УДК 004.9

Бісікало О. В.[1], Висоцька В. А.[2]

[1]Д-р техн. наук, професор, декан факультету комп'ютерних систем і автоматики Вінницького національного технічного університету, Вінниця, Україна

[2]Канд. техн. наук, доцент кафедри «Інформаційні системи та мережі» Національного університету «Львівська політехніка», Львів, Україна

ВИЯВЛЕННЯ КЛЮЧОВИХ СЛІВ НА ОСНОВІ МЕТОДУ КОНТЕНТ-МОНІТОРИНГУ УКРАЇНОМОВНИХ ТЕКСТІВ

Вирішено завдання розробки алгоритмічного забезпечення процесів контент-моніторінгу для розв'язання задачі визначення ключових слів україномовного тексту. Розглянуто формальне обгрунтування методу контент-моніторінгу тексту за допомогою стеммера Портера, в основу модифікації стемінгу покладено відомі результати класифікації морфемної і словотвірної структури дериватів української мови, виявлення закономірностей комбінаторики афіксів, моделювання структурної організації дієслів і суфіксальних іменників, а також морфонологічних модифікацій у процесі словозміни дієслова та словозміні і словотворенні прикметників української мови. Проведено декомпозицію методу та розроблено алгоритмічне забезпечення його основних структурних складових за результатами контент-аналізу тексту. Теоретично виявлено способи покращення показників ефективності пошуку ключових слів, зокрема щільності ключовиків у тексті. На основі розробленого програмного забезпечення отримано результати експериментальної апробації запропонованого методу контент-моніторінгу для визначення ключових слів в наукових текстах технічного профілю. Виявлено, що для обраної експериментальної бази зі 100 робіт найкращих результатів за критерієм щільності досягає метод аналізу статті без початкової обов'язкової інформації і без списку літератури, але із перевіркою уточнених заблокованих слів та уточненого тематичного словника.

Ключові слова: текст, україномовний, алгоритм, контент-моніторінг, ключові слова, контент-аналіз, стеммер Портера, лінгвістичний аналіз, синтаксичний аналіз.

НОМЕНКЛАТУРА

ІТ – інформаційні технології;

СЕКК – система електронної контент-комерції;

е-бізнес – електронний бізнес;

Е-комерція – електронна комерція;

ПЗ – програмне забезпечення;

$X = \{x_1, x_2 \ldots, x_{n_X}\}$ – множина вхідних даних $x_i \in X$ з різних інформаційних ресурсів або від модераторів при $i = \overline{1, n_X}$;

$C = \{c_1, c_2 \ldots, c_{n_C}\}$ – множина комерційного контенту $c_r \in C$ при $r = \overline{1, n_C}$;

C_0 – сформований комерційний контент;

C_1 – відфільтрований комерційний контент;

C_2 – відформатований комерційний контент;

C_3 – комерційний контент з визначеною множиною ключових слів;

$< U_C, U_G, U_K >$ – набір критеріїв для текстового контенту X;

$U_C = \{U_{C1}, U_{C2}, \ldots, U_{Cn_C}\}$ – множина критеріїв створення комерційного контенту;

$U_G = \{U_{G1}, U_{G2}, \ldots, U_{Gn_G}\}$ – множина критеріїв збирання комерційного контенту (фільтри);

$U_K = \{U_{K1}, U_{K2}, U_{K3}, U_{K4}\}$ – множина критеріїв визначення ключових слів в контенті;

U_{K1} – унікальність термів – іменників, словосполучень іменників або прикметника з іменником серед множини слів контенту;

U_{K2} – частота появи ключових слів комерційного контенту;

U_{K3} – кількість знаків без пробілів для $Noun \in U_{K1}$ при $Unicity \geq 80$;

U_{K4} – критерій формування множини ключових слів;

$T = \{t_1, t_2 \ldots, t_{n_T}\}$ – час $t_p \in T$ транзакції формування контенту при $p = \overline{1, n_T}$;

$\alpha_0 : (X, U_C, T) \to C_0$ – оператор створення контенту – відображення даних з різних джерел у контент, який відрізняється актуальністю;

$\alpha_1 : (X, U_G, T) \to C_0$ – оператор збирання контенту – відображення даних від авторів у контент, який відрізняється достовірністю та актуальністю;

$\alpha_2 : (C_0, T, U_B) \to C_1$ – оператор виявлення дублювання контенту – відображення контенту в новий стан, який відрізняється унікальністю;

$\alpha_3 : (C_1, U_{FR}, T) \to C_2$ – оператор форматування контенту – відображення контенту в новий стан, який відмінний від попереднього форматом подання;

$\alpha_4 : (C_2, U_K, T) \to C_3$ – оператор виявлення ключових слів контенту – відображення контенту в новий стан, який відрізняється наявністю множини ключових слів, що загально описують його зміст.

ВСТУП

Активний розвиток мережі Інтернет сприяє зростанню потреб в отриманні оперативних даних виробничого/стратегічного характеру і реалізації нових форм інформаційного обслуговування через сучасні ІТ е-бізнесу [1–3]. Документована інформація, підготовлена відповідно до потреб користувачів, є інформаційним продуктом або комерційним контентом, наприклад, електронний матеріал Інтернет-видавництва, маркетингові дослідження,

DOI 10.15588/1607-3274-2016-1-9

консалтингові послуги тощо. Дії для забезпечення користувачів комерційним контентом є інформаційною послугою. Інтернет-ринок є сукупністю економічних, правових, організаційних і програмних відносин з продажу інформаційних продуктів/послуг між виробниками, постачальниками та користувачами [1–3].

Комерційний контент визначають як:

– вміст інформаційних ресурсів в СЕКК;

– об'єкт бізнес-процесів в СЕКК, наприклад, стаття, ПЗ, книга тощо;

– структурована та логічно завершена множина даних, що є об'єктом взаємовідносин між користувачем та СЕКК;

– набір електронних даних без наперед визначеної структури;

– дані комерційного призначення, що неподільні в часі;

– основний чинник формування області діяльності, функціонування та призначення СЕКК.

Сьогодні е-комерція є об'єктивною реальністю та перспективним бізнес-процесом. Інтернет є бізнес-середовищем, а комерційний контент є товаром з найбільшим попитом у ньому та основним об'єктом процесів електронної контент-комерції. Комерційний контент можна зразу замовити, оформити, оплатити і отримати on-line як товар. Через Інтернет продають весь спектр комерційного контенту – наукові та публіцистичні статті, музика, книги, фільми, фото, ПЗ тощо. Відомими корпораціями, які реалізують електронну контент-комерцію, є Google через Play Market, Apple – Apple Store, Amazon – Amazon.com [1].

Більшість рішень та досліджень зроблено на рівні реальних прикладних проектів, а сучасні СЕКК побудовані за закритим принципом як разові проекти та орієнтовані на реалізацію комерційного контенту, створеного за їх межами. Тому для проектування, створення, впровадження та супроводу СЕКК потребують розробки загальні методи та інформаційні технології формування, управління та супроводу комерційного контенту. З огляду на важливість для функціонування СЕКК ключових слів об'єктом дослідження обрано процес виявлення ключових слів в україномовних текстах у режимі реального часу, предмет дослідження – методи та моделі контент-моніторінгу таких текстів.

1 ПОСТАНОВКА ЗАДАЧІ

Нехай україномовний текстовий контент X з різних джерел інформації у вигляді $X = \{x_1, x_2 \ldots, x_{n_X}\}$ має стати основою відфільтрованого контенту C_1, відформатованого контенту C_2 та його модифікації C_3 з визначеною множиною ключових слів $KeyWords \in U_{K4}$. За відомими критеріями $< U_C, U_G, U_K >$ потрібно визначити оператор виявлення ключових слів комерційного контенту $\alpha_4 : (C_2, U_K, T) \to C_3$ та експериментально перевірити параметр частоти появи ключових слів комерційного контенту U_{K2} за різними режимами роботи алгоритмічного забезпечення запропонованого методу контент-моніторінгу.

