УДК 004.738.5

В.А. Висоцька[1], Л.В. Чирун[2]
Національний університет "Львівська політехніка",
[1]кафедра інформаційних систем та мереж,
[2]кафедра програмного забезпечення

АРХІТЕКТУРА СИСТЕМ ЕЛЕКТРОННОЇ КОНТЕНТ-КОМЕРЦІЇ

З позиції системного підходу застосовано принципи опрацювання інформаційних ресурсів у системах електронної контент-комерції для реалізації життєвого циклу контенту, що дало змогу розробити методи формування, управління та супроводу комерційного контенту. Розроблено комплексний метод формування комерційного контенту для скорочення часу та зменшення ресурсів виробництва контенту, що дає можливість створити засоби опрацювання інформаційних ресурсів та реалізувати підсистему автоматичного формування контенту. Створено оперативний метод управління комерційним контентом для скорочення часу та зменшення ресурсів продажу контенту, що дає можливість реалізувати підсистему управління комерційним контентом. Реалізовано комплексний метод супроводу комерційного контенту для скорочення часу та зменшення ресурсів аналізу цільової аудиторії системи електронної контент-комерції, що дає можливість розробити підсистему супроводу комерційного контенту. Запропоновано модель життєвого циклу контенту в системах електронної комерції. Модель описує процеси опрацювання інформаційних ресурсів у системах електронної контент-комерції та спрощує технологію автоматизації управління контентом. Проаналізовано основні проблеми електронної комерції та функціональних сервісів управління контентом.

Ключові слова: інформаційний ресурс, контент, система управління контентом, життєвий цикл контенту, система електронної контент-комерції

From the perspective of systemic approach, the application of principles of information resources processing in electronic content commerce systems for content lifecycle implementation was conducted, which enabled us to develop methods for the commercial content formation, management and support. An integrated method of commercial content formation for the time and resources reduction of content production was developed. This makes it possible to create a means of information resources processing and implement subsystem of automatically generated content. In this paper operational method of commercial content management for the time and resources reduction of content sales was created, which makes it possible to implement a commercial content management subsystem. A comprehensive method of commercial content support for the time and resource reduction of the target audience analysis in electronic content commerce systems was implemented, which makes it possible to develop a commercial content support subsystem. In the given article content lifecycle model in electronic commerce systems is proposed. The model describes the processes of information resources processing in the electronic content commerce systems and simplifies the content automation management technology. Main problems of e-commerce and content function management services are analyzed in the paper.

Key words: information resources, content, content management system, content lifecycle, electronic content commerce system.

Вступ. Загальна постановка проблеми

Сучасний розвиток Інтернету сприяє зростанню потреб в інформації як виробничого фактора, так і стратегічного ресурсу, і реалізації нових форм інформаційного обслуговування [1–8]. Документована інформація, підготовлена відповідно до потреб користувачів і призначена для їх

задоволення, є інформаційним продуктом або комерційним контентом [1]. Дії для забезпечення користувачів комерційним контентом – інформаційна послуга. Інтернет-ринок – це сукупність економічних, правових, організаційних і програмних відносин з продажу/купівлі інформаційних продуктів та послуг між виробниками/постачальниками і користувачами [1, 8]. Поняття "комерційний контент" визначається як вміст інформаційних ресурсів у системі електронної контент-комерції (СЕКК); об'єкт бізнес-процесів СЕКК, наприклад, стаття, ПЗ, книга тощо; структурована множина, логічно завершена інформація, яка є об'єктом взаємовідносин між користувачем та СЕКК; набір даних без наперед визначеної структури, які існують лише в електронному вигляді; інформація комерційного призначення, неподільна в часі; основний чинник формування області діяльності, функціонування та призначення СЕКК [1].

1. Аналіз останніх досліджень та публікацій

Песпективність розроблення та впровадження СЕКК визначають такими факторами, як швидкі темпи поширення доступу до Інтернету, активний розвиток електронного бізнесу та досліджень в цій галузі (табл. 1), розширення набору інформаційних товарів та послуг, зростання попиту на інформаційні товари та послуги, відсутність теоретичного обґрунтування методів опрацювання інформаційних ресурсів, потреба в уніфікації програмних засобів опрацювання інформаційних ресурсів та активний розвиток досліджень в галузі електронного бізнесу корпораціями Google, AIIM, CM Professionals organization, EMC, IBM, Microsoft Alfresco, Open Text, Oracle, SAP та в наукових роботах Ланде Д.В., Брайчевського С.М., Григор'єва А.Н., Фурашева В.Н., McKeever Susan, Bob Boiko, Gerry McGovern, JoAnn Hackos, Ann Rockley, Russell Nakano, Bob Doyle, Woods Randy, Halverson [1–8].

