

Gathering and Matching Data from the Web: The Bibliographic Data Collection Case Study

Olga Cherednichenko^a, Lubomir Nebesky^b and Marián Kováč^c

Bratislava University of Economics and Management, Furdekova 16, Bratislava, Slovak Republic

Keywords: Information and Communications Technology, Data Collection, Data Consolidation, Matching, Natural Language Processing.

Abstract: As a result of the analysis of existing approaches to consolidating data on research activities we highlight a number of issues. Firstly, the automating the process of data collection which is included the comparing data from different sources. Secondly, the use of external services to obtain bibliographic information which is accompanied by the receipt of erroneous data. The idea of a tracking system for research activity implies that we collect and consolidate data from different web sources and keep them in order to provide relevant bibliographic information. We outline several key points to consider different spellings of the authors' names, data duplication, and filtering out erroneously data. The purpose of the study is to improve the accuracy of comparing bibliographic data from different indexing systems. We propose the framework for gathering and matching bibliographic data from the web. The experimental results show the performance of the proposed algorithm with reaching 0.88 for the F1 metric. The software prototype is developed. The ways to improve the proposed algorithm have been identified, which opens up opportunities for further research.


1 INTRODUCTION


Modern complex organizational systems use a variety of data collection tools to make rational management decisions. However, the existing data collection systems do not fully ensure the relevance of information, which leads to the risks of making wrong decisions and irreversible consequences. This necessitates the study of the problem of creating a distributed system for collecting relevant data on the basis of a unified theoretical and methodological approach by developing appropriate applied information technologies and software products.


The information technology is designed to make human work easier by automating certain processes that need to be performed by humans. As mentioned, for example, in (Cherednichenko et al., 2023) the more such processes are automated, the faster and more efficient human work becomes. Collecting data from open sources on the Internet is not just about e-commerce. In academic research it is also often necessary to collect information, for example,

Stryzhak et al. (2024) analyze the development of travel and tourism by comparing data from 102 countries. Another work (Labunska et al., 2023) studies the investment attractiveness of enterprises and notes that attraction requires the identification of promising investment objects, which cannot be done only on the basis of analyzing the internal reporting of the enterprise. In such situations, researchers often use information resources that aggregate data and, as a result, face problems of accuracy, consistency and deduplication of information.

The research activities are one of the processes that never stops. Recently, they have been paying great attention to electronic systems for accounting and indexing for research results. In particular, Google Scholar, Scopus, Web of Science, and ResearchGate are widely known systems for keeping track of publications. Effective management of research activities should be based on complete, reliable, and up-to-date information. For this purpose, it is necessary to record the results of research activities and conduct regular monitoring.

^a  <https://orcid.org/0000-0002-9391-5220>

^b  <https://orcid.org/0000-0002-2148-7212>

^c  <https://orcid.org/0000-0003-4701-7830>

As Correia et al. (2021) rightly points out, despite a lot of research in the area of named entity resolution, dealing with name ambiguity is still a difficult problem. Since bibliographic data are generated from publications, the growth of which is significant worldwide, many problems arise due to missing data fields, repetitive objects, spelling errors, extra characters, etc. One of the main problems when working with bibliographic data is their comparison from different sources and accounting systems. This can lead to errors and inaccuracies in the collection result, which in turn negatively affects the quality and reliability of the data obtained.

Thus, the subject of the study is devoted to the issues of collecting and consolidating data on the research activities from different web sources. The purpose of the study is to improve the accuracy of comparing bibliographic data from different indexing systems.

The rest of the paper is organised in the following way. The next section is discussed the existed solutions and approaches to tracking, gathering, and collecting bibliographic data. The third section presents methodological basics and describes data sources which are used. The next section suggests the framework for gathering and matching bibliographic data from the web. Then we describe the results of experimenting and the design of the software prototype. And finally, we discuss the contribution of our research and make the conclusion.

2 THE STATE OF THE ART

The task of information retrieval is considered from the point of view of the following components: searching for information required by users, its sorting and extraction. Information retrieval systems are used to search for information. For sorting, classification or clustering is used. And to extract this information, document standards and Data Mining tools are used. An important component of solving information retrieval tasks in this context is information grouping.

