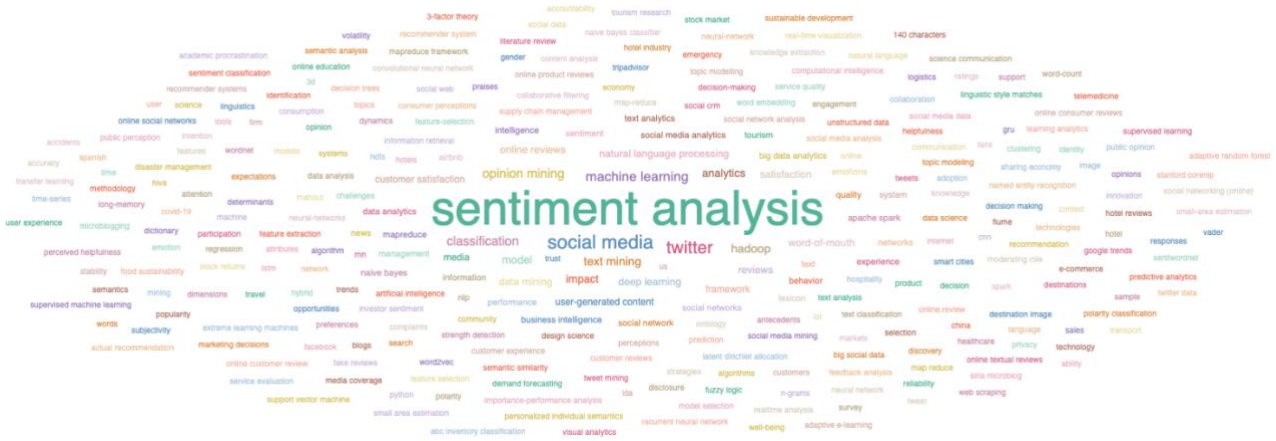
a

b

Fig. 11. Tag clouds of keywords state by the author (a) and of extracted keywords from text (b) by publications from 2011 to 2021 about “internet of things”

Fig. 12. Time-series line graph of the number of publications per 4 weeks by publications from 2011 to 2021 about “smart city”

In certain cases, it is necessary to research the relationship of keywords with each other, namely, to understand how the intra-thematic directions are connected. For example, if we consider research on the theme of "sentiment analysis" (query by the author keywords), we can notice a close connection with "social media" and "twitter", both in the words stated by the author and extracted. The tag clouds are presented in Fig. 13.



a b

Fig. 13. Tag clouds of keywords state by the author (a) and of extracted keywords from text (b) by publications from 2011 to 2021 about “sentiment analysis”

The academic field can be analyzed by specific country or/and by period (using multiple filters). For example, consider the organizations and keyword tag cloud of Indian researchers from 2010 to 2015 (see Fig 14) and from 2016 to 2021 (see Fig 15).

According to the maps (see Fig. 14a, Fig. 15a) we can conclude that two main organizations consider research on the Big Data field in India: Indian Institute of Information Technology (8 articles from 2011 to 2015, and 274 articles from 2016 to 2021) and Institute of Nano Science and Technology (10 articles from 2011 to 2015, and 194 articles from 2016 to 2021). Furthermore, we can see that the number of organizations and publications has increased since 2015. In total, researchers published 117 research papers from 2010 to 2015 by 85 organizations and 3441 papers from 2016 to 2021 by 1036 organizations.

a b

Fig. 14. Map of organizations and keywords tag cloud of India publications from 2010 to 2015

a b

Fig. 15. Map of organizations and keywords tag cloud of India publications from 2016 to 2021

Tag clouds of author keywords demonstrate that the main concerns were about data mining and cloud computing, the Hadoop library, and the MapReduce model from 2010 to 2015 (see Fig. 14b). In the next period (from 2016 to 2021 – see Fig. 15b), the themes are expanding, we see a growth in publications on machine learning, the internet of things, and data analytics.

3.3 Development of the field

Researchers often face the task of following the development in a particular field. By using the collected data, it is possible to consider how the development of the topic took place. Additionally, we created the bar chart race of keywords on Big Data field from 2011 to 2021. The keywords have been unified to create the visualization.

The bar chart race is presented in Fig. 16. A dynamic version of the bar chart race is presented on the published journal site, or it can be found here. It shows the top twenty keywords per year from 2011 to 2021.

