УДК 519.766.48 DOI: 10.30929/1995-0519.2021.5.62-67

ВИКОРИСТАННЯ СТАТИСТИЧНОЇ МОДЕЛІ КОГЕРЕНТНОСТІ ЗВЯ'ЗНОГО ТЕКСТУ В ЯКОСТІ ДОДАТКОВОГО ІНСТРУМЕНТУ КІЛЬКІСНОГО КОНТЕНТ-АНАЛІЗУ

І. В. Шевченко, П. І. Андреєв, М. Г. Дернова
Кременчуцький національний університет імені Михайла Остроградського
ORCID: 0000-0003-3009-8611; 0000-0003-4368-9584; 0000-0003-4545-5247
Н. Ф. Хайрова
Національний технічний університет «Харківський політехнічний інститут»
ORCID: 0000-0002-9826-0286

Метою роботи є дослідження можливостей використання когерентності частотних характеристик абзаців для виявлення ключових слів та слів-сателітів, що оточують ключові слова, тобто контекстних множин. Для досягнення поставленої мети вирішені такі завдання: розроблено модель подання тексту, що відрізняється від наявних тим, що включає множину слів, які найбільш часто зустрічаються, множину ключових слів, множину слів-сателітів, перетин множин абзаців, ключових слів, та слів-сателітів, що дозволяє отримати формальну основу для побудови методу аналізу динаміки відносних частот слів, які найбільш часто зустрічаються у тексті та виявлення ключових слів і контекстних множин; розроблено метод аналізу тексту, якій відрізняється від існуючих тим, що в його основі лежить виявлення позитивних кореляцій між відносними частотами входження підмножини найбільш частих слів в абзацах, що дозволяє виявити ключові слова та контекстні підмножині у текстах, які мають властивість зв'язності та в окремих абзацах тексту, якій має слабку зв'язність. Розроблений метод можна використовувати як допоміжний інструмент контент-аналізу зв'язних текстів.

Ключові слова: контент-аналіз, модель тексту, когерентність, абзаці, відносні частоти, ключові слова, контекстна множина.

АКТУАЛЬНІСТЬ РОБОТИ. Сьогодні завдання автоматичної обробки текстів стає одним із центральних напрямів обробки інформації. Такі давно відомі системи, як автоматичний переклад, питання-відповіді та інформаційно-пошукові системи, поряд із більш сучасними, – чат-боти, системи автоматичної генерації текстів, виявлення плагіату, рекомендацій на основі контенту та іншими, мають величезну кількість як корпоративних, так і індивідуальних користувачів.

Наявні підходи, моделі та методи, що вирішуються на різних етапах аналізу зв'язних текстів в таких системах, можна умовно поділити на три групи. До першої групи відносяться методи, що базуються на граматичному, семантичному та лексичному аналізі та включаються до загального процесора обробки мови, так званого – pipeline Natural Language Processing (NLP). Ці підходи використовуються на етапі Part Of Speech (POS)-тегінга, синтаксичного парсеру, розпізнавання іменованих сутностей корпусів текстів. На етапі семантичного аналізу тут зазвичай використовуються спеціальні словники та лексичні бази даних або знань.

Моделі та методи другої групи використовують традиційні ймовірнісно-статистичні підходи, і сьогодні становлять досить велику частку моделей та алгоритмів машинного навчання та глибокого машинного навчання, які реалізуються при класифікації та кластеризації на будь-яких рівнях мови – від морфологічної розмітки слів до визначення тем повнотекстових документів.

До третьої групи методів можна віднести підходи, що використовують кількісний інструментарій та спеціалізований математичний апарат для розкриття правил функціонування одиниць мови та для встановлення закономірностей побудови тексту [1]. До цієї ж групи підходів можна віднести інструментарій контент-аналізу – методу якісно-кількісної обробки змісту документів з метою виявлення різних фактів та тенденцій, відображених у цих документах. При цьому, можливість застосування кількісних методів контент-аналізу ґрунтується на ймовірнісному характері мови та підтверджується дискретністю та масовістю мовних одиниць, можливістю вибору певного елемента із ряду однорідних.