Таблиця 1

Залежність задач від методів в аналізованих літературних джерелах

Аналіз літературних джерел		Задачі				
		Розроблення життєвого циклу контенту	Розроблення моделі СЕКК	Розроблення моделей опрацювання контенту	Розроблення методів опрацювання контенту	Розроблення архітектури СЕКК
Методи	**системного аналізу**	[1], [5], [6], [7]	[1], [6], [7]	[1], [5], [6], [7]	[1], [4], [6], [7]	[1]
	математичної лігвістики	[1], [2], [3], [5], [6], [7], [8], [10]	[1], [6], [7], [8]	[1], [2], [3], [5], [6], [7], [8]	[1] , [6], [7], [8]	-
	теорію формальних систем	[1], [6], [7]	[1], [6], [7], [9]	[1], [6], [7], [9]	[1], [6], [7], [9]	[1]
	сервісно-орієнтованої архітектури і методологію CMIS	[1], [6], [7]	[1], [6], [7]	[1], [6], [7]	[1], [6], [7]	[1]

Контент – сукупність всіх даних, зосереджених в середовищі інформаційної системи.

Комерційний контент – частина загального контенту, яка є об'єктом споживання користувачем та отримання прибутку його власником; текстовий, візуальний чи звуковий контент як частина досвіду користувача на інформаційному ресурсі (текст, зображення, звуки, відео та анімації); об'єкт бізнес-процесів СЕКК, наприклад, інформаційний продукт або вміст Web-сайту Інтернет-газети, Інтернет-видавництва, маркетингових досліджень, консалтингових послуг тощо [1].

Управління контентом – функції управління для отримання, аналізу, збереження, пошуку і поширення контенту (ISO/IEC/IEEE 24765:2010(E), ISO/IEC 2382-1:1993, Information technology).

Інформаційний ресурс – об'єкт дії засобів і технологій; сукупність документів у інформаційних системах (бібліотеках, архівах, банках даних тощо) (Закони України "Про Національну програму інформатизації", "Про бібліотеки і бібліотечну справу", ст. 1).

Інформаційний продукт – результат застосування дії технології до інформаційного ресурсу; документована інформація, яка підготовлена і призначена для задоволення потреб користувачів (Закон України "Про національну програму інформатизації", ст. 1).

Інформаційний контент – множина метамоделей та моделей примірників у CDIF-переведенні. (ISO/IEC/IEEE 24765:2010(E), 3.1398, ISO/IEC 15474-1:2002, Information technology)

Електронна комерція – галузь цифрової економіки, що містить усі фінансові та торгові транзакції через комп'ютерні мережі та бізнес-процеси, пов'язані з проведенням цих транзакцій.

Електронна контент-комерція – галузь електронної комерції, де об'єктом фінансових/торгових транзакцій та бізнес-процесів є комерційний контент.

Система електронної комерції (СЕК) – система для ведення електронного бізнесу через комп'ютерні мережі, а яка також містить всі фінансові та торгові транзакції та бізнес-процеси, пов'язані з проведенням цих транзакцій, а також пов'язаних з ними організаційних ресурсів, таких як людські, технічні, інформаційні та фінансові, що забезпечують/ розподіляють комерційний контент.

Система електронної контент-комерції (СЕКК) – система опрацювання комерційного контенту, а також пов'язаних з ними організаційних ресурсів, таких як людські, технічні, інформаційні та фінансові, що забезпечують і розподіляють комерційний контент.

Існуючі СЕК не підтримують весь життєвий цикл контентного потоку та не вирішують основних проблем опрацювання інформаційних ресурсів – формування та реалізації контенту (табл. 2).

Таблиця 2

Порівняння особливостей систем електронної комерції та електронної контент-комерції

Назва характеристики функціонування системи	СЕК	СЕКК
Віртуальність (відсутність особистого контакту між суб'єктами процесу купівлі/продажу)	+/–	+
Інтерактивність (інформаційне забезпечення запиту користувача в інтерактивному режимі німого діалогу)	+/–	+
Глобальність (відсутність часових, просторових, асортиментно-товарних, адміністративних, соціально-демографічних меж)	+/–	+
Динамічність (здатність on-line торгівлі до моментальних змін й адаптації до нових умов)	+/–	+
Ефективність (забезпечення попиту, прибутку, економічних вигод, соціального ефекту)	+/–	+
Наявність нематеріального товару (товар як інформаційний продукт без матеріального носія)	–	+
Постійна кількість товару (виробництво товару не вимагає тиражування, а лише розсилку копій)	–	+
Ріст кількості різновиду товару (кожен товар унікальний)	+/–	+
Відсутність складу для збереження товару, наявність лише репозиторію для інформаційний продукту	–	+
Збереження товару в базах даних або репозиторіях	–	+
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	–	+
Автоматичне формування товару	–	+
Автоматичне форматування товару	–	+
Вплив досвіду користувача на збільшення обсягу продажів	+/–	+

