DOI: 10.32347/2412-9933.2021.47.102-
УДК 005.8

Лізунов Петро Петрович
Доктор технічних наук, професор, завідувач кафедри будівельної механіки, *orcid.org/0000-0003-2924-3025*
Київський національний університет будівництва і архітектури, Київ
Білощицький Андрій Олександрович
Доктор технічних наук, професор, проректор з науки та інновацій, *orcid.org/0000-0001-9548-1959*
Astana IT University, Нур-Султан
Кучанський Олександр Юрійович
Доктор технічних наук, доцент, доцент кафедри інформаційних систем та технологій,
orcid.org/0000-0003-1277-8031
Київський національний університет імені Тараса Шевченка, Київ
Андрашко Юрій Васильович
Кандидат технічних наук, доцент кафедри системного аналізу і теорії оптимізації,
orcid.org/0000-0003-2306-8377
ДВНЗ «Ужгородський національний університет», Ужгород
Лященко Тамара Олексіївна
Старший викладач кафедри інформаційних технологій, *orcid.org/0000-0001-9092-0297*
Київський національний університет будівництва і архітектури, Київ

ЗАДАЧА ВСТАНОВЛЕННЯ ПОВНОТИ ВИСВІТЛЕННЯ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЙНИХ ДОСЛІДЖЕНЬ ЗДОБУВАЧАМИ НАУКОВИХ СТУПЕНІВ

***Анотація**. Описано можливості застосування латентного семантичного аналізу для задачі виявлення повноти висвітлення результатів дисертаційних досліджень здобувачами наукових ступенів. Для досягнення мети виконано такі завдання: зроблено огляд ймовірнісної тематичної моделі представлення текстових документів, зокрема наукових документів з використанням специфічних предметних термінів, які представляються n-грамам; наведено формальне описання ймовірнісної тематичної моделі для задачі встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових статтях. Особливістю ймовірнісної тематичної моделі для задачі встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових публікаціях є використання навчання та спеціального регуляризатора. Результатом моделі є матриця належності тем, які визначаються сегментами авторефератів дисертації автора до документів, які визначаються публікаціями автора. Застосування цієї моделі до пропонованої задачі ще не було описано. Розглянута в роботі задача спирається на задачу максимізації функції правдоподібності, яка є некоректно поставленою. Для зведення задачі до коректно поставленої використовуються тільки відповідні регуляризатори. Інші методи зведення задач до коректних не розглядалися. Обмеженням дослідження є проблема канонізації текстів різними мовами. У пропонованому дослідженні використовується текстова інформація українською мовою. У подальшому дослідженні буде запропоновано зведення текстів до однієї мовної бази, оскільки інструменти канонізації текстів англійської мови мають більш широкі можливості, зокрема для наукових публікацій. Також обмеженням є складність отримання повних текстів дисертацій для повноцінної верифікації моделі. Результати дослідження використовуються в комплексі з системою виявлення неповних дублікатів у наукових документах, зокрема дисертаціях на здобуття наукового ступеня.*

Ключові слова: дисертація; наукове дослідження; наукова публікація; латентний семантичний аналіз

Вступ

Наукові результати дисертації на здобуття наукового ступеня мають бути висвітлені у наукових публікаціях, які розкривають основний зміст дисертації. Це визначається п. 2. наказу Міністерства освіти і науки України «Про опублікування результатів дисертацій на здобуття наукових ступенів доктора і кандидата наук» № 1220 від 23.09.2019 р. Відповідальність щодо того,

чи задовольняє дисертаційне дослідження п. 2. наказу № 1220 та вимогам до опублікування результатів дисертацій на здобуття наукових ступенів доктора і кандидата наук несе спеціалізована вчена рада. Спеціалізована вчена рада формує експертну групу, яка перевіряє відповідність цих вимог. Для автоматизації цього процесу в частині встановлення повноти висвітлення наукових та практичних результатів дисертації автора в наукових статтях цього автора можуть бути застосовані відповідні методи латентного семантичного аналізу. Це дає змогу встановити зв'язок кожної частини дослідження з відповідною публікацією автора. Якщо виявиться, що зв'язку немає або він недостатній, то спеціалізована вчена рада може окремо розглянути питання врахування цієї публікації автора в доробок для присудження йому наукового ступеня.