У нашому дослідженні ми одночасно використовуємо два підходи: розглядаємо мову як складну ієрархічну систему та застосовуємо статистичні методі на одному з рівнів цій системи, а саме, для виявлення змістовної структури документу на базі розрахунку коефіцієнтів кореляції відносних частот слів між абзацами зв'язного тексту.

Ми розглядаємо мовну систему, як складну і таку, що погано формалізується. Це змушує досліджувати природну мову як сукупність мовних підсистем, що структуровані у вигляді складної семіотичної ієрархії, в якій зміст одиниць більш високого рівня не повністю зводиться до змістовних складових одиниць нижчого рівня. Отже, сенс одиниць вищого рівня не завжди може бути «обчислений» з урахуванням інформації про сенс одиниць нижчого рівня та інформації про зв'язки між цими одиницями. Одночасно структурна модель мовної системи використовує змістовні, тобто тематичні чи смислові ознаки зв'язності між одиницями одного рівня ієрархії. Це відкриває певні можливості для кількісного контент-аналізу.

Усі дослідження останнього часу в галузі інтелектуального аналізу текстів, комп'ютерної лінгвістики тією чи іншою мірою спираються на системний підхід до природної мови. Все більш поширена тенденція – розглядати зв'язний текст або корпус текстів як системну цілісність. Мовна система при такому підході базується на уявленні у вигляді множини природномовних елементів, що знаходяться у відносинах, пов'язані один з одним і утворюють певну єд-

ність та цілісність. Отже, використання методів системного аналізу [2] щодо моделюванні текстів є доцільним.

Застосування системного підходу виправдане, оскільки мова має всі властивості та характеристики, властиві складним системам. Можна виділити такі властивості та фундаментальні якості природної мови:

– принципова нечіткість значення мовних виразів;
– динамічність мовної системи;
– образність, заснована на метафоричності;
– семантична міць словника, що дозволяє виражати будь-яку інформацію за допомогою кінцевого набору елементів;
– гнучкість у передачі інформації.

Уживаність мовних елементів є проявом їхньої функціональної значущості в мові. Для оцінки цієї значущості необхідно використати певну кількісну міру. З усіх кількісних методів найбільші можливості на вирішення конкретних завдань і охоплення основних фактів мови є у ймовірнісно-статистичного аналізу. У основі використання імовірнісно-статистичного методу аналізу лежить уявлення про текст як послідовність випадкових подій, якими є конкретні вживання лінгвістичних одиниць [3].

Переважна більшість робіт зосереджена на пошуку ключових слів та їх контекстних множин для подальшого автоматичного анотування. У роботі [4] проведено кількісну та якісну оцінку методів відбору значущих слів. Метою дослідження була перевірка застосування ряду критеріїв для редукування множини ключових слів у колекції текстів, до якого згодом мають бути застосовані методи класифікації. У роботі [5] показано, що за допомогою відомих критеріїв Positive Pointwise Mutual Information (PPMI) та TF·IDF можна скоротити множину вхідних значущих слів для класифікації документів без втрати якості для дослідження. У роботі [6] запропоновано спосіб обчислення семантичної близькості між словами, який ґрунтується на зборі контекстних множин обраних слів. Використовується обчислення відносин частот слів у документах. Експериментальні результати свідчать, що метод є ефективним. Однак завдання відбору контекстних множин термів є складним і поки що повністю не вирішене.

У роботах [7, 8, 9] для підвищення інтерпретованості та визначення числа тем (аспектів) аналізованих текстів застосовуються імовірнісні тематичні моделі. В експериментах показано, що з їхньою допомогою вдається визначати семантичну близькість слів та інтерпретувати координати векторного простору як змістовні теми колекції документів.

