The study considers the possibilities of using latent semantic analysis for the tasks of identifying scientific subject spaces and evaluating the completeness of covering the results of dissertation research by science degree seekers.

A probabilistic thematic model was built to make it possible to cluster the publications of scholars in scientific areas, taking into account the citation network, which was an important step for solving the problem of identifying scientific subject spaces. As a result of constructing the model, the problem of increasing instability of clustering the citation graph in connection with a decrease in the number of clusters was solved. This problem would arise when combining clusters built on the basis of citation graph clustering, taking into account the similarity of abstracts of scientific publications.

In the article, the presentation of text documents is described based on a probabilistic thematic model using n-grams. A probabilistic thematic model was built for the task of determining the completeness of covering the materials of an author's dissertation research in scientific publications. The approximate values of the threshold coefficients were calculated to evaluate whether the articles of an author included the research provisions that were reflected in the text of the author's abstract of the dissertation. The probabilistic thematic model for an author's publications was practised on the basis of the BigARTM tool. Using the constructed model and with the help of a special regularizer, a matrix was found to evaluate the relevance of topics specified by the segments of an author's dissertation abstracts to documents that are produced by the author's publications.

Important aspects of the possibilities of using latent semantic analysis were studied to identify tasks of scientific subject spaces and to reveal the completeness of covering the results of dissertation research science degree seekers

Keywords: probabilistic latent semantic analysis, clustering, scientific subject space, thematic model

UDC 005.8
DOI: 10.15587/1729-4061.2020.209886

THE USE OF PROBABILISTIC LATENT SEMANTIC ANALYSIS TO IDENTIFY SCIENTIFIC SUBJECT SPACES AND TO EVALUATE THE COMPLETENESS OF COVERING THE RESULTS OF DISSERTATION STUDIES

P. Lizunov
Doctor of Technical Sciences, Professor
Department of Fundamentals of Informatics
Kyiv National University of Construction and Architecture
Povitroflotskyi ave., 31, Kyiv, Ukraine, 03680
A. Biloshchytskyi
Doctor of Technical Sciences, Professor*
Astana IT University
Turkestan str., Nur-Sultan, Kazakhstan, 020000
A. Kuchansky
PhD, Associate Professor*
E-mail: kuczanski@gmail.com
Yu. Andrashko
PhD, Associate Professor
Department of System Analysis and Optimization Theory
Uzhhorod National University
Narodna sq., 3, Uzhhorod, Ukraine, 88000
S. Biloshchytska
PhD, Associate Professor
Department of Intellectual Technologies**
*Department of Information Systems and Technologies**
**Taras Shevchenko National University of Kyiv
Volodymyrska str., 60, Kyiv, Ukraine, 01033

Received date 07.07.2020
Accepted date 03.08.2020
Published date 31.08.2020

1. Introduction

The area of subjects and objects of scientific activity that are united by similar directions of research or a common subject area will be called a scientific subject space. The subjects of such spaces are scientists, who are usually united in scientific communities. These communities are based on the co-authorship of joint scientific publications. The subjects can also be individual scientists who work in a certain space but do not cite or publish articles shared with other scientists. However, it is common to use such terms and concepts that are specific to each space. The links between the scientists that make up the subject matter space can be established through a network of scientific collaboration or a network of citations. Scientific subject spaces are dynamic structures. This is due to the fact that the list of terms that characterise the area is constantly adjusted due to new research and the loss of relevance of other studies. Accordingly, the composition of scientific subject spaces is changing dynamically. Therefore, to identify the boundaries of these spaces, it is necessary to use tools that can highlight the characteristics of the spaces without pre-setting them. For this task, it is possible to use the methods of probabilistic latent semantic analysis.

The main scientific results of a dissertation for a degree should be covered in scientific publications that reveal the main content of the dissertation. This is established by Clause 2 of the Order of the Ministry of Education and Science of Ukraine "On the publication of the results of dissertations for the degree of doctor and candidate of sciences" No. 1220 dated 23 September 2019. Responsibility as to whether the dissertation research complies with Clause 2 of Order No. 1220 and the requirements for the publication of the results of dissertations for the degree of doctor and candidate of sciences is borne by the Specialized Academic Council. The Specialized Academic Council creates an expert group that verifies compliance with the requirements. This process in terms of evaluating the completeness of covering scientific and practical results of an author's dissertation in the scientific articles of the author can be automated by using appropriate methods of latent semantic analysis. This allows each part of the study to be linked to the corresponding author's publications. If it turns out that there is no connection or it is minimal, the Specialized Academic Council may separately consider whether to include a particular publication of the author among the works for the award of a science degree.

The field of probabilistic thematic analysis is developing intensively. New methods are emerging to be used for a wide range of tasks. Therefore, research on this scientific issue is relevant not only in terms of practical application but also in terms of theoretical justifications used to modify methods of probabilistic semantic analysis.

