КОМП'ЮТЕРНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

УДК 004.738.5

А. Ю. Берко[1], В. А. Висоцька[2], Л. В. Чирун[3]
Національний університет "Львівська політехніка",
[1]кафедра загальної екології та екоінформаційних систем,
[2]кафедра інформаційних систем та мереж,
[3]кафедра програмного забезпечення

ЛІНГВІСТИЧНИЙ АНАЛІЗ ТЕКСТОВОГО КОМЕРЦІЙНОГО КОНТЕНТУ

У цій роботі проаналізовано основні проблеми електронної контент-комерції та функціональних сервісів опрацювання комерційного контенту. Запропонована модель дає можливість створити засоби опрацювання інформаційних ресурсів в системах електронної контент-комерції (СЕКК) та реалізувати підсистеми формування, управління та супроводу комерційного контенту. Процес проектування та створення СЕКК за результатами Інтернет-маркетингу є ітеративним і містить у своєму складі низку етапів від аналізу, проектування, розроблення плану до створення прототипу і експериментальних випробувань, починаючи з формування специфікацій, верстання, створення шаблону контенту, формування контенту та його подальше розміщення згідно з структурою сайта. На початкових етапах до визначення функціональних вимог і початку процесу розроблення до процесу залучають кінцевих користувачів за допомогою листків опитування, альтернатив проектування і прототипів різного ступеня готовності. Без значних зусиль збирають цінну інформацію, одночасно викликаючи у користувачів відчуття безпосередньої участі в процесі проектування та завойовуючи їхню довіру. Проаналізовано способи та моделі послідовності опрацювання інформаційних ресурсів в системах електронної контент-комерції та виділено основні закономірності переходу від процесів формування комерційного контенту до його реалізації. Створено формальну модель систем електронної комерції, що дало змогу реалізувати етапи життєвого циклу комерційного контенту. Розроблено формальні моделі опрацювання інформаційних ресурсів у системах електронної контент-комерції, що дало змогу створити узагальнену типову архітектуру системи електронної контент-комерції. Запропоновано узагальнену типову архітектуру системи електронної контент-комерції, що дало змогу реалізувати процеси формування, управління та реалізації комерційного контенту.

Ключові слова: інформаційний ресурс, комерційний контент, контент-аналіз, контент-моніторинг, контентний пошук, система електронної контент-комерції.

In the given article the main problems of electronic content commerce and functional services of commercial content processing are analyzed. The proposed model gives an opportunity to create an instrument of information resources processing in electronic content commerce systems (ECCS) and to implement the subsystem of commercial content formation, management and support. The process of ECCS design and creation as an Internet marketing result is iterative. It contains in its structure a number of stages (from the analysis, design and development of a plan to a prototype construction and experimental tests). The latter process begins with the specifications and layout formation, content template creation, content formation and its subsequent publishing according to the site's structure. In the initial stages (before setting functional requirements and development process initiation) regular users are

involved into the process through poll letters, alternative design and prototyping of varying degrees of readiness. Thus, valuable information is collected without much effort, along with both evoking users' sense of direct involvement in the design process, as well as winning their trust. The paper analyzes sequence methods and models of information resources processing in electronic content-commerce systems. It also allocates the basic laws of the transition from commercial content formation to its implementation. The formal model of ECCS is created, which allows the implementation in phases of the commercial content lifecycle. The developed formal model of information resources processing in electronic content-commerce systems allows us to create a generalized typical architecture of ECCS. The generalized typical architecture of ECCS is proposed in the paper, which helps implement the processes of commercial content formation, management and realization.

Key words: information resources, commercial content, content analysis, content monitoring, content search, electronic content commerce system.

Вступ. Загальна постановка проблеми

Опрацювання інформаційних ресурсів у системах електронної контент-комерції (СЕКК) дозволяє отримувати оперативні і об'єктивні дані про функціонування системи та для оцінювання рівня конкуренції на сегменті фінансового ринку контенту; оцінювати рівень конкурентів та міри їх конкурентоспроможності на фінансовому ринку з розповсюдження контенту [4, 6, 13–15, 17]. Основні класи користувачів/персонажів інформаційного ресурсу (клієнти, керівники робочих груп і адміністратори) визначають дизайн інформаційного ресурсу і процес ухвалення рішень. СЕКК обов'язково містить Web-вітрину (інформаційний ресурс) з каталогом комерційного контенту (з можливістю пошуку) і необхідними елементами інтерфейсу для введення реєстраційних даних, формування замовлення, здійснення платежів через Інтернет, оформлення доставки (e-mail/on-line), одержання даних про компанію та on-line допомоги. Весь процес управління контентом фіксується в підсистемі супроводу контенту для формування статистики функціонування СЕКК та пропозицій у вигляді списку популярних тематик контенту для підсистеми формування контенту [1, 2].

Аналіз останніх досліджень та публікацій

Напрями використання *автоматичного опрацювання текстів* або АОТ (рис. 1) [20–22].

– Тест Тюрінга: система є інтелектуальною, якщо під час мовного спілкування з нею її не відрізняють від живої людини [20–25, 30–31, 39, 41–44].

– Інтернет-системи (природність мови) → інформаційний пошук [4, 6, 13–15, 17, 32, 46–48].

Рис. 1. Автоматичне опрацювання текстів

Сфери розвитку *інформаційного пошуку* (ІП) [20, 27]

– 1950-ті Information Retrieval (IR) [1].

– 1990-ті WWW (Google > 8 млрд. сторінок, Yandex > 600 млн. сторінок, 2,5 млн. сайтів).

Основні визначення ІП за [1, 3, 16, 20, 27] (рис. 2):

– Подання, зберігання, організація і доступ до інформаційної одиниці.

– Фокусування на інформаційних потребах користувача.

– Акцентування на пошуку інформації (не даних).

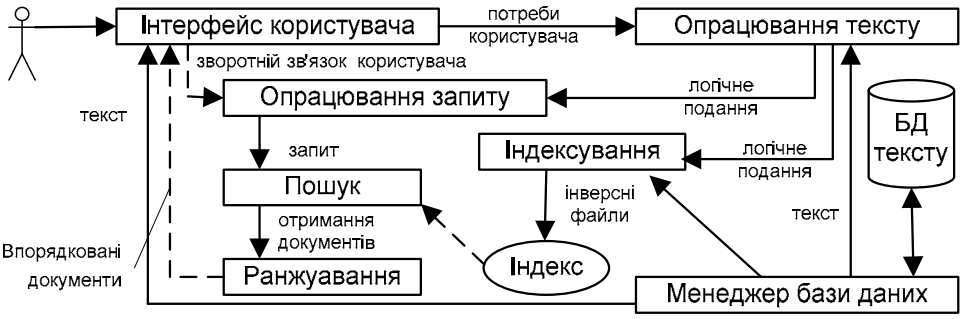


Рис. 2. Процес ІП

Прямий пошук [20]

– Brute force, середня складність якого O(n+m)

Т	А	Т	О		У		С	О	Н	І		Р	О	З	У	М	Н	И	Й

С	О	Н	І

С	О	Н	І

– Dboyer-Moore, середня складність якого O(n/m)

Т	А	Т	О		У		С	О	Н	І		Р	О	З	У	М	Н	И	Й

С	О	Н	І

С	О	Н	І

С	О	Н	І

С	О	Н	І

Індексування – це процес створення пошукового образу документа (логічне подання). Зазвичай – інвертований індекс за [1, 3, 4, 7–12].

$$Dictionary \begin{cases} Brutus & \Rightarrow 2 \rightarrow 4 \rightarrow 8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128 \\ Calpurnia & \Rightarrow 1 \rightarrow 2 \rightarrow 3 \rightarrow 5 \rightarrow 8 \rightarrow 13 \rightarrow 21 \\ Caesar & \Rightarrow \underbrace{13 \rightarrow 16}_{Postings} \end{cases}$$

Етапи попереднього АОТ [20].

– Видобування і/або отримання тексту (HTML, PDF…).
– Визначення кодування та мови.
– Розбиття на слова та речення (tokenization).
– Знищення стоп-слів.
– Лематизація (stemming) – приведення слова до словникової форми.

Tokenization, наприклад, [20]

1. Дати, числа: 13/03/2014, 1415…
2. Прислівники: нарешті, зазвичай, відтоді, потім, наприклад, ….
3. Вступні слова: іншими словами, в підсумок скажемо, між іншим…
4. Прийменники: напередодні, незважаючи на…
5. Частки: все ж таки, немов би, немов як, до того ж, ніби то як…
6. Багатослівні токени: Улан-Уде, Нью-Йорк, Іван Іванович… (collocations).
7. Межі речень: І. І. Іванов приїхав у м. Львів минулої зими.

Визначення стоп-слова [20].

– Текст = неструктурований набір значущих слів («bag of words»).
– Стоп-слова (stop-words) – службові частини мови – прийменники, сполучники, частки…

а, га, ай, ау, ах, ба, без, поблизу, брр, зась, ніби, б, бути, в, ви, ваш, поблизу, вглиб, до того ж, уздовж, адже, замість, замість, поза, усередині, як, біля, навколо, геть, ...

Модель ІП [20].

– Спосіб подання документів.
– Спосіб завдання інформаційних потреб (запитів).
– Спосіб розрахунку близькості між запитом та документом.

Булевська модель ІП [20]

– Документ = множина слів (термів)
– Запит = булевський вираз:

(кішка OR пес) AND корм
птаха ANDNOT військовий

– Опрацювання запиту = операції з множинами, які відповідають словам (термам)

Приклад булевської моделі за [10–12, 20].

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
Mercy	1	0	1	1	1	1
Worser	1	0	1	1	1	0

Особливості булевської моделі ІП [20].

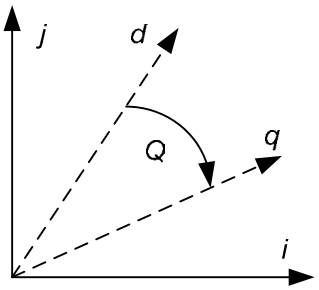
Переваги	Недоліки
Простота	Занадто «контрастно» (як подання документа, так і релевантність)
Зручно для тих, хто знайомий з логічними операторами	

Векторна модель ІП [10–12, 18, 20]

– Документ і запит – вектори в просторі слів (термів); компонент вектора – значимість слова для документа (запиту).
– Міра близькості – косинус кута між векторами (→ ранжування)

$$sim(\bar{d},\bar{q}) = \frac{\sum d_i \cdot q_i}{\left|\bar{d}\right|\left|\bar{q}\right|},$$

де d_i – вага терміну *i* в документі, q_i – вага терміну *i* в запиті [1, 16]:



Вага терміну визначається з врахуванням факторів [20, 26, 45, 49].

1. Як часто зустрічається в документі?
2. Як часто зустрічається в колекції?

Підхід $TF \cdot IDF$, де TF – term frequency, IDF – inverse document frequency, тобто $TF \cdot IDF$ – базовий варіант [1, 16].

$$tf_{ij} = \frac{f_{ij}}{\max\limits_{k} f_{kj}},\ idf_i = \log\frac{N}{n_i},\ w_{ij} = tf_{ij} \cdot idf_i$$

$TF \cdot IDF$, Okapi [20–22].

$$TFIDF_D(l) = \beta + (1-\beta) \cdot tf_D(l) \cdot idf_D(l).$$

$$tf_D(l)\frac{freq_D(l)}{freq_D(l) + 0.5 + 1.5 \cdot \frac{dl_D}{avg\_dl}},\ idf(l) = \frac{\log\left(\frac{|c| + 0.5}{df(l)}\right)}{\log(|c| + 1)}.$$

де avg_dl – середня довжина документа, c – розмір колекції, $\beta = 0...1$

Особливості векторної моделі

Переваги	Недоліки
Добре працює на «чистих» статичних колекціях	Легко атакується (спам)
Припускає часткові співпадіння	Погано працює на коротких текстах

Web [4, 6, 13-15, 17]

- Неконтрольована колекція
- Великі обсяги
- Різні формати
- Різноманітність (мова, теми …)
- Конкуренція (спам)
- Кліки
- Посилання! (PageRank)

Основа при оцінюванні якості пошуку – поняття релевантності (відповідність до інформаційної потреби), тобто точність (precision) $p = a/b$, повнота (recall) $r = a/c$ та F-міра $F = (p + r)/2pr$.
де a – релевантні у відповіді, b – всього у відповіді, c – всього релевантних [20, 27].

