e-ISSN 1607-3274 Радіоелектроніка, інформатика, управління. 2020. № 1
p-ISSN 2313-688X Radio Electronics, Computer Science, Control. 2020. № 1

ПРОГРЕСИВНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

PROGRESSIVE INFORMATION TECHNOLOGIES

ПРОГРЕССИВНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

UDC 004.9

МЕТОД АВТОРИФІКАЦІЇ ТЕКСТУ НАУКОВО-ТЕХНІЧНИХ ПУБЛІКАЦІЙ НА ОСНОВІ ЛІНГВІСТИЧНОГО АНАЛІЗУ КОЕФІЦІЄНТІВ МОВНОЇ РІЗНОМАНІТНОСТІ

Висоцька В. А. – канд. техн. наук, доцент, доцент кафедри «Інформаційні системи та мережі», Національний університет «Львівська політехніка», Львів, Україна.

АНОТАЦІЯ

Актуальність. Авторифікація авторства тексту є технікою визначення автора тексту, коли неоднозначно, хто її написав. Це корисно, коли декілька людей претендують на авторство однієї публікації або у випадках, коли ніхто не претендує на авторство текстового контенту, наприклад, так звані тролі в соціальних мережах під час інформаційної війни. Складність проблеми авторського тексту, очевидно, експоненціально вища, більша кількість вірогідних авторів. Наявність авторських текстових зразків також є суттєвою при просуненні цієї проблеми. Атрибуція авторського тексту включає наступні три проблеми:

– виявлення автора текстового автора з групи імовірних або очікуваних авторів, де автор завжди знаходиться у групі підозрюваних;

– не ідентифікація автора текстового автора з групи вірогідних або очікуваних авторів, де автор може не бути в групі підозрюваних;

– оцінка можливості даного тексту, написаного даним автором чи ні.

Тому задача автоматичного визначення автора текстового контенту науково-технічного спрямування є актуальною й потребує нових (досконаліших) підходів до її розв'язування.

Метою дослідження є розроблення методу визначення автора у україномовних текстах на основі технології лінгвометрії.

Метод. Розроблено лінгвометричний метод алгоритмічного забезпечення процесів контент-моніторінгу для розв'язання задачі автоматичного визначення автора україномовного текстового контенту на основі технології статистичного аналізу коефіцієнтів мовної різноманітності. Проведено декомпозицію методу визначення автора на основі аналізу таких коефіцієнтів мовлення як лексична різноманітність, ступінь (міра) синтаксичної складності, зв'язність мовлення, індекси винятковості та концентрації тексту. Проаналізовані також параметри авторського стилю як кількість слів у певному тексті, загальна кількість слів цього тексту, кількість речень, кількість прийменників, кількість сполучників, кількість слів із частотою 1, та кількість слів із частотою 10 та більше. Особливостями розробленого є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текстів. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частини мови. Для цього провадився аналіз флексій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації.

Результати. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Висновки. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури.

КЛЮЧОВІ СЛОВА: текстовий контент, NLP, контент-моніторінг, стоп-слова, контент-аналіз, статистийний лінгвістичний аналіз, квантитативна лінгвістика.

АБРЕВІАТУРА

ІС – інформаційна система;

ІТ – інформаційна технологія.

DOI 10.15588/1607-3274-2020-1-12

НОМЕНКЛАТУРА

S – система визначення автора;

A – множина авторських статтей;

D – множина статтей для дослідження;

K – множина коефіцієнтів авторського мовлення;

H – множина коефіцієнтів мовлення невідомого автора досліджуваного тексту;

L – словник службових слів;

C – результати порівняння коефіцієнтів мовлення відомих авторів та досліджуванного тексту;

α – оператор парсингу тексту для визначення множини параметрів авторського мовлення;

μ – оператор визначення коефіцієнтів авторського мовлення відомих/досліджуваних публікацій;

ψ – оператор порівняння коефіцієнтів авторського мовлення відомих/досліджуваних публікацій;

χ – оператор формування множини публікацій з подібними значеннями коефіцієнтів авторського мовлення;

ω – оператор машинного навчання системи на основі попередньо зібраної статистики розрахунків коефіцієнтів авторського мовлення;

θ – оператор розрахунку ступення належності досліджуваного тексту конкретному автору з множини потенційних авторів;

K_r – коефіцієнт розповсюдженості;

N_p – відношення кількості підвибірок з певною лінгвістичною одиницею;

N_z – загальна кількість підвибірок;

K_l – коефіцієнт лексичної різноманітності;

W – кількість слів у певному тексті;

N – загальна кількість слів цього тексту;

K_s – коефіцієнт синтаксичної складності;

P – кількість окремих речень;

W – кількість слів у всьому тексті;

K_z – коефіцієнт зв'язності мовлення;

Z – кількість прийменників;

S – кількість сполучників;

I_{wt} – індекс винятковості тексту;

W_1 – кількість слів із частотою 1;

I_{kt} – індекс концентрації тексту;

W_{10} – кількість слів із частотою 10 та більше;

F – частота слова в частотному словнику;

i – ранг слова в частотному словнику;

k – довжина слова у фонемах;

C – стала;

r – ранг слова у фонемах;

m – кількість значень слова;

f – частота слова;

y – середня довжина складових;

x – довжина мовної конструкції;

b – показник, що характеризує динаміку зміни довжини складників (закон діє, якщо $b<0$);

p_x – ймовірність використання слова, яке має x значень;

ω – середня кількість значень слова у словнику.

ВСТУП

Важливими завданнями мовознавства на основі лінгвометрії є створення і порівняння словників (у тому числі частотних та статистичних), автоматичних словників, тезаурусів, систем стенографії, автоматичне визначення мови, інформаційний пошук тощо [1]. Наприклад, для моделювання процесів інформаційного пошуку знаходять статистичні і перехідні ймовірності морфем тексту [2]. На основі побудованих таблиць моделюють перевірку досліджуваного слова на наявність помилки, пропонують кілька найбільш ймовірних варіантів [3]. Лінгвометрія – галузь прикладної лінгвістики, що виявляє, вимірює та аналізує кількісні характеристики одиниць різних рівнів мови чи мовлення. Одним зі способів охарактеризувати літературне багатство тексту є оцінювання характеру використання мовних одиниць на всіх мовних рівнях. Це дає змогу ототожнювати поняття багатство і різноманітність мовлення. У свою чергу стилеметрія як підрозділ прикладної лінгвістики виявляє та аналізує кількісні характеристики певного функціонального стилю мови чи мовлення авторів текстового контенту, тобто авторської атрибуції [4]. Атрибуція полягає у визначенні методома квантитативної лінгвістики достовірності, автентичності авторського твору, його автора, місця й часу створення на основі аналізу технологічних і стилістичних закономірностей та особливостей коефіцієнтів мовної різноманітності конкретного авторач і/або конкрентного текстового твору [5]. Наприклад, однією із відомих мовознавчих проблем є процес визначення авторської атрибуції уривків певного текстового контенту [6]. Для цього обчислють частоти слововживань у запропонованих уривках [7]. Використовуючи частотні словники авторської творчості загалом чи окремих його творів, визначають автора твору (або твір – якщо це дозволяє словник) [8]. Недоліком є збереження або автоматичне генерування великих масивів даних у вигляді частотних словників авторських творів [9]. Опрацювання таких словників вимагає багато часу, а збереження – багато ресурсів [10]. У свою чергу, є автори з малочисельною творчістю, що унеможливлю точне відворення результатів аналізу авторської атрибуції [11]. Відомий метод датування для визначення тривалості роздільного існування двох споріднених мов, ґрунтується на припущенні про те, що основна частина лексичного складу будь-якої мови (ядерна лексика) змінюється з однаковою швидкістю і вимагає підрахунку процентного співвідношення спільних елементів у основному словнику [12]. Модифіковані методи глотохронології застосовують для визначення динаміки зміни авторського мовлення в його текстовому контенті на протязі тривалого часу для датування наближеного періоду, в якому був створений конкртниий текст твору цього автора [13]. Тому задача автоматичного визначення автора текстового контенту є актуальною й потребує нових (досконаліших) підходів до її

DOI 10.15588/1607-3274-2020-1-12

розв'язування, наприклад, на основі статистичного аналізу коефіцієнтів мовної різноманітності [14].

Метою дослідження є розроблення методу визначення автора в україномовних текстах на основі технології лінгвометрії. Для досягнення мети були поставлені такі завдання:

– на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному тексті розробити алгоритми визначення автора тексту;

– розробити програмне забезпечення контент-моніторінгу для визначення автора в україномовних текстах на основі лінгвометричного аналізу визначенних стопових слів текстового контенту;

– здійснити аналіз результатів експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю.

1 ПОСТАНОВКА ПРОБЛЕМИ

Систему визначення автора в україномовних текстах на основі технології лінгвометрії подамо як кортеж $S=<A, D, L, K, H, C, \alpha, \mu, \psi, \chi, \theta, \omega>$. Вагомим значенням у квантитативній лінгвостатистиці є розподіл (дистрибуція) лінгвістичної одиниці у тексті – присутність лінгвістичної одиниці в різних (зазвичай рівних) підвибірках (уривках) [15]. Якщо досліджувана лінгвістична одиниця функціонує тільки в одній нідвибірці, хоча й з високою частотою, то така вибірка є нерепрезентативною стосовно цієї лінгвістичної одиниці [16]. Важливо, коли досліджувана лінгвістична одиниця є рівномірно розподіленою в генеральній сукупності на прикладі науково-технічни статтей україномовних публікацій [17]. Відповідно визначають критерії для розрахунку коефіцієнтів авторського мовлення як $K=\mu(L)°\alpha(A)$ або $H=\mu(L)°\alpha(D)$, де $K=\{ K_r, K_l, K_s, K_z, I_{wt}, I_{kt}\}$ та $H=\{K_r, K_l, K_s, K_z, I_{wt}, I_{kt}\}$ відповідно авторського відомого тексту та досліджуваного тексту статті відповідно. Для цього аналізують коефіцієнт розповсюдженості [18]: $K_r=N_p/N_z$.

Проте характеристики, одержані на матеріалі вибірки, зазвичай відрізнятються від реальних характеристик генеральної сукупності, оскільки завжди присутня в квантитативній лінгвостатистиці відносна неточність дослідження [19]. Розподіл частоти лінгвістичних одиниць мови в текстовому контенті має певну регулярність і утворює його статистичну (частотну, ймовірнісну) структуру [20]. Такий розподіл є відмінним для кожної з мовних елементів – лексем, морфем, фонем тощо [21]. Тому лінгвостатистичні параметри авторських стилів, встановлені на різних рівнях (фонемних, морфемних, *N*-грамних, лексемних тощо), мають неоднакову стилеідентифікованму потужність автоського мовлення для різних пар стилів [22]. Наприклад, споріднені стилі чіткіше розмежовані на синтаксичному рівні, а менш споріднені – на лексичному [23]. Для цього автоматично створюють частотні словники певних лінгвістичних одиниць та завдяки ним аналізують середню повторюваність слова в тексті, коефіцієнт hapax legomena (слова, які мають частоту 1 у досліджуваній вибірці), індекс винятковості, індекс концентрації тощо [1–5, 14, 24].

Розрахунок коефіцієнтів мовної різноманітності повинен припускати взаємозв'язок таких коефіцієнтів, як:

– лексична різноманітність ($K_l=W/N$): відношення кількості слів до загальної кількості словоформ тексту, значення коефіцієнта лежить у межах [0;1];

– ступінь (міра) синтаксичної складності ($K_s=1–P/W$): відношення кількості речень до кількості слів певного тексту [14];

– зв'язність мовлення ($K_z=(Z+S)/(3P)$): відношення кількості прийменників і сполучників до кількості окремих речень;

– індекс винятковості тексту ($I_{wt}=W_1/W$): варіативність лексики, тобто частка тексту, яку займають слова, що трапилися 1 раз;

– індекс концентрації тексту ($I_{kt}=W_{10}/W$): частка тексту, яку займають слова, що трапилися 10 разів і більше [14].

