UDC 004.89

Y. Berlizov, V. Hapiak,
V. Liubchenko, ScD.

THE ALGORITHMS FOR SOFTWARE SYSTEM OF SCIENTIFIC PUBLICATIONS ANALYSIS

Abstract. *The purpose of the work is to define algorithms for the software system of scientific publications analysis, designed to identify research areas and groups of researchers with similar interests within the same university or faculty.*

There are many algorithms for solving information extracting problems, but they have some disadvantages regarding the solved problem. Therefore, we developed a proprietary algorithm that consists of four steps: lexical analysis, terminals normalization, entities combining and filtering.

The results of information extracting are used to solve identification problems of authors groups and keywords groups considered as a clustering problem. The analyzed data are presented in the form of graphs of two types: a weighted graph of authors' interactions and semantic graph of papers. This allows using for the analysis the clustering algorithms based on graph theory and algorithm of stochastic analysis MCL. An analysis of a test articles sample showed that clustering algorithms based on graph theory and algorithm of MCL identified the same clusters, but the algorithm that based on minimum spanning tree was better regarding computational complexity.

Keywords: *Analysis of Scientific Publications, Information Extraction, Terminals, Entities Merging, Entities Filtering, Clustering Algorithms, Graphs, Markov Clustering Algorithm*

Е. В. Берлизов, В. Н. Гапяк,
В. В. Любченко, д-р техн. наук

АЛГОРИТМИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПРОГРАММНОЙ СИСТЕМЫ АНАЛИЗА НАУЧНЫХ ПУБЛИКАЦИЙ

Аннотация*. Рассмотрены алгоритмы, применяемые для автоматизированного извлечения и анализа информации о научных публикациях. Для извлечения информации предложен алгоритм, состоящий из четырех шагов – лексический анализ, нормализация терминалов, объединение и фильтрация сущностей. Анализ информации предложено выполнять с помощью алгоритма, основанного на минимальном покрывающем дереве.*

Ключевые слова: *анализ научных публикаций, извлечение информации, терминалы, объединение сущностей, фильтрация сущностей, алгоритмы кластеризации, графы, алгоритм марковской кластеризации*

Є. В. Берлізов, В. М. Гап'як,
В. В. Любченко, д-р техн. наук

АЛГОРИТМІЧНЕ ЗАБЕЗПЕЧЕННЯ ПРОГРАМНОЇ СИСТЕМИ АНАЛІЗУ НАУКОВИХ ПУБЛІКАЦІЙ

Анотація. *Розглянуто алгоритми, що застосовуються для автоматизованого здобування та аналізу інформації про наукові публікації. Для здобування інформації запропонований алгоритм, що складається з чотирьох кроків – лексичний аналіз, нормалізація терміналів, об'єднання і фільтрація сутностей. Аналіз інформації запропоновано виконувати за допомогою алгоритму базованому на мінімальному кістяковому дереві.*

Ключові слова: *аналіз наукових публікацій, здобування інформації, термінали, об'єднання сутностей, фільтрація сутностей, алгоритми кластеризації, графи, алгоритм марковської кластеризації*

Introduction. The modern world makes high-level demands for the scientific activities of universities. Ongoing researches, both fundamental and applied, should be relevant and naturally connected with educational activity.

The performed research team mainly determines the success and fruitfulness of the research. The team should include not only teachers and postgraduates but also students.

The significant problem of successful research teams' forming is limited awareness of the lines of investigation of the researchers. Often, due to restricted access to scientific journals and personal communications problems, there are happened a situation when a few employees in one university are working independently of each other on one topic.

Research teams are formed mechanically based on particular structural units (such as departments). Also, the evident problem is the identification of potential research topics for students, which causes involvement them into working on the boring topic.

The rapid progress of Internet enables researchers to broaden their access to scientific articles, research results, and conference proceedings. Most publishing houses duplicate texts on their websites. Also, electronic archives of articles, where anyone can upload own publications, are widespread. However, the rapid growth of information quantity on Internet has created significant difficulties with its processing and analysis. It causes the actuality of the task of automated extraction and analysis of information about scientific publications to support the decision on research groups forming.

The purpose of the work is the choice of the effective algorithms for scientific papers analysis system. We focus attention on automated extraction of information from scientific articles and information analysis to identify research areas and groups of researchers with similar interests. One university or faculty limit the scale of the software system.

Information Extraction. The developed information system should convert the source text of articles into structured data to support the automated analysis. In our case, it is necessary to extract from the text of origin authors' names, article title, year of publication, and a set of keywords.

Scientific articles usually have a similar structure; they contain abstracts, author affiliations, a set of keywords, main text, conclusions, and a list of references. Unfortunately, publishers do not establish a common standard template for articles. Each scientific edition usually provides its template, instructions, and requirements on the language of writing [1]. In many articles abstracts, author affiliations, keywords, and references are written in several languages. Listed conditions greatly complicate the task of necessary information extraction.

