УДК 004.738.5

Л.В. Чирун, В.А. Висоцька
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ЗАСТОСУВАННЯ КОНТЕНТ-АНАЛІЗУ ТЕКСТОВОЇ ІНФОРМАЦІЇ В СИСТЕМАХ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

Запропоновано застосовувати метод контент-аналізу текстової інформації в системах електронної комерції для автоматизації електронного бізнесу і підтримки прийняття рішень відповідною особою. Проаналізовано основні проблеми електронної комерції та функціональних сервісів керування контентом. Запропоновано методи вирішення цих проблем.

Ключові слова: контент, контент-аналіз, електронна комерція, кількісний контент-аналіз, якісний контент-аналіз, функціональні сервіси керування контентом.

In the given article authors proposed to apply the technique of content analysis of textual information in electronic commerce systems for automation of e-business and decision making by appropriate person. In the works article main problems of electronically commercial and content management interoperability services are analyzed. New methods for solution of discussed problems are proposed.

Keywords: content, content analysis, electronic commerce, quantitative content analysis, qualitative content analysis, content management interoperability services.

Вступ. Загальна постановка проблеми

Сьогодні рівень складності завдань, які вирішують за допомогою систем електронної комерції, постійно росте: від автоматизованого збору, інтерпретації та реалізації інформації/контенту до управління, проектування, моделювання і прогнозування різноманітних бізнес-процесів. Наприклад, прогнозування системою електронної комерції зміни попиту на визначений вид продукції залежно від тематичного обговорення користувачами на форумах або аналіз масивів коментарів до цієї продукції. Але більшість складних завдань вирішуються не системами електронної комерції автоматично, а особами (модераторами), що супроводжують процес роботи відповідної системи. Виникла необхідність пошуку нетрадиційних підходів використання інформаційних технологій та математичних методів в процесах підготовлення й прийняття рішень в системах електронної комерції на основі отриманої інформації від форумів, коментарів користувачів продукції, електронного листування, результатів роботи пошукових систем та агентів.

Зв’язок висвітленої проблеми із важливими науковими та практичними завданнями

Сьогодні більшість контенту, отриманого системою електронної комерції від користувачів, опрацьовується модераторами. Наприклад, великі текстові масиви коментарів до продукції, представленої на сайті, або масиви контенту на форумі, фільтруються в більшості модераторами за певний період часу визначеним колом довірених осіб. Для підвищення попиту на продукцію необхідно викладати на сайті лише відфільтровані коментарі користувачів, де відсутні нецензурна лексика (такі коментарі блокуються) та реклама продукції конкурентів. Крім того, детальний аналіз сукупності коментарів на продукцію дає можливість виробнику аналізувати ситуацію на ринку та корегувати попит на власну продукцію. Вивчення інформаційних потреб для функціонування аналогічних систем свідчить, що в процесі прийняття рішень поряд із використанням фактографічних матеріалів, важливим стає залучення текстової інформації про відображення взаємовідношень поміж різними фактами, подіями, особами тощо та автоматичне її опрацювання за короткий проміжок часу з мінімальними фінансовими затратами [1–6].

Актуальність теми зумовлена такими чинниками:

- швидкі темпи росту потреб у достовірній інформації для ведення бізнесу;
- необхідність формування множини оперативної інформації/контенту для автоматизації процесів електронного бізнесу;
- необхідність автоматичної фільтрації небажаного контенту на сайтах;
- зменшення часу опрацювання інформаційних ресурсів системи електронної комерції;
- необхідність автоматизованого збору, інтерпретації та реалізації контенту;
- необхідність в автоматизованому управлінні, проектуванні, моделюванні і прогнозуванні різноманітних бізнес-процесів.

Мета і задачі дослідження полягають в застосуванні аналітичного методу опрацювання інформаційних ресурсів системи електронної комерції у вигляді текстової інформації для можливості подальшого керування бізнес-процесами. Керування бізнес-процесами – важливий етап життєвого циклу контенту. Сьогодні на перший план вийшла актуальність та точність контенту (необхідність самої останньої інформації по визначеному питанню), тому необхідне строге керування бізнес-процесами на основі workflow (табл. 1).

Таблиця 1

Особливості застосування workflow та та його властивості

Характеристика	Визначення
повна/часткова автоматизація	контент, документи, інформація або завдання автоматично передаються для виконання необхідних дій від одного учасника до іншого відповідно до набору процедурних правил.
керування потоком робіт	описує, створює і керує бізнес-процесом за допомогою ПЗ, яке інтерпретує опис процесу, взаємодіє з учасниками потоку робіт і за необхідністю викликає відповідні програмні застосування та інструментальні засоби.
автоматизація процесу	автоматизує процес, а не функцію, та реалізує правила взаємодії учасників процесу; ці аспекти є основними центрами втрат через свою невизначеність.
спрощення ведення бізнесу	важливий аспект роботи Інтернет-порталів систем електронної комерції для забезпечення автоматизації складних ділових процесів в організації.

Нові платформи Інтернет-розроблення значно знижують ризик втрати, дублювання контента або створення його знову із-за неможливості відшукати. Засоби контролю версій контенту гарантують, що вміст Інтернет-порталів ніколи не буде втрачений або випадково переписаний. Редактори також з легкістю можуть знаходити необхідні версії сайту. В бізнес-процесах не припустимий хаос та затримки. Побудова бізнес-процесів на основі ролей та груп користувачів означає їх незалежність від затримок виконання окремими особами. Ролі та процеси опрацювання різних інформаційних ресурсів (наприклад, зображень та юридичних документів), як правило, істотно відрізняються.

Аналіз останніх досліджень та публікацій

Одним із відомих методів аналізу текстової інформації є *контент-аналіз* – стандартна методика дослідження в області суспільних наук, предметом якої є аналіз змісту текстових масивів і продуктів комунікативної кореспонденції (наприклад, коментарі, форуми, електронне листування, статті тощо).

Контент-аналіз починався як кількісно-орієнтований метод аналізу текстів для дослідження масових комунікацій [6]. Вперше він був застосований в 1910 році соціологом Максом Вебером (Max Weber) для оцінювання охоплення друком політичних акцій в Німеччині. Американський дослідник засобів комунікації Гарольд Лассвелл (Harold Lasswell) в 30-40 роки використав подібну методику для отримання змісту пропагандних повідомлень під час війни. З появою засобів автоматизації, текстів в електронному вигляді, починаючи з 60-х років XX століття, початковий

розвиток отримав контент-аналіз інформації великих обсягів – баз даних та інтерактивних медіа-джерел. Традиційне політичне використання сучасних інформаційних технологій контент-аналізу було доповнене необмеженим списком рубрик та тематик, які охоплюють виробничу та соціальні сфери, бізнес і фінанси, культуру та науку. Цей процес супроводжувався великою кількістю різноманітних програмних систем. Поняття контент-аналізу не має однозначного визначення [1–6], що породжує проблему: системи, побудовані на основі різних підходів до контент-аналізу, в загальному випадку несумісні (табл. 2).