Оскільки коефіцієнт – величина абсолютна, можна у певних межах нехтувати довжиною порівнюваних текстів [46]. Теоретичний інтерес складає дослідження внутрішньої «динаміки» тексту в частині співставлення коефіцієнтів з різних його ділянок між собою та із загальним для всього тексту коефіцієнтом [1–5, 14, 41, 47]:

– для лексичної різноманітності чим більшим є отримуваний десятковий дріб, тим вищою є лексична різноманітність досліджуваного тексту [14];

– для синтаксичної складності чим більшим є дріб (в межах [0; 1]), тим багатослівнішими загалом є речення такого тексту, а отже, – вища можливість різноманітності синтаксичних відношень між словами в окремому реченні [14];

– для зв'язності мовлення дорівнює одиниці, коли в одному реченні є три сполучні елементи (прийменники і сполучники) [14].

Далі необхідно розрахувати значення ступенів приналежності досліджуваного тексту відповідним авторам із списку подібних за значеннями коефіцієнтами мовлення в допустимих межах:

$$C=\omega(K,H)°\theta(K,H)°\chi(K,H)°\psi(K,H).$$

Результатом функціонування системи буде рангований список потенційних аторів досліджуваного тексту україномовної науково-технічної публікації. Зменшення позицій тексту суттєво залежить від кількості публікації авторів, часового проміжку самих публікації, наявності достовірних даних про приналежність тексту конкретному автору, обсягу статистичних даних для машинного навчання системи для формування частотних словників використання сд\лужбових слів конкретним автором.

DOI 10.15588/1607-3274-2020-1-12

2 АНАЛІЗ ЛІТЕРАТУРНИХ ДЖЕРЕЛ

За даними ЧС обчислюють такі характеристики як багатство словника, індекс різноманітності (K_l) – відношення обсягу словника лексем (W) до обсягу тексту (N), тобто $K_l=W/N$. Згідно табл. 1 найрізноманітніша, найбагатша лексика – у поезії, далі за спадом – у художній прозі, розмовно-побутовому стилі, публіцистиці, науковому та офіційно-діловому стилі [14, 25].

Середня повторюваність слова у тексті A є відношенням обсягу тексту N до обсягу словника лексем W (обернена до індексу різноманітності), тобто $A=N/W$ [26]. За даними ЧС, кожне слово у розмовно-побутовому стилі в середньому вжито 14 разів, а в науковому стилі – 17 [27].

Таблиця 1 – Результати коефіцієнтів мовлення згідно стилів української мови [14]

Стиль	W/N	W_1/N	W_1/W	W_{10}/W	W_{10t}/N
науковий	0,059	0,427	0,025	0,189	0,890
публіцистичний	0,070	0,450	0,031	0,121	0,804
діловий	0,030	0,280	0,0085	0,303	0,935
поетичний	0,103	0,495	0,052	0,098	0,789
художньої прози	0,067	0,430	0,029	0,149	0,821
розмовний	0,073	0,465	0,034	0,161	0,789

Індекс винятковості характеризує варіативність лексики, тобто частку тексту (словника), яку займають слова, що трапилися 1 раз (табл. 1) [28]:

– словника I_{wt} – відношення кількості лексем із частотою 1 W_1 до загальної кількості лексем: $I_{wt}=W_1/W$ [14];

– тексту I_t відношення кількості лексем із частотою 1 W_1 до обсягу тексту N: $I_t=W_1/N$ [14].

Індекс концентрації вказує на частку тексту (словника), яку займають слова, що трапилися 10 разів і більше (табл. 1) [29]:

– словника I_{kt} – відношення кількості слів у словнику з абсолютною частотою 10 і більше (W_{10}) до загальної кількості слів у словнику (W): $I_{kt}=W_{10}/W$ [14];

– тексту I_{tn} – відношення суми абсолютних частот слів з абсолютною частотою 10 і більше W_{10t} до обсягу тексту N: $I_{tn}=W_{10t}/N$ [14].

Мовлення надає перевагу невеликій кількості одиниць, які часто використовують [30]. Формують ядро будь-якої мовленнєвої підсистеми, тоді як переважна кількість одиниць є низькочастотними [31]. Цю закономірність зауважив ще учений Дьюї на поч. XX ст., назвавши її законом переваги [32]. Детальніше дослідив цю закономірність німецький мовознавець Дж. Ціпф, сформулювавши закон Zipf's law, який встановлює залежності [33]:

– частоти слова та його рангу у словнику: у частотнішого слово вищий ранг при $F\cdot i$=const [34];

– частоти слова та його довжини: чим частотніше слово, тим воно коротше при $k=C\lg r$ [35];

– частоти слова та кількості його значень: чим частотніше слово, тим воно багатозначніше при $m=C\sqrt{f}$ [36];

– частоти слова та його походження: чим давніше слово, тим воно частотніше [37].

Згідно закону німецького мовознавця П. Менцерата довжина мовної конструкції (слова, словосполучення, надфразової єдності, речення) обернено пропорційна до довжини її складових (складів, слів, словосполучень і т. д.), тобто чим довша мовна конструкція, тим коротші її складові [14]. Згідно досліджень Ґ. Альтманна $y=ax^b$.

Закон Крилова встановлює залежність між кількістю багатозначних слів та частотою [14]:

$$p_x=1/2^x,\quad px=(\omega-1)^{x-1}/\omega^x.$$

Деякі основні кількісні характеристики мови дуже прості. Наприклад, різниця між кількістю слів (10^4–10^5), кількістю морфем (декілька тисяч), кількістю складів (від декількох сотень до декількох тисяч) і кількістю фонем (від 10 до 80) [31–37]. Висловлюють припущення, що такі співвідношення пов'язані із властивістю людської пам'яті. Зазначимо також, що чим частотніше слово, тим швидше людина його зможе пригадати. Однак відсутні дослідження в галузі залежності змін коефіцієнтів лексичного авторського мовлення на протязі періоду його творчості.

3 МАТЕРІАЛИ ТА МЕТОДИ

Виявлено [14], що текст україномовної казки має K_z=0,77, а текст україномовної наукової статті – 3,0, тобто зв'язність у другому тексті у 3,9 разів сильніша, ніж у першому. Офіційних стандартів для коефіцієнтів різноманітності мовлення для K_l та K_s не існує, але орієнтиром для співставлення та оцінювання якогось тексту в однорідній групі текстів є середньостатистична норма величини коефіцієнта для рівних за довжиною уривків. Мінімальним розміром (довжиною) уривка приймемо 100 слів, вважатимемо, що коефіцієнти тут уже стабілізуються, відображаючи реальні особливості мови автора. Близькість або віддаленість окремого індивідуального коефіцієнта від середнього служить основою для оцінювання різноманітності мовлення у відповідному тексті. Задовільними вважаються тексти, коефіцієнти різноманітності яких потрапляють у зону середніх квадратичних відхилень D від певного середнього

$$D=\sqrt{\frac{\sum_{i=1}^{n}x_i^2}{n}-(\overline{x})^2}\ \text{[14]}.$$

Аналіз та інтерпретація на лінгвістичному рівні стилістичних особливостей і закономірностей письменницького стилю певного автора (або певної літературної епохи) включає подані найосновніші етапи, подані в алг. 1 [14].

Алгоритм 1. Аналіз та інтерпретація на лінгвістичному рівні стилістичних особливостей і

DOI 10.15588/1607-3274-2020-1-12

закономірностей письменницького стилю певного автора

Етап 1. Відбір та первинне опрацювання текстового контенту. Для відбору будують фільтри тексту за параметрами (основна мова тексту, обсяг текстової вибірки, часовий проміжок публікації, джерео публікації, формат тощо). Основними кроками первинного опрацювання тексту є:

– приведення його до єдиного формату (наприклад усунення тегів, якщо попередня публікація є у Інтернет-ресурсі у виглдяі статичної сторінки);

– усунення інформаційного шуму (рисунків, формул, список літеру тари, анотації іншими мовами тощо), який не впливає на результат, але збільшує час опрацювання;

– приведення до єдиного обсягу (скорочення у разі потреби, забираючи неінформативні ділянки початку та закінчення тексту).

Етап 2. Лематизація текстових лінгвістичних одиниць. Об'єднання словоформ під лемою мови [14].

Етап 3. Усунення неоднорідності текстових лінгвістичних одиниць. Розв'язання проблеми неоднорідності текстових лінгвістичних одиниць, наприклад, із погляду відношення до різних видів мови (авторська, не авторська і т. п.).

Етап 4. Побудова системи частотних словників, організація на основі статистичних розподілів у потрібних частотних шкалах. Частотний словник – тип словника, де наведено кількість вживань (частоту) певної лінгвістичної одиниці мови (складу, слова, словоформи, словосполучення, ідіоми, фразеологізму) в різних текстах певного обсягу. Зазвичай, подають абсолютну та відносну частоту вживання мовних одиниць, словникові статті розміщують за спаданням частот.

Етап 5. Пошук параметрів, що адекватно відображають структуру частотного словника. Такі параметри дають змогу сформулювати кілька основних лінгвостатистичних методів дослідження тексту:

– метод опорних слів (підрахунок загальної частоти вживання та знаходження відсоткового складу службових слів: прийменників, сполучників, часток);

– метод розділових знаків (підрахунок лише кількості внутрішніх і зовнішніх розділових знаків);

– метод слів (підрахунок лише слів певної довжини);

– метод речень (підрахунок лише речень визначеної довжини);

– синтаксичний метод (підрахунок розділових знаків, слів і речень певної довжини);

– комбінований (поєднання синтаксичного методу і опорних слів).

Етап 6. Перевірка параметрів на ефективність. Аналіз та порівння отримани результатів на відомих авторських творах для визначення закономірностей впливу авторської стилістики на формування авторської структури частотного словника за цими параметрами.

Етап 7. Математичне моделювання лексикостатистичних розподілів.

Етап 8. Побудова статистичних класифікацій, тобто авторських еталонів, що відображають стилістичні закономірності в межах творів певного автора чи певної літературного стилю з врахуванням літературної епохи та особливостей мови, на якій написані самі аналізовані твори.

Етап 9. Інтерпретація результатів із позицій стилістичних уявлень у визначеному часовому проміжку, загальної й авторської стилістики з врахуванням часових параметрів. Таким чином також вирішемо завдання авторської атрибуції, яке сформулюємо наступним чином. Нехай існує статистично опрацьований доробок автора (еталон). Необхідно оцінити належність певних уривків до еталону із застосуванням відповідних методів. Графічне зображення відносної частоти появи службових слів в Уривку 4 та в еталоні подане на рис. 1. Коефіцієнт кореляції для службових слів у цьому випадку складає R_{e-U4}=0,7326. Наведемо також коефіцієнти кореляції для кожного зі службових слів для уривків 1–4 (табл. 4). Аналізуючи коефіцієнти кореляції для службових слів, приходимо до висновку, що ймовірність належності уривків до досліджуваного еталону найбільшою є для Уривку 4, за ним – Уривок 2, Уривок 1, Уривок 3. Зауважимо, що для всіх чотирьох уривків простежуються стабільно високі коефіцієнти кореляції для часток, що можемо розуміти як відсутність впливу часток на авторський стиль. Додатково для уривків проаналізуємо частотності поав лише прийменників і сполучників, знайдемо відповідні коефіцієнти кореляції та порівняємо результати (табл. 2).