2 ОГЛЯД ЛІТЕРАТУРИ

Сталою сучасною тенденцією можна вважати постійний ріст темпів виробництва текстового контенту в Інтернет-просторі. Цей процес є об'єктивним і позитивним, але виникла проблема – прогрес у галузі виробництва текстового контенту призводить до пониження загального рівня інформованості потенційного користувача Інтернет-простору [1–3]. Крім збільшення обсягів текстового контенту до масштабів, яке унеможливлює його безпосереднє опрацювання та помітно гальмує його поширення виникає низка специфічних проблем (табл. 1).

Негативні чинники у формуванні текстового контенту ускладнюють процес пошуку необхідних даних при скануванні різних джерел інформації. Збільшення фізичного обсягу та змінність співвідношення актуальності/динаміки контентних потоків (наслідок систематичного або нерегулярного оновлення) призводить до виникнення дублювання, інформаційного шуму та надмірності результатів пошуку контенту. Охоплення та узагальнення великих динамічних потоків контенту, які неперервно генерують в Інтернет-джерелах, вимагає якісно нових методів/підходів пошуку – таких як контент-моніторинг (рис. 1) на основі аналізу ключових слів [1–32]. Вхідною інформацією для контент-моніторингу є текст на природній мові як послідовність символів, вихідна інформація – це таблиці розділів, речень і лексем аналізованого тексту. Контент-моніторинг є програмним засобом автоматизації знаходження найбільш важливих складових в потоках контенту за допомогою алгоритмів стемінгу [1–32]. Це змістовний аналіз потоків контенту з метою постійного отримання необхідних якісних/кількісних зрізів на протязі наперед не визначеного проміжку часу.

Мета роботи полягає у створенні алгоритмічного забезпечення методу контент-моніторингу україномовних текстів на основі стеммера Портера та його застосуванні для виявлення значущих ключових слів. Для досягнення

Таблиця 1 – Основні негативні чинники у формуванні текстового контенту

Назва	Основна причина	Рішення
Інформаційний шум	Структурованість масивів контенту.	Фільтри, контент-моніторинг, аналіз сайту, контент-аналіз.
Паразитичний контент	Поява в якості додатків.	Фільтри, контент-моніторинг, контент-аналіз.
Нерелевантність контенту	Невідповідність потребам користувачів.	Створення анотованої бази даних, пошукових образів первинного контенту та їх кластеризація, контент-аналіз.
Дублювання контенту	Дублювання в джерелах.	Контент-аналіз, сканери і фільтри на базі статистики та критеріїв.
Навігація в потоці контенту	Швидкий ріст обсягу і поширення контенту.	Аналіз сайту, фільтри, контент-моніторинг, контент-аналіз.
Надмірність пошуку	Дублювання і нерелевантність.	Анотований пошук, контент-аналіз та реферування.

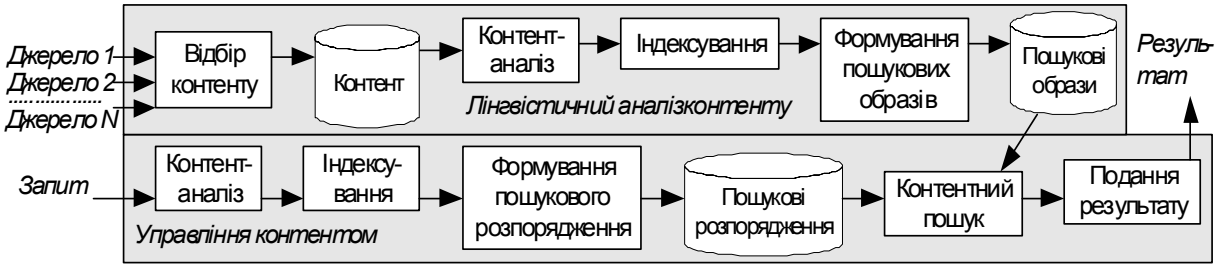


Рисунок 1 – Структурна схема процесу контент-моніторингу текстових масивів даних

мети пропонується розв'язати такі задачі дослідження: провести контент-аналіз текстової інформації; забезпечити визначення множини ключових слів; провести лінгвістичний аналіз текстового контенту; розробити синтаксичний аналізатор текстового контенту.

3 МАТЕРІАЛИ І МЕТОДИ

Головною складовою контент-моніторингу є контентний пошук та контент-аналіз тексту. Контент-аналіз призначений для пошуку контенту в масиві даних за змістовими лінгвістичними одиницями (алг. 1). Одиниця рахунку є кількісною мірою одиниці аналізу, що дозволяє реєструвати частоту (регулярність) появи ознаки категорії аналізу в тексті (кількість певних слів або їх поєднань, рядків, друкованих знаків, сторінок, абзаців, авторських аркушів, площа тексту тощо).

Алгоритм 1. Контент-аналіз текстового контенту.

Етап 1. Визначення набору критеріїв $< U_C, U_G >$ для текстового контенту X.

Крок 1. Формування набору критеріїв як тип джерела (форум, електронна пошта, Інтернет-газета, чат, Інтернет-журнал); тип контенту (стаття, е-лист, банер, коментарій); учасники комунікації (відправник, одержувач, реципієнт).

Крок 2. Визначення розміру (мінімальний обсяг або довжина), частоти появи, способу/місця розповсюдження та час появи контенту.

Крок 3. Фільтрування згідно сформованого набору критеріїв контентного потоку та зберігання ідентифікованого релевантного контенту X.

Етап 2. Контент-аналітичний відбір. Формування вибіркової сукупності контенту X' за критеріями обмеженої вибірки $< U_C, U_G >$ з більшого масиву X.

Етап 3. Виявлення змістовних одиниць аналізу $< U'_C, U'_G >$ текстового комерційного контенту X' (словосполучення, речення, тема, ідея, автор, персонаж, соціальна ситуація, частина тексту, кластеризована за змістом категорії аналізу) за модифікованим алгоритмом Потера. Вимоги до вибору лінгвістичної одиниці аналізу: велика для інтерпретації значення; мала, щоб не інтерпретувати багато значень; легко ідентифікується; кількість одиниць велика для проведення вибірки.

Етап 4. Виділення одиниць рахунку аналізу текстового контенту X'.

Крок 1. Якщо одиниці рахунку $< U_C, U_G >$ збігаються з одиницями аналізу $< U'_C, U'_G >$, то знаходять частоти появи виділеної змістовної одиниці, інакше перейти до кроку 2.

Крок 2. Модератор на основі аналізованого контенту висуває та доповнює одиниці рахунку $< U_C, U_G >$, наприклад, протяжність текстів; площа тексту, заповнена змістовними одиницями; кількість рядків (абзаців, знаків, колонок тексту); розмір/вид файлу; кількість рисунків з певним змістом/сюжетом тощо.

Етап 5. Порівняння змістовних одиниць аналізу $< U'_C, U'_G >$ з одиницями $< U_C, U_G >$.

Крок 1. Класифікація за угрупованнями із оціненням ваги змістовних категорій в загальному обсязі тексту. Класифікатором є загальна таблиця, в яку зведені всі категорії аналізу і одиниці аналізу. Фіксують одиниці виразу категорій.

Крок 2. Статистичні розрахунки зрозумілості та атрактивності контенту.

Етап 6. Розроблення інструменту контент-аналізу.

Крок 1. Створення закодованого протоколу контенту X' для компактності подання даних та швидкого порівняння результатів аналізу різного контенту.

Крок 2. Заповнення протоколу контенту X' властивостями (автор, час видання, обсяг тощо).

Крок 3. Заповнення протоколу контенту X' підсумками його аналізу (кількість вживання в ньому певних одиниць аналізу і висновки щодо категорій аналізу). Протокол кожного контенту X' заповнюється на основі підрахунку даних всіх його реєстраційних карток.

