УДК 004.738.5

В. Висоцька, О. Окрушко
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ІНТЕЛЕКТУАЛЬНА СИСТЕМА РОЗПОДІЛУ ДАЙДЖЕСТІВ МІЖ ПРАЦІВНИКАМИ ЕЛЕКТРОННИХ ЗАСОБІВ МАСОВОЇ ІНФОРМАЦІЇ

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

Ключові слова: контент, інформаційний ресурс, Інтернет-журнал, 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.

Key words: content, information resource, Internet Magazine, systems of electronic content-commerce.

Вступ. Постановка проблеми

Одною з найважливіших проблем в роботі інтернет-видань є надмірне завантаження персоналу опрацюванням великої кількості інформації щодня для постійного пошуку та аналізу актуальних тем для написання статей. Це призводить до необхідності використання частини ресурсів персоналу для розподілу дайджестів, тем, завдань між виконавцями; займає багато часу і ресурсів; не дозволяє розподілити ефективно контент (наприклад, дайджести) через суб’єктивізм та неуважність модераторів під час визначення якості роботи журналістів та редакторів. Тому постає питання про можливість автоматизації розподілу дайджестів між робітниками електронних засобів масової інформації із введенням статистики якості роботи виконавців [1, 2].

Аналіз останніх досліджень та публікацій

Контент (англ. content – зміст) – це інформаційне змістовне наповнення (наприклад, тексти, графіка, мультимедіа) інформаційного ресурсу; множина всіх значень і величин, якими оперує ІС; узагальнене поняття даних без наперед визначеної структури [2]. Комерційний контент – це об’єкт бізнес-процесів систем електронної контент-комерції (табл. 1), наприклад, інформаційний продукт або вміст Web-сайта таких електронних засобів масової інформації, як інтернет-газета, інтернет-журнал, інтернет-видання, інтернет-видавництво тощо [2].

Таблиця 1

Основні характерні риси систем електронної контент-комерції

Назва	Характеристика
Віртуальність	Відсутність особистого контакту між суб’єктами процесу купівлі/продажу.
Інтерактивність	Адекватне інформаційне забезпечення запиту користувача у інтерактивному режимі.
Глобальність	Відсутність часових, просторових, асортиментно-товарних, адміністративних меж.
Динамічність	Спроможність on-line торгівлі до моментальних змін й адаптації з появою нових умов.
Ефективність	Забезпечення попиту, прибутку, економічних вигод, соціального ефекту.

Інтернет-журнали, наприклад, Top Gear (topgear.com), Drive (Drive.ru), AutoDiary (auto-diary.ru) та Автоцентр (autocentre.ua) займають одні з перших місць рейтингів популярності та володіють перевагами використання [1, 2], зокрема: користувач має справу з розширеним контентним тематичним потоком (архіви, інші інформаційні ресурси) через множину гіперпосилань; через форуми/конференції відсутні межі між автором і читачем, а користувач має можливість брати участь у виробництві інформаційного продукту; відсутність цензури та придушення авторської думки; Інтернет-публікації є оперативними та не обмеженими терміном виходу номера; невелика собівартість та децентралізованість сприяє розвитку спеціалізованих видань та забезпечує свободу слова і самовираження. Незважаючи на велику кількість переваг, інтернет-журнали мають недоліки [2]: збереження анонімності автора при публікації матеріалу призводить до зловживань як автором (дезінформація, прихована реклама), так і псевдоавторами (плагіат); невисокий рівень грамотності в інтернеті через відсутність контролю з боку редактора.

При цьому важливим є забезпечення інваріантності 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 зберігання, технічних засобів; вимог користувачів контенту, поява нових вимог та категорій користувачів; порядку розподілу контенту та способів доступу користувачів.

Отже, виникає проблема створення єдиного концептуального опису всього інформаційного ресурсу, який має на меті стабільне підтримання зовнішніх/внутрішніх позначень контенту відповідно до їх завдань, вимог та змін. Тому необхідно класифікувати інформаційні ресурси систем електронної контент-комерції для подальшого дослідження їх природних, технологічних та споживчих якостей з метою виявлення характерних та специфічних властивостей, а також закономірностей та особливостей їх формування та застосування. За основу класифікації взято основні властивості контенту в системах електронної контент-комерції як синтаксис (принципи формального подання контенту), структура (правила побудови та впорядкування контенту) та семантика (формування змісту/розуміння/функцій для визначення основних завдань та порядку застосування контенту). На їх основі обрано основні фактори класифікації [1, 2]: способи подання контенту у системах електронної контент-комерції; способи структурування інформаційного ресурсу; способи доступу до інформаційного ресурсу систем електронної контент-комерції; призначення інформаційного ресурсу систем електронної контент-комерції.

Формують інформаційний ресурс систем електронної контент-комерції різними шляхами: гомогенізацією, розподілом або інтеграцією (рис. 1) [2].