Рисунок 1 – Відносна частота появи службових слів в Уривку 4 та в еталоні

DOI 10.15588/1607-3274-2020-1-12

Таблиця 2 – Коефіцієнти кореляції для службової частини мови та кожного з уривків

№	Прийменник	Сполучник	Частка	R_{e-U}	R'_{e-U}
1	0,72	0,79	1	0,6076	0,6900
2	0,4928	0,5714	0,9580	0,7066	0,4913
3	0,1517	0,1624	0,8800	0,2810	0,2254
4	0,5639	0,9544	0,9594	0,7326	0,6905

Уривок 4 так і залишився найімовірнішим кандидатом щодо належності його до еталону, а наступним із незначним відривом став Уривок 1, далі – Уривок 2. Уривок 3, як і у попередньому дослідженні, має найменшу ймовірність належати до еталону. Для підтвердження результатів звернемося до [1–4], з яких узято уривки для дослідження.

4 ЕКСПЕРИМЕНТИ

Під час дослідження розроблено систему з можливістю обрання мови/мов аналізованого контенту, реалізована на Web-ресурсі Victana (рис. 2). Аналізуючи складові формул для оцінки багатства твору, приходимо до висновку, що треба знайти такі величини як кількість слів і словоформ, речень, сполучників і прийменників, слів із частотою 1 та не меншою за 10. На сервері після запуску процесу розрахунку коефіцієнтів різноманітності тексту запускається алгоритм аналізу цього тексту (алг. 2).

Алгоритм 2. Аналізу стилю автоського мовлення.

Етап 1. Перевірка довжини тексту – лишнє відсікається.

Етап 2. Очищення досліджуваного тексту (цифри, спецсимволи, формули, рисунки).

Етап 3. Визначення кількості речень P.

Етап 4. Визначення кількості слів у тексті N.

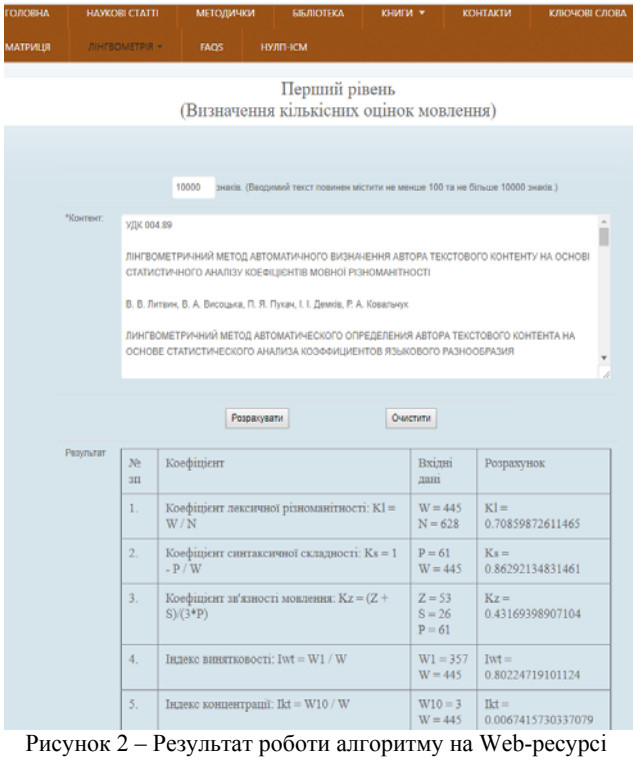

Рисунок 2 – Результат роботи алгоритму на Web-ресурсі Victana (http://victana.lviv.ua/nlp/linhvometriia)

DOI 10.15588/1607-3274-2020-1-12

Етап 5. Визначення кількості слів W (за частотним словником основ слів).

Етап 6. Розрахунок коефіцієнта лексичної різноманітності: $K_l=W/N$.

Етап 7. Розрахунок коефіцієнта синтаксичної складності: $K_s=1–P/W$.

Етап 8. Визначення кількості слів, що зустрілися точно один раз, тобто W_1.

Етап 9. Розрахунок індексу винятковості тексту : $I_{wt}=W_1/W$.

Етап 10. Визначення кількості слів, що зустрілися більше 9 разів, тобто W_{10}.

Етап 11. Розрахунок індексу концентрації тексту: $I_{kt}=W_{10}/W$.

Етап 12. Визначення кількості прийменників Z.

Етап 13. Визначення кількості сполучників S.

Етап 14. Розрахунок коефіцієнта зв'язності мовлення: $K_z=(Z+S)/(3\cdot P)$.

Етап 15. Виведення результатів на ресурсі Victana.

Аналізуючи складові формул для оцінення багатства твору, бачимо, що треба знайти кількість речень, слів і словоформ, прийменників і сполучників, слів із частотою 1 та частотою, не меншою за 10. Для зручності внесемо знайдені дані у таблицю. На інформаційному ресурсі передається сформована таблиця (табл. 3) та отримані результати дослідження виводяться на екран. Спираючись на викладене вище, оцінимо багатство уривків творів одноосібних наукових статей технічного спрямування Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» за період 2001–2017 рр. за допомогою коефіцієнтів різноманітності та зв'язності мовлення, індексів винятковості та концентрації тексту. Для аналізу виберемо частину першу (10000 знаків) кожної статті (алг. 3).

Алгоритм 3. Аналіз статистики функціонування системи виявлення множини стопових слів із 215 наукових статей технічного спрямування

Етап 1. Аналіз 100 наукових статей на визначення діапазону оптимального розміру досліджуваного тексту. Спочатку були проаналізовані тексти в повному обсязі, а потім ці тексти були проаналізовані на різні величини знаків. Результати показали, що

Таблиця 3 – Приклад результату роботи алгоритму аналізу стилю автора публікації на ресурсі Victana

Коефіцієнт	Дані	Розрахунок
лексичної різноманітності: $K_l=W/N$	W=184; N=295	K_l=0,6237
синтаксичної складності: $K_s=1–P/W$	P=18; W=184	K_s=0,902
зв'язності мовлення: $K_z=(Z+S)/(3\cdot P)$	Z=20; S=28; P=18	K_z=0,889
винятковості: $I_{wt}=W_1/W$	W_1=141; W=184	I_{wt}=0,7663
концентрації: $I_{kt}=W_{10}/W$	W_{10}=2; W=184	I_{kt}=0,01

оптимальним дослідженням текстів є діапазон [100;10000] знаків. Менше 100 знаків – неінформативна отримана інформація, часто значення коефіцієнтів різних авторів подібні, а одного ж автора на різних тестах – суттєво різняться. Якщо більше 10000 знаків – суттєво коефіцієнти вже не змінюються, але аналоги для дослідження мають різну довжину і з-за браку різноманітності аналогів великої довжини, було обрано максимальне число для аналізу 10000.

Етап 2. Аналіз понад 200 одноосібних робіт технічного спрямування понад 50 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 3. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу.

Етап 4. Аналіз понад 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення стилів мовлення цих авторів.

Етап 5. Аналіз отриманих коефіцієнтів мовлення біля 100 різних авторів за період 2001–2017 рр. для визначення підмножини авторів з подібним стилем, що і 4 еталоні роботи (колективні роботи, автори яких присутні серед досліджуваних одноосібних робіт).

Етап 6. Аналіз отриманих результатів на етапі 5. Перевірити, чи в отриманих підмножинах присутні справжні автори цих еталонних текстів. Обрати найкращий алгоритм для визначення стилю автора в україномовних науково-технічних текстах на основі технології квантитативної лінвостатистики

5 РЕЗУЛЬТАТИ

Для чистоти дослідження необхідно проаналізувати, чи впливає час публікації робіт на коефіцієнти різноманітності тексту, тобто чи не змінюються ці коефіцієнти з часом на вибірці тих самих авторів та текстів. Спочатку проаналізуємо як змінюється загальний обсяг слів в однакових за розміром уривках в діапазоні 2001–2017 рр. Як бачимо з часом ті ж самі автори частіше вживають коротші слова (рис. 3а).

З часом коефіцієнт лексичної різноманітності K_l суттєво не змінюється (рис. 3б–3г). Аналогічно з часом коефіцієнт синтаксичної складності K_s також суттєво не змінюється. А ось коефіцієнт зв'язності мовлення K_z з часом за 16 років зменшується, хоча не суттєво. На початках (2001 р.) коливається в діапазоні [0,5;1,2], а в в кінці періоду – в діапазоні [0,4; 0,9] (рис. 4).

Аналогічно порівняємо розподіли індексів винятковості та концентрації (рис. 5). Якщо розмах розподілу суттєво не змінюється в часі для I_{wt}, то для I_{kt} є фіксовані значні зміни. З часом автор цих робіт все частіше повторюють деякі терміни в своїх роботах понад 10 разів, звужуючи коло своїх досліджень. На рис. 5г поданий результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр. як мінімальне, максимальне та середнє значення за цей період (визначення коливання значень в цьому часовому проміжку). Більш суттєве коливання спостерігаємо за K_z (рис. 6).

Окремо проаналізуємо розподіл використання всіх словоформ (рис. 6г), слів по одному разу, слів понад 10 разів, вжитих в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр. (рис. 7а). На рис. 7б поданий аналіз вживання прийменників, сполучників та окремих речень в досліджуваних текстах для однакових за розміром уривках в діапазоні 2001–2017 рр., де Z – кількість прийменників, S – кількість сполучників, P – кількість окремих речень. Згідно рис. 7в з часом автори вживають коротші речення для опису предметної області, ніж на початках досліджуваного періоду. Якщо кількість прийменників зменшується, то розподіл вживання сполучників суттєво не зменшується (рис. 7г). На рис. 8а–8б поданий аналіз зміни динаміки вживання слів в досліджуваних текстах за визначений період. На рис. 8в–9г поданий результат аналізу зміни динаміки вживання прийменників, сполучників та речень в досліджуваних текстах за визначений період.

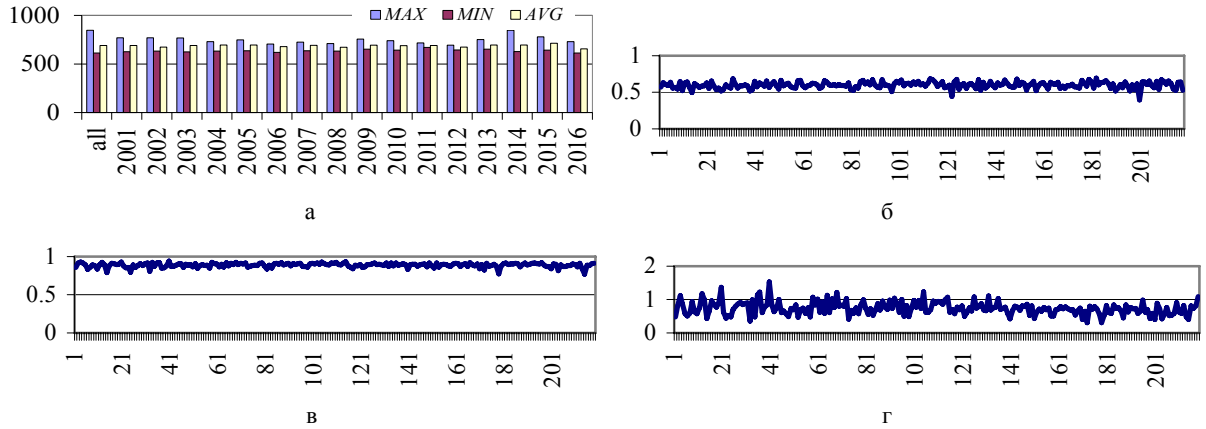

Рисунок 3 – Розподіл: а – слів та коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: б – K_l; в – K_s; г – K_z

