УДК 005.8

[1] **Viktor D. Gogunsky**
Doctor of Engineering, Professor, Head of "Life safety management systems department"
[2] **Volodymyr O. Iakovenko**
Postgraduate, developer of software projects
[1]**Andriy S. Kolyada**
Assistant of "Life safety management systems department"

[1] *Odessa National Polytechnic University, Odessa*
[2] *Company "Tobii Technology" AB, Stockholm, Sweden*

THE DEVELOPMENT OF THE SYSTEM CONCEPT OF SCIENTOMETRIC DATABASES

Анотація. *Проаналізовано принципи роботи наявних науково-метричних баз даних. Запропоновано концепцію автоматизованої інформаційно-аналітичної системи для моніторингу інформаціїї щодо публікацій науковців з України у міжнародних науково-метричних базах даних. Описано проблему розподілення отриманих статей між авторами та наведено її рішення.*

Ключові слова: науково-метричні бази даних; інформаційно-аналітична система; латентно-семантичний аналіз; аналізатор даних

Аннотация. *Проанализированы принципы работы существующих научно-метрических баз данных. Предложена концепция автоматизированной информационно-аналитической системы для мониторинга информациии относительно публикаций ученых из Украины в международных научно-метрических базах данных. Описана проблема распределения полученных статей между авторами и приведено ее решение.*

Ключевые слова: научно-метрические базы данных; информационно-аналитическая система; латентно-семантический анализ; анализатор данных

Abstract. *The article shows an advantage of developing an automated system for monitoring scientific publications of Ukraine in the international scientific-metric databases. The main problems of this system are distribution of articles by author and authors with the same surname, name and patronymic (SNP). It is proposed to use data analyzer to solve these problems. Analyzer will get as an input a list of articles from international scientometric databases according to the specified SNP, and will return adjusted list of publications of authors without the same SNP. Latent semantic analysis is used in an automatic mode to improve system performance and to reduce user interaction with the analyzer. The whole information-analytical system's structure is tightly coupled, but at the same time each of its components executes functions specific only to it relying on the output from other components. The developed system will help to study the structure and evolution of the various branches of science in Ukraine.*

Key words: sciencometric databases; information-analytical system; latent semantic analysis; data analyzer

Introduction

The citation of scientific articles in various scientific studies becomes more and more popular nowadays regardless of a sector. The question of accounting these articles and further processing appears due to the diversity of scientific publications and a large number of authors. There are many scientometric databases, which address this issue, the most popular of them are Web of Science, Scopus, Web of Knowledge, Astrophysics, PubMed, Mathematics, Chemical Abstracts, Springer, Agris, GeoRef and others [1 – 3].

The need to create an automated information system that can monitor and store the information on publications of scientists from Ukraine in international scientometric databases arises because of the new requirements of the Ministry of Education in promoting the publication of scientific articles in national and international journals [4 – 8].

Such automated information system will help to study the structure and evolution of different areas of science in Ukraine. Results can be communicated via tables, graphs, geographic and topic maps. Frontiers emerging across different sciences can be discovered and tracked. Different funding models can be simulated and compared. School children can start to understand the symbiotic relationships among different areas of science [9 – 16].

It is proposed to develop the information system, which will have the structure shown on Fig. 1.

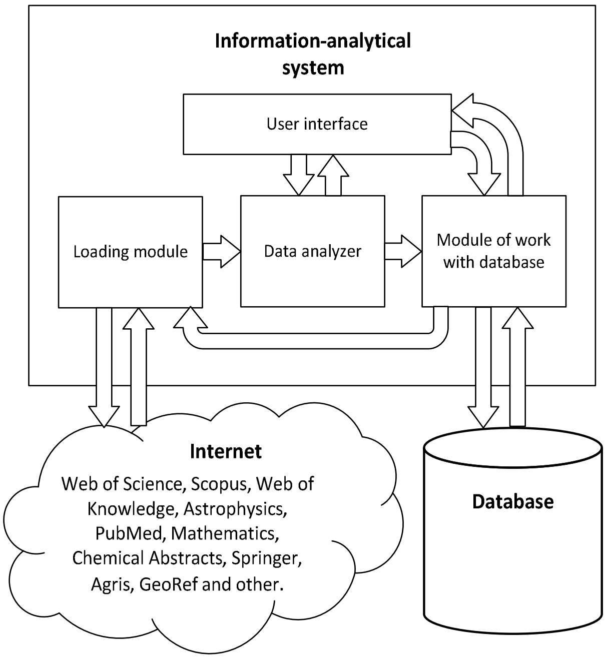
Information-analytical system
User interface
Loading module
Data analyzer
Module of work with database
Internet
Web of Science, Scopus, Web of Knowledge, Astrophysics, PubMed, Mathematics, Chemical Abstracts, Springer, Agris, GeoRef and other.
Database

