

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ МЕТОДУ КОНТЕНТ-АНАЛІЗУ ДЛЯ ОПРАЦЮВАННЯ ІНТЕРНЕТ-ГАЗЕТИ

© Кісь Я. П., Чирун Л. Б., Фольтович В. М., 2014

Запропоновано методи аналізу контенту для інтернет-газети. Модель описує процеси опрацювання інформаційних ресурсів у системах аналізу контенту та спрощує технологію автоматизації управління контентом. Проаналізовано основні проблеми синтаксичного та семантичного аналізу контенту та функціональних сервісів управління контентом.

Ключові слова: контент, аналіз контенту, інформаційний ресурс, система управління контентом.

This article is presented the content analysis techniques for online newspapers. The model describes the processing of information resources in content analysis and automation technology simplifies content management. In this paper are analyzed the basic problem of the syntactic and semantic analysis of content and functionality of content management services.

Key words: content, analysis of content, information resource, content management system.

Вступ. Постановка проблеми

За останні десятиліття людство здійснило значний крок у розробленні та впровадженні новітніх технологій [1]. Розвиток технологій дав можливість вирішити багато складних завдань, з якими стикалося людство, але і породив нові задачі, розв'язання яких є складним. Однією з таких задач є задача аналізу контенту. Методи та системи аналізу контенту використовуються в різних сферах людської діяльності (політика, соціологія, історія, філологія, комп'ютерні науки, журналістика, медицина тощо) [2, 3, 5–11]. Ці системи є доволі успішними і не потребують великих коштів та часу на отримання потрібного результату. Водночас використання систем цього типу підвищує рівень успішності продукту на 60 %. Базова система аналізу контенту надає такі можливості: швидке оновлення інформації, пошук інформації на ресурсі, збирання даних про клієнтів та потенційних клієнтів, формування та редагування опитувань, аналіз відвідування ресурсу. Автоматизація системи за допомогою інформаційної системи аналізу контенту зменшує обсяги роботи та час на опрацювання та отримання необхідної інформації, підвищує продуктивність роботи системи, що, своєю чергою, веде до зменшення затрат коштів і часу на отримання потрібного результату [1]. Актуальність тематики викликана зростанням вимог користувачів цих систем та зумовлена такими чинниками: швидкими темпами зростання потреб в достовірній інформації, необхідністю формування множини оперативної інформації, а також використанням для автоматичної фільтрації небажаної інформації.

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

Розвиток технологій Інтернету та його сервісів дав людству доступ до практично необмеженої кількості інформації, але, як часто буває в цих випадках, – виникла проблема щодо достовірності та оперативності. Саме для того, щоб інформація була оперативною та достовірною, запроваджують технології аналізу контенту [1–3, 5–11]. Застосування цих технологій дає змогу системі отримувати інформацію як результат її функціонування, надає можливість оперативного втручання в діяльність системи для підвищення рівня самої системи, діяльності цього інформаційного ресурсу та для підвищення популярності серед користувачів [1]. У цьому напрямі активно працюють провідні світові виробники засобів опрацювання інформаційних ресурсів, зокрема, Google, АІМ, CM Professionals organization, EMC, IBM, Microsoft Alfresco, Open Text, Oracle, SAP [4, 11].

Аналіз останніх досліджень та публікацій

Аналіз контенту – це якісно-кількісний метод вивчення інформації, який характеризується об'єктивністю висновків і строгістю процедури та полягає у квантифікаційному обробленні з подальшою інтерпретацією результатів [1].

Система управління вмістом (англ. *Content Management System, CMS*) – програмне забезпечення для організації веб-сайтів чи інших інформаційних ресурсів у мережі Інтернет чи окремих комп'ютерних мережах [1].

Сьогодні вже існують сотні доступних CMS. Саме завдяки функціональності їх можна використовувати в різних сферах. Незважаючи на широкий вибір інструментальних та технічних засобів, наявних у CMS, властивості всіх систем управління контентом подібні.



Рис. 1. Найпопулярніші CMS

Система управління Web-вмістом (Web Content Management System, або *WCMS*) – програмний комплекс, з функціями створення, редагування, контролю та організації Web-сторінок. WCMS часто використовуються для створення блогів, особистих сторінок та інтернет-магазинів і націлені на користувачів, які мало знайомі з програмуванням [1].

Виділяють такі стадії аналізу [3]:

1. *Підготовка програми аналізу документів.* На цьому етапі, як правило, формулюється так звана емпірична теорія дослідження. Тобто в ході підготовки до проведення аналізу систематизуються гіпотези, які існують в контексті цієї проблематики, та відкидаються ті з них, які не піддаються верифікації на даних інформаційного масиву.

2. *Відбір джерел аналізу.* Необхідно визначити коло джерел, які містять матеріали та інформацію.

3. *Визначення емпіричних моделей аналізу, формування вибірки* (підбір комунікаційних органів, вибір матеріалів за різні періоди часу, визначення видів повідомлень, типу вибірки).

4. *Розроблення методики конкретного аналізу.*

5. *Пілотажне дослідження, перевірка надійності методики.*

6. *Збір первинної емпіричної інформації.*

7. *Кількісне опрацювання зібраних даних.*

8. *Інтерпретація здобутих результатів, висновки дослідження.*

Аналіз контенту є основою журналістики і масової комунікації, що передбачає застосування техніки в таких емпіричних сферах: психіатрії, психології, історії, антропології, освіті, філології та літературному аналізу, лінгвістиці. Загалом застосування методики контент-аналізу в цих сферах так чи інакше пов'язане із застосуванням в межах соціологічних досліджень. Сьогодні аналіз контенту стрімко розвивається, це пов'язано насамперед з розвитком інформаційних та інтернет-технологій, де цей метод широко застосовується.

Методи аналізу контенту

Під час створення ефективної інформаційної системи потрібно звернути особливу увагу на управління контентом, тому що саме аналіз контенту використовується в системах управління контентом для автоматизації роботи, зменшення затрат часу і коштів.

В управлінні контентом існує декілька етапів, а саме: аналіз контенту, опрацювання та подання контенту. Для ефективної роботи системи спочатку виконують аналіз контенту, після цього опрацьовують відповідні результати і роблять висновки, далі – опрацьовують сам контент. І на кінцевому етапі відбувається подання контенту. Аналіз контенту здійснюється за такими методами: аналіз коментарів, рейтингове оцінювання, аналіз статистики та історії [2].

Аналіз коментарів використовується для аналізу, корегування та спостереження за настроями користувачів системи, які у своїх коментарях пишуть відгуки про систему, недоліки та переваги, або для корегування оперативної та ліквідної інформації.

Аналіз статистики та історії слугує для спостереження і опрацювання результатів, які використовуються для визначення оперативності та ліквідності інформації. Наприклад, якщо одну зі статей відвідало 100 користувачів, а іншу – один, то можна з впевненістю сказати, що інформація з першої статті оперативніша, ніж з другої [4].

Рейтингове оцінювання використовується для визначення рейтингу тих самих статей і проводиться за допомогою опитування, оцінювання користувачами тощо (рис. 2).



Рис. 2. Складові частини рейтингового оцінювання статей

Графічний аналіз контенту зумовлений тим, що здебільшого графічну інформацію користувачі засвоюють швидше, ніж текстову. Це можна простежити, наприклад, на поданні діаграм, графіків та гістограм (в табличному вигляді інформація засвоюватиметься повільніше). Застосовуючи контент-аналіз тексту, використовують відповідні методи, причому в цьому випадку виконують два типи аналізу: кількісний та якісний.

