Analysis of novelty of a scientific text as a basis for assessment of efficiency of scientific activities

Andrei Dynich and Yanzhang Wang
Faculty of Management and Economics, Dalian University of Technology, Dalian, China

Abstract

Purpose – The purpose of this paper is to complement an available system of qualitative analysis of efficiency of scientific activities with assessment of novelty of a subject of research that gives a more complete pattern for evaluating the efficiency of efforts of both scientists and research teams.
Design/methodology/approach – The approach is based on detection of specified linguistic patterns with further evaluation of similarity and novelty scores of obtained definitions at the sentence level.
Findings – This work presents an algorithm of automatic search for a new subject of research in scientific papers on the basis of statistical and linguistic analyses of description of new terms. Application of patterns specified in a given manuscript with further utilization of well-known methods of similarity and novelty detection scores makes it possible to evaluate the degree of novelty of a subject of research.
Practical implications – As a practical application of the proposed algorithm, the algorithm of determination of authority of a scientist will facilitate assessment of personal contributions of certain authors made in a certain field of study.
Originality/value – The main contribution of a given manuscript is in application of linguistic patterns recognition and calculation of similarity and novelty scores to the area of scientific results with further proposition of the method of automatic search for a new subject of research in scientific manuscripts.

Keywords Recognition, Assessment, Linguistic patterns, Python, Scientific novelty

Paper type Research paper

1. Introduction

A huge amount of data available for processing became the distinct feature of the Twenty-first century in different areas of human activities. The world of scientific publications is not an exception. The number of manuscripts published in scientific journals is increasing every year. According to Mark Ware and Michele Wabe, the number of peer reviewed journals published annually has been growing at a very steady rate of about 3.5 percent per year for over three centuries (Ware and Mabe, 2009). While in the recent period from 2002 to 2012, this rate has been slowing down to 2.5 percent per year (Ware and Mabe, 2015).

At present, in the scientific community, activities of a scientist are assessed on the basis of quantitative parameters and qualitative peer-to-peer reviews. According to the results of research demonstrated by Björk *et al.* (2009), the ISI (Web of Science) covered 8,466 scientific journals and the number that was non-covered by the ISI journals was 15,284. As to the number of papers covered by the ISI, it was 70 percent to 30 percent uncovered. The efforts of researchers get published in 64 percent of journals, i.e., 30 percent of papers were still out of proper attention. The quotability of works of a scientist and efficiency of labor are evaluated on the basis of databases such as ISI, Scopus, Google Scholar, and others. There is a big gap between scientific documents (written in English) found on a public web in total and covered by databases like Web of Science or Microsoft Academic Search (Khabsa and Giles, 2014; Orduña-Malea *et al.*, 2014).

Thus, the main part of activities of a scientist goes beyond the scope of labor evaluated on the basis of traditional criteria while the part of work that goes through such kind of assessment allows us to judge about the results of a scientist or a scientific school only partially.

Journal of Organizational Change Management
Vol. 30 No. 5, 2017
pp. 668-682
© Emerald Publishing Limited
0953-4814
DOI 10.1108/JOCM-10-2016-0226

Drawbacks of available approaches from the point of view of assessment of efficiency of professional activities of a scientist or a group of scientists are well-known. In order to assess results of research activities of scientific groups and organizations that operate in a field of fundamental research, the world scientific community started using various objective bibliometric parameters, such as the number of published works in top journals, the overall number of citations, the impact factor of a journal, the maximum quantity of references made to the certain work, and so on.

Quality assessment of the objective value of every single work merely guided by the publication criterion seems to us as impossible to be realized and not really suitable for judgment about the real productivity of a scientist. None of the indices can shed light on semantics or on the nature of a citation. Systems of indices cannot distinguish between a stocked reference and a detailed reference, between mention and development, i.e., consideration of an idea in essence. In addition, systems of indices are not able to recognize informative discussion from bare denial. By a citation index, authors of pseudo-scientific works can easily overmatch reputable scientists as their works are cited quite often also, but, as a rule, for disputing or denying, trying to point out mistakes.

Practices of the forced growth of a citation index and refusal to make a reference of a certain scientist due to non-scientific reasons, such as offence or envy, exist in the world for a long time. An impact factor or prestige of a journal can provide us with some but not exhaustive information about the quality of publication. Appearance of a manuscript in a prestigious journal with a high citation index does not always indicate novelty and the scientific capacity of the manuscript. A great number of manuscripts are still out of attention of the indexing system, since they are not published in high-ranked journals and (or) written in national languages. Quantitative parameters of assessment do not take into account the subject of research and the nature of specialization: humanists and technicians are assessed with the use of the same *cliché*.

Meaningfully, scientific works are evaluated by reviewers after submission of the manuscript to the journal. Quality of such examination depends on a reviewer's awareness of the status quo in the relevant field of science. However, on average, a faculty member is able to "process" approximately 250 scientific papers per year (Tenopir and King, 2007) and become familiar with the negligibly small number of journals. One can object, "The most influential and cited manuscripts are published in the "gold standard" journals (WoS) and we should study them first." However, as everyone knows, not every manuscript is available for free perusal. Open access includes only the part of works and parts of works. In that way the certain portion of results of scientific studies is out of sight of scientists and managers from science for various reasons. Therefore, it is objectively difficult to find the answers to questions that are extremely important for further development of science: which directions in science should be supported? How to appraise the capacity of a young researcher? Who is able to be an expert in a relevant field? and so on.

