DOI: 10.32347/2412-9933.2022.52.39-46
УДК 004.021

Катаєва Євгенія Юріївна
Кандидат технічних наук, доцент, доцент кафедри програмного забезпечення автоматизованих систем,
https://orcid.org/0000-0003-1277-8031
Черкаський державний технологічний університет, Черкаси
Якименко Дмитро Олегович
Аспірант кафедри програмного забезпечення автоматизованих систем,
https://orcid.org/0000-0002-6906-8164
Черкаський державний технологічний університет, Черкаси

МЕТОД ВИЗНАЧЕННЯ АТРИБУЦІЇ ДРУКОВАНИХ ДОКУМЕНТІВ

Анотація. *Швидкий розвиток сучасних інформаційних технологій спонукає організації і підприємства впроваджувати інноваційні технології в інформаційному забезпеченні їхньої діяльності. На сьогодні це має великий вплив на успішну реалізацію управління організацією та прийнятті правильних стратегічних рішень. Ефективне опрацювання постійно зростаючих обсягів інформації, основну частину яких складають документи та звіти різних форматів, можливе тільки за умови автоматизованої перевірки та обробки. Інформаційний пошук у неструктурованому тексті дуже складний, оскільки він містить велику кількість інформації, що вимагає використання специфічних методів і алгоритмів опрацювання для отримання корисних знань. У статті узагальнено результати експериментального дослідження застосування методу визначення атрибуції електронного документа. Проаналізовано роль опрацювання інформації і виявлення в ній моделей і тенденцій, які допомагають приймати рішення, а також принципи інтелектуального аналізу даних. Виокремлено напрями інтелектуального аналізу тексту, такі як: збирання даних, опрацювання web-даних, інформаційний пошук і виймання, комп'ютерна лінгвістика і обробка природної мови. Доведено доцільність реалізації прототипу програмного продукту, який відображає роботу методу, та засвідчує, що метод працює швидко і стабільно.*

Ключові слова: аналіз інформації; модифікований метод; інтелектуальний аналіз даних; обробка даних; програмний продукт

Вступ

Майже кожне підприємство й організація в Україні має постійний набір документів, які циркулюють між різними установами і мають специфічні вимоги до форматування. До таких документів належать звіти різної форми, статті, публікації тощо. Кількість таких паперів у рамках однієї установи може сягати десятків тисяч за рік, а робота над їх створенням займає значну частину роботи такої установи. Зазвичай велика кількість часу витрачається саме на рутинну працю перевірки форматування та відповідності стандартам того чи іншого виду документа.

Швидкий розвиток сучасних інформаційних технологій спонукає організації і підприємства впроваджувати інноваційні технології в інформаційному забезпеченні їхньої діяльності. На сьогодні це має великий вплив на успішну реалізацію управління організацією та прийнятті правильних стратегічних рішень. Управління інформаційним забезпечення діяльності підприємства тісно пов’язане з наявністю сучасних інформаційних ресурсів, а також з можливістю впровадження інноваційних підходів у впорядкуванні та опрацюванні документаційних та інформаційних потоків підприємства. Ефективне опрацювання постійно зростаючих обсягів інформації, основну частину яких складають документи та звіти різних форматів, можливе тільки за умови автоматизованої перевірки та обробки.

Аналіз тексту визначається як процес виявлення прихованого, корисного і цікавого зразка з неструктурованих текстових документів. Аналіз тексту також відомий як процес пошуку знань у текстовому інтелектуальному аналізі даних. Приблизно 80% корпоративних даних знаходиться в неструктурованому форматі. Інформаційний пошук у неструктурованому тексті дуже складний, оскільки він містить велику кількість інформації, що вимагає використання специфічних методів і алгоритмів опрацювання для отримання корисних знань. Оскільки найбільш поширеною формою для зберігання інформації є текст, інтелектуальний аналіз тексту видається більш важливим процесом, ніж інтелектуальний аналіз даних (data mining).

Інтелектуальний аналіз тексту є міждисциплінарною сферою, яка включає збирання

даних, опрацювання web-даних, інформаційний пошук і виймання, комп'ютерну лінгвістику і опрацювання природної мови.

Мета статті

Мета роботи – дослідження наявних методів аналізу текстових документів. Розроблення методу визначення атрибуції електронних документів засобами інтелектуального аналізу даних. Для отримання підтвердження ефективності використання методу розроблення відповідного програмне забезпечення.