Пошуку ключових слів присвячені роботи [10–14]. Зокрема у роботі [13] запропоновано модель пошуку ключових слів, яка базується на інформаційній оцінці результатів парсингу англомовних текстів та враховує результати аналізу граматичних зв'язків між лексичними одиницями, що дозволило формалізувати критерій якості процесу пошуку ключових слів. Запропонований метод дозволяє підвищити чисельні характеристики якості пошуку ключових слів, а саме повноту та точність.

Як видно, тема пошуку ключових слів та контекстних множин активно розробляється із залученням як імовірнісних підходів, так і синтаксичних закономірностей.

Розглядаючи результати відомих робіт, ми звернули увагу, що в жодній з них не використовується аналіз абзаців як самостійних структурних одиниць тексту, що мають надфразову єдність. Як відомо, абзац – компонент композиційної структури зв'язного тексту, який складається, як правило, з кількох речень, пов'язаних за змістом та граматично. В абзаці зазвичай розкривається одна мікротема тексту, що є у розвитку теми всього тексту. Абзац може включати не одну, а кілька мікротем. Тому той самий текст може бути розбитий на абзаци по-різному. Абзацне членування переважно залежить від комунікативного завдання автора, його індивідуального стилю. У науково-технічних текстах здебільшого використовується або аналітико-синтетичний тип абзацу, який містить аналітичну частину (пояснювальну, роз'яснювальну) у першій позиції, а узагальнюючу, підсумкову – у другій або синтетико-аналітичний абзац, що починається з узагальнюючої, стрижневої фрази, зміст якої розкривається у наступних повідомленнях.

Висловлено гіпотезу, що мають існувати певні закономірності у поступовій динаміці частот появи певних слів від одного абзацу до іншого, у разі, якщо досліджуваний текст має властивість когерентності (зв'язності), коли певна тематика грає роль лейтмотиву.

Метою даної роботи є дослідження можливості використання когерентності частотних характеристик абзаців для виявлення ключових слів та співсателітів, що оточують ключові слова, тобто контекстних множин.

Для досягнення поставленої мети мають бути вирішені такі завдання:

– розробка моделі тексту, що враховує завдання поабзацного аналізу динаміки відносних частот;
– розробка методу поабзацного аналізу тексту;
– випробування розробленого методу на колекції документів.

МАТЕРІАЛ І РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ. Для здійснення змістовного аналізу тексту необхідно мати модель тексту. Модель тексту – компактне символічне зображення структурної схеми елементів тексту [15]. Виходячи з цього, для побудови моделі тексту потрібно використовувати певну метамодель онтології. Саме онтологія предметної області (ПрО), що описується у тексті, є природною основою для побудови цієї моделі. З урахуванням цього представимо модель тексту у вигляді:

$$QL^{BO}=<E,AS,ER,EA,F,W(KW,SW),AW,KS,P,PW>, \quad (1)$$

де E – набір сутностей ПрО, ототожнений з множиною образів, які висловлюються набором ключових слів (КС); $ER \subseteq E \times E$ – множина відношень сутно-

стей; *AS* – множина тематичних аспектів опису ПрО; $EA \subseteq E \times AS$ – множина відношень сутностей та тематичних аспектів; $F: E \times ER$ – мовні інтерпретацій відношень сутностей; *W*(*KW, SW)* – множина слів, які найбільш часто зустрічаються та підлягають аналізу; *KW* – підмножина ключових слів; *SW* – підмножина слів-сателітів, що корелюють з КС; $AR \subseteq AS \times AS$ – перетин аспектів; $AW \subseteq AS \times KW$ – перетин аспектів та КС. $KS \subseteq KW \times SW$ – перетин КС та слів сателітів. Відповідно, за кожним аспектом закріплено певну підмножину КС та слів-сателітів; *P* – множина абзаців, що є структурними одиницями тексту; $PW \subseteq P \times KW \times SW$ – перетин абзаців, ключових слів та слів-сателітів.