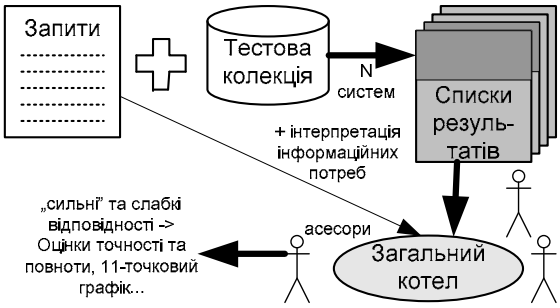


Рис. 3. Метод загального котла

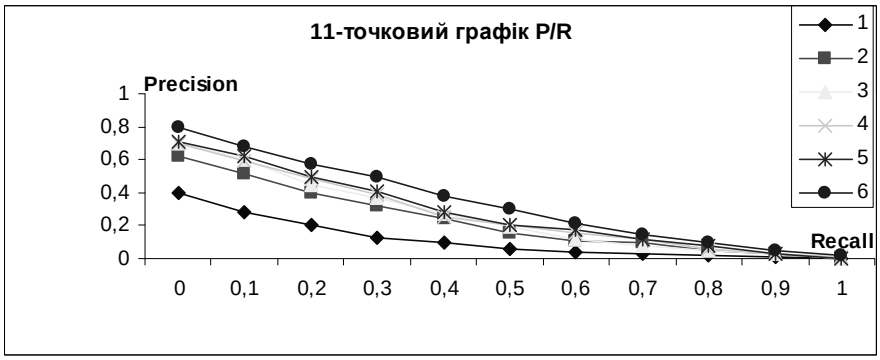


Рис. 4. 11-точковий графік P/R

Потреба в автоматичному морфологічному аналізі [20, 33, 34, 36–38, 50–53].

– Класи еквівалентності ключових слів при пошуку: кішка, кішки, кішку, кішкою, кішці…

– Наступне опрацювання (синтаксичний аналіз, семантичний аналіз…)

Типи аналізу [20]

– Стеммінг (англ. stemming) – виокремлення основи: *лісний, ліс, лісничий, ліса, лісистий → ліс* або *система, системний, систематизувати → систем.*

– Приведення до словникової форми: *лісного, лісному → лісний* або *ліса, лісів → ліс* або *танцюючий → танцювати.*

– POS-tagging (part-of-speech): *танцююче <V> в <PREP> повітрі <N> листя <N>*

– Повна морфологічна інформація: *танцююче <V, дієприкметник, недоконаний, теперішній час, однина, середній рід, називний відмінок> в <PREP> повітрі <N, середній рід, неістотний, однина, давальний відмінок> листя <N, середній рід, неістотний, однина, називний відмінок>*

Стеммінг є процесом скорочення слова до основи шляхом відкидання допоміжних частин як закінчення або/і суфікс . Результати стеммінгу подібні на знаходження кореня слова, але результат стеммінгу (стематизації) може відрізнятися від значення морфологічного кореня слова. Стеммінг застосовують в морфологічному аналізі та в ІП. Більшість пошукових систем використовують стеммінг для процесу злиття, тобто об'єднання слів у множини (синоніми), у яких збігаються форми після стематизації. Під час стеммінгу слова *активно, активний, активні* зведені до форми *активн*. А слова *голосно, голосувати, голосую, голосливий* взагалі до кореня слова *голос*. Активні дослідження в цьому напрямі провів Мартін Портер. Він розробив алгоритм, який набув значного поширення та став де-факто стандартним алгоритмом стеммінгу лише для *англійської мови* [20].

Пошук по таблиці, де зібрані всі можливі варіанти слів та їх форми після стеммінгу [20]. Перевагами є простота, швидкість та зручність опрацювання винятків з мовних правил. До недоліків зараховують те, що таблиця пошуку має містити всі форми слів: алгоритм не буде працювати з новими словами і розміри таблиці можуть бути великими. Для мов з відносно простою морфологією, наприклад, *англійської*, розміри таблиці пошуку малі, але у *турецькій* або *українській*, кількість варіантів слів з одним коренем є великою. Фрагмент таблиці пошуку на прикладі слова *гарний* типу *Слово→Стеммінг*, наприклад, {*гарна, гарне, гарний, гарним, гарними, гарних, гарні, гарній, гарнім, гарного, гарної, гарному, гарною, гарну*} → *гарн*.

Відсікання закінчень та суфіксів базуються на правилах для скорочування слова [20], наприклад, якщо слово закінчується на *льна*, то відсікаємо від слова *ьна*; якщо *льне* – відсікаємо *ьне*; якщо на *льний* – відсікаємо *ьний*; якщо *льним* – відсікаємо *ьним*. Кількість таких правил стеммінгу набагато менша за таблицю з усіма словоформами, а тому алгоритм є доволі компактним та продуктивним. Наведені правила вірно опрацьовують наступні прикметники типу *Слово→Стеммінг*, наприклад, *вільна→віл, мильне→мил, сильний→сил* та *суспільний→суспіл*. Алгоритм може виводити хибні результати. Наприклад, слово *пальне* буде на *пал* замість форми *пальн*. Враховуючи особливості мови, набір правил по відсіканню закінчень та суфіксів є складним, особливо для слов'янських мов. До недоліків належать опрацювання винятків, коли слова мають змінну форму. Наприклад, *криком* та *кричу* повинні мати після стеммінгу вигляд *крик*. Алгоритм має враховувати це, що призводить до ускладнення правил, і негативно впливає на ефективність.

Лематизація базується на визначенні основи слова через POS tagging (визначення частин мови у реченні) [20]. Далі до слова застосовують правила стеммінгу відповідно до частини мови. Тобто слова *пальне* (іменник) та *сильне* (прикметник) мають проходити через різні ланцюжки правил. Ці алгоритми мають високу якість і мінімальний відсоток помилок, якщо правильно та коректно описані правила розпізнавання частин мови.

Стохастичні алгоритми базуються на ймовірності визначення основи слова [20]. Вони мають здатність *навчатися*. База знань для цих алгоритмів – це набір логічних правил та таблиці пошуку. Після опрацювання слова стохастичним алгоритмом може з'явитися декілька варіантів основи слова, з яких алгоритм обере найімовірніший варіант. Наприклад, маємо лише одне логічне правило, за яким від слова відсікаємо останні літери. База знань наведена у множині типу *Слово→*

Стеммінг→ Закінчення, наприклад, {*популярність → популярн → ість*, *хвилини → хвилин → и*, *добрими → добр → ими*}. У значенні *Закінчення* наведений результат *навчання* алгоритму на базі знань. Для ілюстрації спробуємо виконати стеммінг слова *львів'яни* за типом *Слово → Закінчується на? → Результат → Числовий результат*, тобто {*львів'яни → ість → ні →* 0, *львів'яни→и→так→*1, *львівяни→ими→ні→*0}. Маємо один варіант, тому слово після стеммінгу – *львів'ян*. Але якщо передати цьому алгоритму слово *відомими*, то відповідь вже не однозначна при *Слово → Закінчується на? → Результат → Числовий результат* буде { *відомими → ість→ ні→* 0, *відомими → и→ так→* 1, *відомими → ими→ так→* 1}. Ускладнення правила дозволяє розв'язати такі протиріччя: віддають перевагу стеммінгу, який скорочує слово найбільше чи найменше.

Гібридний підхід стеммінгу використовує комбінацію вищенаведених алгоритмів. Наприклад, алгоритм використовує метод відсікання закінчень та суфіксів, але на першому етапі виконує пошук по таблиці. На відміну від пошуку по таблиці ця таблиця містить не всі словоформи, а тільки винятки з правил, які неправильно опрацьовуються алгоритмом, що відсікає закінчення [20].

Відсікання префіксів поряд із відсіканням від слова суфіксів та закінчень. Не можна позбавляти всі слова префіксів, наприклад, від слова *незалежний* утвориться *залежн*, що є протилежним змістом. Але є слова, у яких префікс не змінює значення слова, наприклад, *проголошую*, *наголошувати*, *виголошував* коректно скоротити до *голошу*.

Пошук відповідності використовують базу знань, що містить лише основи слів. Тобто ця база знань складається з тих слів, в які перетворюються звичайні слова після стеммінгу. Якщо порівнювати з пошуком по таблиці, то це слова з другого стовпця. Основна мета цих алгоритмів – через систему внутрішніх правил знайти для слова найвідповіднішу форму з бази знань. Одним з таких внутрішніх правил може бути довжина збігу слова та його основи. Наприклад, у базі знань є основи *чорн* та *чорняв*. При порівнянні зі словом *чорнява* у першому випадку спільна довжина 4 (*чорнява*), а у другому – 6 (*чорнява*), тому алгоритм обере довший варіант.

Стеммінг різними мовами. Якщо перші академічні роботи зі стеммінгу були присвячені лише англійській, то тепер існує доволі багато реалізацій стеммінгу для інших мов. Від особливостей мови залежить складність написання алгоритмів стеммінгу. Так, якщо стеммінг англійської – це доволі тривіальна задача, то стеммінг для арабської чи івриту – задача на порядок складніша. Варіанти стеммінгу для української мови існують і використовують у складі комерційних пошукових систем. Наразі відсутня вільна реалізація подібних алгоритмів.

Помилки стеммінгу. У алгоритмах стеммінгу поширені помилки двох типів.

- *Надстеммінг* (англ. overstemming) – це коли під час стематизації два слова скорочуються до однієї основи, хоча це не мало б статися.
- *Недостеммінг* (англ. understemming) – це протилежна помилка, коли слова отримують різні основи, хоча б мали мати одну спільну.

Алгоритми стеммінгу намагаються мінімізувати подібні помилки, проте скорочення помилок одного типу може призвести до зростання помилок іншого.

Виділення проблем

Інформаційний ресурс в СЕКК – множина даних з набором властивостей (табл. 1), які є об'єктом дій ІТ перетворення їх в контент [2, 5, 6, 19–22, 25–29, 32–40, 50–53].

Таблиця 1

Основні властивості інформаційних ресурсів у СЕКК

Назва	Властивість
Неоднорідність	Наявність складових різного походження, змісту і формату подання.
Узгодженість	Відсутність суперечливих або протилежних значень контенту.
Доступність формату	Доступність для всіх користувачів на основі стандартизованих методів, засобів та інтерфейсів.
Відкритість	Здатність до взаємодії, обміну значеннями та спільного використання з зовнішніми ресурсами.
Динамічність	Швидка актуалізація, відповідно до умов системи чи середовища.
Масштабованість	Можливість зміни логічного/фізичного обсягу контенту (величин/понять та їх позначень).
Контрольованість	Ідентифікація зміни/використання контенту та його впливу на процеси ІС.

Результат застосування однієї ІТ може бути інформаційним ресурсом іншої. Контент у галузі ІТ є формалізованими відомостями і знаннями, розміщеними у середовищі ІС і, на відміну від даних, без детальної специфікації їх властивостей, способів формалізації і впорядкування. Перетворення різних за природою, змістом та походженням даних в узгоджений централізований інформаційний ресурс є однією з важливих проблем побудови та функціонування СЕКК. Порядок формування і використання інформаційних ресурсів в СЕКК (рис. 5) визначають способи відбору даних із первинних джерел, їх фіксацію, фільтрування, перетворення до визначеного формату для формування контенту і розміщення в базі даних.

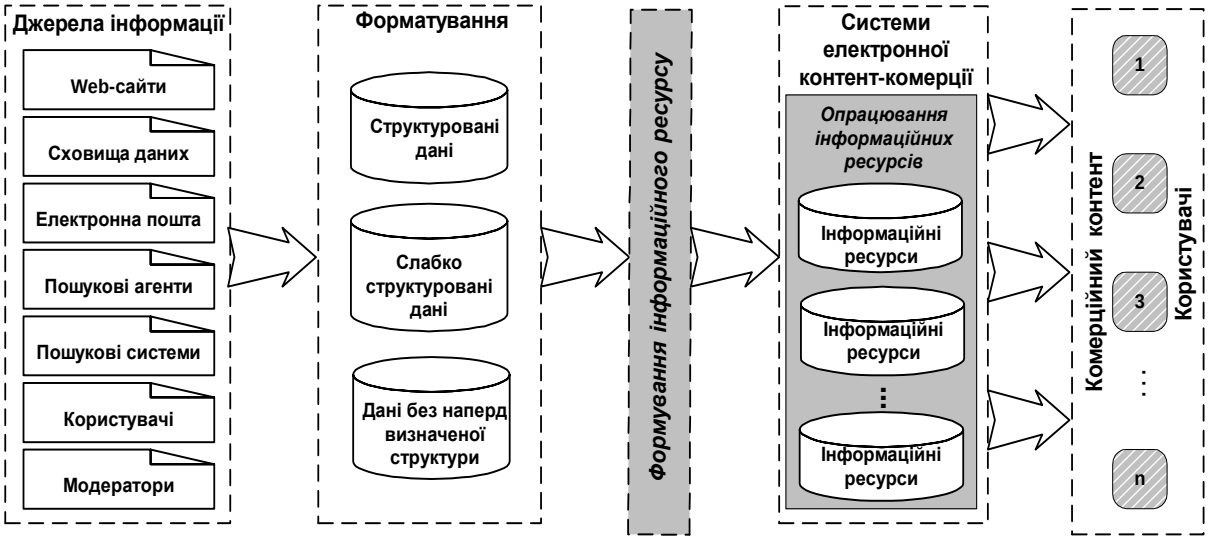


Рис. 5. Порядок формування і використання інформаційних ресурсів у СЕКК

Формулювання мети

Основною задачею проекту створення СЕКК є розроблення інформаційної архітектури інформаційного ресурсу шляхом формування актуального комерційного контенту, яку формують за зворотною реакцією користувачів відповідно до типу поширення комерційної діяльності (рис. 6).