Obviously, a universal criterion for assessment and comparison of all kinds of products of research should reflect something essential and intrinsic to any product of scientific labor. To this requirement meets only novel scientific information, retrieval of which from the object of research and further creative processing are the main and direct aims of science. If it could be managed to find formalized ways of evaluation of contained in each scientific report logically arranged and structured information that these ways might act as the basis for creation of universal and more adequate criteria of significance of scientific works. These techniques could help scientists, reviewers, and managers to automatize retrieval of scientific information, to provide an analysis under conditions of incomplete data, and to make an assessment of publications more objective.

For the beginning, it seems natural to apply methods that have been already used for automatization of search and analysis of novel information, but with taking into account specificity of formation and purpose of scientific texts.

At present, metadata that provide information about the title, the author, and the list of cited sources are retrieved with application of automatized analysis from the available databases of scientific documents, such as Google Scholar, CiteSeerX, Microsoft Academic Search, ResearchGate, and others. Assessment of scientific works is realized with the use of data that include, for example, the web size – the number of pages discovered by search machines, notability – the number of external links to the web page, the number of uploaded documents (pdf, doc, ppt, or ps), the number of published works, and citation indices. Key words, annotations, conclusions, and references are drawn into automatized search. This information, as usual, comes into the relevant bases in the form of manuscripts or parts of them uploaded to the websites of journals, scientific institutes, and libraries. Information retrieval is realized in a way similar to processing the search system query. However, compared with the retrieval of news and facts, search for scientific information is simplified due to the inherent in scientific manuscripts' internal structure and accepted forms of formatting.

The structure of a scientific work is maintained, among other things, with use of various styles of formatting. Variation of styles of formatting can help with division of the manuscript into units. Splitting the manuscript into units facilitates the procedure of retrieval of desirable information. By ignoring footnotes and the italic type inside the sentence, an abrupt change of the style will indicate the beginning of a new section. A unit of a small size and a bigger font size most likely will be the heading of a text of a big size and a smaller font size. Information about the author is located in the unit that starts with the biggest font. Contact information about the author contains the at sign in the e-mail address. Quite often, an e-mail address itself contains the name or/and the surname of the author and the domain name correspondent to the place of the author's work. In that way, with retrieval of information about authors, it is possible to avoid a solution of a task of recognition of named entities.

Annotations and key words are the units that start with relevant words or expressions. The reference unit, as usual, is named in a specific way – references, the list of used literature, and so on. Sources of literature used in the manuscript are usually arranged according to several well-known ways of formatting that makes it possible to employ machine learning for analysis.

There are techniques such as citation analysis and lexical clustering that allow us to find semantically close papers and discover researchers with close scientific interests.

As it is easy to notice, the above-mentioned methods are aimed at acquisition of information that acts as the answer to the strict question. Therefore, for example, this information can be employed (turned into knowledge) by researchers who are already well navigated in the field of the requested information.

Information retrieval is still a complicated task not only for young researchers but also for scientists who work in related areas. In this case, it is necessary to answer the questions that do not have exact answers: what is new happening in a certain area? What kind of tasks are solved? What methods are applied and what the advantages of these methods are? Who is the author of the main concepts? and so on. Therefore, it is necessary to improve the procedure of information retrieval by means of introduction of semantic analysis before formulation of the search query. The author of a scientific work pays special attention to substantiation of applied methods and approaches, to discussion of applicability of obtained results, and so on. Therefore, such patterns as "thus, so, in that way," "let us note that," and other patterns that contribute to simplification of information retrieval take place quite often.

The main value of a scientific manuscript is novelty. Raymond R. Tan in his article draws our attention to scientific novelty saying, "It is based on the fundamental principle that the value of one's research lies not in the degree of effort one puts into it; rather, the value of research lies in the novelty of the results." (www.philstar.com/science-and-technology/2014/04/24/1315251/declaration-novelty-scientific-journal-articles). Whereas some qualitative parameters of assessment are also significant, among which assessment of novelty of a subject of research seems to us as the primary task. Indeed, an analysis of a subject of research will make it possible to compare results of labor of a scientist with results of others in the same field, fix appearance of a new subject of research at an early stage and highlight on this basis promising from the view of a certain scientific institution research works, handle questions related to the priority, and others.

2. Related works

What is "novelty"? A precise definition of novelty detection is hard to arrive at, nor is it possible to suggest what an "optimal" method of novelty detection would be (Marsland, 2002).

Currently, the term "novelty" is comprehended quite different from understanding of this term as scientific novelty that is a subject of research of this manuscript. Novelty detection is the technique used to extract novel information from a set of relevant documents or from the same document in a given topic (query) (Sendhilkumar *et al.*, 2013).

The review of works aimed at novelty detection makes it possible to state that in a majority of works detection of novelty is understood as detection of something that has not been seen before or something that is quite different in comparison with previous documents, sentences, phrases, and objects.

In general, the process of detecting a novel text contains three main steps: preprocessing, categorization, and novelty mining (Zhang and Tsai, 2009a). The stage of preprocessing of text documents/sentences includes processes of taking away stop words, a word stemming, a part of speech tagging, etc. Categorization puts each document or (and) sentence into its relevant topic directory. However, the processes of categorization can be skipped for simplification of the experiment when the search for relevant documents has been already done under the certain topic. Then, within a certain category of relevant documents/sentences, time-sequenced documents or (and) sentences are processed and the "novel" by time information is retrieved. Novelty mining has been performed at three different levels: event, sentence, and document levels (Li and Croft, 2005).