In the most general case, the grouping task is to find, for a given set of objects, such groups that the relationship between objects in the same group is greater than the relationship between objects in different groups. The number of groups in such problems is often assumed to be given, although sometimes it is determined in the process of grouping. The problem of assessing the degree of connection between objects is usually solved by having a table of observations and is usually limited to the case when

the results of measuring these values do not depend on the state of the object under study at the time of previous observations. A general requirement for the data collection methodology to solve the grouping task is to record the values of all parameters at the same time, and if this is not possible, the state of the object should not change during the time of setting the parameter values. The task of object classification is similar. In this case, a set of objects under study is divided into homogeneous classes, each of which is defined by a multidimensional representation, i.e. a set of values of parameters that characterize these objects (Arasu et al., 2003).

Thus, the analysis allows us to conclude that there is a sufficiently developed mathematical framework for solving the problems of grouping information objects collected during monitoring. The choice of a particular method is determined by the peculiarities of the objects to be grouped and the specifics of the task formulation for specific subject areas.

There are a number of issues which we can highlight in terms of consolidating data on research activities. Firstly, the automating the process of data collection which is included the comparing data from different sources is a complex problem (Nurjahan et al., 2023). Secondly, the use of external services to obtain bibliographic information which can be accompanied by the receipt of erroneous data (Correia et al., 2021). In our research we focus on gathering and matching bibliographic data from the web. Thus, in this section we review the existing methods that can be used to solve the problems of comparing bibliographic data in order to develop the algorithm and to solve this problem.

Google Scholar (Orduna-Malea et al., 2017) is a free search service that allows you to find scientific publications, including articles, abstracts, books, reviews, and other types of scientific literature. As for the public API, Google Scholar provides several APIs, but they are only available for use for an additional fee.

There are some Google Scholar scraping tools that allow you to automatically collect data from Google Scholar. The most well-known Google Scholar scraping tool is the Scholarly Python library, which allows you to search for publications, authors, citations, and other data. Other Google Scholar scraping tools include tools such as OutWit Hub, Octoparse, Scrapy, and others. They provide options for automatically collecting information from Google Scholar and saving it in various formats, such as CSV, Excel, JSON, and others.

Google Scholar categorizes data using a system called Google Scholar Metrics (Delgado et al., 2012).

This system uses several deduplication methods to ensure that the same article is not counted multiple times. The main deduplication methods used by Google Scholar Metrics are:

- Article ID: Each article in Google Scholar is assigned a unique article identifier. This identifier is used to identify duplicate articles in different databases.
- Title and author: Google Scholar also uses the title and author name to identify duplicate articles. If two articles have the same title and author name, they are considered duplicates.
- DOI: Google Scholar uses Digital Object Identifier (DOI) to identify duplicate articles in different databases.
- Full-text comparison: Google Scholar can also compare the full text of articles to identify duplicates. Google Scholar uses a proprietary full-text comparison algorithm to identify duplicate articles. The exact details of the algorithm are not disclosed, but it is believed to use a combination of Natural Language Processing (NLP) and machine learning algorithms.

Although Google Scholar is a great tool for finding scientific publications, there are a few downsides to using it.

- Lack of quality control: Google Scholar does not strictly check the quality and reliability of publications, so it is possible to find unverified and unreliable information.
- Incompleteness of the database: Google Scholar may not contain all scientific articles, especially those published in journals with limited access or in other languages.

Google Scholar uses several data processing methods but does not disclose the details of their work. Therefore, it will not be possible to use this algorithm to solve our task.

As a result of the analysis of existing approaches to consolidating data on research activities, a number of problems were identified that need to be addressed. First, the need to constantly compile reports on research activities leads to the issue of automating this process and solving the problem of comparing data from different resources. Secondly, the use of external services to obtain bibliographic information may be accompanied by the receipt of erroneous data, so it is necessary to investigate the issue of their rejection.

3 MATERIALS AND METHODS

Analyzing the identified problems in the process of collecting and processing data from department employees to compile reports on their research activities, we can outline several key points to consider:

- Different spellings of the authors' names.
- Duplication of these publications.
- Filtering out erroneously attributed work.