2. Literature review and problem statement

In [1] it is stated that thematic modelling is one of the areas of natural language processing that analyses a relationship between a set of documents and the terms they contain by creating a set of concepts related to the documents and terms. In [2] it was proposed to use latent semantic analysis to search for information in text files, taking into account the content of the documents without focusing on specific incomplete duplicates. A probabilistic approach to latent semantic analysis was proposed in [3]. The method was further developed in [4], which proposed a paragraph vector model for vector representation of the content of texts. The method has advantages from the point of view of calculations. The main disadvantage of this method is the difficulty of interpreting the obtained numerical results. Besides, modern methods based on latent semantic analysis involve the use of additional data about texts, such as citations of texts between authors, geographical representation, and so on. In particular, in [5] an author topic model was given, which takes into account the co-authorship of publications to determine their topics. Since scientific publications usually use a number of specific terms and concepts, it is appropriate to analyse not just the frequencies of individual words in the text but *n*-grams that reflect specific terms. The use of *n*-gram analysis in combination with latent semantic analysis was described in [6].

Probabilistic latent semantic analysis can be used to compare text documents (clustering and classification tasks) as well as to find similarities between text documents and links between terms. Probabilistic latent semantic analysis is also used to find similarities between small groups of terms. In particular, in [7] the multiple choice questions (MCQ) answering model using probabilistic latent semantic analysis was described. In [8] the use of probabilistic latent semantic analysis for machine learning tasks and text data mining was specified.

We will consider two tasks that can be solved using the tools of latent semantic analysis: to identify scientific subject spaces and to determine the completeness of covering dissertation research results by degree seekers.

Identification of scientific subject spaces is necessary for the following objectives:

1) to determine in which priority areas of research a scientist works individually and with what contribution to these areas;

2) to research the identifiers of subjects of a scientific area;

3) to assess the potential of a scientific subject space, taking into account the dynamics of development of the subjects that make up this space.

Any collection of scientific papers has its own list of subject areas. Accordingly, each scholar published in a collection can receive in his/her own list of subject areas part of the areas that are determined for the collection. This method of identifying areas of research for scientists has disadvantages. In particular, the list of the subject areas of a scientist may include third-party areas that relate to the scientist or indirectly and directly do not correspond to his/her competencies or do not relate to his/her research activity. This is because a collection can be interdisciplinary and cover a wide range of topics. For example, journal [9] has a list of eight subject areas according to the Scopus database. They include Industrial and Production Engineering; Engineering; Energy and Power Technologies; Electrical and Electronic Engineering; Applied Mathematics; Business, Management, and Accounting, in particular, Technology and Innovation Management; Systems Management and Engineering; and Computer Science. That is, the range of the subject areas is quite broad.

Let us assume that an author has sixteen articles published and indexed in Scopus, including eight publications in the Eastern European Journal of Enterprise Technologies. The Scopus database identified the following subject areas for the author: Engineering; Computer Science; Mathematics; Business, Management, and Accounting; Energy Economics; Physics and Astronomy; Social Sciences; and Decision Science. From this list, the author has no article that would even indirectly relate to Energy Economics, Physics, and Astronomy. A similar situation with the formation of subject areas is in the vast majority of scientists whose profiles are maintained in this database. Thus, there is a conclusion that such a method of forming the list of subject areas of a scientist can only be an auxiliary tool. Without direct contact with a scholar to determine his/her primary scientific competencies, it cannot be taken as a basis. Another approach involves the open source software Mendeley [10] for managing bibliographic information. Mendeley is for storing and viewing research papers. In an author's profile, it is possible to view the statistics of reading publications, taking into account information about the degree or status of the person who has viewed a publication and his/her priority subject area.

In [11] the method of clustering publications of scientists was described by scientific areas to define scientific subject spaces. This method proposes two ways to find the distance between publications: the length of the route in the citation

graph between publications and the similarity between the abstracts of publications based on the method of locally sensitive hashing. The use of clustering of publications taking into account *n*-grams of analysis was analysed in [12].

Defining scientific subject spaces is an important task of correct assessment of research activities of universities and scientists [13]. In [14] the method to estimate subjects of scientific areas on the basis of calculating the generalized volume of *m*-simplex was described. The use of this method implies the presence of a significant amount of data on the activities of subjects or objects of a scientific area; in particular, these data can change dynamically and are usually presented in the form of time series [15]. Also, defining the boundaries of scientific subject spaces is a key step in planning and evaluating the results of a subject's research activity. The availability of information about other subjects in a given area allows for comparisons and evaluations of the contribution of this subject to the development of the area as a whole. In works [16, 17] infocommunication systems for evaluation of scientific activity were analysed. A project vector method, which was described in [18], can be used to manage scientific activity. This methodology involves the use of decision-making methods, in particular the multistage decision-making process, which were described in [19], and KPI evaluation [20, 21]. The methodology also involves the consideration of scientific activities through the prism of project activities [22]. Distributed project management information systems were described in [23].

