ІНФОРМАЦІЙНІ СИСТЕМИ, МЕРЕЖІ ТА ТЕХНОЛОГІЇ

УДК 004.738.5

В. А. Андруник[1], Л. В. Чирун[2]
Національний університет "Львівська політехніка",
[1]кафедра інформаційних систем та мереж,
[2]кафедра програмного забезпечення

ФОРМУВАННЯ ТА РУБРИКАЦІЯ ЕЛЕКТРОННИХ ДАЙДЖЕСТІВ

Запропоновано функціонально-логістичний метод опрацювання контенту як етап його життєвого циклу. Метод опрацювання комерційного контенту описує процеси формування та рубрикації електронних дайджестів і спрощує технологію управління комерційним контентом. У роботі проаналізовано основні проблеми функціональних сервісів опрацювання комерційного контенту. Запропонований метод дає можливість створити засоби опрацювання інформаційних ресурсів та реалізувати підсистеми управління комерційним контентом.

Ключові слова: контент, контент-аналіз, контент-моніторинг, контентний пошук.

The functional logistic method of content processing as the content life cycle stage is proposed in the given article. The method of commercial content processing describes the information resources formation and rubrication processes and simplifies the commercial content management. In the given article the main problems of functional services of commercial content processing are analyzed. The proposed method gives an opportunity to create means of information resources processing and to implement the commercial content management subsystems.

Key words: content, content analysis, content monitoring, content search.

Вступ. Загальна постановка задачі

Розв'язання складних задач у будь-якій сфері життєдіяльності потребує інформаційної підтримки. Задоволення інформаційних потреб є необхідною умовою здійснення інновацій. Але складність отримання інформації впливає на оперативність і якість прийняття рішення. Інтернет-середовище є одним із масштабних засобів масової інформації (ЗМІ) у вигляді множини контенту інформаційного ресурсу [1]. Однак хаотичність виникнення та функціонування інформаційних ресурсів, відсутність у більшості з них чіткої періодичності підтримання та оновлення, а також недоліки реалізації ефективного пошуку контенту не дають змоги використовувати інтернет-середовище як єдиний суцільний ЗМІ. Як повноцінні ЗМІ прийнято розглядати окремі елементи мережі. Мережевими ЗМІ є портали новин із визначеною періодичністю оновлення, електронні версії друкованих періодичних видань, електронні газети та журнали. Інтернет-середовище не створює конкуренції традиційним ЗМІ за багатьма показниками, проте за деякими критеріями має переваги завдяки технічним характеристикам. Інтернет-середовище успішно виконує роль джерела і засобу поширення контенту [1–15].

Аналіз останніх досліджень та публікацій

Особливістю інтернет-середовища є постійне зростання темпів виробництва та поширення в ньому контенту різного типу та збільшення його обсягів до масштабів, які унеможливлюють його безпосереднє опрацювання [1]. Існує низка специфічних проблем, пов'язаних зі швидким розвитком

інформаційних технологій (ІТ). З одного боку, є такий потужний інформаційний масив, як інтернет-ресурси для прийняття рішень у різних сферах життя держави, суспільства та фізичної або юридичної особи. А з іншого – брак необхідної для прийняття рішень інформації через її динаміку, обсяги, темпи виробництва, джерела та неструктурованість. Охоплення/узагальнення великих динамічних інформаційних потоків контенту, які безперервно генеруються в ЗМІ, вимагає якісно нових підходів [11–15]. Ситуація різкого зростання темпів виробництва контенту та збільшення його обсягів в інтернет-середовищі призвела до виникнення різних проблем [1–8, 11–15]:

- непропорційне збільшення інформаційного шуму з-за слабкої структурованості контенту;
- поява паразитного контенту (одержуваного як додатки);
- невідповідність формально релевантного контенту (тематично відповідного) справжнім потребам його споживачів;
- багаторазове дублювання контенту (наприклад, публікація в різних виданнях).

Експоненційне зростання темпів виробництва інформації суттєво знижує ефективність опрацювання інформації традиційними методами [1–8]. Важливі дані багаторазово дублюються на інформаційних ресурсах, кількість котрих збільшується за експоненційним законом. На початку комп'ютерної ери створювалися програми автоматизованого опрацювання текстів, що реалізують індексування, анотування, реферування, фрагментованість та інші форми інформаційного аналізу і синтезу. Але більшість процесів створення анотацій та автоматичного реферування неефективні, зберігається необхідність в масштабованих методологіях і програмах.

Існує багато шляхів вирішення проблеми через такі два напрями, як квазіреферування та короткий виклад змісту первинних документів. Квазіреферування основане на екстрагуванні фрагментів документів, тобто виділенні найінформативніших фраз і формуванні з них квазірефератів [9].

Короткий виклад вихідного матеріалу ґрунтується на виділенні з текстів за допомогою методів штучного інтелекту і спеціальних інформаційних мов найсуттєвішої інформації і породження нових текстів, змістовно узагальнювальних первинних документів. Застосовуючи такий підхід, можна одержувати складніші анотації, які можуть містити інформацію, що доповнює вихідний текст. Спираючись на формальне подання семантики вихідного документа, такі системи налаштовані на високий ступінь стиснення, необхідний, наприклад, для розсилання повідомлень на мобільні пристрої. Тому головна відмінність між засобами реферування полягає в формуванні набору витягів або короткому викладі документа. Всі сучасні промислові системи класу Text mining містять засоби автореферування, які є невід'ємними компонентами таких систем.

Однією з базових процедур систем цього класу (автоматичне формування дайджестів) є автореферування на основі великої кількості документів. Для дайджесту відбирають документи, в яких найповніше відбиті тенденції усього вхідного потоку. Такі дайджести найбільшою мірою відповідають інформаційним потребам користувача, за запитом якого формується цей вхідний інформаційний потік. На підставі реферату, що становить за обсягом незначну частину вихідного тексту, користувачі можуть скласти обґрунтований висновок про первинний документ, витративши на це значно менше зусиль, порівняно з його ознайомленням [8]. У разі автореферування обсяг реферату повинен становити від 5 % до 30 % вихідного тексту. Підготовка документів у вигляді анотації з декількох джерел (дайджестів) передбачає ще більший ступінь стиснення.

Квазіреферування зводиться до екстрагування (вилучення) з документів мінімальних релевантних фрагментів з використанням аналізу поверхнево-синтетичних відносин лексичних одиниць у тексті [9, 10]. Квазіреферування акцентує на виділенні характерних фрагментів методом зіставлення фразових шаблонів, в результаті чого виділяються блоки найбільшої лексичної та статистичної релевантності. Автоматичне визначення частот використання окремих слів і поєднань у вихідному документі дає змогу визначати абзаци і пропозиції, в яких тематика документа подана найточніше. Створення результуючого документа виконується простим з'єднанням вибраних фрагментів. Сформований квазіреферат є зв'язним текстом. Якість реферування залежить від жанру опрацьованого тексту. Змістовність квазіреферату залежить і від інших особливостей вхідного тексту. Для великих текстів побудова якісного реферату з фрагментів вихідного документа без урахування семантичних закономірностей практично неможлива. Основою аналітичного етапу

квазіреферування є процедура обчислення вагових коефіцієнтів для кожного блока тексту відповідно до таких характеристик, як розташування цього блока в оригіналі, частота появи в тексті, частота використання в ключових фразах тощо [20].

Text Mining є множиною методів опрацювання тексту, в результаті застосування яких з'являються нові знання [16]. Це міждисциплінарна область, в якій використовують базові технології Data Mining в поєднанні з техніками таких дослідницьких областей, як пошук інформації (Information Retrieval, IR), вилучення інформації (Information Extraction, IE), математична лінгвістика, класифікація, кластеризація, створення онтології [17]. В кожній з цих областей вирішують свої специфічні прикладні завдання, але складно провести чітку межу між Text Mining й іншою областю досліджень: всі вони мають справу з текстами, тому і загальні проблеми, і підходи до їх вирішення перетинаються. Відмінність полягає в кінцевій меті. Наприклад, в IR мета – знайти документи, які хоча б частково збігаються з пошуковим запитом і серед знайдених відібрати ті, для яких збіг є найповнішим [17]. А методи Text Mining спрямовані на виявлення невідомих фактів і прихованих взаємозв'язків під час аналізу семантичних, лексичних і статистичних ознак у масивах текстів, хоча алгоритми для цього використовують однакові. Так, IE відрізняється від Text Mining тим, що в цій області розглядаються способи вилучення специфічної інформації, структурованих даних, таких як імена людей, географічні назви, заголовки книг за попередньо заданими відносинами [18]. В Text Mining наперед не відомо, яка інформація може бути виявлена. Методи Text Mining ефективно застосовують для створення та заповнення баз даних (БД).