Works on novelty mining at the event level are originated from the Topic Detection and Tracking research focused on the online new event detection/first story detection (Allan *et al.*, 1998; Stokes and Carthy, 2001; Spitters and Kraaij, 2001; Yang *et al.*, 2002; Brants *et al.*, 2003; Gabrilovich *et al.*, 2004; to a list of few). Several models, such as the vector space model, language models, lexical chains, and others are usually employed in order to represent new stories/documents. Then documents are grouped into clusters by force of calculating statistic-based or semantic-based similarities. A story that starts a new cluster is marked as the first story about a new topic or it will be marked as an "old" one if there exists a novelty threshold which is smaller than the similarity threshold and the similarity score between the story and its closest cluster is greater than the novelty threshold.

Research works focused on the sentence-level novelty mining aim to find relevant and novel sentences in a stream of documents/sentences. Harman and Soboroff provide us with overviews of TREC novelty tracks, the basic task of which is to detect relevant and novelty sentences (Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004). Novelty of sentences is usually calculated with respect to the number of new words appearing in them. Novelty detection at the sentence level can be conducted mainly in two sequential steps: the relevant sentence retrieval and the novel sentence extraction. Usually, a high similarity score between a sentence and a given query will increase the relevance rank of the sentence, whereas a high similarity score between the sentence and all previously seen sentences will decrease the novelty ranking of the sentence. Research works show that techniques of the sentence- and document-level novelty mining are quite well applicable for the wide range of textual documents and streams: WWW like news (Fu *et al.*, 2015), various internet articles and stories (Breja, 2015), and business blogs (Tsai and Chan, 2010). Some attempts take place to apply novelty mining techniques for documents written not only in English language (Wayne, 2000; Zhang and Tsai, 2009a).

Similarity metrics that can be used for detecting a novel text are the overlap measurement (Zhang *et al.*, 2002; Zhao *et al.*, 2006), the KL distance (Zhang *et al.*, 2002), cosine similarity (Breja, 2015), new words count measure, etc. Some works are statistically oriented using purely statistical analysis, such as Latent Semantic Indexing (Deerwester *et al.*, 1990) and the Hyperspace Analog to Language model (Burgess *et al.*, 1998). Other works utilize ontological knowledge, such as the WordNet database (Fellbaum, 1998; Resnik, 1999; Finkelstein *et al.*, 2002; Schiffman *et al.*, 2002), thesaurus, and others.

To our surprise, the search for research works focused on importance and evaluation of novelty in scientific manuscripts showed us the strong deficiency of such kind of works. One of few attempts in this direction is the Luzon Marco's research work (2000) in which she investigated the construction of novelty in manuscripts from the field of computer science.

3. An algorithm of analysis of a subject of research

3.1 Methodology

Methodology of postmodernism has been shown up in the inclusion of symbolism in determination of the value of a scientific work when the place of carrying out the research or places of publication of results on their own testify the quality of research.

Further development of assessment of research activities is natural to expect in the direction which is typical for advancement of science in the modern world. In modern science, the "open" rationality is opposed to the "close" in-paradigm rationality when a researcher is moving forward within the scopes of the close conceptual frame. As opposed to the "close" rationality, the "open" rationality assumes development (motion) of an idea not only in the frames of the preselected paradigm but also includes the possibility to compare obtained results with those results that were already produced under various cultural traditions. This possibility gives rise to conditions for discovery of points of intersection in diversity of opinions and disciplines, not dissolution of something in another thing and cancellation of unique differences, but establishment of agreement and achievement of understanding among various approaches.

For estimation of the direction of improvement of indicators of assessment, let us focus our attention on the fact that, first of all, quantitative criteria of assessment appeared in a period when the value of science was determined from the attitude of positivism. The basis of assessment in positivism became quantitative methods, primarily statistical ones. Naturally, in this paradigm are solved those tasks that assume statistical analysis.

The research article can be represented as a collection of words. A document can be treated as a random mixture of words with some probabilistic degrees of distribution of them or by using a vector space model, i.e., by calculating weights of words or sentences.

The idea behind the novelty detection approach proposed in a given work is as follows. Employed in this work, sentence-level novelty detection is estimated based on the distance between the pairwise comparison of relevant sentences. Comparison is done according to the cosine similarity metric:

$$\mathrm{Sim}\left(S_k, S_p\right) = \frac{\sum_{i=1}^{m} w_i(S_k)\cdot w_i\left(S_p\right)}{\sqrt{\sum_{i=1}^{m}\left(S_{k,i}\right)^2}\cdot\sqrt{\sum_{j=1}^{x}\left(S_{p,j}\right)^2}}, \qquad (1)$$

where Sim(S_k, S_p) is the numerical result of comparison of the sentence k with the sentence p, $w_i(S_k)$ is the weight of the ith word in the sentence S_k, and $w_i(S_p)$ is the weight of the ith word in the sentence S_p. The denominator is multiplication of the sum of words' weights of the sentence S_k by the sum of word's weights of the sentence S_p.

Weights of words are calculated according to the vector space model:

$$w_i = \log_{10}\left(1+tf_{i,s}\right)\cdot\log_{10}\left(N/df_i\right), \tag{2}$$

where w_i is the weight of the ith word; $tf_{i,s}$ is the number of repetitions of the ith word in the sentence S; i.e., in the sentence which contains definition; N is the overall number of sentences which contain definitions; and df_i is the number of definitions where the ith word occurs.

The novelty score, as the inverse of cosine similarity (Zhang and Tsai, 2009b) is calculated in the following way:

$$N\left(S_k, S_p\right) = 1 - \mathrm{Sim}\left(S_k, S_p\right), \tag{3}$$

where $N(S_k, S_p)$ is the symmetrical novelty score of comparison of k and p sentences.