Отже розроблено модель подання тексту, що відрізняється від існуючих тим, що включає множину слів, які найбільш часто зустрічаються, множину ключових слів, множину слів-сателітів, перетин множин абзаців, ключових слів, та слів-сателітів, що дозволяє отримати формальну основу для побудови методу аналізу динаміки відносних частот слів, що найбільш часто зустрічаються у тексті.

Метод поабзацного аналізу тексту. Нехай *T* – вхідний текст, в якому представлені словоформи $X = \{x_1, \dots, x_m\}$, що мають між собою приховані відношення $R=\{r_1, r_2, \dots, r_v\}$. Задача пошуку цих відношень може бути вирішена при застосуванні деякого оператора $\varphi: X \rightarrow RI$, де *RI* – суперпозиція відношень *R*, що виражені природною мовою.

Оператор φ повинен фіксувати асоціативні зв'язки між окремими словами. Оскільки асоціативні зв'язки мають різний характер, необхідно мати універсальний оператор φ, що працює з різними аспектами. З іншого боку, зміст, прихований в поточному фрагменті тексту X_i, залежить від передісторії, та передбачає деякі події надалі, так само, як значення числового часового ряду, що має приховану закономірність, залежать від попередніх значень ряду.

Розглянемо тепер поняття «абзац», як структурний елемент тексту. Абзац є одиницею членування тексту, проміжною між фразою і главою, і служить для угруповання однорідних одиниць викладу, вичерпуючи або підтримуючи один з його аспектів. Кожний абзац пов'язаний з іншими абзацами через ключові слова. Якщо побудувати матрицю *PW*, можна відстежити зв'язок між абзацами через КС та слова-сателіти.

З огляду на те, що в контексті розв'язуваної задачі значущими елементами тексту є абзаци і множина *W,* позначимо: N_A – кількість абзаців у тексті, що розглядається; f_{ik} – відносна частота входження *i*-го слова в *k*-й абзац; *kw* – ключове слово; *sw* – слово-сателіт.

Далі, виходячи з гіпотези, згідно з якою мають існувати певні закономірності у динаміці відносних частот слів, сформуємо процедурну модель, яка відображає етапи методу аналізу тексту. Для цього визначимо набір та послідовність процедур обробки тексту, необхідних для здійснення аналізу:

1. Процедура *PP* попередньої обробки – реалізує процедури видалення стоп-слів та стемінгу.

2. Процедура *FW* формування множини *W* – реалізує розрахунок частот входження слів у текст, ранжування слів за убуванням частоти та виділення певної частини слів, що мають максимальні частоти входження.

3. Процедура *SP* вторинної обробки – обчислення N_A – кількості абзаців, N_{wk} – кількості слів у кожному абзаці та значень f_{ik} – відносних частот слів з множини *W* у кожному абзаці. Результат – матриця відносних частот слів.

4. Процедура *CCM* розрахунку кореляційної матриці $R=|r_{ij}|$ відносних частот f_{ik}.

5. Процедура *KDP* визначення підмножини *KW* ключових слів. Алгоритм процедури містіть наступні кроки:

5.1 Встановлення порогу ступеню кореляції *T* для виділення пар слів, що мають максимальну поабзацну кореляцію відносних частот:

$$H_{ij} = \{1 \; if \; r_{ij} \geq T\}. \qquad (2)$$

5.2 Формування підмножини *RM* пар слів для яких H_{ij}=1.

5.3 Для кожного слова з підмножини *RM* розрахунок кількості входжень та ранжування слів за кількістю входжень у пари.

5.4 Відбір перших *n* слів у підмножину *KW*. Якщо серед цієї підмножини є слова, які створюють словосполучення, відновлення словосполучень здійснює користувач. У подальшому планується розробити метод автоматичного відновлення словосполучень.

6. Процедура *FCS* формування контекстних підмножин для кожного слова kw_i з підмножини *KW*. Алгоритм процедури містіть наступні кроки:

6.1 Формування початкової підмножини слів-сателітів $SW \subseteq W \backslash KW$.