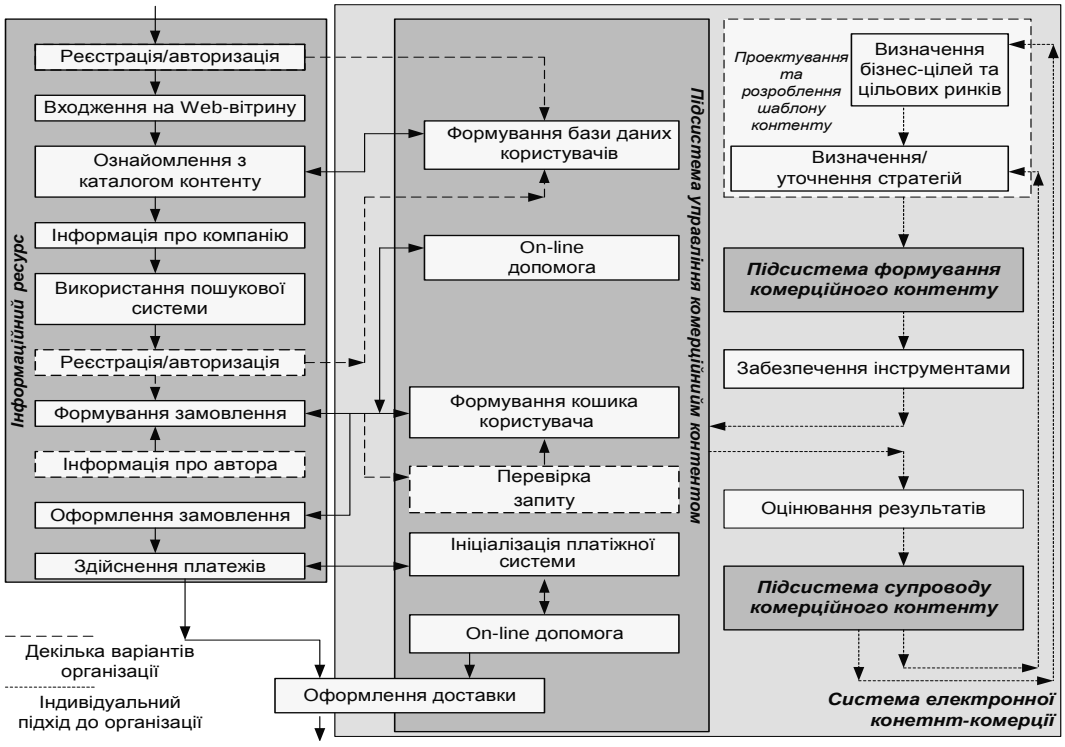


Рис. 6. Схема інформаційних потоків у системах електронної контент-комерції

Аналіз отриманих наукових результатів

Нехай існує деяка попередньо визначена множина n_X первинних джерел контенту з фіксованим або змінним складом. Кожне джерело інформації $Sourse(x_i)$, де x_i – i-й контент з джерела при $i=\overline{1,n_X}$, формує деяку множину значень, що містять відомості/знання/факти з предметної області СЕКК. Результатом звернення певних технологічних засобів СЕКК до джерела $Sourse(x_i)$ є генерування множини значень $X=\{x_1,x_2,...,x_{n_X}\}$, яка сприймається і подається у визначеній формі. В процесі відбору і фіксації генерованих значень, згідно з технологічними особливостями системи, згенеровану кожним джерелом інформації множину значень перетворюють на вхідний набір контенту $C=\alpha(u_f,x_i,t_p)$, визначеного формату – c_r, де $r=\overline{1,n_C}$.

Кожен набір контенту подають у вигляді структурованих, слабкоструктурованих даних або даних без визначеного опису структури та зберігається у БД комерційного контенту $DataBase(C)$. Структурування контенту передбачає формування для кожного набору опису його складу, способів поєднання елементів та їх впорядкування – множини умов $U=\{u_1,u_2,...,u_{n_U}\}$, де u_f – умова формування контенту при $f=\overline{1,n_U}$. Набір даних з джерела є поєднанням множини значень у заданому форматі і множини умов – $\langle X,U\rangle$, при формуванні вхідного набору контенту без опису структури $U=\varnothing$. Отриманий контент перед збереженням проходить верифікацію/валідацію для підтвердження його формальної/змістовної коректності/релевантності щодо вимог системи. У разі невідповідності зазначеним критеріям частина контенту вилучається з подальшого застосування. Відфільтрований контент форматують та зберігають, після чого відповідні відомості і знання $\langle C,H\rangle$ стають доступними користувачам через інформаційний ресурс СЕКК, тобто $Sourse(x_i)\rightarrow x_i\in X\rightarrow X\rightarrow\langle X,U\rangle\rightarrow\alpha(u_f,x_i,t_p)\rightarrow c_r\rightarrow C\rightarrow DataBase(C)\rightarrow\beta(q_d,c_r,h_k,t_p)\rightarrow\langle C,H\rangle$, де $i=\overline{1,n_X}$, де n_X – кількість джерел контенту; $Sourse(x_i)$ –джерело i-го контенту; $x_i\in X$ – i-й контент джерела $Sourse(x_i)$; $X=\{x_1,x_2,...,x_{n_X}\}$ – множина даних як результат відбору з джерела $Sourse(x_i)$; $\langle X_i,U_i\rangle$ – набір даних із множиною умов; $\alpha(u_f,x_i,t_p)$ – оператор формування контенту; c_r – сформований комерційний контент; C – множина сформованого контенту; $DataBase(C)$ – оператор збереження комерційного контенту в базі даних; $\beta(q_d,c_r,h_k,t_p)$ – оператор управління контентом; $\langle C,H\rangle$ – сформований з набору комерційного контенту та умов управління контентом інформаційного ресурсу в СЕКК.

Процес формування комерційного контенту подано такою схемою зв'язків:

$$Sourse(x_i)\rightarrow x_i\in X\rightarrow X\rightarrow\langle X,U\rangle\rightarrow\alpha_1(Downloading(\langle X,U\rangle),T)\rightarrow$$
$$\alpha_2(Verification(\langle X,U\rangle),T)\rightarrow\alpha_3(Conversion(\langle X,U\rangle),T)\rightarrow\alpha_4(\langle X,U\rangle,T)\rightarrow$$
$$\alpha_5(Qualification(\langle X,U\rangle),T)\rightarrow\alpha_6(\langle X,U\rangle,T)\rightarrow\alpha_7(\langle X,U\rangle,T)\rightarrow c_r\in C.$$

де $X=\{x_1,x_2\ldots,x_{n_X}\}$ – множина вхідних даних $x_i\in X$ з різних інформаційних ресурсів або від модераторів при $i=\overline{1,n_X}$; α_1 – оператор збирання контенту з різних джерел; α_2 – оператор виявлення дублювання змісту контенту; α_3 – оператор форматування контенту; α_4 – оператор виявлення ключових слів і понять контенту; α_5 – оператор автоматичної рубрикації контенту; α_6 – оператор формування дайджестів контенту; α_7 – оператор вибіркового поширення контенту; $T=\{t_1,t_2\ldots,t_{n_T}\}$ – час $t_p\in T$ транзакції формування комерційного контенту при $p=\overline{1,n_T}$; $C=\{c_1,c_2\ldots,c_{n_C}\}$ – множина комерційного контенту $c_r\in C$ при $r=\overline{1,n_C}$; $Verification(\langle X,U\rangle)$ – оператор верифікації контенту, $Qualification(\langle X,U\rangle)$ – оператор кваліфікації контенту, $Conversion(\langle X,U\rangle)$ – оператор перетворення контенту, $Downloading(\langle X,U\rangle)$ – оператор завантаження контенту.

Процес формування комерційного контенту для інформаційного ресурсу забезпечує зв'язок між множиною вхідних даних з різних джерел даних та множиною сформованого та збереженого комерційного контенту

$$S(x_i) \rightarrow x_i \rightarrow X \rightarrow \alpha(u_f, x_i, t_p) \rightarrow c_r \rightarrow C \rightarrow D(C), \quad (1)$$

де $S(x_i)$ – джерело даних, $D(C)$ – база даних комерційного контенту.

Типи джерел контенту для підсистеми формування контенту: список адрес інформаційних ресурсів з довірою та необхідними даними; список адрес інформаційних ресурсів з підпискою на контент; множина контенту від модераторів та авторів контенту; список запитів з ключовими словами для пошукових систем. Підсистема формування контенту забезпечує збирання даних з різних інформаційних ресурсів та їх форматування, виявлення ключових слів та дублювання, формування дайджесту, рубрикацію та вибіркове поширення контенту.

Основними етапами процесу формування комерційного контенту в СЕКК є форматування, рубрикація та поширення контенту, які мають таку схему зв'язків:

джерело контенту → збирання/створення контенту → база даних → форматування контенту → база даних → виявлення ключових слів та понять → база даних → рубрикація контенту → база даних → виявлення дублювання контенту → база даних → формування дайджесту контенту → база даних → вибіркове поширення контенту → модератор.

Формування комерційного контенту $\alpha: X \rightarrow C$ подано суперпозицією функцій

$$\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_0, \quad (2)$$

$$\alpha = \alpha_7 \circ \alpha_6 \circ \alpha_5 \circ \alpha_4 \circ \alpha_3 \circ \alpha_2 \circ \alpha_1, \quad (3)$$

де α_0 – оператор створення комерційного контенту; α_1 – оператор збирання контенту з різних джерел; α_2 – оператор виявлення дублювання комерційного контенту; α_3 – оператор форматування комерційного контенту; α_4 – оператор виявлення ключових слів і понять комерційного контенту; α_5 – оператор автоматичної рубрикації комерційного контенту; α_6 – оператор формування дайджестів комерційного контенту; α_7 – оператор вибіркового поширення комерційного контенту. Процес формування комерційного контенту подано як

$$\alpha = \langle X, T, U, C, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7 \rangle. \quad (4)$$

Процес формування комерційного контенту подано як

$$\alpha = \langle X, T, U, C, \alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7 \rangle. \quad (5)$$

1. Оператор створення комерційного контенту – відображення вхідних даних з різних джерел інформації у контент, який відрізняється від попереднього стану контенту актуальністю.

$$\alpha_0 : (X, U_C, T) \rightarrow C_0. \quad (6)$$

2. Оператор збирання комерційного контенту – відображення вхідних даних від авторів або модераторів системи у комерційний контент, який відрізняється від попереднього стану комерційного контенту достовірністю та актуальністю.

$$\alpha_1 : (X, U_G, T) \rightarrow C_0. \quad (7)$$

3. Оператор виявлення дублювання комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього стану відповідно унікальністю.

$$\alpha_2 : (C_0, T, U_B) \rightarrow C_1. \quad (8)$$

4. Оператор форматування комерційного контенту – відображення контенту в новий стан, який відрізняється від попереднього стану форматом подання.

$$\alpha_3 : (C_1, U_{FR}, T) \rightarrow C_2. \quad (9)$$

5. Оператор виявлення ключових слів комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього стану наявністю множини ключових слів, що загально описують його зміст.

$$\alpha_4 : (C_2, U_K, T) \rightarrow C_3. \quad (10)$$

6. Оператор рубрикації комерційного контенту – відображення комерційного контенту в новий стан через його валідацію, який відрізняється від попереднього стану приналежністю комерційного контенту до множини тематичного контенту.

$$\alpha_5 : (C_3, U_{CT}, T) \to C_4 . \tag{11}$$

7. Оператор формування дайджестів комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього стану появою нової частини контенту у вигляді його короткого змісту, що доповнює попередній стан.

$$\alpha_6 : (C_4, U_D, T) \to C_5 . \tag{12}$$

8. Оператор вибіркового поширення комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього стану призначенням комерційного контенту та поширенням серед цільової аудиторії.

$$\alpha_7 : (C_5, U_{Ds}, T) \to C_6 . \tag{13}$$

Множина операторів $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7\}$ є адекватною в процесі формування комерційного контенту. Процес формування комерційного контенту описано оператором вигляду

$$c_{r+1}(t_{p+1}) = \alpha(c_r, t_p, X, u_f), \tag{14}$$

де $u_f = \{u_{1f}, u_{2f}, \ldots, u_{n_{Uf}}\}$ – множина умов формування комерційного контенту c_r як

$$c_r = \left\{ \bigcup_i^{n_X} x_i \middle| \begin{array}{l} \forall x_i \in X_{u_f}, x_i \notin X_{\overline{u_f}}, \exists u_f \in U_{x_i}, u_f \notin U_{\overline{x_i}}, \\ X = X_{u_f} \cup X_{\overline{u_f}}, U = U_{x_i} \cup U_{\overline{x_i}}, f = \overline{1, n_U} \end{array} \right\}, \tag{15}$$

проходячи етапи перетворення даних у множину релевантного, форматованого, класифікованого та валідованого контенту як $x_i \in X \to \alpha_0(X, U_C, T)$ або $\alpha_1(X, U_G, T) \to \alpha_2(C_0, T, U_B) \to \alpha_3(C_1, U_{FR}, T) \to \alpha_4(C_2, U_K, T) \to \alpha_5(C_3, U_{Ct}, T) \to \alpha_6(C_4, U_D, T) \to \alpha_7(C_5, U_{Ds}, T) \to c_r \in C$.

