APPROACHES TO KNOWLEDGE EXTRACTION FROM SCIENTIFIC TEXTS

A O Savelyev, Associate Professor
I. B. Soliev, PhD student group A1-06
Tomsk Polytechnic University
E-mail: ibs2@tpu.ru

Introduction

The tendency towards scientific papers is growing annually, which plays the main route in the research fields. Among the most popular used web pages to investigate scientific papers were: Web of Science, Scopus, Springer, ResearchGate according to the scientific factors (citation bases, tags, research management). In addition, the amount of scientific information made available to the public and converted to the big data category makes it difficult to analyze, refine and adjust scientific and technical priorities at the State level.
Thus, the task of developing mathematical and algorithmic tools to extract knowledge from scientific texts to automate the processes of classification and assessment of the significance of scientific texts, identifying the degree of association and mutual influence of promising areas of research and visualization of the structure of scientific activities in order to support decision making in managing scientific activities.

Methodology

The task of extracting structured data from Web pages would be much simpler if there was a single standard for building sites. However, there are no such standards, all the web pages are diverse by the fantasy of web developers. Only thing that unites them is HTML, which defines the appearance of a Web page element, except cannot describe the rest of the contents.

In the overall outlook, extraction of structured related data with web page comes down to successive decision of five objectives [2]:

- Search and receiving of target pages for data extraction;
- Recognition of sections containing the essential data;
- Finding the structure of the data found;
- To ensure the homogeneity of the data to be extracted;
- Combining data from different sources.

Within the framework of this paper, is an examines the peculiarities of using scientometrics and techniques to extract data for further processing and analysis.

Methods of intelligent scientific data analysis include the use of various metrics based on statistical data, heuristic and iterative algorithms, machine learning algorithms, semantic analysis, recognition algorithms - aired arsenal of mathematical algorithms [3].

Peculiarities of resources and metrics to data extraction from web pages

The feature of digital data is that they are initially, inevitably, somehow structured. Systematization of digital data occurs almost simultaneously with accumulation (sometimes natural, short-range solutions are present). Integration and data update are independent tasks, as this is due to duplication of data and the establishment of accessories (authorship) of information. The consequence of these features is the difficulty of obtaining reliable, actual information to extract knowledge from it to meet the information needs of the user. Issues of selection of search engines, databases, search criteria, etc. In general, it is the problem of preserving and extracting true and reliable information about nature and man. At the same time, it is known that the growth of scientific publications retains exponential nature and for scientific work, it is necessary to cover this data stream. Therefore, an improvement in the search using metrics continues to be one of the priority directions of accumulation and structuring of digital data.

To estimate data and search for information systems are used:

- Scientific characteristics in sections of collections of scientific publications to assess the publication activity [5];
- Metrics of advertising potential in social networks to assess user needs and marketing [6];
- Metrics associated with a professional profile in special information business resources.

As one of the examples of a modern approach to preserving and extracting knowledge, the TechOpedia project can be noted, where an array of scientific and commercial publications and projects is accumulated on various sections of information technologies. This project is at the heart of the data systematization of substantive domains.

It should be mentioned that "knowledge spaces", open and closed encyclopedias [7], built on ontology and subject areas of the thesaurus, hardly use metrics of visits and citations. Figure 1 shows some illustrations of popular metrics indicators in collections (databases) of publications. Main indicators: Citation of publication (above-better), H-index of the author (more-better), journal quartile (Q1 - better than Q4).

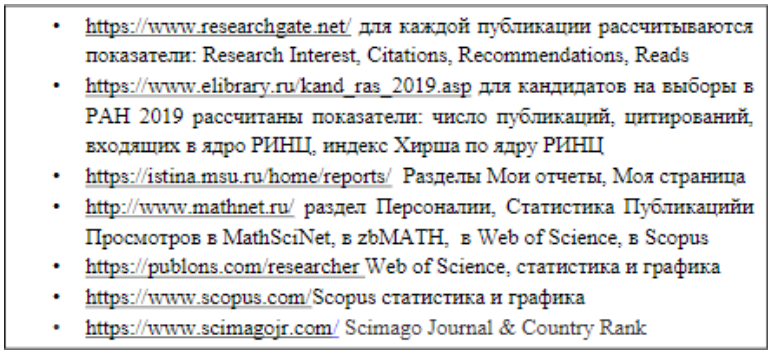

Fig. 1. Examples of scientometric metrics in publications databases

From the point of view of maintaining and retrieving knowledge, it does not matter from which sources information is received, it matters just how the scientific expert community is involved in the process of assessing the quality of the knowledge gained. In this sense, metrics in science and in applications that use artificial intelligence systems are especially interesting [8].