Fig. 1. The concept of the system of scientific publications

The interaction with the system will be possible using the *User interface*. Once the user has made a request for publications by author's name, *Database work module* checks for such publications in the database. The system by the request of user and with the help of the *Download module* receives a list of articles of this author from *Internet*, namely from the international scientometric databases. The list goes further to the *Data analyzer*, which checks the received items to the author's name. The result is passed back to the *Database work module*, which, in turn, provides a list of items for a specific author in the *Database*. The *Database* usage allows keeping records of publications of each scholar in a single local repository, while assigning a unique identification number to the author.

Problem setting

This work is complicated by the following problems:

1. The articles from international databases should be distributed according the author. Doing it manually by the user, not by the system, is too complex and time consuming procedure due to the large set of publications.

2. There are authors with the same surname, name, and patronymic (SNP). This makes the usage of SNP as a unique identifier of the article impossible.

The first problem is explained by the huge amount of scholarly documents in the Internet, only scientific articles in English there is more than 114 millions [17, 18]. The distribution of such huge amount of articles, which are not only written in English, but also written in other languages, by a human is very time consuming or even impossible procedure.

The second problem is quite common, there many authors in Ukraine who have the same SNP. The task is to define to which person with the same SNP each scholarly document belongs. This problem is even exacerbated by the same research area of authors with the same SNP [19 – 21].

Research objective

It is proposed to use data analyzer to solve these problems, such analyzer will get as an input a list of articles from international scientometric databases according to the specified SNP, and will return adjusted list of publications of authors without the same SNP. It is proposed to make analyzer data to be automized, i.e. which will perform analysis by software and by user. The whole process inside the analyzer will take place as follows. Once a list of publications will be received as an input, analyzer will divide them by category using latent semantic analysis and will pass them to the user interface that will display the distributed publications. The user, in turn, using the classification of topics, will select articles, the authorship of which belongs to him, and the analyzer will offer a similar article for each of the selected publications. Detailed process of interaction of data analyzer with the user interface is shown on Fig. 2.

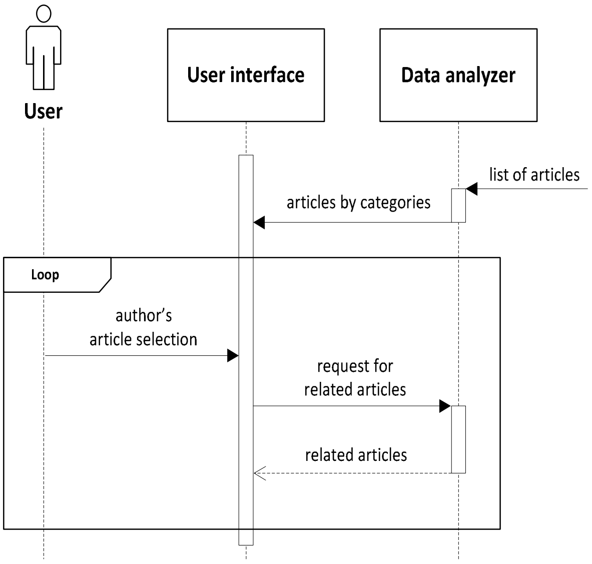

Fig. 2. Interaction of data analyzer with the user interface

Failure to make this process automatic, that means running by a program, is caused by the inability without a user to identify what publications of what specific author are required under the same SNP. Also maybe some inaccuracy in obtaining the results of comparison of publications and their distribution in categories based on latent semantic analysis, because the reaction requires the user to correct such inaccuracies.