The task of assessing the completeness of covering the results of dissertation research is associated with the identification of incomplete duplicates [24], and not in terms of identifying plagiarism but to find fragments of text in the abstract or dissertation. However, since the analysis of a sufficient amount of dissertation materials is a difficult task, due to limited access to these materials, it is convenient for this task to analyse the texts of abstracts, which, according to the requirements, are published for open access. However, in this case there is a difficulty due to the fact that in the abstract the text and wording of the novelty may differ significantly from those that are stated in the author's articles. Therefore, the use of methods for determining incomplete duplicates is incorrect. The authors propose to use probabilistic thematic modelling for the task of assessing the completeness of covering the results of dissertation research.

3. The aim and objectives of the study

The aim of the study is to determine the possibilities of applying latent semantic analysis for the tasks of identifying scientific subject spaces and determining the completeness of covering the results of dissertation research by degree seekers.

To achieve this aim, the following tasks were set and done:

– to perform a systematic analysis of a probabilistic thematic model of presenting text documents, in particular scientific documents, using specific subject terms that are represented by *n*-grams;

– to provide a formal description of the probabilistic thematic model for the problem of clustering publications of scholars in scientific areas, which is an important step for the identification of scientific subject spaces;

– to give a formal description of the probabilistic thematic model for the task of assessing the completeness of covering the materials of the author's dissertation research in his/her scientific articles;

– to verify probabilistic thematic models for the tasks of clustering publications of scholars in scientific spaces and to evaluate the completeness of covering the author's dissertation research in his/her scientific articles.

4. The probabilistic thematic model of presenting text documents taking into account *n*-grams

Suppose a collection of text documents is $Q=\{q_1, q_2, ..., q_m\}$. Then each document is q_j; $j=\overline{1,m}$ is a fragment of a text consisting of words

$$q_j=\left\{w_{1,j}^{\beta_1}, w_{2,j}^{\beta_2}, \ldots, w_{n_j,j}^{\beta_{n_j}}\right\},$$

n_j is the number of words in the document q_j, and β_i, $i=\overline{1,n_j}$ is the word length. A word is represented by a sequence of characters that belong to a finite alphabet $\overline{A}$. If a document contains graphic objects, including drawings, diagrams, charts, mathematical formulae, the probabilistic thematic model does not take them into account.

We will canonise the collection of text documents. For caconization, we first discard all the words in the list of stop words. Then we construct the word sequences of the document q_j in the canonised form. That is, a text document

$$q_j=\left\{\overline{w}_{1,j}^{\beta_1}, \overline{w}_{2,j}^{\beta_2}, \ldots, \overline{w}_{u_j,j}^{\beta_{u_j}}\right\},$$

where $\overline{w}_{i,j}^{\beta_i}$ is the word $w_{i,j}^{\beta_i}$ in the canonised form β_i, $i=\overline{1,u_j}$ is the length of words in the canonised form, and u_j is the number of words in the canonised text.

One of the identifiers to determine to which scientific area a text document belongs is the use of specific terms and concepts. These concepts can be given in one, two or more words. Therefore, according to the authors, it is advisable to consider not individual words of a document but its *n*-grams. Therefore, in the future, the term will be understood as unigrams, bigrams or *n*-grams, which represent those words or expressions that help identify to which scientific area a text document belongs. Let us denote the dictionary of terms for the collection of documents by Ω.

Suppose that there is a finite set of research topics *T*. If the frequency of occurrence of certain terms that define the scientific subject space *B* in the text is higher than the frequency of occurrence of terms of other areas, the text belongs to the scientific subject space *B*. A topic of a document will be understood as the probabilistic discrete distribution on the set of terms Ω, as in [3]. That is, there is a hidden relationship between terms, topics, and a text document. To reflect this dependence, we will present text documents as a set of points (q_i, Ω_i, t_i), $i=\overline{1,Y}$, $Y=|Q|\cdot|\Omega|\cdot|T|$ in a discrete probability area $Q\times\Omega\times T$ with an unknown probability function $p(q, \Omega, t)$. The values of the function $p(q, \Omega, t)$ can be estimated on the basis of a statistical definition of probability:

$$p(q,\omega,t)=\frac{n_{q\omega t}}{\sum\limits_{q=1}^{|Q|}\sum\limits_{\omega=1}^{|\Omega|}\sum\limits_{t=1}^{|T|}n_{q\omega t}},$$

where $p(q, \Omega, t)$ is the probability of using the term Ω in a text document q on the topic t, and $n_{q\Omega t}$ is the number of points (q, Ω, t) in the space $Q\times\Omega\times T$ for a given set of text documents. In other words, $p(q, \Omega, t)$ is the number of uses of the term Ω in the text document q on the topic t.

According to the formula of total probability, the satisfied equality is

$$p(\omega|q)=\sum_{t=1}^{|T|}p(\omega|t,q)\,p(t|q).$$

Let us introduce assumptions about the independence of the use of terms in documents. We will assume that the use of terms depends only on the topic. Let us denote by

$$p(\omega|q)=\sum_{t=1}^{|T|}p(\omega|t)\,p(t|q)=\sum_{t=1}^{|T|}\phi_{\omega t}\,\theta_{tq}. \quad (1)$$

The probabilistic model describes the process of forming a collection of documents based on the known distributions $p(\Omega|t)$ and $p(t|q)$.