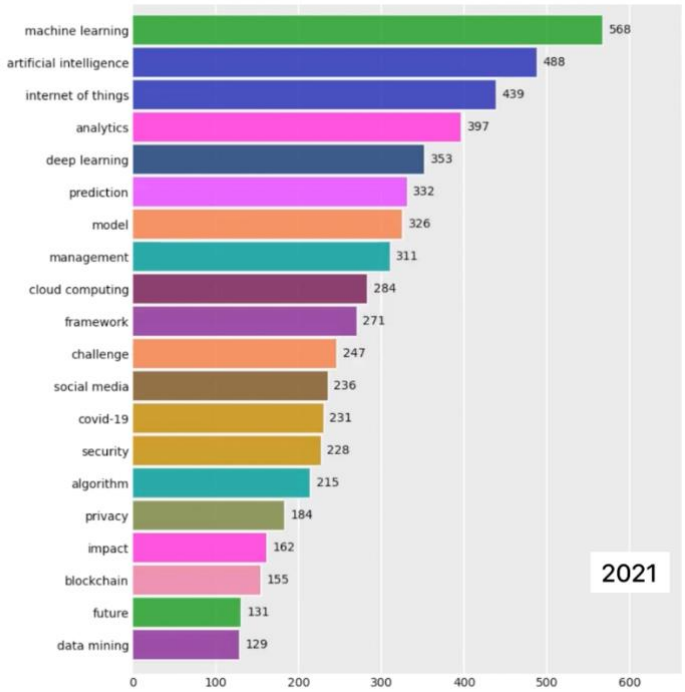


Fig. 16. Bar chart race of keywords on Big Data from 2011 to 2021

The bar chart race reveals that at the origin of the topic, the efforts of researchers were concentrated on hardware issues of working with Big Data technologies, mainly organizing data storages, that causes a high intensity of terms – “Hadoop", “cloud computing”, “storage”.

Since 2016, there has been a change in topic to the development of applied methods of Big Data – “analytics”, “machine learning”, “artificial intelligence”, and the use of these methods for facing issues – “internet of things”, “covid-19”.

The collected data and their visual analytics can help in determining future development. For example, the term "prediction" has been growing since 2018, so the next direction of work may be related to the development of intelligent systems, means of predicting situations and implementing decision support systems. In addition, interdisciplinary researchers have been emerging. For example, we can notice that studies about the interaction of blockchain and Big Data are appearing. Moreover, according to the developed dashboard, we can conclude that these researchers have been working extensively since 2018. For example, Hassani et al. [24] in their work present the term “Big-Crypto” and justify the fusion of Big Data and blockchain technology.

4. Conclusion

To sum up, the study presents an approach to the analysis of a major thematic field. The methods of identifying research trends and competence are proposed. The distributed architecture of the system for automated data collection and analysis, the data extraction workflow, and the interactive dashboard are described and developed. In this paper, research trend analysis of Big Data technologies is carried out. With the help of the developed system and methods, the set of 34062 articles was processed and collected on that topic from 25 selected internet sources. The analysis of the topic on a specific period, country, and theme is shown and discussed.

The developed approach is of interest to various groups of users, with its help it is possible to conduct an examination of applications, and thematic plans, be aware of the development of the industry, implement academic mobility, conduct joint research, etc. The advantage of the developed approach is its operational scalability in various topics. So, within a week, the authors implemented a project to fully analyze the state of the Biomaterials research topic in 2021. The analysis of an array of 1850 papers assisted experts to identify key topics of technological development and determine key research centers.

Based on the implemented approach, the authors plan to develop and implement an adaptive user model for solving the tasks of personalized delivery of thematic pertinent information (content discovery system) and develop a methodology for implementing automated determination of the stages of the technology life cycle.

References

1. Kyvik S., Aksnes D. W. Explaining the increase in publication productivity among academic staff: A generational perspective //Studies in Higher Education. – 2015. – Т. 40. – №. 8. – С. 1438-1453.

2. Kim D., Kim J. Research trend analysis using bibliographic information and citations of cloud computing articles: Application of social network analysis //Journal of intelligence and information systems. – 2014. – Т. 20. – №. 1. – С. 195-211.

3. Lin W., Zhang Z., Peng S. Academic research trend analysis based on big data technology //International Journal of Computational Science and Engineering. – 2019. – Т. 20. – №. 1. – С. 31-39.

4. Kim Y. M., Delen D. Medical informatics research trend analysis: A text mining approach //Health informatics journal. – 2018. – Т. 24. – №. 4. – С. 432-452.