В Text Mining до основних елементів належать сумаризація (summarization), виділення феноменів, понять (feature extraction), кластеризація (clustering), класифікація (classification), відповідь на запити (question answering), тематичне індексування (thematic indexing) і пошук за ключовими словами (keyword searching). Також в деяких випадках набір доповнюють засоби підтримки та створення таксономії (oftaxonomies) і тезаурусів (thesauri). Олександр Лінден, директор компанії Gartner Research, виділив чотири основні види додатків IT Text Mining.

1. Класифікація тексту з використанням статистичної кореляції для побудови правил розміщення документів в обумовлені категорії. Застосовують, наприклад, для вирішення таких завдань: групування документів у Intranet-мережах, розміщення документів у визначені папки, вибіркове поширення новин передплатникам.

2. Кластеризація ґрунтується на ознаках документів з використанням лінгвістичних та математичних методів без використання обумовлених категорій. Застосовують для реферування великих документальних масивів, визначення взаємопов'язаних груп документів, для спрощення візуалізації інформації, виявлення дублікатів або близьких за змістом документів.

3. Семантичні мережі або аналіз зв'язків визначають появу дескрипторів (ключових фраз) у документі для забезпечення навігації. Використана при цьому візуалізація є ключовою ланкою подання схем неструктурованих текстових документів. Застосовують як засіб подання контенту всього масиву документів, а також в реалізації навігаційного механізму для дослідження документів та їх класів.

4. Витяг фактів призначено для одержання фактів з тексту з метою поліпшення класифікації, пошуку та кластеризації.

Існує ще кілька завдань технології Text Mining, наприклад, прогнозування і знаходження винятків (пошук об'єктів з характеристиками, що виділяються із загальної маси) [11].

Стемінг (англ. *Stemming*) – це процес скорочення слова до основи відкиданням допоміжних частин, таких як закінчення чи суфікс. Результати стемінгу схожі на визначення кореня слова, але його алгоритми ґрунтуються на інших принципах. Тому слово після опрацювання алгоритмом стемінгу (стематизації) може відрізнятися від морфологічного кореня слова. Стемінг застосовують в лінгвістичній морфології та в інформаційному пошуку. Багато пошукових систем використовують стемінг для злиття, тобто об'єднання слів, у яких збігаються форми після стематизації (вважають такі слова синонімами). Цей алгоритм використовує принцип пошуку в таблиці, в якій зібрані всі можливі варіанти слів та їх форми після стемінгу. Перевагами цього методу є простота, швидкість та зручність опрацювання винятків з мовних правил. До недоліків зарахуємо те, що таблиця пошуку має містити всі форми слів.

Подання тексту визначається поставленою задачею, від чого залежить легкість та ефективність маніпуляцій даними. Найширше використовуване подання – це модель Vector Space Model (VSM) [14]. Тоді текст описують вектором, вимірювання якого задано кількістю параметрів тексту. Значення цих параметрів – це функції частот, з якими ці параметри з'являються в текстовому корпусі. Цей процес згадується як *модель сумки слів*, тому що порядок і відношення між словами не враховуються. Більшість пропонованих алгоритмів подання тексту є розширенням моделі Vector Space Model. Деякі з них ґрунтуються на фразах, інші враховують семантику слів або відношення між ними, в третіх – використовується ієрархічна структура тексту [18].

Частота, з якою термін з'являється в текстовому корпусі, прояснює значення цього терміна в окремому документі. Частоту визначають двояко: для підкреслення наявності/відсутності терміна вона змінюється в межах [0;1] або задається математичною функцією. Нормалізацію виконують з урахуванням розміру документа та всіх унікальних термінів.

Статистику збирають як для окремих слів, так і для фраз. Фрази надають більше семантичної інформації, ніж окремі слова, бо вони дають загальне уявлення і про контекст. Слово характеризується власним оточенням і через багатозначність більшості слів необхідно знати хоча б одну фразу, яка містить аналізоване слово, щоб визначити його семантичне значення з більшою впевненістю. У табл. 1 сформульовано переваги і недоліки подання документа за окремими словами або цілими фразами. Використання фраз компенсує недоліки аналізу окремих слів і навпаки.

Таблиця 1

Переваги та недоліки слів і фраз у поданні документа

Назва	Переваги	Недоліки
Слова	Визначення синонімів. Наявність розроблених засобів та алгоритмів.	Відсутність контекстної інформації. Проблеми зі знаходженням стійких словосполучень.
Фрази	Наявність контекстної інформації. Можливість стійких словосполучень.	Середнє значення за статистичного опрацювання.

Завдання класифікації документа змінюється залежно від попередньо виявлених зв'язків усередині документа і між документами. Попередньо визначають критерії класифікації, перш ніж вирішити, який алгоритм слід застосувати. Критеріями групування бувають загальна тема робіт, автор, актуальність або міра зацікавленості текстом постійними користувачами. У разі тематичної класифікації фокусуються на іменниках, які можуть характеризувати тему. Розроблення методів автоматичного навчання є основним додатком для реалізації цього типу класифікації. Янг і Лью [19] провели порівняння деяких навчальних алгоритмів, доводячи, наприклад, що методи SVM, k найближчих сусідів і лінійної регресії працюють краще, ніж нейронні мережі та метод Байєса.

Дайджест є анотованим текстом, побудований на основі аналізу кількох документів. Під час складання дайджестів методи автореферування одного документа поширюються на масив з великої кількості документів. Більшість алгоритмів автоматичного реферування документів припускають три основні етапи: аналіз вихідного тексту, визначення вагомих фрагментів (речень або цілих абзаців) і формування висновку. Дайджест є анотованим джерелом гіперпосилань на документи, покладені в його основу. У разі формування дайджестів методами квазіреферування практично неможливо отримати зв'язний текст. Об'єднання рефератів кожного з документів міститиме надмірну незв'язну інформацію. Однак за умови формування автореферату, що складається з певної кількості анонсів вхідних документів і розділеного на підрозділи відповідно до цих документів, описаний вище метод є цілком прийнятним [8].

Контент-моніторинг є змістовим аналізом інформаційних потоків з метою отримання необхідних якісних і кількісних зрізів [1–3]. На відміну від контент-аналізу, він здійснюється неперервно в часі. Безперервне аналітичне опрацювання повідомлень є найхарактернішою рисою цього підходу, який дає змогу видобути факти з тестів, виявляти нові поняття, формувати різноманітні статистичні звіти [3–8]. Названі завдання охоплено двома основними технологіями – видобуванням фактографічної інформації з текстів (Information Extraction) та глибинним аналізом текстів (Text Mining) [2]. Методи контент-моніторингу є адаптацією класичних методів контент-

аналізу до умов динамічних інформаційних масивів, наприклад, потоків інформації з Інтернету. На відміну від систем інтеграції інформації, які реалізують ідею збирання та накопичення всієї доступної інформації, як з внутрішніх, так і з зовнішніх джерел, системи контент-моніторингу дають змогу виявляти неочевидні закономірності в документальних масивах даних або текстів (так звані латентні, або приховані, знання). Системи цього класу дозволяють здійснювати аналіз великих масивів документів і формувати предметні покажчики понять і тем, висвітлених у цих документах. Типова задача контент-моніторингу – це побудова діаграм динаміки появи понять у часі [12].