Таблиця 2

Неоднозначні визначення поняття контент-аналізу різними авторами

Автор	Визначення
Д. Джері, ДЖ. Джері	методика об'єктивного якісного та систематичного дослідження змісту засобів комунікації;
Д. Мангейм, Р. Рич	систематичне кількісне опрацювання, оцінювання та інтерпретація форми і змісту інформаційного джерела;
В. Іванов	якісно-кількісний метод дослідження документів (характеризується об'єктивністю висновків і строгістю процедури) та квантифікаційного опрацювання тексту з подальшою інтерпретацією результатів; предмет дослідження – проблеми соціальної дійсності, які висловлюються і приховуються у документах, та внутрішні закономірності самого об`єкту дослідження;
Б. Краснов	складається із пошуку в тексті визначених змістовних понять (одиниць аналізу), виявлення частоти їх появи та співвідношення із змістом всього документа;
Е. Таршис	техніка дослідження для отримання результатів шляхом аналізу змісту тексту про стан і властивості соціальної дійсності.

У [1] О.М. Алексєєв виділив основні складові контент-аналітичного дослідження (табл. 3).

Таблиця 3

Основні складові контент-аналітичного дослідження

Назва	Властивості основних складових контент-аналітичного дослідження
Спостереження	опрацювання масової сукупності текстів, використовуючи при цьому типові соціологічні процедури суцільного/вибіркового спостереження, з дотриманням вимог репрезентативності.
Структурування	припущення структурування, сегментації, розчленування текстів чи виділення із них змістовних інваріантів (повторення в усіх/ряді текстів) в досліджуваній масовій сукупності.
Формалізація	забезпечення однотипності сегментації і виділення інваріантів, застосування високого ступеня формалізації, суворих операційних правил і формальних алгоритмів в аналітичних процедурах.
Реферування	формалізований поділ цілісних текстів чи виділення окремих елементів їх для наступного збору із застосуванням аналітико-синтетичної процедури.
Аналіз	використання методів теорії ймовірності та математичної статистики для опрацювання текстів.

Отже, *контент-аналіз* – це кількісно-якісний аналіз текстової інформації та текстових масивів з метою подальшої змістовної інтерпретації отриманих кількісно-якісних закономірностей. *Контент-аналіз* застосовують при дослідженні джерел, інваріантних за структурою/змістом, але які існують як не систематизований, безладно організований текстовий матеріал [1–6]. *Метод контент-аналізу* полягає у формуванні з різноманіття текстового матеріалу абстрактної моделі змісту тексту.

Існує два типи методів контент-аналізу: *кількісний* і *якісний* (табл. 4).

Таблиця 4

Типи методів контент-аналізу

Назва	Кількісний (змістовний)	Якісний (структурний)
Визначення	дослідження слів, тем та повідомлень, який зосереджується на змісті контенту.	дослідження, в якому досліджують не зміст контенту, а його форму та структуру.
Приклад	в якості першого кроку дослідник має створити словник, в якому кожне спостереження отримає визначення та буде віднесено до відповідної категорії.	визначення періоду часу або обсягу друкованого простору, який приділено темі в тому чи іншому джерелі, або скільки слів або стовпців приділено кожній темі відповідної категорії.
Особливість	перед проведенням аналізу обраних лінгвістичних одиниць, передбачують їх зміст (створення словника) та визначають кожний можливий результат спостереження у відповідності із очікуванням дослідника.	розраховують питомі ваги P кожної теми і категорії ($P = R/T$, R – кількість одиниць даної категорії, T – загальна кількість одиниць) та проводиться порівняльний аналіз відповідних тем для подальшого прогнозування подій, процесів.

Виділення проблем

Однією із головних особливостей нашого часу є постійний ріст темпів виробництва контенту. Цей процес є об'єктивним і загалом, безумовно, позитивним. Але виникла та існує головна проблема Інтернет-простору: прогрес в галузі виробництва контенту призводить до пониження загального рівня інформованості потенційного користувача. Крім збільшення обсягів контенту до масштабів, яке призводить до неможливості його безпосереднього опрацювання, та швидкості його поширення виник цілий ряд специфічних проблем (табл. 5), пов'язаних із швидким розвитком інформаційних технологій.

Таблиця 5

Основні проблеми, викликані швидкими темпами росту обсягу виробництва контенту

Назва	Основна причина	Рішення
Інформаційний шум	непропорційний ріст інформаційного шуму через структурованості масивів контенту.	Фільтри, контент-моніторинг, технічний аналіз сайту, контент-аналіз.
Паразитивний контент	поява так званого паразитичного контенту, який отримують в якості додатків.	Фільтри, контент-моніторинг, контент-аналіз.
Нерелевантність контенту	невідповідність формально релевантному контенту (тематично відповідному) дійсним потребам його споживачів.	Створення анотованої бази даних, пошукових образів первинного контенту та їх кластеризація, контент-аналіз.
Дублювання контенту	багаторазове дублювання одного контенту в різних джерелах.	Контент-аналіз, сканери і фільтри на базі лінгвостатистики та критеріїв.
Навігація в потоках контенту	швидкий ріст обсягів контенту і швидкості його поширення.	Технічний аналіз сайту, фільтри, контент-моніторинг, контент-аналіз.
Надмірність результатів пошуку	наявність двох якісно різних категорій: дублювання та невідповідності контенту.	Анатований контентний пошук, контент-аналіз та автоматичне реферування.

Причина втрати актуальності традиційних інформаційно-пошукові системи полягає не стільки у фізичних обсягах контентних потоків, скільки в їх актуальності та динаміці (постійне систематичне та не завжди регулярне оновлення контенту). Охоплення та узагальнення великих динамічних контентніх потоків, які неперервно генерують в Інтернет-просторі, вимагає якісно нових методів/підходів для розв'язання поставлених задач та вирішення проблем. Вихід із такої ситуації – застосування засобів автоматизації знаходження найбільш важливих складових в

контентних потоках. Такий перспективний напрям отримав назву *контент-моніторингу*. Його поява була викликана задачами систематичного відстеження тенденцій та процесів в контентному середовищі, що постійно оновлюється [6]. *Контент-моніторинг* – це змістовний аналіз контентних потоків з метою постійного отримання необхідних якісних та кількісних зрізів на протязі не визначеного наперед проміжку часу. Важливішою методологічною складовою *контент-моніторингу* є *контент-аналіз.*

Більшість із наведених вище визначень контент-аналізу конструктивні, тобто процедурні. Через різні початкові підходи вони породжують різноманітні алгоритми, які часом суперечать один одному. Існуючі різноманітні підходи до розуміння контент-аналізу піддають цілком виправданій критиці. Найбільші сумніви викликає ігнорування ролі контексту. Але, не дивлячись на різноманіття трактувань контент-аналізу, практичне значення методу дозволяє уникнути багатьох суперечностей. Об'єднання засобів і методів, їх природний відбір шляхом багатократного оцінювання отриманих результатів відкривають можливість виділення/підтвердження знань і фактичні сили/корисність інструментарію.