According to the description of the data analyzer we can conclude that it will perform two main functions:

1. Distribution of articles on a prescribed list of categories.

2. Finding similar articles to the given one.

It is proposed to implement all these functions basing on a latent-semantic analysis. Latent-semantic analysis (LSA) is a method of information processing in natural language, which allows analyzing the relationship between the collection of documents and terms, which are found in them.

The main task of LSA is to overcome the deficiencies of term-matching retrieval by treating the unreliability of observer term-document association data as a statistical problem. This approach assumes there is some underlying latent semantic structure in the data that is partially obscured by the randomness of word choice with respect to retrieval. LSA uses statistical techniques to estimate this latent structure, and get rid of the obscuring “noise”. A description of terms and documents based on the latent semantic structure is used for indexing and retrieval [21 – 24].

Implementation

The particular "latent semantic indexing" (LSI) analysis uses singular-value decomposition. We take a large matrix of term-document association data and construct a "semantic" space wherein terms and documents that are closely associated are placed near one another. Singular-value decomposition allows the arrangement of the space to reflect the major associative patterns in the data, and ignore the smaller, less important influences. As a result, terms that did not actually appear in a document may still end up close to the document, if that is consistent with the major patterns of association in the data. Position in the space then serves as the new kind of semantic indexing, and retrieval proceeds by using the terms in a query to identify a point in the space, and documents in its neighborhood are returned to the user [22].

The following steps are needed to implement the latent semantic analysis:

1. Excluding of stop symbols.

2. Stemming process.

3. Building of table of word usage.

4. Orthogonal decomposition of created table as a matrix.

5. Getting a two-dimensional matrix from the result of decomposition.

6. Multiplication of two-dimensional matrices.

7. Spearman correlation on the product of matrices.

8. Results analysis.

On the first step we should exclude from the set of documents that arrived as an input, stop symbols, i.e., the most frequently used words that do not have a special meaning. To stop characters in the Ukrainian language belong prepositions, suffixes, participles, interjections, particles, etc. It is easy to find the ready-made list of stop characters in the public domain and basing on it to process documents.

The next step is stemming process. Stemming is the process of finding basis of a word considering the morphology of the original word. Ukrainian language has a complex morphological variability of words, which is a source of error when using stemming. As a solution of this problem together with classical stemming algorithms can be used lemmatization algorithms that lead words to the initial contract base forms. One possible use for the algorithm can be stemmer of Porter. The algorithm consists of several steps. At each step separated word creation suffix and the rest is checked against the rules (for example, the basis of Ukrainian words should have at least one vowel). If the resulting word satisfies the rules, it is moved to the next step [25].

After stop symbols removing and stemming process it is needed to create a table, where the columns will be documents received at the input and the raws will be words that occur in at least two documents. Each cell will show how many times the words are encountered in the document. The created table will be represented as a matrix.

The next step is to complete orthogonal decomposition matrix for selecting of it components and ignoring "noises". Complete orthogonal decomposition matrix A of size NxM by definition has the form A = UKV^T. Here U and V are orthogonal matrices of size NxN and MxM respectively, and K is a matrix of size NxM, which has the following structure: $K = \begin{pmatrix} W & 0 \\ 0 & 0 \end{pmatrix}$, where W is a matrix of size KxK, where K is a rank of the original matrix A. The most famous of orthogonal decompositions is the singular decomposition of form of $A=USV^T$, where S is a diagonal matrix composed of zeros and located on the diagonal of singular values of matrix A [26].

In the next step we reject the last column of matrix A and the last rows of matrix V^T by leaving only the first 2. It is important that optimal results of the next multiplication are guaranteed. The decomposition of this type is called the two-dimensional singular decomposition [27].

The resulting matrix without "noises" will be received after the product of two-dimensional matrices U, S, V^T. There is a possibility to determine semantic correlation between documents after receiving of such matrix. The Spearman rank correlation between the columns of the matrix can be used as one of the criteria of correlation. The larger the value obtained after the correlation, the greater the semantic similarity between documents. The maximum correlation value can be 1 (documents are identical in meaning), and minimal – 1 (completely different documents in meaning). The more words in the document, the more accurate results can be obtained [27].

Conclusions

Basing on the above-described algorithm of latent-semantic analysis we can identify the data required to analyzer as an input for receiving expected result.