Метод формування комерційного контенту – комплекс заходів забезпечення контролю опрацювання даних з різних джерел інформації для створення комерційного контенту з набором додаткових значень таких, як актуальність, достовірність, унікальність, повнота, точність тощо. Створення комерційного контенту описано оператором $C_0 = \alpha_0(X, U_C, T)$, U_C – множина умов створення комерційного контенту. Задачу збирання інформації з джерел описано оператором вигляду $C_0 = \alpha_1(X, U_G, T)$, де U_G – множина умов збирання даних з різноманітних джерел. Задачу виявлення дублювання змісту комерційного контенту описано оператором α_2 вигляду $C_1 = \alpha_2(\alpha_0(X, U_C, T), U_B)$ та $C_1 = \alpha_2(\alpha_1(X, U_G, T), U_B)$, або $C_1 = \alpha_2(C_0, U_B)$, U_B – множина умов виявлення дублювання змісту комерційного контенту. Виявлення дубльованого за змістом комерційного контенту в СЕКК виконують за допомогою лінгвостатистичних методів знаходження загальних термів, ланцюжки яких утворять словесні сигнатури комерційного контенту (текст є унікальним при коефіцієнті унікальності $\geq 80\,\%$). Задачу сканування комерційного контенту та приведення до формату в XML описано оператором α_3 вигляду

$$C_2 = \alpha_3(\alpha_2(C_0, U_B), U_{FR}), \tag{16}$$

де U_{FR} – множина умов форматування комерційного контенту.

Виявлення значущих ключових слів з множини контенту C_2, побудоване на принципі знаходження термів за змістом, базується на законі Зіпфа і зводиться до знаходження слів із середньою частотою появи (найуживаніші слова ігнорують через стоп-словник, а рідкісні слова не враховують). Виявлення ключових слів та понять з використанням словників визначається оператором $\alpha_4(C_2, U_K)$ вигляду

$$C_3 = \alpha_4(\alpha_3(\alpha_2(C_0, U_B), U_{FR}), U_K) \tag{17}$$

при $U_K = \{U_{K1}, U_{K2}, U_{K3}, U_{K4}\}$, де U_K – колекція умов виявлення ключових слів та понять у тексті, U_{K1} – множина всіх термів (термом є основа іменника, іменник, словосполученням іменників або

прикметника з іменником), U_{K2} – множина частот вживання терму в тексті комерційного контенту, U_{K3} – множина коефіцієнтів вживання термів з врахуванням кількості знаків без пробілів (при 2 000–3 000 знаків частота ключових слів у межах 4–6 %, до 2 000 знаків – 6–8 %, за 3 000 знаків – 2–4 %), U_{K4} – множина термів, які відповідають умовам належності до ключових слів.

Задачі класифікації та розподілу контенту реалізують через інформаційно-пошукову систему вибіркового поширення контенту. Комерційний контент аналізують на відповідність запитам з використанням результатів рубрикації. Оператор рубрикації комерційного контенту згідно з виявленими ключовими словами описано як $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ при $U_{CT} = \{U_{CT1}, U_{CT2}, U_{CT3}, U_{CT4}\}$, де U_{CT} – колекція умов рубрикації, U_{CT1} – множина тематичних ключових слів зі словника, U_{CT2} – множина частот вживання ключових слів у комерційному контенті, U_{CT3} – множина залежностей вживання ключових слів різних тематик (коефіцієнти визначає модератор згідно з належністю ключового слова до певної тематики в межах $[0,1]$), U_{CT4} – множина частот вживання тематичних ключових слів у контенті. Множину дайджестів C_5 формують залежністю $C_5 = \alpha_6(C_4, U_D)$, де U_D – множина умов формування дайджестів комерційного контенту, тобто

$$C_5 = \alpha_6(\alpha_5(\alpha_4(C_2, U_K), U_{CT}), U_D). \quad (18)$$

Релевантний контент розсилають користувачам та завантажують у бази даних. Вибіркове поширення контенту описано $C_6 = \alpha_7(C_5, U_{DS})$, де U_{DS} – множина умов поширення контенту.

Виявлення ключових слів тематики комерційного контенту. Текстовий контент C_2 (стаття, коментар, книга тощо) містить значний обсяг даних природною мовою, частина яких є абстрактною. Текст подають як об'єднану за змістом послідовність знакових одиниць, основними властивостями якої є інформаційна, структурна та комунікативна зв'язність/цілісність, що відображає змістовну/структурну сутність тексту. Методом опрацювання тексту є лінгвістичний аналіз змісту (наприклад, коментарі, форуми, статті тощо). Процес опрацювання тексту поділяє контент на лексеми за допомогою кінцевих автоматів (рис. 7).

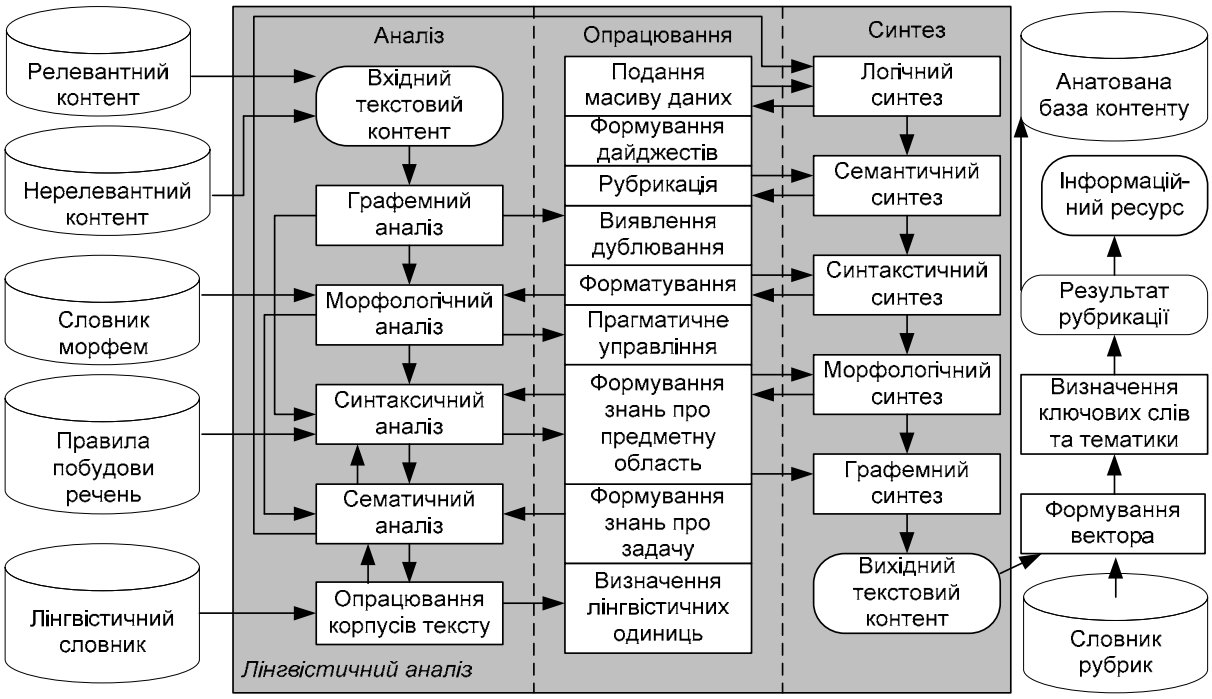


Рис. 7. Структурна схема лінгвістичного аналізу текстового контенту

Як функціонально-семантико-структурна єдність текст має правила побудови, виявляє закономірності змістовного та формального з'єднання складових одиниць. Зв'язність проявляється через зовнішні структурні показники та формальну залежність компонентів тексту, а цілісність – через тематичну, концептуальну та модальну залежність. Цілісність веде до змістовної та комунікативної організації тексту, а зв'язність – до форми, структурної організації. Оператор виявлення ключових слів комерційного контенту $\alpha_4 : (C_2, U_K, T) \to C_3$ є відображенням комерційного контенту C_2 в новий стан, який відрізняється від попереднього стану наявністю множини ключових слів, що загально описують його зміст. При аналізі досліджують багаторівневу структуру контенту: лінійну послідовність символів; лінійну послідовність морфологічних структур; лінійну послідовність речень; мережу взаємопов'язаних єдностей (алг. 1).

Алгоритм 1. Лінгвістичний аналіз текстового комерційного контенту.

Етап 1. Граматичний аналіз текстового контенту C_2.

Крок 1. Поділ текстового комерційного контенту C_2 на речення та абзаци.

Крок 2. Поділ ланцюжка символів контенту C_2 на слова.

Крок 3. Виділення цифр, чисел, дат, незмінних оборотів і скорочень контенту C_2.

Крок 4. Видалення нетекстових символів контенту C_2.

Крок 5. Формування та аналіз лінійної послідовності слів із службовими знаками для контенту C_2 (алг. 3).

Етап 2. Морфологічний аналіз текстового контенту C_2.

Крок 1. Отримання основ (словоформ із відрубаними закінченнями).

Крок 2. Для кожної словоформи формується граматична категорія (колекція граматичних значень: рід, відмінок, відмінювання тощо).

Крок 3. Формування лінійної послідовності морфологічних структур.

Етап 3. Синтаксичний аналіз $\alpha_4 : (C_2, U_K, T) \to C_3$ текстового контенту C_2 (алг. 2).

Етап 4. Семантичний аналіз текстового контенту C_3.

Крок 1. Слова співвідносяться з семантичними класами із словника.

Крок 2. Відбір потрібних для даного речення морфосемантичних альтернатив.

Крок 3. Зв'язування слів у єдину структуру.

Крок 4. Формування упорядкованої множини записів суперпозицій з базисних лексичних функцій і семантичних класів. Точність результату визначається повнотою/коректністю словника.

Етап 5. Референційний аналіз для формування міжфразових єдностей.

Крок 1. Контекстний аналіз текстового комерційного контенту C_3. За його допомогою реалізується дозвіл локальних референцій (цей, який, його) і виділення висловлювання – ядра єдності.

Крок 2. Тематичний аналіз. Поділ висловлювань на тему і рему виділяє тематичні структури, які використовують, наприклад, при формуванні дайджесту.

Крок 3. Визначають регулярну повторюваність, синонімізацію та повторну номінацію ключових слів; тотожність референції, тобто співвідношенням слів з предметом зображення; наявність імплікації, заснованої на ситуативних зв'язках.

Етап 6. Структурний аналіз текстового контенту C_3. Передумовами використання є високий ступінь збігу термінів єдності, дискурсивна одиниця, речення семантичною мовою, висловлювання і елементарна дискурсивна одиниця.

Крок 1. Виявлення базового набору риторичних зв'язків між єдностями контенту.

Крок 2. Побудова нелінійної мережі єдностей. Відкритість набору зв'язків припускає його розширення та адаптацію для аналізу структури текстів C_3.

Синтаксичні аналізатори працюють в два етапи: ідентифікують змістовні лексеми та створюють дерево розбору (алг. 2).

Алгоритм 2. Синтаксичний аналізатор комерційного контенту.

Етап 1. Ідентифікація змістовних лексем $U_{K1} \in U_K$ для комерційного контенту C_2.

Крок 1. Визначення ланцюжка термів у вигляді речення.

Крок 2. Ідентифікація іменної групи за допомогою словника основ.

Крок 3. Ідентифікація дієслівної групи за допомогою словника основ.

Етап 2. Створення дерева розбору зліва направо. Виведення дерева полягає в розгортанні одного з символів попереднього ланцюжка послідовності лінгвістичних змінних, або в заміні його іншим, інші ж символи переписуються без зміни. При розгортанні, замінювані/переписувані символи (*предки*) з'єднують безпосередньо з символами, які виходять в результаті розгортання, заміни або переписування (*нащадками*), та отримують дерево складових, або синтаксичну структуру для змісту комерційного контенту.