The very idea of a system for tracking the results of scientific activity implies that such a system will have its own database where the processed bibliographic information will be stored. Therefore, in order to avoid creating duplicates in the database, it is necessary, in addition to deduplicating data from different sources and filtering out erroneously attributed works, to provide for checking the availability of this information in the database.

One of the most common sources of data inconsistencies is typographical variations in string data. There is a high probability that a dataset collected from heterogeneous sources will have redundant data and will need to be analyzed, redundant copies identified and deleted to keep only one copy.

A process aimed at eliminating redundant copies of data and reducing storage costs is called deduplication. Typically, to identify similar data, a character-based similarity metric is used - a metric that measures the distance ("inverse similarity") between two text strings for approximate string matching or comparison and searching on a fuzzy string. There are five commonly used string similarity metrics.

1. Editing distance. There are three types of editing operations:

- Insert a character into a string.
- Delete a character from a string.
- Replace one character with another.

In its simplest form, each edit operation has a cost of 1. This version of the edit distance is also called the Levenshtein distance (Levenshtein, 1965). Levenshtein distance is the most widely known string metric and is used to correct word errors (in search engines, databases, during text entry, in automatic recognition of scanned text or speech), compare text files, genes, chromosomes, and proteins in bioinformatics.

2. Athenian gap distance. The edit distance metric described above does not work well when comparing strings that have been truncated or shortened (e.g.,

"John R. Smith" vs. "Jonathan Richard Smith"). The affine gap distance metric offers a solution to this problem by introducing two additional editing operations: gap opening and gap widening (Ristad et al., 1998). The cost of gap expansion is usually less than the cost of opening a gap, which results in lower penalties for gap mismatches than the equivalent cost in the edit distance metric.

3. Smith-Waterman distance (Smith et al., 1981). Smith and Waterman (1981) described an extension of the edit distance and affine break distance in which mismatches at the beginning and end of lines have a lower value than mismatches in the middle. This metric improves the local alignment of strings (i.e., substring matching). Therefore, the strings "Professor John R. Smith, University of Calgary" and "John R. Smith, Professor" can match within a small distance using the Smith-Waterman distance, since prefixes and suffixes are ignored. The distance between two strings can be calculated using a dynamic programming based on the Needleman and Wunsch algorithm.

4. Jaro distance metric (Elmagarmid et al., 2007) introduces a string comparison algorithm that is used primarily for comparing last names and first names. Winkler and Thibodeau modified the Jaro metric by giving more weight to prefix matches, as prefix matches are usually more important for matching surnames.

5. Q-grams are short character substrings of the length q of database rows. The intuition behind using q-grams as a basis for approximate string mapping is that when two strings are similar, they have a large number of common q-grams in common. Q-grams, including trigrams, bigrams, and/or unigrams, have been used in various ways for text recognition and spelling correction. One of the natural extensions of q-grams are positional q-grams, which also record the position of q-grams in a string (Elmagarmid et al., 2007).

Analyzing the above similarity metrics, we will use the Levenshtein distance as an approach for duplicating bibliographic data of research papers as the most suitable for calculating the editing distance between two lines. This choice is explained by the fact that the titles of scientific articles most often do not contain first and last names, do not have abbreviations, and the cost of discrepancies in any part of the title is the same.

One of the main tasks of analyzing bibliographic data is to filter out erroneously attributed works. To check whether an article belongs to a certain author, it was decided to analyze the titles of articles and compare their topics with the topics of the author's

previous works. This solution cannot fully provide the necessary filtering, because there is a possibility that the article is really by the author, although it is written on a topic completely distant from the author's main field of work. One way of analyzing data that can help solve this problem is to use the Wu & Palmer Similarity algorithm (Meng et al., 2013).

Wu & Palmer Similarity is a measure of the semantic closeness of two words in a lexical context based on how close the words are in the WordNet hypernym tree (Pratama et al., 2022). WordNet is a lexical database of the English language that describes the relationships between words, including synonyms, antonyms, hypernyms, hyponyms, and others. The Wu & Palmer semantic closeness measure is defined as the height of the Least Common Ancestor (LCA) of two words in the WordNet hypernym tree divided by the sum of the depths of these two words in the tree (Meng et al., 2013). Thus, the closer the words are located in the WordNet hypernym tree, the higher their Wu & Palmer Similarity is. This measure can be used in natural language processing tasks, such as calculating semantic closeness between words, text classification, and other tasks.