Кількість контентних потоків більша, ніж шляхів переміщення товарів на промислових підприємствах [1]. Значна частина контентних потоків складається з легко формалізованих і автоматизованих процедур. Основна проблема – відсутність загального підходу до процесу моделювання, проектування та розроблення СЕКК [1]. Відсутність загальної та детальної класифікації СЕК та СЕКК приводить до проблеми визначення та формування загальних методів проектування та розроблення архітектури та алгоритмів функціонування СЕК і СЕКК [1]. Це обґрунтовує мету, актуальність, доцільність та напрями дослідження. Вихідною інформацією функціонування СЕКК є дані про призначення й умови роботи системи, які визначають основну мету моделювання СЕКК і дають змогу сформулювати вимоги до моделі системи S та моделей управління контентом [9]. Модель СЕКК $S = \langle X, C, V, H, Function, T, Y \rangle$ – це множини величин, що описують функціонування системи і утворюють підмножини, подані в табл. 3. Величини x_i, c_r, v_l, h_k, y_j є елементами підмножин, які містять детерміновані і стохастичні складові. Функціонування СЕКК S описано функцією $y_j(t_i + \Delta t) = Function(x_i, c_r, v_l, h_k, t_i)$ [9], де x_i – це запити відвідувачів/користувачів до СЕКК. Згідно із Google Analytics y_j – це кількість відвідувань за період часу Δt, середній час знаходження 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) [4]. Впливи величин c_r, v_l, h_k, на y_j як результат роботи СЕКК є невідомими та недослідженими [1–10].

Таблиця 3

Складові системи електронної контент-комерції

Множини	Значення	Діапазон	Множина
Вхідні впливи на систему	$x_i \in X$	$i = \overline{1, n_X}$	$X = \{x_1, x_2, \mathbf{K}, x_{n_X}\}$
Впливи потоку контенту на систему	$c_r \in C$	$r = \overline{1, n_C}$	$C = \{c_1, c_2, \mathbf{K}, c_{n_C}\}$
Впливи зовнішнього середовища	$v_l \in V$	$l = \overline{1, n_V}$	$V = \{v_1, v_2, \mathbf{K}, v_{n_V}\}$
Внутрішні (власні) параметри системи	$h_k \in H$	$k = \overline{1, n_H}$	$H = \{h_1, h_2, \mathbf{K}, h_{n_H}\}$
Вихідні характеристики системи	$y_j \in Y$	$j = \overline{1, n_Y}$	$Y = \{y_1, y_2, \mathbf{K}, y_{n_Y}\}$
Час трансакції управління контенту	$t_i \in T$	$i = \overline{1, n_T}$	$T = \{t_1, t_2, \mathbf{K}, t_{n_T}\}$

Величини x_i, c_r, v_l, h_k, y_j є елементами підмножин, які містять детерміновані і стохастичні складові. Функціонування СЕКК S описано функцією $y_j(t_i + \Delta t) = Function(x_i, c_r, v_l, h_k, t_i)$ [9], де x_i – це запити відвідувачів/користувачів до СЕКК. Згідно Google Analytics y_j – це кількість відвідувань за період часу Δt, середній час знаходження 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) [4]. Впливи величин c_r, v_l, h_k, на y_j як результат роботи СЕКК є невідомими та недослідженими [1–10].

Життєвий цикл контенту (англ. Content lifecycle) – це складний процес, який проходить контент під час управління через різні етапи публікації [1]. Існуючі моделі життєвого циклу контенту не містять усіх етапів процесів опрацювання інформаційних ресурсів: формування контенту, управління контентом та супроводу контенту (табл. 4). Дослідження динаміки потоку комерційного контенту та побудова моделей опрацювання інформаційних ресурсів СЕКК є

важливим та актуальним завданням [1–10]. При розгляді динаміки тематичних потоків контенту виявлено обмеженість моделей (табл. 5), що відкриває шлях для подальших досліджень.

Таблиця 4

Порівняння моделей життєвого циклу комерційного контенту

Автор моделі	Формування	Управління	Супровід
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 5

Моделі опрацювання текстового комерційного контенту

Модель	Особливості моделей опрацювання текстового контенту
Бартона-Кеблера	Описує процес старіння контенту, втрати його актуальності, визначення швидкості розвитку окремих тематик або всього контентного простору, має точний розв'язок у вигляді експоненти. Є сумнівною щодо інтерпретації результатів. Монотонно зростає, не описує процесів з локальними екстремумами.
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	Поєднує відносну простоту формулювання задачі з можливістю варіювати розв'язок за допомогою набору параметрів, що мають прозорий фізичний зміст. Вивчення динаміки лише окремого тематичного потоку. Розмірність параметрів та їхній вимір не враховують.
Аналітична	Описує процес старіння контенту, втрати його актуальності. Обов'язкова наявність словника ключових слів.

2. Виділення проблем опрацювання комерційного контенту

Формальна модель СЕКК не розкриває механізми 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 та реалізації контенту і вирішують не всі проблеми управління контентом, наприклад, подання множини контенту кінцевому користувачу згідно з його запитом, історії або інформаційного портфеля, 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 та дублювання змісту, рубрикація, вибіркове поширення контенту. Недолік моделей управління контентом – це відсутність зв'язків між вхідними даними, контентом та вихідними

даними в СЕКК [1]. Існує невідповідність між методами і засобами опрацювання інформаційних ресурсів та принципами побудови систем електронної контент-комерції.