Кількісний контент-аналіз має обов'язково містити стандартизовані процедури підрахунку виділених категорій (табл. 1). Для формулювання висновків вирішальне значення мають кількісні величини, які характеризують ту чи іншу категорію. Показники можуть відрізнятися або, навпаки, бути близькими за абсолютним значенням, яке враховуватиметься в інтерпретації результатів опрацювання. Завдання можна ускладнити, якщо поставити попередню умову – виділення всіх змістових у смисловому відношенні одиниць відповідних текстів, а потім підрахувати відносну значущість цього висловлювання порівняно з іншими. В обох випадках основну частину підрахунків можна виконати із застосуванням простих комп'ютерних програм [2].

Таблиця 1

Етапи кількісного контент-аналізу

| Назва етапу | Характеристика етапу |
|----------------------------|---|
| Виділення одиниць аналізу | Перетворення лінгвістичної одиниці на форму для опрацювання |
| Підрахунок частоти одиниць | Виявлення взаємозв'язків між лінгвістичними одиницями |
| Категоризація | Визначення скінченної та надлишкової сукупностей категорій для отримання кількісних даних їх появи |
| Data Mining | Виявлення в потоці контенту за допомогою кількісних багаторазових оцінок нових знань із подальшою кваліфікацією їх як категорій |
| Інтерпретація результатів | Отримання змістових, семантично наповнених результатів з використанням математичних методів та семантичних формалізаторів |

Якісний контент-аналіз націлений на поглиблене змістове вивчення текстового матеріалу, зокрема з погляду контексту, в якому представлені виділені категорії (табл. 2). Підсумки формулюються тут з урахуванням взаємозв'язків змістових елементів і їх відносної значущості у структурі тексту. Залежно від завдань дослідження якісний контент-аналіз може бути доповнений деякими елементами кількісного контент-аналізу [2].

Етапи якісного контент-аналізу

| Назва етапу | Характеристика етапу |
|-------------------------------|--|
| Розбиття тексту на блоки | Формування інтегрованих змістових одиниць для кодування і опрацювання. |
| Реконструкція потоку контенту | Реконструкція системи значень, думок, поглядів і доказів кожного джерела тексту. |
| Формування висновків | Виведення узагальнень через порівняння індивідуальних системних значень. |

Основні етапи застосування контент-аналізу текстової інформації [5]:

1. *Визначення сукупності джерел*, що досліджуються, або повідомлень відповідно до заданих критеріїв, яким відповідає кожне повідомлення: тип джерела (форум, е-пошта, чат, інтернет-газета, інтернет-журнал); тип повідомлень (стаття, електронний лист, банер, коментар); сторони, що беруть участь в процесі комунікації (відправник, одержувач, реципієнт); розмір повідомлень, що порівнюють (мінімальний обсяг/довжина); частота появи повідомлень; спосіб поширення повідомлень; місце розповсюдження повідомлень; час появи повідомлень тощо.

2. *Формування обмеженої вибіркової сукупності повідомлень*.

3. *Виявлення лінгвістичних одиниць аналізу*. Існують чіткі вимоги до вибору можливої лінгвістичної одиниці аналізу: достатньо велика для інтерпретації значення; достатньо мала, щоб не інтерпретувати багато значень; легко ідентифікується; кількість одиниць достатньо велика для формування вибірки. У разі прийняття за одиницю аналізу теми враховують такі правила: розмір теми не виходить за межі абзацу; нова тема виникає у разі зміни мети теми, призначення, категорії та особи, для якої створюється тема.

4. *Виділення одиниць обчислення*, які можуть збігатися зі змістовими одиницями або мати специфічний характер. У першому випадку процедура аналізу зводиться до підрахунку частоти появи виділеної змістової одиниці, в іншому – дослідник на основі аналізованого матеріалу і цілей дослідження вибирає одиниці обчислення, якими можуть бути: фізична протяжність текстів; площа тексту, заповнена змістовими одиницями; кількість рядків (абзаців, знаків, колонок тексту); розмір та вид файлу; кількість рисунків з певним змістом, сюжетом тощо. В деяких випадках дослідники використовують й інші елементи обчислення. Принципове значення на цьому етапі контент-аналізу має строге визначення його операторів.

5. *Безпосередньо процедура обчислення*. У загальному вигляді схожа на стандартні прийоми класифікації за виділеними групуваннями з формул математичної статистики та теорії ймовірності. Існують також спеціальні процедури підрахунку стосовно контент-аналізу.

6. *Інтерпретація отриманих результатів* відповідно до цілей і завдань конкретного дослідження. На цьому етапі виявляють і оцінюють такі характеристики текстового матеріалу, які дають змогу зробити висновки про те, що хотів підкреслити або приховати його автор [5].

Виділення проблем

Аналіз контенту – це якісно-кількісний метод вивчення інформації, який характеризується об'єктивністю висновків та строгістю процедури і полягає у квантифікаційному опрацюванні з подальшою інтерпретацією результатів. Оскільки за останні декілька десятків років людство зробило значний крок у розробленні та впровадженні новітніх технологій, це породило нові завдання, які складно вирішити. Одним з цих завдань є подання для користувачів достовірної та оперативної інформації.

Оперативність – властивість інформації, яка полягає в тому, що час її збирання та опрацювання відповідає динаміці зміни ситуації [6].

Достовірність – властивість інформації бути правильно сприйнятою, ймовірність відсутності помилок, безсумнівна правильність наведених відомостей, які сприймає людина. Отже, достовірність – не те саме, що істинність. Відомості можуть бути достовірними або недостовірними для того, хто їх сприймає, а не взагалі [6].

Формулювання мети

Контент є основою інтернет-газети, за якою користувач шукає необхідну інформацію. Тому потрібно, щоб інформація була оперативною і доступною для користувача. Наприклад, якщо текст міститиме велику кількість визначень і термінів або формул, користувачеві буде важко і нецікаво читати. З іншого боку, якщо текст міститиме велику кількість непотрібної інформації, користувач витратить більше часу на його читання. Виникає необхідність в розв'язанні актуальної наукової задачі автоматичного опрацювання контенту інтернет-газети для отримання оперативної та доступної для користувача інформації з ліквідацією інформаційного шуму та зменшенням часу на процес формування кінцевого результату пошуку даних.

Аналіз отриманих наукових результатів

Рейтингове оцінювання статей проводиться за допомогою трьох критеріїв. Підраховують кількість звернень, час читання статті та користувацьку оцінку. Ці три критерії формують рейтингове оцінювання статей (рис. 3).

