ІНФОРМАЦІЙНІ СИСТЕМИ ТА МЕРЕЖІ

УДК 004.738.5

А.Ю. Берко, В.А. Висоцька, М.М. Сороковський
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ЗАСТОСУВАННЯ МЕТОДУ КОНТЕНТ-АНАЛІЗУ ДЛЯ ФОРМУВАННЯ ІНФОРМАЦІЙНИХ РЕСУРСІВ У СИСТЕМАХ ЕЛЕКТРОННОЇ КОНТЕНТ-КОМЕРЦІЇ

Розглянуто питання розроблення уніфікованих методів опрацювання інформаційних ресурсів у системах електронної контент-комерції. Розроблено формальну модель та узагальнену типову архітектуру системи електронної контент-комерції. Розроблено методи проектування та реалізації системи електронної контент-комерції на прикладі інтернет-журналу, який відображає результати теоретичних досліджень.

Ключові слова: контент, інформаційний ресурс, системи електронної контент-комерції.

This paper is devoted to the development of unified methods for processing information resources in the systems of electronic content commerce. A formal model and generalized typical architecture of systems of electronic content commerce are declared. Methods of designing and implementation of systems of electronic content commerce on the example of online Magazine, which reflects the results of theoretical research, are developed.

Keywords: content, information resource, electronic content commerce systems.

Вступ. Постановка проблеми

Проблематика опрацювання інформаційних ресурсів у системах електронної контент-комерції є актуальною через активний розвиток досліджень у галузі е-бізнесу і відсутність теоретичного обгрунтування стандартизованих методів й уніфікації програмних засобів формування, управління та супроводу контенту [1–3]. З'являються нові підходи/способи вирішення цієї проблеми, але існує невідповідність між відомими технолгіями опрацювання інформаційних ресурсів та принципами побудови систем електронної контент-комерції. Відсутні загальні підходи до створення цих систем, їх загальнотипова архітектура та уніфіковані методи формування/управління/супроводу контенту [2].

Актуальність тематики зумовлена глобалізацією е-бізнесу; зростанням попиту на комерційний контент та швидкого доступу до нього; нерівномірністю функціонування бізнес-процесів відповідно до регіонів; необхідністю оперативно/регулярно/періодично отримувати необхідний контент; економією часу/ресурсів під час отримання необхідного контенту; персоналізацією у наданні послуг та інтегрованістю систем електронної контент-комерції. Переваги впровадження цих систем полягають у збільшенні оперативності одержання контенту; скороченні циклу виробництва і продажу комерційного контенту; зниженні витрат, пов'язаних з обміном контенту; відкритості стосовно користувачів; автоматичному інформуванні користувачів в інтерактивному режимі; створенні альтернативних каналів продажу [1].

Особливість використання систем електронної контент-комерції полягає у відкритості (доступ для всіх компаній/користувачів), глобальності (доступ з будь-якої точки світу), необмеженності у часі (доступ у будь-який час), відвертості (низький бар'єр для входу на ринок) процесів системи, прямій взаємодії із користувачем (скорочення каналів поширення та ліквідація проміжних ланок); автоматичному опрацюванні запитів та відстеженні інформації про користувачів, скороченні витрат на функціонування е-бізнесу та наданні додаткової інформації в інтерактивному режимі [1–3, 9].

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

Активний розвиток Інтернету сприяє зростанню попиту на комерційний контент та реалізації нових форм інформаційного обслуговування [1, 2]. Документована інформація, підготовлена відповідно до потреб користувачів і призначена для їх задоволення, є інформаційним продуктом (комерційним контентом) та основним об'єктом процесів е-комерції. Розроблення/впровадження систем електронної контент-комерції є одним із стретагічних напрямів розвитку е-бізнесу. Характерною рисою таких систем є автоматичне опрацювання інформаційних ресурсів для збільшення обсягів продажу контенту постійному користувачу, активного залучення потенційних користувачів та розширення меж цільової аудиторії [1, 2]. Зокрема, принципи і технології електронної контент-комерції активно застосовують під час створення програмних продуктів для on-line/off-line продажу контенту, в системах типу інтернет-магазин, cloud storage/computing, аналізу, обміну та збереження контенту (табл. 1). Практичний чинник проблеми опрацювання інформаційних ресурсів у системах електронної контент-комерції пов'язаний із вирішенням проблем швидкого темпу зростання обсягів контенту в Інтернеті та поширення доступності до нього, активного розвитку е-бізнесу, розширення набору та зростання попиту на інформаційні товари/послуги, створення технологій/засобів та розширення областей застосування методів формування, управління та супроводу контенту [2]. У цьому напрямі активно працюють провідні світові виробники засобів опрацювання інформаційних ресурсів, зокрема, Google, AIIM, CM Professionals organization, EMC, IBM, Microsoft Alfresco, Open Text, Oracle, SAP.

Таблиця 1

Напрями розвитку електронної контент-комерції

Напрям	Характеристика напряму
On-line продаж	Розповсюдження контенту через Інтернет-газети, Інтернет-журнали, дистанційне навчання, Інтернет-видання у вигляді словників/довідників, Інтернет-видавництва, розважальні/інформативні/дитячі портали.
Off-line	Продаж контенту через системи типу copywriting services, Marketing Services Shop, RSS Subscription Extension.
Інтернет-магазин	Системи для продажу контенту типу eBooks, Software, video, music, movies, picture, digital art, manuals, articles, certificates, forms, files тощо.
Системи cloud storage/ computing	Системи cloud storage (Amazon S3, EMC Atmos, FilesAnywhere, Google Cloud Storage, iCloud by Apple, Ubuntuone, Windows Azure Storage) та cloud computing (Google, Apple, Windows, Mac, Linux, iPhone, Android, Palm Pre, Micorsoft) призначенні для збереження різного типу контенту або лише медіа-зображення, музики, фільмів. Моделями надання послуг за допомогою cloud computing є інфраструктура як послуга IaaS (Amazon, Microsoft, VMWare, Rackspace, Red Hat), платформа як послуга PaaS (Google Apps) та програмне забезпечення як послуга SaaS (Gmail, Google docs).
Системи аналізу контенту	Сервіси аналізу сайту, верифікації/валідації контенту, визначення частоти пошукових запитів, визначення позицій по ключових словах. Існують спеціальні програми, сервіси та пошукові системи зображень для перевірки унікальності контенту. Пошукові системи Google/Яндекс/ Rambler дозволяють перевірити контент на плагіат при введенні ≤ 100 символів в лапках в рядок пошуку. Системи Web analytics – це відстеження, збирання та вимірювання кількісних/якісних даних про відвідуваність інформаційного ресурсу з подальшим їх аналізом. Основне завдання системи Web analytics є оптимізація сайту та ініціатив Інтернет-маркетингу. Система Google Analytics є потужним інструментом відстеження сайтів будь-якого розміру.
Системи управління контентом (CMS)	CMS (Joomla, Business Catalyst, Magento, PrestaShop, OpenCart, osCSS, Drupal, Plone, Mambo, TYPO3, Xaraya, PHP Fusion, PHP-Nuke, Santafox, e107, Opencms, Impresspages_CMS, VaM Shop, SiMan, LiteDiary, Aurus, Kasseler, Promodo) поділяють на Software as a service, Proprietary software, Open source software for Java, Java packages, Microsoft ASP.NET, Per, PHP, File/flat file, Python, Ruby on Rails, CFML/ColdFusion Markup Language, others. Технології content management framework (CMF), які дозволяють створити CMS.
Програми-клієнти	Обмін файлами за протоколом BitTorrent. Дозволяють організувати мережі з однаковими привілеями учасників (одночасно є клієнтами і серверами). Поділяють на крос-платформені, призначені для GNU/Linux/UNIX, Windows, Mac OS або Android.