The Wu & Palmer Similarity algorithm has the following steps:

- Identify two words for which you want to measure semantic closeness.
- Finding hyperlinks (parent concepts) that contain the two words. A hypernym tree is used to organize concepts in lexical databases such as WordNet.
- Finding the LCA of two words in WordNet, i.e. the parent concept that has the smallest distance from the WordNet root to the two source words. To find the LCA, you can use hypernym tree traversal algorithms, such as the DFS (Depth-First Search) or BFS (Breadth-First Search) algorithm.
- Calculating the height of the LCA (the number of levels up from the root to the LCA).
- Calculating the depth of each word (the number of levels up from the root to each word).
- Calculating semantic closeness.
- Return the semantic closeness value between two words.

Thus, there are some advantages of using the Wu & Palmer Similarity method.

- Hierarchical structure: Wu & Palmer Similarity considers the hierarchical structure of a lexical database such as WordNet, which reflects the inherent relationships between words. This allows

for a more accurate representation of the semantic relationships between words.

- Intuitiveness: Wu & Palmer Similarity is based on the idea that words that share a common ancestor in a hierarchy are more similar in meaning than words that do not. This intuitive approach makes the results easier to understand and interpret.
- Easy to calculate: Wu and Palmer similarity is a relatively simple method that can be computed quickly and efficiently. This makes it a practical choice for many natural language processing applications.
- Widely used: Wu & Palmer Similarity is a well-known method that is widely used and studied in the field of natural language processing. This means that many resources and tools are available to work with it, making it a convenient choice for researchers and developers.

However, it is important to note that Wu & Palmer Similarity is designed to compare semantic similarity between two concepts, not physical proximity. Therefore, while it can be used to compare the similarity of the language used in two strings, it is not suitable for comparing the physical proximity of those strings.

4 DATA COLLECTION FRAMEWORK DEVELOPMENT

The proposed algorithm performs the task of processing records in three stages: the stage of removing duplicate records from the array of data

obtained from different sources, the stage of removing records that are already present in the database, and the stage of checking whether the remaining publications belong to the author. To fully understand the place of Levenshtein's distance and the Wu & Palmer Similarity algorithm in the process of filtering the collected data, Figure 1 shows the process of categorizing these publications.

The stage of removing duplicate records involves checking for matches between publications by title, place of publication (journal, scientific publication, etc.), and year of publication. If these three values match for publications from different sources, we believe that these records can be merged.

The stage of removing publications present in the database is identical to the stage of removing duplicate records. Having received a set of publications after the previous stage, we do the same thing: we compare the title, place of publication, and year of publication with the data of publications already in the database.

The stage of checking whether a publication belongs to the author is based on the Wu & Palmer Similarity algorithm to compare the titles of publications and the interests of authors - the subject categories that interest the author and within which he or she wrote articles.

As a result of this data processing, publications obtained from bibliographic information sources will be divided into 3 categories:

- New publications.
- Publications that require confirmation.
- Previously saved publications.

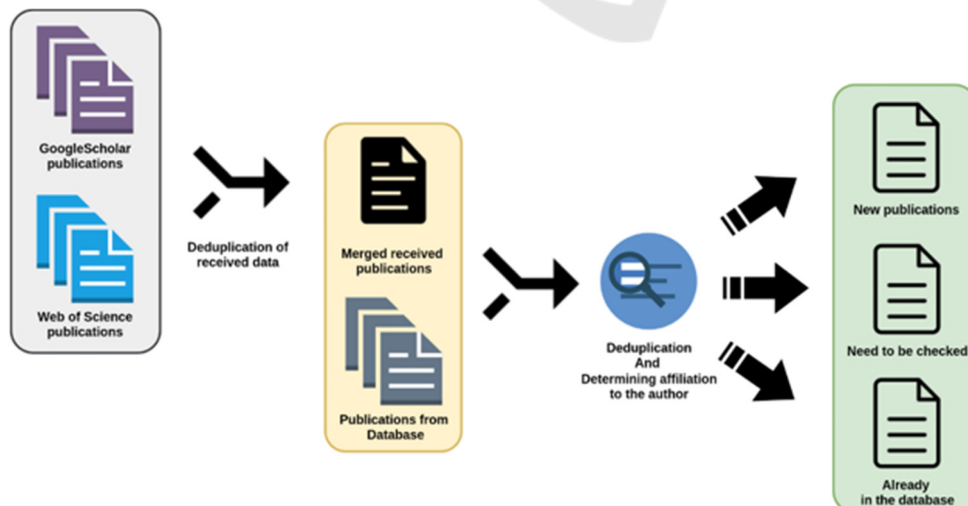


Figure 1: Data categorization process.