Напрям ймовірнісного тематичного аналізу інтенсивно розвивається. З'являються нові методи, що можуть бути використані для широкого спектру задач. Тому дослідження, яке описане в пропонованій роботі, є актуальним не тільки з точки зору практичного застосування, але і з точки зору теоретичних обґрунтувань модифікацій методів ймовірнісного семантичного аналізу. Деякі теоретичні обґрунтування щодо використання латентного семантичного аналізу для встановлення повноти висвітлення результатів дисертаційних досліджень здобувачів наукових ступенів описані авторами в роботах [1; 2].

У роботі [3] вказано, що тематичне моделювання – це один із напрямів опрацювання природної мови, що аналізує взаємозв'язки між набором документів і термів, які вони містять, шляхом створення набору понять, пов'язаних з документами та термами. У роботі [4] було запропоновано використовувати латентний семантичний аналіз для пошуку інформації в текстових документах, враховуючи зміст документу, при цьому без акцентування уваги на конкретних неповних дублікатах. Ймовірнісний підхід до латетного семантичного аналізу було запропоновано в роботі [5]. Метод набув подальшого розвитку в роботі [6], де була запропонована модель «Параграф-Вектор» для векторного представлення змісту текстів. Метод має переваги з точки зору обчислень. Основним недоліком цього методу є складність інтерпретації отриманих числових результатів. Також сучасні методи на основі латентного семантичного аналізу передбачають використання додаткових даних про тексти, наприклад про цитування текстів між авторами, географічне представлення тощо. Зокрема в роботі [7] наведено модель «Автор-Тема», в якій враховується співавторство публікацій для встановлення тем. Оскільки в наукових публікаціях, як правило, використовується низка специфічних термінів і понять, то доречним є аналіз не просто частот появи окремих слів у тексті, а n-грамів, які відображають конкретні терміни. Використання n-грам аналізу в комплексі з латентним семантичним аналізом описано в роботі [8].

Ймовірнісний латентний семантичний аналіз може використовуватись для порівняння текстових документів (задачі кластеризації та класифікації), знаходження подібності між текстовими документами, зв'язків між термами. Також ймовірнісний латентний семантичний аналіз використовується для встановлення подібності між невеликими групами термів. Зокрема, в роботі [9] описано застосування Multi Choice Questions (MCQ) Answering Model з використанням ймовірнісного латентного семантичного аналізу. В роботі [10] описано використання ймовірнісного латентного семантичного аналізу для задач машинного навчання та Text data mining.

Задача оцінювання повноти висвітлення результатів дисертаційних досліджень пов'язана з визначенням неповних дублікатів [11; 12], але не з точки зору ідентифікації плагіату, а для знаходження фрагментів тексту в авторефераті або тексті дисертації. Проте аналіз достатнього обсягу дисертаційних матеріалів є складною задачею через обмежений доступ до даних матеріалів. Тому зручно для цієї задачі аналізувати текст авторефератів, що згідно вимог публікуються для відкритого доступу. Проте в цьому випадку виникає складність, яка пов'язана з тим, що в авторефераті текст та формулювання новизни може суттєво відрізнятися від статей автора. Тому в цьому випадку застосування методів визначення неповних дублікатів є некоректним. Авторами пропонується для задачі оцінювання повноти висвітлення результатів дисертаційних досліджень використовувати ймовірнісне тематичне моделювання [1].

Мета дослідження

Метою дослідження є вивчення можливостей застосування латентного семантичного аналізу для задачі виявлення повноти висвітлення результатів дисертаційних досліджень здобувачами наукового ступеня.

Для досягнення мети поставлено такі завдання:

1. Зробити огляд ймовірнісної тематичної моделі представлення текстових документів, зокрема наукових документів з використанням специфічних предметних термінів, які представляються n-грамами.

2. Навести формальне описання ймовірнісної тематичної моделі для задачі встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових статтях.

Виклад основного матеріалу

Ймовірнісна тематична модель представлення текстових документів з врахуванням n-грамів

Нехай задано колекцію текстових документів $Q=\{q_1,q_2,\ldots,q_m\}$. Тоді кожен документ $q_j,\ j=\overline{1,m}$ являє собою фрагмент тексту, що складається зі слів $q_j=\left\{w_{1,j}^{\beta_1}, w_{2,j}^{\beta_2},\ldots, w_{n_j,j}^{\beta_{n_j}}\right\}$, n_j – кількість слів у документі q_j, а $\beta_i,\ i=\overline{1,n_j}$ – довжина слова. Слово представляється послідовністю символів, які належать до скінченного алфавіту $\overline{A}$. Якщо в документі наявні графічні об'єкти, зокрема рисунки, схеми, діаграми, математичні формули, то ймовірнісна тематична модель не передбачає їх врахування.