Теоретичний чинник проблеми опрацювання інформаційних ресурсів в системах електронної контент-комерції передбачає розроблення уніфікованих методів формування/управління/супроводу

контенту на основі досліджень в галузі е-бізнесу та наукових робіт провідних вчених (табл. 2) [2]. У роботі подано вирішення цієї проблеми у вигляді теоретично обгрунтованої концепції та опису відповідних уніфікованих методів шляхом автоматизації процесів опрацювання інформаційних ресурсів, що ґрунтуються на принципах побудови та функціональних можливостях відповідних систем.

Таблиця 2

Основні теоретичні результати в галузі електронної контент-комерції

Метод	Характеристика
Аналіз потоків інформації	У роботах Д.В. Ланде, С.М. Брайчевського, А.Н. Григор'єва, В.Н. Фурашева досліджено та розвинуто математичні моделі електронних інформаційних потоків та їх оцінювання [9].
Аналіз ключових слів	George Kingsley Zipf запропонував емпіричну закономірність розподілу частоти слів природної мови [2, 9].
Фіксація подій	G. Salton і R. Papka запропонували підходи виявлення нових подій в інформаційних потоках [9].
Життєвий цикл контенту	У роботах McKeever Susan, Bob Boiko, Gerry McGovern, JoAnn Hackos, Ann Rockley, Russell Nakano, Bob Doyle, Woods Randy, Halverson описано моделі життєвого циклу контенту [2].
Контент-аналіз	Започаткували методологію B. Matthews, A. Tenni, D. Spiida, D. Wilcox, J. Kaiser, A. Lindesmith, Glaser, Strauss, D. Robertson та активно розвинули J. Naisbitt, Stemler, H. Lasswell, O. Holsti, E. Babbie. F. Joubish запропонував методологію дослідження текстів для визначення авторства, автентичності або сенсу. K. Neuendorf та K. Krippendorff розробили методи кількісного та якісного аналізу тексту. D. McKeone виявила різницю між prescriptive analysis і open analysis тексту [2].
Управління контентом	Корпорації EMC, IBM, Microsoft Alfresco, Open Text, Oracle і SAP розробили специфікації Content Management Interoperability Services (CMIS) на інтерфейс Web-сервісів, покликаних забезпечити взаємодію між системами ECM. Інструментарій Interoperable Content Application взаємодіє з контентом із різних репозиторіїв за допомогою сервісного інтерфейсу і спеціальної надбудови (CMIS Implementation), яка розробляється кожним учасником CMIS самостійно [2].
Комп'ютерна лінгвістики	Аналіз природномовних текстів складається з послідовних процесів як морфологічний, синтаксичний та семантичний аналіз. Для кожного аналізу створено відповідні моделі та алгоритми для розроблення таких засобів опрацювання природної мови, як системи інформаційно-пошукові, машинного перекладу, анотування, морфологічного/синтаксичного/семантичного аналізу, навчально-дидактичні тощо [2–4, 6].

Аналіз останніх досліджень та публікацій.

Інформаційний ресурс у системах електронної контент-комерції є множиною даних з набором властивостей (табл. 3), які є об'єктом дій технології перетворення їх на комерційний контент. Результат застосування однієї технології є інформаційним ресурсом іншої [2].

Таблиця 3

Властивості інформаційних ресурсів у системах електронної контент-комерції

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

Контент (англ. content – зміст, вміст, суть) у галузі інформаційних технологій є формалізованими відомостями і знаннями в системі без детальної специфікації їх властивостей, способів формалізації і впорядкування [2]. Перетворення різнорідних за природою/змістом/походженням даних на узгоджений централізований інформаційний ресурс є однією з важливих проблем

побудови та функціонування систем електронної контент-комерції. Важливими завданнями є забезпечення інформаційних потреб проблемно-орієнтованих елементів системи, підтримання доступу до контенту різних категорій користувачів, дотримання правил цілісності та несуперечності даних, мінімізація/контроль надлишку контенту, здатність до розвитку/зміни внутрішньої організації інформаційного ресурсу, дотримання вимог якості та ефективності контенту [1]. Необхідно забезпечити інваріантність середовища систем електронної контент-комерції до модифікації інформаційних ресурсів у таких змінах [1–2]: способів подання, форматів та внутрішньої організації контенту; середовища зберігання контенту, фізичних одиниць та технічних засобів; вимог користувачів до контенту; поява нових вимог та категорій користувачів; порядку розподілу контенту та способів доступу користувачів.

Виникає проблема створення єдиного концептуального опису інформаційного ресурсу для підтримання зовнішніх/внутрішніх позначень контенту відповідно до їх завдань/вимог/змін. Тому класифікуємо інформаційні ресурси для дослідження їх природних/технологічних/споживчих якостей з метою виявлення характерних/специфічних властивостей, а також закономірностей/особливостей їх формування/застосування. За основу класифікації взято основні властивості контенту (синтаксис, структура та семантика), на основі яких обрано основні фактори класифікації [1–2]: способи подання контенту; методи структурування ресурсу; шляхи доступу до ресурсу; призначення ресурсу. Загальні принципи формування інформаційних ресурсів у системах електронної контент-комерції (рис. 1) визначають порядок і способи відбору інформації із первинних джерел, її фіксації, фільтрування, перетворення до визначеного формату для формування контенту і розміщення в базі даних.

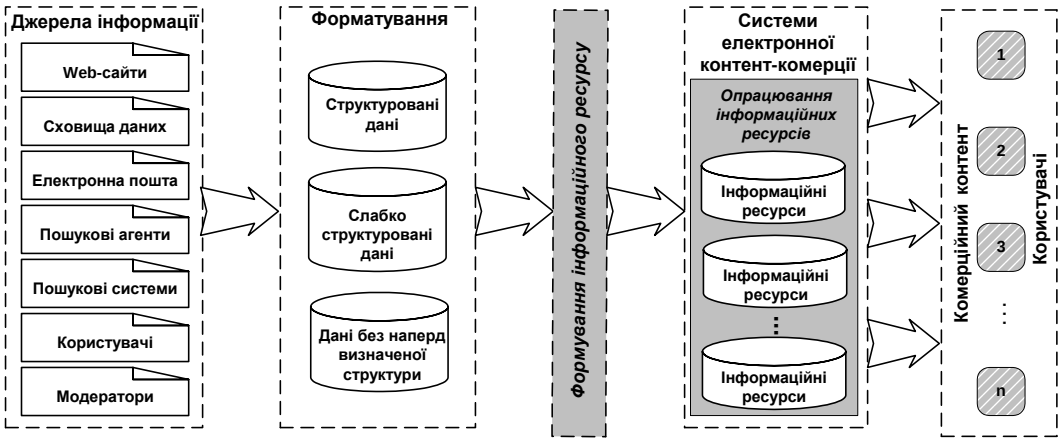


Рис. 1. Порядок формування і використання інформаційних ресурсів у системах електронної контент-комерції

Існує попередньо визначена множина N первинних джерел контенту з фіксованим/змінним складом. Кожне джерело інформації x_i, де $i=\overline{1,N}$, формує множину значень відомостей/знань/ фактів з предметної області системи електронної контент-комерції [2]. Результатом звернення технологічних засобів до джерела x_i є генерування множини значень $X(x_i)$, яку сприймають, фіксують і подають у визначеній формі. Згенеровану кожним джерелом інформації множину значень перетворюють на вхідний набір контенту визначеного формату X_i, де $i=\overline{1,N}$. Кожний набір контенту подається у вигляді структурованих, слабкоструктурованих даних або без визначеного опису структури.