Етап 7. Розроблення таблиці контент-аналізу. Тип таблиці визначають у вигляді системи скоординованих і субординованих категорій аналізу: кожна категорія (питання) передбачає ряд ознак (відповідей), за якими квантифікується зміст тексту X'.

Етап 8. Розроблення кодувальної матриці контент-аналізу.

Крок 1. Якщо обсяг вибірки і 100 одиниць, то аналізується набір матричних листів, інакше виконати крок 2.

Крок 2. Якщо вибірка < 100 одиниць, то проводиться двовимірний аналіз. В цьому випадку для кожного контенту X' формується кодувальна матриця.

Етап 9. Проведення аналізу тексту X' згідно створених кодувальних матриць.

Етап 10. Інтерпретація результатів $\alpha_0 : (X, U_C, T) \to C_0$ та $\alpha_1 : (X, U_G, T) \to C_0$. Виявляють і оцінюють характеристики контенту X' на основі статистичного набору підрахованих коефіцієнтів за певний період часу на визначену категорію. Охоплює всі здобуті фрагменти тексту C_0, висновки спираються не на частину результатів, а враховуються всі без винятку. Фільтрування $\alpha_3 : (C_1, U_{FR}, T) \to C_2$ та форматування $\alpha_3 : (C_1, U_{FR}, T) \to C_2$ комерційного контенту.

Застосування контент-аналізу при моніторингу Інтернет-джерел даних дозволяє автоматизувати процес знаходження найбільш важливих складових в потоках контенту при відборі даних з цих джерел. Це усуває дублювання контенту, інформаційний шум, паразитичний контент, надмірність результатів пошуку тощо. Даний метод застосовують в подальших етапах формування контенту для отримання більш точного релевантного результату – створення унікального комерційного контенту, який користується попитом серед користувачів СЕКК.

З метою реалізації контент-аналізу текстових масивів даних для формування множини ключових слів було розроблено програму на основі стеммера Потера, адаптованого до української мови (алг. 2), а також таблиці основ основних тематичних слів для подальшої рубрикації, текстів, що досліджуються (табл. 2). Блок-схему алгоритму наведено на рис. 2.

Таблиця 2 – Основні складові програми формування ключових слів

№	Назва	Пояснення
1	Filter	список потенційних ключовиків з аналізованого тексту з розрахунком відносної частоти їх появи в тексті
2	Input	вхідний текст для аналізу та визначення ключовиків
3	Format	тестувальний відформатований вхідний текст
4	Parser	парсер, адаптований до української
5	Stemer	правила стеммера Потера, адаптований до української
6	Object	класи об’єктів для стеммера Потера
7	Resvoc	список ключовиків (частоти вживання яких в тексті попали у визначений діапазон згідно алг. 2 на рис. 2 та відповідають тематичному словнику Thematic.txt)
8	Thematic	тематичний словник (формується модератором)
9	Vocab	список слів з аналізованого тексту з розрахунком абсолютної частоти їх появи в тексті

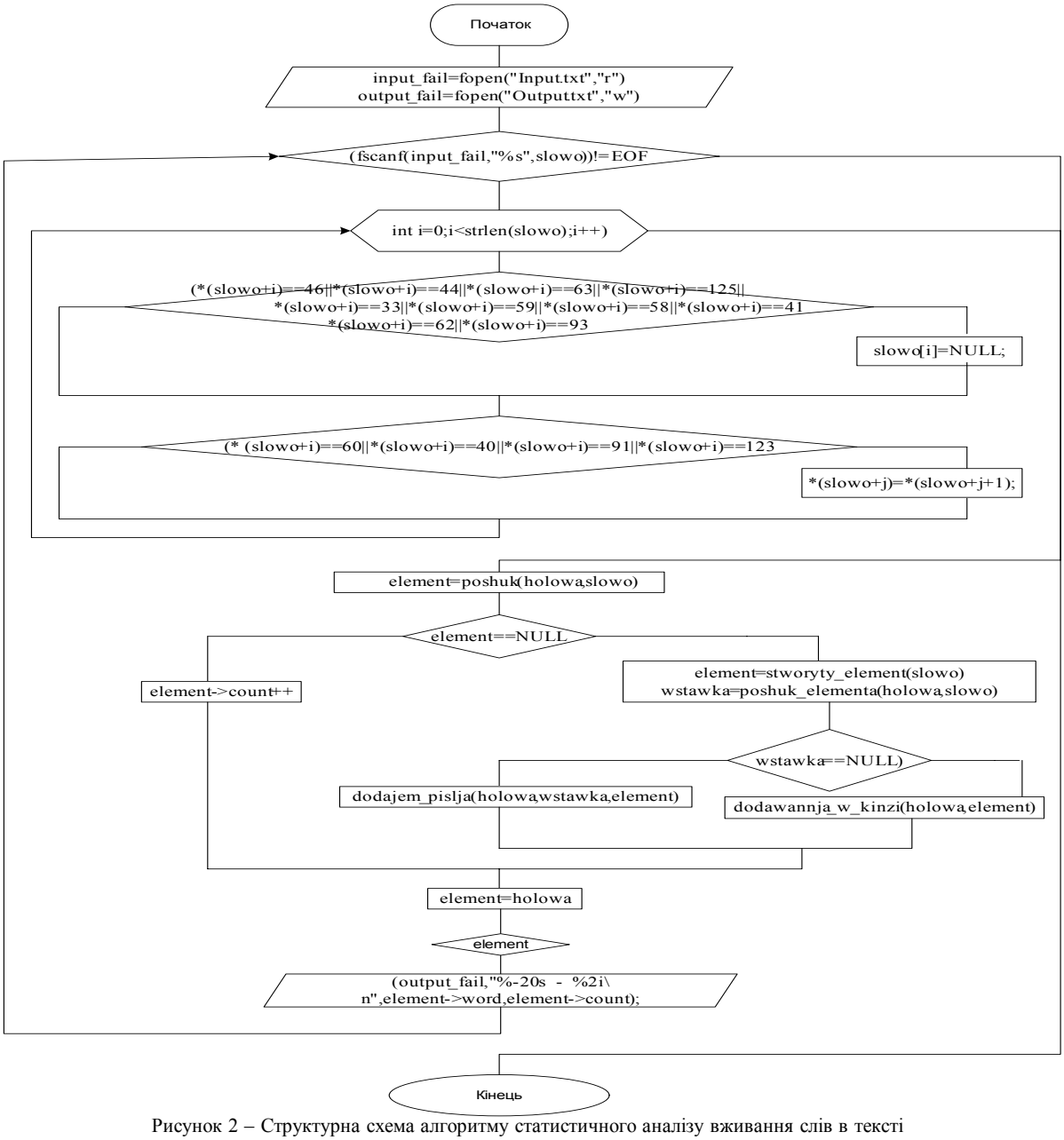


Рисунок 2 – Структурна схема алгоритму статистичного аналізу вживання слів в тексті

Алгоритм 2. Визначення множини ключових слів

Етап 1. В *Input* зберегти текст, який необхідно дослідити.

Етап 2. Відформатувати вхідний текст (однакові апострофи, забрати зайві символи, які не входять в абетку, окрім службових як пробіл, апостроф). В *Format* зберегти текст, який відформатовано.

Етап 3. При необхідності редагувати тематичний словник *Thematic*.

Етап 4. Запустити процес визначення множини ключових слів.

Крок 1. Будуємо спочатку алфавітно-частотний словник (абсолютні частоти) – *Vocab*.

Крок 2. Будуємо потім з *Vocab* алфавітно-частотний словник (відносні частоти) слів, тобто список слів за алфавітом та їх відносні частоти відносно загального обсягу тексту.

Крок 3. Будуємо скорочений список слів, частоти яких відповідають умовам формування ключовиків $U_K = \{U_{K1}, U_{K2}, U_{K3}, U_{K4}\}$, тобто список потенційних ключовиків – *Filter*.

Крок 4. Звіряємо сформований скорочений список з *Filter* зі списком *Thematic* та відповідно формуємо новий список входжень потенційних ключовиків з *Filter* в *Thematic* – список ключовиків в *Resvoc*.