Крок 1. Розгортання іменної групи. Розгортання дієслівної групи.

Крок 2. Реалізація синтаксичних категорій словоформами.

Етап 3. Визначення множини ключових слів $\alpha_4 : (C_2, U_K, T) \to C_3$ для контенту C_2.

Крок 1. Визначення термів $Noun \in U_{K1}$ – іменників, словосполучень іменників або прикметника з іменником серед множини слів текстового контенту.

Крок 2. Розрахунок унікальності $Unicity$ для термів $Noun \in U_{K1}$.

Крок 3. Розрахунок $NumbSymb \in U_{K3}$ (кількість знаків без пробілів) для $Noun \in U_{K1}$ при $Unicity \geq 80$.

Крок 4. Розрахунок $UseFrequency \in U_{K2}$ – частоти появи ключових слів комерційного контенту. Для термів з $NumbSymb \leq 2000$ частота $UseFrequency$ є в межах $(6;8]$%, з $NumbSymb \geq 3000$ – $[2;4)$%, з $2000 > NumbSymb < 3000$ – $[4;6]$%.

Крок 5. Розрахунок $BUseFrequency$ – частота появи ключових слів на початку тексту, $IUseFrequency$ – частота появи ключових слів в середині тексту, $EUseFrequency$ – частота появи ключових слів в кінці тексту комерційного контенту.

Крок 6. Порівняння значень $BUseFrequency$, $IUseFrequency$ та $EUseFrequency$ для розстановки пріоритетів. Ключові слова з більшими значеннями $BUseFrequency$ мають більший пріоритет, ніж ключові слова з більшим значенням $EUseFrequency$.

Крок 7. Сортування ключових слів згідної їх пріоритетів.

Етап 4. Заповнення бази пошукових образів контенту C_3, тобто атрибутів $KeyWords \in U_{K4}$ – ключові слова, $Unicity$ – унікальність ключових слів ≥ 80, $Noun$ – терм, $NumbSymb$ – кількість знаків без пробілів, $UseFrequency$ – частота вживання ключових слів, $BUseFrequency$ – частота вживання ключових слів на початку тексту, $IUseFrequency$ – частота вживання ключових слів в середині тексту, $EUseFrequency$ – частота вживання ключових слів в кінці тексту.

Виявлення ключових слів тематики контенту C_2 з фрагменту тексту виконують за допомогою процесів, поданих на рис. 8.

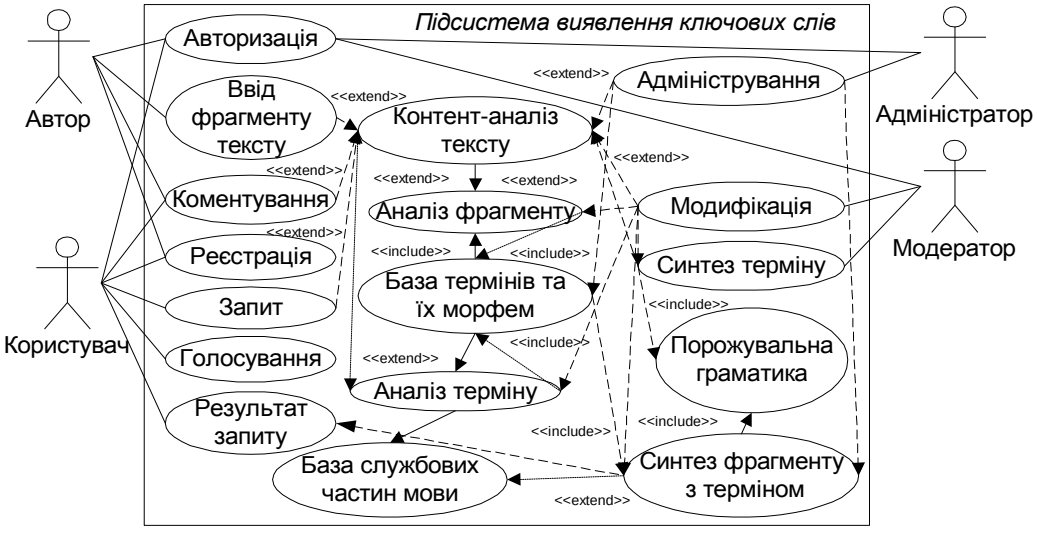


Рис. 8. Діаграма варіантів використання для процесу виявлення ключових слів

Текст реалізує структурно подану діяльність, що передбачає суб'єкт і об'єкт, процес, мету, засоби і результат, які відображаються в змістовно-структурних, функціональних, комунікативних показниках. Одиницями внутрішньої організації структури тексту є алфавіт, лексика (парадигматика), граматика (синтагматика), парадигми, парадигматичні відношення, синтагматичні відношення, правила ідентифікації, висловлювання, міжфразова єдність та фрагменти-блоки. На композиційному рівні виділяють речення, абзаци, параграфи, розділи, глави, підглави, сторінки тощо, які, крім речення, побічно пов'язані з внутрішньою структурою, тому не розглядаються (рис. 9). За допомогою БД (бази термінів/морфем і службових частин мови) та визначених правил аналізу тексту виконують пошук терміну (рис. 10).

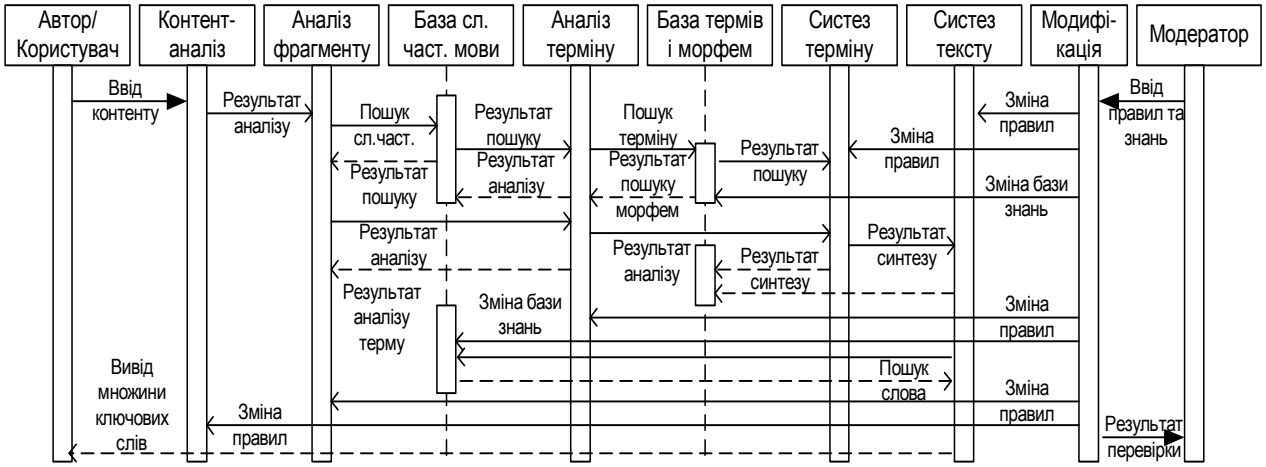
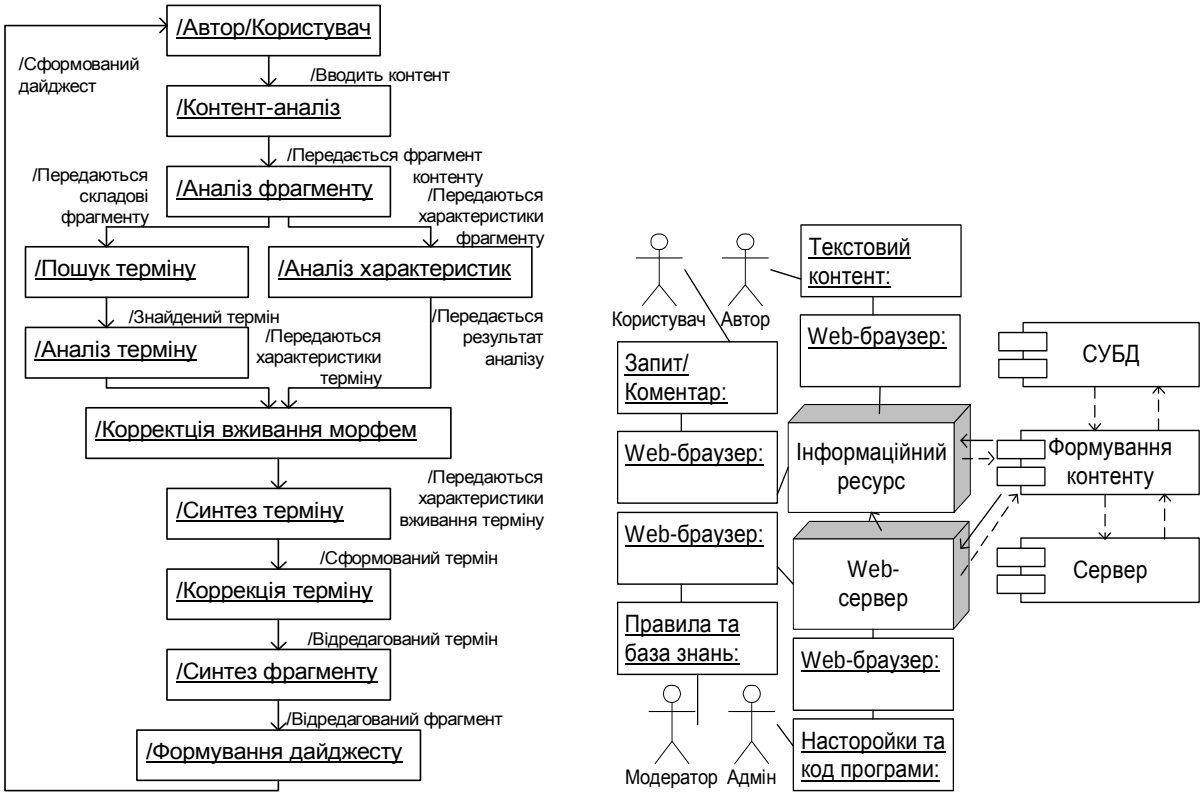


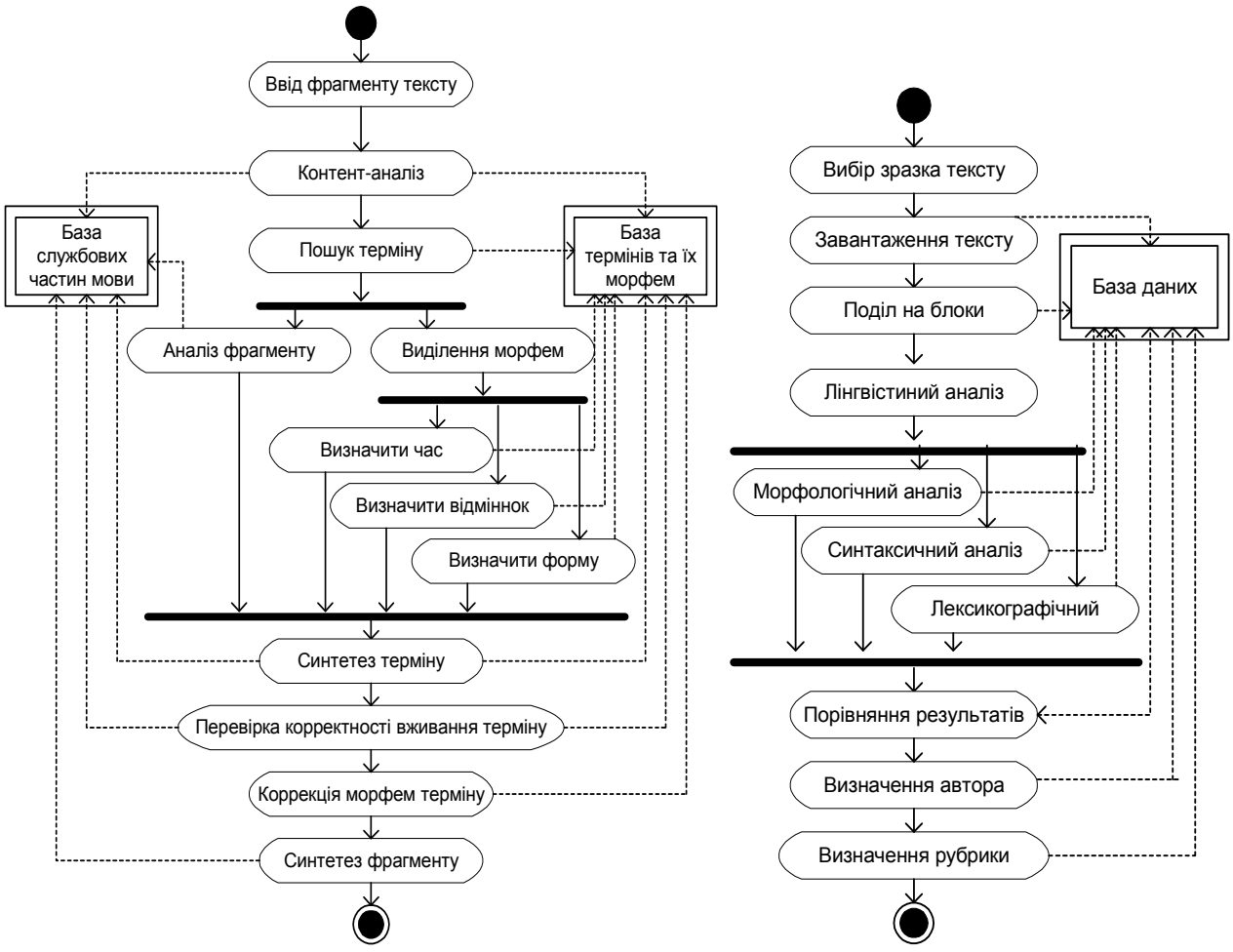
Рис. 9. Діаграма послідовності для процесу виявлення ключових слів контенту