Based on the collection of Q documents, it is necessary to find the frequency estimates of the distributions or the parameters $\varphi_{\Omega t}$ and θ_{tq}. The parameters are defined so that model (1) approximates the estimates of the conditional probabilities

$$p(\omega|q)=\frac{n_{q\omega}}{n_q},$$

where $n_{\Omega q}$ is the number of occurrences of the term Ω in the document q, and n_q is the length of the document q.

Let Φ be a matrix representing the belonging of terms to the topics $\Phi=(\varphi_{\Omega t})_{\Omega\times T}$, and let Θ be a matrix of belonging of the topics to the documents $\Theta=(\theta_{tq})_{T\times Q}$.

The principle of maximum likelihood can be used to estimate the unknown parameters of probabilistic models [25]. The likelihood function is defined as the sampling probability $\left(q_i,\omega_i\right)_{i=1}^{K}$ of the model parameters Φ and Θ, $K=\sum_{q=1}^{|Q|}\sum_{\omega=1}^{|\Omega|}n_{q\omega}$:

$$p\left(\left(q_i,\omega_i\right)_{i=1}^{K};\Phi,\Theta\right)=\prod_{i=1}^{K}p(q_i,\omega_i)=$$

$$=\prod_{q=1}^{|Q|}\prod_{\omega\in q}p(\omega|q)^{n_{q\omega}}\,p(q)^{n_{q\omega}}\to\max_{\Phi,\Theta}. \quad (2)$$

Problem (2) is incorrect because it has an infinite number of solutions. This problem can be solved with the inclusion of the regularizer $R(\Phi, \Theta)$, which helps reduce the problem to the correct one [26].

After logarithmization (2) and taking into account the regularizer $R(\Phi, \Theta)$, we obtain the maximization problem:

$$\sum_{q=1}^{|Q|}\sum_{\omega\in q}n_{q\omega}\ln\sum_{t=1}^{|T|}\phi_{\omega t}\theta_{tq}+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \quad (3)$$

$$\sum_{\omega=1}^{|\Omega|}\phi_{\omega t}=1,\ \ \sum_{t=1}^{|T|}\theta_{tq}=1,\ \ \phi_{\omega t}\geq 0,\ \ \theta_{tq}\geq 0. \quad (4)$$

The regulator defines additional constraints that allow obtaining a single solution from an infinite number of solutions of problem (2). In [27] it was proposed to consider the columns of the matrices Φ and Θ as random vectors with the Dirichlet distribution:

$$R(\Phi,\Theta)=\sum_{t=1}^{|T|}\sum_{\omega=1}^{|\Omega|}(\beta_\omega-1)\ln\phi_{\omega t}+\sum_{q=1}^{|Q|}\sum_{t=1}^{|T|}(\alpha_t-1)\ln\theta_{tq}, \quad (5)$$

$$\alpha_t>0,\ \ \alpha_0=\sum_{t=1}^{|T|}\alpha_t,\ \ \beta_\omega>0,\ \ \beta_0=\sum_{\omega=1}^{|\Omega|}\beta_\omega,$$

$$\phi_{\omega t}>0,\ \ \sum_{\omega=1}^{|\Omega|}\phi_{\omega t}=1,\ \ \theta_{tq}>0,\ \ \sum_{t=1}^{|T|}\theta_{tq}=1.$$

If $\beta_\Omega=1$ and $\alpha_t=1$, then we obtain the problem without a regularizer.

If there is an additional link between documents, such as information about citing some documents in others, it can be assumed that the related documents have similar topics. In this case, the regularizer will look like

$$R(\Theta)=\tau\sum_{q=1}^{|Q|}\sum_{c=1}^{|Q|}n_{qc}\sum_{t=1}^{|T|}\theta_{tq}\theta_{tc}, \quad (6)$$

where n_{qc} is the weight of the relationship between the documents of a collection, for example, the number of citations of a document c in the document q [28], τ is the parameter that affects the convergence of solving the problem by some numerical method.

To numerically solve the problem, it is possible to use an iterative approximate EM-algorithm, which consists of two steps: E-step and M-step. First, the probabilities (E-step) are determined on the approximate values of the parameters:

$$p(t|q,\omega)=\begin{cases}0, & \phi_{\omega t}\cdot\theta_{tq}\leq 0,\\ \dfrac{\phi_{\omega t}\theta_{tq}}{\sum_{t\in T}\max\left\{\phi_{\omega t}\theta_{tq},0\right\}}, & \phi_{\omega t}\cdot\theta_{tq}>0,\end{cases} \quad (7)$$

The M-step is determined by the maximization problem:

$$\sum_{q=1}^{|Q|}\sum_{\omega=1}^{|\Omega|}\sum_{t=1}^{|T|}n_{q\omega}\,p(t|q,\omega)\ln\left(\phi_{\omega t}\theta_{tq}\right)+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \quad (8)$$

where $q\in Q$, $\Omega\in\Omega$ and $t\in T$.