В автоматизованій технології контент-моніторингу є кілька важливих особливостей:

- використання ключового фрагмента публікації як одиниці формування текстового інформаційного масиву;
- формування банку ключових фрагментів публікацій як об'єднання двох взаємозалежних автоматизованих процесів: аналітико-синтетичного опрацювання та багаторівневої процедури контент-аналізу текстів публікацій;
- індексація ключових фрагментів публікацій за допомогою фасетної класифікації.

Унікальність запропонованої технології полягає в об'єднанні змістових і кількісних методів контент-аналізу. Послідовність етапів змістового аналізу проблеми, яку досліджує конкретна інформаційна система, умовно поділяють на змістовий (якісний) аналіз сукупності публікацій і формалізований (кількісний) аналіз інформаційних масивів: індексного, бібліографічного та масиву текстів ключових фрагментів публікацій [6]. На ринку ПЗ є достатня кількість аналітичних систем, орієнтованих на математичний і статистичний аналіз різних кількісних, цифрових показників. Але немає якісного інструментарію для аналізу величезного обсягу текстової інформації в друкованих виданнях, рядках новин інформаційних агентств, тематичних сайтах в Інтернеті. Саме тому набуває актуальності робота у цій галузі.

Кластеризація. В результаті виконання пошукової процедури користувачу подають списки документів, впорядковані за зниженням відповідності запиту. У результаті неминучих неточностей під час ранжування результатів пошуку такий вид подання не завжди є зручним. Тоді використовують кластеризацію результатів пошуку, яка дозволяє подати отримані результати в узагальненому вигляді, що спрощує вибір області, відповідної інформаційним потребам користувача [16]. У цьому випадку використовують два класи методів кластеризації – ієрархічний або неієрархічний. За ієрархічної кластеризації (знизу вверх або зверху вниз) формується дерево кластерів. Методи неієрархічної кластеризації забезпечують якісну кластеризацію за рахунок складніших алгоритмів. Для цих методів є деяка порогова функція якості кластеризації, максимізація якої досягається за рахунок розподілу документів між окремими кластерами.

Тематичне індексування (близькість). Тематика документа визначається його словниковим запасом, а тематична близькість термів характеризується тим, наскільки часто ці терми використовуються в документах тієї самої тематики. Це не завжди означає обов'язкове використання цих термів у тих самих документах.

Позначимо тематичну близькість двох термів w_i і w_j як $FSR(w_i, w_j)$. Розрахунок оцінок тематичної близькості термів і завдання функції $FSR(w_i, w_j)$ виконуються за результатами аналізу використання термів у масиві документів, якими описуються тематики [1–8]. За вихідним масивом документів будується матриця A, рядки якої відображають розподіл термів за документами. Як оцінка тематичної близькості двох термів використовується скалярний добуток відповідних рядків цієї матриці. Для обчислення оцінок близькості між усіма парами термів досить обчислити матрицю A^T. Такий підхід аналогічний до класичних методів подання інформації, які основані на векторно-просторовій моделі. Подальшим розвитком такого підходу є використання так званого латентно-семантичного аналізу. За матрицею A будується її апроксимація A, отримана методом латентно-семантичного аналізу. Функція тематичної близькості двох термів $FSR(w_i, w_j)$ однозначно задається матрицею A^T:

$$FSR(w_i, w_j) = A^T[w_i, w_j]. \quad (1)$$

Зазначимо, що матриця A має розмірність k, де k – це вибрана для апроксимації бажана розмірність простору тематик. За такого підходу трудомісткість обчислення тематичної близькості двох термів становить x^k, тобто вона не залежить від кількості аналізованих документів і розміру загального словника.

Таблиця взаємозв'язаних понять. Як основу для групування документів у інформаційному масиві використовують поняття (не окремі терміни, а деякі семантичні сутності), які, теоретично, можна висловити мовою запитів. Точно так само, як і у випадку окремих термів, кластеризація документів зіставляється з кластеризацією понять; поняття точніше відображають тематичні властивості документів. Це досягається за рахунок ускладнення алгоритмічної частини кластеризації. Побудова таблиць взаємозв'язків понять (ТВП) ґрунтується на мовних засобах інформаційно-пошукової системи, а також методах кластерного аналізу. Семантичне значення понять визначається на основі інформаційно-пошукової мови.

Таблиця взаємозв'язків понять, яка будується як статистичний звіт, відображає близькість (спільне входження в документах) окремих понять з реального світу, – це симетрична матриця $A - \| a_{ij} \|$, елементи якої a_{ij} – це коефіцієнти взаємозв'язку відповідних пар понять. Коефіцієнт a_{ij} відповідає кількості документів вхідного інформаційного потоку, що містять поняття (терміни або словосполучення, подані мовою запитів, що відповідають поняттю i), а коефіцієнт a_{ij}, де $(i \neq j)$ – кількості документів у вхідному потоці, що водночас відповідають поняттям i та j.

Якісні ознаки цілком адекватно виражаються інформаційно-пошуковою мовою. Це рішення здебільшого є ефективним і оперативним. Для перевпорядкування понять з метою виявлення блоків-множин найвзаємозалежніших понять застосовується алгоритм кластерного аналізу.

Булева модель пошуку є класичною і широко використовуваною моделлю подання інформації (ґрунтується на теорії множин) та моделлю інформаційного пошуку (ґрунтується на математичній логіці). Популярність цієї моделі пов'язана з простотою її реалізації, що дає змогу індексувати і виконувати пошук у масивах документів великого обсягу. Нині популярним є об'єднання булевої моделі з алгебраїчною векторно-просторовою моделлю подання даних. Це забезпечує, з одного боку, швидкий пошук з використанням операторів математичної логіки, а з іншого боку – якісне ранжування документів, що базується на вагах ключових слів. У межах булевої моделі документи та запити подаються у вигляді множини морфемних основ ключових слів (термами). В булевій моделі запит користувача є логічним виразом, в якому ключові слова (терми запиту) пов'язані логічними операторами AND, OR і NOT. У різних пошукових системах в Інтернеті за замовчуванням не використовують в явному вигляді логічні операції, а просто перераховують ключові слова. Найчастіше за замовчуванням передбачається, що всі ключові слова з'єднуються логічною операцією AND. В цих випадках в результати пошуку входять тільки ті документи, які містять одночасно всі ключові слова запиту. У тих системах, в яких пробіл між словами прирівнюється до оператора OR, в результатах пошуку містяться документи, в які входить хоча б одне з ключових слів запиту. В разі використання булевої моделі БД містить індекс, який організовується у вигляді інвертованого масиву, в якому для кожного терма зі словника БД існує список документів, в яких цей терм трапляється. В індексі можуть зберігатися також значення частоти входження цього терма в кожному документі, що дає змогу сортувати список за зниженням частоти входження. Класична БД, що відповідає булевій моделі, організована так, щоб за кожним термом можна було швидко отримати доступ до відповідного списку документів. Структура інвертованого масиву забезпечує його швидку модифікацію, якщо в БД входять нові документи. У зв'язку з цими вимогами інвертований масив часто реалізується у вигляді B-дерева.

Латентно-семантичний аналіз, або індексування, є методом вилучення *прихованих* контекстозалежних значень термів і структури семантичних взаємозв'язків між ними шляхом статистичного опрацювання великих наборів текстових даних [8]. Цей метод широко використовується в сфері пошуку та в задачах класифікації інформації. Цей підхід дає змогу автоматично розпізнавати змістові відтінки слів залежно від контекстів їх використання. Він використовує виявлені показ-

ники тематичної близькості термів, які потім застосовуються для обчислення оцінок тематичної близькості документів. Метод широко застосовується в факторному аналізі, завдання якого – виділення головних факторів з простору елементарних.