DOI 10.15588/1607-3274-2020-1-12

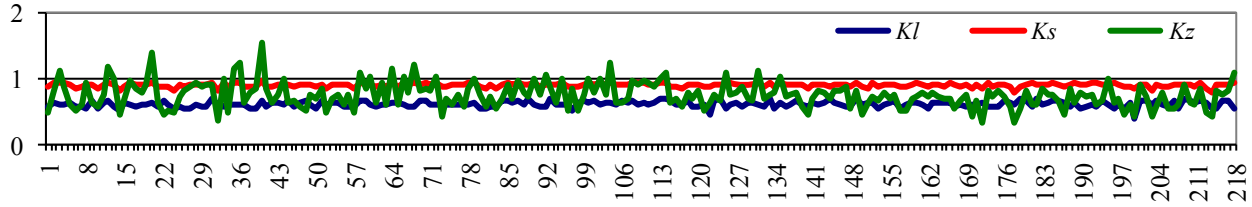

Рисунок 4 – Порівняння розподілу коефіцієнтів мовлення K_l, K_s та K_z

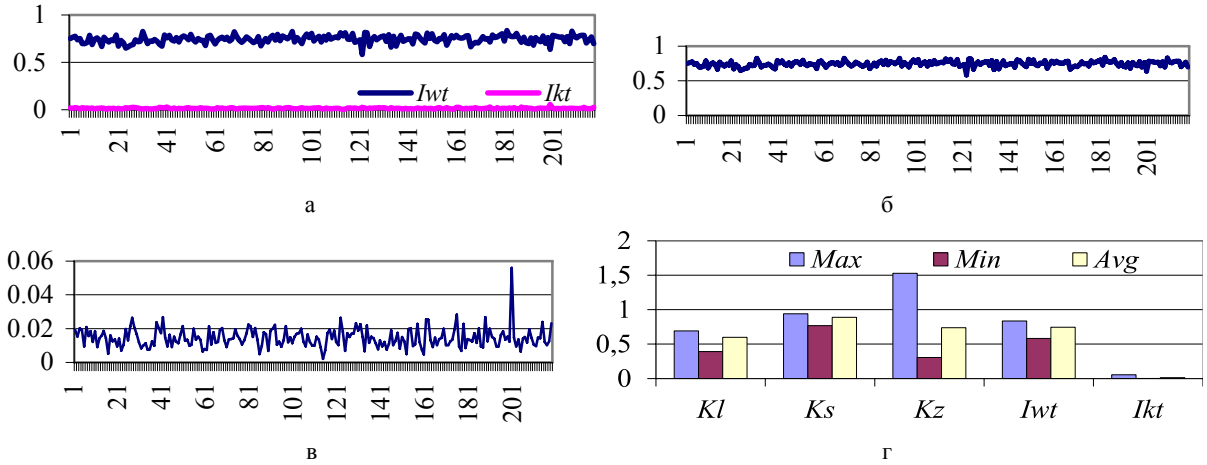

Рисунок 5 – Розподіл індексів мовлення для: а – обох індексів; б – I_{wt}; в – I_{kt}; г – мінімальне, максимальне та середнє значення для всіх коефіцієнтів

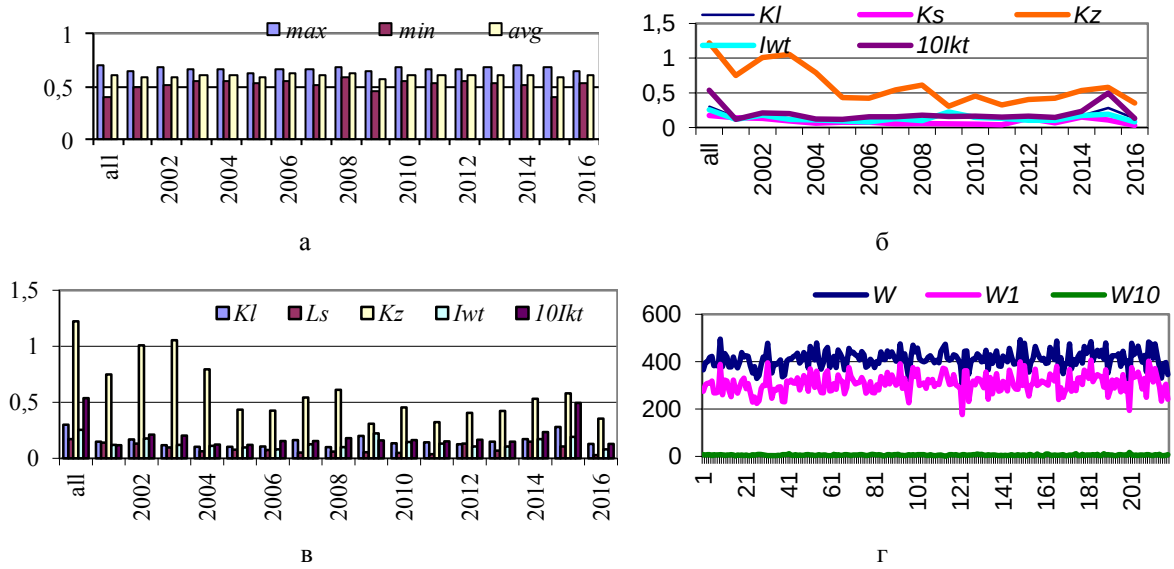

Рисунок 6 – Результат аналізу коефіцієнтів мовлення для однакових за розміром уривках в діапазоні 2001–2017 рр.: а – мінімальне, максимальне та середнє значення за цей період для K_l; б – графік динаміки зміни коефіцієнтів за визначений період; в – гістограма динаміки зміни всі коефіцієнтів за визначений період; г – вживання словоформ (всіх, по 1 разу та понад 10 разів)

DOI 10.15588/1607-3274-2020-1-12

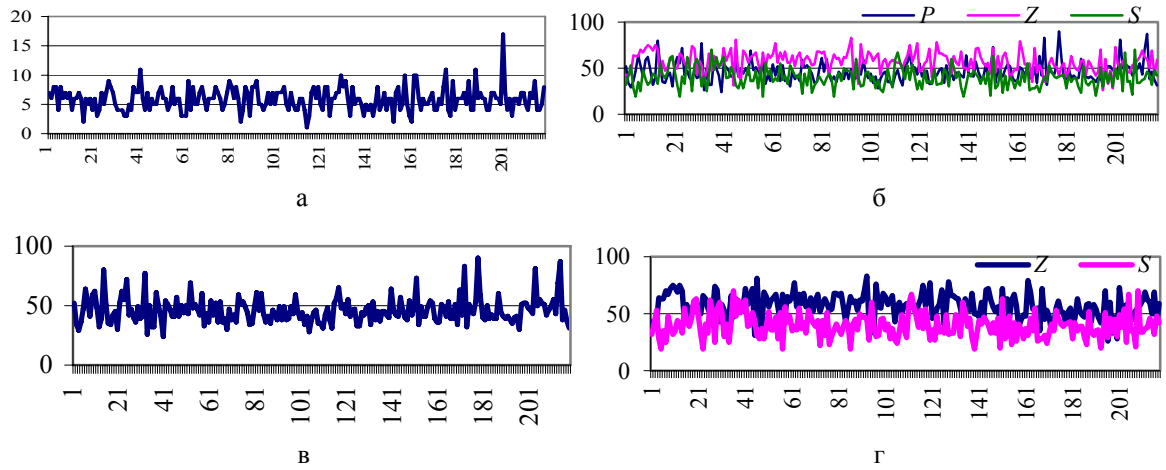

Рисунок 7 – Аналіз частоти вживання слів: *а* – понад 9 разів (W_{10}); *б* – параметрів зв'язності мовлення; *в* – речень; *г* – прийменників, та сполучників

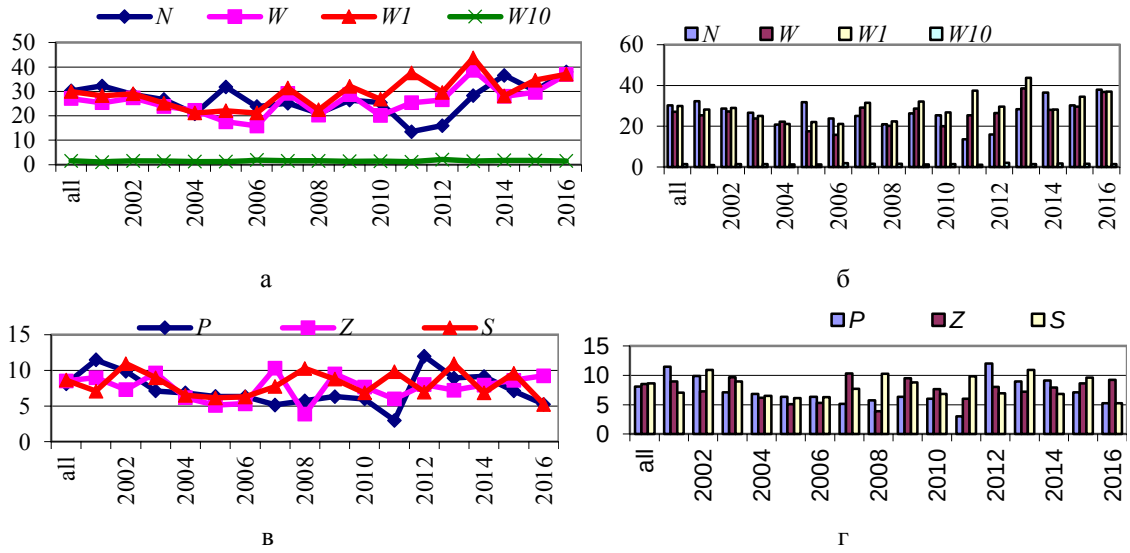

Рисунок 8 – Результат аналізу зміни динаміки вживання слів в досліджуваних текстах за визначений період: а – динаміка зміни параметрів мовлення в часі; б – розподіл значень параметрів мовлення за обумовлений період досліджень; в – динаміка зміни вживання сполучень, прийменників та речень в досліджуваних текстах; г – розподіл значень вживання сполучень, прийменників та речень за визначений період досліджень стилів автора

Довели, що існує динаміка зміни не лише коефіцієнтів мовлення авторського тексту за визначений період його творчості. Також є динаміка зміни і окремих складових, як кількість вживання словоформ на загальну кількість слів, сполучників та прийменників, речень у визначеному обсязі уривку, словоформ, які вживані лише один раз, та які вживані понад 10 разів.

6. ОБГОВОРЕННЯ

Для більш точного визначення величини приросту кожного із досліджуваного параметру необхідно провести більш суттєве дослідження на більшій вибірці як самих одноосібних творів, так збільшити діапазон дослідження творчості різних авторі на більший часовий проміжок творчості.

Далі проаналізуємо вибірку за авторським стилем та оберемо найкращий алгоритм для визначення стилю автора. На рис. 9а графік відображає визначення стилю автора по коефіцієнтах мовлення. На рис. 9б графік із накопиченням відображає зміни загальної суми за коефіцієнтами мовлення. На рис. 9в нормований графік відображає зміну вкладення кожного значення за коефіцієнтами мовлення.

Як бачимо, коефіцієнти авторського мовлення окрім *Kz* значно не змінюються в залежності від стилю конктерного автора для україномовних науково-технічних текстів. Або зміни є в млих межах, що ускладнює процес ідентифікації особливостей стилю мовлення конкретного автора в множині аналізованих авторських стилів. І чим більшою є така множина, тим складнішою буде процес ідентифікації

DOI 10.15588/1607-3274-2020-1-12

стилю конкретного автора без додаткових параметрів аналізу. Тоді проаналізуємо вибірку за авторським стилем за додатковими параметрами як загальна кількість речень в однакох за обсягом уривків, кількість слів у вибірці, частотність та пояка прийменників та сполучників. На рис. 10 графік відображає визначення стилю автора по додаткових параметрах авторського мовлення.