Thus, the processing stages and the algorithm for processing bibliographic data were formed, as a result of which the resulting set of research papers can be saved to the database and used for further reporting.

The high-level architecture of developed software is shown in Figure 2.

Python was chosen as the programming language. This choice is due to the fact that Python is one of the most popular programming languages for working with neural networks, providing a large number of libraries with implementations of various methods of data analysis, processing, and filtering. The Wu-Palmer algorithm is also implemented within the Python.

5 EXPERIMENTS

The proposed algorithm involves deduplication of data and verification of their belonging to the author. It is presupposed to use 6 threshold values. They are for comparing the author's name, for calculating the Wu & Palmer similarity, two for checking the sources of publications and two for checking the titles of publications based on the Levenshtein distance. In order to choose the threshold values an experiment is conducted.

Four datasets were created as input data for the experiment:

- Bibliographic data of publications of an arbitrary author obtained from Google Scholar (100 records).
- Bibliographic data of the author's publications obtained from Web of Science (30 records).

- Personal data of the author and his publications previously obtained and stored in the database.
- The expected result in the form of a set of publications that will be marked as "New".

The purpose of this experiment is to find the parameters of the algorithm to obtain the best accuracy of categorization of the obtained data of the author's publications.

Metrics based on the confusion matrix were chosen as evaluation criteria. Such a matrix contains generalized information about the model's performance, including the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) classifications. The matrix is usually presented in the form of a table, with the predicted labels on one axis and the actual data labels on the other. The cells in the table contain the counts of the various outcomes. In the experiment, the cells of the confusion matrix are defined as follows:

- TP is the number of publications that were attributed to the author correctly.
- TN is the number of publications that were not attributed to the author.
- FP is the number of publications that were attributed to the author by mistake.
- FN is the number of publications that were not attributed to the author by mistake.

The confusion matrix is a useful tool for evaluating the performance of classification models and can be used to calculate such metrics as accuracy, precision, recall, and F1 score.

Unlike optimizing the performance of single-variable algorithms, optimizing a multi-variable algorithm requires considering many possible values

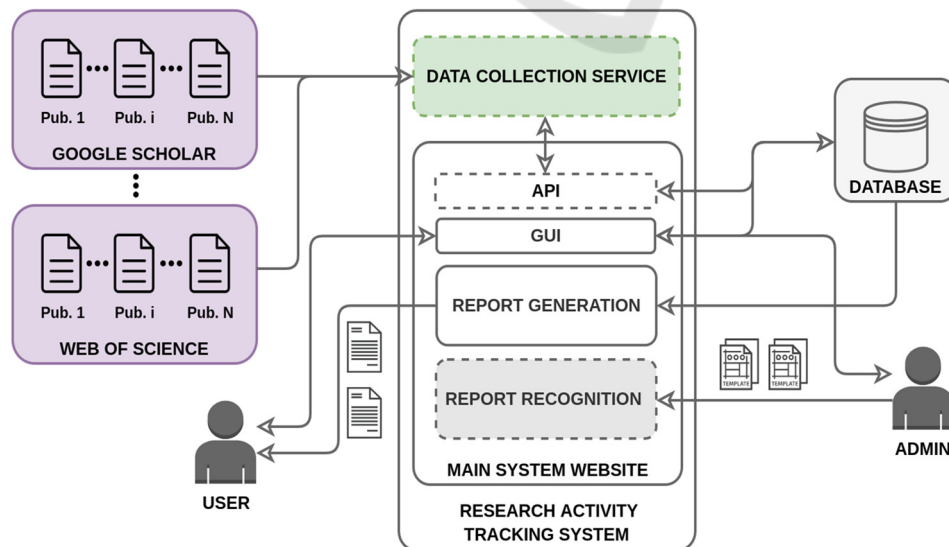


Figure 2: The high-level architecture.

for each variable and their interactions, which can be time-consuming and require specialized optimization methods such as gradient descent or evolutionary algorithms. Taking this into account, it was decided to create a mathematical model of the algorithm as a function of several variables and use special optimization methods to find the optimal values of the variables. These methods make it possible to find the minimum or maximum of a function in a given range of variables with high accuracy and efficiency, which makes optimization of complex algorithms possible.