Аналіз останніх досліджень

Для вилучення знань з текстової інформації використовуються різноманітні методи автоматичного аналізу Data Mining. Такі методи використовують алгоритми та засоби штучного інтелекту для дослідження і вилучення з великих об’ємів інформації знань, які будуть практично корисні та доступні для інтерпретації людиною [1]. До основних методів Data Mining належать: класифікація, кластеризація, регресія, пошук асоціативних правил, анотування та автореферування.

Задача класифікації зводиться до визначення класу об’єкта за його характеристиками, причому множина класів задається завчасно.

Класифікація – використовує статистичні кореляції для побудови правил розміщення документів у наперед заданій категорії; задача класифікації – це задача розпізнавання, коли система відносить новий об’єкт до тієї чи іншої категорії. У Data Mining задачу класифікації розглядають як визначення значення одного з параметрів об’єкта на основі значення інших параметрів [2].

Задача регресії подібна до задачі класифікації і дає змогу визначити за відомими характеристиками об’єкта значення деякого його параметра. Тут значенням параметра є не кінцева множина класів, а множина дійсних чисел.

Класифікація та регресія передбачають здійснення двох обов’язкових етапів. Перший етап – виділення набору об’єктів, для яких відомі значення залежних і незалежних змінних. На основі отриманого набору будується модель визначення значення залежної змінної (функція класифікації або регресії). На другому етапі побудовану модель застосовують до об’єктів, які аналізуються. Недоліком класифікації та регресії є те, що розробник системи повинен фіксувати кількість класів і характеристик, за якими буде проводитись дослідження. Це означає, що якщо система не виявить ознаки або класу, до якого можна віднести, наприклад, текстовий документ, він не буде коректно оброблений.

Анотування – це процес створення коротких повідомлень про електронний текст, які дають змогу робити висновки щодо доцільності його докладного вивчення [3]. Сучасні системи аналітичного опрацювання текстової інформації володіють засобами автоматичного складання анотацій, при цьому існує два підходи до вирішення цієї проблеми.

У першому підході програма-анотатор вилучає з першоджерела невелику кількість фрагментів, у яких найбільш повно представлено зміст документа. При другому підході анотація являє собою синтезований документ у вигляді короткого змісту. Анотація, сформована згідно з першим підходом, якісно поступається анотації, одержаній при синтезі. Для підвищення якості анотування необхідно вирішити проблему орієнтування на вузьку предметну область. Тоді у такому процесі необхідна участь людини.

Автоматичне реферування – являє собою створення коротких викладів матеріалів, анотацій, дайджестів, тобто вилучення найбільш важливих відомостей з одного або декількох документів, і генерації на їх основі лаконічних та інформаційно-ємних звітів. На сьогодні існує два основних напрями автореферування:

– квазіреферування (засноване на виділенні найбільш інформативних фраз і формування з них квазірефератів);

– коротке викладення змісту первинних документів [4].

Автоматичне реферування та анотування використовуються переважно для економії часу користувачам, створення каталогів інформаційних ресурсів, використання словників-тезаурусів загального та спеціального призначення. Застосовується автоматичне реферування й анотування в корпоративних системах документообігу, пошукових машинах та каталогах ресурсів Інтернет, автоматизованих інформаційно-бібліотечних системах, каналах зв’язку, службах розсилки новин і т.д.

Пошук асоціативних правил являє собою метод пошуку часткових залежностей між об’єктами та суб’єктами. Знайдені залежності представляються у вигляді правил та використовуються для кращого розуміння природи даних, що аналізуються. Тобто з великої кількості наборів об’єктів визначаються такі набори, що найчастіше зустрічаються. При виявленні закономірностей можна з певною ймовірністю передбачити появу подій у майбутньому, що допомагає приймати рішення. Така задача є різновидом задачі пошуку асоціативних правил і називається сиквенційним аналізом.

Інтелектуальний аналіз тексту можна поділити на дві фази [5]:

1. Фільтрація тексту (очищення від «сміття»).
2. Вилучення знань.

Фаза очищення тексту перетворює вихідну форму текстових документів в обрану проміжну. Вилучення знань, як видно з назви, отримує шаблони або знання з проміжної форми. Проміжна форма (Intermediate Form) може бути частково структурованою або структурованою. Проміжна форма може являти собою документ, де кожна сутність є іншим документом або певним поняттям, у якому кожна сутність являє собою об'єкт або набір даних з певної предметної області.

Аналіз проміжної форми документів надає зразки та взаємозв'язки серед всіх документів. Прикладом є кластеризація, візуалізація і категоризація документів [6].