Проведемо канонізацію колекції текстових документів. Для каконізації спочатку відкинемо всі слова, що входять до переліку стоп-слів. Потім побудуємо послідовності слів документа q_j в канонізованій формі. Тобто текстовий документ $q_j=\left\{\overline{w}_{1,j}^{\beta_1}, \overline{w}_{2,j}^{\beta_2},\ldots, \overline{w}_{u_j,j}^{\beta_{u_j}}\right\}$, де $\overline{w}_{i,j}^{\beta_i}$ – це слово $w_{i,j}^{\beta_i}$ в канонізованій формі $\beta_i,\ i=\overline{1,u_j}$ – довжини слів у канонізованій формі, а u_j – кількість слів у канонізованому тексті.

Одним з ідентифікаторів, який визначає до якої галузі належить науковий текстовий документ, є використання у ньому специфічних термів і понять. Ці поняття можуть задаватися одним, двома і більше словами. Тому, на думку авторів, доцільним є розгляд не окремих слів документу, а його n-грамів. Отже, надалі під термом будемо розуміти уніграми, біграми або n-грами, які представляють усталені слова або вирази, що дають змогу ідентифікувати, якому науковцю належить текстовий документ. Позначимо словник термів для колекції документів через Ω.

Нехай існує скінченна множина тем досліджень T. Якщо частота появи певних термів, які визначають предметний науковий простір B, у тексті вищий, у порівнянні з частотою появи термів інших напрямів, то текст належить до предметного наукового простору B. Темою документу будемо розуміти ймовірнісний дискретний розподіл на множині термів Ω, як в роботі [5]. Тобто існує прихована залежність між термами, темами та текстовим документом. Для відображення цієї залежності представимо текстові документи як множину точок $(q_i,\omega_i,t_i),\ i=\overline{1,Y}$, $Y=|Q|\cdot|\Omega|\cdot|T|$ у дискретному ймовірнісному просторі $Q\times\Omega\times T$ з невідомою функцією ймовірності $p(q,\omega,t)$. Значення функції p(q,ω,t) можуть бути оцінені на основі статистичного означення ймовірності:

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

де p(q,ω,t) – ймовірність використання терму ω в текстовому документі q за темою t; $n_{q\omega t}$ – кількість точок (q,ω,t) в просторі $Q\times\Omega\times T$ для заданої множини текстових документів. Іншими словами p(q,ω,t) – це кількість використань терма ω в текстовому документі q за темою t.

Згідно з формулою повної ймовірності виконується рівність

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

Введемо припущення щодо незалежності застосування термів у документах. Будемо вважати, що застосування термів залежить тільки від теми. Позначимо через

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

ймовірнісну модель, що описує процес формування колекції документів на основі відомих розподілів p(ω|t) та p(t|q).

На основі колекції документів Q необхідно знайти частотні оцінки розподілів або параметри $\varphi_{\omega t}$ та θ_{tq}. Пареметри визначені в такий спосіб, щоб модель (1) наближала оцінки умовних ймовірностей

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

де $n_{\omega q}$ – це число входжень терма ω в документ q; n_q – довжина документа q.

Нехай Φ – матриця, що представляє належність термів до тем $\Phi=(\varphi_{\omega t})_{\Omega\times T}$, а Θ – матриця належності тем до документів $\Theta=(\theta_{tq})_{T\times Q}$.

Для оцінювання невідомих параметрів ймовірнісних моделей можна використати принцип максимума правдоподібності. Функція правдоподібності визначається як ймовірність вибірки $\left(q_i,\omega_i\right)_{i=1}^{K}$ від параметрів моделі Φ та Θ,

$$K=\sum_{q=1}^{|Q|}\sum_{\omega=1}^{|\Omega|}n_{q\omega}$$

$$p\left((q_i,\omega_i)_{i=1}^{K};\Phi,\Theta\right)=\prod_{i=1}^{K}p(q_i,\omega_i)=$$
$$=\prod_{q=1}^{|Q|}\prod_{\omega\in q}p(\omega|q)^{n_{q\omega}}\,p(q)^{n_{q\omega}}\to\max_{\Phi,\Theta}. \quad (1)$$

Задача (1) є некоректно поставленою, оскільки має нескінченну кількість розв'язків. Для вирішення цієї проблеми можна ввести регуляризатор R(Φ,Θ), що допомагає звести задачу до коректно поставленої [13].