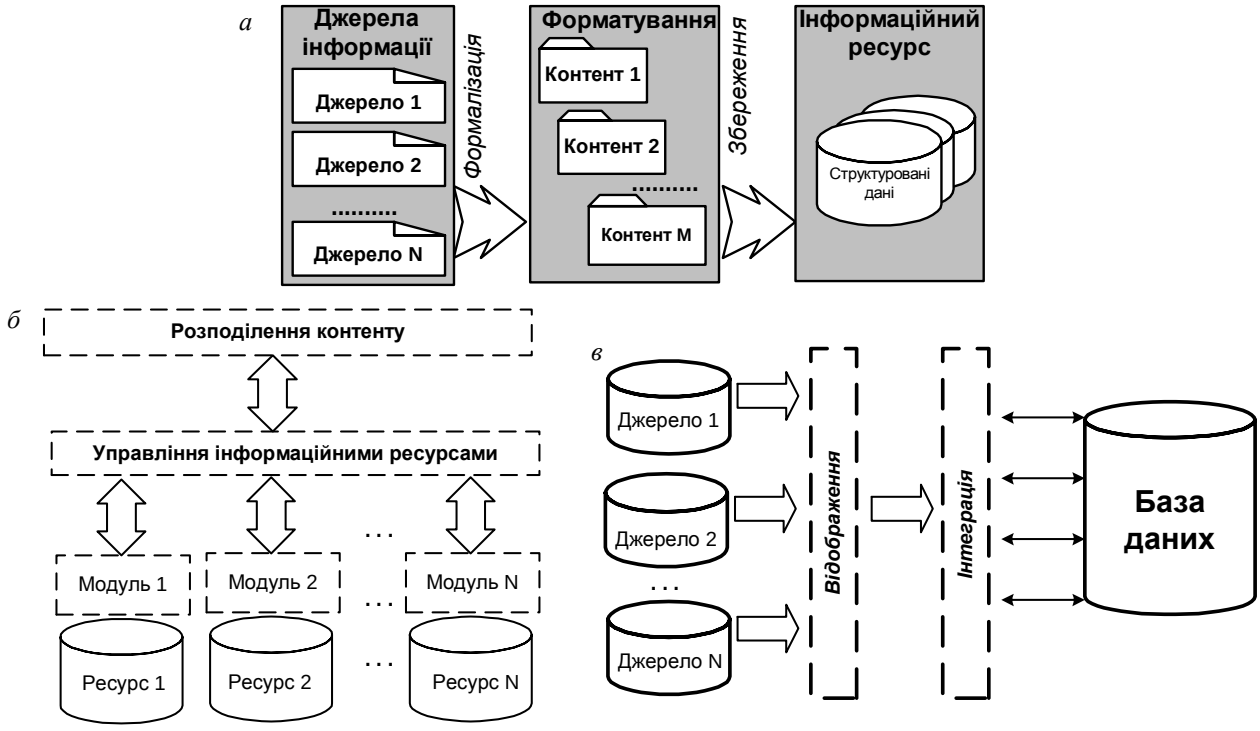


Рис. 1. Побудови інформаційного ресурсу шляхом:
а – гомогенізації; б – розподілу; в – інтеграції контенту

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 електронної контент-комерції [2]:

1) підтримка інтеграції даних (забезпечення роботи нових застосувань із наявними репозиторіями даних і накопиченим в них контентом);

2) незалежність управління контентом різних репозиторіїв за допомогою Web-сервісів;

3) підтримка Web 2.0 (забезпечення загальних Web-сервісів та інтерфейсів для спрощення розроблення застосувань);

4) незалежність платформи (створення платформи, незалежної від мови контенту);

5) формування контенту (підтримка розроблення композитних застосувань і колажів, тобто нового контенту, складеного із контенту декількох джерел, який виглядає як єдине ціле).

Web 2.0 позначають ряд технологій та послуг інтернету та описують зміну сприйняття інтернет-користувачами [1, 2, 4]. Зміни полягають у посиленні комунікативності, співробітництва, безпечного використання контенту та загальному розвитку функціональності мережі [9, 10]. Термін не означає поновлення технічних специфікацій, а лише зміни у шляхах використання мережі розробниками програмного забезпечення та кінцевими користувачами (табл. 2) [9, 10].

Таблиця 2

Порівняння параметрів Web 1.0 та Web 2.0

Назва	Web 1.0	Web 2.0	Нові властивості
Актори	Розробник-користувач/автор контенту і читач	Користувач як співрозробник, читач як співавтор або товариство	Право на участь; скасування модерації
ПЗ	ПЗ створювалось для ПК, це товар; закриті вихідні коди; ліцензійний продаж; прив'язка ПЗ до обладнання; націленість на винахід; запланований реліз; для перегляду контенту є браузер.	ПЗ для Web; це сервіс/застосування; відкриті вихідні коди, open-source software; безкоштовне ПЗ; ПЗ над обладнанням; вічна бета; пошук застосування винайденому; альтернативні засоби сприйняття.	Web-платформа; зняття/розмиття бар'єрів/обмежень: доступність, універсальність, спрощення.
Контент	Поповнення БД: плата автору або наймання модераторів; таксономічна організація даних (ієрархія рубрик); засоби збереження даних – каталог, бібліотека, сховище; односторонні посилання; форма подання – персональні сторінки; статичний сайт; адресу у сторінки сайта; джерело – розум автора; меню навігації сайта для роботи з даними сайта; копірайт; для сприйняття контенту мандрують сайтом по посиланню чи закладці;	Поповнення баз даних – те, що має один, відразу стає доступне кожному; дані організовують фолксономічно; засоби використання даних – API-інтерфейси; автоматичні двосторонні посилання; форма подання – блоги; динамічний сайт; адресу має мікроелемент контенту; джерело – колективний розум; інтерфейс для роботи з даними по всій мережі; "вільна" ліцензія GNU FDL; для сприйняття контенту не потрібно відвідувати сайт – можливо читати RSS-стрічки.	Мережа як єдиний колективний розум, атомізація контенту, агрегація, синдикація.
Події	Замовлення та виготовлення ПЗ; публікація контенту авторами і сприйняття його читачами; звертання до третьої особи – посередника для задіяння його ресурсів; великі, нечисленні угоди;	Співпраця через відділ технічної підтримки ПЗ; взаємодія, додавання властивостей, цінності, створення спільного контенту кожним учасником; самообслуговування, яке засноване на партнерській архітектурі сервісу – сервіс лише посередник між користувачами, які використовують їхні ресурси; дрібні численні транзакції.	Співпраця; самодіяльність; масові одиничні взаємовідносини.
Цінність	Цінність в ПЗ; Інтернет цінний як джерело інформації;	Цінність в базах даних та сервісах роботи з ними; Інтернет – інструмент комунікацій.	Робота з базою даних; економія часу/уваги.