3. Формулювання мети статті

Необхідно розв'язати задачу розроблення методів та програмних засобів формування, управління та супроводу інформаційних продуктів у вигляді теоретично обґрунтованої концепції шляхом автоматизації опрацювання інформаційних ресурсів у системах електронної контент-комерції для збільшення обсягів продажу контенту постійному користувачу, активного залучення потенційних користувачів та розширення меж цільової аудиторії. Метою роботи є розроблення архітектури СЕКК, уніфікованих методів та програмних засобів для опрацювання інформаційних ресурсів у СЕКК. Метою роботи визначено необхідність вирішення таких завдань: розроблення формальної моделі СЕКК для визначення недоліків існуючих методів та засобів опрацювання ресурсів; розроблення уніфікованих методів опрацювання інформаційних ресурсів у системах СЕКК для створення узагальненої типової архітектури СЕКК; розроблення архітектури СЕКК для реалізації етапів життєвого циклу комерційного контенту; розроблення програмних засобів опрацювання інформаційних ресурсів СЕКК для скорочення часу і затрат на формування контенту, управління контентом та реалізації контенту, підвищення якості опрацювання інформаційних ресурсів шляхом використання випробуваних вирішень. Об'єкт досліджень – процеси організації життєвого циклу інформаційних продуктів у СЕКК. Предмет досліджень – уніфіковані методи та програмні засоби формування, управління та реалізації інформаційних продуктів. Дослідження, виконані під час роботи над статтею, ґрунтуються на методах системного аналізу (проектування системи електронної контент-комерції), елементах загальної теорії систем (проектування системи електронної контент-комерції), апараті теорії реляційних баз даних (розроблення баз даних), теорії проектування за допомогою CASE-засобів (проектування системи електронної контент-комерції), теорії множин (розроблення моделей та методів опрацювання інформаційних ресурсів), теорії ймовірності (розроблення методів, визначення тональності контенту), математичній лінгвістиці (розроблення методів контент-аналізу текстової інформації), математичній статистиці (розроблення методів аналізу роботи системи) та теорії імітаційного моделювання (розроблення моделі та архітектури системи електронної контент-комерції).

4. Аналіз отриманих наукових результатів

Запропоновані загальні принципи проектування архітектури СЕКК дають можливість ефективно реалізовувати технології опрацювання інформаційних ресурсів СЕКК (рис. 1).