Після логарифмування (1) та врахування регуляризатора R(Φ,Θ) отримаємо задачу максимізації:

$$\sum_{q=1}^{|Q|}\sum_{\omega\in q}n_{q\omega}\ln\sum_{t=1}^{|T|}\phi_{\omega t}\theta_{tq}+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \quad (2)$$

$$\sum_{\omega=1}^{|\Omega|}\phi_{\omega t}=1,\quad \sum_{t=1}^{|T|}\theta_{tq}=1,$$
$$\phi_{\omega t}\ge 0,\qquad \theta_{tq}\ge 0. \quad (3)$$

Регуляризатор визначає додаткові обмеження, що допомагають з нескінченної кількості розв'язків задачі (1) отримати єдиний. У роботі [14] запропоновано вважати стовпці матриць Φ та Θ випадковими векторами з розподілом Діріхле:

$$R(\Phi,\Theta)=\sum_{t=1}^{|T|}\sum_{\omega=1}^{|\Omega|}(\beta_\omega-1)\ln\phi_{\omega t}+\sum_{q=1}^{|Q|}\sum_{t=1}^{|T|}(\alpha_t-1)\ln\theta_{tq}, \quad (4)$$

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

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

Якщо $\beta_\omega=1$, $\alpha_t=1$, то отримаємо задачу без регуляризатора.

У випадку, якщо наявний додатковий зв'язок між документами, наприклад інформація про цитування одних документів у інших, то можна вважати, що пов'язані документи мають подібні теми. Тоді регуляризатор матиме вигляд:

$$R(\Theta)=\tau\sum_{q=1}^{|Q|}\sum_{c=1}^{|Q|}n_{qc}\sum_{t=1}^{|T|}\theta_{tq}\theta_{tc}, \quad (5)$$

де n_{qc} – вага зв'язку між документами колекції, наприклад кількість цитувань документа c у документі q [15]; τ – параметр, від якого залежить збіжність задачі за чисельним методом.

Для чисельного розв'язання задачі можна використати ітеративний наближений ЕМ-алгоритм, який складається з двох кроків: Е-крок та М-крок. Спочатку на основі наближених значень параметрів визначаються ймовірності (Е-крок):

$$p(t|q,\omega)=\begin{cases}0, & \phi_{\omega t}\cdot\theta_{tq}\le 0\\ \dfrac{\phi_{\omega t}\theta_{tq}}{\sum_{t\in T}\max\{\phi_{\omega t}\theta_{tq},0\}}, & \phi_{\omega t}\cdot\theta_{tq}>0,\end{cases} \quad (6)$$

М-крок визначається задачею максимізації

$$\sum_{q=1}^{|Q|}\sum_{\omega=1}^{|\Omega|}\sum_{t=1}^{|T|}n_{q\omega}p(t|q,\omega)\ln(\phi_{\omega t}\theta_{tq})+R(\Phi,\Theta)\to\max_{\Phi,\Theta}, \quad (7)$$

де $q\in Q$, $\omega\in\Omega$, $t\in T$.

Кроки Е та М виконуються послідовно. Умовою завершення алгоритму є виконання умов $\exists\varepsilon>0$, що $\left\|\Phi^k-\Phi^{k-1}\right\|<\varepsilon$ і $\left\|\Theta^k-\Theta^{k-1}\right\|<\varepsilon$, де Φ^k – матриця, що представляє належність термів до тем, отримана на ітерації k; Θ^k – матриця, що представляє належність тем до документів, отримана на ітерації k, $k\in N$; ε – наперед визначена константа, яка визначає точність обчислення.