Життєвий цикл контенту (англ. Content lifecycle) – це складний процес, який проходить контент під час управління через різні етапи публікації [2]. Існуючі моделі життєвого циклу контенту не містять всіх етапів процесів опрацювання інформаційних ресурсів: формування,

управління та реалізація контенту (табл. 3) [4]. Кількість контентних потоків більша, ніж шляхів переміщення товарів на промислових підприємствах. Значна частина контентних потоків складається з легко формалізованих і автоматизованих процедур. Основна проблема – відсутність загального підходу до процесу моделювання, проектування та розроблення систем електронної контент-комерції (СЕКК) [2]. Відсутність загальної та детальної класифікації систем електронної контент-комерції, що приводить до проблеми визначення і формування загальних методів проектування та розроблення архітектури та алгоритмів функціонування цих систем. Це обґрунтовує мету, актуальність, доцільність та напрями дослідження. Наявні системи електронної комерції (СЕК) не підтримують всього життєвого циклу контентного потоку та не вирішують основних проблем опрацювання інформаційних ресурсів – формування та реалізації контенту (табл. 4).

Таблиця 3

Порівняння моделей життєвих циклів контенту

Автор моделі Content lifecycle	Формування	Управління	Реалізація
McKeever Susan	+/–	–	+/–
Bob Boiko	+/–	+/–	+/–
Gerry McGovern	+/–	–	+/–
JoAnn Hackos	+/–	–	+/–
Ann Rockley	+/–	+/–	+/–
Russell Nakano	+/–	–	+/–
The State government of Victoria	+/–	–	+/–
AIIM	+/–	+/–	+/–
CMP organization	+/–	+/–	–
Bob Doyle	+/–	+/–	+/–
Woods Randy	+/–	+	+
Halverson	+	+/–	+/–

Таблиця 4

Порівняння особливостей систем

Назва характеристики	СЕК	СЕКК
Нематеріальність товару	–	+
Постійна кількість товару	–	+
Ріст кількості різновиду товару	+/–	+
Відсутність складу	–	+
Збереження товару в базах даних	–	+
Ефективність просування товару за ключовими словами	+/–	+
Ефективність пошуку товару за ключовими словами	+/–	+
Автоматичне виявлення та ліквідація дублювання товару	–	+
Автоматичне визначення старіння товару за змістом	–	+
Автоматичне визначення актуальності товару	+/–	+
Автоматичний аналіз аудиторії	+/–	+
Автоматичне формування дайджестів	–	+
Автоматичний розподіл товару між учасниками	+/–	+
Автоматичний розподіл дайджестів між працівниками	–	+
Автоматичне формування товару	–	+
Автоматичне форматування товару	–	+
Вплив досвіду користувача на збільшення обсягу продажів	+/–	+

Автоматичний розподіл контенту передбачає декілька етапів: формування списку об’єктів розподілу (наприклад, статей, ПЗ, книг або дайджестів); визначення критеріїв (ознак) розподілу контенту з отриманого списку (процент унікальності контенту; кількість звернень до контенту; користувацька оцінка; час перегляду); оцінювання певних параметрів з метою використання в процесі розподілу. Наведені критерії не можуть вважатися однаковими за значенням та важливістю при аналізі їх загалом та обчисленні зведеної оцінки якості роботи. Крім цього, оцінювати кожну публікацію необхідно не поодинці, а в комплексі з іншими публікаціями, які підлягають оціню-

ванню в цей момент. Одним з найважливіших критеріїв розподілу дайджестів між працівниками електронних засобів масової інформації є процент унікальності контенту в попередніх публікаціях кожного окремого журналіста [3]. Зауважимо, що проста перевірка на дублювання контенту загалом чи окремих його блоків з уже наявними в мережі зразками не завжди дозволяє виявити плагіат, оскільки існує метод рерайтингу (від англ. rewrite – «переписувати»). Хоча дублікат може виглядати стилістично гіршим за оригінал, проте процент унікальності дуже високий і звичайні сервіси з перевірки унікальності не знайдуть у цьому випадку плагіату. Проблема рерайтера полягає тільки в тому, що більшість контенту містить деякі терміни, до яких важко/неможливо підібрати синоніми. Відомими системами визначення унікальності контенту є Praide Unique Content Analyser 2, FIndCopy та Miratools. Існують такі види запозичення [8]: повне або часткове копіювання тексту з одного джерела; копіювання і компонування тексту з декількох джерел; копіювання тексту з іншого джерела і зміна послідовності слідування частин тексту.

Для приховування факту запозичень застосовують такі підходи.

1. Коригування родів, чисел та часів слів текстової інформації.
2. Незначна зміна запозиченого тексту.
3. Скорочення запозиченого тексту (видалення речень, абзаців, рисунків, формул тощо).
4. Заміна кириличних символів на аналогічні за накресленням латинські та навпаки.
5. Здійснення ручної або автоматичної синонімізації тексту.

Це враховують, перевіряючи унікальність текстів та визначаючи якість роботи. Отримавши процент унікальності тексту в публікації працівника електронного видання, оцінюють якість цієї публікації та заносять отриману оцінку в таблицю рейтингів. Із збільшенням кількості публікації можна точніше оцінити якість та продуктивність кожного працівника електронного видання. Із збільшенням кількості критеріїв оцінювання можна охопити ширший спектр аспектів роботи працівника видання. Існує велика кількість методів вирішення проблеми комплексного оцінювання будь-чого. Одним з найвідоміших є метод аналізу ієрархій (MAI) – математичний інструмент системного підходу до складних проблем прийняття рішень [6]. MAI, який розробив американський математик Томас Сааті, дає змогу раціонально структурувати складну проблему прийняття рішень у вигляді ієрархії, порівняти їх і кількісно оцінити альтернативні варіанти рішення [5, 6].