Формулювання мети

Використання контент-аналізу текстової інформації в системах електронної комерції дозволяє визначити поширеність тієї чи іншої ознаки досліджуваної сукупності текстів. При цьому важливо не стільки абсолютне, скільки відносне значення ознаки, тобто характеристика її місця (частки) серед інших ознак. Наприклад, відсоток обговорення користувачами форуму економічних питань відносно політичних, або відсоток позитивних коментарів щодо продукції відносно негативних та відносно всіх коментарів на дану категорії продукції в Інтернет-магазині. Вимірювання співвідношення між ознаками в текстах дає емпіричний матеріал для розуміння функціональних зв`язків між елементами відображеної в текстах дійсності, наприклад, визначення настрою аудиторії форуму щодо економічної або політичної ситуації в країні/світі. За наявності текстів, що мають хронологічну послідовність, отримують низку фіксованих у часі “портретів” досліджуваної реальності (зміна попиту на категорію продукції залежно від сезону, наприклад, фантастику читають більше взимку, а детективи – влітку) або “портретів” цільової аудиторії (зміна попиту на категорію продукції залежно від статі, наприклад, попит на жіночій одяг у вересні більший, ніж в березні), що дає змогу висувати гіпотези прогностичного характеру про функціонування елементів системи.

Аналіз отриманих наукових результатів

Під час дослідження механізмів породження текстової інформації в роботі [4] виявлено, що від того, як побудоване ймовірнісно-лінгвістичне випробування та організовано вибір з тексту окремих його одиниць, залежить вибір тієї чи іншої моделі опису тексту. Ймовірнісне моделювання текстової інформації та його складових є вступним, підготовчим етапом до опису функцій лінгвістичних одиниць в тексті. Вивчення функціонування мови і мовлення за допомогою ймовірнісного моделювання текстової інформації ґрунтується на моделях квантитативної лігвістики і методів теорії ймовірності, математичної статистики, теорії інформації та комбінаторики. У лінґвістичних дослідженнях, особливо під час реалізації алгоритмів контентного пошуку [3], постійно виникають завдання, пов’язані з прогнозуванням появи в сеґменті заданої довжини певної кількості словоформ/словосполучень, що належать до певних класів. Ймовірнісне моделювання тексту і складів, словосполучень, граматичних класів дає змогу визначати обсяг вибірки, необхідної для забезпечення із заданою ймовірністю появи хоча б одного разу відповідної лінґвістичної одиниці [4].

Кількісні оцінки змістовної інформації в тексті та утворюваних його словах/ словосполученнях (рис. 1) отримують, спираючись на значення синтаксичної інформації та користуючись ідеєю контекстної обумовленості. Під час експерименту із вгадування літер невідомого тексту в роботі [4] зауважено, що свої гіпотези про найбільш ймовірнісні продовження тексту учасники експерименту будують, виходячи із двох типів комбінаторних обмежень: комбінаторики фіґур (літер та складів) та комбінаторики знаків (морфем, слів, словосполучень). Експеримент показує, що вже на 4-му або 5-му літерних кроках тексту комбінаторика літер та

складів пригнічується обмеженнями, пов'язаними зі сполучуваністю морфем та слів. У міру розгортання тексту на комбінаторику слів нашаровуються обмеження зі сполученням словосполучень та речень, з'являються обмеження, пов'язані з комбінаторикою параграфів, розділів, частин книги або статті. Отже, при вгадуванні літер, що розташовані на достатній відстані від початку тексту, учасник експерименту спирається не на статистичну комбінаторику літер та складів, а на змістовну (лексико-граматичну) побудову тексту. Якщо контент, вилучений із початкової ділянки тексту, виступає як кількісна оцінка дистрибуції (розподілу) та статистики літер, то синтаксична інформація, яка отримується з віддалених від початку тексту ділянок, слугує відображенням змістовної (семантико-прагматичної) інформації. Ці міркування дають змогу запропонувати метод контент-аналізу для кількісного оцінювання змістовної інформації в тексті та його сеґментах.

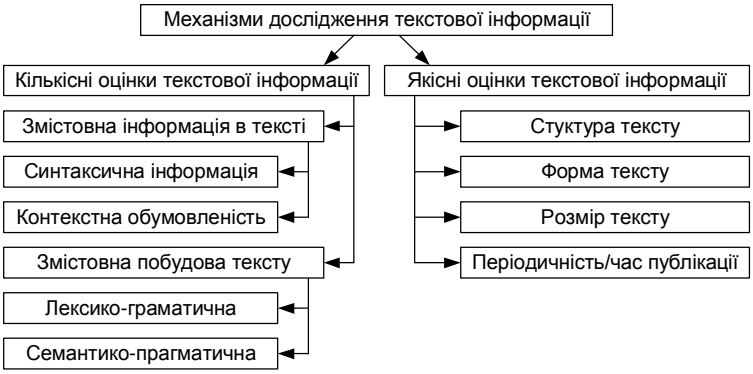

Рис. 1. Механізми дослідження текстової інформації

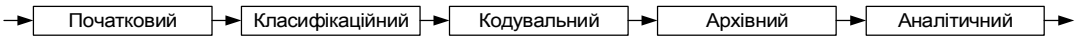

Рис. 2. Етапи роботи модуля опрацювання текстової інформації

Аналіз лексико-граматичної та семантико-прагматичної побудови тексту використовується насамперед в модулях опрацювання текстової інформації. Різновидом таких модулів є інформаційно-пошукові – сукупність методів і засобів, призначених для зберігання та пошуку документів, відомостей про них чи певних фактів. Головне завдання автоматизованих інформаційно-пошукових модулів полягає в тому, щоб з сукупності даних, які належать до системи, за допомогою контент-аналізу знайти і обрати ті, які найбільше відповідають інформаційним потребам споживача. Виконання перерахованих на рис. 2 етапів призводить до формування тематично підібраних масивів текстової інформації (табл. 6), в яких акумулюється інформація про висвітлення всіх аспектів досліджуваної проблеми, враховуючи різноманітність думок і поглядів.