Novelty scores are calculated for all pairs of sentences in a similar way. Then takes place summation of all comparative novelty scores for all sentences in order to rank them.

Novelty of a scientific work may be found in one of three following points (Figure 1): a new category with description of connections to other categories; a new subject of research with description of relations with other subjects and with description of properties of a given subject; and new properties of a subject with description of them and with establishment of connections to already discovered properties.

The given work concerns a central part of cognitive activities of a scientist – introduction of a new subject. Exactly this "degree" of novelty is typical to the key works of key scientists and exactly on the basis of existence of the "own" subject colleagues interpret certain researchers as "influential scientists." Introduction of a new category in science takes place relatively rare, and these cases do not need for special search. Discovery of new characteristics of defined earlier subjects is a "standard" of scientific activities. That is, in what researchers are engaged from their diploma works and further.

Introduction of a new subject in the scientific turnover is implemented by introduction of a definition which specifies a subject constructively in relation to the content and in a system form. In other words, a published work should contain an author's definition with description of a method of discovery (acquisition, development) of a subject and with description of a place of this subject in a system of scientific concepts. Such definition, usually, does not appear in the first published work on a certain topic while the history of formation of a subject can be traced retrospectively in series of publications.

Definitions take different forms. Their forms and application at a large degree depend on a certain language. Syntactic and contextual definitions are more frequent in scientific texts due to variety of their forms. Different patterns are typical for various types of definitions (Table I).

In a text, definitions can be recognized by their grammatical patterns (Table II).

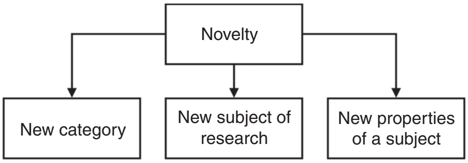

Figure 1.
Degrees of novelty of a scientific work

Table I.
Types of definitions

1. Definitions with taking into account the content of the concept	Description of physical objects discovers their physical parameters
	Description of an abstract concept discovers its essence by reduction to those more specific characteristics which are generalized by a given concept
	Description of an action/process discovers its main operations/stages, the producer and the object of actions, the required result if such result flows out from the essence of action, time for realization in a situation when time is a matter of principal importance and could be set
	In some cases, description of one of the essential characteristics of the studied phenomenon might be needed. Description of an essential characteristic discovers it by reduction to the aggregation of elementary characteristics
	Description of genus-species concepts through their functions stipulates indication a common characteristic for all types related to a given kind. The nature of the characteristic is determined by the essence of the concept
	Description of elements, tools, means, methods, etc., reveals their essential characteristics and indicates purpose
2. By the method of expansion of a concept	A “classical definition” that explains “what is what”
	Definition that contains a reference to collective opinion
	Author’s definition
	Functional definition
	Definition-description
	Definition-enumeration
	Definition-comparison
	Definition in a form of interpretation of a word
3. By the volume of definitions	Definition in the scope of the whole sentence
	Definition in the scope of a part of a sentence
	Definition in the scope of two and more sentences

3.2 The proposed algorithm

A corpus of scientific texts usually passes through the word stemming and the POS tagging at the stage of preprocessing of the entire document. However, to the contrast of the majority of works aimed at novelty detection, the word stemming is implemented at the stage of comparison of obtained sentences which contain definitions. The POS tagging also can be implemented at the stage when all sentences with definitions are acquired.

The order of words has been kept both at the stage of detection of definitions and at the stage of comparison of sentences which contain definitions. A step of categorization of documents is skipped as scientific manuscripts are selected under the same topic. The algorithm of novelty detection is proposed to be applied to every document where there is a need for evaluation of such degree of scientific novelty.

The proposed algorithm of processing of a scientific text for the purpose of novelty detection (Figure 2) includes the following steps:

(1) generation of the corpus of scientific texts downloaded from the scientific databases;

(2) input of data by insertion of the downloaded text into the string variable keeping the words order; the text is spitted into sentences with use of the function *sent_tokenize()* (Table II, line 4);

(3) programming patterns of definitions specified in the Table II;

(4) retrieval of definitions from the corpus of scientific texts (Table III, lines 7-11);

(5) assigning weights to all words of sentences which contain definitions according to the Formula 2 (punctuation is eliminated at this stage, all capital letters are substituted to lower-case, words stemming is performed);

1.	With the meaning of "qualification of the subject and the method of explanation"	Who-what is who-what? What is what? What is called by what? Who-what is called by who-what? What represents what? What is understood as what? What is considered as what? What is taken for what? What acts as what? What is included in what? What is determined by what? What can be determined as what? What we determine as what? What is manifested in what? What occurs when something happens? What is characterized by what? What is named by what?
2.	With the meaning of "classification of a subject and methods of expression of classification"	For indication of division of subjects and processes into groups What is divided into what What is subdivided into what For indication of relations of subjects, processes, and the certain group What belongs to what What is attributed to what What composes what For indication of the structure of subjects and processes What consists of what What contains what What is included in what What is in the structure of what

Source: Kuznetsova *et al.* (2013)

Table II. The most generally used grammatical patterns of definitions

(6) pairwise comparison of sentences which contain definitions and calculation of their scores of similarity with use of the Formula 1;

(7) calculation of the novelty score for every element of the similarity matrix (the matrix that contains similarity scores of sentences-definitions) according to the Formula 3;