Формулювання цілі статті

Вихідною інформацією процесу функціонування СЕКК є дані про призначення й умови роботи системи, які визначають основну мету моделювання СЕКК і дають змогу сформулювати вимоги до формальної моделі системи S та моделей управління контентом [2]. Формальна модель СЕКК $S = \langle X, C, V, H, Function, T, Y \rangle$ – це множини величин, що описують процес функціонування системи і утворюють такі підмножини: вхідні впливи на систему $x_i \in X$ ($i = \overline{1, n_X}$, $X = \{x_1, x_2, \ldots, x_{n_X}\}$), впливи потоку контенту на систему $c_r \in C$ ($r = \overline{1, n_C}$, $C = \{c_1, c_2, \ldots, c_{n_C}\}$), впливи зовнішнього середовища $v_l \in V$ ($l = \overline{1, n_V}$, $V = \{v_1, v_2, \ldots, v_{n_V}\}$), внутрішні параметри системи $h_k \in H$ ($k = \overline{1, n_H}$, $H = \{h_1, h_2, \ldots, h_{n_H}\}$), вихідні характеристики системи $y_j \in Y$ ($j = \overline{1, n_Y}$, $Y = \{y_1, y_2, \ldots, y_{n_Y}\}$) та час транзакції опрацювання інформаційного ресурсу $t_i \in T$ ($i = \overline{1, n_T}$, $T = \{t_1, t_2, \ldots, t_{n_T}\}$) [7].

Величини x_i, c_r, v_l, h_k, y_j є елементами непересічних підмножин і містять детерміновані і стохастичні складові [2, 4, 7]. Процес функціонування S описується функцією $y_j(t_i + \Delta t) = Function(x_i, c_r, v_l, h_k, t_i)$, де x_i – це запити відвідувачів/користувачів до СЕКК. Згідно з Google Analytics y_j – це кількість відвідувань за період часу Δt, середній час перебуття на сайті (хв:с), показник відмовлень (%), досягнута мета; динаміка (%), кількість всього перегляду сторінок, кількість перегляду сторінок за одне відвідування; нові відвідування (%); абсолютно унікальні відвідувачі; джерело трафіка у % (пошукові системи, прямий трафік або інші сайти). Впливи

величин c_r, v_l, h_k, на y_j як результат роботи СЕКК є невідомими та недослідженими [2]. Вивчення динаміки потоку комерційного контенту та побудова моделей опрацювання інформаційних ресурсів СЕКК є важливими та актуальними. Формальна модель СЕКК не розкриває механізмів управління контентом. Формальні моделі управління контентом призначені лише для визначення процесів старіння (актуальності) контентного потоку, а деякі із них (логістична, аналітична) – і для тематичного потоку. Вони не вирішують проблем формування, розподілу та реалізації контенту і вирішують не всі проблеми управління контентом, наприклад, подання множини контенту кінцевому користувачу згідно із його запитом, історією або інформаційним портфелем, автоматичне виявлення тематичних сюжетів, автоматичне формування дайджестів, інформаційних портретів, побудова таблиць взаємозв'язку понять, розрахунок рейтингів понять, збирання інформації з різних джерел та її форматування, виявлення ключових слів/понять та дублювання змісту, автоматична рубрикація, вибіркове поширення контенту. Недолік моделей управління контентом – це відсутність зв'язків між вхідними даними, контентом та вихідними даними в СЕКК [2]. Під час розгляду динаміки тематичних потоків контенту виявлено обмеженість моделей (табл. 5), що відкриває шлях для подальших досліджень [4].

Таблиця 5

Складові систем електронної контент-комерції

Модель	Перевага
Бартона–Кеблера	Описує процес старіння контенту, втрати його актуальності, швидкості розвитку окремих тематик або всього контентного простору; має точний розв'язок у вигляді експоненти.
Просторово-векторна	Визначення значущого терму в потоці контенту та найактуальнішого контенту із множини наявних. Обов'язкове ранжування контенту, використання параметричних множників, що залежать від часу.
Лінійна	Визначення інтенсивності потоку в часі при лінійній динаміці управління тематичного контенту.
Експонентна	Описує процес старіння контенту, втрати його актуальності. Кореляція між окремим контентом несуттєва.
Логістична	Вивчення динаміки окремого тематичного потоку. Розмірність параметрів та їхній вимір не враховують.
Аналітична	Описує процес старіння контенту та втрати його актуальності з використанням словника ключових слів.

Із врахуванням вже описаних задач виникає потреба в розробленні інтелектуальної системи розподілу дайджестів між працівниками електронних засобів масової інформації, яка буде здатна оцінювати якість роботи та розподілу завдань. При цьому система повинна отримувати готовий список дайджестів і надавати до нього доступ усім працівникам окремо за чергою, яка формується відповідно до таблиці рейтингів працівників видання. Таблицю рейтингів формують за методом аналізу ієрархій, оцінюючи готові публікації за декількома критеріями одночасно.

Виклад основного матеріалу

Система електронної контент-комерції має таку структурну схему:

$$\textit{Вхідні дані} \rightarrow \textit{Модуль формування контенту} \rightarrow \textit{Модуль управління контентом} \rightarrow \textit{Модуль реалізації контенту} \rightarrow \textit{Інформаційний ресурс},$$

а формальна модель – це шістка $S = \langle X, Formation, C, Management, Realization, Y \rangle$, де $X = \{x_1, x_2, \ldots, x_{n_X}\}$ – множина вхідної інформації, *Formation* – функція формування контенту, $C = \{c_1, c_2, \ldots, c_{n_C}\}$ – множина контенту, *Management* – функція управління контентом, *Realization* – функція реалізації контенту та $Y = \{y_1, y_2, \ldots, y_{n_Y}\}$ – множина вихідної інформації.