The steps E and M are performed sequentially. The condition for completing the algorithm is the fulfilment of the conditions $\exists\varepsilon>0$, that $\left\|\Phi^k-\Phi^{k-1}\right\|<\varepsilon$ and $\left\|\Theta^k-\Theta^{k-1}\right\|<\varepsilon$, where Φ^k is a matrix representing the belonging of terms to topics obtained by the iteration k, Θ^k is a matrix representing the belonging of topics to documents obtained by the iteration k, $k\in N$, and ε is a predefined constant that determines the accuracy of calculation.

5. Application of the probabilistic thematic model to the problem of clustering publications of scientists by subject spaces

Let $A=\{a_1, a_2,..., a_n\}$ be a set of scientists with n as the number of scientists and $Q=\{q_1, q_2,..., q_m\}$ be a set of publications by these scientists with m as the number of publications. Let us denote by $V=\{\eta_1, \eta_2, ..., \eta_\psi\}$ a set of scientific subject spaces with ψ as the number of areas. Identification of scientific subject spaces is a process of establishing a correspondence between a particular scientist and the scientific spaces in which this scientist works and publishes scientific articles within these areas. That is, it is necessary to find the

mapping $\Lambda: A \to V$. One of the ways to identify scientific subject spaces is to use information on the publishing activity of scientists.

In [11] a two-stage clustering method was proposed, which was based on two ways of finding the distance between publications. In the first stage, the distance between publications was calculated based on the length of the minimum route on the citation graph. In the second stage, the distance between publications was calculated based on the degree of similarity in the content of the abstracts of these publications by the Hamming distance based on the method of locally sensitive hashing. After applying the method of clustering this graph and taking into account the specifics of the input data, it was proposed to combine the clusters according to the criterion of proximity of centres of gravity. The use of each stage separately leads to the emergence of isolated clusters of publications that may belong to the same scientific subject space. In particular, there are groups of scholars who study one topic in parallel, but there are few or no citations between their publications. Similarly, given the different authors' styles of writing, there are abstracts of publications that focus on different aspects of the same problem, thus being quite distant in content. The use of both methods of estimating the distance between abstracts makes it possible to identify more accurately publications that belong to a common area of research. These stages do not take into account the thematic area of a text, the author's style and other characteristics for qualitative identification of the research area. Therefore, it is proposed to increase the efficiency of clustering publications in combination with probabilistic latent semantic analysis.

Let the set $C \subset Q \times Q$ specify the citation ratio between scientists' publications. The relationship between publications and their citations can be represented as an oriented graph (Q, C) where publications from the set Q are vertices and citations C are arcs of the graph. To find the scales $n_{q\omega}$, it is possible to use the value inverse to the length of the minimum route between the corresponding vertices of the graph (Q, C). If there is no route between the vertices, then $n_{q\Omega}=0$. From formulae (3), (4), and (6), we obtain the following:

$$\sum_{q=1}^{|Q|}\sum_{\omega\in q} n_{q\omega} \ln \sum_{t=1}^{|T|} \phi_{\omega t}\theta_{tq} + \tau \sum_{q=1}^{|Q|}\sum_{c=1}^{|Q|} n_{qc} \sum_{t=1}^{|T|} \theta_{tq}\theta_{tc} \to \max_{\Phi,\Theta}, \tag{9}$$

$$\sum_{\omega=1}^{|\Omega|} \phi_{\omega t} = 1, \quad \sum_{t=1}^{|T|} \theta_{tq} = 1, \quad \phi_{\omega t} \geq 0, \quad \theta_{tq} \geq 0. \tag{10}$$

In contrast to strict clustering [11], on the basis of problem (9)–(10), a fuzzy distribution of publications belonging to the clusters will be obtained, which is described by the matrix Θ:

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \dots & \theta_{1m} \\ \theta_{21} & \theta_{22} & \dots & \theta_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{\psi 1} & \theta_{\psi 1} & \dots & \theta_{\psi m} \end{pmatrix}. \tag{11}$$

Let us denote by $\overline{Q}_z(a_i)$ the set of publications of the author a_i, $i=\overline{1,n}$, included in the cluster y_z, $z=\overline{1,\psi}$:

$$\overline{Q}_z(a_i) = \left\{ q_j \in Q \,\middle|\, (a_i, q_j) \in U, q_j \in y_z \right\},$$

$i=\overline{1,n}$, $z=\overline{1,\psi}$, $U \subset A \times Q$,

which reflects the authorship of the publications $q \in Q$.

Then the membership of the scientist a_i, $i=\overline{1,n}$ in each of the clusters y_z, $z=\overline{1,\psi}$ will be defined as follows:

$$R_z(a_i) = \frac{\sum\limits_{q\in\overline{Q}_z(a_i)} \theta_{zq}}{card\left(\overline{Q}_z(a_i)\right)}, \quad z=\overline{1,\psi}, \tag{12}$$

where $R_z(a_i)$ is the belonging of the scientist a_i to the cluster y_z.