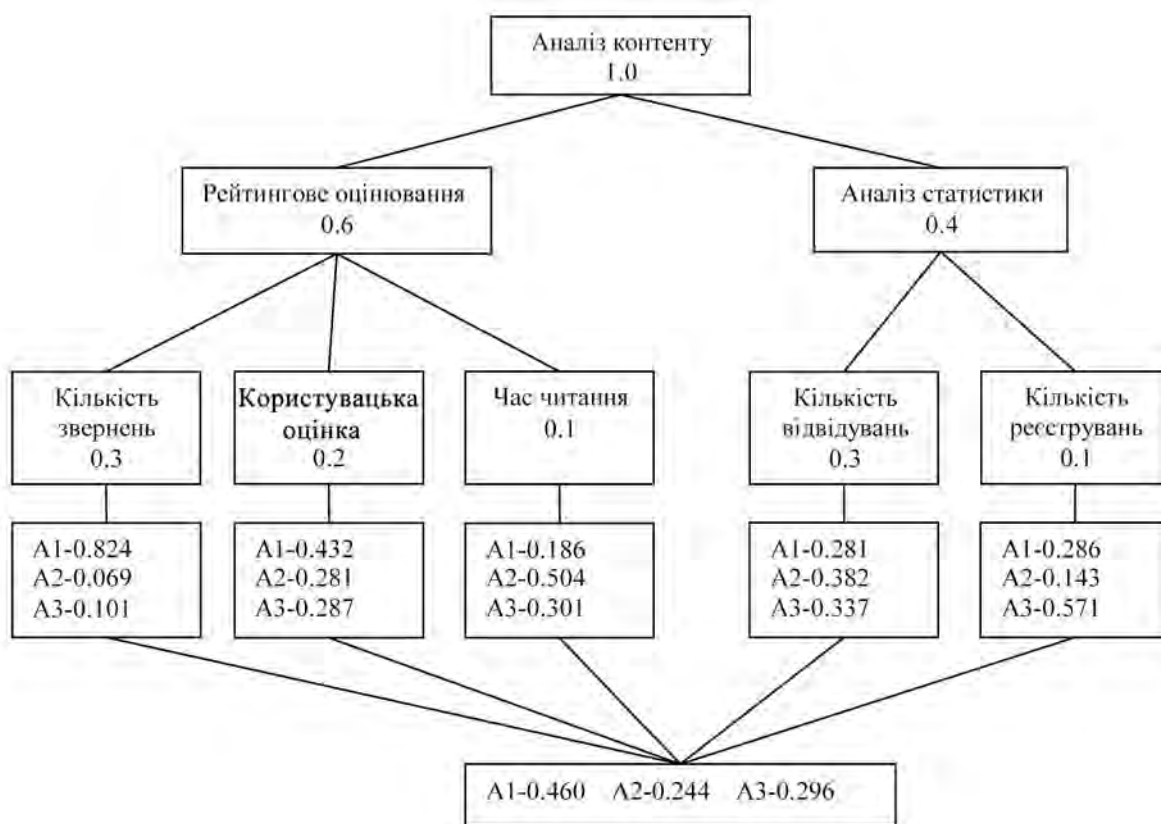


Рис. 3. Ієрархія у випадку трьох альтернатив для аналізу контенту

На рис. 3 продемонстровано ієрархію у випадку трьох альтернатив аналізу контенту. Ієрархія містить чотири рівні: мета, критерії вищого та нижчого рівнів та альтернативи. Числа на рис. 3 показують пріоритети елементів ієрархії з погляду мети, які обчислюються за допомогою МАІ на основі парних порівнянь елементів кожного рівня щодо пов'язаних з ними елементів вищого рівня. Пріоритети альтернатив обчислюють на завершальному етапі методу лінійним згортанням локальних пріоритетів усіх елементів. Відомі пріоритети критеріїв вищого рівня надають найбільшої ваги критерію *Рейтингове оцінювання*, оскільки він об'єктивно відображає якість роботи. Критерію *Аналіз статистики* надають меншу вагу, оскільки він не повною мірою відображає якість роботи. У критеріях нижчого рівня надають найбільшої ваги критерію *Кількість звернень* та *Кількість відвідувань*, які відображають міру зацікавленості користувачів у конкретному матеріалі. Наступним за вагою критерієм є *Користувацька оцінка*, який відображає

оцінку користувачів для конкретного матеріалу. Найменшої ваги надають критеріям *Час читання* та *Кількість реєструвань*. Так оцінюють всі альтернативи окремо за кожним з критеріїв. З розрахунків випливає, що перша альтернатива найкраща за вибраними критеріями оцінювання. Дуже важливою обставиною для функціонування системи є наявність вхідних даних. У нашому випадку вхідними даними є статті. Стаття – це публіцистичний чи науковий твір, що на підставі розгляду та зіставлення великої групи фактів чи ситуацій, ґрунтовно й глибоко, з науковою точністю трактує, осмислює й теоретично узагальнює проблеми соціальної дійсності. Додання, редагування та видалення статей здійснює адміністратор. Джерелом інформації для статей слугують енциклопедії, періодичні видання, книги, інші статті тощо. Наступною важливою обставиною є наявність користувачів. Користувачі – це фізичні особи, які користуються інформаційним ресурсом, здійснюють пошук статей, читають їх та голосують. Без користувацького аспекту неможлива перевірка якості контенту.

Діаграма варіантів використання аналізу контенту інтернет-газети

Першим етапом проектування концептуальної моделі інформаційної системи є побудова діаграми варіантів використання. На наступних етапах будуть побудовані діаграма послідовностей, кооперації, діяльності, розгортання та компонентів. Мова UML являє собою загальноцільову мову візуального моделювання, яка розроблена для специфікації, візуалізації, проектування і документування компонентів програмного забезпечення, бізнес-процесів та інших систем (рис. 4).

На рис. 4 відображено діаграму варіантів використання аналізу контенту інтернет-газети. Суть цієї діаграми така: проектована система подається у вигляді множини сутностей, або акторів, що взаємодіють із системою за допомогою так званих варіантів використання. Ця діаграма відображає інформаційну систему аналізу контенту інтернет-газети “Акваріумістика”. Як можна побачити з діаграми, у нашій системі є два актори: актор *Користувач* і актор *Адміністратор*. Актор *Користувач* призначений для моделювання користувача системи як фізичної особи, яка здійснює певні дії. Актор *Користувач* реєструється/авторизується в системі, голосує за переглянуті статті, переглядає статті, здійснює пошук статей. Актор *Адміністратор* – це фізична особа, яка аналізує та систематизує інформацію, додає, редагує та видаляє статті, а також здійснює інші дії. Актор *Адміністратор* здійснює адміністрування нашої системи. Адміністрування передбачає вилучення некоректних статей та додавання нових статей. Варіанта використання *реєстрація* використовується для реєстрації користувачів у системі. Варіанта використання *авторизація* використовується у системі для авторизації користувачів. Варіанти використання *авторизація* та *реєстрація* розширюють варіанту використання *аналіз статистики*.

Як можна побачити на рис. 4, варіанта використання *аналіз статистики* використовується для проведення статистичного оцінювання, щоби можна було відображати кількість користувачів системи під час виконання цієї варіанти. Можна зробити висновки, наприклад: що більше користувачів користується системою, то популярніша система. Кожен користувач може переглядати статті за допомогою варіанти використання *перегляд статей*. Також користувач може виконувати пошук потрібних йому статей за допомогою варіанти використання *пошук статей*, проводити голосування за допомогою варіанти використання *голосування*. До складу варіанти використання *голосування* входять також варіанта використання *користувацька оцінка*, яка застосовується для рейтингового оцінювання статей.