Формальна модель формування інформаційних ресурсів – це

$$Formation = \langle X, C, Gathering, Formatting, KeyWords, Categorization, Backup, Dissemination \rangle,$$

де X – множина вхідної інформації з Web-сайтів або від модераторів; C – множина контенту; *Gathering* – функція збирання інформації з джерел; *Formatting* – функція форматування

інформації, перетворення на множину контенту; *KeyWords* – функція виявлення ключових слів, понять; *Categorization* – функція автоматичної рубрикації; *Backup* – функція виявлення дублювання змісту контенту; *Dissemination* – функція вибіркового поширення контенту.

Моделі управління інформаційними ресурсами.

1. **Генерація сторінок за запитом**, структурна схема якої така:

$$Вхідні\ дані \rightarrow Модуль\ редагування \rightarrow База\ даних \rightarrow Модуль\ подання \rightarrow Інформаційний\ ресурс,$$

а формальна модель – $Management_Q = \langle X, C, Q, R, Edit, Y \rangle$, де X – множина вхідної інформації; C – множина контенту; Y – множина сформованих сторінок; Q – множина запитів; R – функція формування та подання сторінки; $Edit$ – функція редагування та модифікації контенту.

2. **Генерація сторінок під час редагування**, структурна схема якої така:

$$Вхідні\ дані \rightarrow Модуль\ редагування \rightarrow База\ даних \rightarrow Інформаційний\ ресурс,$$

а формальна модель – $Management_E = \langle C, Edit, Y \rangle$, де C – множина контенту; Y – множина статичних сторінок; $Edit$ – функція редагування та модифікації контенту. Процес формування сторінок описується функцією $\vec{y}(t) = Edit(\vec{c}, Weight, t)$. У разі внесення змін до змісту сайту створюють набір статичних сторінок, t не враховується, тому є інтерактивність між відвідувачем і вмістом сайту.

3. **Змішаний тип**, структурна схема якої така:

$$Вхідні\ дані \rightarrow Модуль\ редагування \rightarrow База\ даних\ інформаційних\ блоків \rightarrow Модуль\ збору \rightarrow Кеш \rightarrow Модуль\ подання \rightarrow Інформаційний\ ресурс,$$

а формальна модель – $Management_M = \langle X, C, Q, R, Edit, Caching, Y \rangle$, де X – множина вхідної інформації; C – множина контенту; Y – множина сформованих сторінок; Q – множина запитів; R – функція формування та подання сторінки; $Edit$ – функція редагування та модифікації контенту, $Caching$ – функція формування кешу. Ця модель поєднує переваги перших двох та реалізується шляхом кешування – модуль подання генерує сторінку один раз, надалі її в декілька разів швидше завантажують з кешу, який оновлюють автоматично (після закінчення деякого терміну часу або при внесенні змін до певних розділів сайта) або вручну адміністратором. Інший підхід – збереження інформаційних блоків під час редагування сайту і збирання сторінки з блоків за запитом користувача.

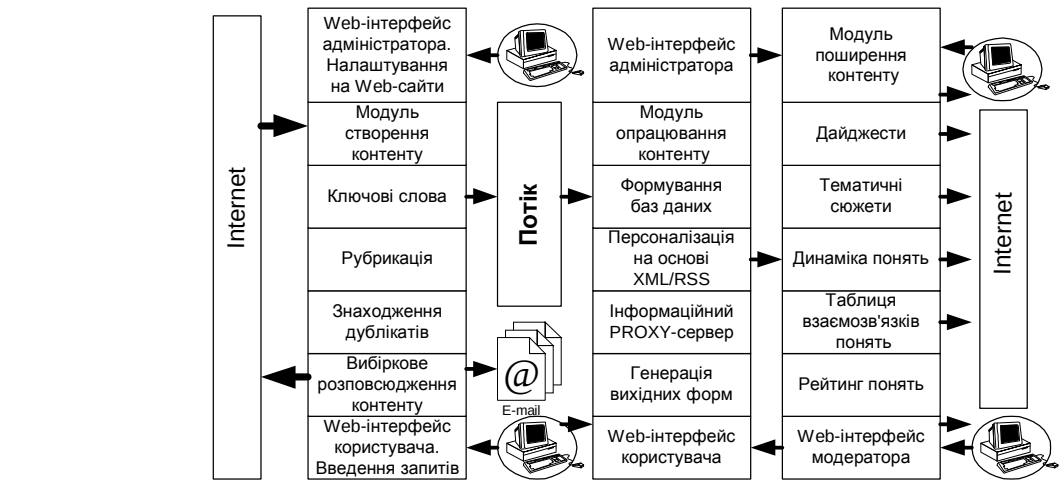
На рис. 2 подано типову схему СЕКК, що забезпечує ознайомлення, вибір категорії контенту, оформлення замовлення, здійснення взаєморозрахунків, відстеження виконання замовлення [4].

Аналіз лексико-граматичної та семантико-прагматичної побудови тексту використовують в модулі автоматичної рубрикації контенту та формування дайджестів. Виконання перерахованих в табл. 7 етапів призводить до формування тематично підібраних масивів контенту, в яких акумулюється інформація про висвітлення всіх аспектів досліджуваної проблеми, враховуючи різноманітність думок і поглядів.