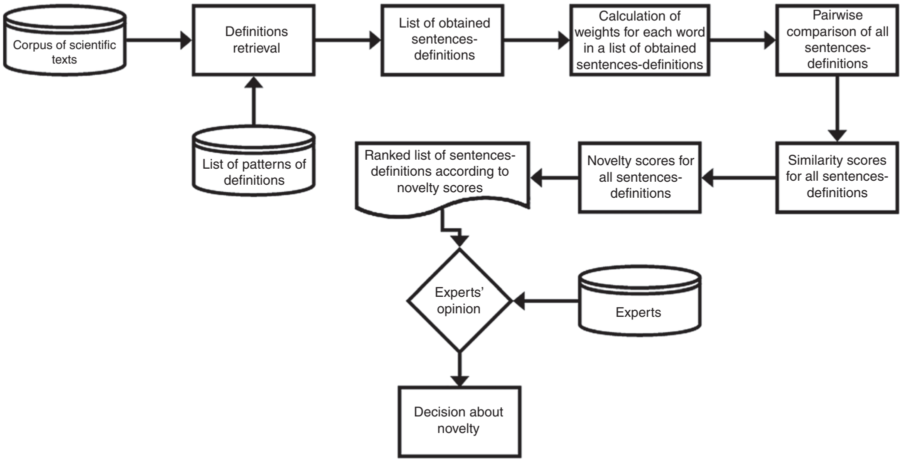
Corpus of scientific texts
Definitions retrieval
List of obtained sentences-definitions
Calculation of weights for each word in a list of obtained sentences-definitions
Pairwise comparison of all sentences-definitions
List of patterns of definitions
Ranked list of sentences-definitions according to novelty scores
Novelty scores for all sentences-definitions
Similarity scores for all sentences-definitions
Experts' opinion
Experts
Decision about novelty

Figure 2. An algorithm of processing of a scientific text for the purpose of novelty detection

(8) calculation of the overall novelty score for every sentence-definition by summing up novelty scores obtained from comparison of a certain sentence-definition with all others;

(9) rank the overall novelty scores of sentences; and

(10) display definitions in ascending/descending orders of novelty scores or save them into the file for further analysis that will be performed by experts.

Three key parts of the proposed algorithm of identification of novelty in a scientific work are as follows: detection of specified patterns of definitions in the corpus of texts; pairwise comparison of all sentences which contain definitions; and judgment about the degree of novelty of sentences-definitions by ranking the sentences according to novelty scores and with taking into account decisions made by experts.

4. Results of the experiment

Analysis of the content of a scientific manuscript for the purpose of novelty detection with use of computational processing assumes availability of specific software tools. Recently, for the purpose of natural language processing (NLP), the high-level general-purpose programming language Python has been actively employed, which is different from many other programming languages by its high flexibility and dynamism. In order to solve problems formulated in a given manuscript, the programming language Python in combination with the Natural Language Toolkit (NLTK) is applied. The NLTK is a suite of libraries and programs for symbolic and statistical NLP for English.

Let us put the problem of automatic detection of sentences in scientific manuscripts which contain definitions of concepts that have been formulated earlier or, what are more valuable, totally new definitions. Detection of lexical patterns frequently used in scientific manuscripts is applied for this purpose. An example given in this manuscript is based on the pattern "be (is/are) characterized by." The process of retrieval of a specified pattern from a scientific text is implemented by the following program (Table III).

Description of operation of the program and demonstration of its capabilities for patterns retrieval are provided as follows: the first step is loading the library; then open the text file of the manuscript text.txt located in the local disk; load the whole text of the manuscript in the string variable *var2*; the process of text tokenizing when the text is divided into certain parts – tokens is implemented; in our case, it seems as proper to split the text into sentences with use of the standard function *sent_tokenize()* from the library *nltk*; a list of sentences is inserted in the variable *varSent*; a loop "for" is used in order to take into account every

Table III. The program code for retrieval of the lexical pattern "be (is, are) characterized by" in the natural language mode written in the programming language Python with use of the NLTK

```
Line    Program body
1       import nltk
2       var1 = open("d:/text.txt")
3       var2 = var1.read()
4       varSent = nltk.sent_tokenize(var2)
5       for i in range(0, len(varSent)):
6         var3 = nltk.word_tokenize(varSent[i])
7         if len(var3) > 3:
8           for k in range(0, len(var3)-3):
9             if (var3[k] = = "be") or (var3[k] = = "is") or (var3[k] = = "are"):
10              if var3[k+1] = = "characterized":
11                if var3[k+2] = = "by":
12                  print (varSent[i])
13      var1.close()
```

sentence of the whole text; perform tokenization of the text into separate words with use of the standard function *word_tokenize()* from the library *nltk*; lines 7-11 stand for retrieval of the specified pattern "be (is, are) characterized by" in every sentence; the sentence with the specified patter is displayed; and at the end of the session, the file is closed.

As an example, a given algorithm has been applied to two scientific manuscripts: "Utilization efficiency of spherical metal nanoparticles that increase light absorption in absorbing media" (Dynich, 2011); and "Energy breathing of nanoparticles" (Dynich, 2015). At first, texts of manuscripts were converted from the pdf format into the txt format. Then the above-descripted program was applied to process them. The result is shown in the Table IV.

The obtained result shows that detection of a specified lexical pattern is realized absolutely precisely. The proposed program detected all of entries of a given lexical pattern. A small drawback of a program is in incorrect tokenization of the text, i.e., in division of the text into sentences. This took place due to some uncertainties in operation of a function *sent_tokenize()*.

5. Evaluation of an impact exerted by a certain work of a certain scientist on development of research of other scientists

A fact of publication in a scientific journal always reflects existence of influence of an author on the status quo in science and on scientists themselves, at least on a reviewer or on the editorial board of a journal.