6.2 Встановлення порогу ступеню кореляції *T* для виділення пар слів kw_i, sw_j, що мають позитивну поабзацну кореляцію відносних частот.

6.3 Формування контекстної підмножини SSW_i для слова kw_i з таких слів $sw_j \in SW$, для яких позитивна кореляція зі словом kw_i перевищує поріг *H*.

Кожна сформована підмножина SSW_i згідно з робочою гіпотезою є аспектною підмножиною.

7. Процедура *IKD* визначення інформаційного ядра абзацу. Алгоритм процедури містіть наступні кроки:

7.1 Обчислити добутки відносних частот f_{ik} та створити матрицю *LMA* зв'язків між частими словами.

7.2 Сформувати підмножину *RMA* пар слів для яких елементи матриці *LMA* є ненулевими.

7.3 Для кожного слова з підмножини *RMA* розрахувати кількість входжень та ранжувати слова за кількістю входжень у пари.

7.4 Відібрати перших *m* слів у підмножину ядра *IC*. Якщо серед цієї підмножини є слова, які створюють словосполучення, відновлення словосполучень здійснює користувач.

Зазначений набір процедур об'єднаємо в проце-

дурну модель *PM*, яка відображає етапи методу аналізу тексту:

$$PM = PP \rightarrow FW \rightarrow SP \rightarrow CCM \rightarrow KDP \rightarrow FCS \rightarrow IKD. \quad (3)$$

Отже, розроблено метод аналізу тексту, якій відрізняється від існуючих тим, що в його основі лежить виявлення позитивних кореляцій між відносними частотами входження підмножини найбільш частих слів в абзацах, що дозволяє виявити ключові слова та контекстні підмножині у текстах, які мають властивість зв'язності та в окремих абзацах тексту, якій має слабку зв'язність.

Експериментальні дослідження. Для перевірки працездатності методу аналізу тексту було сформовано набір україномовних, російськомовних та англомовних науково-технічних текстів – усього 20 текстів. В набір увійшли науково-технічні статті різної тематики і фрагменти навчальних посібників. Середній обсяг текстів складав близько 2200 слів. Середня кількість абзаців – близько 30. Результати машинного аналізу щодо виявлення ключових слів порівнювалися із авторськими наборами ключових слів у науково-технічних статтях. Для визначення наборів ключових слів фрагментів навчальних посібників залучалися експерти.

Порівняння авторських та експертних наборів ключових слів з наборами, які було сформовано за пропонованим методом показало його працездатність. Збіг складав від 50 % до 90 % з урахуванням того, що в авторських наборах були присутні словосполучення, а в машинних наборах елементи цих словосполучень показувалися окремо.

Достатньо очевидно, що результат виявлення ключових слів сильно залежить від змістовної зв'язності тексту. Якщо у тексті є «лейтмотив» – тема, яка «червоною ниткою» проходить скрізь абзаці, то набір адекватних ключових слів і, відповідно, їх контекстні підмножини, виявляються дуже успішно. Якщо текст має розмиту тему, наприклад, огляд досягнень штучного інтелекту, в якому мова їде про різні, мало пов'язані один з одним аспекті, набір ключових слів також є розмитим, а ключові слова, що мають велику кореляцію та об'єднують весь текст, складають малу частину виявлених слів. Втім, існує можливість в рамках запропонованого підходу здолати цей недолік. Для цього розроблено процедуру *IKD,* яка визначає інформаційне ядро кожного абзацу. У межах абзацу можливо виявити значущі слова, які не увійшли до підмножини *KW*. На основі цього, як нам здається, можна зробити висновки щодо важливості цих слів у даному абзаці та зробити окремі підмножини ключових слів для кожного абзацу. Але дослідження цього підходу тільки починається. Тим не менш, розроблений метод можна використовувати як допоміжний інструмент контент-аналізу зв'язних текстів.