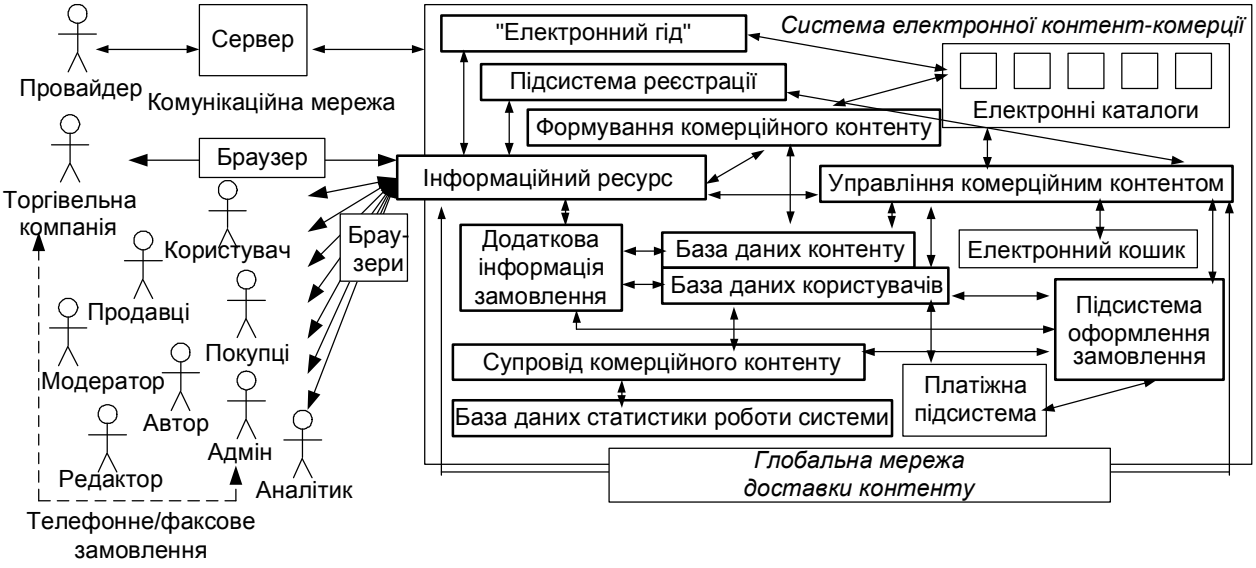

Рис. 1. Загальна структура системи електронної контент-комерції

Формальна модель систем електронної контент-комерції – це шістка

$$S = \langle X, Formation, C, Management, Realization, Y \rangle,$$

де $X = \{x_1, x_2, \mathbf{K}, x_{n_X}\}$ – множина вхідної інформації, *Formation* – оператор формування контенту, $C = \{c_1, c_2, \mathbf{K}, c_{n_C}\}$ – множина контенту, *Management* – оператор управління контентом, *Realization* – оператор супроводу контенту та $Y = \{y_1, y_2, \mathbf{K}, y_{n_Y}\}$ – множина вихідної інформації.

На рис. 1 подано типову схему СЕКК, що забезпечує ознайомлення, вибір категорії контенту, оформлення замовлення, здійснення взаєморозрахунків, відстеження виконання замовлення (рис. 2).

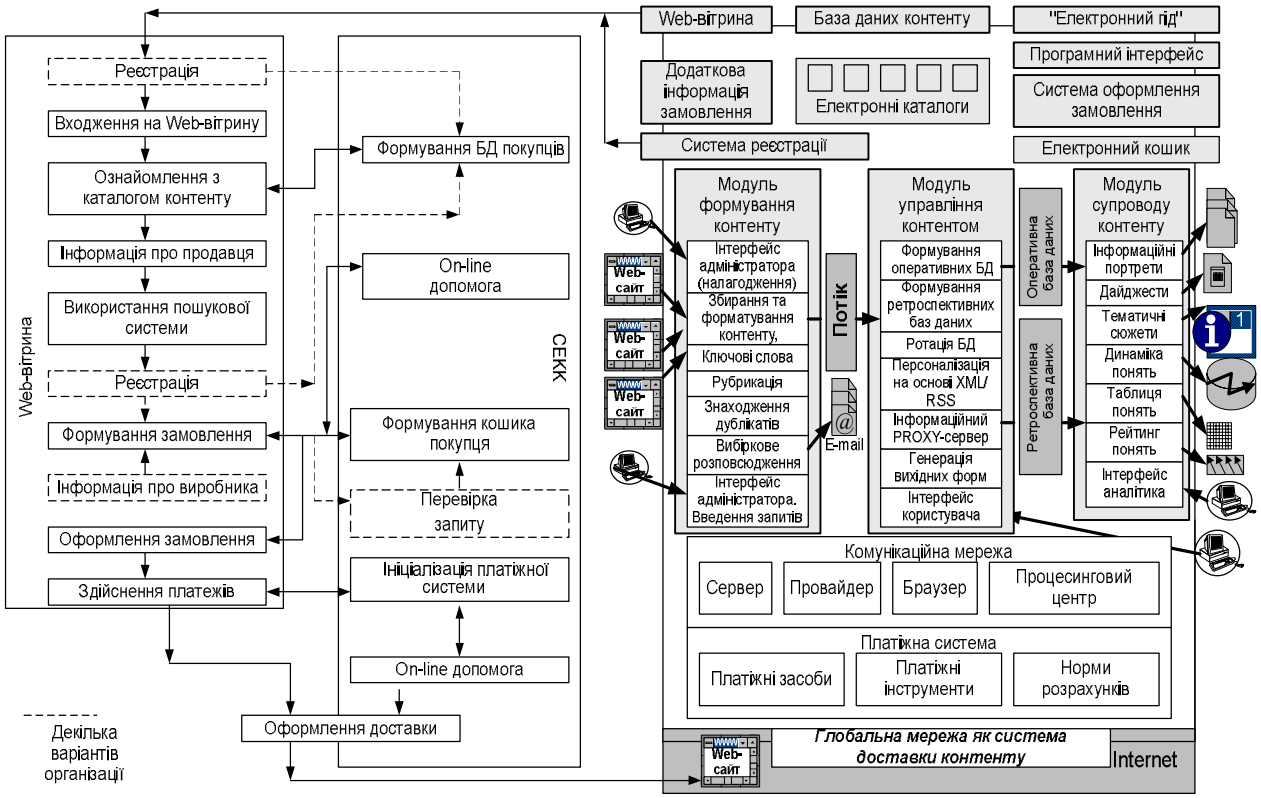

Рис. 2. Структурна схема функціонування системи електронної контент-комерції

Оператор формування комерційного контенту – відображення комерційного контенту в новий стан, який відрізняється від попереднього додатковими значеннями як актуальність, достовірність, унікальність, повнота, точність тощо.

Оператор управління комерційним контентом – відображення комерційного контенту в новий стан, який відрізняється від попереднього відповідно значеннями визначальних параметрів (актуальність, повнота, релевантність, автотичність, достовірність), що задовольняють наперед визначені вимоги.

Оператор супроводу комерційного контенту – відображення комерційного контенту в колекцію значень у результаті статистичного аналізу відвідуваності інформаційного ресурсу та активної діяльності постійних користувачів на цьому ресурсі.