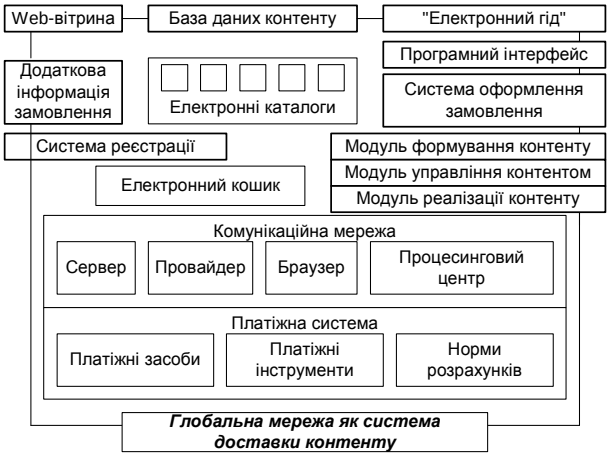
Таблиця 7

Основні етапи роботи модуля рубрикації контенту

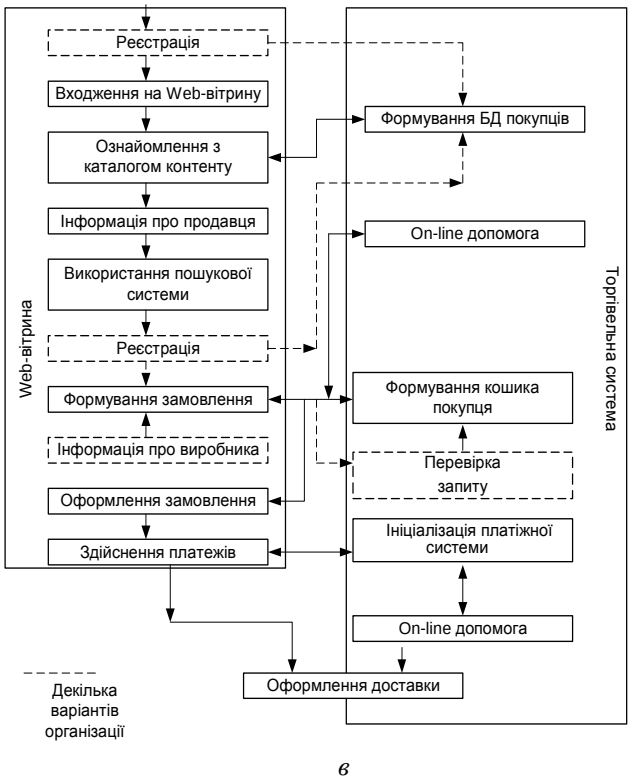
Назва	Призначення етапу
Підготовка	Визначення тематики, мети та об`єкту аналізу, його хронологічні та географічні рамки, принципи відбору.
Класифікація	Формування класифікатора відбору ключових цитат та інструкції для кодувальника.
Кодування	Кодування фрагментів текстової інформації.
Архівація	Збереження фрагментів текстової інформації в базі даних.
Аналіз	Автоматичне опрацювання фрагментів текстової інформації.



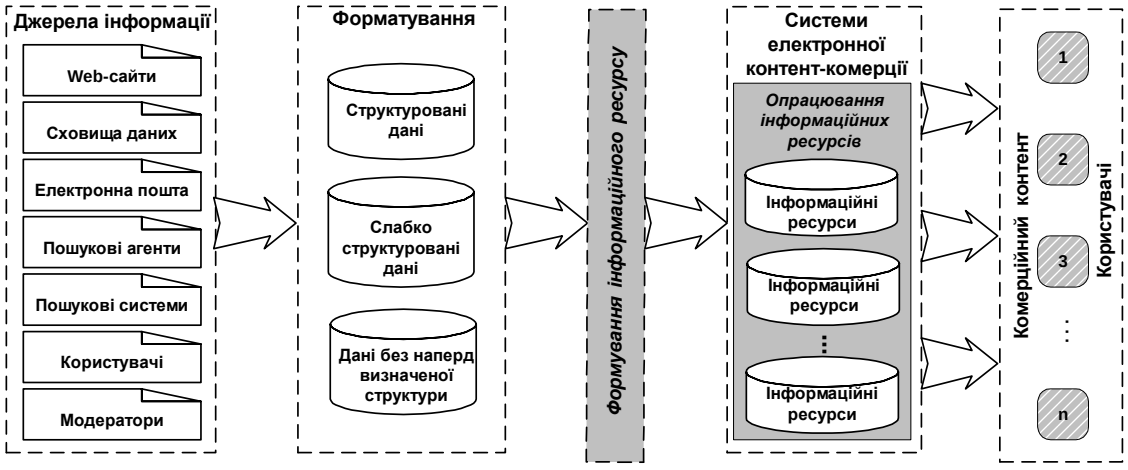
а



б



в



г

Рис. 2. Схема взаємодії модулів (а); архітектури (б); алгоритму роботи (в); функціонування СЕКК (г)

Основною формою подання і збереження інформації є природна мова, тому ефективність процесу рубрикації контенту залежить від розв'язання проблеми автоматичного опрацювання текстів, кінцева мета якого – розпізнавання їх змісту (табл. 8).

Таблиця 8

Основні етапи опрацювання текстів для розпізнавання змісту

Опрацювання	Призначення етапу
Формальне	Перетворення фрагментів тексту без звернення до аналізу його змісту. Морфологічні дані забезпечують доступ до змісту, опосередкованого через співвідношення одиниць змісту з одиницями виразу.
Змістове (семантичне)	Розпізнавання змісту окремих елементів і логіко-семантичних відношень між ними для подання семантики контенту.
Синтаксичне	Автоматично за наявності лексико-граматичних та граматичних даних до кожного слова синтаксично прив'язують словоформи у реченні.
Морфемне	Сегментування тексту, де виділення префіксів можливе без знання частин мови, а суфіксів – ні: потрібні різні їх набори та процедури відсікання суфіксів для іменників, дієслів, прикметників, прислівників.