Цей алгоритм не враховує пошук ключовиків по основах, у зв'язку з цим – результати дослідження текстів на формування ключових слів були негативні, зокрема:

– відсутні взагалі ключові слова у вихідному файлі (знайдені слова не відповідали вимогам до ключових слів – не попадали в діапазон частоти вживання в тексті);

– в список ключових слів попали службові слова, дієприкметники, дієслова, як ніяк не можуть бути ключовими словами (некоректно прописана база правил заблокованих слів);

– були присутні декілька слів з одною основою, але з різними флексіями (некоректно прописана база правил визначення основ, наприклад, пошук – пошуковими, користувач – користувачам, рейтинг – високорейтингового, рейтингу, контент – контентного, інформація – інформаційний, або були присутні граматичні помилки).

Тому був розроблений інший алгоритм знаходження множини ключових слів з врахуванням основ тематичних слів (рис. 3), та розміщений у відкритому доступі за адресою http://victana.lviv.ua/index.php/kliuchovi-slova.

4 ЕКСПЕРИМЕНТИ

Лінгвістичною базою для експериментального дослідження обрано 100 наукових публікацій Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (http://science.lp.edu.ua/sisn), двох номерів 783 (http://science.lp.edu.ua/SISN/SISN-2014) та 805 (http://science.lp.edu.ua/sisn/vol-cur-805-2014-2). Аналіз статистики функціонування системи виявлення множини ключових слів із 100 наукових статей було проведено у два етапи, зокрема:

1. Проаналізувати всі статті із перевіркою загальних заблокованих слів та тематичного словника.

2. Проаналізувати всі статті із перевіркою уточнених заблокованих слів та уточненого тематичного словника (з більшою кількістю запуску системи формується множина невідомих слів (відсутніх і в тематичному словнику і в множині заблокованих).

Окрім того на кожному етапі перевірка відбувалась в два кроки для кожної статті: аналіз всієї статті (рис. 4а) та аналіз статті без початку (назва, автори, удк, анотації двома мовами, авторські ключові слова двома мовами, місце роботи авторів) і без списку літератури (рис. 4б) для того, щоб визначити похибки точності формування множини ключових слів.

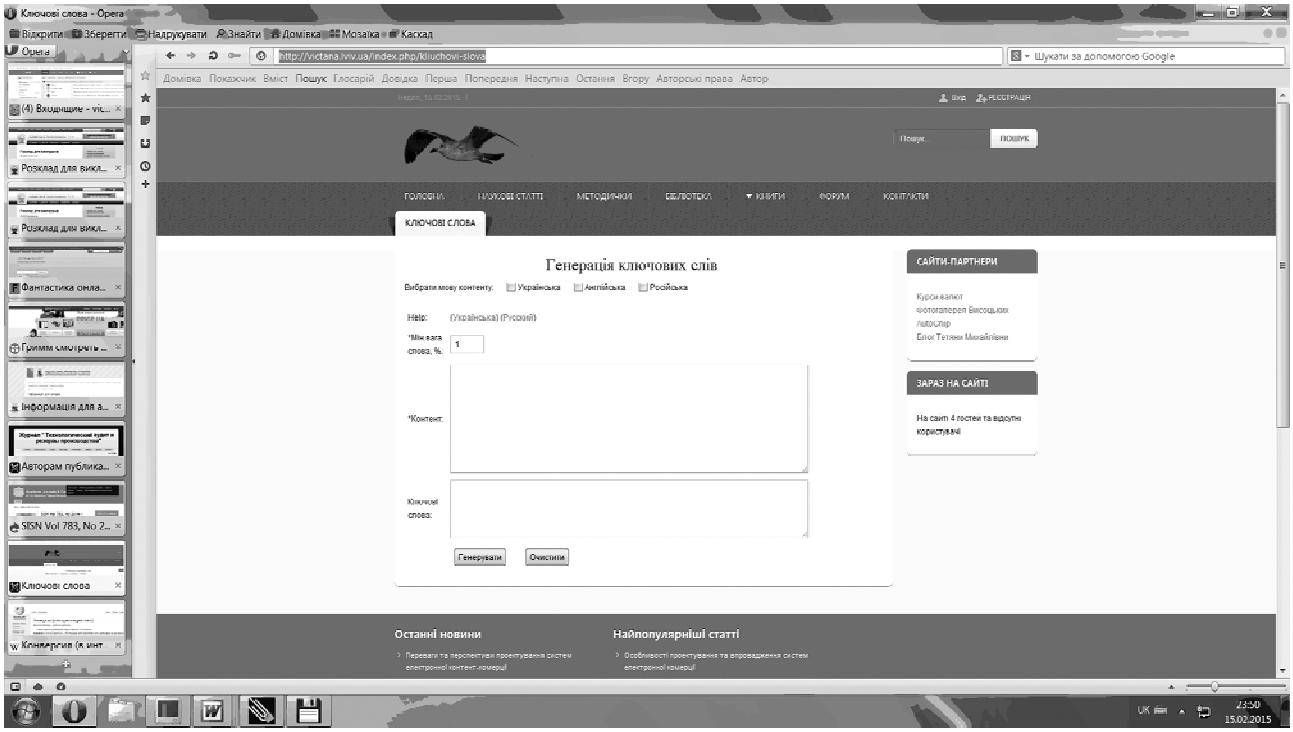
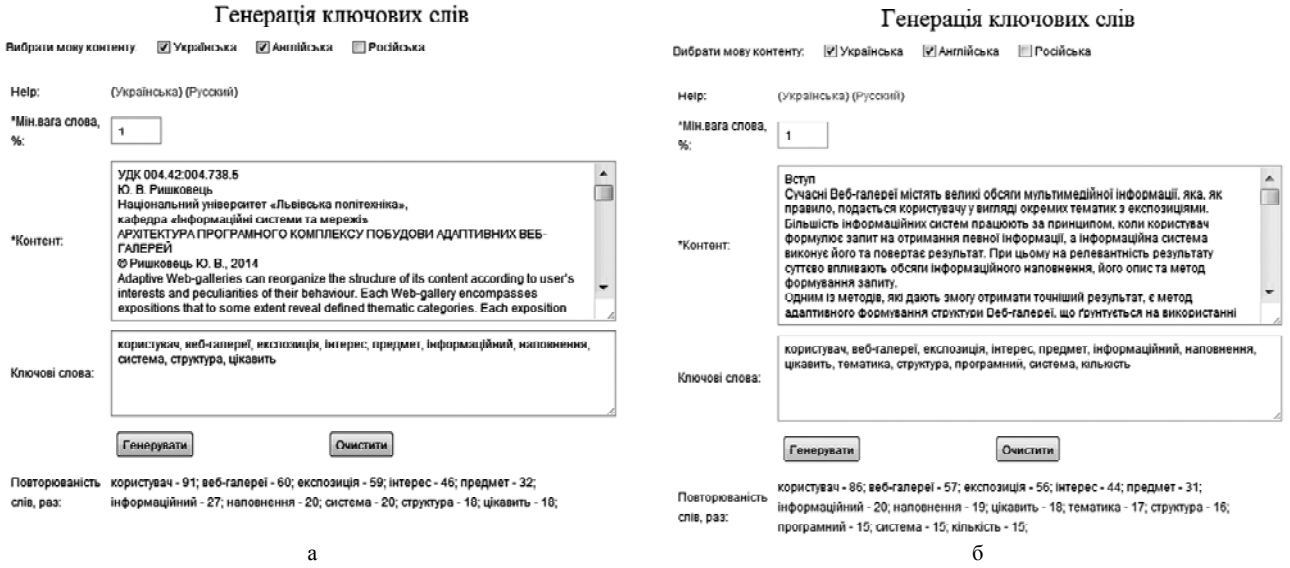


Рисунок 3 – Інформаційний ресурс визначення ключових слів з тексту



а б

Рисунок 4 – Результати перевірки статті: а – приклад аналізу всієї статті, б – приклад аналізу статті без початку і без списку літератури

5 РЕЗУЛЬТАТИ

Аналіз статистики здійснювався за принципом порівняння множини авторських ключових слів (визначені та прописані в статті самими авторами цих робіт), множини ключових слів визначених за першим та другим етапами з різними вагами слів (але більше, за визначене в опції **Мін.вага слова, %* в межах [1,5]) з повними та скороченими текстами робіт (табл. 3) при середньому арифметичному значенні авторських ключових словосполучень / слів біля 5 (4,77), які в середньому утворені з 10 (9,82) слів. Вага слова розраховується як відносна частота появи основи цього слова у всьому тексті. В табл. 4 присутні такі позначення, як *A* (всього ключових слів, визначених системою при заданій вазі слова), *B* (змістовних слів зі списку утворених, тобто без невідомих абревіатур, дієслів, службових слів тощо), *C* (збіг слів з визначеними автором статті), *D* (точність збігу знайдених ключовиків з авторським ключовими словами), *E* (додаткові ключові слова, визначені системою, але не визначені автором статті).