Вхідні дані з різних джерел інформації $x_i \in X \quad i = \overline{1, n_X} \quad X = \{x_1, x_2 \mathbf{K}, x_{n_X}\}$

Запити користувачів $q_d \in Q \quad d = \overline{1, n_Q} \quad Q = \{q_1, q_2 \mathbf{K}, q_{n_Q}\}$

Оператор формування контенту $x_i \to c_{r+1} = Formation(x_i, c_r, u_F, t, \Delta t) = x_i + c_r + \Delta c$

Внутрішні параметри системи $h_k \in H \quad k = \overline{1, n_H} \quad H = \{h_1, h_2, \mathbf{K}, h_{n_H}\}$

Комерційний контент	$c_r \in C$	$r = \overline{1, n_C}$	$C = \{c_1, c_2 \mathbf{K}, c_{n_C}\}$
Параметри зовнішнього середовища	$v_l \in V$	$l = \overline{1, n_V}$	$V = \{v_1, v_2, \mathbf{K}, v_{n_V}\}$
Оператор управління контентом	$q_d \rightarrow z_{w+1} = Management\,(z_w, c_r, h_k, u_M, t, \Delta t)$		
Оператор супроводу контенту	$z_w \rightarrow y_i = Support\,(c_r, q_d, v_l, u_S, t, \Delta t) = \{a_1, a_2, ..., a_g\}$		
Елементи інформаційного ресурсу	$z_w \in Z$	$z = \overline{1, n_Z}$	$Z = \{z_1, z_2 \mathbf{K}, z_{n_Z}\}$
Час транзакції	$t_p \in T$	$p = \overline{1, n_T}$	$T = \{t_1, t_2 \mathbf{K}, t_{n_T}\}$
Статистичні дані роботи системи	$y_j \in Y$	$j = \overline{1, n_Y}$	$Y = \{y_1, y_2 \mathbf{K}\, y_{n_Y}\}$

Незважаючи на різні класи СЕКК і широкі можливості реалізації моделей, виділено основні закономірності переходу від процесів формування інформаційних ресурсів до їх реалізації. Відповідно запропоновано формальні моделі формування, управління та реалізації інформаційних ресурсів, які дають змогу оптимально реалізувати архітектуру СЕКК. Комплексний метод формування контенту забезпечує збирання інформації з різноманітних Web-сайтів та її форматування; виявлення ключових слів і понять контенту; автоматичну рубрикацію контенту; виявлення дублювання змісту контенту; вибіркове поширення контенту. Оперативний метод управління контентом забезпечує формування БД і забезпечення доступу до неї; формування оперативних і ретроспективних БД; ротацію БД; персоналізацію роботи користувачів; збереження персональних запитів і джерел; ведення статистики роботи; забезпечення пошуку в БД; генерацію вихідних форм; інформаційну взаємодію з БД інших підсистем. Комплексний метод супроводу контенту забезпечує формування інформаційних портретів; формування дайджестів; виявлення тематичних сюжетів; побудова таблиць взаємозв'язку понять; розрахунок рейтингів понять, виявлення нових подій, їхнє відстеження й кластерізація. Для повнофункціональної СЕКК характерна складна система взаємозв'язаних операцій, методів, прийомів поданих на рис. 2. Створення бази анотацій – це створення бази даних пошукових образів первинного контенту та їх кластерізація (формування груп контенту із близькими за деякими критеріями пошуковими образами). Бази анотацій (пошукових образів кластерів, які використовують у процесі пошуку) пов'язані з базою кластерів, кожен запис якої відповідає визначеному кластерові і містить його опис, виконаний методами автоматичного реферування (рис. 2).

Модель формування комерційного контенту

$$Formation = \left\langle \begin{matrix} X, Gathering, Formatting, KeyWords, Backup, \\ Caterization, BuDigest, Dissemination, T, C \end{matrix} \right\rangle$$

Вхідні дані з різних джерел	$x_i \in X$
Оператор збирання/створення контенту	$C_0 = Gathering\,(X, U_G, T)$
Оператор виявлення дублювання контенту	$C_1 = Backup\,(C_0, T, U_B)$
Оператор форматування контенту	$C_2 = Formatting\,(C_1, U_{FR}, T)$
Оператор виявлення ключових слів	$C_3 = KeyWords\,(C_2, U_K, T)$
Оператор автоматичної рубрикації контенту	$C_4 = Categoriza\,tion\,(C_3, U_{Ct}, T)$
Оператор формування дайджестів контенту	$C_5 = BuDigest\,(C_4, U_D, T)$
Оператор вибіркового поширення контенту	$C_6 = Dissemination\,(C_5, U_{Ds}, T)$
Час транзакції формування контенту	$t_p \in T$
Комерційний контент	$c_r \in C$

Задача виявлення нових тематик з потоку контенту припускає, що на вхід СЕКК послідовно надходить новий контент. Він надходить безпосередньо від засобів сканування (політематичний потік) або від контентного роутера, системи вибірного поширення контенту, які відібрані за тематичним запитом. Далі виявляються нові тематики, що описуються в контенті, для яких за

допомогою окремих програмних модулів у тимчасовій ретроспективі формуються ланцюжки подібного контенту (сюжетні ланцюжки). Контент, що відображає різні нові тематики, є основою нових груп взаємозалежного контенту (кластерів). Технологія синдикації контенту містить навчання програм збирання даних структурним особливостям окремих джерел, безпосереднє сканування контенту, приведення до загального формату в XML, рубрикацію.