ВИСНОВКИ. У проведеному дослідженні показано, що для виявлення ключових слів та контекстних (аспектних) підмножин слів можна використовувати показник кореляції від абзацу до абзацу відносних частот слів що найбільш часто зустрічаються.

Розроблено модель подання тексту, що відрізняється від наявних тим, що включає множину слів, які найбільш часто зустрічаються, множину ключових слів, множину слів-сателітів, перетин множин абзаців, ключових слів, та слів-сателітів, що дозволяє отримати формальну основу для побудови методу аналізу динаміки відносних частот слів, які найбільш часто зустрічаються у тексті та виявлення ключових слів і контекстних множин.

Розроблено метод аналізу тексту, якій відрізняється від існуючих тим, що в його основі лежить виявлення позитивних кореляцій між відносними частотами входження підмножини найбільш частих слів в абзацах, що дозволяє виявити ключові слова та контекстні підмножини у текстах, які мають властивість зв'язності та в окремих абзацах тексту, якій має слабку зв'язність.

Розроблений метод можна використовувати як допоміжний інструмент контент-аналізу зв'язних текстів.

ЛІТЕРАТУРА

1. Суркова А. С. Идентификация текстов на основе информационных портретов. *Вестник Нижегородского университета им. Н. И. Лобачевского*. 2014. № 3 (1). С. 145–149.

2. Згуровский М. З., Панкратова Н. Д. Системный анализ. Проблемы, методология, приложения. Киев : Наукова думка, 2011. 729 с.

3. Bolshakov I. A., Gelbukh A. Computational linguistics: models, resources, applications. Mexico, 2004. 186 p.

4. Калабин А. Л., Корнеева Е. И. Анализ информационных критериев отбора значимых признаков в методах text mining. *Proceedings of VSU, Series: Systems analysis and information technologies*. 2020. № 2. С. 150–159.

5. Pantel P., Lin D. Document clustering with committees. *Proceedings of the 25th Annual International ACM SIGIR Conference*. 2002. P. 199–206.

6. Бондарчук Д. В. Определение семантической близости термов с помощью контекстного множества. *Компьютерный анализ изображений: Интеллектуальные решения в промышленных сетях (CAI-2016) : сборник научных трудов по материалам I Международной конференции*. 2016. С. 175–179.

7. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования. *Машинное обучение и анализ данных.* 2013. Т. 1. № 6. С. 657–686.

8. Sidorov G., Gelbukh A., Gómez-Adorno H., Pinto D. Soft similarity and soft cosine measure: Similarity of features in vector space mode. *Computación y Sistemas*. 2014. Vol. 18. No. 3. P. 491–504.

9. Khairova N., Kolesnyk A., Mamyrbayev O., Petrasova S. Applying VSM to Identify the Criminal Meaning of Texts. *Proceedings of the Conference*

Computational Linguistics and Intelligent Systems, CoLInS 2020. 2020. P. 20–31.

10. Palshikar G. K. Keyword Extraction from a Single Document Using Centrality Measures. *Lecture Notes in Computer Science*. 2007. P. 503–510. DOI: https://doi.org/10.1007/978-3-540-77046-6_62.

11. Воронина И. Е., Кретов А. А., Попова И. В. Алгоритмы определения семантической близости ключевых слов по их окружению в тексте. Вестник ВГУ, Серия: Системный анализ и информационные технологии. 2010. № 1. С. 148–153.

12. Ванюшкин А. С., Гращенко Л. А. Методы и алгоритмы извлечения ключевых слов. *Новые информационные технологии в автоматизированных системах*. 2016. № 2. С. 85–93.

13. Бісікало О. В., Яхимович О. В. Метод визначення ключових слів англомовного тексту на основі DKPro Core. *Технологический аудит и резервы производства: Информационные технологии*. 2015. Том 1. № 2 (21). С. 26–30. DOI: https://doi.org/10.15587/2312-8372.2015.37274.