6. Application of the probabilistic thematic model to evaluate the completeness of the dissertation coverage in publications

Let the dissertation text or dissertation abstract be divided into m segments, $Q=\{q_1, q_2, ..., q_m\}$, which correspond to the described results (divided into paragraphs, chapters, sections, etc.). It is not known in advance what result is described in which of the segments. In the presentation of probabilistic latent semantic analysis, the author's publications will be determined by the probability distribution of the frequencies of the terms and can be used to define specific topics $T=\{t_1, t_2, ..., t_{|T|}\}$. The set of terms Ω is based on analysing the text of the author's abstract or dissertation. The degree of belonging of the abstract segment to the corresponding author's publication can be found on the basis of solving problem (3)–(5). The solution of the problem will be the matrix Θ. If it has a column whose values are close to zero, $\forall j\ \theta_{tj} \leq H_1$ for $t=\overline{1,|T|}$, the result described in the text segment corresponding to this column is not covered in any of the author's publications. If there is a line in it the values of which are close to zero, $\forall j\ \theta_{tq} \leq H_2$ for $q=\overline{1,m}$, the author's publication contains results that do not correspond or are not covered in the dissertation research or abstract. H_1 and H_2 are small numbers the values of which are determined as a result of statistical observations. The results of the analysis are passed on for examination.

The distribution of terms by topic, that is, the matrix Φ, is approximately known and determined in advance because for a particular author, it is easy to find this distribution. For example, the terms of an author's scientific publications may be their relevant keywords.

To identify the elements of the matrix Φ, partial training can be conducted, during which experts can note in the topics those terms and segments of the text that are relevant. This will increase the stability of the model. The enquiry is made for the semantic core of one or more topics [29]. It is necessary to adjust the formula for the regularizer (5) because the columns of the matrices Φ and Θ are not independent:

$$R(\Phi,\Theta) = \sum_{t=1}^{|T|}\sum_{\omega=1}^{|\Omega|} \beta_{\omega t} \ln \phi_{\omega t} + \sum_{q=1}^{|Q|}\sum_{t=1}^{|T|} \alpha_{tq} \ln \theta_{tq}, \tag{13}$$

where $\beta_{\Omega t}$ is a numerical estimate determined by the number of relevant terms, and α_{tq} is a numerical estimate determined by the number of relevant text segments.

7. Verification of the probabilistic thematic model for the set tasks

To verify the results of the study, the analytical information system "Database of Scientists of Ukraine" was finalized; it is a database with information about 215,082 scientific publications and 58,834 Ukrainian authors of these publications. To cluster the scientific publications, the method of clustering a citation graph taking into account the similarity of abstracts of scientific publications was used for the task to identify scientific subject spaces. The functionality of the system was also expanded in terms of using probabilistic latent semantic analysis for the problem of clustering scientific publications based on the probabilistic thematic model (9), (10), taking into account *n*-grams. To calculate the matrices Φ and Θ, a tool for building thematic models with the open source BigARTM was used [30]. The probabilistic thematic models that can be used in this tool were described in [31]. The input for thematic modelling in this tool is two files based on a collection of publications. One file consists of a list of all words in the canonised texts of publications; stop words, all nouns in the nominative singular form, verbs in the infinitive form, etc. are not taken into account. The second file is represented by a table with fields such as a publication index, a word index in the first file, and the number of word occurrences in the publication. The result of the tool is the matrices Φ and Θ. Next, formula (12) was used for the constructed matrix Θ to make a conclusion about the belonging of the author of the publication to a relevant scientific subject space.

The solution of the problem of assessing the completeness of covering the dissertation materials in scientific articles required two steps:

1) to study the probabilistic thematic model based on an author's publications on the basis of the BigARTM tool; the training outcomes are stored in a separate file in binary representation;

2) to find the matrix Θ for the segments of dissertation abstracts using the practised model and the special regularizer (13).

It was established that the values of H_1 and H_2 should be close to $5 \cdot 10^{-5}$.

8. Discussion of the results of applying the probabilistic thematic model

The verification results show the possibilities of applying latent semantic analysis to the tasks of identifying scientific subject spaces and assessing the completeness of covering the results of dissertation research by degree seekers. A feature of the probabilistic thematic model for clustering publications of scientists in scientific spaces, taking into account the citation network (9), (10), and (12), is the problem of increasing instability of clustering the citation graph due to a smaller number of clusters. This problem arises when combining clusters based on the clustering of the citation graph, taking into account the similarity of the abstracts of scientific publications. A feature of the probabilistic thematic model for the problem of evaluating the completeness of covering the author's dissertation research materials in his/her scientific publications (9), (10), (13) is the use of training and the special regularizer (13). The result of applying the model is a matrix of belonging of segments of an author's dissertation abstract to the documents determined by the author's publications. The application of this model to this problem has not been described yet.

For the task of clustering scientific publications, due to the properties of latent semantic analysis to detect hidden links, it was possible to reduce the number of clusters of scientific subject spaces and solve the problem of increasing instability of clustering graph citation due to reducing the number of clusters. The model (9), (10), and (12) is used for this purpose. For the task of determining the completeness of covering the results of dissertation research, degree seekers have the opportunity to study the probabilistic thematic model (9), (10), and (13) according to the author's publications. As a result, the matrix Φ is recorded, and the special regularizer (13) is used to find the matrix Θ.