5. Koshlan, D. I., Tretyakov, E. S., Korenkov, V. V., Onykij, B. N., Artamonov, A. A. Agent technology situational express analysis in assessment of technological development level of the BRICS countries //CEUR Workshop Proceedings. – 2018. – Т. 2267. – С. 436-440.

6. Xu M. et al. Research trend analysis of composting based on Web of Science database //Environmental Science and Pollution Research. – 2021. – Т. 28. – №. 42. – С. 59528-59541.

7. Herrera-Franco G., Montalván-Burbano N., Carrión-Mero P., Apolo-Masache B., Jaya-Montalvo M. Research trends in geotourism: A bibliometric analysis using the scopus database //Geosciences. – 2020. – Т. 10. – №. 10.

8. Singh V. K., Singh P., Karmakar M., Leta J., Mayr P. The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis //Scientometrics. – 2021. – Т. 126. – №. 6. – С. 5113-5142.

9. Shtanko A. N., Kulik S. D., Kondakov A. A. Effective scientific personnel training in the field of modern computer technologies for the implementation of advanced research projects of the Megascience class //Journal of Physics: Conference Series. – IOP Publishing, 2020. – Т. 1685. – №. 1.

10. Park S., Lee M. ARTAS: automatic research trend analysis system for information security //Proceedings of the 35th Annual ACM Symposium on Applied Computing. – 2020. – С. 361-368.

11. Onykiy B., Antonov E., Artamonov A., Tretyakov E. Information analysis support for decision-making in scientific and technological development //International Journal of Technology. – 2020. – Т. 11. – №. 6. – С. 1125-1135.

12. Mitchell R., Pottier L., Jacobs S., Ferreira da Silva R., Rynge M., Vahi K., Deelman E. Exploration of workflow management systems emerging features from users perspectives //2019 IEEE International Conference on Big Data (Big Data). – IEEE, 2019. – С. 4537-4544.

13. De Mauro A., Greco M., Grimaldi M., Ritala P. Human resources for Big Data professions: A systematic classification of job roles and required skill sets //Information Processing & Management. – 2018. – Т. 54. – №. 5. – С. 807-817.

14. Antonov, E., Lopatina, E., Ionkina, K., Tretyakov, E. Agent data merging //Procedia Computer Science. – 2020. – Т. 169. – С. 473-478.

15. Ananieva, A., Onykiy, B., Artamonov, A., Ionkina, K., Galin, I., Kshnyakov, D. Thematic thesauruses in agent technologies for scientific and technical information search //Procedia Computer Science. – 2016. – Т. 88. – С. 493-498.

16. Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword extraction from single documents using multiple local features //Information Sciences. – 2020. – Т. 509. – С. 257-289.

17. Grin D., Grigorieva M., Artamonov A. Visual Analysis Application for the Error Messages Clustering Framework //Procedia Computer Science. – 2021. – Т. 190. – С. 274-283.

18. Shtanko A., Kulik S. Increasing the effectiveness of intelligent module by enlarging training dataset from real data //Procedia Computer Science. – 2021. – Т. 190. – С. 712-716.

19. Kadhim A. I. Survey on supervised machine learning techniques for automatic text classification //Artificial Intelligence Review. – 2019. – Т. 52. – №. 1. – С. 273-292.

20. Elmsheuser J., Di Girolamo A. Overview of the ATLAS distributed computing system //EPJ Web of Conferences. – EDP Sciences, 2019. – Т. 214. https://doi.org/10.1051/epjconf/201921403010

21. Shah N., Willick D., Mago V. A framework for social media data analytics using Elasticsearch and Kibana //Wireless networks. – 2018. – С. 1-9.

22. Aulov V. A., Golosova M. V., Grigorieva M. A., Klimentov A. A., Padolski S., Wenaus T. Data Knowledge Base for HENP Scientific Collaborations //Journal of Physics: Conference Series. – IOP Publishing, 2018. – Т. 1085. – №. 3. https://doi.org/10.1088/1742-6596/1085/3/032013

23. Beer D. How should we do the history of Big Data? //Big Data & Society. – 2016. – Т. 3. – №. 1. https://doi.org/10.1177/2053951716646135

24. Hassani H., Huang X., Silva E. Big-crypto: Big Data, blockchain and cryptocurrency //Big Data and Cognitive Computing. – 2018. – Т. 2. – №. 4. – С. 34.