На рис. 10б графік із накопиченням відображає зміни загальної суми за параметрами. На рис. 10в нормований графік відображає зміну вкладення кожного значення за параметрами. Як бачимо введення додаткови параметрів зменшить множину авторів, стилі мовлення яких подібні для україномовного науково-технічного стилю публікацій. Введмо ще додаткові параметри як кількість речень, сполучників та прийменників (рис. 11) та проаналізуємо динаміку (табл. 4).

В табл. 4 наведені результати аналізу стилю 94 авторів на одноосібних працях (понад 200 одноосібних робіт) технічного спрямування за період 2001–2017 рр. Для кожного автора виведено середньоарифметичне значення кожного коефіцієнта та параметра мовлення на основі аналізу декілької його робіт за цей визначений період. Також проаналізовані стилі 4-х статтей одного авторського колективу під № 1–4 (в таблиці виділено жирним), частина авторів яких є в табл. 4 під № 69 та 93 (в таблиці виділено курсивом). Однак замала вибірка текстів для аналізу (понад 200) та кількості авторів (94) не гарантує точних результатів. Дослідження має бути продовжене на більшій кількості текстів, до яких незавжди маємо доступ. В подальшому необхідно також вдосконалити метод зарахунок аналізу текстів методами стилем атрії та глотохронології.

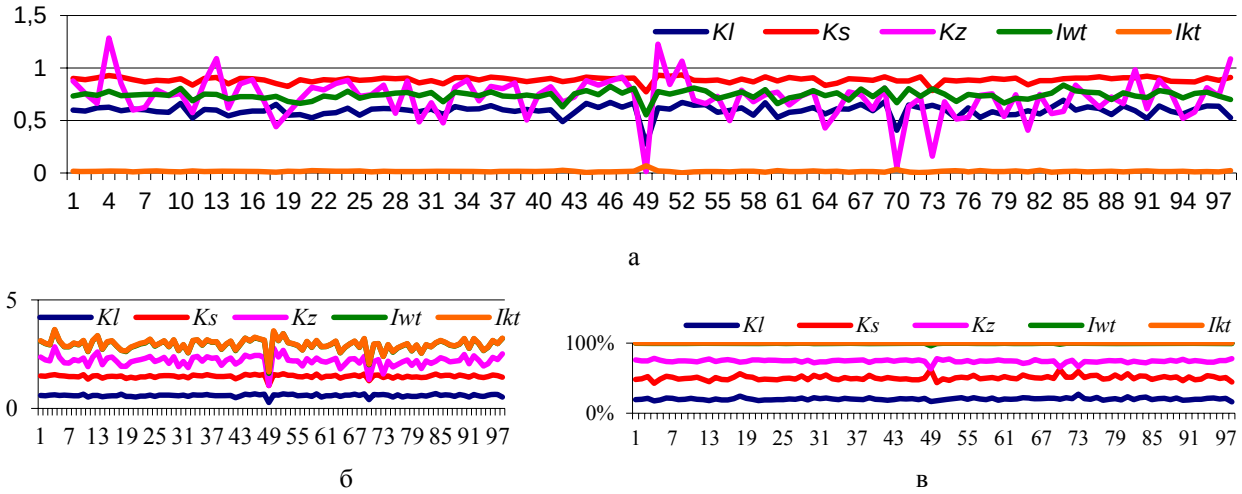

Рисунок 9 – Детальний аналіз: а – процесу у часі визначення стилю автора по коефіцієнтах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення

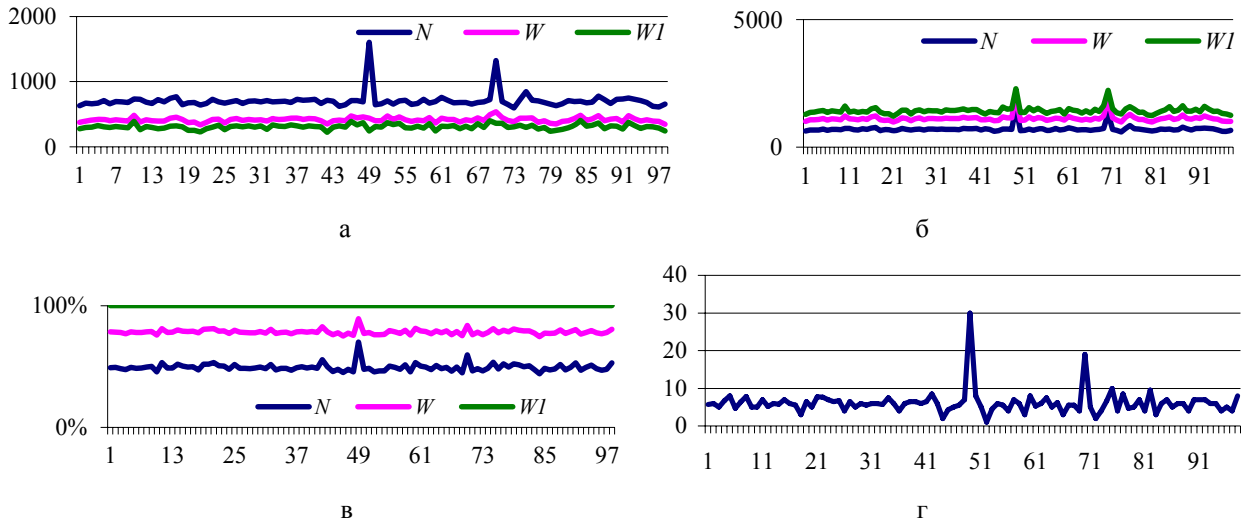

Рисунок 10 – Детальний аналіз: а – процесу визначення стилю автора по параметрах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення; г – зміни параметру як частота появи слова понад 10 разів (W_{10})

DOI 10.15588/1607-3274-2020-1-12

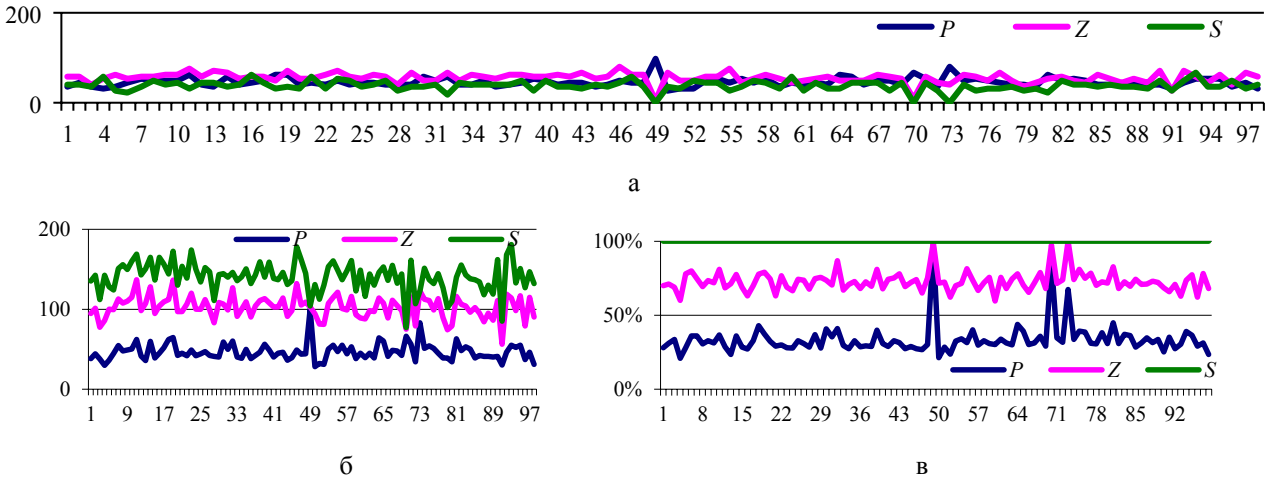

Рисунок 11 – Детальний аналіз: а – процесу визначення стилю автора по параметрах мовлення; б – зміни загальної суми за коефіцієнтами мовлення; в – зміни вкладення кожного значення за коефіцієнтами мовлення

Таблиця 4 – Результат роботи алгоритму аналізу стилю автора публікації на інформаційному ресурсі Victana

№	N	S	P	Z	W	W_1	W_{10}	I_{kt}	I_{wt}	K_z	K_s	K_l
1.	**631,3**	**40,7**	**38**	**56,7**	**377,7**	**277,7**	**5,7**	**0,015**	**0,73**	**0,88**	**0,9**	**0,6**
2.	**658**	**42**	**31**	**59**	**345**	**241**	**8**	**0,023**	**0,7**	**1,07**	**0,91**	**0,52**
3.	**614**	**32**	**46**	**69**	**391**	**287**	**4**	**0,01**	**0,73**	**0,73**	**0,88**	**0,64**
4.	**622**	**48**	**37**	**42**	**397**	**305**	**5**	**0,013**	**0,77**	**0,81**	**0,91**	**0,64**
5.	680	34	55	62	414	314	4	0,01	0,76	0,58	0,87	0,6
6.	709	35	52	46	398	285	6	0,015	0,72	0,52	0,87	0,56
7.	732	67	55	59	429	329	6	0,014	0,77	0,76	0,87	0,59
8.	749	49	46	73	478	375	7	0,015	0,78	0,88	0,9	0,64
9.	734	29	30	26	381	273	7	0,018	0,72	0,61	0,92	0,52
10.	730	51	41	70	433	317	7	0,016	0,73	0,98	0,91	0,59
11.	665	33	40	46	425	324	4	0,009	0,76	0,66	0,91	0,64
12.	723	35	41	54	401	280	6	0,015	0,7	0,72	0,9	0,55
13.	780	34	41	43	479	366	6	0,013	0,76	0,63	0,91	0,61
14.	685	39	42	53	432	333	5	0,012	0,77	0,73	0,9	0,63
15.	674	35	39	63	404	316	7	0,017	0,78	0,84	0,9	0,9
16.	700	42	50	46	485	406	6	0,012	0,84	0,59	0,9	0,69
17.	695	39	53	51	436	332	3	0,007	0,76	0,57	0,88	0,63
18.	709,5	49,5	48	58	399	292,5	9,5	0,024	0,73	0,75	0,88	0,56
19.	661	24	63	53	391	275	4	0,01	0,7	0,41	0,84	0,59
20.	631	31	34	45	350	249	7	0,02	0,71	0,75	0,9	0,55
21.	654	28	39	35	361	240	5	0,014	0,66	0,54	0,89	0,55
22.	682,3	37,7	39	50,3	398,7	296,3	4,7	0,012	0,74	0,75	0,9	0,58
23.	706	31	45	68,5	374	275	8,5	0,023	0,73	0,74	0,88	0,53
24.	712,5	33	51	48	442,5	331,5	4	0,009	0,75	0,53	0,88	0,62
25.	846	26	54	57	440	299	10	0,023	0,68	0,51	0,88	0,52
26.	726,3	39	51	61,3	441,3	332,3	6,7	0,015	0,75	0,68	0,88	0,6
27.	598	0	83	40	386	309	4	0,01	0,8	0,16	0,78	0,65
28.	652	28	34	45	405	296	2	0,005	0,73	0,72	0,92	0,62
29.	697	46	56	59,5	450	361,5	5	0,011	0,8	0,63	0,88	0,65
30.	1325	2	66	9	538	360	19	0,035	0,67	0,06	0,88	0,4
31.	726	46	42	56	493	399	4	0,008	0,81	0,81	0,91	0,68
32.	689,5	28	47,5	57	407,5	296	5,5	0,014	0,73	0,61	0,88	0,59
33.	683	43,5	48,5	63	446	357	5,5	0,012	0,8	0,74	0,89	0,65
34.	658	47	41	48	399	277	3	0,008	0,69	0,78	0,9	0,6
35.	682,6	45	60	47,8	416,2	318	6,2	0,015	0,76	0,59	0,86	0,6
36.	679	32	64	50	381	280	5	0,013	0,73	0,43	0,83	0,56
37.	673,5	33	39	58	419	329	7,5	0,018	0,79	0,78	0,91	0,62
38.	717	46	45	53	422	310	6	0,014	0,73	0,73	0,89	0,59
39.	761	28,3	39,3	48,5	440	315,8	5,3	0,012	0,71	0,65	0,91	0,58
40.	693	60	45	44	366	242	8	0,022	0,66	0,77	0,88	0,53
41.	670	30	38	55	449	356	3	0,007	0,79	0,75	0,92	0,67
42.	732	45	53	63	402	290	6	0,015	0,72	0,68	0,87	0,55
43.	666	49	44	55	412	318	7	0,017	0,77	0,79	0,89	0,62
44.	652	36	55	46	389	287	4	0,01	0,74	0,5	0,86	0,6
45.	716	27,5	47	74,5	413,5	293	5,5	0,013	0,71	0,73	0,89	0,58
46.	704,8	45,8	54,8	60	458,8	360	6	0,013	0,78	0,66	0,88	0,65