Актор *Адміністратор* виконує адміністрування цього інформаційного ресурсу за допомогою варіанти використання *адміністрування* (додає, редагує та видаляє статті). Варіанта використання *адміністрування* містить дві варіанти: *вилучення та розміщення статей*. Варіанти використання *вилучення та розміщення статей* містять варіанту використання *зміна даних про статті*. Варіанта використання *зміна даних про статті* використовується для того, щоб в системі завжди були оперативні дані про останні статті. Варіанта використання *виведення останніх статей* застосовується для виведення останніх статей, а також розширює варіанту використання *пошук статей*.

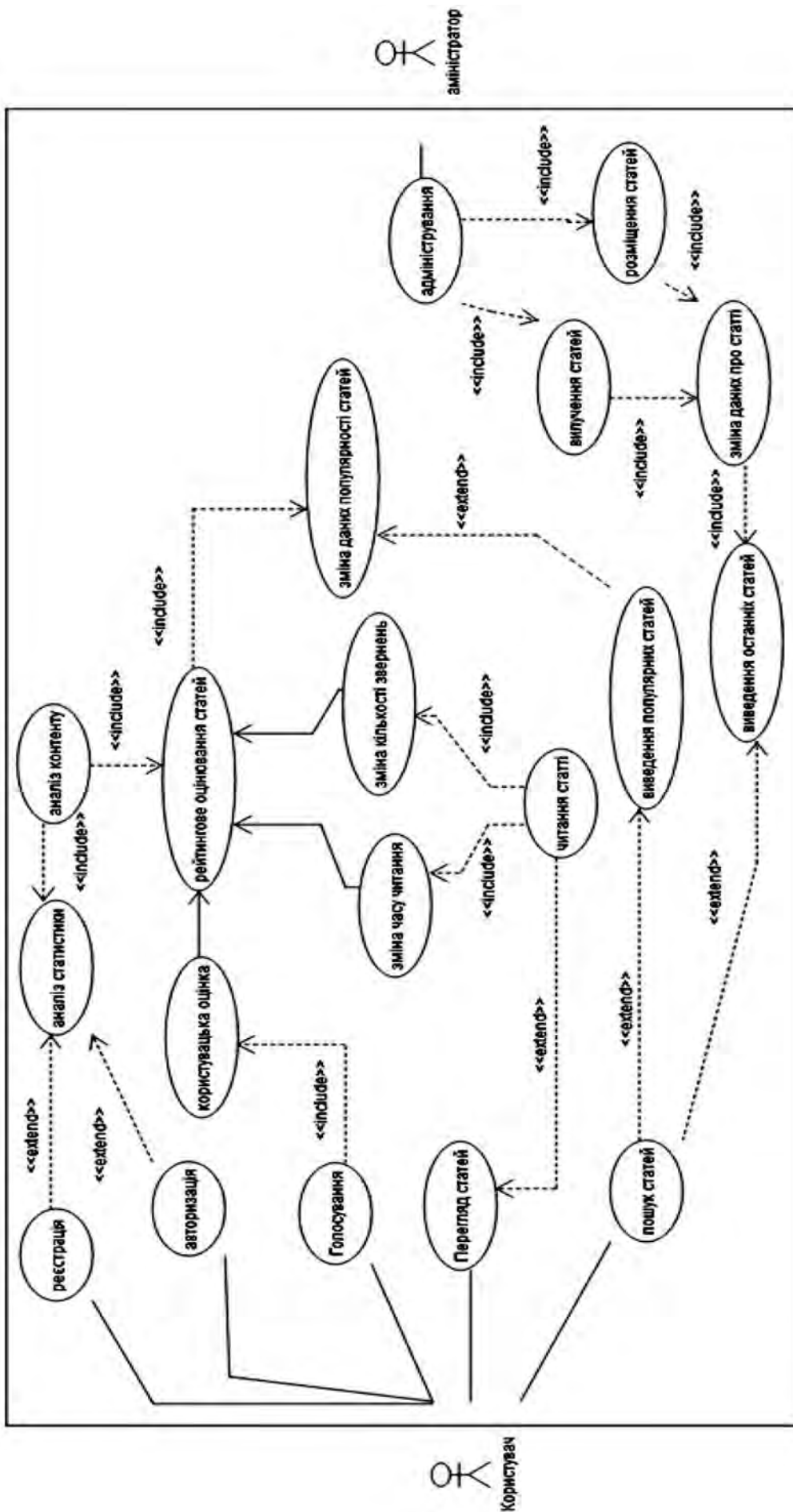


Рис. 4. Діаграма варіантів використання

Варіанта використання *аналіз контенту* застосовується для аналізу контенту інтернет-газети. Ця діаграма містить дві варіанти: *аналіз статистики* та *рейтингове оцінювання*. У варіанту використання *рейтингове оцінювання* входить варіанта використання *зміна даних популярності статей*. Після проведення рейтингового оцінювання вносяться зміни щодо популярності статей. Варіанта використання *виведення популярних статей* розширює варіанту використання *зміна даних популярності статей*. Варіанта використання *пошук статей* розширює варіанту використання *виведення популярних статей*. Варіанта використання *читання статей* застосовується для читання статей користувачем. Як видно з рис. 4, варіанта використання *читання статей* містить варіанту використання *зміна часу читання та зміна кількості звернень*. Варіанта *Зміна часу читання* застосовується для відображення часу читання статті. Варіанта *зміна кількості звернень* застосовується для відображення кількості відкривань статті. Ці дві варіанти використання входять до складу варіанти використання *рейтингове оцінювання*. Варіанта використання *читання статті* розширює варіанту використання *перегляд статей*.

Діаграма послідовності

На рис. 5 відображена діаграма послідовності. Ця діаграма демонструє взаємодію об'єктів, яка впорядкована за часом їх виконання. Такі діаграми відображають задіяні об'єкти та послідовність відправлених повідомлень. Система аналізу контенту інтернет-газети "Акваріумістика" містить велику кількість варіантів використання. Об'єкт *користувач* використовується на діаграмі послідовності для відображення дій користувача як фізичної особи. Користувач здійснює авторизацію за допомогою об'єкта *авторизація*. Під час виконання авторизації об'єкт *користувач* пересилає дані для авторизації. Після цього об'єкт *авторизація* звертається до об'єкта *база даних* і відсилає запит на отримання даних. Після одержання даних об'єкт *авторизація* виконує дії та надсилає результат. Якщо результат правильний, то авторизація успішна, якщо ж результат негативний, то авторизація не вдалася. Як можна побачити з рис. 5, після того, як користувач провів авторизацію, він звертається до об'єкта *статті* та виконує читання статей. Інформацію отримують у вигляді результату читання статті. Після читання статей інформація від об'єкта *статті* пересилається до об'єкта *рейтингове оцінювання* для проведення оцінювання. Після закінчення операції рейтингового оцінювання оновлені дані про популярність статей заносять у базу даних.