Таблиця 3 – Статистичні дані досліджених обсягів текстів статей

Назва обсягу статті	Крок 1		Крок 2	
	Всього	Середнє арифметичне	Всього	Середнє арифметичне
Сторінок	956	9,56	828	8,28
Абзаців	16497	164,97	15263	152,63
Рядків	42553	425,53	36965	369,65
Слів	345580	3455,8	291247	2912,47
Знаків	2327209	23272,09	1974773	19747,73
Знаків та пробілів	2674889	26748,89	2265917	22659,17

6 ОБГОВОРЕННЯ

На рис. 5 наведена порівняльна діаграма відсотків вживання знайдених системою ключових слів в відфільтрованому тексті (без початку (назва, автори, удк, анотації двома мовами, авторські ключові слова двома мовами, місце роботи авторів) і без списку літератури) Per_f та первинному авторському тексті Per_0 без уточнення модератором тематичного словника через поповнення заблокованих слів.

Отримані середні значення для 100 текстів $\overline{Per_f} = 0{,}28$ та $\overline{Per_0} = 0{,}19$ показують, що така фільтрація наукових статей покращує щільність ключовиків у 1,48 раз або на 47,83 відсотка. На рис. 6 наведена порівняльна діаграма відсотків вживання знайдених системою ключових слів в відфільтрованому тексті (без початку (назва, автори, удк, анотації двома мовами, авторські ключові слова двома мовами, місце роботи авторів) і без списку

Таблиця 4 – Статистичні дані досліджених змісту текстів статей

Назва	Вага слова	Етап 1					Етап 2				
		A	*B*	*C*	*D*	*E*	*A*	*B*	*C*	*D*	*E*
Крок 1	≥ 1	5,46	3,92	2,51	2,08	1,74	7,43	7,03	3,27	3	4,18
	≥ 2	1,08	0,88	0,63	0,59	0,26	2,67	2,64	1,65	1,54	1,12
	≥ 3	0,41	0,38	0,22	0,21	0,16	1,21	1,2	0,85	0,79	0,41
	≥ 4	0,15	0,13	0,09	0,09	0,04	0,46	0,45	0,33	0,31	0,15
	≥ 5	0	0	0	0	0	0	0	0	0	0
Крок 2	≥ 1	6,51	5,02	2,68	2,23	2,37	8,35	7,78	3,25	2,91	4,99
	≥ 2	1,34	1,11	0,74	0,72	0,39	3,12	3,07	1,81	1,67	1,43
	≥ 3	0,51	0,45	0,29	0,27	0,17	1,42	1,4	0,93	0,85	0,54
	≥ 4	0,19	0,17	0,12	0,12	0,05	0,73	0,72	0,45	0,42	0,31
	≥ 5	0,11	0,1	0,06	0,06	0,04	0,33	0,32	0,25	0,23	0,1

літератури) Per_f^v та первинному авторському тексті Per_0^v з врахуванням уточнення модератором тематичного словника через поповнення заблокованих слів. Отримані середні значення для 100 текстів $\overline{Per_f^v} = 0{,}34$ та $\overline{Per_0^v} = 0{,}25$ показують, що фільтрація з одночасною модерацією тематичного словника покращує щільність ключовиків у 1,35 раз або на 35,44%.

На рис. 7 наведена порівняльна діаграма відсотків вживання знайдених системою ключових слів в початковому первинному авторському тексті без уточнення модератором тематичного словника через поповнення заблокованих слів (Per_0) та з врахуванням уточнення модератором тематичного словника через поповнення заблокованих слів (Per_0^v).

Порівняння значень $\overline{Per_0} = 0{,}19$ та $\overline{Per_0^v} = 0{,}25$ демонструє ефективність модерації тематичного словника у початковому тексті – щільність ключовиків збільшується у 1,34 раз або на 34,33 відсотка. На рис. 8 наведена порівняльна діаграма відсотків вживання знайдених системою ключових слів в відфільтрованому авторському тексті без уточнення модератором тематичного словника через поповнення заблокованих слів (Per_f)

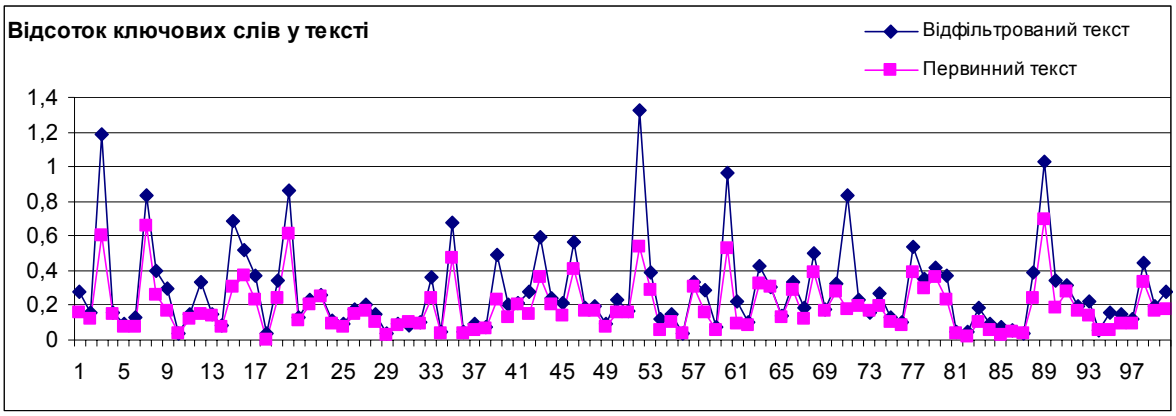


Рисунок 5 – Результати перевірки статей без уточнення модератором тематичного словника

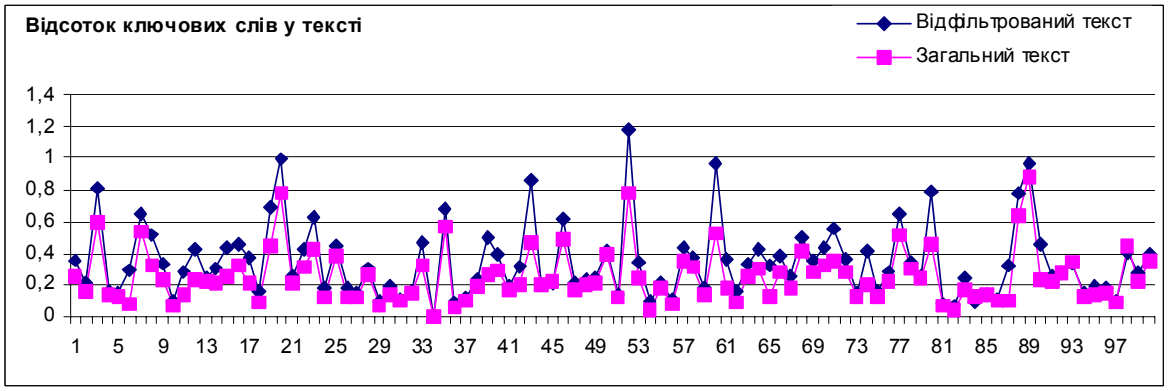


Рисунок 6 – Результати перевірки статей з врахуванням уточнення модератором тематичного словника