а *б*

Рис. 10. Діаграма а – кооперації та б – компонентів процесу виявлення ключових слів контенту

Ґрунтуючись на правилах породжувальної граматики, виконують корекцію терміну згідно з правилами його вживання у контексті (рис. 11, *а*). Речення задають межі дії знаків пунктуації, анафоричних і катафоричних посилань. Семантика тексту зумовлена комунікативним завданням передавання даних. Структура тексту визначається внутрішньою організацією одиниць тексту і закономірностями їх взаємозв'язку. Через синтаксичний аналіз текст оформляють у структуру даних, наприклад, в дерево, яке відповідає синтаксичній структурі вхідної послідовності, і найкраще підходить для подальшого опрацювання. Після аналізу фрагменту тексту і терміну синтезують новий термін як ключове слово тематики контенту, використовуючи базу термінів та їх морфем (рис. 11, *а*). Далі синтезують терміни для формування нового ключового слова, використовуючи базу службових частин мови.



а *б*

Рис. 11. Діаграма діяльності для процесу:
а – виявлення ключових слів; б – рубрикації контенту

Принцип виявлення ключових слів за термами базується на законі Зіпфа і зводиться до вибору слів із середньою частотою появи (найуживаніші слова ігнорують через стоп-словники, а рідкісні слова не враховують). Модуль виявлення ключових слів реалізований на ресурсі Victana та доступний за адресою http://victana.lviv.ua/index.php/kliuchovi-slova (рис. 12).

Процес рубрикації контенту. Аналіз лексико-граматичної та семантико-прагматичної побудови тексту використовують для автоматичної рубрикації контенту та формування дайджестів, що призводить до формування тематично підібраних масивів контенту (табл. 2).

Таблиця 2

Основні етапи рубрикації контенту

Назва	Призначення етапу
Підготовка	Визначення тематики/мети/об`єкту аналізу, хронологічні та географічні рамки, принципи відбору.
Збір даних	Формування класифікатора відбору ключових цитат та інструкції для кодувальника.
Формальний аналіз	Перетворення фрагментів тексту без аналізу його змісту. Морфологічні дані забезпечують доступ до змісту, опосередкованого через співвідношення одиниць змісту з одиницями виразу.
Змістовий аналіз	Аналіз елементів і логіко-семантичних відношень між ними для подання семантики контенту.
Синтаксичний аналіз	Автоматично за наявності лексико-граматичних та граматичних даних до кожного слова синтаксично прив'язують словоформи у реченні.
Морфемний аналіз	Сегментування тексту, де виділення префіксів можливе без знання частин мови, а суфіксів – ні: потрібні різні їх набори та процедури відсікання суфіксів для кожної частини мови окремо.
Класифікація	Автоматичне опрацювання фрагментів текстового контенту для розпізнавання змісту.
Кодування	Кодування фрагментів текстового контенту.
Архівація	Збереження фрагментів текстового контенту в базі даних.

Рис. 12. Результат виявлення ключових слів

Оператор рубрикації α_5 комерційного контенту є відображенням контенту C_3 в новий стан C_4 через його валідацію, який відрізняється від попереднього стану його приналежністю до множини тематичного контенту $\alpha_5 : (C_3, U_{CT}, T) \to C_4$. За змістовний аналіз контенту відповідає процес витягування граматичних даних зі слова через графемний аналіз та корегування результатів морфологічного аналізу через аналіз граматичного контексту лінгвістичних одиниць (алг. 3).

Алгоритм 3. Рубрикація текстового комерційного контенту

Етап 1. Поділ комерційного контенту C_3 на блоки.

Крок 1. Подання на вхід блоку побудови дерева блоки комерційного контенту C_3.

Крок 2. Створення нового блоку в таблиці блоків.

Крок 3. Накопичення символів до символу нового рядка.

Крок 4. Перевірка на наявність крапки перед символом нового рядка. Якщо є, то перехід до кроку 5, якщо ні, то збереження послідовності у таблиці, розбір нового блоку контенту C_3 та перехід до крок 3.

Крок 5. Перевірка наявності кінця тексту для контенту C_3. Якщо кінець тексту, то перехід до кроку 6, якщо ні, то зберігається накопичена послідовність у таблицю, розбір нового блоку контенту C_3 та перехід до кроку 2.

Крок 6. Отримання на виході дерево блоків контенту C_3 у вигляді таблиці $U_{CT}^B \in U_{CT}$.

Етап 2. Поділ блоку на речення зі збереженням структури контенту C_3.

Крок 1. На вхід подається таблиця блоків $U_{CT}^B \in U_{CT}$. Створення таблиці речень $U_{CT}^R \in U_{CT}$ із зв'язком за полем Код_розділу типу *n-to-*1 із таблицею блоків контенту C_3.

Крок 2. Створення нового речення в таблиці речень $U_{CT}^R \in U_{CT}$.

Крок 3. Накопичення символів до крапки, крапки з комою або символу нового рядка.

Крок 4. Перевірка на наявність скорочення. Якщо скорочення, то перехід до кроку 5, якщо ні, то збереження послідовності у таблиці $U_{CT}^R \in U_{CT}$, розбір нового речення та перехід до кроку 2.

Крок 5. Перевірка наявності кінця тексту блоку для контенту C_3. Якщо кінець тексту, то перехід до кроку 6, якщо ні, то збереження послідовності у таблиці $U_{CT}^R \in U_{CT}$, розбір нового речення та перехід до кроку 2.

Крок 6. Отримують на виході дерево речень у вигляді таблиці $U_{CT}^R \in U_{CT}$.

Крок 7. Перевірка наявності кінця тексту для контенту C_3. Якщо кінець тексту, то перехід до кроку 8, якщо ні, то розбір нового блоку та перехід до кроку 1.

Крок 8. Отримання на виході дерева речень у вигляді таблиць $U_{CT}^R \in U_{CT}$.

Етап 3. Поділ речень на лексеми із вказанням належності до речень $U_{CT}^L \in U_{CT}$.

Крок 1. Формування на основі таблиці речень таблиці лексем $U_{CT}^L \in U_{CT}$ із полями Код_лексеми (унікальний ідентифікатор), Код_речення (число, рівне коду речення з лексемою), Номер_лексеми (число, рівне номеру лексеми в реченні), Текст (текст лексеми).

Крок 2. Подання на вхід для розбору на лексеми речення з таблиці речень $U_{CT}^R \in U_{CT}$.

Крок 3. Створення нової лексеми в таблиці лексем $U_{CT}^L \in U_{CT}$.

Крок 4. Накопичення символів до крапки, пропусків або кінця речення та збереження в таблиці лексем.

Крок 5. Перевірка кінця речення. Якщо так, то перехід до кроку 6, якщо ні, то збереження накопиченої послідовності у таблицю $U_{CT}^L \in U_{CT}$, розбір нової лексеми та перехід до кроку 3.

Крок 6. Проведення синтаксичного аналізу на основі вихідних даних (алг. 2).

Крок 7. Проведення морфологічного аналізу на основі даних, одержаних на виході.

Етап 4. Визначення тематики комерційного контенту $U_{CT}^T \in U_{CT}$.

Крок 1. Побудова ієрархічної структури властивостей $U_{CT}^T \in U_{CT}$ кожної лексичної одиниці тексту, що містить граматичну та семантичну інформацію.

Крок 2. Формування лексикону з ієрархічною організацією типів властивостей, де кожен тип-нащадок успадковує і перевизначає властивості предка.

Крок 3. Уніфікація – базовий механізм побудови синтаксичної структури.

Крок 4. Визначення ключових слів *KeyWords* комерційного контенту $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ при $U_{CT} = \{U_{CT1}, U_{CT2}, U_{CT3}, U_{CT4}\}$, де U_{CT} – колекція умов рубрикації, U_{CT1} – множина тематичних ключових слів зі словника, U_{CT2} – множина частот вживання ключових слів в комерційному контенті, U_{CT3} – множина залежностей вживання ключових слів різних тематик (коефіцієнти визначає модератор згідно належності ключового слова до певної тематики в межах $[0,1]$), U_{CT4} – множина частот вживання тематичних ключових слів в контенті. (алг.2).

Крок 5. Визначення $U_{Ct}^{T} \in U_{Ct}$ з *TKeyWords* – тематичні ключові слова в множині *KeyWords* для . *Topic* – тема контенту та *Category* – категорія контенту.

Крок 6. Визначення *FKeyWords* – частота вживання ключових слів та *QuantitativeryTKey* – частота вживання тематичних ключових слів в тексті комерційного контенту.

Крок 7. Визначення *Comparison* – порівняння появи ключових слів різних тематик Розрахунок *CofKeyWords* – коефіцієнт тематичних ключових слів контенту, *Static* – коефіцієнт статистичної важливості термів, *Addterm* – коефіцієнт наявності додаткових термів. Порівняння множини ключових слів контенту з ключовими поняттями тем. Якщо є збіг, то перехід до кроку 9, якщо ні, то перехід до кроку 8.

Крок 8. Формування нової рубрики з набором ключових понять аналізованого контенту C_4.

Крок 9. Присвоєння визначеній рубриці аналізованого комерційного контенту C_4.

Крок 10. Розрахунок *Location* – коефіцієнт розташування контенту C_4 в тематичній рубриці.

Етап 4. Заповнення бази пошукових образів для атрибутів *Topic* – тема контенту, *Category* – категорія контенту, *Location* – коефіцієнт розташування контенту в тематичній рубриці, *CofKeyWords* – коефіцієнт тематичних ключових слів в контенті, *Static* – коефіцієнт статистичної важливості термів, *Addterm* – коефіцієнт наявності додаткових термів, *TKeyWords* – тематичні ключові слова, *FKeyWords* – частота вживання ключових слів, *Comparison* – порівняння появи ключових слів різних тематик, *QuantitativeryTKey* – частота вживання тематичних ключових слів в тексті комерційного контенту C_4.

Побудова тексту контенту C_4 визначається темою, вираженою інформацією, умовами спілкування, завданням повідомлення та стилем викладення. Із семантичною, граматичною та композиційною структурою контенту C_4 пов'язані його стильові/стилістичні характеристики, залежні від індивідуальності автора та підпорядковані тематичній/стильовій домінанті тексту. Процес рубрикації контенту C_4 у вигляді діаграми варіантів подано на рис. 13.

Рис. 13. Діаграма варіантів використання для процесу рубрикації контенту в СЕКК

Основні етапи визначення морфологічних ознак U_{CT} одиниць тексту C_4: визначення граматичних класів слів – частин мови і принципів їх класифікаційного виділення; виокремлення частини семантики слова як морфологічної; обґрунтування набору морфологічних категорій та їх природи; опис сукупності формальних засобів, закріплених за частинами мови та їх морфологічними категоріями. Процес рубрикації $C_4 = \alpha_5(\alpha_4(C_2, U_K), U_{CT})$ через автоматичне індексування складових комерційного контенту C_3 розбито на послідовні блоки: морфологічний аналіз, синтаксичний аналіз, семантико-синтаксичний аналіз лінгвістичних конструкцій та варіювання змістовного запису текстового контенту (рис. 14).