Аналіз проміжної форми понять надає зразки та взаємозв'язок об'єктів або інших понять. Такі операції, як прогнозуюче моделювання та асоціативне дослідження, належать до цієї категорії аналізу. Проміжна форма документа може бути перетворена у проміжну форму поняття шляхом вилучення релевантної інформації, що стосується об'єктів з певної галузі інтересів. З цього випливає, що проміжна форма документа зазвичай не залежить від певної предметної області. Наприклад, набір новинних стрічок при виконанні фільтрації тексту перетворює кожен документ у придатну проміжну форму у вигляді документа. Потім можна виконати обробку знань з метою організації статей згідно з їхнім змістом з метою візуалізації і навігації. Для вилучення знань у певній предметній області проміжна форма документа може бути перетворена у проміжну форму поняття залежно від висунутих вимог. Наприклад, можна отримати інформацію, пов'язану з певним «товаром», з проміжної форми документа і сформувати базу даних товарів для надання знань про них [7].

Загальний процес інтелектуального аналізу даних представлено на схемі на рис. 1.

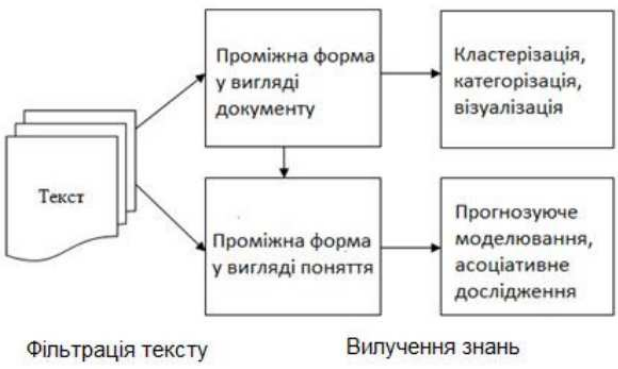

Рисунок 1 – Загальний процес інтелектуального аналізу даних

Кроки, які виконуються при інтелектуальному аналізі тексту

Кроки виконання аналізу тексту представлені нижче.

1. Попереднє опрацювання тексту. Попереднє опрацювання розділене на такі кроки [8]:

Перший крок попереднього опрацювання тексту далі розділений на токенізацію та видалення «стоп-слів».

Токенізація. Текстові документи містять набір сутностей. На цьому кроці виконується поділ тексту на окремі слова з видаленням пустих місць і знаків пунктуації.

Видалення «стоп-слів». На цьому кроці проводиться видалення артиклів (наприклад, для англійської мови це «a», «is», «of» та ін.) та сполучників, які не несуть самостійного смислового навантаження тексту (й, і, а, ні, та, то, бо, що, зате, однак, або, якщо, якби та ін.).

2. Перетворення тексту. Текстовий документ видається словами, з яких він складається, та інформацією про їх походження. Є два підходи, які використовуються для представлення документа: мішок слів і векторні простори слів [9].

3. Пошук ознак. Це також відомо як пошук змінних. Це – процес відбору підмножини важливих ознак для використання у створенні моделей. Ця фаза, в основному, виконує видалення надлишкових та неважливих ознак. Вибір ознак є підмножиною більш загальної області вилучення ознак [10].

4. Методи аналізу тексту. У цьому пункті інтелектуальний аналіз тексту стає збором даних. Методи розпізнавання даних, такі як кластеризація, класифікація, інформаційний пошук і т.д., можуть використовуватися також і для інтелектуального аналізу тексту [11].

5. Інтерпретація/Оцінка. На цьому кроці відбувається аналіз результатів залежно від поставлених цілей.

Аналіз програмного забезпечення для обробки текстів

Наявні відкриті системи порівняльного аналізу текстової інформації, такі як «Advego Plagiatus», «Shingles Expert», «Compare It!», «IsEqual», «Cognitive Dwarf», а також системи, що здійснюють повнотекстовий пошук та аналітичне опрацювання текстів, містять в своїй основі спільні механізми вилучення знань з текстової інформації та базуються на статистичних методах [12 – 15].

Порівняння – це зіставлення об’єктів з метою виявлення спільних рис або різниці між ними. Прийом порівняння використовується в процесі узагальнення, коли необхідно виявити тотожності, збіги та протиріччя в об’єктах дослідження. Тут тотожність – це повноцінний збіг усіх ознак;

збіг – узгодження ознак, починаючи з однієї; протиріччя – коли ознаки одних об'єктів відсутні в інших. Для здійснення порівняння необхідні ознаки, що визначають можливі відношення між об'єктами.

Одним із методів, що застосовується для виявлення кластерів документів, які мають схожі властивості лише за деякими ознаками, наприклад, словами чи зображеннями, є бікластеризація. Метод застосовується для здійснення запитів та індексації повнотекстових систем. Початкові дані являють собою матрицю, в якій рядки відповідають за слова, а стовпчики – за документи. Для кластеризації документів враховується кількість входжень слова до документа, загальна кількість документів та кількість документів, що містить певне слово [16]. Тобто, слова можуть бути кластеризовані на основі документів, в яких вони зустрічаються. Кластери зручні для автоматичної побудови статистичних тезаурусів, уточнення запитів та автоматичної класифікації документів, проте здійснити змістовний аналіз тексту з використанням кластерів неможливо. Дослідження показали, що ані коди бібліотечних класифікаторів, ані назва текстового документа, ані множина слів, що найчастіше зустрічаються у тексті, у більшості випадків недостатньо адекватні або зовсім неадекватні його змісту [17]. Тому при їх використанні як критерію добору текстів стандартний пошуковий сервер видає величезний обсяг інформації, більша частина якої немає ніякого відношення до тематики тексту, що підлягає аналізу.

Аналіз роботи існуючих відкритих систем порівняльного аналізу електронних документів дав змогу здійснити їх порівняльну характеристику (таблиця).

Із таблиці видно, що діапазон зміни відсотка збігу для різних систем і для однакових текстів великий і становить приблизно 20%. Це говорить про неточність роботи існуючих систем порівняльного аналізу та необхідність розроблення якісно нових алгоритмів екстракції знань із текстових документів.

Викладення основного матеріалу

Для розроблення модифікованого методу, першим кроком, була опрацьована загальна схема класифікації.

Загальна схема класифікації складається з таких етапів:

1. Попереднє опрацювання та індексація.
2. Зменшення розмірності множини ознак.

На етапі попереднього опрацювання та індексації документів формуються ознаки документа, в якості яких виступають всі його значущі слова або словосполучення. Цей етап включає в себе розбиття тексту на більш дрібні об'єкти, наприклад, речення, фрази або слова; видалення розмітки, пунктуації, цифр, перетворення всіх букв у прописні; видалення функціональних слів (широковживаних слів, які не несуть самостійного сенсу, але мають значну частоту використання в текстах, – це сполучники, прийменники, частки, займенники, артиклі тощо).

Зменшення розмірності множини ознак – це процес надання ваги словам, залежно від їх важливості для класифікації тексту, та подальше видалення маловагомих термів з множини ознак.

Таблиця – Характеристика систем порівняльного аналізу текстових документів

Параметр / Назва програми	Антиплагіат	StrikePlagiarism	Advego Plagiatus	Cognitive Dwarf
Концепція отримання результатів порівняльного аналізу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу, розширена оцінка відсотку збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу	Показує ступінь унікальності тексту, джерела дублювання тексту, відсоток збігу
Функціональні можливості	Пошук серед існуючих документів у базі даних	Порівняння текстових документів з існуючими у доступних пошукових системах	Порівняння текстових документів з існуючими у доступних пошукових системах	Порівняння текстових документів з існуючими у доступних пошукових системах
Мова	укр., рос., англ., франц., нім., чеська, польська	укр., рос., англ., франц., нім., чеська, польська	укр., рос., англ.	укр., рос., англ.
Швидкодія	15–20 хв	4 хв – декілька год	15 – 20 хв	15 – 20 хв
% збігу	82,3 %	76,1%	65%	61,5%

Для видалення термів встановлюється порогова вага, терми нижче якої вважаються не важливими.

За рахунок зменшення розмірності множини термінів можна знизити ефект перенавчання – явище, за якого класифікатор орієнтується на випадкові або помилкові характеристики навчаючих даних, а не дійсно важливі.

Основним підходом до попереднього опрацювання текстових документів, для так званого визначення атрибуції документа використовується статистичний підхід. Методи цього підходу, як правило, полягають в аналізі частот зустрічальності слів у текстах у тій чи іншій його варіації і у використанні цієї інформації в процесі виявлення та відбору представницьких ознак документів.

Процес надання ваги словам відбувається за рахунок використання методу визначення ваг ознак документа – TF-IDF.

TF-IDF – статистичний показник, що використовується для оцінки важливості слів у контексті документа, що є частиною колекції документів. Значущість слова пропорційна кількості вживань цього слова в документі і обернено пропорційна частоті вживання слова в інших документах колекції. Міра TF-IDF часто використовується для подання документів колекції у вигляді числових векторів, що відображають важливість використання кожного слова з деякого набору слів у кожному документі.

Першим кроком в обробці текстів є розрахунок ваг TF-IDF для кожного слова ω в кожному документі Φ_P^* [18]:

$$\mathrm{tf}(\omega, \Phi_P^*) = \frac{n_{\omega\Phi_P^*}}{n_{\Phi_P^*}}, \qquad (1)$$

де $n_{\omega\Phi_P^*}$ – кількість входжень слова ω документ; $n_{\Phi_P^*}$ – загальна кількість слів у документі.

IDF (обернена частота документа) – ін-версія частоти, з якою слово зустрічається в документах колекції. Використання IDF зменшує вагу широковживаних слів [18]

$$\mathrm{idf}(\omega) = \frac{|\Phi_P^*|}{|\Phi_P^* \supset \omega|}, \qquad (2)$$

де $|\Phi_P^*|$ – загальна кількість проаналізованих документів; $|\Phi_P^* \supset \omega|$ – кількість документів, у яких зустрічається слово ω (коли $n_{\Phi_P^*} \neq 0$).

Отже, $\mathrm{tfidf} = \mathrm{tf} * \mathrm{idf}$ (3)

більшу вагу TF-IDF отримають слова з високою частотою появи в межах документа та низькою частотою вживання в інших документах колекції [18].

Результати

За результатами досліджень було розроблено два додатки: один додаток призначений для демонстрації результатів роботи методу з текстом (додаток з user friendly інтерфейсом), другий додаток з консольним інтерфейсом для проведення тестувань роботи методу без впливу інтерфейсної частини на швидкодію.

Додаток з user friendly інтерфейсом наведено на рис. 2.

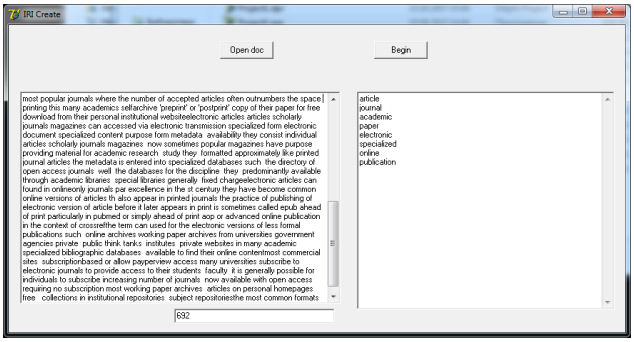

Рисунок 2 – Додаток з user friendly інтерфейсом

Додаток з консольним інтерфейсом для тестування швидкодії роботи методу (рис. 3).

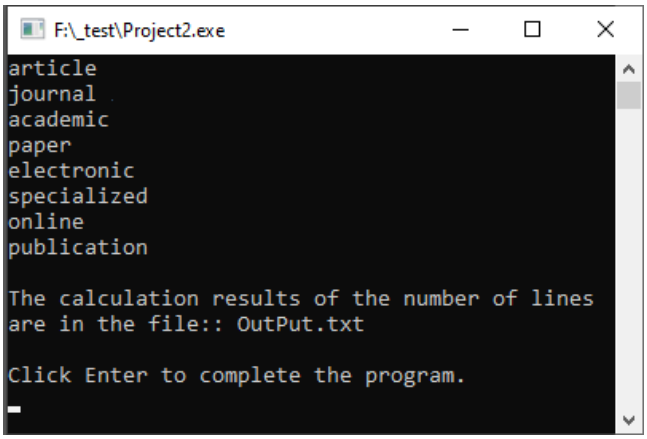

Рисунок 3 – Додаток з консольним інтерфейсом для тестування швидкодії роботи методу

Графік залежності часу обробки документа від кількості слів у документі (рис. 4).

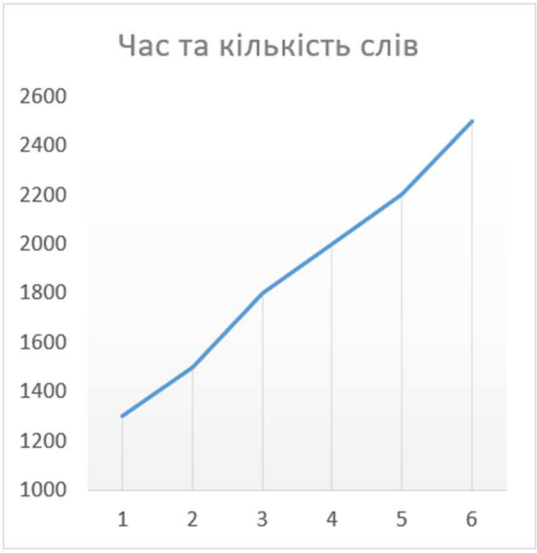

Рисунок 4 – Графік залежності часу обробки документа від кількості слів у документі

У результаті було отримано дані, на основі яких побудовано графіки ефективності роботи модифікованого методу у порівнянні з аналоговим (рис. 5 – 7).

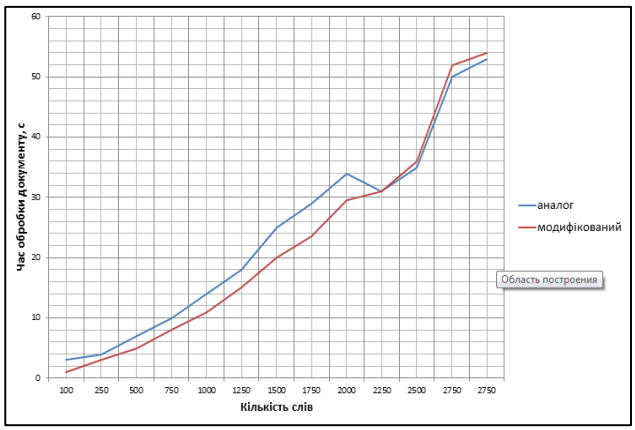

Рисунок 5 – Графік зміни часу роботи від розміру документа

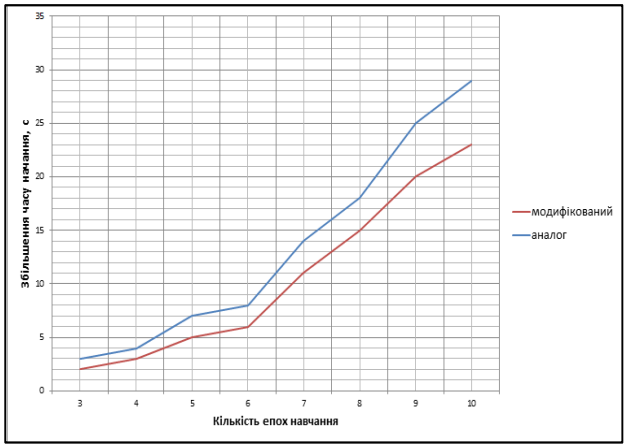

Рисунок 6 – Графік зміни часу роботи від кількості епох навчання

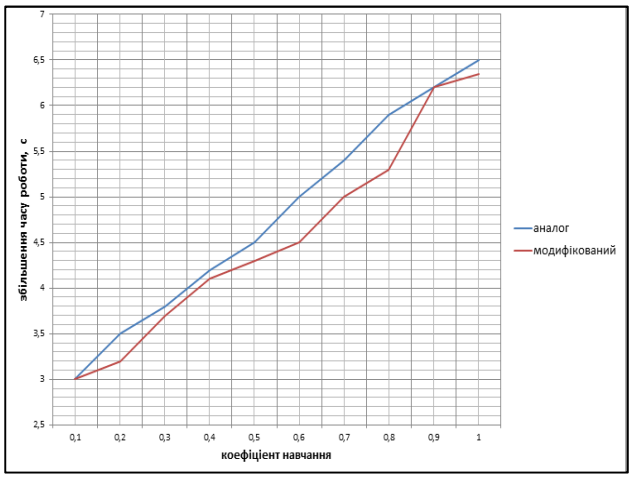

Рисунок 7 – Графік зміни часу роботи від значень коефіцієнта навчання

На графіках показано криві, які відповідають значенням часу роботи аналогового та створеного модифікованого методу. Можна побачити, що ефективність модифікованого методу порівняно з аналогом вища, однак, як свідчать графіки, великий об'єм документа, дуже сповільнює швидкість роботи модифікованого методу, в результаті чого швидкість обробки даних обох методів збігається, а на значні обсяги документа модифікований метод витрачає більше часу на опрацювання документа.

Висновок

Розглянуто поняття аналізу тексту. Визначено, що аналіз тексту є процесом виявлення прихованих і корисних шаблонів у неструктурованих текстових документах. З іншого боку, інтелектуальний аналіз тексту є міждисциплінарною сферою, що включає збирання та опрацювання даних, пошук інформації, комп'ютерну лінгвістику та обробку природньої мови.

Для отримання даних з текстової інформації використовуються різні методи автоматичного інтелектуального аналізу даних. Такі методи використовують алгоритми та засоби штучного інтелекту для пошуку та отримання даних з великого обсягу інформації, які будуть практично корисні та доступні для інтерпретації людиною. В основі основних методів передачі даних лежать категорійний, кластерний, регресійний пошук за асоціативними правилами, анотування та самореференція. Відомі відкриті системи порівняльного аналізу текстової інформації, такі як Advego Plagiatus, Shingles Expert, Compare It!, IsEqual, Cognitive Dwarf, а також системи, що полегшують повнотекстовий пошук та аналітичне опрацювання текстів, у тому числі на основі єдиних механізмів отримання знань з текстової інформації та на основі статистичних методів. Основний інструмент попереднього опрацювання текстових документів називається створенням документа з використанням статистики. Ці методи, як правило, спрямовані на аналіз частоти слів у текстах в одній або декількох варіаціях.

У результаті розроблено та представлений метод визначення атрибуції електронних документів засобами інтелектуального аналізу даних. Отримано дані, на основі яких було побудовано графіки ефективності роботи методу в порівнянні з аналоговим. На основі цих даних можна зробити висновок, що ефективність модифікованого методу, порівняно з аналогом, вища.

Список літератури

1. Piatetsky-Shapiro G. Data Mining and Knowledge Discovery – 1996 to 2005: Overcoming the Hype and moving from "University" to "Business" and "Analytics": Data Mining and Knowledge Discovery journal, 2007. 365 p.
2. Барсегян А. А., Куприянов М. С., Степаненко В. В., Холод И. И. Технологии анализа данных: Data Mining, Viual Mining, Text Mining, OLAP: Спб.: БХВ-Петербург, 2007. 384с.
3. Башмаков А. И. Интеллектуальные информационные технологии: Учеб. пособие. Москва : Изд-во МГТУ им. Баумана, 2005. 304 с.
4. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа, Москва : ООО "Вильямс", 2005. 272 с.
5. Walter L., Radauer A., Moehrle M. The beauty of brimstone butterfly: novelty of patents identified by near environment analysis based on text mining. Scientometrics (en), 2017. 111 (1). p. 103–115.
6. Roll U., Correia R. A., Berger-Tal O. Using machine learning to disentangle homonyms in large text corpora. Conservation Biology (en), 2018. 32 (3). p. 716–724.
7. Ramiro H. G., Agustín G. Assessing the usefulness of online message board mining in automatic stock prediction systems. Journal of Computational Science, 2017. 19.
8. Renganathan V. Text Mining in Biomedical Domain with Emphasis on Document Clustering. Healthcare Informatics Research, 2017. 23 (3). p. 141–146.
9. Chang W. L., Tay K. M., Lim C. P. A New Evolving Tree-Based Model with Local Re-learning for Document Clustering and Visualization. Neural Processing Letters, 2017. 46. p. 379–409.
10. Paltoglou G., Thelwall M. Unsupervised Sentiment Analysis in Social Media. ACM Transactions on Intelligent Systems and Technology (TIST), 2012. p. 66.
11. Advego Plagiatus – перевірка унікальності тексту. URL: www.advego.com/plagiatus/ (дата звернення: 25.08.2022).
12. Он-лайн сервіс провірки тексту на унікальність. URL:www.text.ruwww.advego.com/plagiatus/.(дата звернення: 25.08.2022).
13. Антиплагіатна Інтернет-система. URL: http://strikeplagiarism.com/ua/antiplagiarism-system/(дата звернення: 25.08.2022) 14.
14. Програма для порівняння текстів «Shingles Expert». URL: http://makebusiness.ru/seo/37 (дата звернення: 25.08.2022)
15. Zubrytskyi A. Yu. Intellectual system of text research and analysis . Master's thesis - national technical university of Ukraine "Ihory Sikors'koho Kyiv polytechnic institute", 2019. № 31.
16. Ландэ Д.В. Поиск знаний в Internet. Профессиональная работа. Москва : ИД “Вильямс”, 2005. 272 с.