The class of expert methods is mostly used for the task to identify scientific subject spaces. The disadvantage of this class for this task is that the number of publications of scientists to be distributed among a sufficiently large number of clusters is constantly growing. In the presented probabilistic thematic model (9), (10), and (12), the problem of identifying scientific subject spaces is solved by clustering scientific publications without taking into account opinions of experts.

The task to identify the completeness of covering the results of dissertation research by seekers of a science degree is poorly understood. The given model (9), (10), and (13) helps receive a matrix of belonging of an author's segments of the dissertation abstract to documents that are defined by publications of the author for a small volume of input data. The model is sensitive to the emergence of new publications, so in this case, to fully solve the problem, it is necessary to modify the model so that it does not require retraining.

A limitation of the study is the problem of canonisation of texts in different languages. This study uses textual information in the Ukrainian language. In further research, restriction of texts to one language database will be offered, especially due to the fact that the tools to canonise English texts have more opportunities, in particular for scientific publications. Besides, a limitation of the second task is the difficulty of obtaining full texts of dissertations for complete verification of the model.

The peculiarity of latent semantic analysis is that it is used for a wide range of text processing tasks. This paper has considered two specific problems that were solved on the basis of probabilistic thematic models with appropriate regularizers. The considered tasks were based on the problem of maximizing the plausibility function that used to be set incorrectly. Only appropriate regularizers were used to reduce the task to the correct one. Other methods of reducing tasks to the correct ones were not considered.

9. Conclusion

1. A systematic analysis was performed for the probabilistic thematic model of presenting text documents, in particular scientific documents with the use of specific subject terms that were represented by *n*-grams. It has been established that this model can be effectively used to solve the tasks to identify scientific subject spaces and to determine the completeness of covering the results of dissertation research by degree seekers.

2. A formal description of the probabilistic thematic model was given for the problem of clustering publications

of scholars by scientific areas. This is an important step in identifying scientific subject spaces. The verification of the described method on the basis of this thematic model produced the following results:

– the problem of increasing the instability of clustering the citation graph was solved due to the reduction in the number of clusters;

– the number of obtained clusters of scientific subject spaces in other studies used to be about 500. In the case of clustering by the probabilistic thematic model with the regularizer $R(\Phi, \Theta)$ it is about 250.

3. A formal description of the probabilistic thematic model was made for the problem of evaluating the completeness of covering the materials of an author's dissertation research in his/her scientific publications. Since there is not a large enough database of dissertation abstracts for training, it is difficult to specify the optimal values of the threshold coefficients H_1 and H_2.

4. When checking the recognised completeness of covering the dissertation research materials of an author in his/her scientific publications (the volume of 20 texts), it was found that the coefficients H_1 and H_2 should be close to $5 \cdot 10^{-5}$.

Acknowledgment

The study was performed within the research work Development of Combined Methods to Identify Incomplete Duplicates and Determine the Completeness of Covering Scientific Results of Dissertation Research Published by an Author, No. 0119U002579. The study has also solved problems that can be considered in the research work Development of Methods for Analysing the Quality of Research Work of Scientists, HEIs of the Ministry of Education and Science of Ukraine, and Its Departments, No. 0119U100187.