Матричний латентно-семантичний аналіз. Математичний апарат цього методу ґрунтується на сингулярному розкладанні матриць. Метод дозволяє виявити приховані семантичні зв'язки у разі опрацювання великих масивів документів. Як вихідну інформацію латентно-семантичний аналіз використовує ту саму матрицю, що і у векторно-просторовій моделі. Елементи цієї матриці містять значення частоти використання окремих термів у документах. З матричного аналізу відомо, що будь-яку прямокутну матрицю A можна розкласти на добуток трьох матриць: $A = UXV^T$ таких, що матриці U і V складаються з ортонормованих колонок, а X – діагональна матриця сингулярних значень, діагональні елементи яких є сингулярними числами матриці A, тобто невід'ємними квадратними коренями власних чисел матриці. Найпоширеніший варіант оснований на використанні розкладу матриці за сингулярними значеннями, завдяки чому вихідна матриця розкладається в множину з k ортогональних матриць, лінійна комбінація яких є непоганим наближенням вихідної матриці.

Виділення проблем

Внаслідок перелічених обставин традиційні інформаційно-пошукові системи поступово втрачають актуальність [1]. Причина цього криється не стільки у фізичних обсягах інформаційних потоків, скільки в їх динаміці, тобто в постійному систематичному оновленні інформації, яка далеко не завжди має очевидну регулярність. Охоплення і узагальнення великих динамічних інформаційних потоків, які безперервно генеруються в ЗМІ, потребує якісно нових підходів [1–3].

Вихід можна знайти тільки в засобах автоматизації виявлення найважливіших складових в інформаційних потоках. Протягом останніх років все частіше використовують системи моніторингу ресурсів. Цей перспективний напрям отримав назву контент-моніторинг. Його поява зумовлена завданнями систематичного відстеження тенденцій і процесів у інформаційному середовищі, котре постійно оновлюється. Під контент-моніторингом найчастіше розуміють змістовий аналіз інформаційних потоків з метою отримання необхідних якісних і кількісних зрізів, який ведеться постійно протягом проміжку часу, не визначеного заздалегідь [1–9]. Найважливішою методологічною складовою контент-моніторингу є контент-аналіз [8].

Технологію ефективного (глибинного) аналізу тексту Text mining найчастіше використовують для отримання зрізів інформаційних потоків. Використовуючи обчислювальні потужності, він дає змогу виявити відношення, які можуть призводити до отримання нових знань. Завдання Text mining – вибрати ключову і найбільш значущу інформацію для користувача. Користувачу не потрібно переглядати величезну кількість неструктурованої інформації. Розроблені на основі статистичного і лінгвістичного аналізу, а також методів штучного інтелекту, технології Text mining призначені для проведення змістового аналізу, забезпечення навігації та пошуку в неструктурованих текстах. Застосовуючи системи класу Text mining, користувачі отримують нову цінну інформацію у вигляді знань. Технології глибинного аналізу тексту історично передували створенню технології глибинного аналізу даних, методологія та підходи якої широко використовуються в методах Text mining. Як і більшість когнітивних технологій, Text mining – це алгоритмічне виявлення раніше не відомих зв'язків та кореляцій у вже наявних текстових даних.

Експоненційне зростання обсягу інформації в Інтернеті є причиною все більшого утруднення пошуку необхідних документів та організації їх у вигляді структурованих за змістом сховищ. Користувачу стає все важче знайти необхідну інформацію, традиційні механізми пошуку малоефективні. Тому актуальність теми спричинена експонентним зростанням кількості документів, що унеможливлює опрацювання даних традиційними методами без втрати якості.

Формулювання мети

Мета дослідження – проектування та розроблення інтелектуальної компоненти автоматичного формування та рубрикації електронних дайджестів.

Для досягнення поставленої мети необхідно розв'язати такі задачі:

1. Провести системний аналіз предметної області.

2. Розробити модуль формування та рубрикації дайджестів.

Об'єктом дослідження є аналіз процесів формування (опрацювання, рубрикування) інформаційних потоків у ЗМІ. Предметом дослідження є алгоритми опрацювання тексту та формування дайджестів. Наукова новизна одержаних результатів дослідження зумовлена сукупністю поставлених завдань, полягає у комплексному дослідженні та систематизації теоретичних і прикладних проблем отримання необхідних якісних і кількісних інформаційних зрізів у вигляді дайджестів. У результаті розв'язання поставлених завдань вперше запропоновано контекстну модель системи автоматичного формування та рубрикації дайджестів публікацій ЗМІ.

Аналіз отриманих наукових результатів

Кінцевою метою системи є автоматизоване складання коротких викладів матеріалів – дайджестів електронних публікацій у ЗМІ, тобто видобування найважливіших відомостей з одного або декількох документів та генерація на їх основі лаконічних та інформаційно насичених звітів. Система повинна проводити інформаційний моніторинг, отримуючи великі обсяги даних, аналізувати, систематизувати дані за допомогою автоматичного рубрикатора, акумулювати інформацію, індексуючи матеріал та зберігаючи його в БД, вирішувати завдання тематичної фільтрації та формувати дайджести в автоматичному режимі. Вибір кінцевого єдиного компромісного рішення з урахуванням різноманітних критеріїв є достатньо складним завданням під час планування та прийняття рішень, тому доцільно подати задачу відбору вагомої інформації в ієрархічній формі за допомогою методу аналізу ієрархії. Інформація відбирається за допомогою алгоритму, котрий опрацьовує вхідний текст та виділяє найбільш інформаційно вагомі його частини, фрагменти. Отже, на верхньому рівні ієрархії ціль – вибір вагомої інформації. На другому рівні – критерії, що уточнюють ціль: основа тексту, повнота словника термінів та кількість ключових термінів. На третьому рівні містяться альтернативи вибору – фрагменти вхідного тексту (рис. 1).

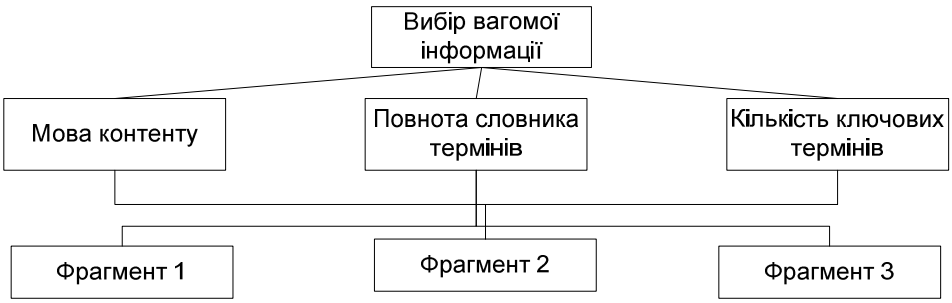


Рис. 1. Ієрархічне подання задачі

Діаграма прецедентів відображає різноманітні сценарії взаємодії між акторами (користувачами) і прецедентами (випадками використання); описує функціональні аспекти системи. На рис. 2 подано діаграму прецедентів системи автоматичного формування та рубрикації дайджестів для електронних ЗМІ. Сформована журналістом стаття опрацьовується системою, внаслідок чого визначаються статистичні показники термінів. Тематична класифікація дає змогу зарахувати статтю до певної рубрики. Після цього за допомогою алгоритмів Text Mining формується дайджест. Для візуалізації статистичних аспектів системи будується діаграма класів. На рис. 3, а подана діаграма класів, що описує систему. Клас *Контент* поданий як частина класу *Аналіз*, котрий є частиною класу *Рубрикатор*. Клас *Словник* поданий як частина класів *Рубрикатор* та *Аналіз*. На рис. 3, б зображена діаграма станів системи автоматичного формування та рубрикації дайджестів для електронних ЗМІ. На ній подано кінцевий автомат з простими станами й переходами.

Метою розроблення є дієздатна, готова до застосування інтелектуальна система автоматичного формування та рубрикації електронних дайджестів. Модуль Web моніторингу дає змогу обходити зазначені користувачем сторінки і виконувати завантаження оновлень Web сторінок. Дані, що надходять, зчитуються спеціальними модулями, які орієнтовані на приймання інформації

цього типу. Після отримання та попереднього опрацювання усі матеріали опрацьовуються модулем рубрикації. Спочатку потрібна побудова і навчання рубрикатора експертом. Суть навчання – в експертному аналізі навчальних матеріалів із зарахуванням їх до тієї чи іншої рубрики, експерт повинен вказати ступінь відношення цього тексту до тієї чи іншої теми.