14. Бісікало О. В., Лісовенко А. І., Яхимович О. В., Траченко С. С. Визначення змістовних ознак тексту на основі аналізу зв'язків між лексичними одиницями. *Вісник НТУ «ХПІ». Серія «Механіко-технологічні системи та комплекси»*. 2015. № 21 (1130). С. 83–89.

15. Лосев А. Ф. Введение в общую теорию языковых моделей. Москва : Эдиториал УРСС, 2010. 296 с.

USE OF THE STATISTICAL MODEL OF COHERENCE OF CONNECTED TEXT AS AN ADDITIONAL TOOL OF QUANTITATIVE CONTENT ANALYSIS

I. Shevchenko, P. Andreev, M. Dernova
Kremenchuk Mykhailo Ostrohradskyi National University
ORCID: 0000-0003-3009-8611; 0000-0003-4368-9584; 0000-0003-4545-5247
N. Khairova
National Technical University «Kharkiv Polytechnic Institute»
ORCID: 0000-0002-9826-0286

Purpose. We consider the language system as a set of subsystems, structured in the form of a semiotic hierarchy, in which the content of higher-level units is not completely reduced to the substantive components of lower-level units. Therefore, the meaning of higher-level units cannot always be «calculated» taking into account information about the meaning of lower-level units and information about the relationships between these units. At the same time, the structural model of the language system uses thematic or semantic features of connectivity between units of one level of the hierarchy. This opens up certain possibilities for quantitative content analysis. **Methodology.** Considering the results of known works, we noticed that none of them uses the analysis of paragraphs as independent structural units of the text. The paragraph usually reveals one micro-theme of the text, which is in the development of the theme of the whole text. It is hypothesized that there should be certain patterns in the gradual dynamics of the frequencies of certain words from one paragraph to another, if the studied text has the property of coherence, when a certain topic plays the role of leitmotif. The aim of this work is to study the possibility of using the coherence of the frequency characteristics of paragraphs to identify keywords and satellite words surrounding the keywords – context sets. **Results.** To achieve this goal the following tasks are solved: development of a text model that takes into account the task of paragraph-by-paragraph analysis of the dynamics of relative frequencies; development of a method of paragraph-by-paragraph text analysis; testing of the developed method on a collection of documents. **Originality.** A text representation model has been developed that differs from the existing ones in that it includes a set of the most common words, a set of keywords, a set of satellite words, the intersection of sets of paragraphs, keywords, and satellite words. This provides a formal basis for building a method of analyzing the dynamics of relative frequencies of words that are most common in the text and identifying keywords and context sets. A method of text analysis has been developed, which differs from the existing ones in that it is based on the detection of positive correlations between the relative frequencies of occurrence of a subset of the most frequent words in paragraphs. This allows you to identify keywords and context subsets in texts that have some coherence and in individual paragraphs of text that have weak coherence. **Practical value.** A set of Ukrainian-language, Russian-language and English-language scientific and technical texts was formed to test the efficiency of the text analysis method. The set includes scientific and technical articles on various topics and fragments of textbooks. The results of machine analysis for keyword detection were compared with the author's sets of keywords in scientific and technical articles. Experts were involved to determine the keyword sets of the textbook fragments. Comparison of author's and expert sets of keywords with sets that were formed by the proposed method showed its efficiency. The match ranged from 50 % to 90 %, taking into account the fact that in the author's sets there were phrases, and in the machine sets the elements of these phrases were shown separately. The developed method can be used as an auxiliary tool for content analysis of related texts. References: 15.

Keywords: content analysis, text model, coherence, paragraphs, relative frequencies, keywords, context set.