Модель супроводу комерційного контенту

$$Support = \left\langle \begin{matrix} Q, C, H, V, T, BuInfPort, IdThemTop, \\ ConCorrTablConc, CalRankConc, Z, Y \end{matrix} \right\rangle$$

Запити користувачів	$q_d \in Q$
Комерційний контент	$c_r \in C$
Оператор формування портретів комерційного контенту та користувачів	$Y_{Pc} = BuInfPort(V_{Pc}, C, H, Q, T)$ $Y_{Pq} = BuInfPort(V_{Pq}, Q, H, Z, T)$
Оператор виявлення тематичних сюжетів	$Y_T = IdThemTop(C, H, Q, V_T, T)$
Оператор побудови взаємозв'язку контенту	$Y_C = ConCorrTablConc(C, V_c, T)$
Оператор розрахунку рейтингів комерційного контенту та користувачів	$Y_{Rc} = CalRankConc(C, Q, H, Y_C, V_{Rc}, Spam, Tonality, T)$ $Y_{Rm} = CalRankConc(C, Q, H, Y_C, V_{Rm}, T)$
Внутрішні параметри системи	$h_k \in H$
Параметри зовнішнього середовища	$v_l \in V$
Елементи інформаційного ресурсу	$z_w \in Z$
Час транзакції	$t_p \in T$
Статистичні дані роботи системи	$y_j \in Y$

Формування комерційного контенту – комплекс заходів забезпечення контролю опрацювання даних з різних джерел інформації для створення комерційного контенту з набором а додаткових значень – таких, як актуальність, достовірність, унікальність, повнота, точність тощо (рис. 3).

Управління комерційним контентом – комплекс заходів забезпечення підтримки значень таких визначальних параметрів комерційного контенту, як актуальність, повнота релевантність, автотичність, достовірність до визначених вимог за набором критеріїв (рис. 3).

Супровід комерційного контенту – комплекс заходів забезпечення функціонування системи електронної контент-комерції згідно із визначеними вимогами і будь-які подальші зміни в цих вимогах (рис. 3).

Формування інформаційних портретів. Використання контент-аналізу текстової інформації в СЕКК дає змогу визначити поширеність ознаки досліджуваного контенту. При цьому важливо не стільки абсолютне, скільки відносне значення ознаки, тобто характеристика її місця (частки) серед інших ознак. Наприклад, відсоток обговорення користувачами форуму економічних питань відносно політичних, або відсоток позитивних коментарів щодо статей відносно негативних та відносно всіх коментарів щодо цієї категорії статей в Інтернет-газеті. Вимірювання співвідношення між ознаками в текстах дає емпіричний матеріал для розуміння функціональних зв`язків між елементами відображеної в текстах реальності, наприклад, визначення настрою аудиторії Інтернет-газети щодо економічної або політичної ситуації в країні/світі. За наявності текстів, що мають хронологічну послідовність, отримують низку фіксованих у часі "портретів" досліджуваної реальності (зміна попиту на категорію контенту залежно від сезону, наприклад, фантастику читають більше взимку, а детективи – влітку) або "портретів" цільової аудиторії (зміна попиту на категорію контенту залежно від статті, наприклад, попит на політичні статті перед виборами), що дає змогу висувати гіпотези прогностичного характеру про функціонування елементів системи.