Various visual information that people can perceive (such as the font style, text layout, etc.) often accompany different data formats. Many extraction algorithms require prior conversion of source text of articles into text with a markup of visual information. However, if the analysed documents do not have such information, the quality of results produced by extraction algorithms is significantly reduced.

An additional difficulty for the extraction algorithm is the necessity of analysing the texts in different languages, in our case in English, Russian and Ukrainian.

There are many algorithms for solving information extraction problems by using neural networks [2] or hidden Markov models [3]. The percentage of errors in such algorithms is quite low, but they have several disadvantages.

1. The algorithms require a significant amount of labeled articles for training. Accordingly, the training takes a lot of time. At the same time, it is hard to control and is highly dependent on size and quality of the sample.

2. Most algorithms are focused on the English language only. If the algorithm supports multiple languages, it becomes much more complicated, especially in the case of information extraction from texts in English and Slavic languages.

3. Now, neural networks do not cope well with the entities merge. The neural network can understand that the word indicates the name in the text, but they are not able to combine this name in one entity in case it is repeated in the article. For example, in case the name is written in several languages in one article.

These disadvantages caused the development of own algorithm consisted of four steps:

1) lexical analysis;
2) terminals normalization;
3) entities combining;
4) entities filtering.

Lexical analysis is the process of analytical parsing of the input symbol sequence to obtain the output sequence of typed atomic units (terminals).

To develop a list of rules defined the transformation of article text to the terminals we used a visual analysis of the available scientific articles.

The rules should not conflict with each other. To eliminate one of these conflicts, we

decided to recognize the last names with initials only as the names at lexical analysis level. Full names that content of the first, middle and last name, are challenging for identification without a dictionary because there are often placed more than two consecutive words begun with a capital letter in the articles. For example, names of universities usually consist of a few words written in Title Case style. It is possible to search the full name of authors in the text after identifying their last name and initials.

The terminals may have various spellings styles. For example, the terminal described the last name may be written with a first capital letter, or with all capital letters. It immensely complicates the further work with the entities. Because of it, we have to do normalization, i.e. to reduce all data to a common standard form.

It is common practice to use transliteration for the proper names instead of its translation into another language. For most languages, there are provided transliteration standards to transcribe the words with the Latin alphabet. We propose to use the transliteration to transfer the entities into space for comparison. Such solution is suitable for comparison of authors' names in various languages. To combine keywords or other entities, we recommend using transliteration as well as translator or training base of vector space model of words [4].

After combining entities, we have to filter them. For example, we have to extract the names of article authors. The names placed in the section for authors at the approximately same distance between them will be considered as names of the article authors. Also, the repeated references to the names in any language will be recognized as the author's name and combined with the respective entity. All other names in the text of the article will be discarded.

The result of the algorithm is a list of entities extracted from the text of the article, with links to their use in the text.

Algorithms for Analysis of Connections

Entities derived from articles organize the data system with the following elements: authors, article names, and keywords. The primary objectives of analysing science articles are finding groups of authors that often publish together and finding the key research areas in the given set of materials.

Therefore, in this case, we are interested in the following relations in described data system:

1) a common article of two authors (in case the article has more than two authors we consider all possible pairs);

2) using the keyword by the author;

3) using two keywords in one article (in case the article has more than two keywords we count possible pairs).

We propose to examine the problem of finding groups of authors and groups of keywords as the clustering problem. The clustering criteria for finding author groups are the number of joint articles. The corresponding rule for keyword groups is the number of articles where these words are used simultaneously.

The classical clustering algorithm is an iterative *k*-means algorithm that creates a predetermined number of clusters [5]. This algorithm is based on a minimization of the square error deviation of the object from the centre of mass of the cluster to which it belongs in the current iteration. The drawbacks of this algorithm are predefined the number of clusters and difficulties with selecting first centres of mass of the clusters.

To avoid specifying the number of clusters, we can use an algorithm based on the graph theory. In this case, we must represent analysing data as graphs of two types.

The graph of the first type is a weighted graph of authors' collaboration $G_i=(V_i,E_i)$. The set of vertices matches the set of authors. Two vertices are connected by an edge if authors have a joint article. The weight of an edge is defined as the number of articles that were written by these authors together.

The graph of the second type is a semantic graph of articles $G_a=(V_a,E_a)$. The set of vertices matches the set of keywords. Two vertices are connected by an edge if these keywords are present in the same article. The weight of an edge is defined as the number of articles where two keywords present simultaneously.

The most common clustering algorithms based on the graph theory are the following [6-9]:

1. Allocation of connected components. The algorithm relies on removing edges with weight less than some defined coefficient R from the graph. After this graph splits into several connected components representing clusters.