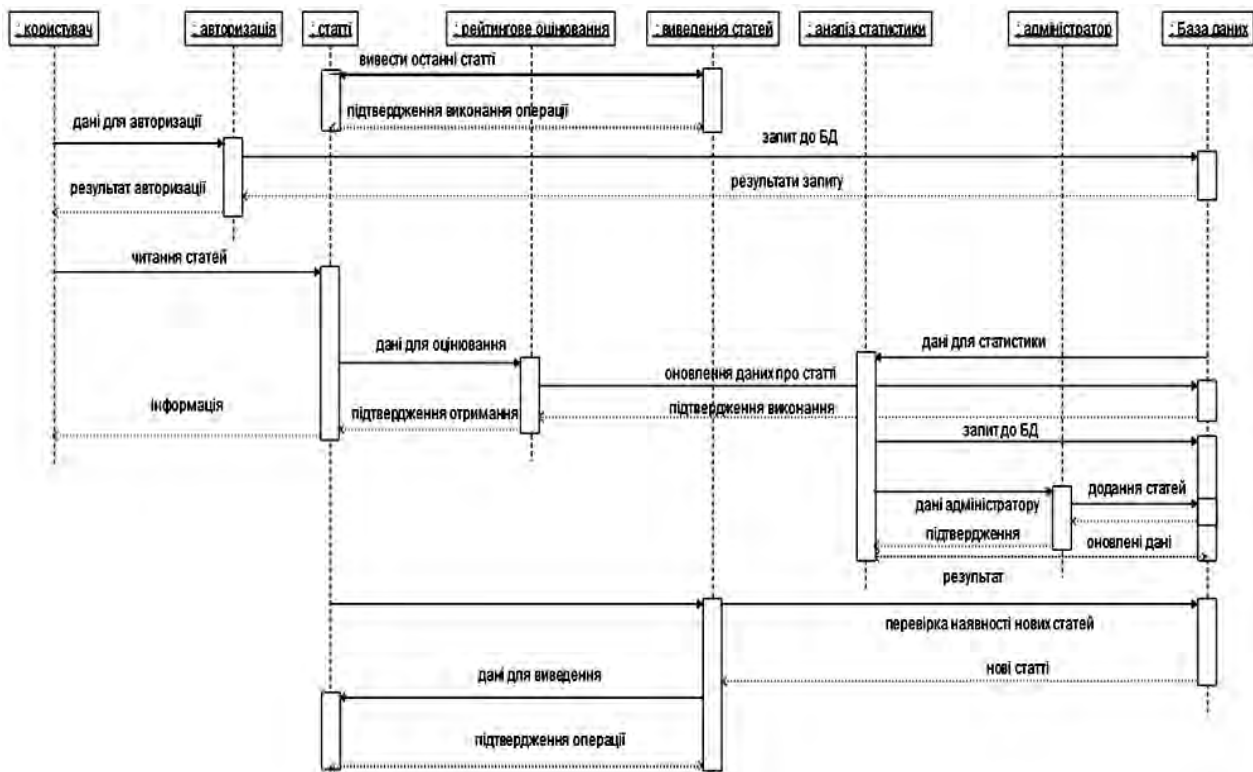


Рис. 5. Діаграма послідовності

Об'єкт *аналіз статистики* використовується для проведення статистичного оцінювання. Для цього об'єкту надсилають дані з бази даних. Після проведення оцінювання результати пересилаються об'єкту *адміністратор* для того, щоб цей об'єкт зміг зробити висновки, які будуть потрібні для додавання нових статей. Після додавання нових статей або виведення статей об'єкту передають дані, які потрібні для виведення останніх та популярних статей. Також цей об'єкт виконує перевірку наявності нових статей. Результатом проведення перевірки наявності нових статей є нові статті, які передаються об'єкту для виведення. Цей об'єкт виводить нові статті й отримує результат – підтвердження операції.

Діаграми послідовності та кооперації

На рис. 6 відображена діаграма кооперації. За допомогою діаграми кооперації можна описати повний контекст взаємодій як своєрідний часовий зріз сукупності об'єктів, що взаємодіють між собою для виконання певного завдання або бізнес-мети програмної системи. На відміну від діаграми послідовності, на діаграмі кооперації зображені тільки відношення між об'єктами, що відіграють певні ролі у взаємодії. З іншого боку, на цій діаграмі не вказується час у вигляді окремого виміру. Тому послідовність взаємодій і паралельних потоків можна визначити за допомогою порядкових номерів. Діаграма відображає детальні дії, які виконує кожен об'єкт стосовно іншого об'єкта, а також вказує тип зв'язку між цими об'єктами. Об'єкт *користувач* використовується на діаграмі кооперації для відображення дій користувача як фізичної особи. У цій системі є дві фізичні особи: об'єкт *користувач* та об'єкт *адміністратор*, які виконують притаманні їм дії. Користувач здійснює авторизацію за допомогою об'єкта *авторизація*. Виконуючи авторизацію, об'єкт *користувач* пересилає дані для авторизації. Після цього об'єкт *авторизація* звертається до об'єкта *база даних* і відсилає запит на отримання даних. Після отримання даних об'єкт *авторизація* виконує дії та відсилає результат. Якщо результат правильний, то авторизація успішна, якщо ж негативний, то авторизація не вдалася. Як можна побачити з рис. 6, після того, як користувач здійснив авторизацію, він звертається до об'єкта *статті* та виконує читання статей, отримуючи інформацію у вигляді результату читання статті. Після читання статей дані для оцінювання від об'єкта *статті* пересилаються об'єкту *рейтингове оцінювання* для проведення оцінювання. Після закінчення операції рейтингового оцінювання оновлені дані про популярність статей заносяться в базу даних.

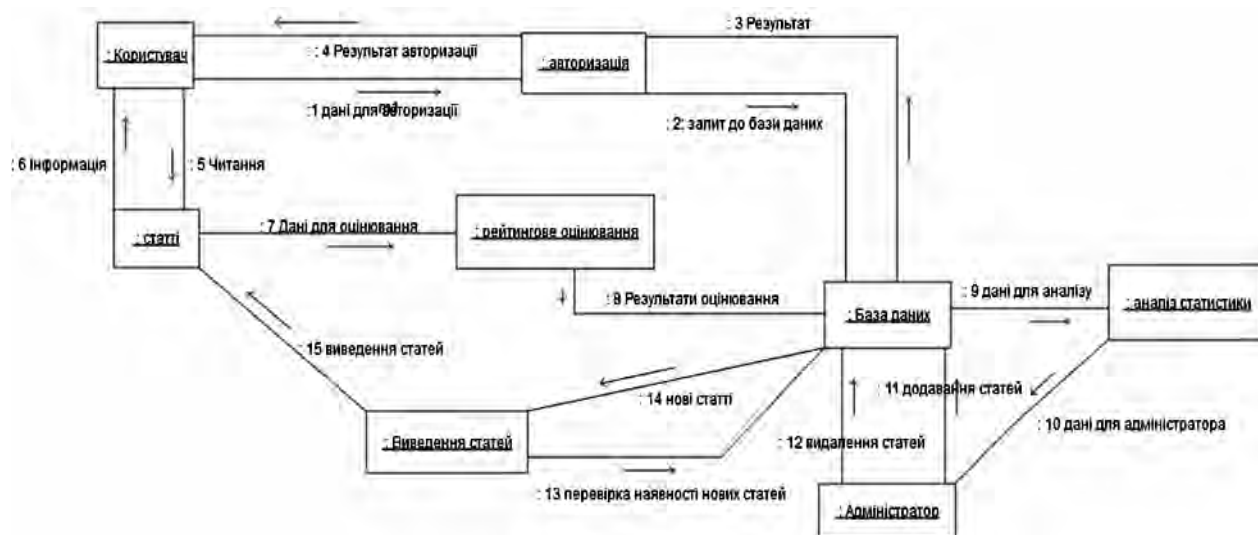


Рис. 6. Діаграма кооперації

Об'єкт *аналіз статистики* використовують, щоб провести статистичне оцінювання, для цього об'єкту надсилають дані з бази даних. Після проведення оцінювання результати пересилаються об'єкту *адміністратор* для того, щоб цей об'єкт зміг зробити висновки, які будуть потрібні для додавання нових статей або видалення непотрібних. *Адміністратор* – єдиний об'єкт у системі, який може додавати або видаляти статті. Після додавання нових статей або видалення об'єкту *виведення статей* передаються дані, які потрібні для виведення останніх та популярних статей. Також об'єкт виконує

перевірку наявності нових статей, якщо йому не надані дані. Результатом проведення перевірки наявності нових статей є нові статті, які передаються об'єкту для виведення. Цей об'єкт виконує виведення нових статей. Під час побудови діаграми кооперації визначено об'єкти системи, проведено аналіз їх взаємодії. Як видно з діаграми кооперації (рис. 6), у системі існує п'ятнадцять паралельних потоків. Кожен потік переправляє відповідні повідомлення та дані від об'єкта до об'єкта.