Стаття надійшла до редколегії 10.12.2022

Kataieva Yevheniia
Phd, associate professor at the Department of software for automated systems,
https://orcid.org/0000-0002-9668-4739
Cherkasy State Technological University, Cherkasy
Yakymenko Dmytro
Postgraduate student, Department of software for automated systems,
https://orcid.org/0000-0002-6906-8164
Cherkasy State Technological University, Cherkasy

METHOD OF DETERMINING THE ATTRIBUTION OF PRINTED DOCUMENTS

Abstract. *The rapid development of modern information technologies encourages organizations and enterprises to introduce innovative technologies in the information provision of their activities. Currently, this has a great impact on the successful implementation of organizational management and the adoption of correct strategic decisions. Effective processing of ever-growing amounts of information, the main part of which consists of documents and reports of various formats, is possible only under the condition of automated verification and processing. Information search in unstructured text is very difficult, because it contains a large amount of information, which requires the use of specific processing methods and algorithms to obtain useful knowledge. The article summarizes the results of an experimental study of the method of determining the attribution of an electronic document. The role of information processing and the identification of models and trends in it, which help to make decisions, as well as the principles of intelligent data analysis, are analyzed. Areas of intellectual text analysis, such as data collection, web data processing, information search and extraction, computer linguistics and natural language processing, are singled out. The expediency of implementing a prototype software product that reflects the operation of the method and shows that the method works quickly and stably has been proven.*

Keywords: Information analysis; modified method; intelligent data analysis; data processing; software product

References

1. Piatetsky-Shapiro, G. (2007). Data Mining and Knowledge Discovery – 1996 to 2005: Overcoming the Hype and moving from "University" to "Business" and "Analytics". *Data Mining and Knowledge Discovery journal*, 365.
2. Barsehian, A. A., Kupryianov, M. S., Stepanenko, V. V., Kholod, Y. Y. (2007). Tekhnolohiiof data analisys: Data Mining, Viual Mining, Text Mining, OLAP: Spb.: BKhV-Peterburh, 384.
3. Bashmakov, A. Y. (2005). Intellectual information technologies: Textbook. Moscow: Publ. MHTU im. Baumana, 304.
4. Lande, D. V. (2005). Data search in the internet. Professional work. Moscow: OOO "Williams", 272.
5. Walter, L., Radauer, A., Moehrle, M. (2017). The beauty of brimstone butterfly: novelty of patents identified by near environment analysis based on text mining. *Scientometrics*, 111 (1), 103–115.
6. Roll, U., Correia, R. A., Berger-Tal, O. (2018). Using machine learning to disentangle homonyms in large text corpora. *Conservation Biology*, 32 (3), 716–724.
7. Ramiro, H. G., Agustín, G. (2017). Assessing the usefulness of online message board mining in automatic stock prediction systems. *Journal of Computational Science*, 19.
8. Renganathan, V. (2017). Text Mining in Biomedical Domain with Emphasis on Document Clustering. *Healthcare Informatics Research*, 23 (3), 141–146.
9. Chang, W. L., Tay, K. M., Lim, C. P. (2017). A New Evolving Tree-Based Model with Local Re-learning for Document Clustering and Visualization. *Neural Processing Letters*, 46, 379–409.
10. Paltoglou, G., Thelwall, M. (2012). Unsupervised Sentiment Analysis in Social Media. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 66.
11. Advego Plagiatus – perevirka unikalnosti tekstu [electronic source]. www.advego.com/plagiatus/
12. On-lain servis provirky tekstu na unikalnist [electronic source]. www.text.ruwww.advego.com/plagiatus/
13. Antyplahiatna Internet-systema [electronic source]. http://strikeplagiarism.com/ua/antiplagiarism-system/
14. Prohrama dlia porivniannia tekstiv «Shingles Expert» [electronic source]. http://makebusiness.ru/seo/37
15. Zubrytskyi, A. Yu. (2019). Intellectual system of text research and analysis. Master's thesis – national technical university of Ukraine "Ihor Sikors'ky Kyiv polytechnic institute", 31.
16. Lande D. V. (2005). Search for knowledge on the Internet. Professional work. Moscow: Publishing House “Williams”, 272 p.

Посилання на публікацію

APA Kataieva, Yevheniia & Yakymenko, Dmytro. (2022). Method of determining the attribution of printed documents. *Management of Development of Complex Systems*, 52, 39–46. [in Ukrainian], dx.doi.org\10.32347/2412-9933.2022.52.39-46.

ДСТУ Катаєва Є. Ю., Якименко Д. О. Метод визначення атрибуції друкованих документів. *Управління розвитком складних систем*. Київ, 2022. № 52. С. 39 – 46, dx.doi.org\10.32347/2412-9933.2022.52.39-46.