Застосування ймовірнісної тематичної моделі для встановлення повноти висвітлення матеріалів дисертації у публікаціях

Нехай текст дисертації або автореферат дисертації розділяється на m сегментів, $Q=\{q_1,q_2,\ldots,q_m\}$, які відповідають описаним результатам (розділяється на абзаци, параграфи, розділи тощо). Наперед невідомо, який результат в якому із сегментів описано. У представленні ймовірнісного латентного семантичного аналізу публікації автора будуть визначатися розподілом ймовірностей частот термів і можуть бути використані для означення конкретних тем $T=\{t_1,t_2,\ldots,t_{|T|}\}$. Множина термів Ω будується на основі аналізу тексту автореферата або дисертації автора. Ступінь належності сегменту автореферата відповідній публікації автора можна знайти на основі розв'язування задачі (2), (3), (4). Розв'язком задачі буде матриця Θ. Якщо в ній наявний стовпець, значення якого близькі до нуля, $\forall j\ \theta_{tj}\le H_1$ для $t=\overline{1,|T|}$, то результат, який описано в сегменті тексту, що відповідає цьому стовпцю, не висвітлено в жодній із публікацій автора. Якщо в ній наявний рядок, значення якого близькі до нуля, $\forall j\ \theta_{tq}\le H_2$ для $q=\overline{1,m}$, то публікація автора містить результати, які не відповідають або не висвітлені в дисертаційному дослідженні або авторефераті. H_1 та H_2 малі числа, значення яких визначаються в результаті проведення статистичних спостережень. Результати аналізу передаються далі для проведення експертизи.

Розподіл термів по темах, тобто матриця Φ наближено відома і визначається заздалегідь. Отже, для конкретного автора можна легко визначити даний розподіл. Наприклад, термами наукових публікацій деякого автора можуть бути їх відповідні ключові слова.

Для визначення елементів матриці Φ можна провести часткове навчання, в ході якого експерти можуть відзначати в темах терми і сегменти тексту, які є релевантними. Це дасть змогу збільшити стійкість моделі. Як запит задається семантичне ядро однієї чи кількох тем [16]. Необхідно скоригувати формулу для регуляризатора (5), бо стовпці матриць Φ, Θ не є незалежними:

$$R(\Phi,\Theta)=\sum_{t=1}^{|T|}\sum_{\omega=1}^{|\Omega|}\beta_{\omega t}\ln\phi_{\omega t}+\sum_{q=1}^{|Q|}\sum_{t=1}^{|T|}\alpha_{tq}\ln\theta_{tq}, \quad (8)$$

де $\beta_{\omega t}$ – числова оцінка, яка визначається на основі кількості релевантних термів; α_{tq} – числова оцінка, яка визначається на основі кількості релевантних сегментів тексту.

Для обчислення матриць Φ та Θ було використано інструмент для побудови тематичних моделей з відкритим кодом BigARTM [17]. Ймовірнісні тематичні моделі, які можуть бути застосовані в цьому інструменті описані в роботі [18]. Вхідними даними для тематичного моделювання в цьому інструменті є два файли, що будуються на основі колекції публікацій. Один файл складається з переліку всіх слів у канонізованих текстах публікацій: не враховуються стоп-слова, всі іменники у називному відмінку однини, дієслова у інфінітиві тощо. Другий файл представляється таблицею з полями: індекс публікації, індекс слова у першому файлі та кількість входжень слова у публікацію. Результатом роботи інструменту є матриці Φ та Θ.

Для розв'язання задачі встановлення повноти висвітлення матеріалів дисертації в наукових статтях потрібно зробити два кроки:

1. Провести навчання ймовірнісної тематичної моделі за публікаціями автора на основі інструменту BigARTM. Результати навчання зберігаються в окремому файлі в бінарному представленні.

2. Далі, використовуючи навчену модель та спеціальний регуляризатор (5), знаходимо матрицю Θ для сегментів авторефератів дисертацій.

Встановлено, що значення H_1 та H_2 мають бути близькі до $5\cdot10^{-5}$. Результати верифікації показують можливості застосування латентного семантичного аналізу для задач виявлення повноти висвітлення результатів дисертаційних досліджень здобувачами наукового ступеня [1].

Особливістю ймовірнісної тематичної моделі для задачі встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових публікаціях є використання навчання та спеціального регуляризатора. Результатом моделі є матриця належності тем, які визначаються сегментами авторефератів дисертації автора до документів, які визначаються публікаціями автора.