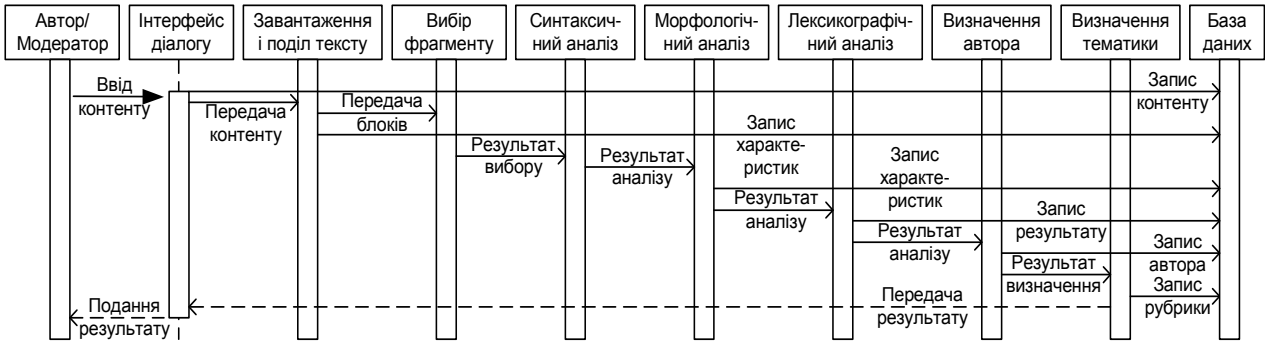


Рис. 14. Діаграма послідовності для процесу рубрикації контенту в СЕКК

Використано такі способи виразу граматичного значення: синтетичний, аналітичний, аналітико-синтетичний та суплетивний. Граматичні значення узагальнені через однотипні характеристики та підлягають поділу на часткові значення. Для позначення класів однотипних граматичних значень використано поняття граматичної категорії. До морфологічних значень належать категорії роду, числа, відмінка, особи, часу, способу, стану, виду, об'єднувані у парадигми для класифікації частин тексту. Об'єктом морфологічного аналізу є структура слова, форми словозміни, способи виразу граматичних значень. Морфологічні ознаки одиниць тексту – це інструменти дослідження зв'язку між лексикою, граматикою, використанням їх у мовленні, парадигматикою (відмінкові форми відмінюваних слів) і синтагматикою (лінійні зв'язки слів, сполучення). Реалізація автоматичного кодування слів тексту, тобто приписування їм кодів граматичних класів, пов'язане з граматичною класифікацією. Морфологічний аналіз містить такі етапи: виділення основи у словоформі; пошук основи у словнику основ; порівняння структури словоформи з даними у словниках основ, коренів, префіксів, суфіксів, флексій. У процесі аналізу ідентифікують значення слів та синтагматичних відношень між словами контенту. Інструментами аналізу є словники основ/флексій/омонімів та статистичних/синтаксичних словосполучень, зняття лексичної омонімії, семантичний аналіз іменних безприйменникових конструкцій, таблиці семантико-синтаксичного сполучення іменників/прикметників та компонентів прийменникових конструкцій, алгоритми аналізу для визначення послідовностей перевірок і звертань до словника і таблиць; система поділу слів тексту на флексію й основу; тезаурус еквівалентностей для заміни еквівалентних слів одним/кількома номерами понять, які слугують ідентифікаторами змісту замість основ слів; тезаурус у вигляді ієрархії понять для забезпечення пошуку для даного поняття загального/асоційованого з ним поняття; система обслуговування словників. Процес індексування залежить від дескрипторного словника або інформаційно-пошукового тезауруса (рис. 11, *б*). Дескрипторний словник має структуру таблиці з колонками: основи слів; набори дескрипторів, приписані кожній основі; граматичні ознаки дескрипторів. Індексування складається з виділення інформативних словосполучень з тексту; розшифрування абревіатури; заміна слів з основами-дескрипторами на код дескриптора; зняття омонімії.

Формування дайджестів комерційного контенту. Дайджест – це короткий зміст публікації в СЕКК, для формування якого використовуються контент-аналіз з врахуванням частотних ваг слів із сформованого словника понять. Оператор формування дайджестів комерційного контенту $\alpha_6 : (C_4, U_D, T) \to C_5$ є відображенням комерційного контенту C_4 в новий стан C_5, який відрізняється від попереднього стану появою нової частини контенту у вигляді його короткого змісту, що доповнює попередній стан. Процес формування дайджестів складається з алгоритмів формування словника понять (алг. 4) та створення дайджесту (алг. 5).

Алгоритм 4. Формування словника понять.

Етап 1. Формування словника понять.

Крок 1. Послідовне виділення всіх лінгвістичних одиниць з вхідного контенту.

Крок 2. Побудова алфавітно-частотного словника.

Крок 3. Нормалізація слів через автоматичний морфологічний аналіз.

Крок 4. Модифікація алфавітно-частотного словника.

Крок 5. Приписування словам ваги w (частоти появи).

Крок 6. Вилучення зі словника незначних слів ($W \le k$, де k – значення порогу вилучення).

Етап 2. Вибір тематичного словника відповідно до запиту.

Етап 3. Коригування алфавітно-частотного словника з урахуванням термів тематичного словника (коригування значень ваг окремих одиниць).

Етап 4. Вибір $N = n$ слів із більшою вагою вагомих із алфавітно-частотного словника, де $n = const$ і задається модератором.

Алгоритм 5. Створення дайджесту.

Етап 1. Вибір контенту з урахуванням його ваги.

Крок 1. Завдання розміру дайджесту C_4.

Крок 2. Виконання алгоритму 1.

Крок 3. Послідовне визначення ваги контенту як суми значень ваг окремих лінгвістичних одиниць $W = \sum_i w_i$.

Крок 4. Сортування вхідного потоку контенту за величинами ваг.

Крок 5. Визначення змістовних дублів за статистичним критерієм унікальності тексту $U_D \ge 0{,}9$ (алг. 1).

Крок 6. Фільтрування контенту, непридатного для формування дайджестів (при $W \le l$, де l – значення порогу вилучення контенту, за допомогою правил структуризації та модерації контенту із самонавчанням) та статистично змістових дублів.

Крок 7. Вибір $V = q$ контенту із більшою вагою, де $q = const$ і задається модератором.

Етап 2. Побудова тексту дайджесту з відібраного контенту.

Крок 1. Побудова словника з відібраного контенту (алг. 4).

Крок 2. Застосування контент-аналізу до тексту (алг. 1).

Крок 3. Фільтрування речень, що не відповідають семантичним правилам структуризації та модерації контенту.

Крок 4. Автоматичне формування гіпертекстового подання дайджесту, його змісту і гіперпосилання на вихідні джерела.

Етап 3. Редагування сформованого тексту дайджесту c_{i4}, де $C_5 = \{c_{i4}, C_4\}$.

Крок 1. Перевірка обсягу c_{i4} сформованого контенту C_5. Якщо $c_{i4} < C_4$, то виконання кроку 2, інакше виконання етапу 4.

Крок 2. Видалення з вхідного потоку контенту C_4 з сформованим дайджестом c_{i4}.

Крок 3. Виконання етапів 1-2.

Крок 4. Дописування до сформованого дайджесту отриманого та перехід до кроку 1.

Етап 4. Форматування тексту дайджесту як окремий контент C_5 та збереження в бази даних дайджестів із посиланням на джерело.

Процес формування дайджестів комерційного контенту C_5 формує множину коротких анотацій та основних положень контенту за певний період. Це зручно для швидкого ознайомлення з основними змістом певної тематики/рубрики, дослідження та пошуку необхідного контенту.

Процес розподілу комерційного контенту. Процес розподілу контенту (рис. 15) реалізує розподілення навантаження між авторами/модераторами СЕКК при активному збільшенні постійної аудиторії читачів та обсягу затребуваного комерційного контенту.

Рис. 15. Діаграма варіантів використання для процесу розподілу контенту СЕКК

Релевантний контент розсилають користувачам та завантажують в БД. Вибіркове поширення контенту описано $C_6 = \alpha_7(C_5, U_{DS})$, де U_{DS} – множина умов вибіркового поширення контенту. На рис. 16 діаграма кооперації ілюструє взаємодію модератора із модулями оцінювання матеріалу, рейтингування авторів та розподілу дайджестів. Оцінювання кожного контенту проводять у комплексі з іншим тематичним контентом (рис. 17). Одним із важливих критеріїв розподілу дайджестів між авторами є процент унікальності опублікованого контенту кожного автора. Системами визначення унікальності контенту є Praide Unique Content Analyser 2, FIndCopy та Miratools.

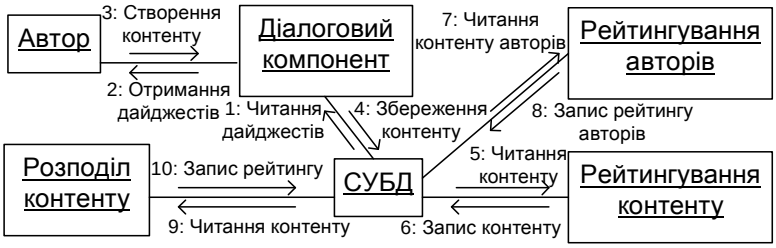


Рис. 16. Діаграма кооперації для процесу розподілу комерційного контенту

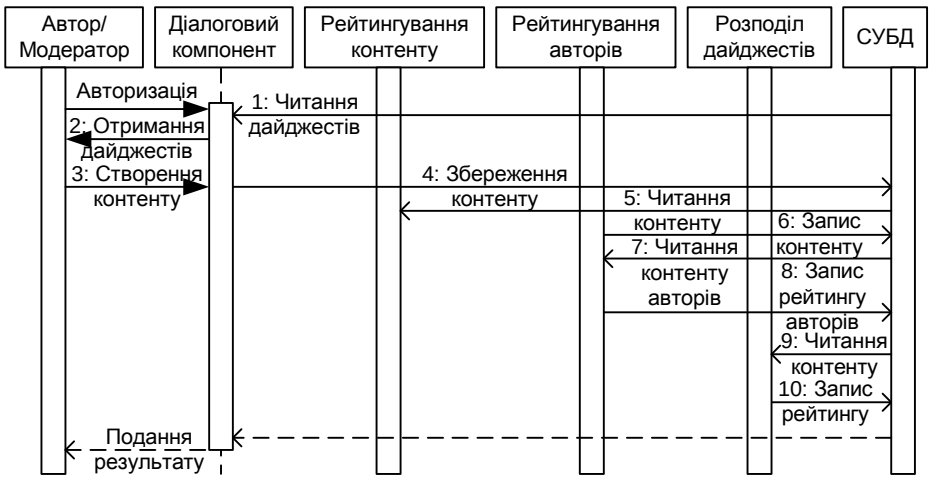


Рис. 17. Діаграма послідовності для процесу розподілу комерційного контенту

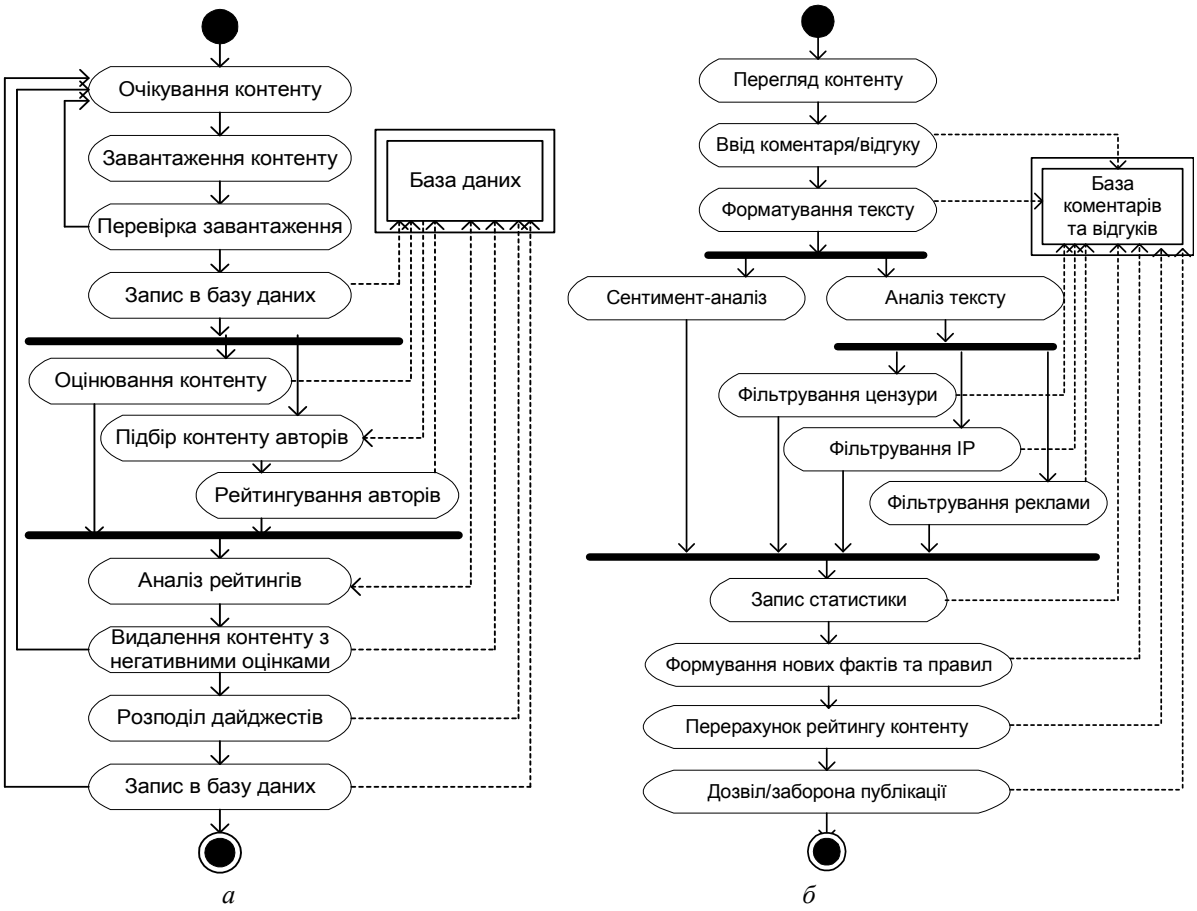


Рис. 18. Діаграма діяльності для процесу а) розподілу контенту та б) управління контентом