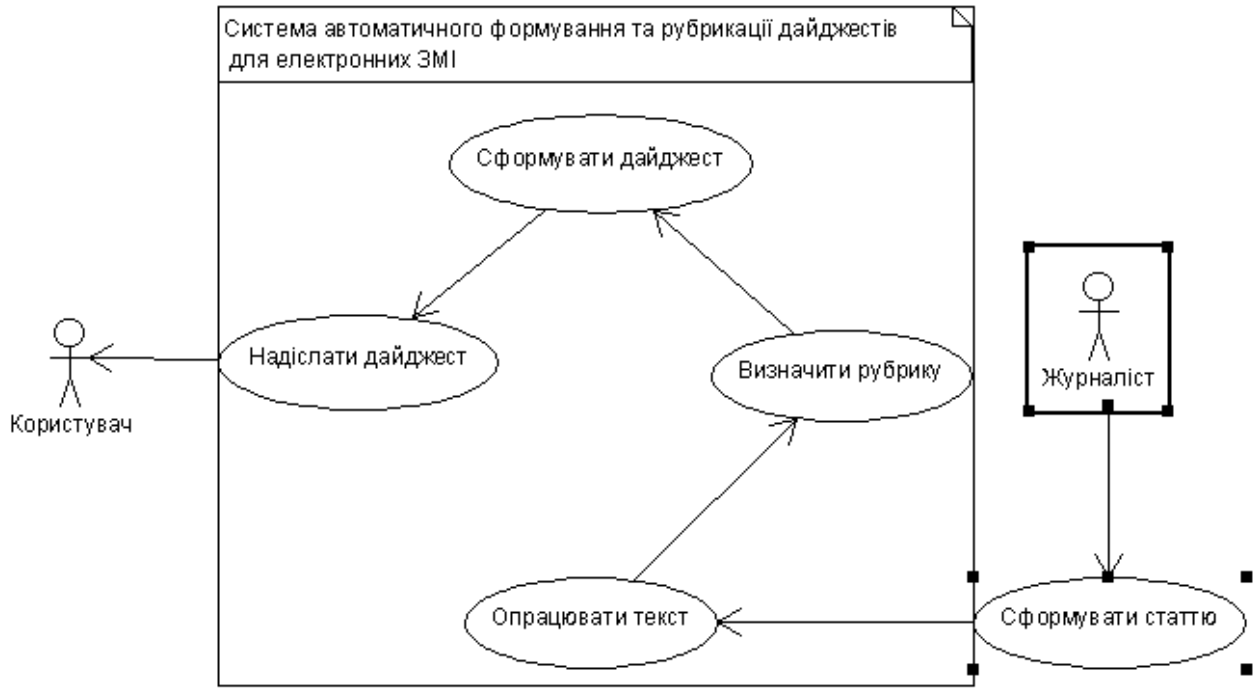
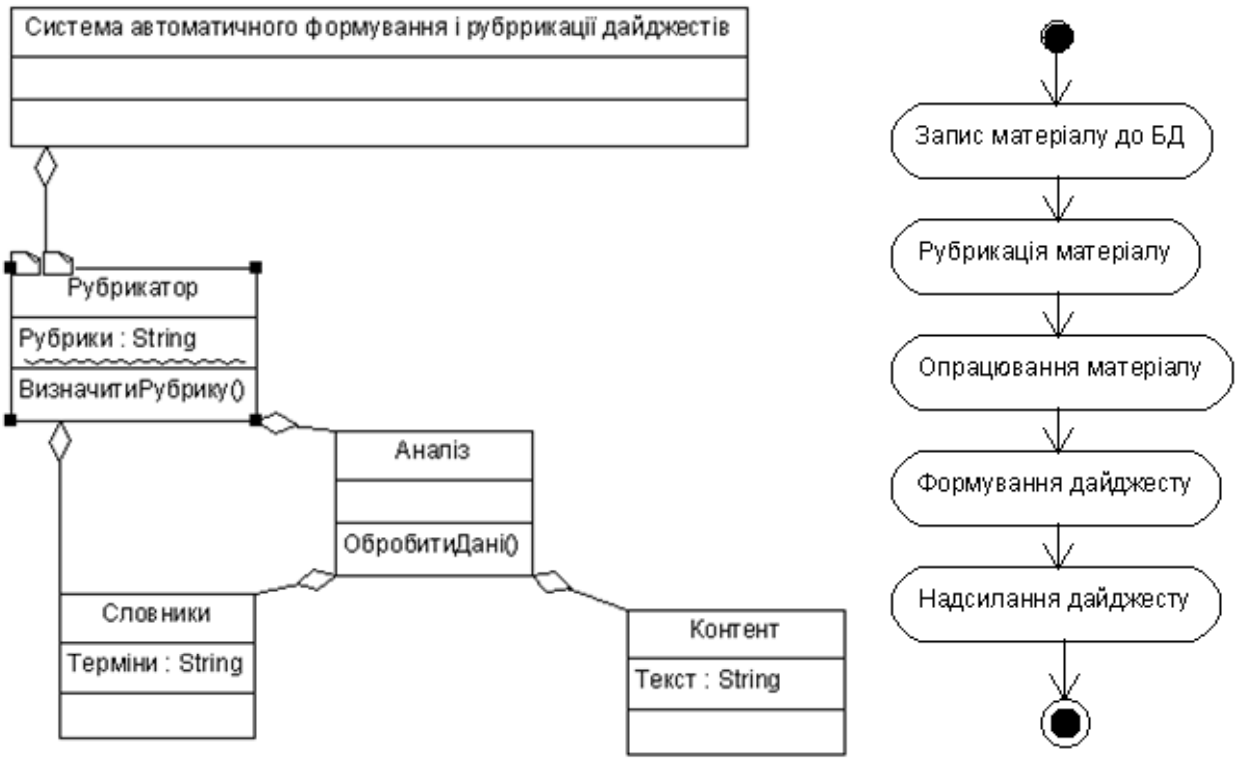


Рис. 2. Діаграма прецедентів системи



а *б*

Рис. 3. Діаграма класів (а) та станів (б)

Діаграма діяльності (Activity diagram) – діаграма, на якій подано розклад деякої діяльності на її складові частини (рис. 4). Під діяльністю (activity) розуміємо специфікацію поводження, що виконується у вигляді координованого послідовного й паралельного виконання підлеглих елементів – вкладених видів діяльності й окремих дій (action), з'єднаних між собою потоками, які йдуть від виходів одного вузла до входів іншого. Аналогом діаграм діяльності є схеми алгоритмів.

На діаграмі послідовностей показано обмін повідомленнями (тобто виклик методів) між декількома об'єктами у окремій обмеженій часом ситуації. Об'єкти є екземплярами класів. Основний наголос в діаграмах послідовностей робиться на порядок і моменти часу, у які повідомлення надсилаються об'єктам. На діаграмах послідовностей об'єкти показано вертикальними штриховими лініями з назвою об'єкта над ними. Вісь часу також має вертикальний напрямок, її спрямовано вниз, повідомлення, які надсилаються від одного об'єкта до іншого, позначено стрілками з назвами операції і параметрів (рис. 5).