DOI 10.15588/1607-3274-2020-1-12

№	*N*	*S*	*P*	*Z*	*W*	W_1	W_{10}	I_{kt}	I_{wt}	K_z	K_s	K_l
47.	656	46	50	57,5	422,5	341,5	4,5	0,011	0,81	0,69	0,88	0,64
48.	705	49	31	50	474	369	1	0,002	0,78	1,06	0,93	0,67
49.	661,5	31	32	49,5	402,5	302	5	0,012	0,75	0,84	0,92	0,6
50.	644	37	28	66	400	310	8	0,02	0,78	1,23	0,93	0,62
51.	1602	1	100	3	442	245	30	0,068	0,55	0,01	0,77	0,28
52.	689	36	44	65	458	369	7	0,015	0,81	0,77	0,9	0,66
53.	708	56,5	43,5	62	442,5	336,5	5,5	0,012	0,76	0,91	0,9	0,63
54.	708	46	49	83	475	392	5	0,011	0,83	0,88	0,9	0,67
55.	645	37,7	39,3	58,7	403	302,3	4,3	0,011	0,74	0,84	0,9	0,62
56.	620	40	36	55	411	323	2	0,005	0,79	0,88	0,91	0,66
57.	699	32	46	68	401	302	6	0,015	0,75	0,72	0,89	0,57
58.	715,5	34	45	58	352	223,5	8,5	0,024	0,63	0,68	0,87	0,49
59.	666	35,5	40	63	401,5	305	6,5	0,016	0,76	0,82	0,9	0,6
60.	728	51	49	59	430	313	6	0,014	0,73	0,75	0,89	0,59
61.	717,5	26,5	56	57,5	433,5	321,5	6,5	0,015	0,74	0,5	0,87	0,6
62.	714,5	48,5	46	65	418,5	304,5	6,5	0,016	0,73	0,86	0,89	0,59
63.	730	39	42	62	440	323	6	0,014	0,73	0,8	0,9	0,6
64.	683	42	38	52	438	339	4	0,009	0,77	0,82	0,91	0,64
65.	699	41	49,5	60	427	314	6	0,014	0,74	0,69	0,88	0,61
66.	695	41	38,5	61,3	422,5	318,3	7,5	0,018	0,75	0,89	0,91	0,6
67.	691	44,7	40	51	436,7	336,7	5,7	0,013	0,77	0,82	0,91	0,63
68.	711	19	60	67	396	268	6	0,015	0,68	0,48	0,85	0,56
69.	*688,8*	*41,3*	*49,7*	*49,3*	*416,8*	*321,9*	*6*	*0,016*	*0,77*	*0,67*	*0,88*	*0,6*
70.	704,5	38	59	47,5	412	303,5	5,5	0,013	0,74	0,49	0,86	0,58
71.	700	35	40	68,5	418,5	320,5	6	0,014	0,77	0,88	0,9	0,6
72.	665	28	41	42	406	309	5	0,012	0,76	0,57	0,9	0,61
73.	708,5	47,5	42	57,5	434	323,5	6,5	0,015	0,75	0,84	0,9	0,61
74.	691	40	47	65	421	311	4	0,01	0,74	0,74	0,89	0,6
75.	668,8	34,5	44	55,8	368,3	262,5	6,8	0,018	0,71	0,73	0,88	0,55
76.	691,7	50	41,8	58,2	425,7	331,3	6,5	0,015	0,78	0,88	0,9	0,62
77.	731	54	49	71	420	301	7	0,017	0,72	0,85	0,88	0,57
78.	665	32,3	41,7	65	376	275,7	7,7	0,02	0,73	0,79	0,89	0,57
79.	642	56,8	44,8	52,3	337,5	230,3	7,8	0,023	0,68	0,81	0,87	0,52
80.	680	33	42	55	379	251	5	0,013	0,66	0,7	0,89	0,56
81.	677,5	36	64,5	72	373,5	255	6,5	0,018	0,68	0,57	0,86	0,55
82.	647	32	62	50	422	308	3	0,007	0,73	0,44	0,85	0,65
83.	768	47	51,5	58	452,5	323	5,5	0,012	0,71	0,68	0,89	0,59
84.	745	61	45	59	439	319	6	0,014	0,73	0,89	0,9	0,59
85.	691	42,3	39	55,3	396,7	289	7	0,018	0,73	0,85	0,9	0,57
86.	724,2	36,8	59,6	68,4	394,2	278,8	5,8	0,015	0,71	0,61	0,85	0,55
87.	665,5	43	35,5	72	399	299	6	0,015	0,75	1,09	0,91	0,6
88.	686,5	45	41,1	56,9	414,5	312,6	5,9	0,012	0,75	0,86	0,9	0,6
89.	729	32	62	75	380	261	7	0,018	0,69	0,58	0,84	0,52
90.	733,5	45	50	65	486,5	392	5	0,01	0,8	0,76	0,9	0,66
91.	682,5	39,7	49	61	394,2	291	5	0,013	0,74	0,74	0,88	0,58
92.	691,8	47,8	47,8	60	403,4	301,6	7,8	0,019	0,75	0,79	0,88	0,58
93.	*694,5*	*38,1*	*54,3*	*58,5*	*417,4*	*313,1*	*6,4*	*0,015*	*0,75*	*0,62*	*0,87*	*0,6*
94.	661,1	24,8	44,7	54,7	402,7	299,7	4,7	0,012	0,74	0,6	0,89	0,61
95.	708	28	36	64	419	309	8	0,019	0,74	0,85	0,91	0,59
96.	668,8	57	29,8	56	418,3	325,8	6,8	0,016	0,78	1,28	0,93	0,63
97.	662,5	34,8	37,8	39,8	410,3	303	5	0,012	0,74	0,67	0,9	0,61
98.	671,3	41,1	44,2	57,1	395,6	299	6	0,015	0,76	0,76	0,89	0,59

ВИСНОВКИ

Розроблено метод визначення автора тексту на основі аналізу коефіцієнтів лексичного авторського мовлення в еталонному уривку авторського тексту. Розроблено алгоритм лексичного аналізу україномовних текстів та алгоритм синтаксичного аналізатора текстового контенту на основі аналізу кожного слова з врахуванням його частини мови та відмінювання. Тобто при аналізі лінгвістичних одиниць типу слів, враховувалась належність до частини мови та відмінювання в межах цієї частни мови. Для цього прводився аналіз флекцій цих слів для класифікації, виділення основи для формування відповідних алфавітно-частотних словників. Наповнення цих словників в подальшому враховувалися на наступних кроках визначення авторства тексту як розрахунок параметрів та коефіцієнтів авторського мовлення. Для індивідуального стилю письменника показовими є саме службові (стопові або опорні) слова, оскільки вони ніяк не пов'язані з темою і змістом публікації. Розроблено алгоритм визначення стопових слів текстового контенту на основі лінгвістичного аналізу текстового контенту. Його особливостями є адаптація морфологічного та синтаксичного аналізу лексичних одиниць до особливостей конструкцій україномовних слів/текс-

DOI 10.15588/1607-3274-2020-1-12

тів. Наведено теоретичне та експериментальне обгрунтування методу контент-моніторінгу та визначення стопових слів україномовного тексту. Метод спрямовано на автоматичне виявлення значущих стопових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації парсингу текстового контенту науково-технічного спрямування. Запропоновано підхід до розроблення програмного забезпечення контент-моніторінгу для визначення автора в україномовних науко-технічних текстах на основі NLP, стилементії та Web Mining. Проаналізовано розробленою системою понад 200 одноосібних наукових публікацій зі всіх номерів Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (Україна) за період 2001–2017 рр. Досліджено внутрішню «динаміку» цих текстів довільно обраних авторів через аналіз коефіцієнтів зв'язності мовлення, лексичної різноманітності та синтаксичної складності, а також індексів концентрації та винятковості для перших *k*, *n* та *m* (без заголовка) слів авторського уривку та аналізованого.

Досліджено результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення автора в україномовних наукових текстах технічного профілю. Проведено порівняння результатів на множині 200 одноосібних робіт технічного спрямування біля 100 різних авторів за період 2001–2017 рр. для визначення чи змінюються і як коефіцієнти різноманітності тексту цих авторів в різні проміжки часу. На основі розробленого програмного забезпечення отримано результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення та аналізу стопових слів в україномовних наукових текстах технічного профілю на основі технології Web Mining. Виявлено, що для обраної експериментальної бази з понад 200 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації як анотації та ключові слова різними мовами, а також списку літератури. Подальшого експериментального дослідження потребує апробація запропонованого методу для визначення стилю автора з інших категорій текстів – наукових гуманітарного профілю, художніх, публіцистичних тощо.