As for its first function, namely the distribution of articles by category, you need to get as an input an actual article and thesaurus of categories. The meaning of thesaurus refers to keywords of each category (for example, if as categories is a list of specialties, the thesaurus can be a passport of each specialty). The data analyzer will compare the article with all thesauruses during executing of this function and will return a category as an output, thesaurus of which is most semantically similar to the publication.

To perform the second function, such as finding similar articles, it is needed to give the original article to the input of analyzer, set of items for comparison and an amount of related articles that will be received at the output. It is proposed to provide only sets of articles from the same category, from which is original article, to save the calculation time.

The example of such functions executions by the analyzer is shown on Fig. 3.

The data analyzer is the core element of the proposed information-analytical system. The main its drawback is dependency on a user, which should control analyzer's work. The inability to abandon this dependency sets a goal to reduce user interaction with the analyzer to minimum. The usage of latent semantic analysis provides the ability to achive this goal by implementing two core functions of the data analyzer.

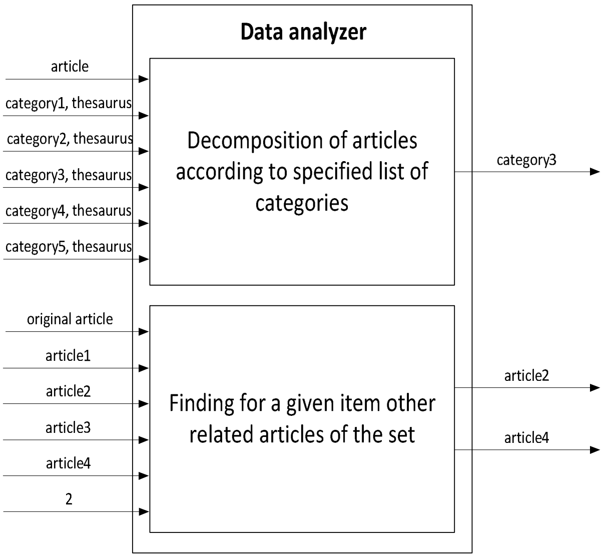

Fig. 3 The example analyzer's data input and output

The whole information-analytical system's structure is tightly coupled, but at the same time each of its components executes functions specific only to it relying on the output from other components. The sytem can work offline using only local data from the database. It is a great advantage, but at the same time it requires big amount of local memory space.

The most simple and widely used approach to reduce complexitiy of the information-analytical system, especially of the data analyser component, is the usage of the unique identificator for each author from Ukraine. This solution will help to dispose time that will be spent on decomposition of articles between scientists with the same SNP. This approach can be implemented inside the database of the proposed information-analytical system basing on some local identificators for each author or the system can use already created identificators for authors by Ministry of Education. The drawback of the second solution is that such identificators are not widely used by the authors from Ukraine.

References

1. Burkov, V. N., Beloschitsky, A. A., & Gogunsky, V. D. (2013). Options citation of scientific publications in scientometric databases. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 15, 134 - 139.

2. Gogunsky, V. D., Kolyada, A. S., & Iakovenko, V. O. (2014). Scientometric data scientific publication "Management of development of difficult systems. Management of development of difficult systems. Kyiv, Ukraine: KNUCA, 19, 6 – 11.

3. Bushuev, S. D., Beloschytsky, A. A., & Gogunsky, V. D. (2014). Scientometric database: characteristics, opportunities and challenges. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 18, 145 – 152.

4. Bui, D., Beloschytsky, A., & Gogunsky, V. (2014). Scopus and other scientometric database: simple questions and vague answers. High School. Kyiv, Ukraine: 4, 37 - 40.

5. Beloschitsky, A. A. (2012). Management problems in the methodology of design vector control of the educational environment. Management of development of difficult systems. № 9, 104 - 107.