Таблиця 6

Основні етапи роботи модуля опрацювання текстової інформації

№	Назва етапу	Призначення етапу
1	Початковий	визначення тематики дослідження, мети та об`єкту аналізу, його хронологічні та географічні рамки, принципи відбору.
2	Класифікаційний	формування класифікатора для відбору ключових цитат та складання інструкції для кодувальника.
3	Кодувальний	кодування фрагментів текстової інформації.
4	Архівний	збереження фрагментів текстової інформації в базі даних.
5	Аналітичний	автоматичне опрацювання фрагментів текстової інформації.

Побудова модулів опрацювання текстової інформації, наприклад, контентного пошуку (рис. 3), залежить від наявності та дотримання правил проведення вищезазначених етапів (табл. 7).

Таблиця 7

Правила проведення етапів опрацювання текстової інформації на базі контент-аналізу

Назва правила	Характеристика правила	Етап
контент-аналітичний відбір	процедура із набором точно визначених дій для опрацювання без будь-яких змін усіх об`єктів дослідження;	1
формування класифікатора	розділи класифікатора для кодування фрагментів тексту мають бути ясними і атомарними, не допускати двозначності тлумачення;	2
інтерпретація результатів	охоплює всі здобуті фрагменти тексту, висновки спираються не на частину результатів, а враховуються всі без винятку.	5

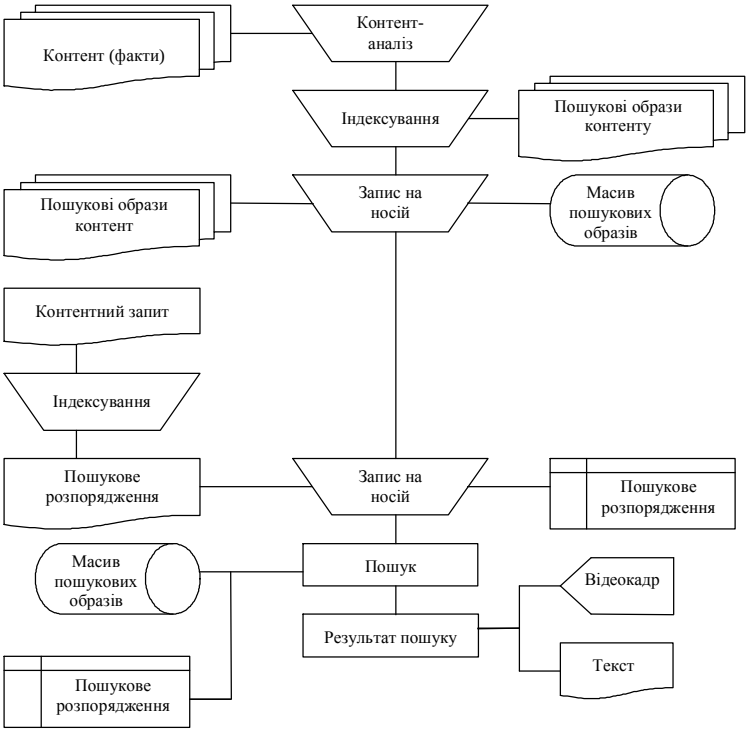

Рис. 3. Структурна схема модуля контентного пошуку

Таблиця 8

Класифікація механізмів контентного пошуку

Тип	Класифікація	Характеристика
за тематикою та змістом контенту	галузевий	наявність словника галузевого напряму, зберігання контенту наперед визначеного зразка, пошук з використанням словника.
	політематичний	відсутність загального правила пошуку.
	вузькоте-матичний	наявність словника тематичного напряму, зберігання контенту наперед визначеного зразка, пошук з використанням словника.
за типом контенту	документальний	об'єктом зберігання та пошуку є документи.
	фактографічний	зберігання та пошук окремих даних про будь-які події, процеси чи явища.
за режимом автоматизації	вибірковий	пошук за постійним набором запитів для певного континґенту користувачів у масиві поточних надходжень контенту через певні інтервали часу, при цьому змінюється зміст масивів модуля, а запити лишаються без змін.
	ретроспективний	пошук за всіма масивами, що зберігаються в модулі, згідно з разовими запитами, які змінюються залежно від інформаційних потреб користувача.

Контентний пошук – це сукупність наперед визначених операцій (табл. 8), необхідних для відшукання в системі документів, текстів, відомостей, фактів і даних, які відповідають на запит користувача. Для автоматизованого контентного пошуку та опрацювання текстової інформації велике значення має наявність/відсутність та частота появи тієї чи іншої категорії лінґвістичної одиниці в досліджуваному контентному масиві. Кількісний підрахунок дає змогу зробити об`єктивні висновки щодо спрямованості матеріалів за кількістю уживань одиниць аналізу (ключових цитат) у досліджуваних контентних масивах, наприклад, кількість позитивних/негативних відгуків на певний вид продукції. Якісний аналіз дає змогу зробити об`єктивні висновки щодо наявності в контентному масиві шуканої лінґвістичної одиниці та напрям ії контексту. *Операція пошуку* – це основна операція в автоматизованому модулі контентного пошуку. Сутність контентного пошуку полягає в ідентифікації відомостей, які зберігаються в такому модулі з інформаційним запитом.

Найефективніший засіб контентного пошуку – це перегляд кожного контенту і перевірка його на відповідність інформаційному запиту. Проте зберігання контенту в повному обсязі потребує значного обсягу пам’яті, а сам контентний пошук стає вельми тривалим. Тому в модулі контентний пошук виконується не за текстом контенту, а за його стислими характеристиками (табл.9), тобто пошуковими образами (ПОб) – це поданий у термінах інформаційно-пошукової мови (ІПМ) основний зміст контенту для зберігання в автоматизованому модулі контентного пошуку. Процедура визначення ПОб – це індексування, семантичний аналіз основного змісту контенту й переклад його на ІПМ. У модулі зберігають не тексти контенту, а його ПОб. Для пошуку контенту, поданого на ІПМ (проіндексованого) підлягають й інформаційні запити. Інформаційний запит, перекладений на ІПМ і доповнений для пошуку допоміжною інформацією (серія, рік видання тощо), є пошуковим розпорядженням (ПР). Рішення про успішне закінчення пошуку не обов’язково приймають в разі повного збігу ПОб та ПР. Для прийняття рішення про успішне закінчення пошуку достатньо збігу ПР і ПОб у межах $(0,7;1]$ або $(0,5;1]$. Усе залежить від критерію пошуку, який змінюють за бажанням користувача. Результатом пошуку текстової інформації може бути не один контент, а множина, з якої споживач обирає той контент, який найбільше відповідає його інформаційній потребі.