Найпростіша ієрархія містить три рівні: мета, критерії та альтернативи. Числа на рис. 3 показують пріоритети елементів ієрархії з погляду мети, які обчислюються в MAI на основі парних порівнянь елементів кожного рівня щодо пов'язаних з ними елементів вищого рівня. Пріоритети альтернатив обчислюються на завершальному етапі методу шляхом лінійного згортання локальних пріоритетів всіх елементів. У цьому випадку відомі пріоритети критеріїв: найбільшої ваги надають критерію *Унікальність*, оскільки він об'єктивно відображає якість роботи журналіста. Наступним за вагою є *Час читання*, який відображає міру зацікавленості користувачів в конкретному матеріалі. Критерії *Користувацька оцінка* та *Кількість звернень* мають меншу вагу, оскільки не можуть користуватися довірою через легкість фальсифікації їх показників. Так оцінюють всі альтернативи за кожним з критеріїв окремо. З розрахунків випливає, що друга альтернатива є найкращою за обраними критеріями оцінювання.

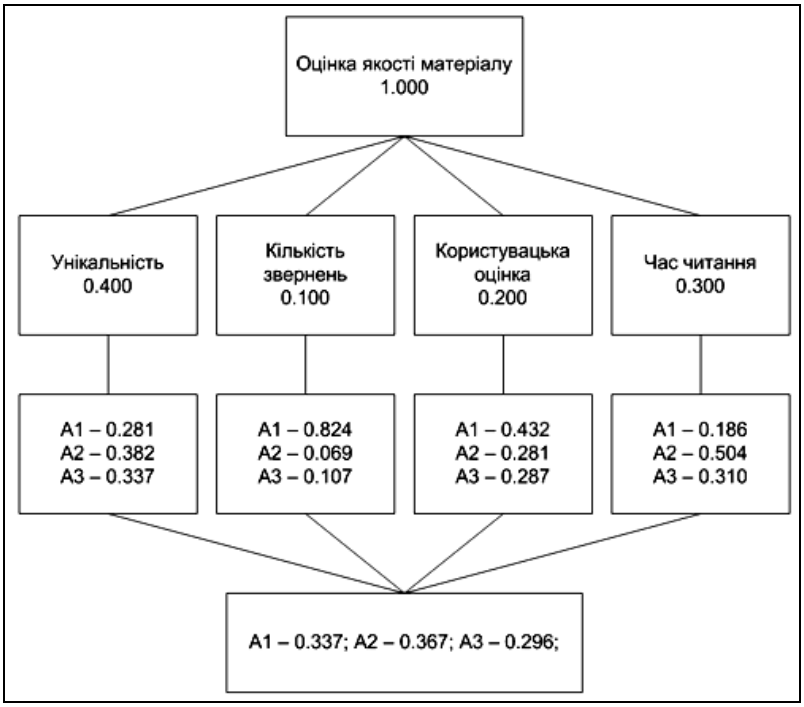
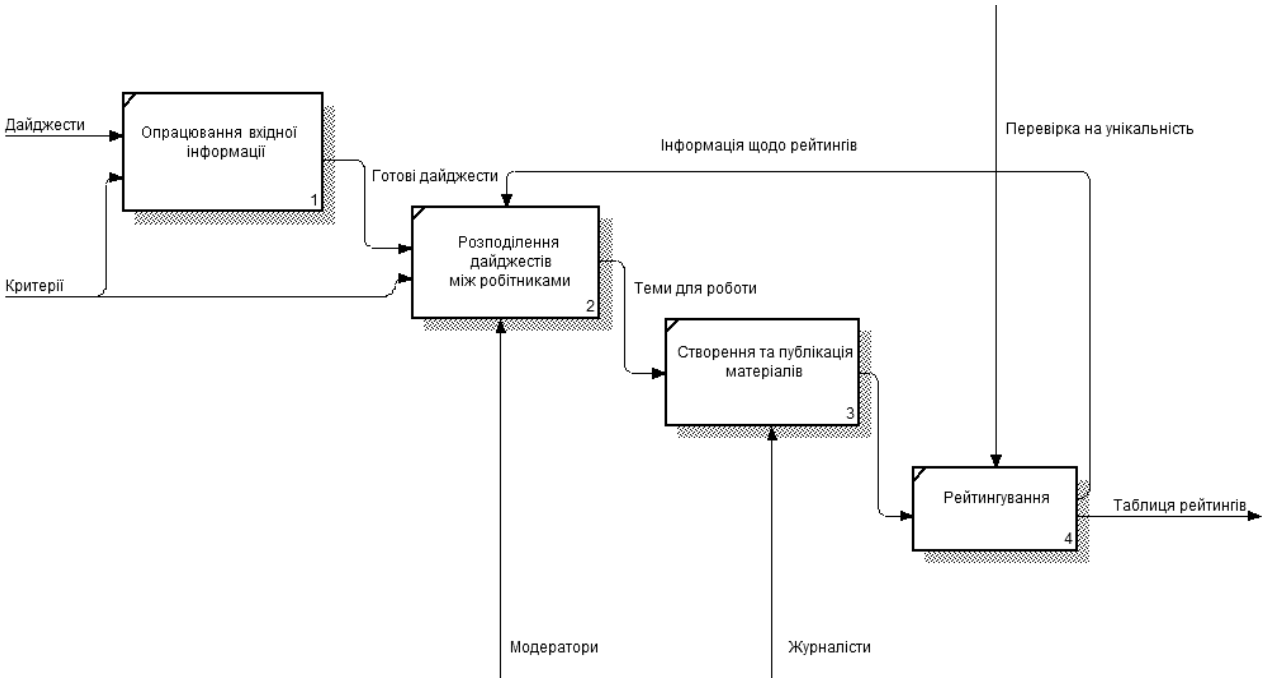


Рис. 3. Ієрархія у випадку трьох альтернатив розподілу дайджестів

Спроектуємо систему автоматичного розподілу дайджестів за допомогою методології IDEF, яку використовують для вирішення завдань моделювання складних систем. IDEF дає змогу відображати та аналізувати моделі діяльності широкого спектра складних систем у різних аспектах. Модель «To be» відображає функціонування системи загалом (рис. 4, а) та основні дані, змінювані під час роботи системи, правила її роботи, вихідні дані та необхідні для роботи ресурси. На рис. 4, б подано декомпозицію, яка поділяє роботу системи розподілу дайджестів на основні етапи.