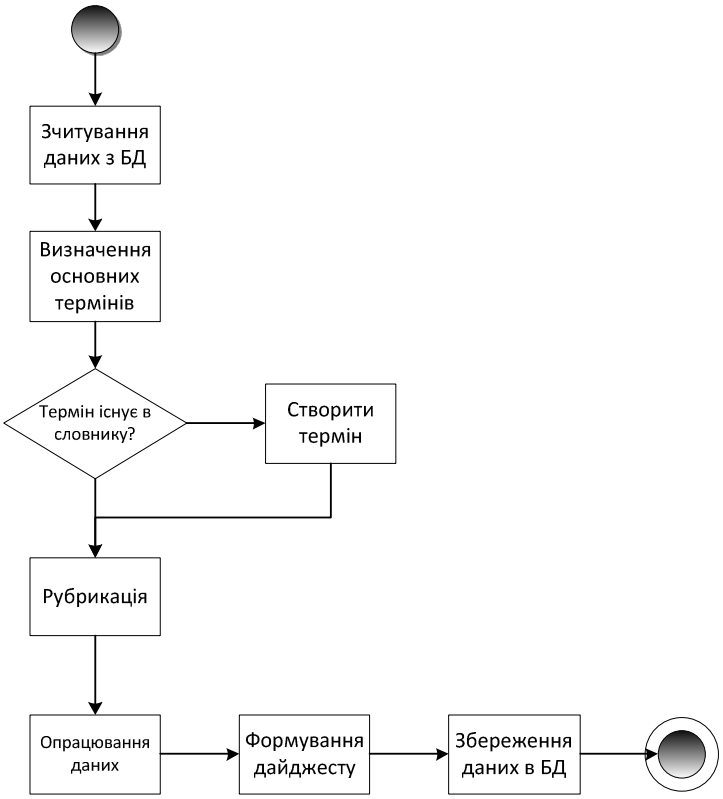


Рис. 4. Діаграма діяльності

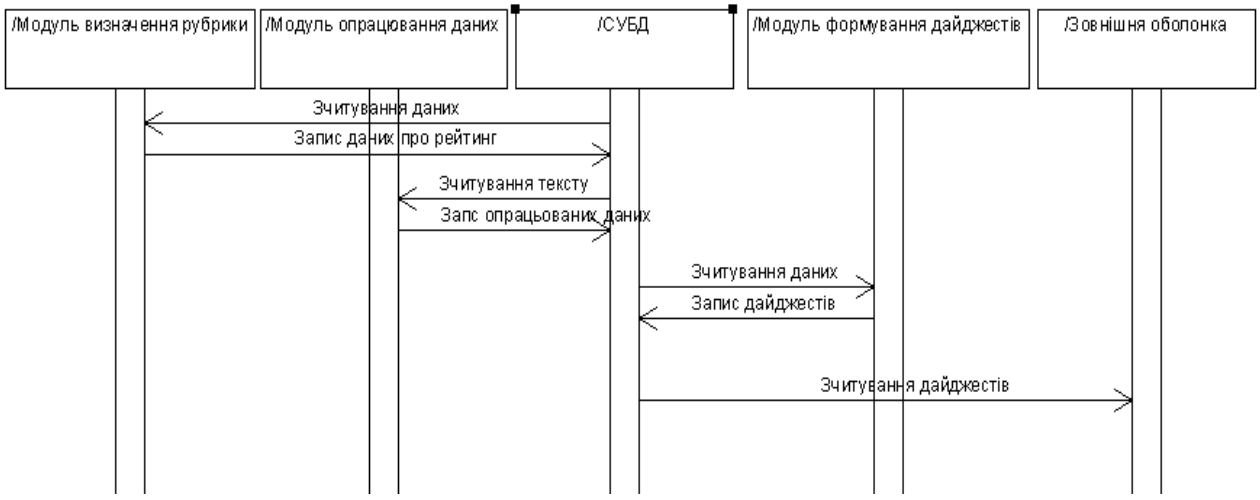
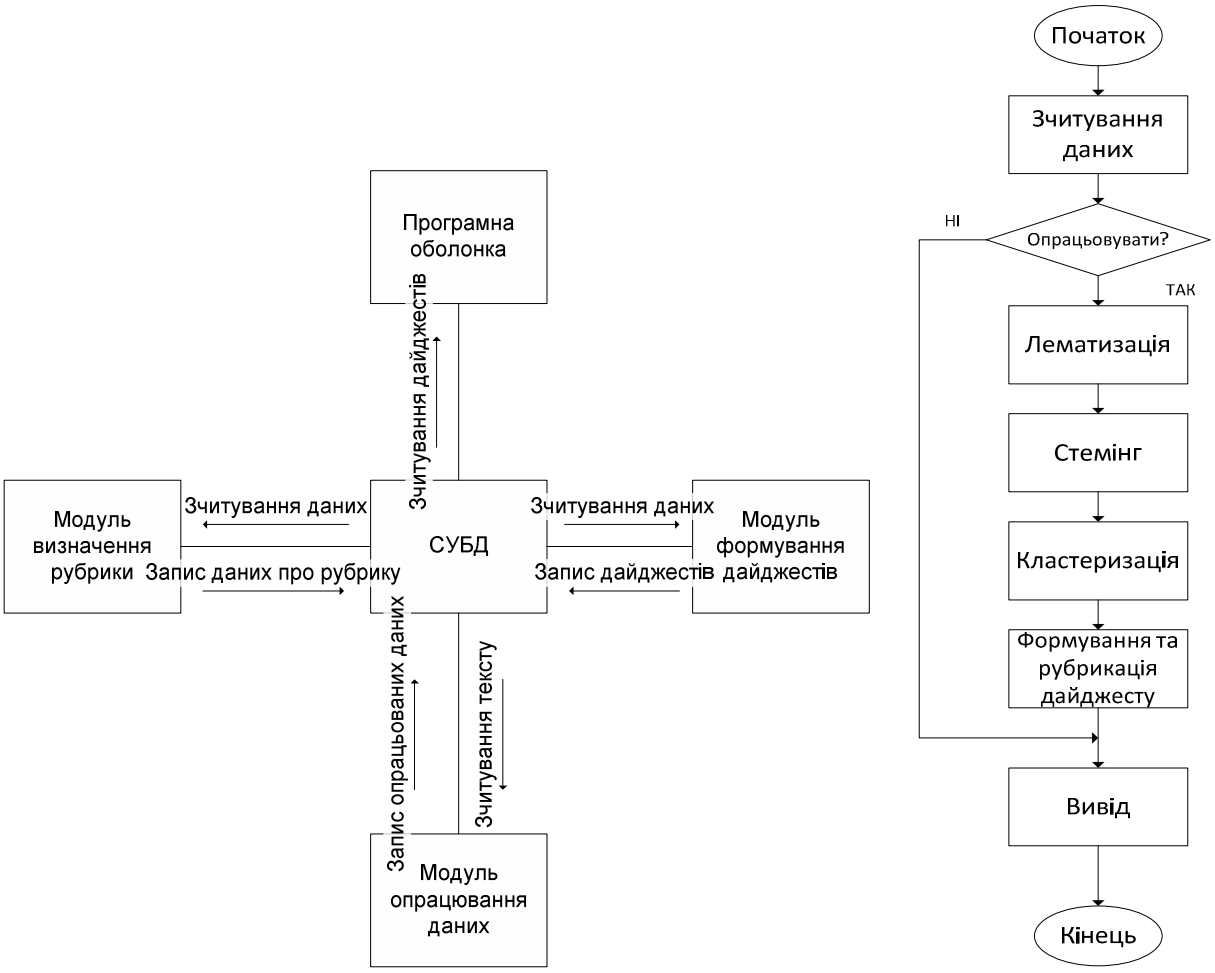


Рис. 5. Діаграма послідовності

Діаграма кооперації (рис. 6, а) призначена для специфікації структурних аспектів взаємодії системи. Кооперації зручні для моделювання шаблонів проектування. На підставі експертних оцінок модуль рубрикації виконує морфологічний і семантичний аналіз текстів, виділяючи основні тематичні поняття та аналізуючи структуру їх розміщення у тексті. Систематизація даних в автоматичному режимі ведеться на підставі результатів навчання. Вхідний текст зараховується до однієї або декількох рубрик з проставляням ступеня відношення. Після розподілу за рубриками матеріали повинен проходити індексацію за всіма словами свого змісту. Ця процедура забезпечує гнучкі можливості пошуку за ознаками матеріалів та за їх змістом. Результати роботи системи подаються у вигляді тематичних дайджестів, які створюються автоматично. Виконуючи опрацювання тексту, необхідно брати до уваги проблеми формування словників, визначення частини мови та людський фактор. По-перше, всі словники різні й не еквівалентні один одному. Найчастіше завдання відрізнити сенс слів один від одного не викликає труднощів, однак у деяких випадках різні значення слова можуть бути близькими семантично (наприклад, якщо кожен з них є метафорою або метонімією), і в таких ситуаціях поділ за сенсом у різних словниках і тезаурусах значно різниться. Вирішенням цієї проблеми є загальне використання того самого джерела даних: одного загального словника. Якщо говорити глобально, то результати досліджень, що використовують узагальненішу систему поділу за сенсом, ефективніші. По-друге, в деяких мовах проблема визначення частини мови (англ. Part-of-speech tagging) слова дуже тісно пов'язана з проблемою дозволу багатозначностей, в результаті чого ці два завдання один одному заважають. Немає єдиної думки, чи варто розділяти їх на дві автономні складові, проте перевага на боці тих, хто вважає, що це необхідно [7]. Третя проблема полягає в людському факторі. Систему оцінюють, порівнюючи результати з результатом роботи експертів. А у випадку стилістичного написання статей завдання побудови змістових дайджестів та коректної їх рубрикації виконує експерт.