Особливістю латентного семантичного аналізу є те, що він використовується для широкого кола задач опрацювання текстової інформації. Розглянута в роботі задача ґрунтується на задачі максимізації функції правдоподібності, яка є некоректно поставленою. Для зведення задачі до коректно поставленої використовуються тільки відповідні регуляризатори. Інші методи зведення задач до коректних не розглядалися.

Обмеженням дослідження є проблема канонізації текстів різними мовами. У пропонованому дослідженні використовується текстова інформація українською мовою. В подальшому дослідженні буде запропоновано зведення текстів до однієї мовної бази, оскільки інструменти канонізації текстів англійської мови мають більш широкі можливості, зокрема для наукових публікацій. Також обмеженням є складність отримання повних текстів дисертацій для повноцінної верифікації моделі [1].

Результати дослідження використовуються при розробці системи виявлення неповних дублікатів у межах науково-дослідної роботи «Розробка комбінованих методів ідентифікації неповних дублікатів та виявлення повноти висвітлення наукових результатів дисертаційних досліджень, опублікованих автором», Київський національний університет будівництва і архітектури. № 0119U002579. Система інтегрує в собі також комбіновані методи виявлення неповних дублікатів, які описані в роботі [19].

Висновки

Здійснено огляд ймовірнісної тематичної моделі представлення текстових документів, зокрема наукових документів з використанням специфічних предметних термінів, які представляються n-грамами. Встановлено, що така модель може бути ефективно використана для розв'язання задачі виявлення повноти висвітлення результатів дисертаційних досліджень здобувачами наукового ступеня. Також здійснено формальне описання ймовірнісної тематичної моделі для задачі встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових статтях. Оскільки немає достатньо великої бази авторефератів дисертацій для проведення навчання, складно визначити оптимальні значення порогових коефіцієнтів H_1 та H_2. Проте при перевірці встановлення повноти висвітлення матеріалів дисертаційних досліджень автора в його наукових статтях (обсяг 20 матеріалів) [1] виявлено, що коефіцієнти H_1 та H_2 мають бути близькі до $5\cdot10^{-5}$.

Список літератури

1. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Yu., Biloshchytska, S. (2020). The use of probabilistic latent semantic analysis to identify scientific subject spaces and to evaluate the completeness of covering the results of dissertation studies. *Eastern-European Journal of Enterprise Technologies*, 4/4 (106), 14–20.

2. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Yu., 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.

3. Dumais, S. T. (2005). Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38, 188–230. doi: https://doi.org/10.1002/aris.1440380105.

4. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by Latent Semantic Analysis. *JASIS*, 41, 391–407.

5. Hofmann, T. (1999). Probabilistic Latent Semantic Indexing. In Proceedings of the Twenty-Second *Annual International SIGIR Conference on Research and Development in Information Retrieval (SIGIR-99)*, 289–296. arXiv:1301.6705

6. Dai, A. M., Olah, C., Le, Q. V. (2015). Document embedding with paragraph vectors. NIPS Deep Learning Workshop. arXiv:1507.07998v1

7. Rosen-Zvi, M., Gri ths, T., Steyvers, M., Smyth, P. (2004). The author-topic model for authors and documents. Proceedings of the 20th conference on Uncertainty in articial intelligence, 487–494.

8. Pagliardini, M., Gupta, P., Jaggi, M. (2018). Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics, 528–540. doi: https://doi.org/10.18653/v1/N18-1049

9. Lifchitz, A., Jhean-Larose, S., Denhiere, 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. PMID 19897829.

10. Galvez, R. H., Gravano, A. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*, 19, 1877–7503. doi: https://doi.org/10.1016/j.jocs.2017.01.001.

11. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Biloshchytska, S., Chala, L. (2016). Detection of near duplicates 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

12. Biloshchytskyi A., Kuchansky A., Biloshchytska S., Dubnytska A. (2017). Conceptual Model of Automatic System of Near Duplicates Detection on Electronic Documents. IEEE "The Experience of Designing and Applications of CAD Systems in Microelectron." (CADSM), P. 381-384.

13. Rossi, R. J. (2018). Mathematical Statistics: An Introduction to Likelihood Based Inference. New York: John Wiley & Sons.

14. Tikhonov, A., Arsenin, V. (1986). Methods for solving ill-posed problems. M: Nauka.

15. Blei, D. M., Ng, A. Y., Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of Machine Learning Research, 3, 993–1022.