In order to assess a promising degree, it seems to us as important to determine the level of influence that leads to alteration of notions of colleagues about the subject field or at least of one scientist who carries out research work and publishes results on one's own.

An algorithm (Figure 3) can be described as follows: detection of scientists-authors who introduced a new subject of research; evaluation of the number of scientists who

Manuscript number	The result of application of the program
1	The ratio of energy absorbed by the particle to the energy incident on the particle per unit time is characterized by absorption efficiency Qabs
2	Every point in space, which has been affected by an electromagnetic wave, can be characterized by particular values of the strength of the electric E and magnetic field H. For any time moment, we see quite a definite spatial distribution pattern of the electromagnetic field The energy flow density is characterized by the Pointing vector S. This directed quantity allows us to define not only the energy characteristic for every space point, but also the direction of the further energy transfer

Table IV. The result of application of the program (Table III) to scientific manuscripts

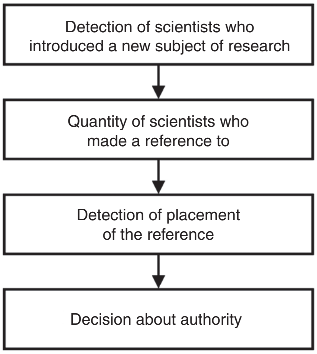
Detection of scientists who introduced a new subject of research
Quantity of scientists who made a reference to
Detection of placement of the reference
Decision about authority

Figure 3. An algorithm of determination of authority of a scientist

made a reference to the work in which a new subject was introduced; detection of a part of work in which the reference is situated, as placement of the reference in tasks or conclusions testifies to authority; and conclusion about authority of a scientist in the profile area or in other areas.

Then it is possible to assess influence and the level of authority of a scientist in the traditional style (using references) with taking into account not all of works of a given author, but only those works in which the definition of a new subject takes place.

Furthermore, it is possible to evaluate the scope of influence of an author. As the proposed procedure plans to record names of authors who made references, so in respect to them it is possible to implement retrieval of two types of relations. First, the proposed algorithm makes it possible to specify influence of an author exerted on members of a research team to which the author belongs. Second, it is possible to evaluate influence of an author on views of colleagues from adjacent or remote subject fields that will make it possible to discover existence of the interdisciplinary influence. In order to do this, works that contain references to the author's definition should be attributed to the certain subject field. It seems interesting to set the task of evaluation of an institutional impact of an author, i.e., an impact of an expert and other organizations engaged in certification of scientific works and in determination of scientific and technical policies.

To top it all, the proposed method makes it possible to evaluate a degree of credibility of a scientific organization – from the research institute and the university to the board of the fund by means of integrative assessment of influence of its members.

6. Reflecting upon opportunities

Proposing a new tool for assessment of results of scientific activities, we would like to rest shortly on possible prospects of adoption of such instrument. First of all, it is worth to be noted that the authors of a given manuscript do not see any special barriers from the point of view of technical implementation. The current level of modern informational technologies, and they keep on developing dynamically, makes it possible to progress in automatization of the process of assessment of novelty of scientific works.

A greater obstacle in a way of introduction of the proposed approach toward evaluation of labor of scientists could be the natural distrust of reviewers and all those who are in charge of organization of transfer of new scientific knowledge from a researcher to the professional scientific subculture. The well-established centuries-ago system of administration of science and assessment of results keeps on functioning. The persisting significant attention to bibliometric indices of evaluation of labor of a scientist clearly shows the credence given to this system nowadays. The authors of a given work expect that in these conditions the additional opportunities provided by the proposed modernization of evaluation of results of scientific labor will draw some attention. In this way, even nowadays a shortage of reviewers takes place against the background of the growing number of scientific journals and published works; therefore, the prospect of optimization of labor of enthusiasts who for the glory of science carry on the process of reviewing can facilitate the promotion of ideas expressed in this work.

At the same time, it is difficult to expect that the considered tool in this work in the very near future will "master the minds" of a wide range of researchers, although exactly with this aim, i.e. with self-organization of the system of the scientific community in the field of evaluation of its labor this tool is proposed for discussion.

The point is that the special feature of advancement of science at the current stage is a diversity of ways of manifestation of scientific ideas and results.

Scientific advancement in the Twenty-first century involves the formation of an open space for scientific publications. This includes the open access movement and online scientific infrastructure. These are hallmarks of Science 2.0.

Blogs, for example, help scientists to engage others in the discussion of their research results. Such discussions in blogs also encourage the growth of open expertise. With the use of various technologies for voting, many scientific resources on the internet can rank their publications to assess the degrees of novelty, interest, and urgency in their published works. After registration on the websites of e-content services like that found in arXiv.org, for instance, authors are allowed to submit their own manuscripts to the archives. The publication of an author's material can occur as soon as several hours after its submission. Authors can also update their own manuscripts by submitting corrected or more well-developed versions, and have the right to delete their publications if they think it is necessary. Services for the exchange of scientific video content afford scientists with opportunities to present visualizations that explain the essence of their work. Also possible is the development of reports about studies that not only will be understood by experts but also by the general public – in other words, research in translation. Social networks for scientific researchers and their research areas also make it possible to build chains of connections on the basis of professional interests and acquaintances. Every participant has an opportunity for professional growth due to the network effects of participation by others in their community. In addition, online exchange markets can operate in the scientific world not only for trading instruments and reagents but also as markets for knowledge and technologies.