Структурування контенту передбачає формування для кожного набору опису його складу, способів поєднання елементів та їх впорядкування (множини умов U_i, де $i=\overline{1,N}$). Кожний набір контенту є поєднанням множини значень у заданому форматі та множини умов $\langle X_i, U_i \rangle$. У випадку

формування вхідного набору контенту без опису структури $U_i = \varnothing$. Отриманий контент перед форматуванням та збереженням проходить процедуру верифікації/валідації для підтвердження формальної/змістовної коректності/релевантності щодо вимог системи. За невідповідності зазначеним критеріям вилучається.

Формально таку послідовність технологічних процесів подамо як ланцюжок:

$$\begin{aligned} & x_i \to X\left(x_i\right) \to X_i \to \left\langle X_i, U_i \right\rangle \to Verification\left(\left\langle X_i, U_i \right\rangle\right) \to Qualification\left(\left\langle X_i, U_i \right\rangle\right) \to \\ & \to Conversion\left(\left\langle X_i, U_i \right\rangle\right) \to Downloading\left(\left\langle X_i, U_i \right\rangle\right) \to \left\langle X, U \right\rangle, \quad при\ i = \overline{1, N}, \end{aligned} \tag{1}$$

де N – кількість джерел, x_i – i-те джерело, $X\left(x_i\right)$ – множина згенерованих значень, X_i – набір відібраного контенту, $\left\langle X_i, U_i \right\rangle$ – набір контенту із множиною умов, $Verification\left(\left\langle X_i, U_i \right\rangle\right)$ – верифікація, $Qualification\left(\left\langle X_i, U_i \right\rangle\right)$ – кваліфікація , $Conversion\left(\left\langle X_i, U_i \right\rangle\right)$ – перетворення контенту та $Downloading\left(\left\langle X_i, U_i \right\rangle\right)$ – завантаження контенту, $\left\langle X, U \right\rangle$ – ресурс системи.

У табл. 4 подано архітектурні принципи побудови інформаційних ресурсів у системах електронної контент-комерції, [1–2, 5–6, 9–10]. Формують інформаційний ресурс методом гомогенізації/ розподілу/інтеграції (табл. 5) [2].

Таблиця 4

Архітектурні принципи побудови інформаційних ресурсів у системах електронної комерції

Принцип	Властивість інформаційного ресурсу
Системність	Побудова ресурсу вимагає узгодження з вимогами/функціями інших складових системи.
Комплексність	За своїм складом, форматами подання та змістом ресурс є неоднорідним, тому під час його формування поєднують різнотипні елементи в цілісний набір.
Повнота	Склад інформаційного ресурсу забезпечує виконання всіх завдань і функцій системи.
Цілісність	Дотримання структури (забезпечує ізоморфізм станів одиниць даних та зв'язків між ними під час їхнього оновлення), семантики (формує релевантну інтерпретацію кожної одиниці/агрегату даних в усіх станах ресурсу) та функціональності (забезпечує виконання всіх функцій системи у разі зміни станів її ресурсу).
Відкритість	Здатність ресурсу взаємодіяти та обмінюватися контентом з ресурсами інших систем/джерел.
Інтероперабельність	Зміст, способи сприйняття, застосування, інтерпретації та функціонування ресурсу не залежать від платформ і технологій реалізації та є незмінними у разі зміни середовища функціонування системи.
Неперервність	Новий стан інформаційного ресурсу при оновленні напряму залежить від попереднього (є функцією від нього та набору операцій над ним).

Таблиця 5

Способи формування інформаційних ресурсів у системах електронної комерції

Назва	Характеристика
Гомогенізація	Формування єдиного однорідного набору контенту, а саме: отримання знань і відомостей з різноманітних попередньо визначених джерел у довільному неформалізованому вигляді; формалізація/стандартизація отриманих із джерел та подання у вигляді тематичного контенту; верифікація контенту (перевірка на відповідність синтаксичним правилам, змісту, на повноту, відсутність повторень тощо); завантаження контенту (ручне/автоматизоване розміщення контенту в базі даних з додатковим перетворенням із проміжного формату на визначений).
Розподіл	Автономне формування і використання локальних складових (множин контенту визначеного змісту, структури і призначення із набором технологій для створення/підтримання/управління/доступу) єдиного глобального набору контенту.
Інтеграція	Інтегрований ресурс утворює узгоджену множину значень/відомостей/знань/фактів, придатних для спільного використання. Всі складові інтегрованого ресурсу є доступними для використання в межах єдиного набору методів, засобів і технологій. Ефект від використання глобального інтегрованого інформаційного ресурсу перевищує сумарний ефект від застосування всіх локальних ресурсів (складових процесу інтеграції).

У результаті виконання послідовності кроків формування ресурсу за методом гомогенізації утворюють/поповнюють множину контенту. Завданням комплексу засобів управління розподіленими ресурсами є організація узгодженої роботи і взаємодії локальних технологічних засобів для забезпечення спільного використання всіх/частини локальних ресурсів. Перші два методи мають низку недоліків (табл. 6). Основними чинниками розвитку методу інтеграції є зростання частки слабко структурованих даних та без попереднього опису структури у разі збільшення обсягів ресурсів.

Таблиця 6

Порівняння способів формування інформаційних ресурсів у системах електронної комерції

Назва	Переваги	Недоліки
Гомогенізація	Застосовують у таких випадках: існує єдиний спосіб подання контенту та набір засобів/технологій для його підтримання; джерела є попередньо визначеними, ідентифікованими і специфікованими; кількість джерел є невеликою; інформація з джерел придатна для приведення до єдиних уніфікованих форматів.	Вимоги формування ресурсів є жорсткими, їх виконання є проблемним і часто неможливим.
Розподіл	Основної сферою застосування розподілених ресурсів є Інтернет, хоча підхід часто застосовують при побудові корпоративних систем великого масштабу (розподілені бази даних).	Зростає частка розподілених ресурсів слабкоструктурованого типу чи без наперед визначеної структури, тому застосовують спеціальні системи.
Інтеграція	Дозволяє поєднати контент різного формату, змісту і походження у єдиному узгодженому наборі; об'єднувати контент без узагальненого форматування; будувати віртуальні користувацькі зображення контенту, що не залежать від їх реального вигляду; оперувати різним контентом у його поєднанні; динамічно доповнювати, змінювати і перетворювати контент та його описи; забезпечити єдині технології сприйняття та застосування великої кількості різного контенту.	Немає

Особливістю методу інтеграції є однорідний/неоднорідний результат, який потребує фізичного переміщення даних або використовує їх віртуальні зображення, єдиний глобальний опис даних чи множину локальних описів. У такий спосіб вдається сформувати інформаційний ресурс як єдину цілісну узгоджену множину даних, придатну для розв'язання широкого кола різнопланових проблем/задач у різних напрямах розвитку галузі електронної контент-комерції (табл. 7).

Таблиця 7

Набір актуальних задач у різних напрямах розвитку галузі електронної контент-комерції

Назва	Пояснення
Інтернет як платформа	Найбільших успіхів досягають в проектах, які використовують Інтернет як платформу для е-комерції [6, 12–19].
Надання послуг/сервісів	Надання послуг/сервісів в Інтернеті, а не продаж ПЗ із встановленням на комп'ютері користувача [6]. Тут спостерігається найбільша активність користувачів, а купівля чи вільне завантаження локального ПЗ знаходиться на периферії. Стрімкий успіх Google (надає послуги) порівнянно з Netscape (продає ПЗ) свідчить про правильно обрану стратегію напередодні домінування Web 2.0 [6, 12–19].
Багатомовність	Під час обслуговування користувачів з усього світу зростає необхідність оперативного перекладу інтерфейсу та вмісту контенту/сайту. Популярні/поширені інтернет-сервіси підтримують багатомовність [12–19].
Незалежність від користувача	Знімає необхідність регулярних офіційних релізів/оновлень [16]. Розвивається безкоштовне ПЗ з відкритим кодом для самостійного створення користувачами сайтів/порталів згідно з вимогами Web 2.0 [6, 12–19].
Лінгвістичні складові системи	Активний розвиток застосувань для редагування інтерфейсу в реальному часі редакторами/користувачами [6]. Зростає попит на електронні словники, системи автоматичної перевірки орфографії/граматики, автоматичного переклад із вмонтуванням в адмінівську/редакторську частину систем [2–9].

Продовження табл. 7

Назва	Пояснення
Нежорстка категоризація	Групування/структуризація контенту за семантикою/змістом та ключовими полегшують реалізацію алгоритмів контентного пошуку [2–3, 6–9].
Контентний пошук	Впроваджуються нові підходи до подання результатів пошуку для користувачів, які більше спираються на семантику. Створюються різноманітні лінгвістичні фільтри для відсіювання випадкових сторінок та спаму, пошуку за синонімами, перевірки наявності вірусів тощо [2, 6-9, 11].
Використання баз даних	Доступ до унікальних баз даних (термінологічні/орфографічні, парадигматичні, для автоматичного перекладу, корпуси текстів тощо) за оплату [6, 16].
Співпраця між мережевими проектами	Технології RSS дозволяють отримувати дані з сайту без його відвідування, відображати ці дані на інших сайтах чи локально в програмі [1–3]. Взаємоінтеграція проектів вимагає прикладних лінгвістичних програм, зокрема, семантичного аналізу (наприклад, у сфері автоматичного чи автоматизованого підбору і групування новин з різних Інтернет-ресурсів) [1–9].
Користувач як активний учасник проекту	Популярність і/або комерційна успішність прямо залежать від кількості залучених до проекту користувачів, наприклад, як автора текстів (Вікіпедія, Живий журнал) або медіа (YouTube, Flickr, Photo.net), коментарів/відгуків на товар (Amazon), продавця/покупця (eBay), поповнювача бази даних проекту у певній сфері знань (Open Library), власника контенту (BitTorrent, Вікіпедія) тощо [1–3, 5–6]. Увага до користувача сприяє зростанню ваги лінгвістичного наповнення на сайті. Участь користувача у проекті передбачає надання йому певного лінгвістичного інструментарію (програми для редагування тексту з автоматичною перевіркою орфографії, термінологічними та перекладними словниками з відповідної галузі знань тощо) [13].
Зацікавлення користувачів	Застосовують різні психологічні/психолінгвістичні прийоми, максимально спрощують інтерфейс та роблять його зручним [15], відсутність реклами, прискорення швидкості завантаження сторінок (AJAX), технології перетворення сайту на аналог звичайної програми з великим набором функцій і зручністю користування, наприклад, поштові системи (Gmail), Web-редактори текстів/зображень, офісні програми, електронні словники та програми автоматичного перекладу тексту тощо [1, 2, 6].
Співпраця з користувачами	Вибудовують певну ієрархію, де користувачі, яким довіряють, є модераторами/адміністраторами проекту. Використовують напрацювання психолінгвістики та інших суміжних з лінгвістикою дисциплін [2, 6].
Електронні бібліотеки	Google Books та Open Library містять повнотекстові книги користувачіі [6], які використовують системи автоматичного/автоматизованого опрацювання/розпізнавання текстів, бібліографічні і каталогізації.
Дистанційне навчання	Прикладні лінгвістичні програми, пов'язані з лінгвістикою, наприклад, дистанційне навчання природним мовам (Mova.info) [1–3, 6].
Блоги	Присутній комунікативний аспект (користувачі коментують записи, знайомляться, спілкуються тощо) [6], що призводить до геометричного зростання різного тексту в Інтернет і потреби у нових принципах його автоматичного опрацювання пошуковими системами та розвитку програм редагування тексту [2].
Соціальні мережі	Закрита для сторонніх і захищена спільнота (Facebook, ВКонтакте, Одноклассники) за професійними або іншими інтересами [6, 19]. Існують проекти, коли створюється віртуальне робоче місце користувача, яке дає змогу утримувати потрібні йому сайти в одному місці без потреби їх відвідування [2].
Комп'ютерна лінгвістика	Зростає комунікативний аспект Інтернет, який напряму пов'язаний з мовою, тому зростає і вага досліджень та розроблень з комп'ютерної лінгвістики у функціонуванні систем електронної контент-комерції [6, 11].

Виділення проблем

Комерційний контент – це об'єкт купівлі/продажу між учасниками е-комерції [2], наприклад, інформаційні блоки, які поділяють на синдикати (наприклад, блок погоди), анонси матеріалів з інших розділів сайту або інших сайтів (з посиланням), довідкова інформація (наприклад, святкові дати, анонс заходу, розклад руху поїздів), розважальна інформація (наприклад, анекдот дня), реклама, кнопки і посилання інформаційних партнерів, кнопки статистики.

Формальна модель життєвого циклу комерційного контенту – це шістка

$$S = \langle X, Creation, C, Processing, Distribution, Y \rangle, \quad (2)$$

де $X=\{x_1,x_2,\ldots,x_{n_X}\}$ – множина вхідної інформації, $C=\{c_1,c_2,\ldots,c_{n_C}\}$ – множина контенту, *Creation* – функція створення контенту, *Processing* – функція опрацювання контенту, *Distribution* – функція поширення контенту та $Y=\{y_1,y_2,\ldots,y_{n_Y}\}$ – множина вихідної інформації (рис. 2).

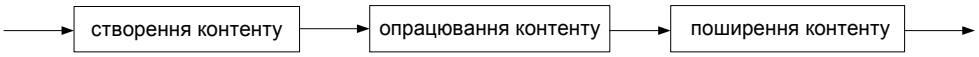


Рис. 2. Модель системи розподілення контенту в системах електронної контент-комерції

Формулювання мети

Одним із видів системи електронної контент-комерції є інтернет-журнал [5], тобто множина структурованого контенту (наприклад, електронні видання у вигляді статей, анонсів, дайджестів, книг, репортажів, блогів, коментарів тощо) в інформаційному ресурсі, призначена для задоволення потреб цільової аудиторії та сприйняття за допомогою відповідних програмних засобів через Інтернет (рис. 3). Переваги впровадження Інтернет-журналу: компактний; економічно вигідний; миттєво доставляє будь-який контент у будь-яку точку світу та в будь-який час; не прив'язаний до проблеми тиражування; зменшена до мінімуму відстань між автором і читачем; нові шляхи донесення інформації до читача. Автоматичне опрацювання електронних видань зменшує час пошуку необхідного контенту лише під час введення відповідної адреси або ключових слів до пошукової системи [2].

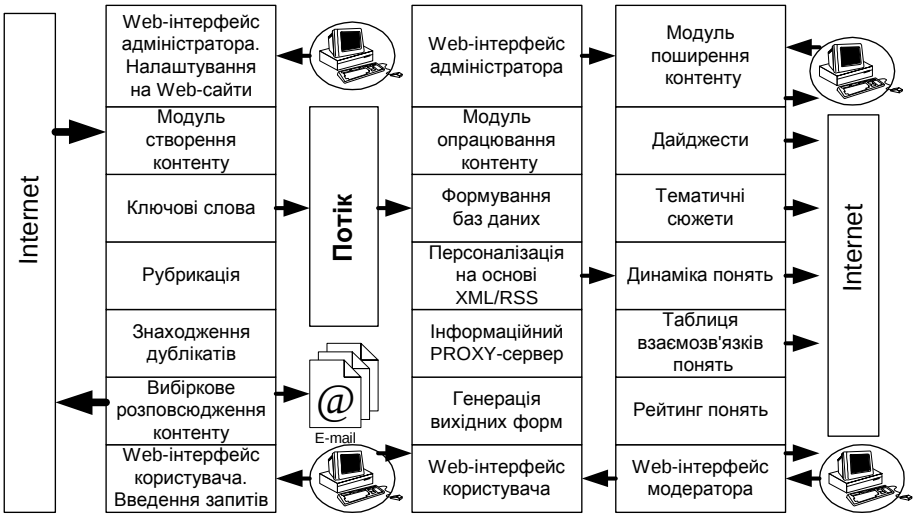
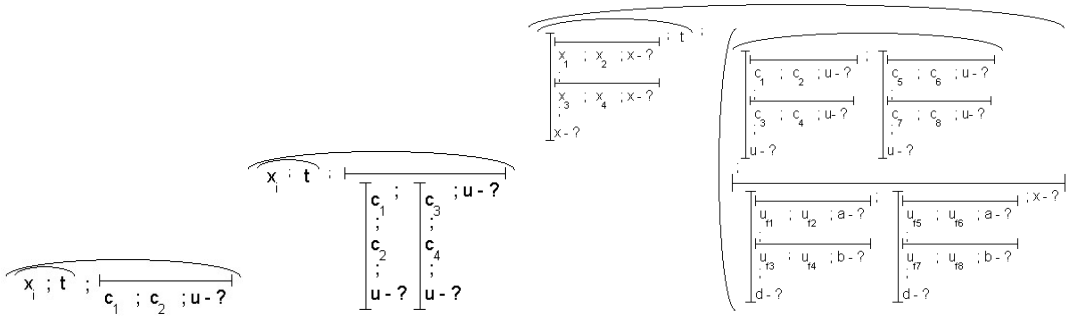
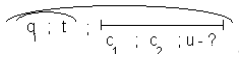


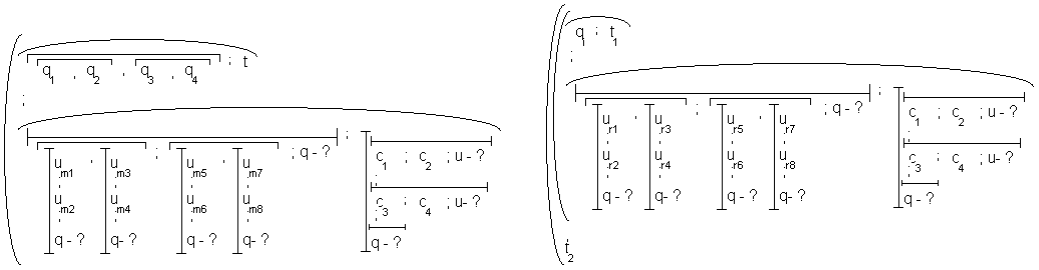
Рис. 3. Схема взаємодії модулів систем електронної контент-комерції, розроблена за [9]

Процес створення контенту описується функцією вигляду $\vec{c}(x_i,t)=Creation(\overrightarrow{u_C},x_i,t)$, де $\overrightarrow{u_C}(x_i)$ – множина умов створення контенту, тобто $\overrightarrow{u_C}(x_i)=\left(u_{c_1}(x_i),u_{c_2}(x_i),\ldots,u_{c_{n_{Uc}}}(x_i)\right)$. Контент створюють як $c_j=\left\{\bigcup u_{c_k}\,\middle|\,(x_i\in X)\wedge(\exists u_{c_k}\in U_F),U_C=U_{Cx}\vee U_{C\bar{x}},i=\overline{1,m},k=\overline{1,n}\right\}$, тобто



(3)

Процес опрацювання контенту описується як $\vec{c}(q_i,t) = Processing\left(\overrightarrow{u_P}, q_i, t\right)$, де $Q = \left\{q_1, q_2, \ldots, q_{n_Q}\right\}$ – множина запитів користувачів [2], $\overrightarrow{u_P}(q_i)$ – множина умов опрацювання контенту, тобто $\overrightarrow{u_P}(q_i) = \left(u_{p_1}(q_i), u_{p_2}(q_i), \ldots, u_{f_{nUp}}(q_i)\right)$. Опрацювання контенту відбувається як $c_j = \left\{\bigcup u_{p_k} \middle| (q_i \in Q) \wedge \left(\exists u_{p_k} \in U_P\right), U_P = U_{Mq} \vee U_{P\bar{q}}, i = \overline{1,m}, k = \overline{1,n}\right\}$, тобто  .



(4)

Процес поширення контенту описується як $\vec{y}(t + \Delta t) = Distribution\left(\overrightarrow{u_D}, \vec{c}, q_i, t, \Delta t\right)$, де $\overrightarrow{u_D}\left(q_i, \vec{c}\right)$ – множина умов поширення контенту $\overrightarrow{u_D}\left(q_i, \vec{c}\right) = \left(u_{d_1}\left(q_i, \vec{c}\right), u_{d_2}\left(q_i, \vec{c}\right), \ldots, u_{d_{n_{Ud}}}\left(q_i, \vec{c}\right)\right)$ при

$$y_j = \left\{\bigcup u_{d_k} \middle| (q_i \in Q) \wedge \left(\vec{c} \in C\right) \wedge \left(\exists u_{r_k} \in U_R\right), U_D = U_{Dc} \vee U_{D\bar{c}}, i = \overline{1,m}, k = \overline{1,n}\right\}. \quad (5)$$

Відомим методом опрацювання тексту є контент-аналіз, тобто дослідження змісту текстових масивів/продуктів комунікативної кореспонденції (наприклад, коментарі, форуми, електронне листування, статті тощо). Контент-аналіз полягає в пошуку інформації [7, 8, 11] за змістовими одиницями (словосполучення, речення, тема, ідея, автор, персонаж, соціальна ситуація, частина тексту, кластеризована за змістом категорії аналізу) та інтерпретація результату (табл. 8).