Діаграми діяльності для процесу контент-аналізу

На рис. 7 відображена діаграма діяльності, що є окремим випадком діаграми станів. Вона дає змогу реалізувати особливості процедурного і синхронного керування, зумовленого завершенням внутрішніх процесів системи. На цій діаграмі відображена діяльність системи після авторизації користувача. На рис. 7 бачимо, що після авторизації користувач розпочинає перегляд статей, що має на меті продемонструвати користувачу системи статті, які наявні в цій системі. Після проведення дії *перегляд статей* система переходить у наступний з двох станів дій *виведення популярних статей* та *виведення останніх статей* залежно від вибору клієнта системи. Після того, як користувач переглянув статті, здебільшого він здійснює вибір конкретної статті. За це відповідає стан дії *вибір конкретної статті*. Після вибору конкретної статті користувач переходить до стану дії *читання статті*. Це основний стан дії на цьому етапі, тому що саме в цьому стані дій користувач читає статтю, отримуючи інформацію. Після переходу до стану дії *читання статті* система переходить в два наступні стани. Це стан *виведення вибраної статті* та стан дії *рейтингове оцінювання*. Стан дії *рейтингове оцінювання* є одним з найважливіших станів. Цей стан має три підстани, в які переходить система після стану *рейтингове оцінювання*. Це стан *зміна кількості звернень*. У цьому стані змінюється інформація про кількість звернень до цієї статті. Наступним станом дії є стан дії *зміна часу читання*, у цьому стані змінюється час читання статті. Наступним станом дії є стан *користувацького оцінювання*, в якому користувач голосує за відповідну статтю. Ці три стани дії використовують для рейтингового оцінювання. Виконавши потрібні дії, система переходить у стан дії *закриття статті*. Цей стан характеризується збереженням даних та результатів. Після цього система переходить в кінцевий стан.

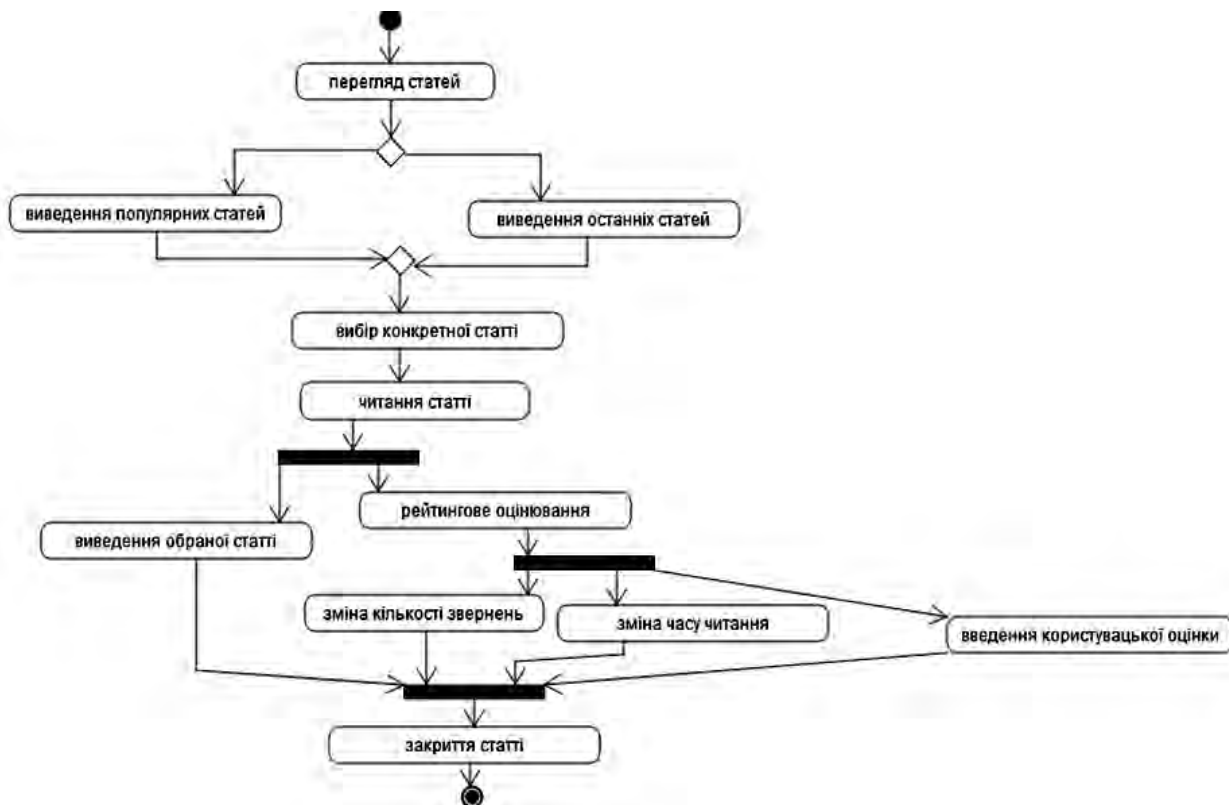


Рис. 7. Діаграма діяльності

На рис. 8 подано процес контент-аналізу. Після введення адміністратором статті система переходить в стан дії *контент-аналіз*, виконавши який, система переходить у наступний стан *розбиття тексту на блоки*. В цьому стані дії виконується розподіл контенту на складові частини (блоки) та виділяється одиниця аналізу (перетворення лінгвістичної одиниці на форму для опрацювання). Далі виконується аналіз фрагмента, підрахунок частоти одиниць та виявлення взаємозв'язків між лінгвістичними одиницями. Після виконання всіх дій виконується інтерпретація результатів контент-аналізу.