Any open system demonstrates in its development two models of action which supersede each other. The first is a linear or an evolutionary phase when development of the system takes place in an already established mode presenting itself as relatively deterministic and predictable by its immediate past.

The second phase is associated with the fact that at the certain stage the condition of the system becomes unstable: residing, according to the definition of "openness" in a constant interaction with the surrounding environment, receiving its signals, and replying to them in a nonlinear mode.

The present system of evaluation of labor of scientists is based exactly on a linear model of dissemination and assessment of scientific knowledge. This model resembles the laminar quasi-stationary process: a researcher – a reviewer – the "inner circle" of scientists engaged in a similar scientific area – the "outer circle," and so on. Administration of science in the open stage assumes application of a model that will take into account a turbulent (nonlinear) mode of dissemination of scientific ideas.

Change of models, in its turn, is a complicated process and its analysis could act as a subject of a detached study. Yet some peculiarities of the course of such process could be predicted relying on the already developed models of operation of open systems. As an analog could act, for example, economic, physical, and other phenomena.

Employing the analogy with the physical process it is worth to suppose that for adoption of the proposed approach of quality assessment of results of labor of a scientist, first of all, the "viscosity," i.e. current interaction in a hierarchical structure of science, should be reduced. In our case, this means the acknowledgment of a potential value of a scientific idea regardless of how (in what way) or where this idea has been published. Second, not only "from top to down" conditions for quality assessment of scientific labor, i.e. a journal (a reviewer), heads of academic organizations, the professional scientific community, but also "from down to top" conditions should be provided. This means the possibility of assessment of scientific labor relying not only on competence and professionalism of an expert but also on immediacy and accuracy of incoming information to the researcher.

7. Conclusion

Authors of a given manuscript hope that this attempt initiates the direction of novelty detection in scientific articles, as novelty is a necessary attribute of any scientific work that stipulates the quality of results. The proposed algorithm of novelty detection on the basis of identification of definitions in the text of the scientific article makes it possible to analyze novelty brought by a certain author by comparing them with definitions of other authors. As a practical application of the proposed algorithm, the algorithm of determination of authority of a scientist will facilitate assessment of personal contributions of certain authors made in a certain field of study.

References

Allan, J., Papka, R. and Lavrenko, V. (1998), "On-line new event detection and tracking", in Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. (Eds), *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Melbourne, Australia*, ACM, New York, NY, pp. 37-45.

Björk, B.-C., Roos, A. and Lauri, M. (2009), "Scientific journal publishing: yearly volume and open access availability", *Information Research*, Vol. 14 No. 1, pp. 1-14.

Brants, T., Chen, F. and Farahat, A. (2003), "A system for new event detection", in Clarke, C.L.A., Cormack, G.V., Callan, J., Hawking, D. and Smeaton, A.F. (Eds), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Toronto, Canada*, ACM, New York, NY, pp. 330-337.

Breja, M. (2015), "A novel approach for novelty detection of web documents", *International Journal of Computer Science and Information Technologies*, Vol. 6 No. 5, pp. 4257-4262.

Burgess, C., Livesay, K. and Lund, K. (1998), "Explorations in context space: words, sentences, discourse", *Discourse Processes*, Vol. 25 Nos 2-3, pp. 211-257.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990), "Indexing by latent semantic analysis", *Journal of the American Society for Information Science*, Vol. 41 No. 6, pp. 391-407.

Dynich, R.A. (2011), "Utilization efficiency of spherical metal nanoparticles that increase light absorption in absorbing media", *Journal of the Optical Society of America A*, Vol. 28 No. 2, pp. 222-228.

Dynich, R.A. (2015), "Energy breathing of nanoparticles", *Journal of Nanoparticle Research*, Vol. 17 No. 6, pp. 1-10.

Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. and Ruppin, E. (2002), "Placing search in context: the concept revisited", *ACM Transactions on Information Systems*, Vol. 20 No. 1, pp. 116-131.

Fu, X., Ch'ng, E., Aickelin, U. and Zhang, L. (2015), "An improved system for sentence-level novelty detection in textual streams", *Proceedings of the IET International Conference on Smart and Sustainable City and Big Data (ICSSC 2015) in Shanghai, China*, IET, Stevenage, pp. 1-6.

Gabrilovich, E., Dumais, S. and Horvitz, E. (2004), "Newsjunkie: providing personalized newsfeeds via analysis of information novelty", in Feldman, S.I., Uretsky, M., Najork, M. and Wills, C.E. (Eds), *Proceedings of the 13th International Conference on World Wide Web in New York, NY*, ACM, New York, NY, pp. 482-490.

Harman, D. (2002), "Overview of the TREC 2002 novelty track", in Voorhees, E.M. and Buckland, L.P. (Eds), *Proceedings of the 11th Text Retrieval Conference, NIST Special Publication 500-251, in Gaithersburg, MD*, NIST, Gaithersburg, MD, pp. 46-55.

Khabsa, M. and Giles, C.L. (2014), "The number of scholarly documents on the public Web", *PLOS ONE*, Vol. 9 No. 5, pp. 1-6, doi: 10.1371/journal.pone.0093949.

Kuznetsova, J., Osipov, G. and Chudova, N. (2013), "Intellectual analysis of scientific publications and the current state of science", *Proceedings of the V.A. Trapeznikov Institute of Control Sciences of Russian Academy of Science, Big Systems Management, the Special Issue "Scientometrics and Assessment in Administration of Science*, No. 44, pp. 106-138 (in Russian).