ПОДЯКИ

Роботу виконано в рамках держбюджетної теми «Методи та засоби функціонування систем підтримки прийняття рішень на основі онтологій» (ID:839 2017-05-15 09:20:01 (2459-315)). Дослідження провадилось в межах спільних наукових досліджень кафедри інформаційних систем та мереж НУ «Львівська політехніка» на тему «Дослідження, розроблення і впровадження інтелектуальних розподілених інформаційних технологій та систем на основі ресурсів баз даних, сховищ даних, просторів даних та знань з метою прискорення процесів формування сучасного інформаційного суспільства». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ НУ «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

ЛІТЕРАТУРА / ЛИТЕРАТУРА

1. Mobasher B. Data mining for web personalization / B. Mobasher // The adaptive web. – 2007. – Vol. 4321. – P. 90–135.
2. Dinucă C. Web Content Mining. In: University of Petroşani / C. Dinucă, D. Ciobanu // Economics. – 2012. – Vol. 12. – P. 85–92.
3. Xu G. Web content mining / G. Xu, Y. Zhang, L. Li // Web Mining and Social Networking. – 2011. – Vol. 6. – P. 71–87.
4. Khribi M. K. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval / M. K. Khribi, M. Jemni, O. Nasraoui // Advanced Learning Technologies : International Conference, 1–5 July 2008 : proceedings. – Santander, Cantabria, Spain : IEEE, 2008. – P. 241–245.
5. Automatic web content personalization through reinforcement learning / [S. Ferretti, S. Mirri, C. Prandi, P. Salomoni] // Journal of Systems and Software. – 2016. – Vol. 121. – P. 157–169.
6. User attitudes towards news content personalization / [T. Lavie, M. Sela, I. Oppenheim et al] // International journal of human-computer studies. – 2010. – Vol. 68(8). – P. 483–495.
7. Fredrikson M. Repriv: Re-imagining content personalization and in-browser privacy / M. Fredrikson, B. Livshits // Symposium on Security and Privacy: Conference, 22–25 May 2011 : proceedings. – Berkeley, CA, USA : IEEE, 20011. – P. 131–146.
8. Application of neural networks and Kano's method to content recommendation in web personalization / [C. C Chang, P. L. Chen, F. R. Chiu, Y. K. Chen] // Expert Systems with Applications. – 2009. – Vol. 36(3). – P. 5310–5316.
9. Pat. US7,571,226B1 US Content personalization over an interface with adaptive voice character / [H. Partovi, R. Brathwaite, A. Davis et al.] (US) ; TellMe Networks, Inc., Mountain View, CA (US). – No.: 09/523,853 ; Marz 14, 2009; August 4, 2009, Patent and Trademark Office. – 20 p.
10. Pat. US2009/0171968A1 US Widget-assisted content personalization based on user behaviors tracked across multiple web sites / F. J. Kane, C. Hicks (US) ; Amazon Technologies Inc (US). – No.: 11/966,817 ; December 28, 2007; Jule 2, 2009, Google Patents. – 24 p.
11. Mirri S. Experiential adaptation to provide user-centered web content personalization / S. Mirri, C. Prandi, P. Salomoni // Advances in Human oriented and Personalized Mechanisms, Technologies, and Services : The Sixth International Conference, October 27 – November 1, 2013: proceedings. – Venice, Italy : IARIA, 2003. – P. 31–36.
12. Fernandez-Luque L. Review of extracting information from the Social Web for health personalization / L. Fernandez-Luque, R. Karlsen, J. Bonander // Journal of medical Internet research. – 2011. – Vol. 13(1). – P. 15.

DOI 10.15588/1607-3274-2020-1-12

13. Pat. US8,019,777B2 US Digital content personalization method and system / E. Hauser (US) ; CRICKET MEDIA Inc (US). –No.: 12/795,419 ; June 7, 2010; September 13, 2011, Patent and Trademark Office. – 15 p.
14. Ho S. Y. Timing of adaptive web personalization and its effects on online consumer behavior / S. Y. Ho, D. Bodoff, K. Y. Tam // Information Systems Research. – 2011. – Vol. 22(3). – P. 660–679.
15. Uchyigit G. Personalization techniques and recommender systems / G. Uchyigit, M. Y. Ma. – Singapore : World Scientific, 2008. – 322 p.
16. Pat. US2006/0020883A1 Web page personalization / [N. Kothari, M. Harder, R. Howard et al.] (US) ; Microsoft Technology Licensing LLC (US). – No.: 10/857,724 ; May 28, 2004; Januar 26, 2006, Patent and Trademark Office. – 18 p.
17. Zhang H. Construction of ontology-based user model for web personalization / H. Zhang, Y. Song., H. T. Song // Lecture Notes in Computer Science. – 2007. – Vol. 4511. – P. 67–76.
18. Pat. US 8,254,892 B2 US Methods and apparatus for anonymous user identification and content personalization in wireless communication / H. Chien (US) ; AT&T Mobility II LLC (US). – No.: 12/468,708 ; September 10, 2009; August 28, 2012, Patent and Trademark Office. – 9 p.
19. Pat. US7,970,664B2 US Content personalization based on actions performed during browsing sessions / G. D. Linden, B. R. Smith, N. K. Zada (US) ; Amazon Technologies Inc (US). – No.: 11/009,732 ; December 10, 2004; June 28, 2011, Patent and Trademark Office. –36 p.
20. Web personalization using web mining: concept and research issue / [P. Mehtaa, B. Parekh, K. Modi, P. Solanki] // International Journal of Information and Education Technology. – 2012. – Vol. 2(5). – P. 510–512.
21. Zhezhnych P. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism Documentation Objects / P. Zhezhnych, O. Markiv // Advances in Intelligent Systems and Computing. – 2018. – Vol. 689. – P. 656–667.
22. Basyuk T. The main reasons of attendance falling of internet resource / T. Basyuk // Computer Sciences and Information Technologies : Xth International Scientific and Technical Conference, 14–17 September 2015 : proceedings. – Lviv : IEEE, 2015. – P. 91–93.
23. Uniform Method of Operative Content Management in Web Systems / [A. Gozhyj, L. Chyrun, A. Kowalska-Styczen, O. Lozynska] // CEUR Workshop Proceedings. – 2018. – Vol. 2136. – P. 62–77.
24. Kravets P. The control agent with fuzzy logic / P. Kravets // Perspective Technologies and Methods in MEMS Design : VIth International Conference, 20–23 April 2010 2015 : proceedings. – Lviv : IEEE, 2015. – P. 40–41.
25. Davydov M. Linguistic Models of Assistive Computer Technologies for Cognition and Communication / M. Davydov, O Lozynska // Computer Science and Information Technologies : XIth International Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 171–175.
26. Mykich K. Algebraic model for knowledge representation in situational awareness systems / K. Mykich, Y. Burov // Computer Sciences and Information Technologies : nternational Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 165–167.
27. Mykich K. Uncertainty in situational awareness systems / K. Mykich, Y. Burov // Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Conference, 623–26 Februar 2016 : proceedings. – Lviv : IEEE, 2016. – P. 729–732.
28. Mykich K. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness / K. Mykich, Y. Burov // Advances in Intelligent Systems and Computing. – 2017. – Vol. 512. – P. 217–227.
29. Mykich K. Research of uncertainties in situational awareness systems and methods of their processing / K. Mykich, Y. Burov // EasternEuropean Journal of Enterprise Technologies. – 2016. – Vol. 1(79). – P. 19–26.
30. Vysotska V. Linguistic Analysis of Textual Commercial Content for Information Resources Processing / V. Vysotska // Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Scientific and Technical Conference, 23–26 February 2016 : proceedings. – Lviv : IEEE, 2016. – P. 709–713.
31. Information resources processing using linguistic analysis of textual content / [J. Su, V. Vysotska, A. Sachenko, V. Lytvyn, Y. Burov] // Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications : 9th International Conference, 21–23 September 2017 : proceedings. – Bucharest, Romania: IEEE, 2017. – P. 573–578.
32. Content Linguistic Analysis Methods for Textual Documents Classification / [V. Lytvyn, V. Vysotska, O. Veres, I. Rishnyak, H Rishnyak] // Computer Science and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings. – Lviv : IEEE, 2016. – P. 190–192.
33. Bisikalo O. V. Identifying keywords on the basis of content monitoring method in ukrainian texts / O. V. Bisikalo, V. A. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 1(36). – P. 74–83.
34. Bisikalo O.V. Sentence syntactic analysis application to keywords identification Ukrainian texts / O. V. Bisikalo, V. A. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 3(38). – P. 54–65.
35. Lytvyn V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic / V. Lytvyn, I. Bobyk, V. Vysotska // Radio Electronics, Computer Science, Control. – 2016. – Vol. 4(39).– P. 54–67.
36. Alieksieieva K. Technology of commercial web-resource management based on fuzzy logic / K. Alieksieieva, A. Berko, V. Vysotska // Radio Electronics, Computer Science, Control. – 2015. – Vol. 3(34). – P. 71–79.
37. Application of Sentence Parsing for Determining Keywords In Ukrainian Texts / [Vasyl Lytvyn, Victoria Vysotska, Dmytro Dosyn, Roman Holoschuk, Zoriana Rybchak] // Computer Science and Information Technologies : 12th International Scientific and Technical Conference, 5–8 September 2017 : proceedings. – Lviv : IEEE, 2017. – P. 326–331.

Received 25.10.2019.
Accepted 09.02.2020.

DOI 10.15588/1607-3274-2020-1-12

УДК 004.9

МЕТОД АВТОРИФИКАЦИИ ТЕКСТА НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ НА ОСНОВЕ ЛИНГВИСТИЧЕСКОГО АНАЛИЗА КОЭФФИЦИЕНТОВ ЯЗЫКОВОГО РАЗНООБРАЗИЯ

Высоцкая В. А. – канд. техн. наук, доцент, доцент кафедры «Информационные системы и сети», Национальный университет «Львовская политехника», Украина.

АННОТАЦИЯ

Актуальность. Авторификация авторства текста является техникой определения автора текста, когда неоднозначно, кто ее написал. Это полезно, когда несколько человек претендуют на авторство одной публикации или в случаях, когда никто не претендует на авторство текстового контента, например, так называемые тролли в социальных сетях во время информационной войны. Сложность проблемы авторского текста, очевидно, экспоненциально выше, большее количество возможных авторов. Наличие авторских текстовых образцов также является существенным при продвижении этой проблемы. Атрибуция авторского текста включает следующие три проблемы:

– выявление автора текстового автора из группы возможных или ожидаемых авторов, где автор всегда находится в группе подозреваемых;

– не идентификация автора текстового автора из группы возможных или ожидаемых авторов, где автор может не быть в группе подозреваемых;

– оценка возможности данного текста, написанного данным автором или нет.

Поэтому задача автоматического определения автора текстового контента научно-технического направления актуальна и требует новых (более совершенных) подходов к ее решению.

Целью исследования является разработка метода определения автора в украиноязычных текстах на основе технологии лингвометрии.

Метод. Разработано лингвометричний метод алгоритмического обеспечения процессов контент-мониторинга для решения задачи автоматического определения автора русскоязычного текстового контента на основе технологии статистического анализа коэффициентов языкового разнообразия. Проведения декомпозиции метода определения автора на основе анализа таких коэффициентов речи как лексическая разнообразие, степень (мера) синтаксической сложности, связность речи, индексы исключительности и концентрации текста. Проанализированы также параметры авторского стиля как количество слов в определенном тексте, общее количество слов этого текста, количество предложений, количество предлогов, количество союзов, количество слов с частотой 1, количество слов с частотой 10 и больше. Особенностями разработанного является адаптация морфологического и синтаксического анализа лексических единиц к особенностям конструкций украиноязычных слов / текстов. То есть при анализе лингвистических единиц типа слов, учитывалась принадлежность к части речи и склонение в пределах этой части речи. Для этого проводился анализ флексий этих слов для классификации, выделение основы для формирования соответствующих алфавитно-частотных словарей. Наполнение этих словарей в дальнейшем учитывались на следующих шагах определения авторства текста как расчет параметров и коэффициентов авторской речи. Для индивидуального стиля писателя показательны именно служебные (стоп или опорные) слова, поскольку они никак не связаны с темой и содержанием публикации.

Результаты. Проведено сравнение результатов на множестве 200 самостоятельных работ технического направления около 100 различных авторов период 2001–2017 гг. Для определения меняются и как коэффициенты разнообразия текста этих авторов в разные промежутки времени.

Выводы. Выявлено, что для выбранной экспериментальной базы из более 200 работ лучших результатов по критерию плотности достигает метод анализа статьи без начальной обязательной информации как аннотации и ключевые слова на разных языках, а также список литературы.

КЛЮЧЕВЫЕ СЛОВА: текстовый контент, NLP, контент-мониторинг, стоп-слова, контент-анализ, статистический лингвистический анализ, квантитативных лингвистика.

УДК 004.9

THE SCIENTIFIC AND TECHNICAL PUBLICATIONS TEXT AUTHORITICATION METHOD BASED ON LINGUSTICAL ANALYSIS OF LANGUAGE DIVERSITY COEFFICIENTS

Vysotska V. – PhD, Associate Professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine.