А



б

Рис. 4. Функціонування системи розподілу дайджестів загалом (а) та декомпозиція (б)

IDEF3 діаграма системи розподілу дайджестів (рис. 5) відображає послідовність дій з умовами, що виконуються під час роботи системи для досягнення поставленої мети. Вона складається з чотирьох основних робіт, двох розгалужень та чотирьох об’єктів посилань.

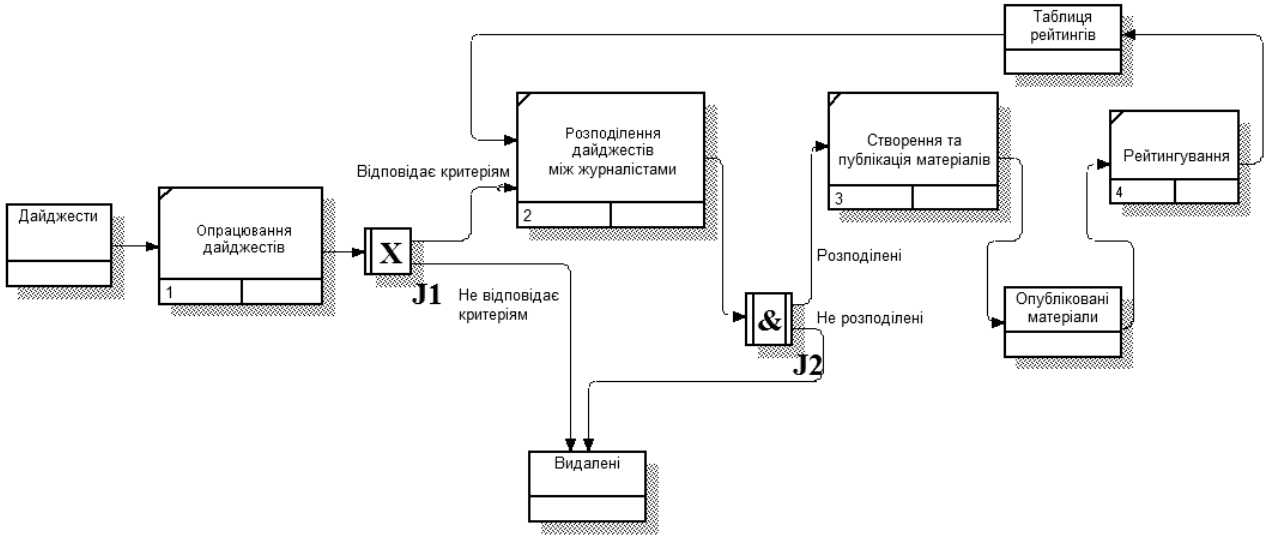


Рис. 5. Загальний вигляд IDEF3 діаграми системи розподілу дайджестів

На DFD діаграмі системи автоматичного розподілу дайджестів (рис. 6) відображено потоки даних та етапи перетворення інформації у системі. В системі використано 3 сховища даних (Джерело дайджестів, готові дайджести та таблиця рейтингів), 4 роботи (опрацювати дайджести, розподілити дайджести між працівниками, створити та опублікувати матеріали, визначити рейтинг за якістю публікації) та 2 зовнішні посилання (Модератори та журналісти). Початковим продуктом є джерело дайджестів, кінцевим – таблиця рейтингів, яку використовують в процесі розподілу дайджестів між працівниками.

Рис. 6. Загальний вигляд DFD діаграми системи розподілу дайджестів

Висновки

Досліджено проблему керування розподілом робочої сили та робочого часу працівників засобів масової інформації. Подано структуру інтелектуальної системи розподілу дайджестів. Отже, за допомогою поданих вище діаграм стає можливим розроблення програмного продукту, який відповідав би поставленим вимогам. Такий програмний продукт можливо реалізувати у вигляді як самостійної системи, так і додаткового модуля для популярних систем керування вмістом. Варіант з реалізацією як додаткового модуля універсальніший, оскільки робить продукт універсальнішим, та дає можливість працювати з різними системами керування вмістом та їх конфігураціями.

1. Береза А.М. Електронна комерція / А.М. Береза. – К.: КНЕУ. – 2002. 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. Бондаренко С. Обнаружение плагиата / С. Бондаренко // 3DNews. – Режим доступу: http://www.3dnews.ru/software/plagiarism_detection. 4. 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. 5. Подиновский В.В. О некорректности метода анализа иерархий / В.В. Подиновский // CONTROL SCIENCES. – Режим доступу: http://pu.mtas.ru/archive/ Podinovski.pdf. 6. Саати Т.Л. Принятие решений при зависимостях и обратных связях: Аналитические сети / Т.Л. Саати. – М.: Издательство ЛКИ, 2008. – 360 с. 7. Советов Б.Я. Моделирование систем (2-е изд.) / Б.Я. Советов, С.А. Яковлев. – М.: Высшая школа, 1998 р. 8. Шарапов Р.В. Система проверки текстов на заимствования из других источников / Р.В. Шарапов // Муромський інститут ГОУ ВПО. – Режим доступу: http://ceur-ws.org/Vol-803/paper16.pdf. 9. 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. 10. Graham P. Web 2.0 / P. Graham. – Nov. 2005. – http://www.paulgraham.com/web20.html.