Таблиця 9

Основні операції контентного пошуку

Назва операції	Характеристика операції
Формування ПОб	створення, введення, зберігання в модулі ПОб контенту чи ПОб із контентом;
Формування запитів і ПР	створення, введення та зберігання в модулі запитів і ПР;
Пошук контенту	порівняння ПОб контенту з ПР;
Контент-аналіз	кількісний та якісний аналіз текстової інформації;
Прийняття рішення	прийняття рішення про видавання контенту відповідно до результату застосування контент-аналізу текстової інформації в межах $(0,7;1]$ або $(0,5;1]$;
Подання контенту	видавання контенту, що відповідає інформаційному запиту.

Процес індексування, семантичного аналізу і визначення основного змісту контенту виконується вручну модератором або автоматично за допомогою контент-аналізу. Під час індексування модератор вивчає зміст контенту, відокремлює його центральну тему та описує її в термінах ІПМ. Для деякого контенту їх назви розкривають центральну тему і предмет, але за назвою не завжди можна ідентифікувати контент. Ступінь докладності подання контенту в ПОб його центральній темі чи предмета, а також супутніх тем і предметів є глибиною індексування. Кожний модератор один і той самий контент може індексувати суб’єктивно, тому автоматизація цього процесу дала б змогу забезпечити його уніфікацію, звільнивши частину персоналу від важкої непродуктивної праці з індексування контенту. Кожний автоматизований модуль пошуку містить у своєму складі певний набір семантичних засобів: ІПМ, методи індексування документів та запитів,

методи пошуку. Основу семантичних засобів становить ІПМ – це спеціалізована штучна мова, яка призначена для опису центральних тем/предметів і формальних характеристик контенту, а також для опису інформаційних запитів і подальшого виконання пошуку (табл. 10). Іноді в модулях контентного пошуку одну мову використовують для індексації контенту, а іншу – для індексації інформаційних запитів. Природна мова не може бути використана як ІПМ через недостатню структуризацію, численні граматичні включення, неоднозначність і велику надлишковість (надлишковість української мови досягає 75–80 %). В ІПМ серед основних елементів (табл. 11) не використовують характерні для природної мови синоніми та омоніми через надання мові семантичної неоднозначності.

Таблиця 10

Вимоги до інформаційно-пошукової мови

Назва вимоги	Характеристика вимоги
забезпечення достатнього набору лексико-граматичних засобів	для точного вираження центральних тем чи змісту будь-якого контенту і теми довільного інформаційного запиту;
однозначність	кожний запис цією мовою має лише одне семантичне тлумачення;
зручність і компактність	для співставлення та повного або часткового порівняння ПОб і ПР;
відкритість і невелика вартість	для можливого розширення та внесення змін.

Таблиця 11

Основні елементи інформаційно-пошукової мови

Назва елементу	Характеристика елементу мови
Алфавіт	система графічних знаків для фіксації слів і висловлювань мови.
Лексика	сукупність використовуваних у мові слів.
Граматика	сукупність дієвих засобів мови та правил побудови висловлювань.
Парадигматичні (базові, аналітичні) відношення	відношення між словами, що не залежать від контексту, в якому вони використовуються, і породжені не мовними, а логічними зв'язками.
Синтагматичні відношення	лінійні відношення між словами, які безпосередньо встановлюються при об'єднанні слів у словосполучення та фрази; наприклад, коли до того чи іншого поняття приєднується ще одне, яке пояснює перше і вказує на те, що це поняття є власним ім'ям, визначенням чи певним видом діяльності (в контенті про винахід поняття, пов'язані з винаходом, і поняття, пов'язані з винахідником, мають різні назви для уникнення плутанини при пошуку).
Парадигмами	об'єднання в лексико-семантичні групи слів завдяки предметно-логічним зв'язкам на основі тієї чи іншої семантичної ознаки.
Правила побудови індексів та їх ідентифікація	парадигматика (лексика) мови спирається на певну сукупність (перелік) пов'язаних між собою мовних одиниць; синтагматика (граматика) мови потребує певних правил об'єднання цих одиниць у словосполучення.

При розробленні ІПМ треба звернути увагу на такі моменти:

- специфіка галузі чи предмета, для якого ця мова розробляється;
- особливості документів або текстів, які утворюють пошуковий масив контенту;
- характер інформаційних потреб користувачів модуля контентного пошуку.

Доцільність використання тієї чи іншої мови (табл. 12) багато в чому залежить від призначення автоматизованого модуля контентного пошуку, рівня її оснащення технічними засобами, рівня автоматизації інформаційних процедур і ланки керування.

Контент-аналіз розподілу текстової інформації користувачів дає можливість якісно оцінити контентний потік в системах електронної комерції для подальшого прийняття рішень відповідною особою. Діапазон основних методів і процедур контент-аналізу, наступний:

1) опис проблемної ситуації, пошук мети дослідження;

2) точне визначення об'єкта і предмета дослідження;
3) попередній аналіз об'єкта;
4) змістовне уточнення і емпірична інтерпретація понять;
5) опис процедур реєстрації властивостей і явищ;
6) визначення загального плану дослідження;
7) визначення типу вибірки, кола джерел тощо.

Таблиця 12

Класифікація інформаційно-пошукових мов

Назва	Передкоординатні мови	Посткоординатні мови
Тип мови	класифікаційний	дескрипторний
Властивість	словниковий склад задають через фіксований список слів і словосполучень.	словник обмежують окремими ізольованими словами/словосполученнями за алфавітом.
Класифікація	ієрархічні, фасетні або алфавітно-предметні традиційні бібліотечно-бібліографічні мови	дескрипторні (від фр. description – описання) та синтагматичні мови, а також семантичні коди.
Призначення	зберігання контенту в технічних архівах і бібліотеках лише у вузькоспеціалізованих системах.	подання змісту контенту через установлення відповідності між його текстом і деякою множиною ключових слів, або дескрипторів.
Особливість	дуже обмежений словниковий запас, який важко поповнювати.	ґрунтується на методі координатного індексування.
Недоліки	з огляду на зростання інформаційних потоків та інформаційних потреб не задовольняють вимоги сучасних пошукових систем.	недостатньо дослідження методи застосування аналогічних мов в сучасних пошукових системах.

У будь-якій фазі методу якісного контент-аналізу для оцінювання результатів можна залучити експерта (табл. 13). Цей метод покликаний забезпечити експерта необхідними засобами для аналізу висновків і результатів. Експерт за допомогою таких засобів може виявити певні властивості частини контенту і перевірити їх щодо загального контентного потоку, а загальні властивості контентного потоку розповсюдити на його конкретну тематичну частину. Метод кількісного контент-аналізу, як правило, складається з трьох основних етапів (табл. 14).

Контент-аналіз (табл. 15) застосовують для автоматичного формування дайджестів, автоматичного виявлення взаємозв'язку понять (категорій), автоматичної кластеризації взаємозв'язків для виявлення найбільш важливих, автоматичного виявлення забарвлення взаємозв'язків (наприклад, позитивних і негативних). Одною з найважливіших проблем в контент-аналізі є процес категоризації, який задає концептуальну сітку, в термінах якої аналізується контентний потік (табл. 16). За будь-яким з двох підходів відбувається генерація нових категорій.

Таблиця 13

Етапи якісного контент-аналізу

Назва етапу	Характеристика етапу
розбиття тексту на блоки	зведення множини контенту до скінченної кількості інтегрованих змістовних одиниць (категорія, послідовність, тема) для кодування і опрацювання;
реконструкція суб'єктивних складових контентного потоку	реконструкція системи значень, думок, поглядів і доказів кожного джерела тексту;
формування висновків	виведення узагальнень шляхом порівняння індивідуальних системних значень.

Таблиця 14

Етапи кількісного контент-аналізу

Назва етапу	Характеристика етапу
виділення одиниці аналізу	перетворення лінгвістичної одиниці у форму, прийнятну для опрацювання;
підрахунок частоти одиниць аналізу	застосування різноманітного математичного апарату для виявлення взаємозв'язків між лінгвістичними одиницями;
інтерпретація отриманих результатів	отримання змістовних, семантично наповнених результатів з використанням математичних методів без залучення штучного інтелекту, об'ємних семантичних формалізаторів, експертів.

Таблиця 15

Етапи контент-аналізу текстової інформації

№	Назва етапу	Характеристика етапу
1	Визначення сукупності джерел або контенту	за допомогою набору заданих критеріїв, яким відповідає кожний контент.
2	Формування вибіркової сукупності контенту	формування за критеріями обмеженої вибірки з більшого масиву інформації.
3	Виявлення лінґвістичних одиниць	дотримання чітких вимог до вибору можливої лінґвістичної одиниці аналізу (слова або теми – вислів про який-небудь предмет).
4	Виділення одиниць обчислення	одиниці обчислення можуть збігатися із змістовними одиницями або мати специфічний характер
5	Безпосередньо процедура обчислення	стандартні прийоми класифікації за виділеними угрупуваннями із формул математичної статистики та теорії ймовірності.
6	Інтерпретація отриманих результатів відповідно до цілей і завдань конкретного дослідження	виявляються і оцінюються такі характеристики текстового матеріалу, які дозволяють робити висновки про те, що хотів підкреслити або приховати його автор; або, на основі статистичного набору підрахованих коефіцієнтів Яніса за певний період часу на визначену категорію продукції можна спрогнозувати зміни щодо попиту на цю саму продукцію.

Таблиця 16

Напрями дослідження достатньо великого контентного потоку

Назва	Характеристика напряму
Категоризація	визначення скінченної, але свідомо надлишкової, сукупності категорій для отримання кількісних даних поява деяких з них, при цьому передбачається і автоматична або напівавтоматична кластеризація (поділ на групи і класи) неврегульованої послідовності категорій і, відповідно, отримання на її основі нових узагальнених категорій;
Data Mining	виявлення в контентному потоці за допомогою кількісних багаторазових оцінок нових знань із подальшою кваліфікацією їх як категорій.

Етап визначення сукупності джерел, що досліджуються, або повідомлень за допомогою набору заданих критеріїв, яким відповідає кожне повідомлення:

- заданий тип джерела (форум, електронна пошта, Інтернет-газета, чат, Інтернет-журнал);
- один тип повідомлень (стаття, електронний лист, баннер, коментар);
- задані сторони, що беруть участь в процесі комунікації (відправник, одержувач, реципієнт);
- зіставлений розмір повідомлень (мінімальний обсяг або довжина);
- частота появи повідомлень;

- спосіб розповсюдження повідомлень;
- місце розповсюдження повідомлень;
- час появи повідомлень тощо.

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 проведення вибірки.

У разі прийняття за одиницю аналізу теми, враховують такі правила:

- розмір теми не виходить за межі абзацу;
- нова тема виникає, якщо проводять заміну того, хто сприймає або діє, цілі, категорії.

Етап виділення одиниць обчислення, які можуть збігатися із змістовними одиницями або мати специфічний характер. У першому випадку процедура аналізу зводиться до підрахунку частоти зустрічання виділеної змістовної одиниці, в іншому — дослідник на основі аналізованого матеріалу і цілей дослідження висуває одиниці обчислення, якими можуть бути:

- фізична протяжність текстів;
- площа тексту, заповнена змістовними одиницями;
- кількість рядків (абзаців, знаків, колонок тексту);
- розмір та вид файла;
- 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.

Існують також спеціальні процедури підрахунку результату контент-анализу, наприклад, формула розрахунку коефіцієнта Яніса c, призначеного для обчислення співвідношення позитивних і негативних (щодо вибраної позиції) оцінок, думок, аргументів. Коефіцієнт Яніса можна застосовувати, наприклад, для розрахунку співвідношення позитивних і негативних думок, висвітлених в коментарях користувачів щодо продукції, яка реалізується через систему електронної комерції. У разі, якщо кількість позитивних оцінок перевищує кількість негативних, коефіцієнт Яніса підраховується за формулою $c = \frac{f^2 - f \cdot n}{r \cdot t}$, де f – кількість позитивних оцінок; n – кількість негативних оцінок; r – об'єм змісту тексту, що має пряме відношення до проблеми, яка досліджується; t – загальний об'єм аналізованого тексту. У разі, коли кількість позитивних оцінок менша за негативну, коефіцієнт Яніса знаходиться за формулою $c = \frac{f \cdot n - n^2}{r \cdot t}$.

Отже, у простому вигляді ідею контент-моніторингу сформулюємо як постійне виконання вузько обкресленого своїми завданнями контент-аналізу безперервних контентних потоків. Підкреслимо, що саме безперервне відтворення в часі процесу опрацювання текстової інформації є найхарактернішою особливістю контент-моніторингу. Власне контент-аналіз виступає тут як складова, а контент-моніторинг має власну проблематику і власні шляхи вирішення прикладних завдань, наприклад, технічний аналіз контенту сайта (табл. 17).

Застосування контент-аналізу текстової інформації в системах електронної комерції, на думку авторів, дає ряд переваг для спрощення ведення бізнесу та вирішує низьку проблем, що стоять перед учасниками бізнес-процесів. Наведемо декілька основних переваг застосування контент-аналізу текстової інформації в системах електронної комерції:

1) автоматизація процесу фільтрації текстової інформації, яку розміщує користувач на сайті системи електронної комерції;

2) можливість автоматичного створення “портрету” постійного користувача на основі його коментарів;

3) можливість автоматичного створення “портрету” цільової аудиторії на основі аналізу “портретів” постійних користувачів;

4) скорочення кількості модераторів, які обслуговують систему електронної комерції;

5) скорочення часу для розміщення текстової інформації постійного користувача на сайті внаслідок автоматичного опрацювання цієї інформації та відсутності проміжної ланки у вигляді модератора;

6) ліквідація мовного бар'єру внаслідок автоматичного формування словників постійного користувача та використання система автоматичного перекладу.

Бізнес-процес системи електронної комерції схематично поданий на рис. 4 у вигляді послідовності операцій (робіт, функцій), які виконуються окремими модулями з використанням контенту і відповідно до деяких правил, що диктують порядок виконання роботи, визначають маршрути руху документів, терміни виконання окремих функцій. Ефективним є використання програмних агентів у ланцюгах постачання та збуту в системах електронної комерції (рис. 5).

У табл. 18 подані основні етапи керування контентом в системах електронної комерції.

Таблиця 17

Методи технічного аналізу контенту сайту

Напрям аналізу	Характеристика операції	
на користувача	перевірка навігаційних елементів;	повноцінності та ефективності
	перевірка пошуку контенту;	правильності функціонування пошуку по сайту;
	перевірка можливості замовлення;	товарів і/або послуг
	перевірка способів оплати;	товарів і/або послуг
	перевірка технічної сторони роботи сторінок спілкування із аудиторією;	форумів;
		гостьових книг;
		відповідей на питання;
		підписки на новини;
		реєстрації відвідувачів;
		авторизації на сайті;
	перевірка можливості змін;	сканування/редагування прайс-аркушів та каталогів
на пошукового робота/систему	аналіз доступності сайту;	перевірка надійності та швидкості роботи хостингу;
		аналіз захищеності від несанкціонованого доступу;
		визначення ваги файлів, коректності відображення сайту в браузерах (Internet Explorer, Opera, Mozilla);
		оцінка якості html-коду з погляду індексації сайту пошуковими роботами;
		оцінка функціонування елементів мультимедіа;
	аналіз засобів керування контентом сайту;	оцінка можливості редагування сайту: • зміни існуючих сторінок та розділів сайту; • створення нових сторінок, розділів, різноманітних інформаційних елементів; • зміни структури сторінок;можливості додавання довільних блоків коду;
		оцінка зручності користування засобами керування сайтом;
		оцінка надійності роботи всіх елементів керування сайтом та опрацювання помилок;
	контент-аналіз тексту;	кількісний та якісний.

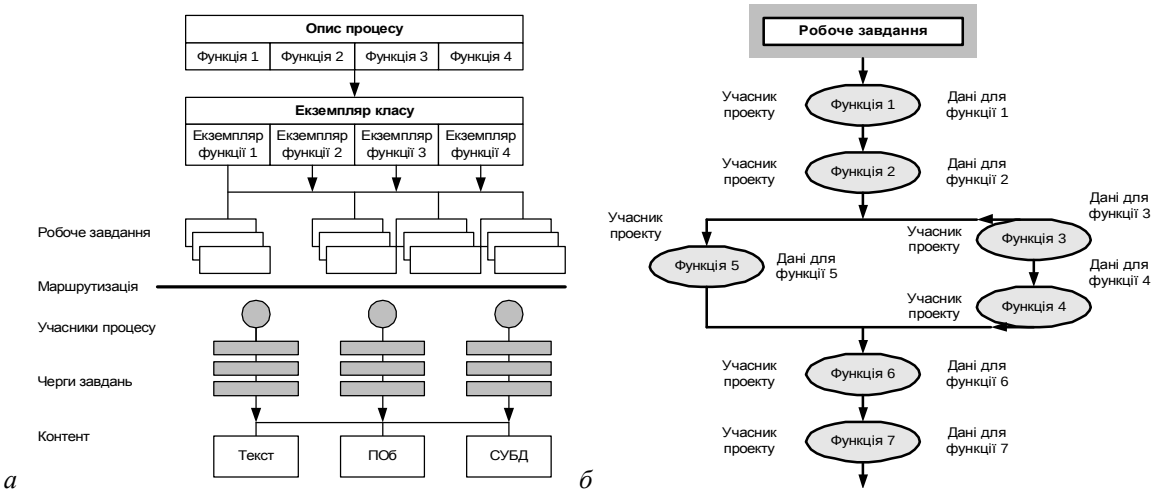

Рис. 4. Структура бізнес-процесу системи електронної комерції

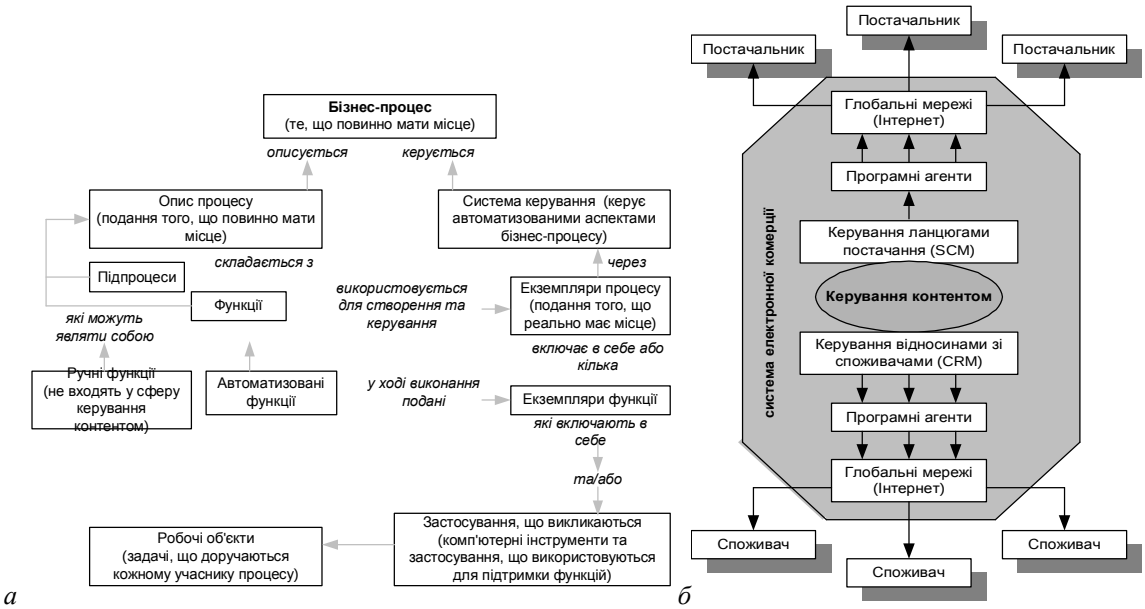

Рис. 5. Схема бізнес-процесу (а)
та програмних агентів системи електронної комерції (б)

Персоналізація на базі правил – це надання контента певним користувачам або їх групам із застосуванням умовної бізнес-логіки. Наприклад, правило, згідно з яким всі, хто цікавиться дитячими книгами, потрапляють до групи, до якої постійно прямує реклама дитячого одягу. Правила зазвичай розробляють на основі контенту, який вводиться користувачем в реєстраційну картку про себе і свої переваги. При персоналізації за допомогою фільтрів (інтелектуальні агенти) використовують складні алгоритми категоризації і надання контента на основі аналізу поведінки користувача, а саме того, до якого контенту він звертається, які сайти відвідує тощо, тобто ведеться історія зареєстрованого користувача та історія групи користувачів з переважною кількістю пріоритетів щодо зацікавлень.