16. Dietz, L., Bickel, S., Scheffer, T. (2007). Unsupervised prediction of citation influences. In Proceedings of the 24th international conference on Machine learning. ICML '07. New York, NY, USA: ACM, 233–240.

17. BigARTM. (2015). Retrieved from https://bigartm.readthedocs.io/en/stable/intro.html

18. Vorontsov, K. V. (2013). Probabilistic topic modeling. Retrieved from http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf

19. Lizunov, P., Biloshchytskyi, A., Kuchansky, A., Andrashko, Y., Biloshchytska, S. Serbin, O. (2021). Development of the combined method of identification of near duplicates in electronic scientific works. *Eastern-European Journal of Enterprise Technologies*, 4(4(112), 57–63. https://doi.org/10.15587/1729-4061.2021.238318.

Стаття надійшла до редколегії 22.09.2021

Lizunov Petro
DSc (Eng.), Professor, Head of the Department of Structural Mechanics, *orcid.org/0000-0003-2924-3025*
Kyiv National University of Construction and Architecture, Kyiv
Biloshchytskyi Andrii
DSc (Eng.), Professor, Vice-Rector for Science and Innovation, *orcid.org/0000-0001-9548-1959*
Astana IT University, Nur-Sultan
Kuchansky Alexander
DSc (Eng.), Associate Professor, Department of Information Systems and Technologies, *orcid.org/0000-0003-1277-8031*
Taras Shevchenko National University of Kyiv, Kyiv
Andrashko Yurii
PhD, Associate Professor, Department of Systems Analysis and Optimization Theory, *orcid.org/0000-0003-2306-8377*
Uzzhorod National University, Uzzhorod

Liashchenko Tamara
Lecturer, Department of Information Technology, *orcid.org/0000-0001-9092-0297*
Kyiv National University of Construction and Architecture, Kyiv

THE PROBLEM OF ESTABLISHING THE COMPLETENESS OF THE COVERAGE OF THE DISSERTATION RESEARCH RESULTS BY GRADUATES

Abstract. *The paper describes the possibilities of applying latent semantic analysis to identify the completeness of the coverage of the results of dissertation research by applicants for scientific degrees. To achieve this goal, the following tasks were set and achieved: a review of the probabilistic thematic model of presentation of text documents, in particular, scientific papers using specific subject terms, which are represented by n-grams; a formal description of the probabilistic thematic model for the problem of establishing the completeness of the coverage of the author's dissertation research materials in his scientific articles is given. A feature of the probabilistic thematic model for the problem of establishing the completeness of the coverage of the author's dissertation research materials in his scientific publications is training and a special regularizer. The result of the model is a matrix of belonging of the topics, which are determined by the segments of the author's dissertation abstracts to the documents, which are determined by the author's publications. The application of this model to this problem has not yet been described. The problem considered in the paper is based on the issue of maximizing the likelihood function, which is incorrectly posed. Only the appropriate regularizers are used to reduce the task to the correct one. Other methods of reducing tasks to the correct ones were not considered. A limitation of the study is the problem of the canonization of texts in different languages. This study uses textual information in the Ukrainian language. In further research, the reduction of texts to one language base will be offered. In particular, because the tools of canonization of English texts have more opportunities, particularly for scientific publications. Also, a limitation is the difficulty of obtaining full texts of dissertations for complete verification of the model. The research results are combined with the system of detection of incomplete duplicates in scientific documents, particularly dissertations for the degree.*

Keywords: dissertation; scientific research; scientific publication; latent semantic analysis

---

Посилання на публікацію

APA Lizunov, Petro, Biloshchytskyi, Andrii, Kuchansky, Alexander, Andrashko, Yurii & Liashchenko, Tamara. (2021). The problem of establishing the completeness of the coverage of the dissertation research results by graduates. *Management of Development of Complex Systems*, 47, 102–108, dx.doi.org\10.32347/2412-9933.2021.47.102-108.

ДСТУ Лізунов П. П., Білощицький А. О., Кучанський О. Ю., Андрашко Ю. В., Лященко Т. О. Задача встановлення повноти висвітлення результатів дисертаційних досліджень здобувачами наукових ступенів. *Управління розвитком складних систем.* Київ, 2021. № 47. С. 102 – 108, dx.doi.org\10.32347/2412-9933.2021.47.102-108.