а *б*

Рис. 6. Діаграма кооперації (а) та алгоритм роботи системи (б)

Побудова дайджесту. Найпростіша стратегія побудови квазіреферату полягає у фіксації моментів зміни профілю, тобто переходів з низьких щаблів на вищі, і навпаки. Це означає, що фіксують початок і кінець кожного кластера, відбираючи для квазіреферату його перше і останнє речення. Така стратегія характерна для позиційних методів реферування [10], суть яких зводиться до обліку початкових і кінцевих фрагментів у структурі тексту, що задає автор. У тексті наукової статті найважливішими є (поряд із заголовком) вступ і висновок, в кожному з підрозділів – початковий і кінцевий абзац, в кожному абзаці – початкове і кінцеве речення. Не всі наукові статті, не кажучи вже про тексти інших жанрів, розбиті на розділи та підрозділи, що є додатковим аргументом на користь розвиваючого підходу [9]. Друга стратегія побудови квазіреферату полягає у призначенні ваг кожному реченню тексту. Вага речення визначається кількістю входжень у нього словоформ, що демонструють кластеризацію в довільному місці тексту, тобто речення може і не входити до складу відповідного кластера. Варіант, коли вага речення фіксує різноманітність поданих у ньому кластеризованих словоформ, а не повну їх кількість, є привабливішим [9]. Перевага методів квазіреферування полягає в простоті їх реалізації. Проте виділення текстових блоків не враховує відношень між ними, часто призводить до формування беззв'язних рефератів. Деякі речення можуть виявитися пропущеними або в них можуть міститися слова або фрази, які неможливо зрозуміти без попереднього, але пропущеного в авторефераті тексту. Спроби вирішити цю проблему зводяться до вилучення таких речень з рефератів. Рідше роблять спроби здійснення посилань за допомогою методів лінгвістичного аналізу. В деяких людино-машинних підходах створюються спеціальні інтерфейси, за допомогою яких визначають наявність змістового розриву. Такий підхід не підходить для довільного масового опрацювання текстів.

Як і у випадку квазіреферування одного текстового документа, на першому етапі формування дайджесту відбувається відбір найвагоміших лексичних одиниць, що входять у масив вихідних документів (вхідний інформаційний потік), на підставі яких будується словник системи. Вихідні документи з вхідного масиву побудови дайджесту вибирають також з урахуванням їх ваг. Вага кожного документа визначається з урахуванням нормованої за довжиною документа суми ваг окремих слів, що входять в цей документ. Етап вибору документів для дайджесту складається з таких кроків, як визначення ваги кожного документа, сортування вхідного потоку документів за вагами, визначення змістових дублів документів за статистичними критеріями, відкидання документів, непридатних для побудови дайджестів (недопустимих типів документів, наприклад, оглядів), а також змістовних дублів (виявляються за частотними алгоритмами). Останній етап вибору документів для формування дайджесту полягає у виборі заздалегідь певної кількості найвагоміших документів з відсортованого і відфільтрованого на попередніх етапах масиву [8]. Відібрані документи подані в дайджесті заздалегідь заданою кількістю вагомих речень. У разі формування дайджестів на основі інформації, що динамічно змінюється, з Інтернету, автоматично формується гіпертекстове подання дайджесту, який розглядають як самостійний документ, що містить посилання на документи-першоджерела в мережі. Наведена процедура забезпечує формування дайджесту, що відображає основні тенденції, подані у вихідному інформаційному масиві. Має сенс формування *віялового* багатоаспектного дайджесту, що відображає, поряд з головною тенденцією, кілька інших аспектів, ігнорованих у дайджестах першого типу. Багатоаспектний дайджест можна побудувати на основі технологічних рішень, що застосовуються у попередньому підході, реалізуючи такий алгоритм [8].

1 етап. Побудова дайджесту, що відображає основну тенденцію.

2 етап. Видалення з вхідного інформаційного потоку документів, що відповідають тенденції, яка визначена на попередньому кроці.

3 етап. Побудова дайджесту, що відображає основну тенденцію решти інформаційного потоку.

4 етап. Об'єднання отриманих дайджестів.

5 етап. За необхідності (на основі необхідних обсягів результуючого дайджесту) виконується перехід до етапу 2.

Розглянемо алгоритм роботи системи автоматичного формування та рубрикації електронних дайджестів (рис. 6, б). Система отримує дані з БД у вигляді масиву статей, після цього перевіряє, чи задав користувач запит на опрацювання тексту статті. Якщо існує стаття, котрій необхідне опрацювання, то послідовно проводиться лематизація тексту статті для визначення лексем, стемінг для відкидання стоп-слів та кластеризація. В результаті такого опрацювання будуть зібрані дані про положення та ваги словоформ, котрі дадуть змогу зарахувати статтю до визначеної рубрики та побудувати дайджест.

Логічна схема бази даних. Всі характеристики системи подано такими інформаційними відношеннями в SQL:

- jos_content – інформація та метадані статей, розміщених на інформаційному ресурсі;
- digest – інформація про створені дайджести (ідентифікатор, назва, текст дайджесту, дата створення, ідентифікатор рубрики);
- rubrick – інформація про рубрики (ідентифікатор, назва рубрики, вага);
- lexema – таблиця лексем словника (ідентифікатор, лексема, лічильник);
- stem – таблиця нормальних форм словника (ідентифікатор, стема, лічильник);
- stopword – таблиця стоп-слів словника (ідентифікатор, стоп-слово, вага).

Структура БД для інтелектуальної системи автоматичного формування та рубрикації дайджестів подана засобами адміністраторського інтерфейсу phpMyAdmin на рис. 7.