We use the F1 metric as the value to maximize, as it is a good indicator of model quality, combining Precision and Recall into one value. The determination of the appropriate optimization method depends on many factors, such as the number of variables, the type of function, and the presence of constraints. To determine the type of our function, we created a graph of the change in metrics depending on the threshold value (Fig. 3).

As we can highlight over the graph, it is a step function. There are several optimization methods that can be used to optimize multi-criteria step functions, i.e. Nelder-Mead method, genetic algorithms, linear programming methods, methods of cluster analysis. Each of these methods has its own advantages and disadvantages. We choose the Nelder-Mead algorithm to optimize the function.

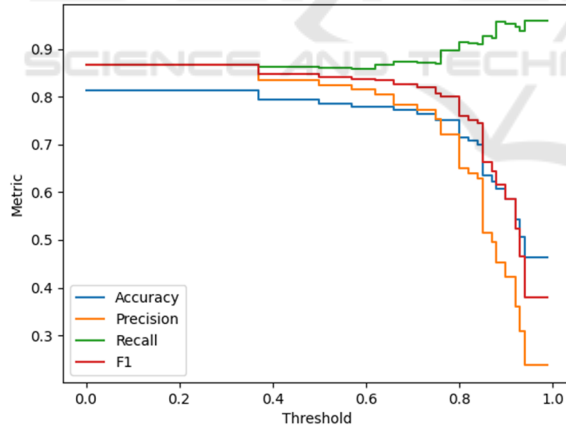


Figure 3: Variation of a function by one variable.

The Nelder-Mead optimization algorithm is a commonly used numerical optimization algorithm for minimizing nonlinear, unconstrained objective functions (Bazaraa et al., 2005). It belongs to the class of direct search methods that do not require information about the gradient of the objective function. Instead, the algorithm iteratively searches for the minimum of the objective function, moving from one point to another in the search space.

As a result of applying the Nelder-Mead algorithm to optimize the F1 function the necessary values of the thresholds are obtained.

- the author's name threshold = 0.88
- the first publication source threshold = 0.93
- the second publication source threshold = 0.71
- the first publication title threshold = 0.95
- the second publication title threshold = 0.92
- Wu & Palmer similarity = 0.62

Based on the obtained thresholds the developed algorithm gives the following values of key metrics.

- Accuracy=0.84
- Recall= 0.81
- Precision=0.96
- F1 Score=0.88

We can conclude that the proposed algorithm showed good results, reaching 0.88 for the F1 metric. Thus, in this section, we test the developed algorithm for comparing and classifying publications and, by performing an experiment, established the optimal values of the thresholds used in it.

6 DISCUSSION AND CONCLUSIONS

Thus, to compare similar records from different sources, we suggested using edit distance to compare titles, source, and year of publications. To determine whether the bibliographic data belonged to the author, we suggested using Wu & Palmer Similarity algorithm to compare the topics of the author's publications with the topics that have already been used by him before. All together is the basis of the framework suggested in this research. We have experimented with algorithm to find out the optimal values of the thresholds and developed the software prototype.

To summarize the obtained results, we can underline the following. To answer the first research question, we propose the deduplication algorithm with the Levenshtein distance. The editing distance between two lines is the most appropriate approach because of the fact that the titles of scientific articles most often do not contain first and last names, do not have abbreviations, and the cost of discrepancies in any part of the title is the same. Thus, these features make it inappropriate to use the other similarity metrics to eliminate redundant copies of data.