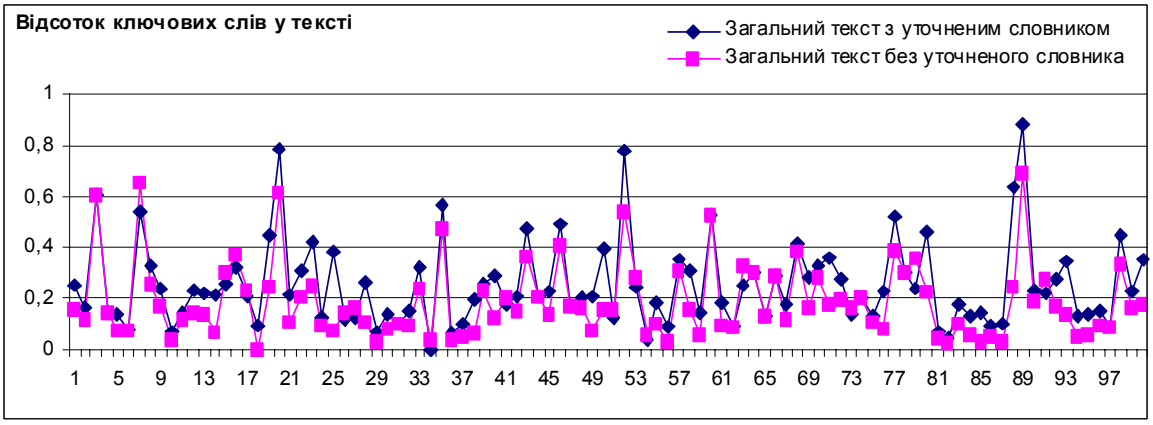


Рисунок 7 – Результати перевірки первинних авторських статей з різними словниками

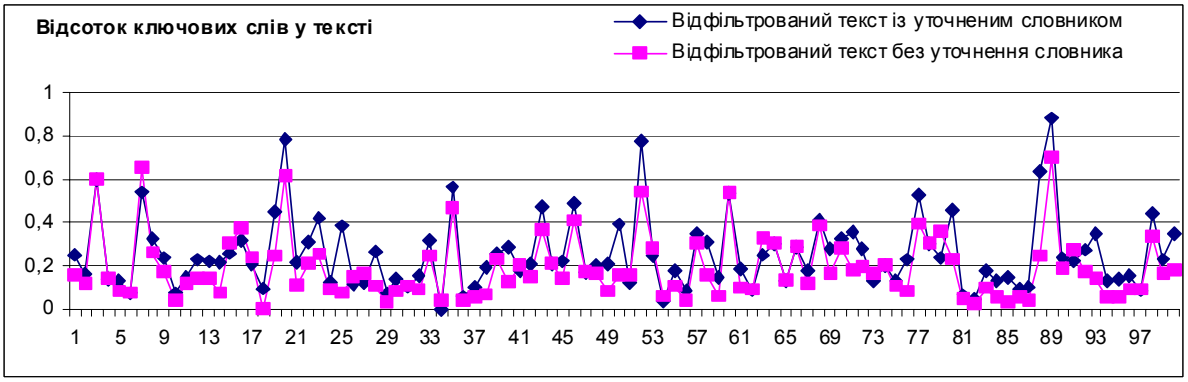


Рисунок 8 – Результати перевірки відфільтрованих статей з різними словниками

та з врахуванням модерації тематичного словника (Per_f^v).Порівняння значень $\overline{Per_f} = 0{,}28$ та $\overline{Per_f^v} = 0{,}34$ демонструє ефективність модерації тематичного словника у відфільтрованому тексті – щільність ключовиків збільшується у 1,23 раз або на 23,14 відсотка.

ВИСНОВКИ

У статті наведено теоретичне та експериментальне обґрунтування методу контент-моніторінгу україномовного тексту на основі стемінгу Портера. Метод спрямовано на автоматичне виявлення значущих ключових слів україномовного тексту за рахунок запропонованого формального підходу до реалізації стемінгу україномовного контенту. Проведено декомпозицію методу контент-моніторінгу на взаємопов'язані складові контент-аналізу текстової інформації та визначення множини ключових слів. Розроблено алгоритмічне забезпечення основних структурних складових запропонованого методу, а основу якого покладено адаптований до української мови алгоритм (стеммер) Портера. Теоретично виявлено способи покращення показників ефективності пошуку ключових слів, зокрема щільності ключовиків у тексті. Експериментальне дослідження 100 наукових публікацій з двох номерів (783 та 805) Вісника Національного університету «Львівська політехніка» серії «Інформаційні системи та мережі» (http://science.lp.edu.ua/sisn) продемонструвало позитивний вплив фільтрації тексту статті та модерації тематичного словника на визначення ключових слів. Виявлено, що для технічних наукових текстів експериментальної бази найкращих результатів досягає метод аналізу статті без початку (назва, автори, удк, анотації двома мовами, авторські ключові слова двома мовами, місце роботи авторів) і без списку літератури із перевіркою уточнених заблокованих слів та уточненого тематичного словника – для нього середнє значення щільності ключовиків у тексті досягає $\overline{Per_f^v} = 0{,}34$, що на 81% більше за аналогічне значення щільності первинного тексту $\overline{Per_0} = 0{,}19$. Потребує подальшого експериментального дослідження визначення ключових слів для інших категорій текстів – наукових гуманітарного про-філю, художніх, публіцистичних тощо.

ПОДЯКИ

У статті розв'язана науково-практична задача автоматичного виявлення значущих ключових слів та рубрикації україномовного контенту в Інтернет-системах на основі попереднього опрацювання відповідної текстової інформації. Роботу виконано в рамках спільних наукових досліджень кафедри інформаційних систем та мереж Національного університету «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», а також кафедри автоматики та інформаційно-вимірювальної техніки Вінницького національного технічного університету у межах діяльності науково-дослідного центру прикладної та комп'ютерної лінгвістики. Результати досліджень здійснювались у рамках держбюджетних науково-дослідних робіт за темами «Розробка методів, алгоритмів і програмних засобів моделювання, проектування та оптимізації інтелектуальних інформаційних систем на основі Web-технологій «ВЕБ» та «Інтелектуальна інформаційна технологія образного аналізу тексту та синтезу інтегрованої бази знань природно-мовного контенту». Наукові дослідження провадилися також в рамках ініціативної тематики досліджень кафедри ІСМ Національного університету «Львівська політехніка» на тему «Розроблення інтелектуальних розподілених систем на основі онтологічного підходу з метою інтеграції інформаційних ресурсів».

СПИСОК ЛІТЕРАТУРИ

1. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л. : НУЛП, 2009. – 612 с.
2. Математична лінгвістика / [В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич]. – Л. : «Новий Світ-2000», 2012. – 359 с.
3. Найефективніші методи залучення потенційних клієнтів [Електронний ресурс] / Центр ресурсів якості трафіку оголошень, Google AdWords. – Режим доступу: http://www.google.com/intl/uk_ALL/ads/adtrafficquality/advertisers/best-practices-for-generating-leads.html. – Назва з титул. екрану.
4. Нечеткий поиск в тексте и словаре [Електронний ресурс]. – Режим доступу: http://habrahabr.ru/post/114997/. – Назва з титул. екрану.
5. Реализации алгоритмов. Расстояние Левенштейна [Електронний ресурс]. – Режим доступу: http://ru.wikibooks.org/wiki/Реализации_алгоритмов/Расстояние_Левенштейна. – Назва з титул. екрану.
6. Задача о расстоянии Дамерау-Левенштейна [Електронний ресурс]. – Режим доступу: http://neerc.ifmo.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B4%D0%B0%D1%87%D0%B0_%D0%BE_%D1%80%D0%B0%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%B8%D0%B8_%D0%94%D0%B0%D0%BC%D0%B5%D1%80%D0%B0%D1%83-%D0%9B%D0%B5%D0%B2%D0%B5%D0%BD%D1%88%D1%82%D0%B5%D0%B9%D0%BD%D0%B0. – Назва з титул. екрану.
7. Насонов Д. Функция Левенштейна [Електронний ресурс] / Д. Насонов. – Режим доступу: http://rain.ifmo.ru/cat/data/theory/unsorted/levenshtein-2006/article.pdf. – Назва з титул. екрану.