Таблиця 8

Основні процедури формалізованого методу контент-аналізу

Назва	Особливості процедури формалізованого методу контент-аналізу
Виявлення одиниць	Залежно від змісту/цілей/завдань/гіпотез дослідження формують множину змістовних одиниць.
Виділення одиниць	Одиниці рахунку можуть збігатися (підрахунок частоти згадки виділеної смислової одиниці) або не збігатися (формує дослідник на основі аналізованого матеріалу) з одиницями аналізу.
Процедура підрахунку	Класифікації за виділеними угрупованнями із застосуванням спеціальних формул (наприклад, оцінення питомої ваги змістовних категорій у тексті), статистичних розрахунків зрозумілості/атрактивності тексту.
Розроблення протоколу	Класифікатором є загальна таблиця із категоріями/одиницями аналізу, яка гранично чітко фіксує одиниці виразу кожної категорії. Наприклад, категоріями аналізу в анкеті є запитання, а одиниці аналізу – відповіді. Контент-аналіз містить: відомості про контент (автор, час видання, обсяг тощо); підсумки його аналізу (кількість вживання в ньому певних одиниць аналізу і висновки щодо категорій аналізу). Протокол заповнюють у закодованому вигляді для компактності подання інформації та швидкого порівняння результатів аналізу різного контенту на основі підрахунку даних всіх реєстраційних карток (кодувальні матриці, де зазначається кількість одиниць рахунку та характеризуються одиниці аналізу).
Розроблення таблиці	Тип таблиці визначається етапом дослідження, наприклад, у вигляді системи скоординованих і субординованих категорій аналізу, яка ззовні нагадує анкету: кожна категорія (питання) передбачає ряд ознак (відповідей), за якими квантифікується зміст тексту.
Розроблення матриці	Якщо обсяг вибірки ≥ 100 одиниць, то аналізують набір матричних листів. Інакше проводиться двовимірний/багатомірний аналіз, де для кожного тексту будується кодувальна матриця.

Одиниці аналізу розглядають на тлі ширших лінгвістичних/змістовних структур, що вказують на характер сегментування тексту, в межах якого ідентифікується присутність/відсутність контекстуальних одиниць. Наприклад, для одиниці аналізу *слово* контекстуальна одиниця – *речення*. Одиниця рахунку є кількісною мірою одиниці аналізу, що дає змогу реєструвати частоту (регулярність) появи ознаки категорії аналізу в тексті (кількість певних слів або їх поєднань, рядків, друкованих знаків, сторінок, абзаців, авторських аркушів, площа тексту тощо). Параметри вибірки визначаються завданнями і масштабами дослідження.

Аналіз отриманих наукових результатів

Контент є основою інтернет-журналу, за якою користувач шукає необхідну інформацію. Але текстів, які переповнені ключовими словами, не завжди достатньо для того, щоб користувач отримав потрібну інформацію. До того ж виділення ключових слів для кожної статті є довготривалим і трудомістким процесом. За допомогою формалізованого методу контент-аналізу процес є повністю автоматизованим і відбувається у разі додавання автором нової статті. Визначають статті, подібні до тих, які переглядав користувач, та виводять їх для перегляду. Використовуючи ключові слова, користувач отримує інформацію, не зовсім подібну або зовсім не подібну на ту, яка його цікавить (для підвищення популярності статті автори додають ключові слова з різних тем). Перевагою використання контент-аналізу є визначення наявності контенту під конкретний запит, наприклад, визначення відсутності контенту з певної тематики і спрямування роботи авторів на розвиток цього питання.

В інтернет-журналі розподіляють статті. На вхід подають неопрацьовані статті, які скеровують на блок опрацювання статей. За правилами у вигляді параметрів пошуку, дати та ключових слів переглянутих статей статті розподіляють на категорії (популярні, подібні, відібрані в результаті пошуку або останньо переглянуті) та розміщують їх у базі даних. Процес контролює адміністратор. Опрацьовані статті подають на блоки для пошуку статей популярних (з найбільшою кількістю переглядів), останньо переглянутих (за датою публікації в межах періоду від заданої до поточної дати), параметричних (за введеними користувачем параметрами) та подібних (за ключовими словами з раніше переглянутих користувачем статей) і виведення результату.

Опрацьовані статті потрапляють на блок пошуку збігів з ключовими словами, переглянутими користувачем раніше. Сформований список статей сортується за кількістю збігів, щоби відібрати ті, які є найподібніші на переглянуті автором. Відбираються перші статті, кількість яких задається в налаштуваннях системи. Далі роботи автора статті, яка переглядається в цей час, переміщаються на початок списку і виводяться для перегляду користувачем. На рис. 4 подано діаграму класів, яка описує сукупність об’єктів інтернет-журналу (*користувацький інтерфейс*, *Web-сторінка* та групу класів *Сервер*, яка містить клас *СУБД*, *Web-сервер* і компоненти пошуку за параметрами, подібних нових та популярних статей) та їх відносин з погляду керівника проекту.

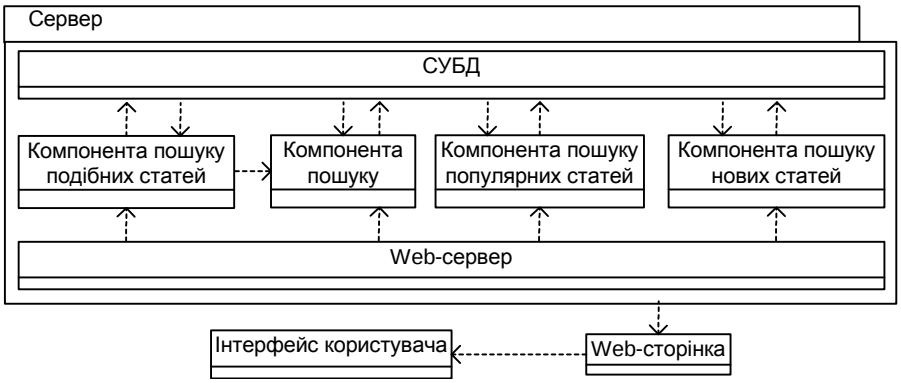


Рис. 4. Діаграма класів системи електронної контент-комерції типу інтернет-журнал

На рис. 5 подано діаграму прецедентів та сценарій покрокового виконання завдань інтернет-журналу. Авторську статтю опрацьовують та зберігають у базі даних. Параметри пошуку користу-

вача потрапляють у блок пошуку статей за параметрами, який формує запит в базу даних і список статей, які відповідають параметрам пошуку, та передає їх користувачу і в блок формування списку статей, переглянутих користувачем. Останній відправляє сформований список у блок формування списку подібних статей, який видає результат користувачу. Блоки формування списку нових і/або популярних статей формують запити в базу даних та виводять сформовані списки для перегляду.

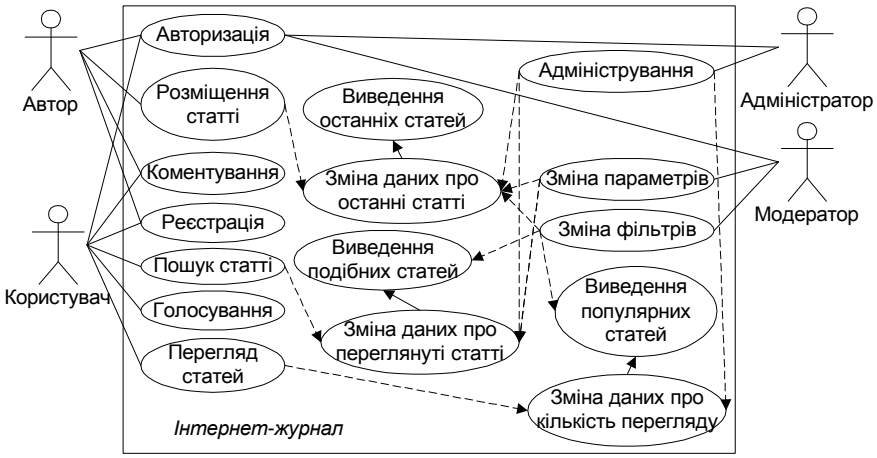


Рис. 5. Діаграма прецедентів системи електронної контент-комерції типу Інтернет-журнал

Актор *Автор* реєструється/авторизується в системі, коментує/розміщує статті. Актор *Користувач* реєструється/авторизується в системі, оцінює, переглядає, здійснює пошук та коментує статті. Перегляд статей користувачем приводить до зміни даних про кількість переглядів статті, що передбачає виведення оновленого списку популярних статей. Також перегляд та пошук за параметрами змінюють дані про переглянуті статті, що передбачає виведення оновленого списку подібних статей. Актор *Модератор* формує/модифікує параметри пошуку статей. Система формує та виводить список переглянутих користувачем статей. На рис. 6, а подано діаграму розгортання інтернет-журналу.

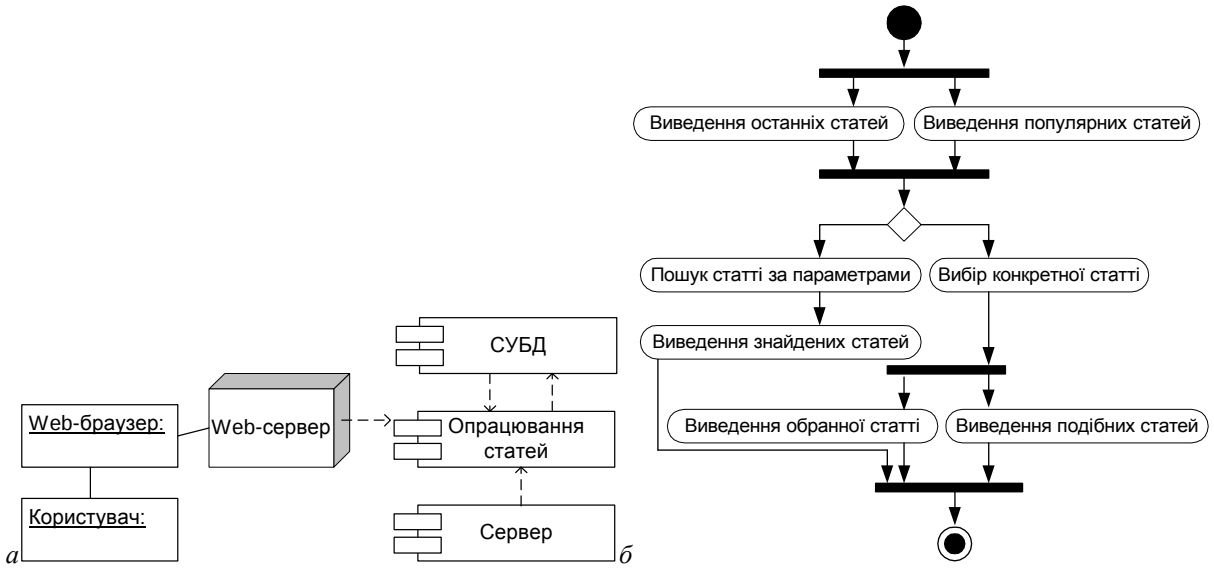


Рис. 6. Даграма розгортання (а) та діяльності (б) системи електронної контент-комерції типу інтернет-журнал

Після авторизації автор (користувач) відправляє статтю на опрацювання. Періодично з базою даних відбувається обмін інформацією для формування списку нових/популярних, таких, що відповідають параметрам пошуку, і подібних на переглянуті користувачем. Всі ці чотири списки

статей відправляють для перегляду користувачу. На основі отриманих списків користувач робить новий запит, і механізм повторюється. Після виконання операцій неопрацьована стаття автора перетворюється на опрацьовану статтю і зберігається в базі даних. Її використовують для пошуку за розміром, ключовими словами, автором та популярністю. На рис. 6, б подано діаграму активності Інтернет-журналу. Для користувача видаються списки останніх або популярних статей, які мають найбільше переглядів. Користувач може обрати конкретну статтю для перегляду або здійснити пошук за параметрами, що приводить до виведення обраної/подібних/знайдених статей. На рис. 7 подано діаграму послідовностей для інтернет-журналу, а на рис. 8 – діаграму компонентів системи та залежностей між ними.

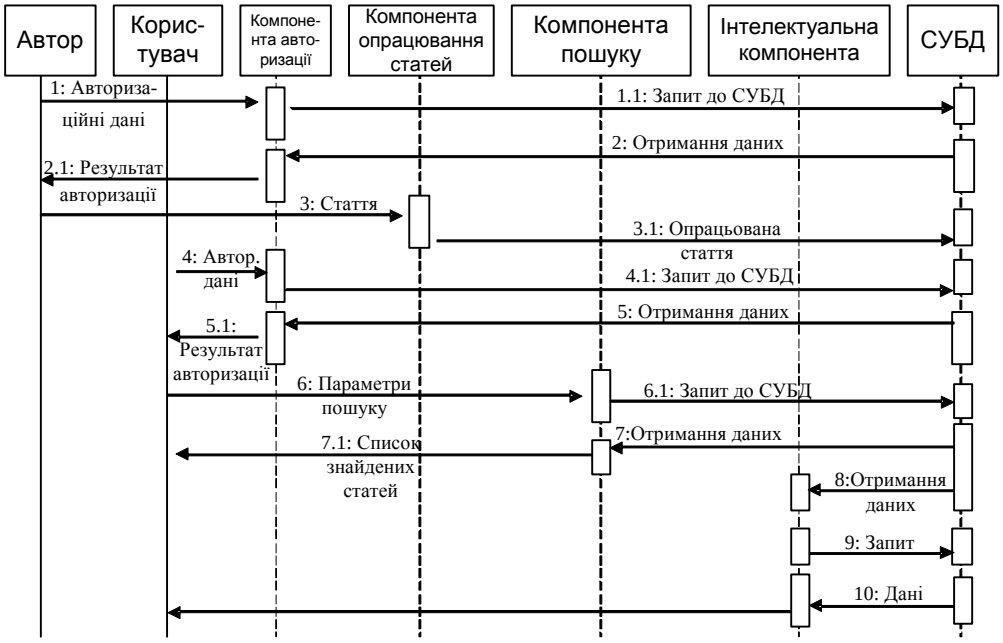


Рис. 7. Діаграма послідовностей системи електронної контент-комерції типу інтернет-журнал

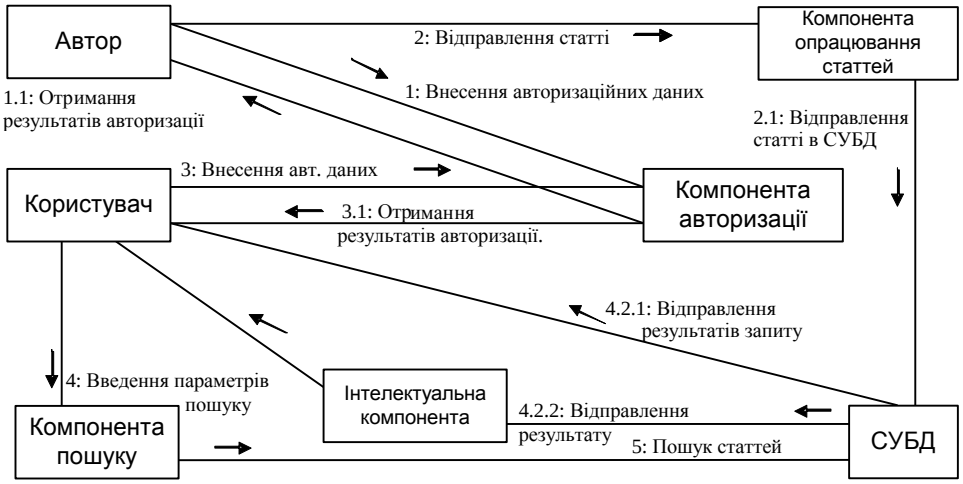


Рис. 8. Діаграма компонентів системи електронної контент-комерції типу інтернет-журнал

Висновки і перспективи подальших наукових розвідок

Інтернет-журнал є типовим прикладом системи електронної контент-комерції, яка належить до керованих кібернетичних систем і має визначений набір властивостей. Контент у вигляді статей є основою інтернет-журналу, за яким користувач шукає необхідну йому інформацію. Але текстів, які переповнені ключовими словами, не завжди достатньо для отримання користувачем потрібної інформації. Виділення ключових слів до кожної статті є довготривалим і трудомістким процесом.

За допомогою формалізованого методу контент-аналізу процес є повністю автоматизованим і відбувається у разі додавання автором нової статті. За цим методом визначають статті, подібні до тих, які переглядав користувач. Перевагою використання контент-аналізу є визначення наявності контенту під конкретний запит, наприклад, визначення відсутності контенту за певною тематикою і спрямування роботи авторів на розвиток цього питання.

1. Береза А.М. Електронна комерція / А.М. Береза // К.: КНЕУ. – 2002 р. 2. Берко А.Ю. Системи електронної контент-комерції: Монографія / А.Ю. Берко, В.А. Висоцька, В.В. Пасічник. – Львів: Вид-во Нац. ун-ту "Львівська політехніка", – 2009. – 612 с. 3. Дарчук Н.П. Комп'ютерна лінгвістика (автоматичне опрацювання тексту) / Н.П. Дарчук – К.: Видавничо-поліграфічний центр "Київський університет", 2008. – 351 с. 4. Дерба С.М. Словник з української термінології прикладної (комп'ютерної) лінгвістики / Дерба С.М. – К.; 2007. – 325 с. 5. Костишин О.М. Інструментальна система для створення онлайнових електронних журналів: Мат-ли міжн. наук. конф. "Інтелектуальні інформаційні технології у бібліотечній справі": (український мовно-інформаційний фонд НАН України, Київ) / О.М. Костишин, Н.М. Сидорчук. – К.: Національна бібліотека України ім. В.І. Вернадського, – 2005. – [Електронний ресурс] Режим доступу: http://www.nbuv.gov.ua /new/05_kiev/05komoez.html. 6. Кузьменко Д. Комп'ютерна лінгвістика і Web 2.0 / Д. Кузьменко // Studia Linguistica. Vol. II. – К.: ВПЦ "Київський університет", 2009. – С. 214–219. [Електронний ресурс] – Режим доступу: http://kuzmenko.org.ua/uk/web20. 7. Манаєв О.Т. Контент-аналіз як метод дослідження / Манаєв О.Т. // Псі-фактор. – Режим доступу: http://psyfactor.org/lib/content-analysis3.htm. 8. Назаров М.М. Контент-аналіз медіа текстів: за матеріалами книги "Массовая коммуникация и общество" / М.М. Назаров // ψ-фактор. – 2004. – Режим доступу: http://psyfactor.org/lib/content-analysis2.htm. 9. Ландэ Д.В. Основы моделирования и оценки електронных информационных потоков: монография / Д.В. Ландэ, В.М. Фурашев, С.М. Брайчевский, О.М. Григорьев // К.: ТОВ "Інжиніринг", 2006. – 348 с. 10. Советов Б.Я. Моделирование систем (2-е изд.) / Б.Я. Советов, С.А. Яковлев – М.: Высшая школа, 1998. 11. Хорошилова Т. Зміст методики "контент-аналіз" / Т. Хорошилова // Прикладна лінгвістика. – Режим доступу: http://studentstpl.ucoz.ru/publ/teorija_vozdejstvija/ metodika_kontent_analiza/ soderzhanie_ metodiki_kontent_analiz/12-1-0-116. 12. Critical Perspectives of Web 2.0. Special issue of First Monday. – Vol. 13, #3, 2008. – Режим доступу: http://www.uic.edu/htbin/ cgiwrap/bin/ojs/index.php/fm/issue/view/263/ showToc. 13. Graham P. Web 2.0 / P. Graham. – Nov. 2005. – http://www.paulgraham.com/web20.html. 14. Hinchcliffe D. The State of Web 2.0 / D. Hinchcliffe. – 04.02.2006. – Режим доступу: http://web2.wsj2.com/the_state_of_web_20.htm. 15. MacManus R. Web 2.0 for Designers / R. MacManus, J. Porter // Digital Web Magazine. – May 4, 2005. – Режим доступу: http://www.digital-web.com/articles/web_2_for_designers. 16. O'Reilly T. What Is Web 2.0. – 09.30.2005. – Режим доступу: http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html. 17. Scholz T. Market Ideology and the Myths of Web 2.0 // First Monday. – Vol. 13, #3, 2008. – Режим доступу: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2138/1945. 18. Singel R. Are You Ready for Web 2.0? / R. Singel. – 10.06.2005. – Режим доступу: http://www.wired.com/news/technology/ 0,1282,69114,00.html. 19. Vickery G. Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking / G. Vickery, S. Wunsch-Vincent // OECD. – 2007. – Режим доступу: http://www.oecd.org/ document/40/0,3343,en_2649_201185_ 39428648_ 1_1_1_1,00.html.