Спочатку підсистема отримує готові дайджести з джерел через RSS. Потім дайджести розподіляють між авторами за їх рейтингом: першими отримують дайджести для опрацювання автори з найвищим рейтингом (рис. 18). Після виконання усіх дій підсистема переходить у стан очікування до появи нового контенту. Рейтинг авторів вказує на продуктивність/результативність роботи кожного з них окремо. Впливають на нього такі критерії U_{DS}, як процент унікальності контенту (якість роботи автора), кількість переглядів контенту (вага пошукових та прямих переходів), оцінка користувача (активність користувачів) та час перебування на сторінці з контентом (міра зацікавленості користувачів в контенті). Підсистема рейтингування оцінює якість роботи за сукупністю критеріїв, що забезпечує об'єктивність та стимулювання якісної роботи. Частка роботи модератора як розподіл однотипних даних, її сортування, оцінювання та аналіз зменшується. Це зменшує застосування ресурсів, скорочує час створення контенту та покращує характеристики контенту із-за об'єктивності оцінювання якості виконаних завдань. З процентом унікальності тексту оцінюють якість роботи автора цього контенту та заносять отриману оцінку в таблицю рейтингів.

Висновки і перспективи подальших наукових розвідок

Зазвичай розподіл виконують модератори. Підсистема розподілу контенту скорочує час та зменшує ресурси для подальшого функціонування СЕКК. Процес розподілу передбачає декілька етапів: формування списку об'єктів розподілу (наприклад, статей, програмного забезпечення, книг або дайджестів); визначення критеріїв/ознак розподілу контенту з отриманого списку (процент унікальності контенту; кількість звернень до контенту; користувацька оцінка; час перегляду); рейтингування авторів контенту; оцінювання параметрів контенту з метою використання в процесі розподілу. Наведені критерії не є однаковими за значенням та важливістю під час аналізу їх загалом та обчислення зведеної оцінки якості роботи авторів контенту. Контент містить тему та дайджест. Підсистема розподілу контенту вибірково розсилає дайджести між авторами згідно з

рейтингуванням якості їх роботи. Збільшення обсягу контенту призводить до точнішого оцінювання якості та продуктивності кожного автора контенту. Збільшення кількості критеріїв оцінки дозволяє охопити ширший спектр аспектів роботи автора/модератора.

1. Baeza-Yates R. Modern Information Retrieval / Ricardo Baeza-Yates, Berthier Rebeiro-Neto // Menlo Parl, California, New York : ACM Press, Addison-Wesley, 1999. Access mode: http://people.ischool. berkeley.edu/~hearst/irbook/print/chap10.pdf. 2. Boiko B. Content Management Bible. – Hoboken, 2004. – 1176 p. 3. Braslavski P.Style-Dependent Document Ranking / P. Braslavski, A. Tselishchev // In Proc. RCDL'2005. Access mode: http://www.rcdl2005.uniyar.ac.ru/ru/RCDL2005/papers/sek7_1_paper.pdf. 4. Brin S. The Anatomy of a Large-Scale Hypertextual Web Search Engine / S. Brin, L. Page // Статья с WWW7. Access mode: http://www-db.stanford.edu/pub/papers/google.pdf. 5. CM Lifecycle Poster / Content Management Professionals. – Режим доступу: http://www.cmprosold.org/resources/poster/. 6. CMIS. Part I – Introduction, General Concepts, Data Model, and Services / EMC, IBM and Microsoft Corporation. –2008. – 76 p. 7. Grefenstette G. Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques / G. Grefenstette // Proceedings of SIGIR, 1995. 8. Hearst M.A. Automatic Acquisition of Hyponyms from Large Text Corpora / M.A. Hearst //Proc. of the 14th International Conference on Computational Linguistics, Nantes, France, 1992. Access mode: http://acl.ldc.upenn.edu/ C/C92/C92-2082.pdf. 9. Karlgren J. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis / J. Karlgren, D. Cutting // In Proceedings of the 15th International Conference on Computational Linguistics (COLING), Kyoto, 1994, vol. 2, p. 1071–1075. Access mode: http://www.sics.se/~jussi/Papers/1994_Coling_Kyoto_l/cmplglixcol.ps. 10. Manning C.D. Foundations of Statistical Natural Language Processing / C.D. Manning, H. Schutze // Chapter 5: Collocations. Эл. версия главы. 11. Manning D.C. Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schutze // Cambridge University Press. 2007. Главы будущей книги. 12. Manning C. An Introuction to information retrieval / Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze // Cambridge University Press, Cambridge, England.. 2008, 482 pp, ISBN: 978-0-521-86571-5, Online edition © 2009 Cambridge Up. Access mode: http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf. 13. Sato S. Automatic collection ofrelated terms from the web / S. Sato, Y. Sasaki // In Proc. 41st ACL, 2003. p. 121–124. 14. Sebastiani F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys, Vol. 34, No. 1, March 2002, pp. 1–47. 15. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine, Access mode: http://company.yandex.ru/articles/iseg-las-vegas.html + обсуждение в форуме. 16. Takkinen J. IRI: Introduktion och IR-systemgrunder (modellering och utvardering) / Juha Takkinen // IISLAB/ADIT/IDA, Linkopings universitet, 2006-01-24. Access mode: https://www.ida.liu.se/~TDDC08/tddc08-ir1.pdf. 17. Zamir O. Web Document Clustering: A Feasibility Demonstration / O. Zamir, O. Etzioni // In Proc. SIGIR'98. 18. Айвазян С. А. Прикладная статистика: Классификация и снижение размерности: Справ. изд. / Под. ред. С. А. Айвазяна. – М.: Финансы и статистика, 1989. 19. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л: НУЛП, 2009. – 612 с. 20. Браславский П. И. Интеллектуальные информационные системы / П. И. Браславский. – Режим доступу: http://www.kansas.ru/ai2006/. 21. Браславский П.И. Фасетная организация интернет-каталога и автоматическая жанровая классификация документов / П. И. Браславский, Е. А. Вовк, М. Ю. Маслов // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. семинара "Диалог-2002". Т. 2. – М.: Наука, 2002. – С. 83–93. Режим доступу:: http://company.yandex.ru/ articles/article8.html. 22. Браславский П. И. eXtragon: экспериментальная система для автоматического реферирования веб-документов / П. Брасласвский, И. Колычев // Труды РОМИП-2005. – СПб., 2005. – С. 40–53. Режим доступу: http://www.romip.narod.ru/romip2005/03_extragon.pdf. 23. Гаврилова Т. А. Извлечение и структурирование знаний для экспертных систем / Т. А. Гаврилова, К. Р. Червинская. – М.: Радио и связь, 1992. 24. Гаврилова Т. А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб.: Питер, 2000. 25. Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения / А. В. Гладкий. – М.: Наука, 1985. 26. Добров Б. Н. Формирование базы терминологических свловосочетаний по

текстам предметной области / Б. Н. Добров, Н. В. Лукашевич, С. В. Сыромятников // Электронные библиотеки: Труды конференции RCDL'2003. – СПб, 2003. – С. 201–210. Режим доступу: http://rcdl2003.spbu.ru. 27. Добрынин В. Теория информационно-логических систем. Информационный поиск. (Методичекие указания к курсу) / В. Добрынин. – СПб., 2002. Режим доступу: http://ir.apmath.spbu.ru/publications/dobrynin_ir_intro/. 28. Іванов В. Контент-аналіз: Методологія і методика дослідження ЗМК / В. Іванов. – К., 1994. – 112 с. 29. Иванов С. Статистический анализ документальных информационных потоков / С. Иванов, Н. Круковская // Научно-техническая информация. – 2004. – № 2. – С. 11–14. 30. Искусственный интеллект: Справочник: Кн.1: Системы общения и экспертные системы. – М.: Радио и связь, 1990. 31. Искусственный интеллект: Справочник: Кн.2: Модели и методы. – М.: Радио и связь, 1990. 32. Клифтон Б. Google Analytics / Б. Клифтон. – М.: ООО "И. Д. Вильямс", 2009. – 400 с. 33. Коваленко А. Вероятностный морфологический анализатор русского и украинского языков / А. Коваленко. Режим доступу: http://www.keva.ru/stemka/stemka.html. 34. Кукушкина О. В. Определение авторства текста с использованием буквенной и грамматической информации / О. В. Кукушкина, А. А. Поликарпов, Д. В. Хмелёв // Проблемы передачи информации, 2001, Т. 37, Вып. 2. – С. 96–108. Режим доступу: http://www.math.toronto.edu/dkhmelev/PAPERS/published/gramcodes/gramcodes.pdf. 35. Ландэ Д. Основы моделирования и оценки електронных информационных потоков / Д. Ландэ, В. Фурашев, С. Брайчевский, О. Григорьев. – К.: Інжиніринг, 2006. – 348 с. 36. Леоньтьева Н. Н. Автоматическое понимание текстов: системы, модели, ресурсы / Н. Н. Леоньтьева. – М.: Издательский центр "Академия", 2006. 37. Нейл К. Web-инструмент для выявления плагиата / К. Нейл, Г. Шанмагантан //Открытые системы. 2005. – № 01. – С. 40–44. Режим доступу: http://www.osp.ru/os/2005/01/040_print.htm. 38. Некрестьянов И. Системы текстового поиска для Веб / И. Некрестьянов, Н. Пантелеева // Программирование. – 2002. – № 28(4). – С. 207–225. Режим доступу: http://meta.math.spbu.ru/~nadejda/papers/web-ir/web-ir.html. 39. Осуга С. Обработка знаний / С. Осуга. – М.: Мир, 1989. 40. Пасічник В. Математична лінгвістика / В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич. – Львів: Новий Світ, 2012. – 359 с. 41. Пенроуз Р. Новый ум короля: О компьютерах, мышлении и законах физики / Р. Пенроуз. – М.: УРСС, 2003. 42. Перспективы развития вычислительной техники в 11 кн. Кн. 2. Интеллектуализация ЭВМ. – М.: Высшая школа, 1989. 43. Попов Э.В. Общение с ЭВМ на естественном языке / Э. В. Попов. – М.: Наука, 1982. 44. Попов Э. В. Статические и динамические экспертные системы / Э. В. Попов, И.Б. Фоминых, Е. Б. Кисель, М. Д. Шапот. – М.: Финансы и статистика, 1996. 45. Рао С. Р. Линейные статистические методы и их применения / С. Р. Рао. М.: Наука, 1968. 46. Сегалович И. В. Как работают поисковые системы / И. В. Сегалович // Мир Internet, – 2002. – № 10. Режим доступу: http://www.dialog-21.ru/directions/Segalovich_vorprint.doc. 47. Сокирко А. В. Морфологические модули на сайте www.aot.ru / А. В. Сокирко //Материалы конференции "Диалог-2004". Режим доступу: http://www.dialog-21.ru/Archive/2004/Sokirko.htm. 48. Солтон Д. Динамические библиотечно-информационные системы / Д. Солтон. – М.: Мир, 1979. – 560 с. 49. Факторный, дискриминантный и кластерный анализ: Пер. с англ. – М.: Финансы и статистика, 1989. 50. Федорчук А. Контент-мониторинг информационных потоков / А. Федорчук. – БНАН. – К., 2005. – № 3. Режим доступу: www.nbuv.gov.ua/articles/2005/05fagmip.html. 51. Хан У. Системы автоматического реферирования / У. Хан, И. Мани // Открытые системы, 2000. – №12. Режим доступу: http://www.osp.ru/os/2000/12/067_print.htm. 52. Хмелев Д. Распознавание автора текста с использованием цепей А.А. Маркова / Д. Хмелев // Вестник МГУ, сер. 9: Филология, № 2, 2000. – С. 115–126. Режим доступу: http://www.rusf.ru/books/analysis/vestnik2000win.htm. 53. Храмцов П. Информационно-поисковые системы Internet / П. Храмцов // Открытые системы. – 1996. – № 3. Режим доступу: http://www.osp.ru/os/1996/03/46_print.htm.