8. Левенштейн, который сравнивает строки [Електронний ресурс] / Веб-разработка. – Режим доступу: http://dayte2.com/levenshtein. – Назва з титул. екрану.
9. Вычисление расстояния Левенштейна между двумя строками [Електронний ресурс]. – Режим доступу: http://wm-help.net/lib/b/book/827961078/78. – Назва з титул. екрану.
10. Стеммер Потера [Електронний ресурс]. – Режим доступу: http://labs.abcvg.com/stemmer/index.php. – Назва з титул. екрану.
11. Moseichuk V. Porter stemming algorithm for Ukrainian languages [Electronic resource] / V. Moseichuk. – Access mode: http://www.marazm.org.ua/document/stemer_ua/. – Title from the screen.
12. Стемінг [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/Стемінг. – Назва з титул. екрану.
13. Russian stemming algorithm [Electronic resource]. – Access mode: http://snowball.tartarus.org/algorithms/russian/stemmer.html. – Title from the screen.
14. Porter stemmer – реализация алгоритма стеммера Портера для русского языка на чистом функциональном языке Clojure [Електронний ресурс]. – Режим доступу: https://github.com/allaud/porter-stemmer. – Назва з титул. екрану.
15. The Porter Stemming Algorithm – Porter's homepage. [Електронний ресурс]. – Режим доступу: http://tartarus.org/~martin/PorterStemmer/. – Назва з титул. екрану.
16. The Porter Stemming Algorithm – Project «Snowball» [Electronic resource]. – Access mode: http://snowball.tartarus.org/algorithms/porter/stemmer.html. – Title from the screen.
17. The English (Porter2) stemming algorithm – Project «Snowball» [Electronic resource]. – Access mode: http://snowball.tartarus.org/algorithms/english/stemmer.html. – Title from the screen.
18. Porter M. F. An algorithm for suffix stripping [Electronic resource] / M. F. Porter // Program. – 1980. – Т. 14, № 3. – С. 130–137. – Access mode: http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html. – Title from the screen.
19. Willett P. The Porter stemming algorithm: then and now [Electronic resource] / P. Willett // Program: Electronic Library and Information Systems. – 2006. – В. 3, Т. 40. – С. 219–223. – ISSN 0033-0337. – Access mode: http://eprints.whiterose.ac.uk/1434/. – Title from the screen.
20. Сеник М. Вільний алгоритм стемінгу для української мови [Електронний ресурс] / М. Сеник. – Режим доступу: http://www.senyk.poltava.ua/projects/ukr_stemming/stemming_about.html. – Назва з титул. екрану.
21. Сеник М. Інструмент для пошуку слів з однаковими закінченнями [Електронний ресурс] / М. Сеник. – Режим доступу: http://www.senyk.poltava.ua/projects/ukr_stemming/word_by_ending.html. – Назва з титул. екрану.
22. Сеник М. Статичне дерево закінчень [Електронний ресурс] / М. Сеник. – Режим доступу: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html#dyn. – Назва з титул. екрану.
23. Сеник М. Демо стемінгу для української мови [Електронний ресурс] / М. Сеник. – Режим доступу: http://www.senyk.poltava.ua/projects/ukr_stemming/demo.html. – Назва з титул. екрану.
24. Вероятностный морфологический анализатор русского и украинского языков [Електронний ресурс]. – Режим доступу: http://www.keva.ru/stemka/stemka.html. – Назва з титул. екрану.
25. Стемминг [Електронний ресурс]. – Режим доступу: https://ru.wikipedia.org/wiki/Стемминг. – Назва з титул. екрану.
26. Lovins J. B. Development of a stemming algorithm / J. B. Lovins // Mechanical Translation and Computational Linguistics 11:22–31. – 1968.
27. Jongejan, B. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike [Electronic resource] / B. Jongejan, H. Dalianis // In the Proceeding of the ACL-2009, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, August 2–7, 2009, pp. 145–153. – Access mode: http://www.aclweb.org/anthology/P/P09/P09-1017.pdf. – Title from the screen.
28. Вірогідний морфологічний аналізатор російської та української [Електронний ресурс]. – Режим доступу: http://www.keva.ru/stemka/stemka.html. – Назва з титул. екрану.
29. Модуль Drupal для стемінга українською. Новий модуль для алгоритму Стема для Українського пошуку з виділенням коренів [Електронний ресурс]. – Режим доступу: http://drupal.ua/node/1170. – Назва з титул. екрану.
30. Стемінг Портера для української мови [Електронний ресурс]. – Режим доступу: http://www.marazm.org.ua/document/stemer_ua/. – Назва з титул. екрану.
31. Hardcoded stemmer for Ukrainian [Electronic resource]. – Access mode: https://github.com/vgrichina/ukrainian-stemmer. – Title from the screen.
32. Perestoronin P. Стеммер Портера для русского языка [Електронний ресурс] / P. Perestoronin. – Режим доступу: http://blog.eigene.in/post/49598738049/snowball. – Назва з титул. екрану.

Стаття надійшла до редакції 23.12.2015.
Після доробки 04.01.2016.

Бисикало О. В.[1], Высоцкая В. А.[2]
[1]Д-р техн. наук, профессор, декан факультета компьютерных систем и автоматики Винницкого национального технического университета, Винница, Украина
[2]Канд. техн. наук, доцент кафедры «Информационные системы и сети» Национального университета «Львовская политехника», Львов, Украина

ВЫЯВЛЕНИЕ КЛЮЧЕВЫХ СЛОВ НА ОСНОВЕ МЕТОДА КОНТЕНТ-МОНИТОРИНГА УКРАИНОЯЗЫЧНЫХ ТЕКСТОВ

Решена задача разработки алгоритмического обеспечения процессов контент-мониторинга для решения задачи определения ключевых слов русскоязычного текста. Рассмотрено формальное обоснование метода контент-мониторинга текста с помощью Стеммер Портера, в основу модификации стемминг положены известны результаты классификации морфемнои и словообразовательной структуры дериватов украинского языка, выявление закономерностей комбинаторики аффиксов, моделирование структурной организации глаголов и суфиксальних существительных, а также морфонологичных модификаций в процессе словоизменения глагола и словоизменении и словообразовании прилагательных украинского языка. Проведения декомпозиции метода и разработано алгоритмическое обеспечение его основных структурных составляющих по результатам контент-анализа текста. Теоретически обнаружены способы улучшения показателей эффективности поиска ключевых слов, в том числе плотности ключевиков в тексте. На основе разработанного программного обеспечения получены результаты экспериментальной апробации предложенного метода контент-мониторинга для определения ключевых слов в научных текстах технического профиля. Выявлено, что для выбранной экспериментальной базы из 100 работ лучших результатов по критерию плотности достигает метод анализа статьи без начальной обязательной информации и без списка литературы, но с проверкой уточненных заблокированных слов и уточненного тематического словаря.

Ключевые слова: текст, украиноязычный, алгоритм, контент-мониторинг, ключевые слова, контент-анализ, Стеммер Портера, лингвистический анализ, синтаксический анализ.