2. Minimum spanning tree. In the initial graph, the minimum spanning tree is built. Then, as in algorithm for allocating connected components, we remove edges with weight less than the defined coefficient *R*. If we remove edges iteratively, from weak to strong, we will build the cluster hierarchy.

It is worth noting that in the second algorithm we build the minimal spanning tree, and as we want to obtain clusters with the maximum sum of edges weight, we must change the weights of edges in the appropriate way.

For graphs clustering the algorithms based on stochastic analysis are often used. One of such algorithms is MCL (Markov Clustering aLgorithm). In this algorithm, we build the normalized by columns adjacency matrix A of graph G and iteratively inflate it until matrix does not become stable [10]. An inflation operator Γ_r with power coefficient r is defined as

$$\left(\Gamma_r A\right)_{pq} = \left(A_{pq}\right)^r \Big/ \sum_{i=1}^{k} \left(A_{iq}\right)^r ,$$

where k is a number of vertices in the analysed graph.

The matrix is stable if applying the inflating operator to the matrix does not change it. For finding clusters, we must analyse the stable matrix row by row. If the row has a non-zero element, the numbers of columns that contain these non-zero elements determine the number of vertices that form a cluster.

It should be noted that MCL is a fuzzy clustering algorithm because sometimes a vertex can simultaneously belong to several clusters. This property is especially valuable in the case of interdisciplinary research.

Properties of graph clustering algorithm for finding groups of co-authors and research directions on the defined set of scientific articles were analyzed on a test set of articles. Clustering algorithms based on graph theory and MCL algorithm showed the same result for allocating clusters, but the algorithm based on the minimum spanning tree required less computational costs.

Conclusion. The software system for the analysis of scientific publications is developed to identify research areas and research groups within one university or department. Due to the scale of the problem, existing algorithms for text mining and dependency analysis in scientific networks are redundant. In the paper, there are examined algorithms that are sufficient to meet the challenges of the developed system. In particular, the algorithm for extracting information from scientific articles designed to extract only the necessary information rather than parse all articles. At the same time, proposed algorithm allows working with articles in several languages. To identify research areas and research groups we considered clustering algorithms based on the graph theory for solving problems of the developed system. Experimental results on a test articles sample showed the advisability of using an algorithm based on the minimal spanning tree.

References

1. Sarawagi S. Information Extraction, (2007), *Foundations and Trends in Databases*, Vol. 1, No. 3, pp. 261 – 377.

2. Sarkar K., Nasipuri M., and Ghose S. (2010), A New Approach to Keyphrase Extraction Using Neural Networks *International Journal of Computer Science Issues*, Vol. 7, Issue 2, No. 3, pp. 16 – 25.

3. Skounakis M., Craven M., and Soumya R., (2003), Hierarchical Hidden Markov Models for Information Extraction, *IJCAI'03 Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 427 – 433.

4. Zou W.Y., Socher R., Cer D.M., and Manning C.D., (2013), Bilingual Word Embeddings for Phrase-Based Machine Translation, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1393 – 1398.

5. Kanungo T., Mount D.M., Netanyahu N. S., Piatko C.D., Silverman R., and Wu A.Y., (2002), An Efficient *k*-means Clustering Algorithm: Analysis and Implementation, *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, Issue 7, pp. 881 – 892.

6. Tarjan R.E., (1983), An Improved Algorithm for Hierarchical Clustering using Strong Components, *Information Processing Letters*, Vol. 17, Issue 1, pp. 37 – 41.

7. Flakea G.W., Tarjan R.E., Tsioutsiouliklisc K., (2004), Graph Clustering and Minimum Cut Trees, *Internet Mathematics*, Vol. 1, Issue 4, pp. 385 – 408.

8. Schaeffer S.E., (2007), Graph Clustering, *Computer Science Review*, Vol. 1, Issue 1, pp. 27 – 64.

9. Zhou Y., Cheng H., and Yu J.X., (2009), Graph Clustering Based on Structural/attribute Similarities, *Proceedings of the VLDB Endowment*, Vol. 2, Issue 1, pp. 718 – 729.

10. Brandes U., Gaertler M., and Wagner D., (2003), Experiments on Graph Clustering Algorithms, *Lecture Notes in Computer Science*, Vol. 2832, pp. 568 – 579.

Received 05.06.2016

Berlizov
Yevgen, Student of Odessa National Polytechnic University,
tel. (091) 988 21 44.
E-mail: berlizov@me.com

Hapiak
Viktor, Student of Odessa National Polytechnic University,
tel. (063) 415 00 70.
E-mail: viktor.hapiak@gmail.com

Liubchenko
Vira, D.Sc., Assoc. Prof.
Head of System Software Department, Odessa National Polytechnic University,
tel. (048) 705 8675,
E-mail: lvv@edu.opu.ua