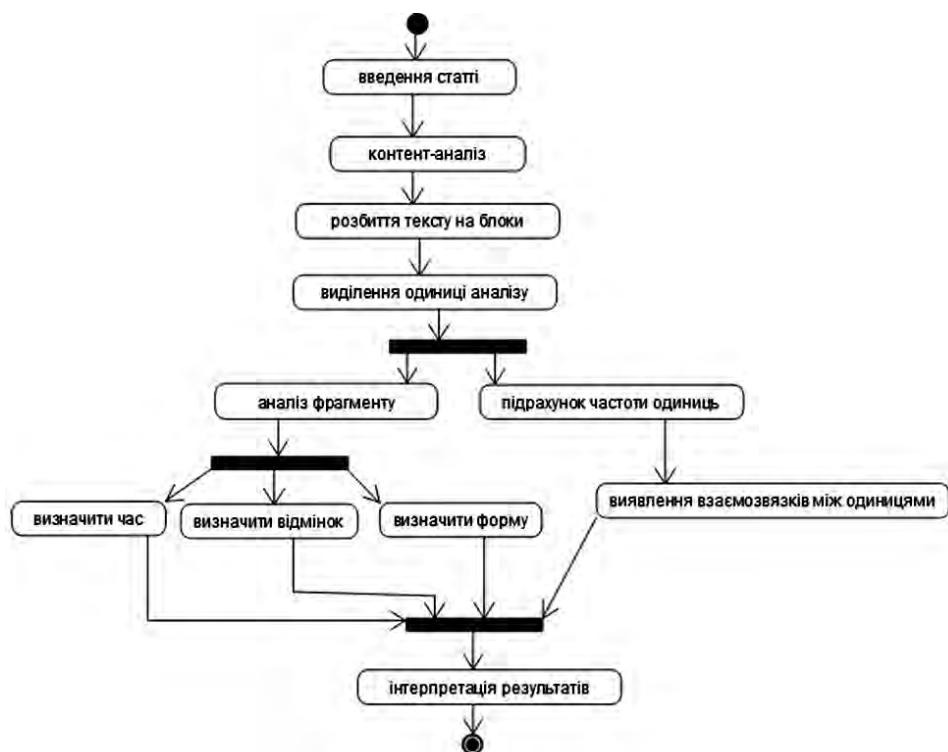


Рис. 8. Діаграма діяльності для процесу контент-аналізу

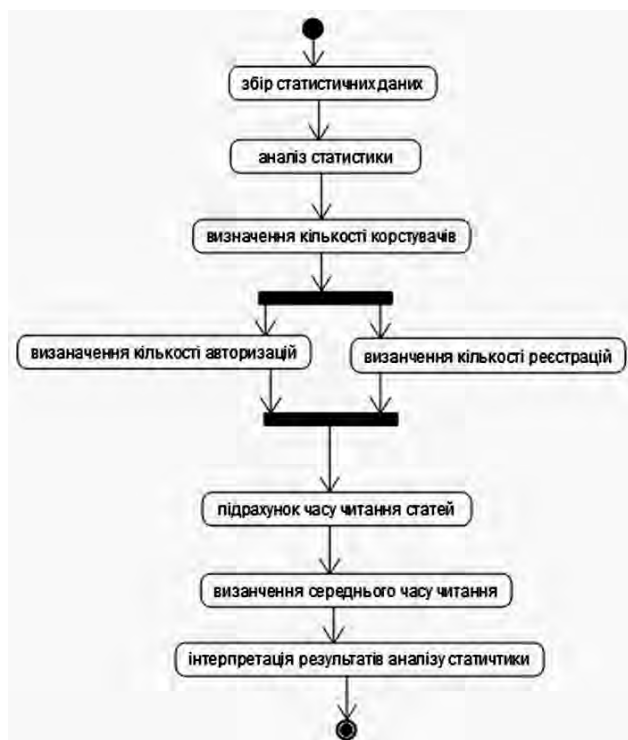


Рис. 9. Діаграма діяльності для процесу аналізу статистики

На рис. 9 зображена діаграма діяльності для процесу аналізу статистики. Аналіз статистики виконується після того, як система здійснить збір даних для статистики. Основною з дій є дія *визначення кількості користувачів системи*. Також визначається кількість авторизацій та реєструвань. Далі система визначає кількість часу читання статей та середній час читання. Після завершення всіх дій виконується *інтерпретація результатів аналізу статистики*.

Діаграма компонентів

На рис. 10 відображена діаграма компонентів, що дає змогу визначити архітектуру системи, яка розробляється, встановивши залежності між програмними компонентами, якими може виступати початковий, бінарний і виконуваний код. За допомогою інтерфейсу користувач здійснює свою авторизацію у системі, після цього відбувається запуск

файлів і бібліотек, які необхідні користувачу. Однією з компонент інформаційної системи аналізу контенту інтернет-газети є компонента, яка виконує аналіз контенту за допомогою відповідних класів. Наприклад, клас `login` використовується для реєстрації/авторизації користувача в системі, а клас `view` – для виведення статей. Клас `category` слугує для того, щоб можна було розділити статті за категоріями для кращого відображення інформації та полегшення її пошуку. Клас `ACore` є логічним ядром системи. Наступною компонентою є компонента *база даних*, яка відображає базу даних і використовується для з'єднання з базою даних, а також для обміну інформацією. Компонента *опрацювання статей* використовується для виконання потрібних дій над статтями. Так, за допомогою класу `statti` виконується додавання, редагування та видалення статей, а за допомогою класу `main` здійснюється передавання нових статей та їх виведення.