Bisikalo O. V.[1], Vysotska V. A.[2]
[1]F.D., professor, Dean of Faculty for Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, Ukraine
[2]Phd, associate professor of Information Systems and Networks Department, Lviv Polytechnic National University, Lviv, Ukraine

IDENTIFYING KEYWORDS ON THE BASIS OF CONTENT MONITORING METHOD IN UKRAINIAN TEXTS

The task of developing algorithmic providing processes of content monitoring for the problem solution of determining a keyword in Ukrainian text is solved. The formal justification of content monitoring in text using Porter stemmer is considered. The basis of the stemming

modification is the known results of morpheme and word building structure derivatives classification in Ukrainian language, affix combinatorics patterns identification, modeling the structural organization of verbs and suffixal nouns and morphonological modifications in the verb inflection and word formation and inflection of adjectives in Ukrainian language. The method decomposition is conducted and the algorithmic software of its basic structural components of the text content analysis results is developed. Theoretically means to improve the performance indicators of keywords search are identified, including keyword density in text. Based on the software obtained results of experimental testing of the proposed method of content monitoring to keywords identification in scientific texts of technical profile are developed. It is detected that the chosen experimental base of 100 works the article analysis method the without the initial required information and without the reference list reaches the best results for the density criterion, but with the specified blocked words and qualifying thematic dictionary verification.

Keywords: text, a Ukrainian, algorithm, content monitoring, keywords, content analysis, Porter stemmer, linguistic analysis, parsing.

REFERENCES

1. Berko A., Vysotska V., Pasichnyk V. Systemy elektronnoyi kontent-komertsiyi. Leningrad, NULP, 2009, 612 p.
2. Vysotska V., Pasichnyk V., Scherbyna J., Shestakevych T. Matematychna linhvistyka. Leningrad, Novyy Svit-2000, 2012, 359 p.
3. Nayefektyvnishi metody zaluchennya potentsiynyh kliyentiv [Electronic resource]. Tsentr resursiv yakosti trafiku oholoshen, Google AdWords. Access mode: http://www.google.com/intl/uk_ALL/ads/adtrafficquality/advertisers/best-practices-for-generating-leads.html. Title from the screen.
4. Nechetkyy poysk v tekste y slovare [Electronic resource]. Access mode: http://habrahabr.ru/post/114997/. Title from the screen.
5. Realyzatsyy alhorytmov. Rasstoyanye Levenshteyna [Electronic resource]. Access mode: http://ru.wikibooks.org/wiki/Реализации_алгоритмов/Расстояние_Левенштейна. Title from the screen.
6. Zadacha o rasstoyanyy Damerau-Levenshteyna [Electronic resource]. Access mode: http://neerc.ifmo.ru/wiki/index.php?title=%D0%97%D0%B0%D0%B4%D0%B0%D1%87%D0%B0_%D0%BE_%D1%80%D0%B0%D1%81%D1%81%D1%82%D0%BE%D1%8F%D0%BD%D0%B8%D0%B8_%D0%94%D0%B0%D0%BC%D0%B5%D1%80%D0%B0%D0%B0%D1%83-%D0%9B%D0%B5%D0%B2%D0 %B5%D0%BD%D1%88%D1%82%D0%B5%D0%B9%D0%BD%D0%B0. Title from the screen.
7. Nasonov D. Funktsyya Levenshteyna [Electronic resource]. Access mode: http://rain.ifmo.ru/cat/data/theory/unsorted/levenshtein-2006/article.pdf. Title from the screen.
8. Levenshteyn, kotory sravnivaet stroki [Electronic resource]. Web development. Access mode: http://dayte2.com/levenshtein. – Title from the screen.
9. Vychislenie rasstoyaniya Levenshteyna mezhdu dvumya strokami [Electronic resource]. Access mode: http://wm-help.net/lib/b/book/827961078/78. Title from the screen.
10. Porter stemmer [Electronic resource]. Access mode: http://labs.abcvg.com/stemmer/index.php. Title from the screen.
11. Moseichuk V. Porter stemming algorithm for Ukrainian languages [Electronic resource]. Access mode: http://www.marazm.org.ua/document/stemer_ua/. Title from the screen.
12. Steming [Electronic resource]. Access mode: https://uk.wikipedia.org/wiki/Стемінг. Title from the screen.
13. Russian stemming algorithm [Electronic resource]. Access mode: http://snowball.tartarus.org/algorithms/russian/stemmer.html. Title from the screen.
14. Porter stemmer-realizatsiya algoritma stemmera Portera dlya russkogo yazyka na chistom funktsionalnom yazyke Clojure [Electronic resource]. Access mode: https://github.com/allaud/porter-stemmer. Title from the screen.
15. The Porter Stemming Algorithm-Porter's homepage [Electronic resource]. Access mode: http://tartarus.org/~martin/PorterStemmer/. Title from the screen.
16. The Porter Stemming Algorithm – Project «Snowball» [Electronic resource]. Access mode: http://snowball.tartarus.org/algorithms/porter/stemmer.html. – Title from the screen.
17. The English (Porter2) stemming algorithm – Project «Snowball» [Electronic resource]. Access mode: http://snowball.tartarus.org/algorithms/english/stemmer.html. Title from the screen.
18. Porter M. F. An algorithm for suffix stripping [Electronic resource], Program, 1980,Vol. 14, No. 3, pp. 130–137. Access mode: http://telemat.det.unifi.it/book/2001/wchange/download/stem_porter.html. Title from the screen.
19. Willett P. The Porter stemming algorithm: then and now [Electronic resource], Program: Electronic Library and Information Systems, 2006, B. 3, Vol. 40, pp. 219–223. ISSN 0033-0337. – Access mode: http://eprints.whiterose.ac.uk/1434. Title from the screen.
20. Senyk M. Vilnyy alhorytm steminhu dlya ukrayinskoyi movy [Electronic resource]. Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/stemming_about.html. Title from the screen.
21. Senyk M. Instrument dlya poshuku sliv z odnakovymy zakinchennyamy [Electronic resource], Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/word_by_ending.html. Title from the screen.
22. Senyk M. Statychne derevo zakinchen [Electronic resource]. Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/ukr_endings.html#dyn. Title from the screen.
23. Senyk M. Demo steminhu dlya ukrayinskoyi movy [Electronic resource]. Access mode: http://www.senyk.poltava.ua/projects/ukr_stemming/demo.html. Title from the screen.
24. Veroyatnostny morfologichesky analizator russkogo i ukrainskogo yazykov [Electronic resource]. Access mode: http://www.keva.ru/stemka/stemka.html. Title from the screen.
25. Steming [Electronic resource]. Access mode: https://ru.wikipedia.org/wiki/Стемминг. Title from the screen.
26. Lovins J. B. Development of a stemming algorithm, Mechanical Translation and Computational Linguistics, 11:22–31. – 1968.
27. Jongejan B., Dalianis H. Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike [Electronic resource], In the Proceeding of the ACL-2009, Joint conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Singapore, August 2–7, 2009, pp. 145–153. Access mode: http://www.aclweb.org/anthology/P/P09/P09-1017.pdf. Title from the screen.
28. Virohidnyy morfolohichnyy analizator rosiyskoyi ta ukrayinskoyi [Electronic resource]. – Access mode: http://www.keva.ru/stemka/stemka.html. – Title from the screen.
29. Modul Drupal dlya steminha ukrayinskoyu. Novyy modul dlya alhorytmu Stema dlya Ukrayinskoho poshuku z vydilennyam koreniv [Electronic resource]. – Access mode: http://drupal.ua/node/1170. – Title from the screen.
30. Steminh Portera dlya ukrayinskoyi movy [Electronic resource]. – Access mode: http://www.marazm.org.ua/document/stemer_ua/. – Title from the screen.
31. Hardcoded stemmer for Ukrainian [Electronic resource]. – Access mode: https://github.com/vgrichina/ukrainian-stemmer. – Title from the screen.
32. Perestoronin, P. Stemmer Portera dlya russkogo yazyka [Electronic resource]. – P. Perestoronin // Access mode: http://blog.eigene.in/post/49598738049/snowball. – Title from the screen.