As to regards of the second research question we suggest using the Wu & Palmer Similarity algorithm. To check whether an article belongs to a certain

author we compare their topics with the topics of the author's previous works. As Wu & Palmer Similarity is a measure of the semantic closeness of two words in a lexical context, it helps to filter the data by topic related to the author. It is important to notice that this solution cannot fully provide the necessary filtering, because there is a possibility that the article is really by the author, although it is written on a topic completely distant from the author's main field of work.

Analyzing the results of the algorithm, we identified several weaknesses that need to be improved in the future.

In cases where an author contributes to an article that is not related to his or her field of interest, a problem arises when filtering out falsely attributed works, because the algorithm we developed will classify such articles as falsely attributed or those that need to be clarified.

Analyzing the data of publications by different authors, it was determined that often the same author can have several different articles with identical or almost identical titles. Therefore, it is necessary to consider alternative ways of comparing publication titles in order to avoid mistakenly combining two different publications. Thus, further ways to improve the system have been identified.

ACKNOWLEDGEMENTS

The research study depicted in this paper is partially funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under project No. 09I03-03-V01-00078.

REFERENCES

- Arasu, A., Garcia-Molina, H. (2003). Extracting structured data from web pages. *Proceedings of the 2003 ACM SIGMOD international conference on Management of data*. 337–348.
- Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2005). *Nonlinear Programming: Theory and Algorithms*. *Nonlinear Programming: Theory and Algorithms* (pp. 1–853). John Wiley and Sons.
- Cherednichenko, O., Ivashchenko, O., Lincényi, M., & Kováč, M. (2023). Information technology for intellectual analysis of item descriptions in e-commerce. *Entrepreneurship and Sustainability Issues*, 11(1), 178–190.
- Correia, A., Guimaraes, D., Paulino, D., Jameel, S., Schneider, D., Fonseca, B., & Paredes, H. (2021). AuthCrowd: Author Name Disambiguation and Entity Matching using Crowdsourcing. In *Proceedings of the 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2021* (pp. 150–155). Institute of Electrical and Electronics Engineers Inc.
- Delgado López-Cózar E., Cabezas-Clavijo, Á. (2012). Google Scholar Metrics: an unreliable tool for assessing scientific journals. *El Profesional de la información*. Vol. 21. 4.
- Elmagarmid, A. K., Panagiotis, G.I., Vassilios, S.V. (2007) Duplicate Record Detection: A Survey. *IEEE Transactions On Knowledge And Data Engineering*. Vol.19. No.1.
- Labunska, S., Cibák, L., Sidak, M., & Sobakar, M. (2023). The role of internally generated goodwill in choosing areas and objects of investment. *Investment Management and Financial Innovations*, 20(2), 215–231.
- Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10, 707-710.
- Meng, L., Huang, R., & Gu, J. (2013). A Review of Semantic Similarity Measures in WordNet. *International Journal of Hybrid Information Technology*, 6(1), 1–12.
- Nurjahan, V. A., & Jancy, S. (2023). Dual Cloud Bibliographic Network Model for Citation Recommendation Systems. In *2023 Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems, AICERA/ICIS 2023*. Institute of Electrical and Electronics Engineers Inc.
- Orduna-Malea, E., Martín Martín, A., Delgado Lopez-Cozar, E. (2017). Google Scholar as a source for scholarly evaluation: A bibliographic review of database errors. *Revista Española de Documentación Científica*, Vol. 40. 4.
- Pratama, M. A., & Mandala, R. (2022). Improving Query Expansion Performances with Pseudo Relevance Feedback and Wu-Palmer Similarity on Cross Language Information Retrieval. In *2022 9th International Conference on Advanced Informatics: Concepts, Theory and Applications, ICAICTA 2022*. Institute of Electrical and Electronics Engineers Inc.
- Ristad, E.S., Yianilos, P.N. (1998) Learning String Edit Distance. *IEEE Trans. Pattern Analysis and Machine Intelligence*. Vol. 20. 5. 522-532.
- Smith, T. F. , Waterman, M.S. (1981) Identification of Common Molecular Subsequences. *J. Molecular Biology*. Vol. 147. 195- 197.
- Stryzhak, O., Cibák, L., Sidak, M., & Yermachenko, V. (2024). Socio-economic development of tourist destinations: A cross-country analysis. *Journal of Eastern European and Central Asian Research (JEECAR)*, 11(1), 79–96.