REFERENCES

1. Surkova, A. S. (2014). Identifikatsiya tekstov na osnove informatsionnyh portretov [Text authorship attribution on the basis of information portraits]. *Vestnik of Lobachevsky University of Nizhni Novgorod*. No. 3 (1), pp. 145–149. [in Russian]

2. Zgurovskiy, M. Z., Pankratova, N. D. (2011).

Sistemnyy analiz: problemy, metodologiya, prilozheniya [System analysis. Problems, methodology, applications]. Kyiv, 728 p. [in Russian]

3. Bolshakov, I. A., Gelbukh, A. (2004). Computational linguistics: models, resources, applications. Mexico, 186 p.

4. Kalabin, A. L., Korneeva, Y. I. (2020). Analiz informatsionnyh kriteriev otbora znachimyh priznakov v metodah text mining [Analysis of information criteria of relevant feature selection in text mining methods]. *Proceedings of VSU, Series: Systems analysis and information technologies*. No. 2, pp. 150–159. [in Russian]

5. Pantel, P., Lin, D. (2002). Document clustering with committees. *Proceedings of the 25th Annual International ACM SIGIR Conference*. pp. 199–206.

6. Bondarchuk, D. V. (2016). Opredelenie semanticheskoy blizosti termov s pomoschyu kontekstnogo mnozhestva [Calculating the semantic relatedness of terms with the context set]. *Computer image analysis: Intelligent solutions in industrial networks (CAI-2016): collection of scientific papers based on the materials of the I International conference*. pp. 175–179. [in Russian]

7. Vorontsov, K. V., Potapenko, A. A. (2013). Modifikasii EM-algoritma dlya veroyatnostnogo tematiceskogo modelirovaniya [EM-like algorithms for probabilistic topic modeling]. *Machine Learning and Data Analysis*. T. 1. No. 6, pp. 657–686. [in Russian]

8. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space mode. *Computación y Sistemas*. Vol. 18. No. 3, pp. 491–504.

9. Khairova, N., Kolesnyk, A., Mamyrbayev, O., Petrasova, S. (2020). Applying VSM to Identify the Criminal Meaning of Texts. *Proceedings of the Conference Computational Linguistics and Intelligent Systems, CoLInS 2020*, pp. 20–31.

10 Palshikar, G. K. (2007). Keyword Extraction from a Single Document Using Centrality Measures. *Lecture Notes in Computer Science*. pp. 503–510. DOI: https://doi.org/10.1007/978-3-540-77046-6_62.

11. Voronina, I. E., Kretov, A. A., Popova, I. V. (2010). Algoritmy opredeleniya semanticheskoy blizosti klyuchevyh slov poi h oruzheniyu v tekste [Algorithms of semantic proximity assessment based on the lexical environment of the keywords in a text]. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*. No. 1, pp. 148–153. [in Russian]

12. Vanyushkin, A. S., Grashchenko, L. A. (2016). Metody i algoritmy izvlecheniya klyuchevyh slov [Methods and algorithms for extracting keywords]. *New information technologies in automated systems*. No 2, pp. 85–93. [in Russian]

13. Bisikalo, O., Yahimovich, A. (2015). Metod vyznachennia kliuchovykh sliv anhlomovnoho tekstu na osnovi DKPro Core [Method of assigning key words to English text based on DKPro Core]. *Technology audit and production reserves: Information Technology*. Vol. 1. No. 2 (21), pp. 26–30. DOI: https://doi.org/10.15587/2312-8372.2015.37274 [in Ukrainian]

14. Bisikalo, O. V., Lisovenko, A. I., Yakhymovych O. V., Trachenko S. S. (2015). Vyznachennia zmistovnykh oznak tekstu na osnovi analizu zviazkiv mizh leksychnymy odynytsiamy [Determining the content of the text based on the analysis of relationships between lexical items]. *Bulletin of the National Technical University «KhPI». Series «Mechanical-technological systems and complexes»*. No. 21 (1130), pp. 83–89. [in Ukrainian]

15. Losev, A. F. (2010). Vvedenie v obshhuju teoriju jazykovyh modelej [Introduction to the general theory of language models]. Moskva, 296 p. [in Russian]

Стаття надійшла 10.09.2021