Most of the modern approaches to extract knowledge and their structuring is to optimize the metric, however, if the measurement becomes the goal, then it ceases to be a measurement. This is a remark about the influence that is becoming a goal-based, based on achieving some metric indicators. In each dimension, there is a certain need. Obtaining a quantitative characteristic becomes the goal for a process or phenomenon, when a qualitative characteristic is unclear and subjective, which can be fully attributed to the scientometric metrics.

Based on the structured data and the knowledge extracted from them, you can form text comparison tools to identify authors and subject areas using various measures such as:

- Measure similarity of articles (for different authors).
- Measure of crossing subject areas.
- Measure of the proximity publication of a subject area.
- Measures connectedness of terms, etc.

As a result, you can estimate, a new result was obtained or the old one is rewritten, i.e. apply the listed tools for:

- Preservation of priorities in science;
- Evidence of the reliability of the results and facts;
- Structuring knowledge and subject areas;
- Saving and extracting knowledge.

Also, having a terminological description of the subject area in the form of thesaurus, you can "teach" thesaurus to automatically expand based on of new bonds obtained [9].

Techniques to extracting data from web pages

There are many techniques which are used to extract data from web pages. Web pages mostly contain semi- structured or unstructured data, in which information follows a nested structure. To extract data from web pages, use the following techniques described below and their combinations.

Regular expressions - a regular expression is very common and powerful formal language. It is used to identify string's unstructured text based on some criteria. In this process, making rules manually can be

complicated and require lots of time and efforts. On the opposite hand regular expression based mostly wrappers mechanically and dynamically generate the principles to extract desired knowledge from websites.

Machine Learning - the best data analytical technique to extract domain-specific information from the web sources. Machine learning depends on the training sessions in between, this system achieves a domain expertise. During training sessions, domain experts produce some manually labeled web pages, that are collected from different web pages but also in the same websites [1].

Web Scraping - is the process of extracting data from the websites. With the help of web browsers, web scraping and data extraction techniques can directly access and extract data from the World Wide Web. In the early year's, the data scraping technique that existed was the manual human-copy-paste. But because of the rapid growth and changes on the web, the web page and its data also change dynamically. Hence, because of the dynamic nature of the web world, the traditional method is not feasible. Therefore, automatic techniques are used in the web scraping process.

HTTP Protocol - is the protocol used to transfer data over the web. It is part of the Internet protocol suite and defines commands and services used for transmitting webpage data. HTTP uses a server-client model. A client, for example, may be a home computer, laptop, or mobile device.

Conclusions

The development of the information system will provide information support for such research processes of research, as an analysis, clarification and adjustment of scientific and technical priorities. As part of the study, an analysis of using metrics and techniques for extracting knowledge from web pages was performed. Further developments are aimed at extracting information from scientific publications according to the aforecited metrics and techniques.

References

1. Parvez, M. S., Tasneem, K. S. A., Rajendra, S. S., & Bodke, K. R. (2018). Analysis of Different Web Data Extraction Techniques. 2018 International Conference on Smart City and Emerging Technology, ICSCET 2018. https://doi.org/10.1109/ ICSCET.2018.8537333
2. O.O. Demidova, & A.O. Savelyev. (2018). Comparative analysis of data extraction technician from web pages when solving the task of clustering scientific publications. XIV International Scientific and Practical Conference, November 28-30, 2018
3. Tuchkova, N. P., & Ataeva, O. M. (2020). Approaches to Knowledge Extraction in Scientific Subject Domains. Information and mathematical technologies in science and management, 2 (18).
4. Herrouz A., Khentout C., Djoudi M. Overview of Web Content Mining Tools // The International Journal of Engineering, Overview of Web Content Mining Tools, The International Journal of Engineering and Science (IJES). – June 2013. – No. 6. – P. 106–110.
5. Tsyganov A.V. A brief description of the scientometric indicators based on citation // Management of large systems. Special issue 44: "Scientometrics and expertise in science management». 2013. http://ubs.mtas.ru/archive/search _results_new.php?publication_id=19061. (accessed 15.08.2020).
6. Brodovskaya E.V. Digital citizens, digital society and digital citizenship // Power.2019. T.27. NO4. C.65-69. DOI: https: //doi.org/10.31171/ vlast. v27i4.6587. (accessed 08/15/2020).
7. Ataeva O.M., Sererbryakov V.A., Tuchkova N.P. Query Expansion Method Application for Searching in Mathematical Subject Domains //CEUR Workshop Proceedings, M. Jeusfeld c/o Redaktion Sun SITE, Informatics V, RWTH Aachen (Aachen, Germany). Vol. 2543. Pp. 38-48.
8. The problem with metrics is a big problem for AI. Available at: https://www.fast.ai/2019/09/24/ metrics/ (accessed 15.08.2020)
9. Garfield E. Citation Indexes for Science // Science. 1955. Vol.122. No3159. Pp. 108–111.