References

1. Dumais, S. T. (2005). Latent semantic analysis. Annual Review of Information Science and Technology, 38 (1), 188–230. doi: https://doi.org/10.1002/aris.1440380105
2. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41 (6), 391–407. doi: https://doi.org/10.1002/(sici)1097-4571(199009)41:6<391::aid-asi1>3.0.co;2-9
3. Hofmann, T. (1999). Probabilistic latent semantic indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '99. doi: https://doi.org/10.1145/312624.312649
4. Dai, A. M., Olah, C., Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv. Available at: https://arxiv.org/pdf/1507.07998v1.pdf
5. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P. (2004). The Author-Topic Model for Authors and Documents. Conference: UAI '04, Proceedings of the 20th Conference in Uncertainty in Artificial Intelligence.
6. Pagliardini, M., Gupta, P., Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 528–540. doi: https://doi.org/10.18653/v1/n18-1049
7. Lifchitz, A., Jhean-Larose, S., Denhière, G. (2009). Effect of tuned parameters on an LSA multiple choice questions answering model. Behavior Research Methods, 41 (4), 1201–1209. doi: https://doi.org/10.3758/brm.41.4.1201
8. Gálvez, R. H., Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. Journal of Computational Science, 19, 43–56. doi: https://doi.org/10.1016/j.jocs.2017.01.001
9. Scopus Preview. Eastern-European Journal of Enterprise Technologies. Available at: https://www.scopus.com/sourceid/21100450083
10. Mendeley. Available at: https://www.mendeley.com/?interaction_required=true
11. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Kuzka, O., Shabala, Y., Lyashchenko, T. (2017). A method for the identification of scientists' research areas based on a cluster analysis of scientific publications. Eastern-European Journal of Enterprise Technologies, 5 (2 (89)), 4–11. doi: https://doi.org/10.15587/1729-4061.2017.112323
12. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S. (2019). Improvement of the method for scientific publications clustering based on n-gram analysis and fuzzy method for selecting research partners. Eastern-European Journal of Enterprise Technologies, 4 (4 (100)), 6–14. doi: https://doi.org/10.15587/1729-4061.2019.175139
13. Bykov, V. Y., Kuchanskyi, O. Y., Biloshchytskyi, A. O., Andrashko, Y. V., Dikhtiarenko, O. V., Budnik, S. V. (2019). Development of information technology for complex evaluation of higher education institutions. Information Technologies and Learning Tools, 73 (5), 293–306. doi: https://doi.org/10.33407/itlt.v73i5.3397
14. Kuchansky, A., Andrashko, Yu., Biloshchytskyi, A., Danchenko, O., Ilarionov, O., Vatskel, I., Honcharenko, T. (2018). The method for evaluation of educational environment subjects' performance based on the calculation of volumes of msimplexes. Eastern-European Journal of Enterprise Technologies, 2 (4 (92)), 15–25. doi: https://doi.org/10.15587/1729-4061.2018.126287
15. Kuchansky, A., Biloshchytskyi, A., Andrashko, Y., Biloshchytska, S., Shabala, Y., Myronov, O. (2018). Development of adaptive combined models for predicting time series based on similarity identification. Eastern-European Journal of Enterprise Technologies, 1 (4 (91)), 32–42. doi: https://doi.org/10.15587/1729-4061.2018.121620
16. Biloshchytskyi, A., Biloshchytska, S., Kuchansky, A., Bielova, O., Andrashko, Y. (2018). Infocommunication system of scientific activity management on the basis of project-vector methodology. 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET). doi: https://doi.org/10.1109/tcset.2018.8336186

17. Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S., Danchenko, O. (2018). Development of Infocommunication System for Scientific Activity Administration of Educational Environment's Subjects. 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). doi: https://doi.org/10.1109/infocommst.2018.8632036
18. Biloshchytskyi, A., Kuchansky, A., Paliy, S., Biloshchytska, S., Bronin, S., Andrashko, Y. et. al. (2018). Development of technical component of the methodology for projectvector management of educational environments. Eastern-European Journal of Enterprise Technologies, 2 (2 (92)), 4–13. doi: https://doi.org/10.15587/1729-4061.2018.126301
19. Mulesa, O., Snytyuk, V., Myronyuk, I. (2019). Optimal alternative selection models in a multi-stage decision-making process. EUREKA: Physics and Engineering, 6, 43–50. doi: https://doi.org/10.21303/2461-4262.2019.001005
20. Ostakhov, V., Artykulna, N., Morozov, V. (2018). Models of IT Projects KPIs and Metrics. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). doi: https://doi.org/10.1109/dsmp.2018.8478464
21. Ostakhov, V., Morozov, V. (2019). Models and Methods of IT and Infocommunications Portfolio Management Using the System of Metrics and KPIs. 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T). doi: https://doi.org/10.1109/picst47496.2019.9061328
22. Kolesnikov, O., Gogunskii, V., Kolesnikova, K., Lukianov, D., Olekh, T. (2016). Development of the model of interaction among the project, team of project and project environment in project system. Eastern-European Journal of Enterprise Technologies, 5 (9 (83)), 20–26. doi: https://doi.org/10.15587/1729-4061.2016.80769
23. Morozov, V., Kalnichenko, O., Liubyma, I. (2017). Managing projects configuration in development distributed information systems. 2017 2nd International Conference on Advanced Information and Communication Technologies (AICT). doi: https://doi.org/10.1109/aiact.2017.8020088
24. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Biloshchytska, S., Chala, L. (2016). Detection of near dublicates in tables based on the locality-sensitive hashing method and the nearest neighbor method. Eastern-European Journal of Enterprise Technologies, 6 (4 (84)), 4–10. doi: https://doi.org/10.15587/1729-4061.2016.86243
25. Rossi, R. J. (2018). Mathematical Statistics: An Introduction to Likelihood Based Inference. John Wiley & Sons. doi: https://doi.org/10.1002/9781118771075
26. Tihonov, A. N., Arsenin, V. Ya. (1986). Metody resheniya nekorrektnyh zadach. Moscow: Nauka, 287.
27. Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.
28. Dietz, L., Bickel, S., Scheffer, T. (2007). Unsupervised prediction of citation influences. Proceedings of the 24th International Conference on Machine Learning - ICML '07. doi: https://doi.org/10.1145/1273496.1273526
29. Andrzejewski, D., Zhu, X. (2009). Latent Dirichlet Allocation with topic-in-set knowledge. Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing - SemiSupLearn '09. doi: https://doi.org/10.3115/1621829.1621835
30. BigARTM. Available at: https://bigartm.readthedocs.io/en/stable/intro.html
31. Vorontsov, K. V. (2013). Veroyatnostnoe tematicheskoe modelirovanie. Available at: http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf