УДК 004.738.5

Я. П. Кісь, В. А. Висоцька, Л. Б. Чирун, В. М. Фольтович
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж.

ЗАСТОСУВАННЯ КОНТЕНТ-АНАЛІЗУ ДЛЯ ОПРАЦЮВАННЯ ТЕКСТОВИХ МАСИВІВ ДАНИХ

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

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

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

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

Вступ. Загальна постановка проблеми

Розроблення технології опрацювання текстових масивів даних на Web-ресурсах є актуальним з огляду на такі фактори, як недостатність теоретичного обґрунтування методів опрацювання потоків контенту і потреба в уніфікації програмних засобів опрацювання інформаційних ресурсів Інтернет [1, 4]. Практичний чинник опрацювання інформаційних ресурсів у системах електронної контет-комерції (СЕКК) пов’язаний з вирішенням завдань формування, управління та супроводу зростаючих обсягів комерційного контенту в Інтернеті, активним розвитком електронного бізнесу, швидкими темпами поширення доступності до Інтернету, розширенням набору інформаційних товарів та послуг, зростанням попиту на комерційний контент [1–4]. Принципи та ІТ електронної контент-комерції застосовують для створення інтернет-магазинів (продаж eBooks, Software, video, music, movies, picture), систем on-line (газети, журнали, дистанційне навчання, видавництва) та off-line продаж контенту (copywriting services, Marketing Services Shop, RSS Subscription Extension), cloud storage та cloud computing [1]. У цьому напрямі працюють такі провідні світові виробники засобів опрацювання інформаційних ресурсів, як Apple, Google, Intel, Microsoft, Amazon [1-4].

Аналіз останніх досліджень та публікацій

Теоретичний чинник опрацювання інформаційних ресурсів у СЕКК пов’язаний із розробленням ІТ опрацювання комерційного контенту. В наукових роботах Д. Ланде, С. Брайчевського, А. Григор’єва та В. Фурашева досліджено та розвинуто математичні моделі електронних інформаційних потоків [1, 5, 6]. Г. Зіпф (G. Zipf) запропонував емпіричну закономірність розподілу частоти слів природної мови [1], а Дж. Селтон (G. Salton) та Р. Папка (R. Papka) – виявлення нових подій в потоках контенту. В роботах Б. Дойл (B. Doyle), Б. Бойка (B. Boiko), С. Макківер (S. McKeever), Дж. Макговерн (G. McGovern), Дж. Хаскос (J. Hackos), Е. Роклі (A. Rockley), Р. Накано (R. Nakano), Р. Вудс (R. Woods), Халверсон (Halverson) описані моделі життєвого циклу контенту [1, 9, 10]. Методологію контент-аналізу започаткували А. Тенні (A. Tenni), Б. Метьюз (B. Matthews), Д. Спіїда (D. Spiida), Ж. Кайзер (J. Kaiser), Б. Гласер (B. Glaser),

. Стросс (A. Strauss) та активно розвинули Г. Лассуел (H. Lasswell), О. Холсті (O. Holsti), В. Іванов, М. Сорока, А. Федорчук [1–8]. Ф. Джобіш (F. Joubish) запропонував методологію дослідження текстів для визначення авторства, автентичності або сенсу. К. Нойендорф (K. Neuendorf) та К. Кріпендорф (K. Krippendorff) розробили методи кількісного та якісного аналізу текстового контенту [6]. В роботах В. Корнєєва, А. Ф. Гарєєва, С. В. Васютіна, В. В. Райха запропоновано методи інтелектуального опрацювання текстової інформації. Корпорації EMC, IBM, Microsoft Alfresco, Open Text, Oracle і SAP розробили специфікації Content Management Interoperability Services на інтерфейс Web-сервісів для взаємодії систем управління контентом е-бізнесу [11].

Виділення проблем

З наукового погляду цей сегмент ІТ є малодослідженим. Кожний окремий проект реалізують практично з початку, фактично на основі своїх ідей та рішень. У літературі надзвичайно мало висвітлено суттєві теоретичні обґрунтування, дослідження, висновки, рекомендації, узагальнення для проектування СЕКК та опрацювання інформаційних ресурсів у таких системах. Виникла потреба в аналізі, узагальненні та обґрунтуванні підходів до реалізації електронної комерції та побудови СЕКК. Актуальною є задача створення комплексу технологічних засобів на основі теоретичного обґрунтування методів, моделей і принципів опрацювання інформаційних ресурсів в СЕКК, побудованих за принципом відкритих систем, які дають змогу керувати процесом збільшення обсягів реалізації комерційного контенту. Аналіз наведених чинників дає змогу зробити висновок про існування певного протиріччя між активним розвитком і поширенням ІТ та СЕКК, з одного боку, та порівняно незначним обсягом наукових досліджень з цієї тематики та їх локальністю – з іншого. Це протиріччя породжує проблему стримування інноваційного розвитку сектора електронної контент-комерції через створення і запровадження відповідних новітніх прогресивних ІТ, що негативно впливає на темпи зростання цієї частки ринку.

Формулювання мети

У межах загальної проблеми актуальною є задача розроблення науково обґрунтованих методів опрацювання інформаційних ресурсів електронної контент-комерції та побудови на їх основі технологічних програмних засобів для створення, поширення і сталого розвитку СЕКК. У роботі проведено дослідження з метою визначення закономірностей, особливостей та залежностей у процесах опрацювання інформаційних ресурсів в СЕКК.

Аналіз отриманих наукових результатів

Збільшення обсягу контенту і швидкості його поширення (рис. 1) сприяє формуванню контентних потоків, аналіз яких вимагає використання нового інструментарію опрацювання інформаційних ресурсів у СЕКК для формування, управління та супроводу комерційного контенту.

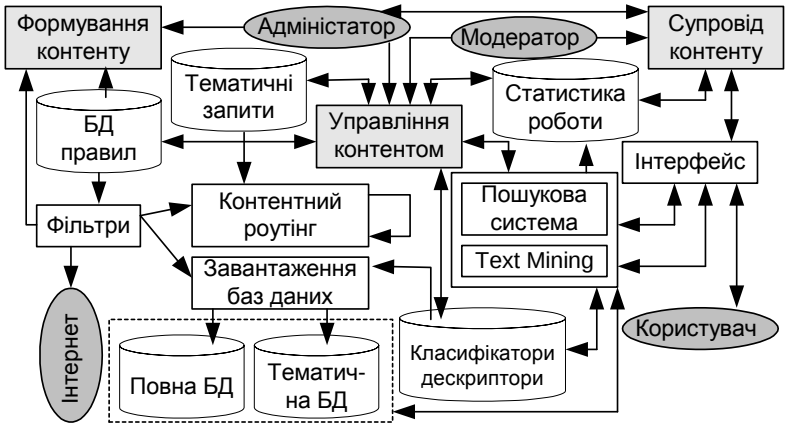
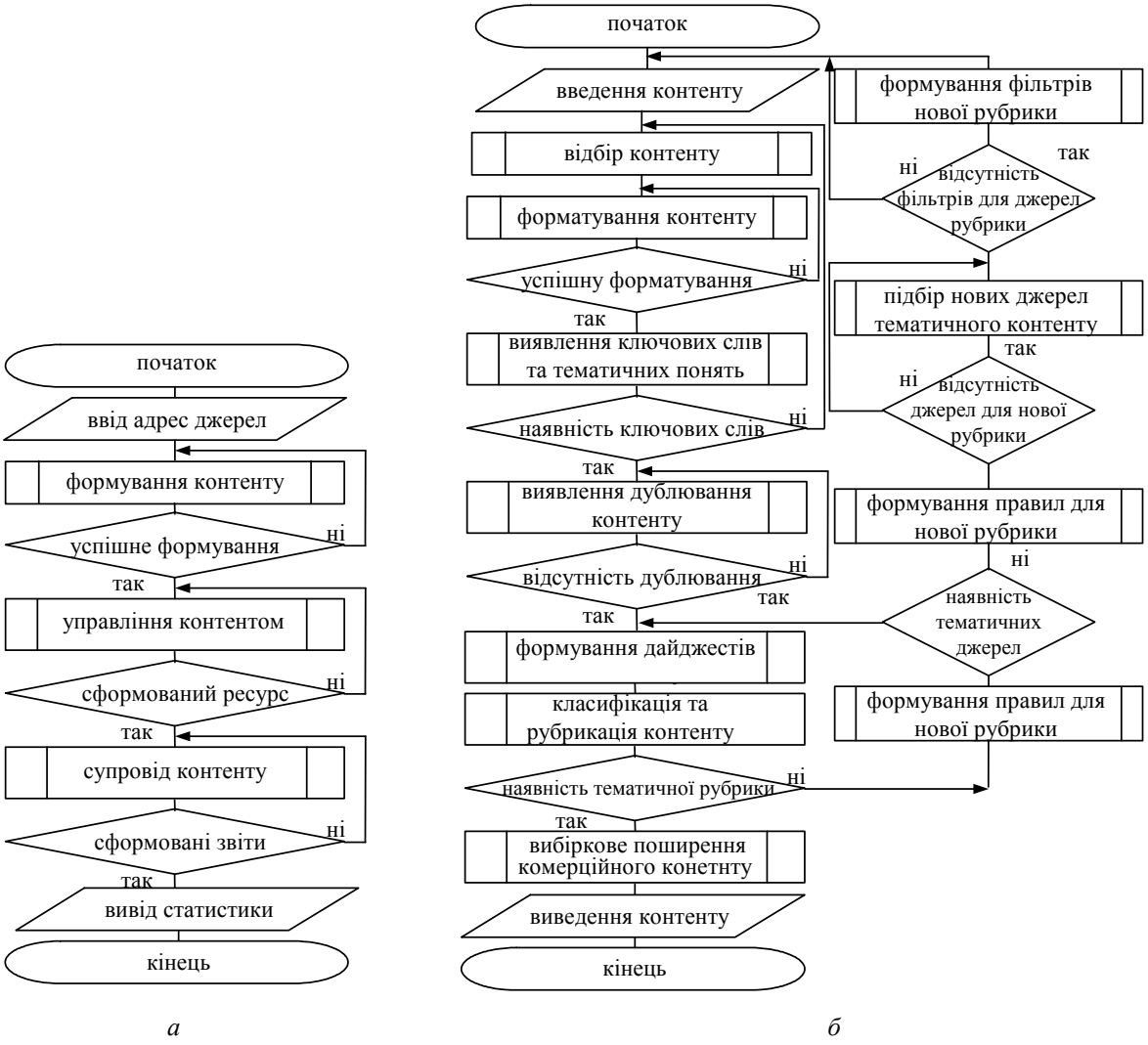


Рис. 1. Функціональна схема процесу опрацювання інформаційних ресурсів у системах електронної контент-комерції

Класичний математичний апарат та поширені інструментальні засоби не здатні адекватно відображати аналіз контентного масиву фіксованого розміру та навігацію в ньому. Структура СЕКК має рівні ієрархії, які забезпечують незалежність збережених даних від програм, що їх використовують, та можливість розвитку системи без руйнування існуючих застосувань.

На рис. 2, а подано схему процесу опрацювання інформаційних ресурсів в СЕКК, а на рис. 2, б – етапи формування комерційного контенту в системах електронної контент-комерції.



а *б*

Рис. 2. Процес опрацювання (а) та формування інформаційних ресурсів (б)

СЕКК працюють за такою схемою зв’язків:

Контент менеджер→підсистема редагування→ база даних →ядро→користувач.

Система електронної контент-комерції містить ядро адміністрування, підсистему авторизації/аутентифікації, менеджер шаблонів і менеджер контенту для розв’язування задач з позиції користувача (рис. 3). При створенні ядра системи електронної контент-комерції використовують об’єктно-орієнтовану модель та абстрактні об’єкти зі властивостями/методами (рис. 4). Взаємодію з основними об’єктами ядра системи реалізують через інкапсуляцію. Для цього в класах реалізують інтерфейсні методи, призначені для маніпуляцій всередині об’єкта з даними/властивостями. Процесами опрацювання інформаційних ресурсів у системах електронної контент-комерції є формування, управління та супровід контенту (рис. 5).

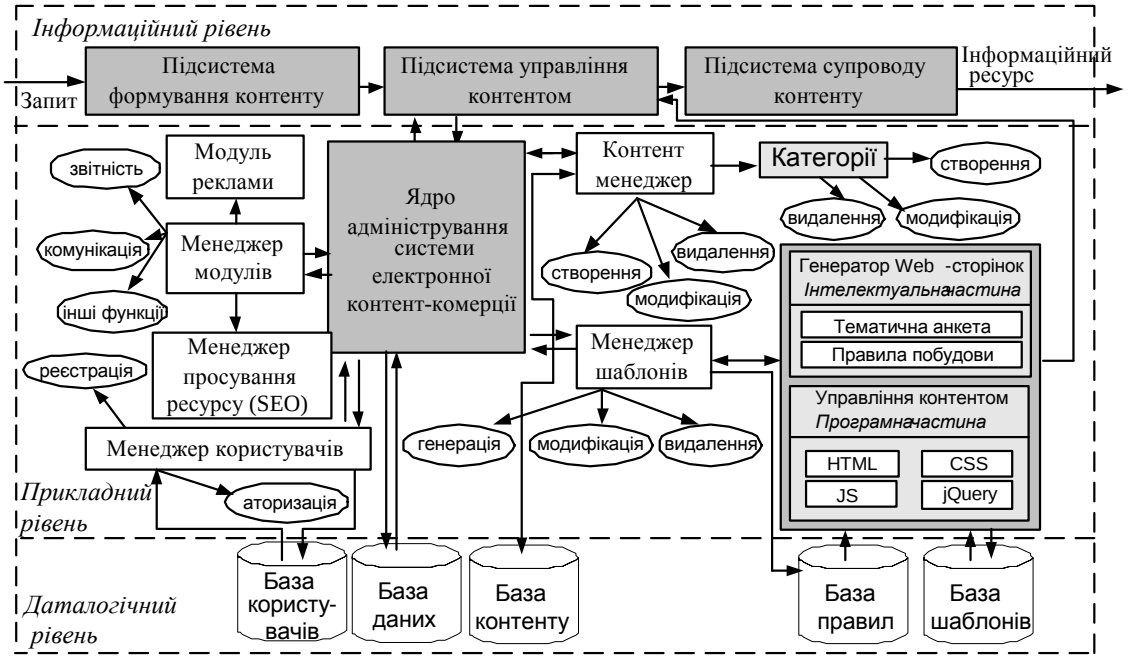


Рис. 3. Структура системи електронної контент-комерції

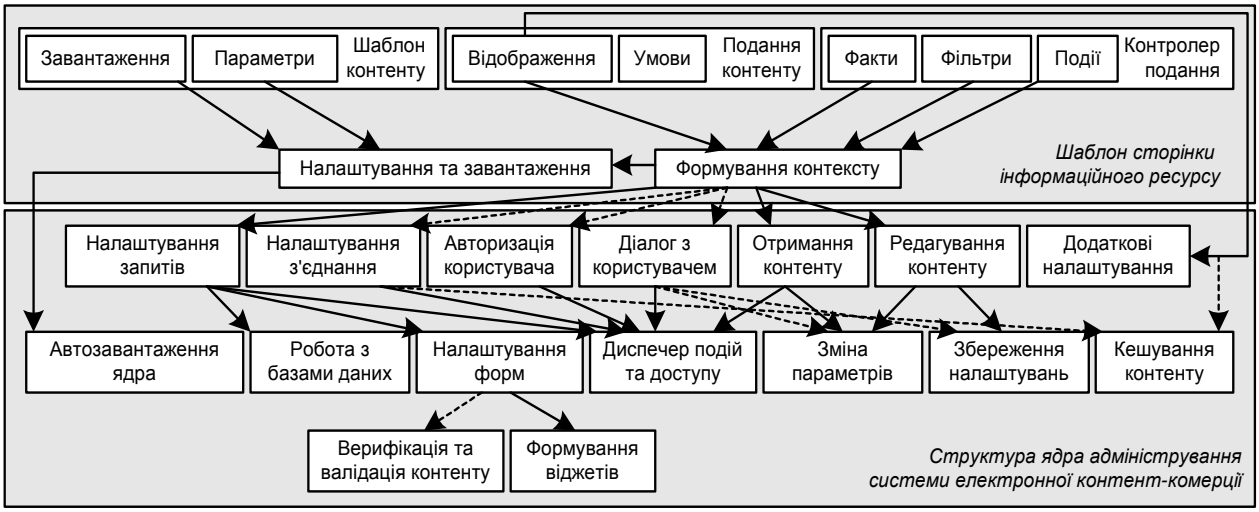


Рис. 4. Діаграма структури системи електронної контент-комерції

Отримані дані враховують при створенні або оновленні інформаційного ресурсу та удосконаленні архітектури СЕКК. Раніше модератор самостійно шукав та опрацьовував необхідний контент: збір контенту з різних джерел даних, аналіз та фільтрація контенту, формування комерційного контенту як кінцевого продукту згідно з індивідуалізованими даними користувача системи. Реалізація процесу опрацювання інформаційних ресурсів в СЕКК полегшує роботу модератора, автоматизуючи збирання контенту з джерел, аналіз та фільтрацію контенту.

І. **Процес формування контенту** складається із декількох етапів:

Модератор → створення контенту → база даних → систематизація контенту → база даних → поширення контенту → редактор, або *Інформаційний ресурс (джерело) → збирання контенту → база даних → систематизація контенту → база даних → поширення контенту → модератор*

і реалізується у вигляді контент-моніторингу контенту та створення бази даних відповідно до інформаційних потреб споживачів (рис. 6, а). Після збирання і первинного опрацювання контент приводять до єдиного формату, класифікують відповідно до визначеного рубрикатора та йому приписують дескриптори, враховуючи ключові слова. При застосуванні інтернет-маркетингу (рис. 6, б) етапи систематизації контенту забезпечують постійне поповнення бази даних

оперативними даними, ефективний одночасний доступ багатьох користувачів до бази даних, зручні засоби пошуку необхідного контенту.

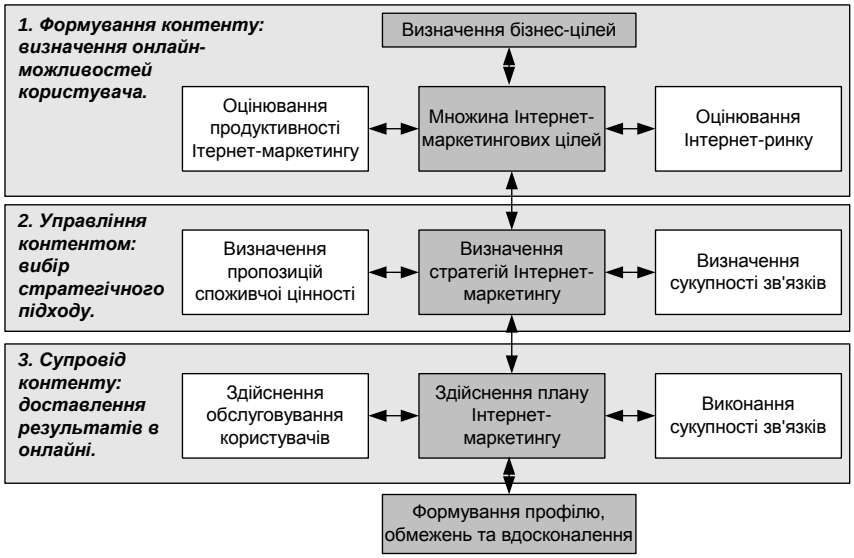


Рис. 5. Основні процеси опрацювання інформаційних ресурсів у системах електронної контент-комерції

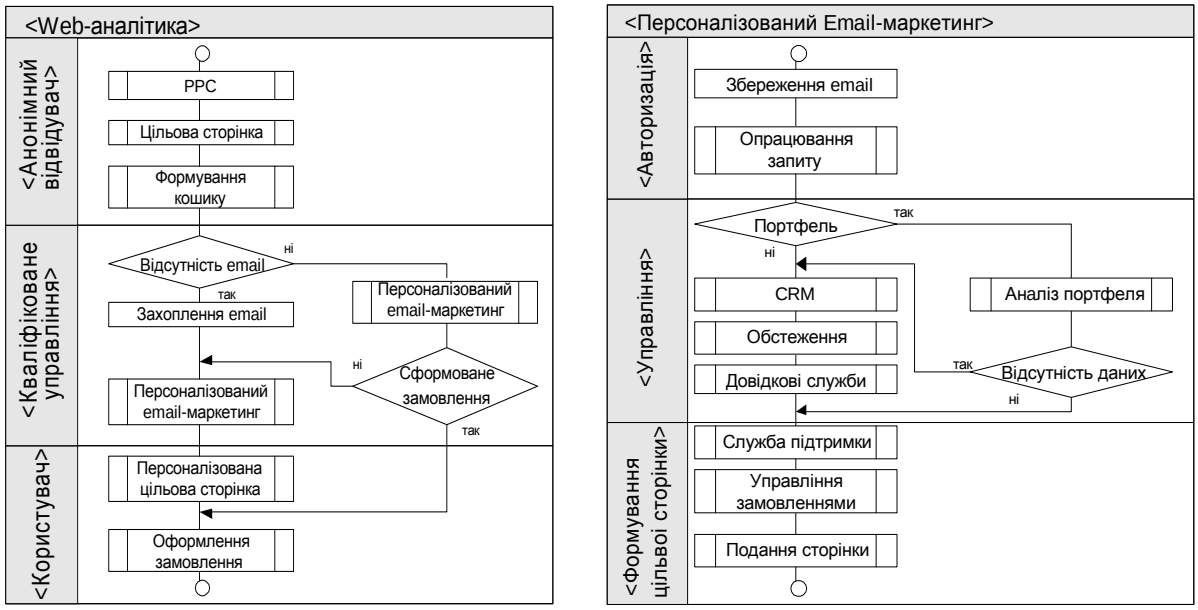
а *б*

Рис. 6. Методи збирання/створення (а) контенту та маркетингу (б)

Поділ на етапи формування комерційного контенту в СЕКК підвищує ефективність при адмініструванні системи; забезпечує економію ресурсів, інтернет-трафіку та анонімність користувачів; автоматичне сканування джерел даних.

II. **Процес управління контентом** складається із декількох етапів (рис. 7):

Користувач → опрацювання контенту → база даних → аналіз контенту → база даних → подання контенту → користувач.



а *б*

Рис. 7. Схема аналізу контенту відвідувачів (а) та інформаційного ресурсу (б)

Управління контентом інформаційного ресурсу та відвідувачів (рис. 8), їх моделювання є методом кількісного дослідження динаміки окремих тематичних напрямів та технічного аналізу інформаційного ресурсу.

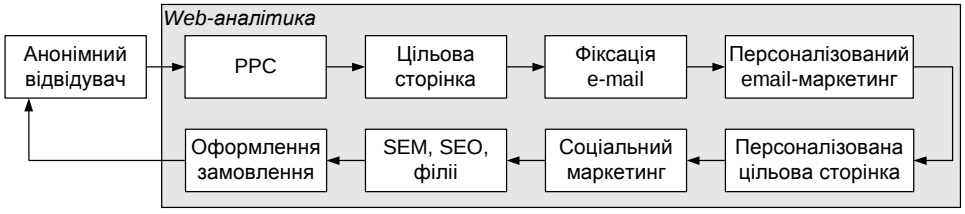


Рис. 8. Схема процесу аналізу відвідування інформаційного ресурсу

Аналіз результатів управління контентом впливає на швидкість розвитку тематичних напрямів та контентного простору. Стійкі статичні зв'язки між контентом свідчать про кореляцію тематик, ефективність посилань на публікації джерел, більш ранні цитування, републікації тощо. Механізми, основані на узагальнених методах кластерного аналізу, виявляють контент у потоках, що формує навколо себе нові тематичні напрями. Кластерний аналіз, теорія фракталів і автомодельних процесів за їхнього коректного застосування кількісно оцінюють ступінь зв'язку в тематичних контентних потоках. Оперативний аналіз контенту відвідувачів сприяє реалізації процесу управління контентом за допомогою генерації сторінок через інформаційні блоки, яка поділяється на типи: тематична; за останніми зверненнями; комбінована.

III. **Процес супроводу контенту** – це оперативні етапи узагальнення, модерації та структурування комерційного контенту (рис. 9), тобто

Користувач → структурування контенту → база даних → модерація контенту → база даних → узагальнення контенту → модератор.

Із підсистемою супроводу контенту СЕКК має такі можливості: формування рейтингу комерційного контенту; формування інформаційного портрету постійного користувача СЕКК; аналіз характеристик (коментарі, відгуки, побажання тощо) на комерційний контент з боку користувача СЕКК; збирання, накопичення та опрацювання інформації про потреби кінцевого/потенційного користувача СЕКК та споживача контенту; формування інформаційного портрету контентних потоків.

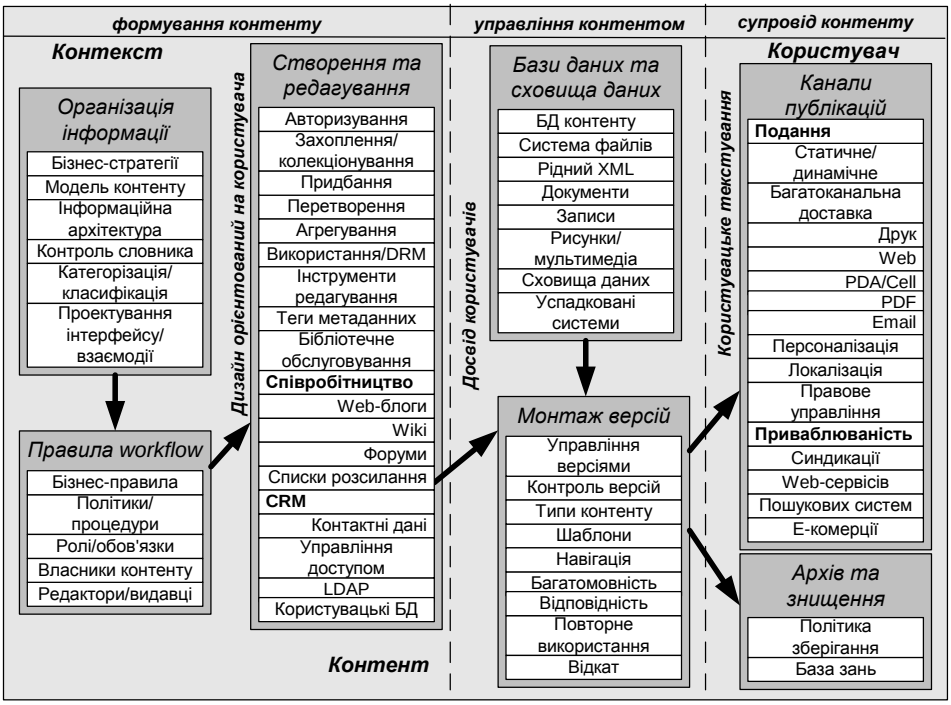


Рис. 9. Підсистеми формування, управління та супроводу контенту в системах електронної контент-комерції

Система електронної контент-комерції значно полегшує роботу модератора з формування, управління та супроводу комерційного контенту (рис. 10). Вона передбачає такі основні етапи опрацювання контенту інформаційних ресурсів.

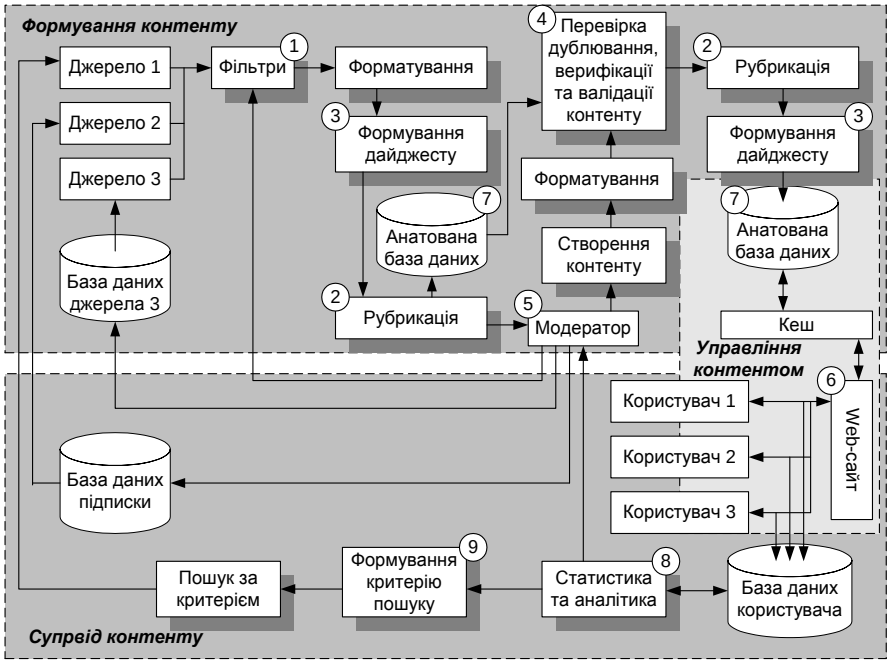


Рис. 10. Схема взаємодії підсистем опрацювання інформаційних ресурсів

1. Збирання контенту з різних джерел даних системою.
2. Фільтрація створеного і/або зібраного контенту системою.
3. Аналіз (визначення ключових слів, рубрикація, формування дайджестів) контенту системою.

4. Аналіз отриманих даних для концентрування уваги модератора на необхідній тематиці. Підготовка даних для оперативної роботи з ними.
5. Аналіз потреб потенційних та постійних користувачів системи.
6. Формування кінцевого продукту згідно із проаналізованими даними та потребами потенційних/постійних користувачів.
7. Поповнення списку джерел даних згідно із аналізом запитів користувачів.
8. Поповнення правил та словників фільтрів згідно аналізу контенту.
9. Виявлення нових тематичних рубрик потоку контенту.

На рис. 11 подано класифікацію розроблених методів опрацювання інформаційних ресурсів у СЕКК з детальним переліком реалізованих етапів формування, управління та супроводу комерційного контенту в цих системах.

Рис. 11. Методи опрацювання інформаційних ресурсів у системах електронної контент-комерції

Розроблені методи опрацювання інформаційних ресурсів у СЕКК дають можливість сформувати вимоги до підпрограм опрацювання інформаційних ресурсів. Функціонування СЕКК описується такими схемами зв'язків основних компонентів цієї системи:

1) для процесу формування інформаційного ресурсу системи схема така

контент → формування контенту → база даних контенту → управління контентом → інформаційний ресурс системи;

2) для процесу формування відповіді на запит користувача схема така

запит користувача → управління контентом → інформаційний ресурс системи → супровід контенту → база даних користувачів;

3) для процесу формування звіту роботи системи для модератора схема така

запит модератора → супровід контенту → база даних користувачів → управління контентом → звіт для модератора;

4) для процесу модерації внутрішніх параметрів системи схема така

запит модератора → формування контенту → база правил → супровід контенту → база правил → управління контентом → результат.

Основними підсистемами опрацювання інформаційних ресурсів в СЕКК є формування, управління та супровід контенту (рис. 12), схема зв'язків яких є такою: *формування контенту* → *управління контентом* → *супровід контенту.*

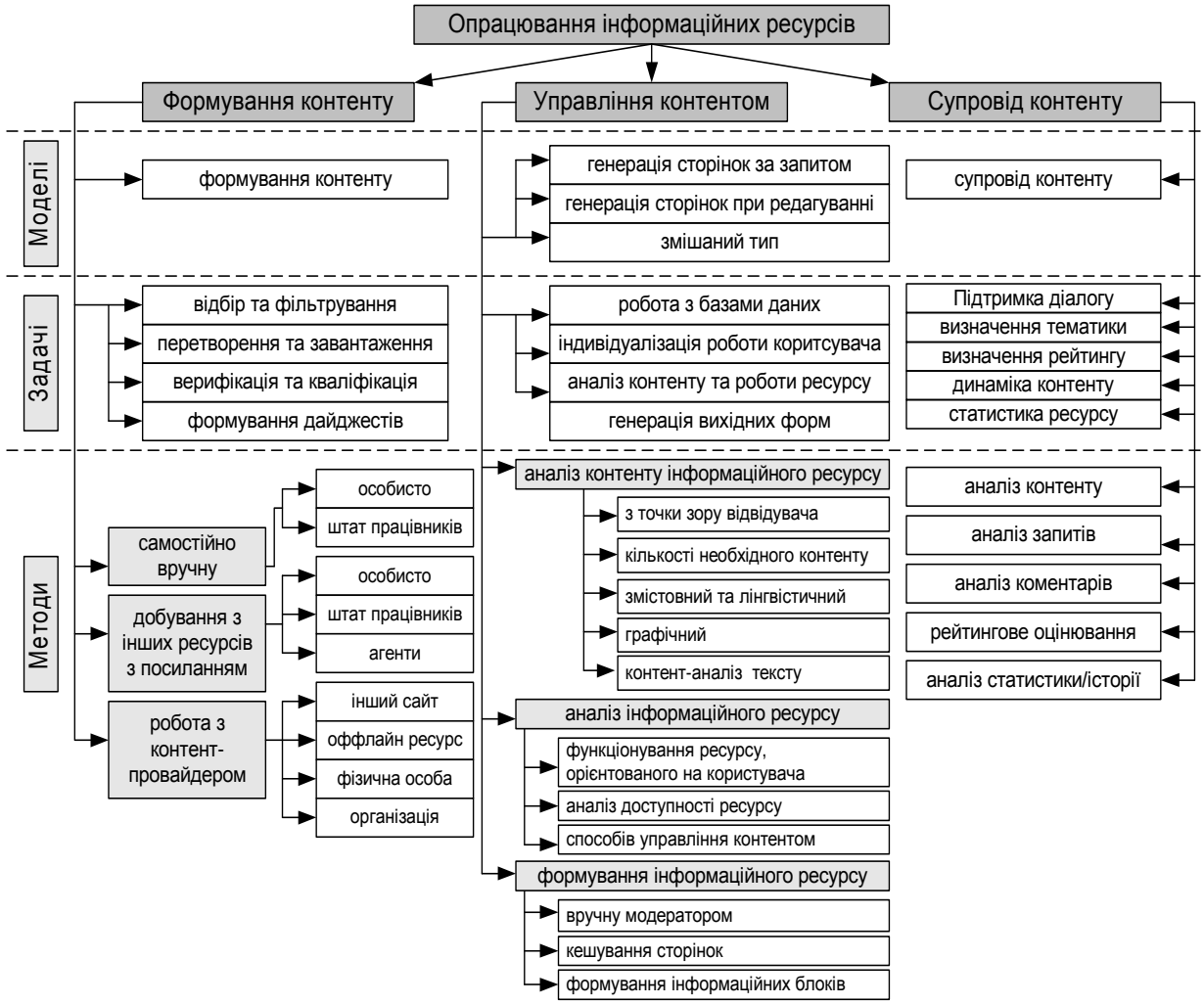


Рис. 12. Особливості процесу опрацювання інформаційних ресурсів СЕКК

Систему електронної контент-комерції подано як

$$Y = \langle X, Q, H, C, V, Z, T, \alpha, \beta, \gamma \rangle, \quad (1)$$

де величина $X = \{x_1, x_2 \ldots, x_{n_X}\}$ – множина вхідних даних $x_i \in X$ з різних джерел інформації при $i = \overline{1, n_X}$; величина $Q = \{q_1, q_2 \ldots, q_{n_Q}\}$ – множина запитів $q_d \in Q$ користувачів при $d = \overline{1, n_Q}$; величина $H = \{h_1, h_2 \ldots, h_{n_Y}\}$ – множина внутрішніх параметрів $h_k \in H$ СЕКК при $k = \overline{1, n_H}$; величина $C = \{c_1, c_2 \ldots, c_{n_C}\}$ – множина комерційного контенту $c_r \in C$ при $r = \overline{1, n_C}$; величина $V = \{v_1, v_2 \ldots, v_{n_V}\}$ – множина параметрів впливу $v_l \in V$ зовнішнього середовища на СЕКК при $l = \overline{1, n_V}$; величина $Z = \{z_1, z_2 \ldots, z_{n_Z}\}$ – множина сторінок $z_w \in Z$ інформаційного ресурсу в СЕКК при $w = \overline{1, n_Z}$; величина $T = \{t_1, t_2 \ldots, t_{n_T}\}$ – час $t_p \in T$ транзакції опрацювання інформаційного ресурсу в СЕКК при $p = \overline{1, n_T}$; величина $Y = \{y_1, y_2 \ldots y_{n_Y}\}$ – колекція статистичних даних $y_j \in Y$ роботи СЕКК при $j = \overline{1, n_Y}$; величина α – оператор формування контенту, β – оператор управління контентом, γ – оператор супроводу контенту. Тоді $\delta : X \to Y$ подано суперпозицією функцій

$$\delta = \gamma \circ \beta \circ \alpha, \quad (2)$$

Величина $y_j = \{y_{1j}, y_{2j}, ..., y_{gj}\}$ є колекцією даних за визначений період часу, де y_1 – кількість відвідувань; y_2 – середній час відвідування інформаційного ресурсу (хв:с); y_3 – показник відмовлень (%); y_4 – досягнута мета пошуку; y_5 – динаміка контенту (%); y_6 – загальна кількість переглянутих сторінок; y_7 – кількість переглянутих сторінок за одне відвідування; y_8 – нові відвідування (%); y_9 – абсолютно унікальні відвідувачі; y_{10} – джерело трафіка (прямі переходи, переходи з пошукових систем, переходи з інших сайтів тощо) у % тощо [107].

Оператор формування комерційного контенту α є відображенням комерційного контенту c_r в новий стан c_{r+1}, що відрізняється від попереднього стану появою нової частини контенту Δc, яка доповнює попередній стан $c_{r+1} = c_r + \Delta c$ при

$$\alpha : (c_r, t_p, X, u_f) \to (c_{r+1}, t_{p+1}), \tag{3}$$

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

Комерційний контент c_r подано як

$$c_r = \left\{ \bigcup_i^{n_X} x_i \middle| \begin{array}{l} \forall x_i \in X_{u_f}, x_i \notin X_{\overline{u_f}}, \exists u_f \in U_{x_i}, u_f \notin U_{\overline{x_i}}, \\ X = X_{u_f} \cup X_{\overline{u_f}}, U = U_{x_i} \cup U_{\overline{x_i}}, f = \overline{1, n_U} \end{array} \right\}, \tag{4}$$

де множину умов u_f формування контенту c_r визначають як

$$u_f = \left\{ \bigcup_j^{k} u_{jf} \middle| \begin{array}{l} \forall u_{jf} \in U_{x_i}, \exists x_i \in X_{u_f}, u_{jf} \notin U_{\overline{x_i}}, \\ U = U_{x_i} \cup U_{\overline{x_i}}, X_{u_f} \subseteq X, f = \overline{1, n_U}, i = \overline{1, m} \end{array} \right\}. \tag{5}$$

Оператор управління контентом β є відображенням контенту c_r в новий стан c_r', який відрізняється від попереднього стану значеннями визначальних параметрів $h_k \to h_k'$ (актуальність, старіння, повнота, точність, релевантність, автентичність, достовірність), що задовольняють наперед визначені вимоги

$$\beta : (q_d, z_w, c_r, h_k, u_M, t_p) \to (c_r', h_k', z_{w+1}, t_{p+1}), \tag{6}$$

де $q_d \in Q$, $h_k \in H$, $h_k = \{h_{1k}(c_r, q_d), \ldots, h_{n_H k}(c_r, q_d)\}$ – множина умов управління контентом c_r.

Управління комерційним контентом подано як

$$z_w = \left\{ \bigcup_{r=1}^{n_C} c_r \middle| \begin{array}{l} \forall c_r \in C_{q_d}, \exists q_d \in Q, \exists h_k \in H_{c_r}, c_r \notin C_{\overline{q_d}}, h_k \notin H_{\overline{c_r}}, \\ C = C_{q_d} \cup C_{\overline{q_d}}, H = H_{c_r} \cup H_{\overline{c_r}}, d = \overline{1, n_Q}, k = \overline{1, n_H} \end{array} \right\}, \tag{7}$$

де множину значень визначальних параметрів формують як $h_k' = h_k + \Delta h$.

Оператор супроводу контенту γ є відображенням контенту c_r в колекцію значень y_i, які утворюються у результаті аналізу, моніторингу, оцінювання взаємодії з користувачем, пошуковими системами та іншими інформаційними ресурсами, що є основою для прийняття рішень щодо формування та управління контентом:

$$\gamma : (c_r, q_d, v_l, h_k, z_w, u_S, t_p) \to y_i, \tag{8}$$

де $v_l = \{v_{1l}(q_i, h_k, c_r, z_w, t_p), \ldots, v_{n_V l}(q_i, h_k, c_r, z_w, t_p)\}$ – множина умов супроводу контенту та впливів середовища на систему. Вихідні дані реалізовано

$$y_j = \left\{ \bigcup_l^{n_V} v_l \middle| \begin{array}{l} \forall v_l \in V_{q_d} \cup V_{z_w}, \exists q_d \in Q, \exists z_w \in Z, \exists h_k \in H_{c_r}, v_l \notin V_{\overline{q_d}}, v_l \notin V_{\overline{z_w}}, \\ V_{q_d} \subset V, V_{z_w} \subset V, d = \overline{1, n_Q}, w = \overline{1, n_Z}, r = \overline{1, n_C}, k = \overline{1, n_H} \end{array} \right\}. \tag{9}$$

Колекція $y_j = \{y_{1j}, y_{2j}, ..., y_{gj}\}$ описує процес функціонування СЕКК з такими основними процесами опрацювання інформаційних ресурсів, як формування, управління та супровід контенту. Аналізують статистику роботи СЕКК згідно з результатами аналізу реакцій на цю систему

постійного/потенційного користувача (відвідування, запити, пошук за ключовими словами тощо). Це сприяє ефективному аналізу реакції цільової та потенційної аудиторії на функціонування СЕКК.

Висновки і перспективи подальших наукових розвідок

Обгрунтовано доцільність впровадження СЕКК з огляду на такі переваги: збільшення оперативності одержання контенту; скорочення циклу виробництва і продажу контенту; зниження витрат, пов'язаних з обміном контентом; відкритість СЕКК стосовно клієнтів; автоматичне інформування користувачів про контент; створення альтернативних каналів продажу через інтернет-газети. У роботі вибрано підходи послідовності опрацювання інформаційних ресурсів в СЕКК для проектування та створення таких систем, що дало змогу виділити основні закономірності переходу від процесів формування контенту до його супроводу, а також визначити основні компоненти СЕКК та їх взаємозв'язки. Вдосконалено структуру СЕКК для спрощення етапів розроблення таких систем, що дало змогу інтерпретувати процеси формування, управління та супроводу контенту. Побудовано формальний опис процесу опрацювання інформаційних ресурсів в СЕКК для реалізації етапів життєвого циклу контенту розробленням методів формування, управління та супроводу контенту, що дало змогу виділити зв'язки між компонентами СЕКК. Розроблено формальний опис процесу формування контенту для виділення закономірностей та параметрів управління контентом, що дало змогу сформувати множину критеріїв, які впливають на значення актуальності, релевантності, старіння, повноти, зростання попиту та прибутковості контенту. Розроблено формальний опис процесу управління контентом для формування множини параметрів супроводу контенту, що дало змогу розробити вимоги та рекомендації для проектування СЕКК.

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