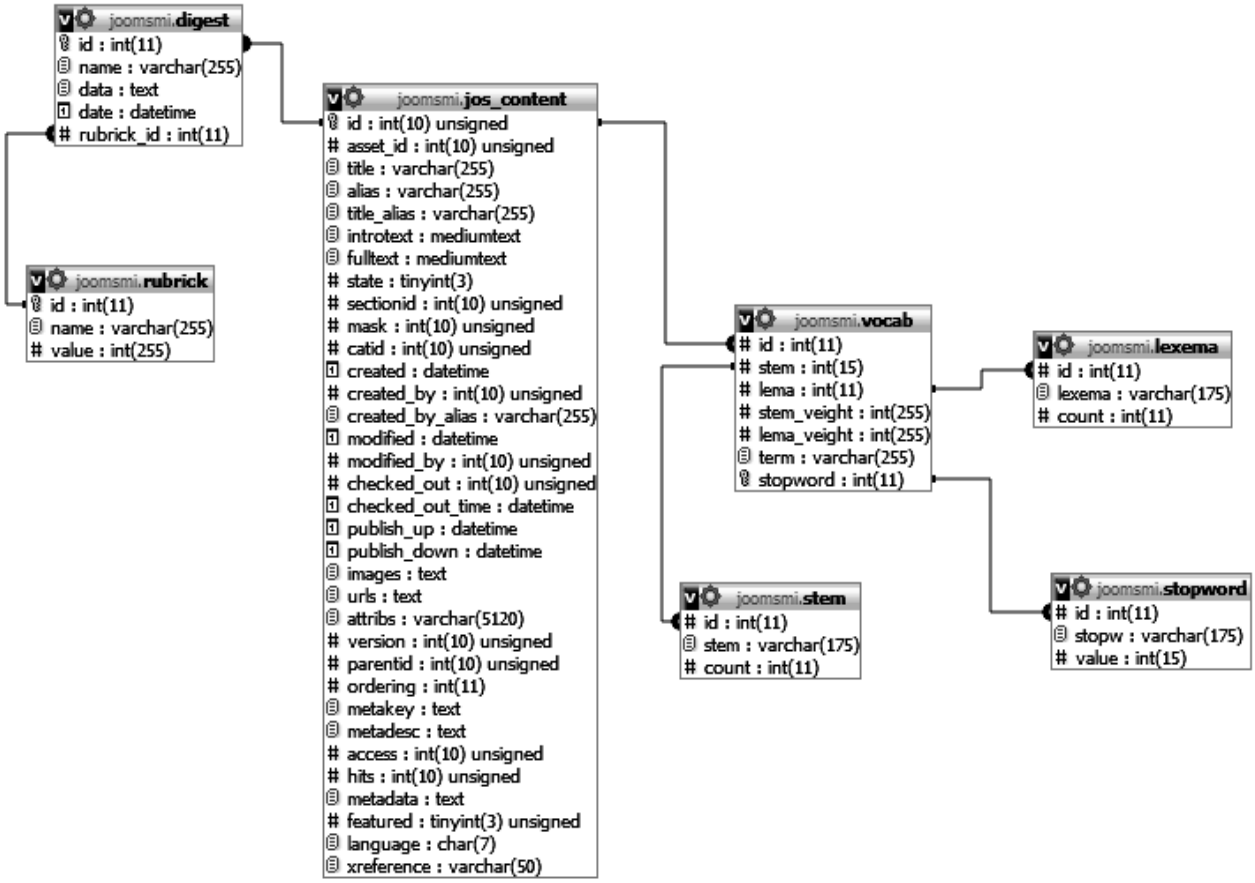


Рис. 7. Структура бази даних

До основних програмних одиниць, що забезпечують функціонал системи, належать файли:

- default.php призначений для формування тексту дайджесту та виконує функції вибору основних коротких тез з повного тексту аналізованої статті;
- stemmer.php призначений для стемінгу тексту – відсікання від слова закінчень і суфіксів, щоб частина, що залишилася, звана stem, була однаковою для всіх граматичних форм слова (в такому вигляді стемер працює тільки з мовами, які реалізують словозміни через афікси);
- lemmatizer.php призначений для лематизації тексту – приведення окремих слів до нормальних словоформ;
- rubrick.php – автоматичний рубрикатор тексту, виконує функції ідентифікації рубрики для певного тексту, використовуючи для цього кількісну статистику появи слів у тексті, отриману за допомогою lemmatizer.php;
- content.php призначений для виведення інформації на сторінку.

Висновки і перспективи подальших наукових розвідок

У роботі інформаційних та аналітичних служб та підприємств доводиться стикатися з великою різноманітністю джерел інформації. Це електронні газети та інші інтернет-ресурси. У цій роботі розглянуто електронні ЗМІ, їхні недоліки, переваги, сервіси. Проведені дослідження електронних ЗМІ дають підстави зробити висновок про недоцільність використання людської праці в процесах, які стосуються формування та рейтингування дайджестів. Ключовою частиною цієї роботи є розроблення методів формування та рубрикації електронних дайджестів. Досвід упровадження системи в різних організаціях показав ефективність і простоту адаптації системи, завдяки розробленому інструменту автоматизованого формування дайджестів та їх рубрикації. Універсальний модуль збирання даних дає змогу повністю автоматизувати введення електронної інформації з різних джерел з її приведенням до єдиного внутрішнього формату, тобто звести до мінімуму рутинну роботу – введення текстових даних. Вбудована система автоматичного спостереження за оновленням вказаних сторінок на інформаційних сайтах в Internet дає змогу автоматизувати і цю частину діяльності інформаційних та аналітичних служб підприємств.

1. Берко А. Ю. Системи електронної контент-комерції: монографія / А. Ю. Берко, В. А. Висоцька, В. В. Пасічник. – Львів: Видавництво Львівської політехніки, 2009. – 612 с. 2. Додонов А. Г. Виявлення категорій і їх взаємозв'язків у рамках технології контент-моніторингу / Додонов А. Г., Ланде Д. В. // Вісник Державної служби України. – 2006. – № 4. – С. 45–52 с. 3. Григорьев А., Ландэ Д. Контент-мониторинг сетевых информационных потоков / Официальный ежемесячный www-регистр бизнес-ресурсов Украины и зарубежья. – Режим доступу: infostream.com.ua/infostream/publ/cbr/index.shtml. *4. Ланде Д. В. Програмно-апаратна система інформаційної підтримки прийняття рішень: наук.-метод. посіб. / Д. В. Ланде, В. М. Фурашев, О. М. Григор'єв. – Київ, 2006. – 48 с. 5. Леліков Г. І. Моніторинг діяльності органів виконавчої влади із застосуванням комп'ютерної системи контент-аналізу електронних ЗМІ / Г. І. Леліков, В. М. Сороко, О. М. Григор'єв, Д. В. Ланде // Вісник Державної служби України. – 2002. – № 2. – С. 72–78 с. 6. Федорчук А. Г. Контент-мониторинг информационных потоков / Б-ки Нац. акад. наук: пробл. Функционирования, тенденции развития. – К., 2005. – Вып. 3. – Режим доступу:* www.nbuv.gov.ua/articles/2005/ 05fagmip.html. *7. Григорьев А. Н.* InfoStream. *Мониторинг новостей из Интернет: технология, система, сервис: научно-метод. пособие / Григорьев А. Н., Ландэ Д. В., Бороденков С. А., Мазуркевич Р. В., Пацьора В. Н. – К., ООО "Старт-98", 2007. – 40 с. 8. Ландэ, Д. В. Поиск знаний в* Internet. *Профессиональная работа: пер. С англ. – М.: Вильямс, 2005. – С. 161 с., 203–205 с.: ил. – Парал. Тит. Англ. 9. Гусев В. Д., Мирошниченко Л. А., Саломатина Н. В. Тематический анализ и квазиреферирование текста с использованием сканирующих статистик / Международная конференция по компьютерной лингвистике – Режим доступу: /www.dialog-21.ru/Archive/2005/Gusev%20Miroshnichenko%20Salomatina/GusevVD.htm. 10. Гиндин С. И. Позиционные методы автоматического фрагментирования текста, их теоретико-текстовые и психолингвистические предпосылки // Семиотика и информатика. – М.: ВИНИТИ, 1978. – Вып. 10. – С. 32–3. 11. Ландэ Д.В. Ловцы данных // СШР/Украина. – 2004. – № 3. – С. 72–75. 12. Ландэ Д. В. Мобильный информатор // СШР/Украина. – 2004. – № 2. – С. 80–83. 13. Ландэ Д. В. На границе стихий // СШР/Украина. – 2003. – 15. – С. 72–77. 14. Ландэ Д. В. Навигация в Сети: каталоги – поисковики – порталы // 1п1ешеША. – 2000. – № 1. – С. 43–47. 15. Ландэ Д. В. О чем говорят запросы пользователей к поисковым серверам // Сети и телекоммуникации. – 1999. – № 4. – С. 19–21. 16. Hearst M. A. Untangling text data mining // Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, 1999. – P. 3–10. 17. Fan W., Wallace L., Rich S. And Zhang Z. Tapping the power of text mining // Communications of the ACM, 2006. – P. 76–82. 18. Christopher D. Manning, Prabhakar Ragavan, Hinrih Schutze. Introduction to Information Retrieval – Cambridge University Press, 2008. – C. 496. 19. Yang Y., Liu X. A re-examination of Text Categorization Methods // Proceedings of the ACM SIGIR, 1999. – P. 42–49. 20. Регулярні вирази і спеціальні символи // Режим доступу:* http://uk.shram.kiev.ua/hacker/regular.shtml.