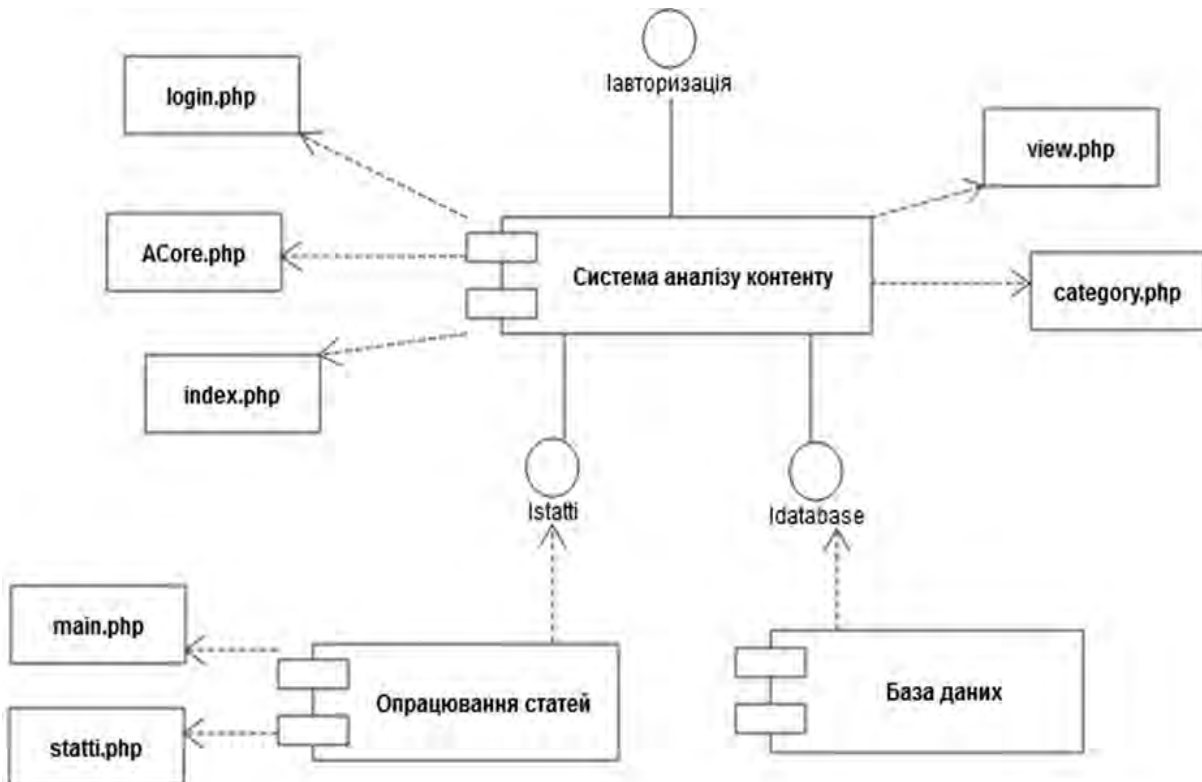


Рис. 10. Діаграма компонентів

Діаграма розгортання

На рис. 11 відображена діаграма розгортання, призначена для візуалізації елементів та компонентів програми, що існують на етапі її виконання. Ця діаграма показує, з яких частин складається система аналізу контенту інтернет-газети. У системі наявні такі вузли: користувач, інформаційний ресурс, система аналізу контенту та Web-браузер. Вузол *користувач* представляє фізичну особу, яка виконує певні дії на вузлі *інформаційний ресурс*, використовуючи вузол *Web-браузер*. Вузол *система аналізу контенту* виконує аналіз даних інформаційного ресурсу. Вузол *система аналізу контенту* має такі компоненти: *опрацювання статей*, СУБД, аналіз контенту, сервер. Компонента *аналіз контенту* використовується для проведення аналізу, а компонента *опрацювання статей* – для додавання, редагування та видалення статей. Компонента *СУБД* відображає систему управління баз даних, яка використовується для збереження даних. Компонента *аналіз контенту* призначена для проведення аналізу з використанням різних методів для отримання відповідних результатів. Компонента *сервер* застосовується як ресурс, який надає свої обчислювальні ресурси для системи.

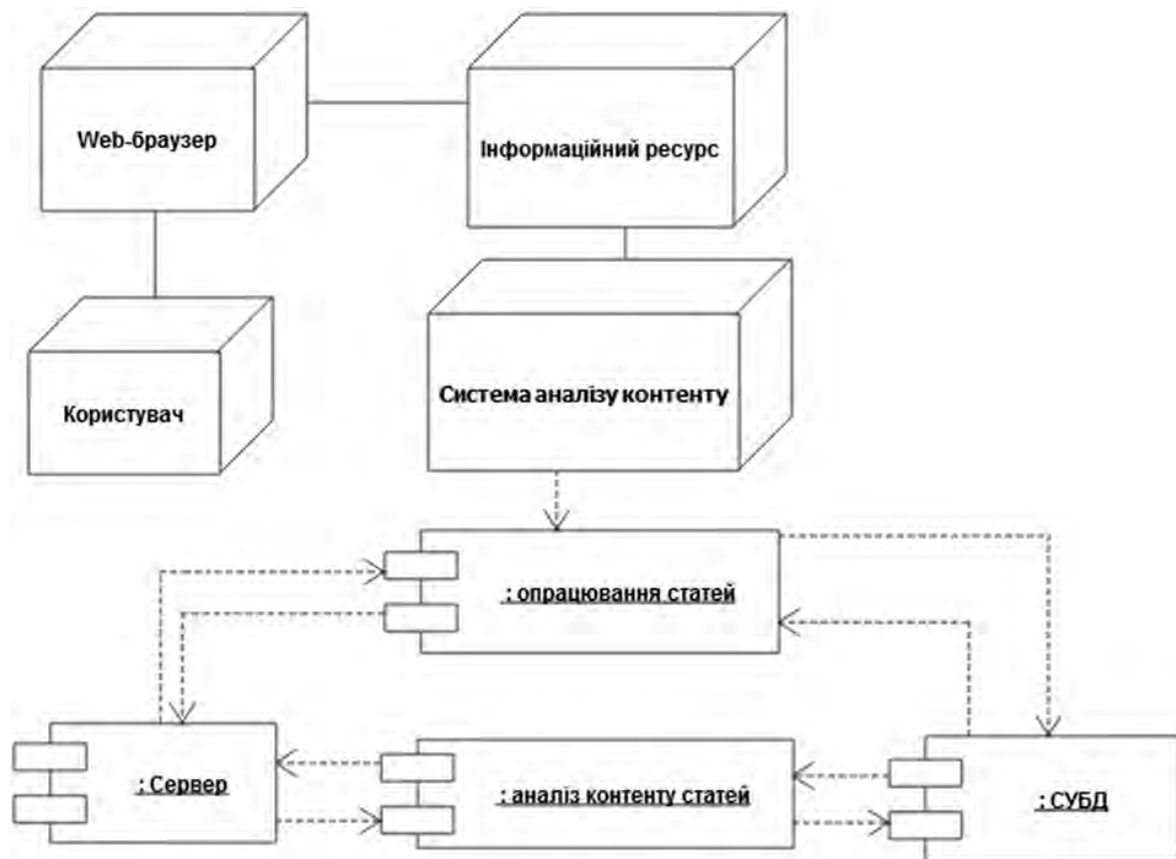


Рис. 11. Діаграма розгортання

Висновки і перспективи подальших наукових розвідок

Контент у вигляді статей є основою інтернет-газети, за якою користувач шукає необхідну йому інформацію. За допомогою методу аналізу контенту власник системи може визначити достовірність та оперативність інформації, що розміщена в статтях інтернет-газети на інформаційному ресурсі. Завдяки цьому можна визначити популярність газети та виконати певні дії для збільшення кількості користувачів, підвищення популярності тощо. Розроблено загальні рекомендації з проектування архітектури систем аналізу контенту, які відрізняються від тих, що існують, більшою деталізацією етапів та наявністю модулів опрацювання інформаційних ресурсів, що дають змогу ефективно та просто опрацьовувати інформаційні ресурси на рівні розробника систем.

1. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л: НУЛП, 2009. – 612 с.
2. Иванов В. Контент-аналіз: Методологія і методика дослідження ЗМК / В. Иванов. – Київ, 1994. – 112 с.
3. Иванов С. Статистический анализ документальных информационных потоков / С. Иванов, Н. Круковская // Научно-техническая информация. – 2004. – № 2. – С. 11–14.
4. Клифтон Б. Google Analytics / Б. Клифтон. – М: ООО "И. Д. Вильямс", 2009. – 400 с.
5. Ландэ Д. Основы моделирования и оценки электронных информационных потоков / Д. Ландэ, В. Фурашев, С. Брайчевский, О. Григорьев. – К.: Інжиніринг, 2006. – 348 с.
6. Пасічник В. Математична лінгвістика / В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич. – Л: Новий Світ, 2012. – 359 с.
7. Солтон Д. Динамические библиотечно-информационные системы / Д. Солтон. – М.: Мир, 1979. – 560 с.
8. Федорчук А. Контент-мониторинг информационных потоков // БНАН. – Киев, 2005. – № 3. Режим доступу: www.nbuv.gov.ua/articles/2005/05fagmip.html.
9. Voiko V. Content Management Bible. – Hoboken, 2004. – 1176 p.
10. CM Lifecycle Poster / Content Management Professionals. – Режим доступу: <http://www.cmprosold.org/resources/poster/>.
11. CMIS. Part I – Introduction, General Concepts, Data Model, and Services / EMC, IBM and Microsoft Corporation. – 2008. – 76 p.