Глобалізації і локалізації – одна з вимог сучасного ринку, зокрема, для успішних маркетингових досліджень необхідна обширна статистична база. Глобалізація – це більше, ніж просто перетворення Web-сторінок з однієї мови на іншу, але і локалізація (просування місцевих торговельних марок). В рамках глобалізації і локалізації сайту повинні також розглядатися стратегія контенту, ієрархія інформації і навігаційна структура. Для відповідних систем необхідно вирішувати такі проблеми:

1) створення контенту авторами на різних мовах(список мов заздалегідь визначений);
2) збереження глобального/локального контенту в різних базах даних;
3) підтримка служб автоматичного перекладу текстової інформації;
4) відстеження змін контенту та його затвердження.

Таблиця 18

Етапи керування контентом

Назва етапу	Назва процесу	Особливості процесу
створення контенту	визначення теми контенту;	мета створення, зміст та структура контенту;
	визначення форми подання контенту:	• графічна інформація; • текст (стаття, прес-реліз, посадові інструкції); • шаблони HTML; • код back-end тощо;
	вибір інструментів/редакторів для створення контенту:	• редактори HTML; • текстові процесори; • візуальні редактори та засоби створення об'єктів;
керування автоматизованими бізнес-процесами	призначення прав доступу;	повний або обмежений доступ до контента;
	визначення набору процесів	стандартні процеси створення та публікації нового інформаційного наповнення;
	збереження контенту;	в базі даних або сховищі (репозиторії)
	протоколювання процесів;	процеси створення, передачі та збереження;
	автоматичне інформування;	інформування про контент наступного виконавця;
	ведення аудиту подій;	збереження версій контенту;
	контент-аналізу тексту;	кількісний або якісний;
	доступ до попередніх версій;	підтримка можливості звернення користувачів до попередніх версій контенту;
	створення бізнес-процесів:	• визначення мети, ролі та задач; • задання для ролей групи користувачів; • розроблення різних бізнес-процесів для різного контенту;
розповсюдження контента	статичне ;розповсюдження;	без застосування будь-якої логіки поведінки
	динамічне розповсюдження;	• персоналізація (правила/фільтри); • глобалізація; • локалізація.

Зауважимо, що від передового закінченого рішення по керуванню контентом очікується врахування не лише технічних вимог системи, але і людських і творчих аспектів роботи.

В системах керування контентом мають бути реалізовані наступні функції та модулі:

1) контроль зсередини системи – призначення користувачів, яким доступний той або інший опублікований контент;

2) інтеграція контенту – можливість перенести готовий контент в нове рішення;

3) підтримка контенту різного типу – зберігання і сортування будь-якого контенту, включаючи графіку, аудіо і відео, в центральному репозиторії;

4) детальна якісна документація і контекстно-інтелектуальна довідка;

5) рейтингова система оцінки статей сайту;

6) шаблонні зміни – загальні зміни форматування контенту однієї частини сайту відображаються на весь сайт;

7) підтримка workflow – створення своїх автоматизованих бізнес-процесів для конкретного контенту (зображень, статей тощо);

8) маркування контенту – можливість додавати нові категорії і маркери до контенту до і після їх розміщення в репозиторії;

9) контроль версій – створення нових версій, перегляд і повернення до попередніх версій контенту;

10) контент-аналіз контентних потоків в системі;

11) інструмент візуальної адміністрації – легке керування контентом авторами, не удаючись до програмування, зазвичай реалізується за допомогою HTML-форм.

Висновки і перспективи подальших наукових розвідок

Цікавою особливістю контент-аналізу є і те, що цю методологію до останнього часу пов'язували з певною сферою людської діяльності (політикою і соціологією). Проте, сьогодні контент-аналіз все ширше застосовується в багатьох сферах політичного і економічного життя, що сприяє більшому прикладному значенню використаних в методології контент-аналізу філософських категорій, соціології і лінгвістики. Контент-аналіз в рамках дослідження контентних потоків – новий напрям, який передбачає аналіз масиву текстових документів, – результатів моніторингу інформаційного простору. Загальновизнаним є розподіл методології контент-аналізу на дві гілки: якісну і кількісну. Основа кількісного контент-аналізу – частота появи в документах певних характеристик змісту. Метод якісного контент-аналізу ґрунтується на самому факті присутності або відсутності в тексті однієї або декількох характеристик змісту.

1. Алексеев А.Н. Контент-анализ в социологии и точки соприкосновения с другими отраслями знания / А.Н. Алексеев // Проблемы контент-анализа в социологии: М-лы Сибирского социологического семинара – Новосибирск, 1970. – С.11–12. 2. Іванов В.Ф. Контент-аналіз: Методологія і методика дослідження ЗМК: Навч. посібник / Наук. ред. А.З. Москаленко / В.Ф. Іванов. – К.,1994. – 112 с. 3. Литвин В. В. Інтелектуальні системи / Литвин В. В., Пасічник В. В., Яцишин Ю.В. – Львів: "Новий Світ – 2000", 2009. – 406 с. 4. Пиотровский Р. Г. Математическая лингвистика: Учебное пособие / Пиотровский Р. Г., Бектаев К. Б., Пиотровская А. А. – М. : Высшая школа, 1977. – 384 с. 5. Сорока М.Б. Використання методу контент-аналізу при створенні автоматизованих інформаційних систем / М.Б. Сорока, Н.В. Танатар // Бібліотека. Наука. Культура. Інформація: Наукові праці НБУВ. – 1998. – Вип. 1. – С. 318–322. 6. Федорчук А. Г. Контент-мониторинг информационных потоков [Електронний ресурс] // Б-ки нац. акад. наук: пробл. функционирования, тенденции развития. – Электрон. дан. (1 файл). – К., 2005. – Вып. 3. – Режим доступа: http://www.nbuv.gov.ua/articles/2005/05fagmip.html. – Назва з екрана.