ABSTRACT

Context. Authorization of the authorship of the text is a technique for determining the author of the text, when it is ambiguous who wrote it. It is useful when several people claim to be the authors of one publication or in cases where nobody claims to authorship of text content, for example, so-called trolls in social networks during an information warfare. The complexity of the problem of the author's text, obviously, is exponentially higher, more likely authors. The presence of author's text samples is also significant in advancing this problem. The attribution of the author's text includes the following three problems:

– author discovery of text from probable or expected authors group, where the author is always in a suspects group;

– not identification of the author of a text author from a group of probable or expected authors, where the author may not be in a group of suspects;

– assessment of the possibility of this text, written by the author or not.

Therefore, the task of automatically determining the author of text content of scientific and technical direction is relevant and requires new (more perfect) approaches to its solution.

Objective of the study is to develop a method for determining the author in Ukrainian texts based on the technology of lingometry.

DOI 10.15588/1607-3274-2020-1-12

Method. Lingvometric method of algorithmic provision of content monitoring processes for solving the problem of automatic determination of the author of Ukrainian-language text content on the basis of technology of statistical analysis of linguistic diversity coefficients is developed. A decomposition of the method of determination of the author on the basis of analysis of such broadcasting factors as lexical diversity, degree (degree) of syntactic complexity, speech connectivity, singularity indexes and text concentrations is made. Also, author's style parameters are analyzed as the number of words in a particular text, the total number of words in this text, the number of sentences, the number of prepositions, the number of conjunctions, the number of words with the frequency of 1, and the number of words with a frequency of 10 or more. The features of the developed is the adaptation of the morphological and syntactic analysis of lexical units to the features of the designs of Ukrainian-language words / texts. That is, in the analysis of linguistic units of the type of words, the affiliation with the part of speech and declarations within this part of the language was taken into account. To do this, an analysis of the flexion of these words was carried out for classification, the allocation of the basis for the formation of the corresponding alphabet-frequency dictionaries. The filling of these dictionaries was further taken into account in the subsequent steps of determining the authorship of the text as the calculation of parameters and coefficients of copyright broadcasting. For the individual style of a writer, it is precisely service (stop or reference) words that are indicative because they are not related to the topic and content of the publication.

Results. A comparison of results on a plurality of 200 individual technical works of about 100 different authors over the period 2001–2017 has been made to determine whether the coefficients of the diversity of the text of these authors are different at different intervals.

Conclusions. It has been found that for the chosen experimental base with over 200 works of the best results, the method of analysis of the article without initial obligatory information as annotations and keywords in various languages and the list of literature achieves the density criterion.

KEYWORDS: text content, NLP, content monitoring, stop words, content analysis, statistical linguistic analysis, quantitative linguistics.

REFERENCES

1. Mobasher B. Data mining for web personalization, *The adaptive web*, 2007, Vol. 4321, pp. 90–135.
2. Dinucă C., Ciobanu D. Web Content Mining. In: University of Petroşani, *Economics,* 2012, Vol. 12, pp. 85–92.
3. Xu G. Zhang Y., Li L. Web content mining, *Web Mining and Social Networking*, 2011, Vol. 6, pp. 71–87.
4. Khribi M. K., Jemni M., Nasraoui O. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval, *Advanced Learning Technologies : International Conference, 1–5 July 2008 : proceedings.* Santander, Cantabria, Spain, IEEE, 2008, pp. 241–245.
5. Ferretti S., Mirri S., Prandi C., Salomoni P. Automatic web content personalization through reinforcement learning, *Journal of Systems and Software*, 2016, Vol. 121, pp. 157–169.
6. Lavie T., Sela M., Oppenheim I., Inbar O., Meyer J. User attitudes towards news content personalization, *International journal of human-computer studies,* 2010, Vol. 68(8), pp. 483–495.
7. Fredrikson M., Livshits B. Repriv: Re-imagining content personalization and in-browser privacy, *Symposium on Security and Privacy: Conference, 22–25 May 2011 : proceedings*. Berkeley, CA, USA, IEEE, 2011. pp. 131–146.
8. Chang. C., Chen P., Chiu F., Chen Y. Application of neural networks and Kano's method to content recommendation in web personalization, *Expert Systems with Applications,* 2009, Vol. 36(3), pp. 5310–5316.
9. Partovi H., Brathwaite R., Davis A., McCue M., Porter B., Giannandrea J., Li Z. (US) Pat. US7,571,226B1 US Content personalization over an interface with adaptive voice character, U.S. ; TellMe Networks, Inc., Mountain View, CA (US). No.: 09/523,853 ; Marz 14, 2009; August 4, 2009, Patent and Trademark Office, 20 p.
10. Kane F. J., Hicks C. (US) Pat. US2009/0171968A1 US Widget-assisted content personalization based on user behaviors tracked across multiple web sites; Amazon Technologies Inc (US). No.: 11/966,817; December 28, 2007; Jule 2, 2009, Google Patents, 24 p.
11. Mirri S., Prandi C., Salomoni P. Experiential adaptation to provide user-centered web content personalization, *Advances in Human oriented and Personalized Mechanisms, Technologies, and Services : The Sixth International Conference, October 27 – November 1, 2013: proceedings*. Venice, Italy, IARIA, 2003, pp. 31–36.
12. Fernandez-Luque L., Karlsen R., Bonander J. Review of extracting information from the Social Web for health personalization, *Journal of medical Internet research*, 2011, Vol. 13(1), P. 15.
13. Hauser E. (US) Pat. US8,019,777B2 US Digital content personalization method and system; CRICKET MEDIA Inc (US). No.: 12/795,419 ; June 7, 2010; September 13, 2011, Patent and Trademark Office, 15 p.
14. Ho S. Y., Bodoff D., Tam K. Y. Timing of adaptive web personalization and its effects on online consumer behavior, Information Systems Research, 2011, Vol. 22(3), pp. 660–679.
15. Uchyigit G., Ma M. Y.. Personalization techniques and recommender systems. Singapore, World Scientific, 2008, 322 p.
16. Kothari N., Harder M., Howard R., Sanabria A., Schackow S. (US) Pat. US2006/0020883A1 Web page personalization; Microsoft Technology Licensing LLC (US). No.: 10/857,724 ; May 28, 2004; Januar 26, 2006, Patent and Trademark Office. – 18 p.
17. Zhang H., Song Y., Song H. T. Construction of ontology-based user model for web personalization, *Lecture Notes in Computer Science,* 2007, Vol. 4511, pp. 67–76.
18. Chien H. (US) Pat. US 8,254,892 B2 US Methods and apparatus for anonymous user identification and content personalization in wireless communication; AT&T Mobility II LLC (US). No.: 12/468,708 ; September 10, 2009; August 28, 2012, Patent and Trademark Office. – 9 p.
19. Linden G. D., Smith B. R., Zada N. K. (US) Pat. US7,970,664B2 US Content personalization based on actions performed during browsing sessions; Amazon Technologies Inc (US). No.: 11/009,732 ; December 10, 2004; June 28, 2011, Patent and Trademark Office, 36 p.
20. Mehtaa P., Parekh B., Modi K., Solanki P. Web personalization using web mining: concept and research issue, *International Journal of Information and Education Technology*, 2012, Vol. 2(5), pp. 510–512.
21. Zhezhnych P., Markiv O. Linguistic Comparison Quality Evaluation of Web-Site Content with Tourism

DOI 10.15588/1607-3274-2020-1-12

Documentation Objects, *Advances in Intelligent Systems and Computing*, 2018, Vol. 689, pp. 656–667.
22. Basyuk T. The main reasons of attendance falling of internet resource, *Computer Sciences and Information Technologies : Xth International Scientific and Technical Conference, 14–17 September 2015 : proceedings*. Lviv, IEEE, 2015, pp. 91–93.
23. Gozhyj A., Chyrun L., Kowalska-Styczen A., Lozynska O. Uniform Method of Operative Content Management in Web Systems, *CEUR Workshop Proceedings,* 2018, Vol. 2136, pp. 62–77.
24. Kravets P. The control agent with fuzzy logic, *Perspective Technologies and Methods in MEMS Design : VIth International Conference, 20–23 April 2010 2015 : proceedings.* Lviv, IEEE, 2015, pp. 40–41.
25. Davydov M., Lozynska O. Linguistic Models of Assistive Computer Technologies for Cognition and Communication, *Computer Science and Information Technologies : XIth International Scientific and Technical Conference, 6–10 September 2016 : proceedings.* Lviv, IEEE, 2016, pp. 171–175.
26. Mykich K., Burov Y. Algebraic model for knowledge representation in situational awareness systems, *Computer Sciences and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings.* Lviv, IEEE, 2016, pp. 165–167.
27. Mykich K., Burov Y. Uncertainty in situational awareness systems, *Modern Problems of Radio Engineering, Telecommunications and Computer Science : 13th International Conference, 623–26 Februar 2016 : proceedings.* Lviv, IEEE, 2016, pp. 729–732.
28. Mykich K., Burov Y. Algebraic Framework for Knowledge Processing in Systems with Situational Awareness, *Advances in Intelligent Systems and Computing*, 2017, Vol. 512, pp. 217–227.
29. Mykich K., Burov Y. Research of uncertainties in situational awareness systems and methods of their processing, *EasternEuropean Journal of Enterprise Technologies,* 2016, Vol. 1(79), pp. 19–26.
30. Vysotska V. Linguistic Analysis of Textual Commercial Content for Information Resources Processing, *Modern Problems of Radio Engineering, Telecommunications and Computer Science : International Scientific and Technical Conference, 23–26 February 2016 : proceedings*. Lviv, IEEE, 2016, pp. 709–713.
31. Su J., Vysotska V., Sachenko A., Lytvyn V., Burov Y. Information resources processing using linguistic analysis of textual content, *Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications : 9th International Conference, 21–23 September 2017 : proceedings.* Bucharest, IEEE, 2017, pp. 573–578.
32. Lytvyn V., Vysotska V., Veres O., Rishnyak I., Rishnyak H. Content Linguistic Analysis Methods for Textual Documents Classification, *Computer Science and Information Technologies : 11th International Scientific and Technical Conference, 6–10 September 2016 : proceedings*. Lviv, IEEE, 2016, pp. 190–192.
33. Bisikalo O. V., Vysotska V. A. Identifying keywords on the basis of content monitoring method in ukrainian texts, *Radio Electronics, Computer Science, Control*, 2016, Vol. 1(36), pp. 74–83.
34. Bisikalo O. V., Vysotska V. A. Sentence syntactic analysis application to keywords identification Ukrainian texts, *Radio Electronics, Computer Science, Control,* Vol. 3(38), 2016, pp. 54–65.
35. Alieksieieva K., Berko A., Vysotska V. Technology of commercial web-resource management based on fuzzy logic *Radio Electronics, Computer Science, Control,* 2015, Vol. 3(34). pp. 71–79.
36. Lytvyn V., Bobyk I., Vysotska V. Application of algorithmic algebra system for grammatical analysis of symbolic computation expressions of propositional logic, *Radio Electronics, Computer Science, Control,* 2016, Vol. 4(39), pp. 54–67.
37. Lytvyn Vasyl, Vysotska Victoria, Dosyn Dmytro, Holoschuk Roman, Rybchak Zoriana Application of Sentence Parsing for Determining Keywords In Ukrainian Texts, *Computer Science and Information Technologies : 12th International Scientific and Technical Conference, 5–8 September 2017 : proceedings*. Lviv, IEEE, 2017, pp. 326–331.

DOI 10.15588/1607-3274-2020-1-12