Li, X. and Croft, W.B. (2005), "Novelty detection based on sentence level patterns", in Herzog, O., Schek, H.-J., Fuhr, N., Chowdhury, A. and Teiken, W. (Eds), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management in Bremen, Germany*, ACM, New York, NY, pp. 744-751.

Marco, M.J.L. (2000), "The construction of novelty in computer science papers", *Revista Alicantina de Estudios Ingleses*, Vol. 13, pp. 123-140.

Marsland, S. (2002), "Novelty detection in learning systems", *Neural Computing Surveys*, Vol. 3, pp. 1-39.

Orduña-Malea, E., Ayllón, J.M., Martín-Martín, A. and López-Cózar, E.D. (2014), "About the size of Google Scholar: playing the numbers", EC3 Working Papers No. 18, Granada, p. 23.

Resnik, P. (1999), "Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language", *Journal of Artificial Intelligence Research*, Vol. 11, pp. 95-130.

Schiffman, B., Nenkova, A. and McKeown, K. (2002), "Experiments in multidocument summarization", in Marcus, M. (Ed.), *Proceedings of the Second International Conference on Human Language Technology Research in San Diego, CA*, Morgan Kaufmann Publishers Inc., San Francisco, CA, pp. 52-58.

Sendhilkumar, S., Nandhini, N.S. and Mahalakshmi, G.S. (2013), "Novelty detection via topic modeling in research articles", *Computer Science & Information Technology (CS & IT)*, Vol. 3 No. 5, pp. 401-410.

Soboroff, I. (2004), "Overview of the TREC 2004 novelty track", *Proceedings of the 13th Text Retrieval Conference, NIST Special Publication 500-261*, *November 16-19*, NIST, Gaithersburg, MD.

Soboroff, I. and Harman, D. (2003), "Overview of the TREC 2003 novelty track", *Proceedings of the 12th Text Retrieval Conference, NIST Special Publication 500-255*, NIST, Gaithersburg, MD, pp. 38-53.

Spitters, M. and Kraaij, W. (2001), "TNO at TDT2001: language model-based topic detection", Topic Detection and Tracking Workshop Report, NIST, Gaithersburg, MD.

Stokes, N. and Carthy, J. (2001), "First story detection using a composite document representation", *Proceedings of the First International Conference on Human Language Technology Research*, Morgan Kaufmann, San Francisco, CA, pp. 134-141.

Tenopir, C. and King, D.W. (2007), "Perceptions of value and value beyond perceptions: measuring the quality and value of journal article readings", *Serials: The Journal for the Serials Community*, Vol. 20 No. 3, pp. 199-207.

Tsai, F.S. and Chan, K.L. (2010), "Redundancy and novelty mining in the business blogosphere", *The Learning Organization*, Vol. 17 No. 6, pp. 490-499.

Ware, M. and Mabe, M. (2009), *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*, International Association of Scientific, Technical and Medical Publishers, Oxford.

Ware, M. and Mabe, M. (2015), *The STM Report: An Overview of Scientific and Scholarly Journal Publishing*, 4th ed., International Association of Scientific, Technical and Medical Publishers, The Hague.

Wayne, C.L. (2000), "Multilingual topic detection and tracking: successful research enabled by corpora and evaluation", *Proceedings of the 2nd International Conference on Language Resources and Evaluation in Athens*, *May 30-June 2*, available at: www.lrec-conf.org/proceedings/lrec2000/pdf/168.pdf

Yang, Y., Zhang, J., Carbonell, J. and Jin, C. (2002), "Topic-conditioned novelty detection", in Zaïane, O.R., Goebel, R., Hand, D., Keim, D. and Ng, R. (Eds), *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in Edmonton, AB, Canada*, ACM, New York, NY, pp. 688-693.

Zhang, M., Song, R., Lin, C., Ma, S., Jiang, Z., Jin, Y., Liu, Y. and Zhao, L. (2002), "Expansion-based technologies in finding relevant and new information: THU TREC2002 novelty track experiments", in Voorhees, E.M. and Buckland, L.P. (Eds), *Proceedings of the 11th Text Retrieval Conference*, NIST, Gaithersburg, MD.

Zhang, Y. and Tsai, F.S. (2009a), "Chinese novelty mining", in Koehn, P. and Mihalcea, R. (Eds), *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing in Suntec, Singapore*, Association for Computational Linguistics, Stroudsburg, PA, pp. 1561-1570.

Zhang, Y. and Tsai, F.S. (2009b), "Combining named entities and tags for novel sentence detection", in Alonso, O., Zaragoza, H., Amatriain, X., Castells, P., Gertz, M., Jackson, P., Kaplan, A., Mika, P. and de Vries, A.P. (Eds), *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval in Barcelona, Spain*, ACM, New York, NY, pp. 30-34.

Zhang, Y., Callan, J. and Minka, T. (2002), "Novelty and redundancy detection in adaptive filtering", in Jävelin, K., Beaulieu, M., Baeza-Yates, R.A. and Myaeng, S.-H. (Eds), *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Tampere, Finland*, ACM, New York, NY, pp. 81-88.

Zhao, L., Zhang, M. and Ma, S. (2006), "The nature of novelty detection", *Information Retrieval*, Vol. 9 No. 5, pp. 521-541.

Further reading

Allan, J., Wade, C. and Bolivar, A. (2003), "Retrieval and novelty detection at the sentence level", in Clarke, C.L.A., Cormack, G.V., Callan, J., Hawking, D. and Smeaton, A.F. (Eds), *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in Toronto, Canada*, ACM, New York, NY, pp. 314-321.

Corresponding author
Andrei Dynich can be contacted at: andreydynich@gmail.com