6. *Bushuev, S. D., Gogunsky, V. D., & Koshkin, K. V. (2012). Areas of dissertation research in the specialty "Program and Project Management." Management of development of difficult systems. № 12, 6 - 9.*

7. *Lizunov, P. P., Beloschitsky, A. A., & Beloschitskaya, S. V. (2011). Design vector control higher education institutions / Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 6, 135 - 139.*

8. *Maslennikova, K. S., & Kolesnikova, K. V. (2013). Components behavioral competence of project team members on the basis of competency approach. Management of development of difficult systems. Kyiv, Ukraine, KNUCA: 14, 48 - 51.*

9. *La Rowe, Gavin, Ambre, Sumeet, Burgoon, John, Ke, Weimao and Börner, Katy. (2007) The Scholarly Database and Its Utility for Scientometrics Research. In Proceedings of the 11th International Conference on Scientometrics and Informetrics, Madrid, Spain, June 25-27, 2007, pp. 457-462.*

10. *Lizunov, P. P., & Biloschytsky, A. A. (2007). Create information-educational environment of higher educational institution. Journal of East-Ukrainian National University V.I. Dahl. № 5 (111), part 1, 205 - 210.*

11. *Rach, V., Rossoshans'ka, O., & Medvedeva, O. (2011). Building a terminological system of scientific knowledge. Scientific world. No.4, 13 - 16.*

12. *Teslya, Yu. M., Beloschytsky, A. A., & Teslya, N. Yu. (2010). Information Technology Project Management based ERPP (enterprise resources planning in project) and APE (administrated projects of the enterprise) systems. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 1, 16 - 20.*

13. *Kolesnikova, K. V. (2013). The development of the theory of project management: project initiation study law. Management of development of difficult systems. Kyiv, Ukraine, KNUCA: 17, 24 - 30.*

14. *Vlasenko, O. V., Lebed' V. V., & Gogunsky, V. D (2012). Markov model of communication processes in international projects. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 12, 35 - 39.*

15. *Kolesnikova, K. V. (2013). The development of the theory of project management: Explanation law K.V Koshkin to complete projects. Management of development of difficult systems. Kyiv, Ukraine, KNUCA: 16, 38 - 45.*

16. *Lizunov, P., Biloschytsky, A. (2007). Models and means of forming complex information-educational environment of the institution. Information processing systems. Kharkiv, Ukraine: 6(63), 2-7.*

17. *Khabsha, M., & Giles, C.L, (2014). The Number of Scholarly Documents on the Public Web. PLoS ONE 9(5): e93949.*

18. *Mazaraki, A., Prytulska, N., Melnichenko S. (2011). Integration of domestic science to the world through scientometric database. Bulletin KNTEU. Kyiv, Ukraine: 6, 5 - 13.*

19. *Biloschytsky, A. A., & Dikhtyarenko, O. V. (2013). Effectiveness of methods to search for matches in the texts. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 14, 144 – 147*

20. *. Biloschytsky, A.A., Dikhtyarenko, O.V., & Lyaschenko, T.O. (2013). Conversion of different types of files to one format. Management of development of difficult systems. Kyiv, Ukraine: KNUCA: 18, 140 – 144.*

21. *Gogunsky, V. D., Iakovenko, V. O., & Kolyada, A. S. (2014). Application of Latent Dirichlet allocation for the analysis of scientometric publications database. Proc. of Odes. Polytechnic. Univ. Odessa, Ukraine, ONPU: 1 (43), 186 – 191.*

22. *Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman (1990). «Indexing by Latent Semantic Analysis» (PDF). Journal of the American Society for Information Science 41 (6): 391–407.*

23. *Susan, T. Dumais (2005). "Latent Semantic Analysis". Annual Review of Information Science and Technology 38: 188.*

24. *Markovsky, I. (2012) Low-Rank Approximation: Algorithms, Implementation, Applications, Springer.*

25. *Lovins, Julie Beth (1968). "Development of a Stemming Algorithm". Mechanical Translation and Computational Linguistics 11: 22–31.*

26. *DeAngelis, G C, Ohzawa I, & Freeman R. D. (October 1995). "Receptive-field dynamics in the central visual pathways". Trends Neurosci. 18 (10): 451–8.*

27. *Chris Ding and Jieping Ye. "Two-dimensional Singular Value Decomposition (2DSVD) for 2D Maps and Images". Proc. SIAM Int'l Conf. Data Mining (SDM'05), pp. 32–43, April 2005.*

28. *Lehman, Ann (2005). Jmp For Basic Univariate And Multivariate Statistics: A Step-by-step Guide. Cary, NC: SAS Press. p. 123.*

Стаття надійшла до редколегії 29.10.2014

Reviewer: д-р техн. наук, проф. А.Л. Становський, Одеський національний політехнічний університет, Одеса.