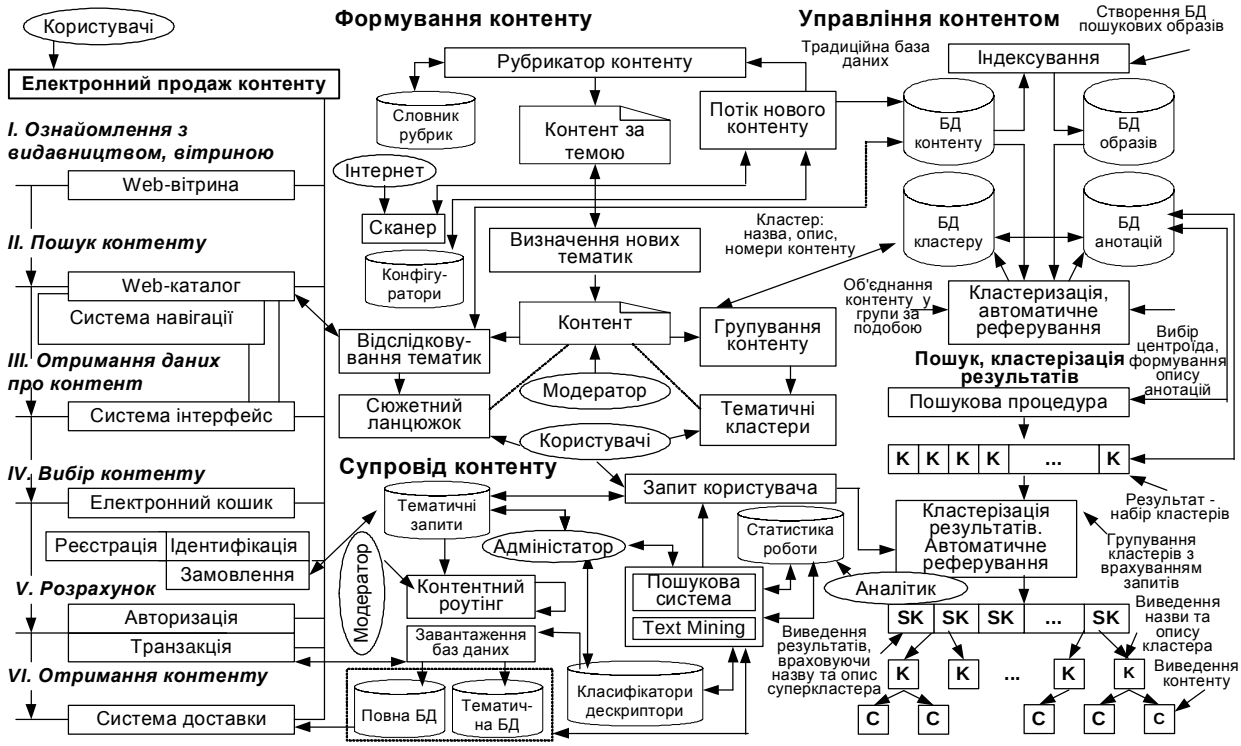

Рис. 3. Структурна схема опрацювання інформаційних ресурсів у СЕКК

Формування дайджестів – коротких змістів публікації, для створення яких використовують контент-аналіз із врахуванням частотних ваг слів із сформованого словника понять. Процес формування дайджестів складається з алгоритмів формування словника понять (алг. 1), визначення дублювання контенту (алг. 2) та створення дайджесту (алг. 3).

Алгоритм 1. Формування словника понять.

Етап. 1. Формування словника понять.

Крок 1. Послідовне виділення всіх слів вхідного потоку контенту.

Крок 2. Побудова алфавітно-частотного словника.

Крок 3. Нормалізація слів через автоматичний морфологічний аналіз.

Крок 4. Модифікація алфавітно-частотного словника.

Крок 5. Приписування словам ваги w (частоти вживання).

Крок 6. Вилучення зі словника незначних слів ($W \le k$, де k – значення порогу вилучення слова).

Етап. 2. Вибір тематичного словника відповідно до запиту.

Етап. 3. Коригування ваг слів алфавітно-частотного словника з урахуванням тематичного словника.

Етап. 4. Вибір $N = n$ слів із більшою вагою із алфавітно-частотного словника, де $n = const$ задається модератором.

Алгоритм 2. Визначення дублювання контенту.

Етап. 1. Задання початкових даних.

Крок 1. Задання модератором кількості слів у ланцюжку $m = const$.

Крок 2. Задання коефіцієнта унікальності ланцюжків $U = const$.

Крок 3. Задання меж коефіцієнта вживання ключових слів $K = [a_1, a_2]$, де $a_1 = const$ та $a_2 = const$.

Крок 4. Розбиття контенту на n ланцюжків по m слів.

Крок 5. Розрахунок частот вживання ключових слів k_i.

Етап. 2. Визначення дублів.

Крок 1. Порівняння між собою ланцюжків слів для всього контенту.

Крок 2. Розрахунок коефіцієнтів унікальності ланцюжків u_i.

Крок 3. Порівняння коефіцієнтів унікальності ланцюжків u_i із U. При $\frac{1}{n}\sum_{i}^{n} u_i < U$ контент маркують як непридатний.

Крок 4. Порівняння частоти k_i із коефіцієнтом K. Якщо $k_i < a_1$ або $k_i > a_{21}$, то контент маркувати як непридатний.

Алгоритм 3. Створення дайджесту.

Етап. 1. Вибір контенту з урахуванням його ваги.

Крок 1. Задання розміру дайджесту C.

Крок 2. Виконання алгоритму 1.

Крок 3. Послідовне визначення ваги кожного контенту як суми ваг окремих його слів, тобто $W = \sum_i w_i$.

Крок 4. Сортування вхідного потоку контенту за величинами ваг.

Крок 5. Визначення змістовних дублів контенту за статистичним критерієм унікальності тексту $U \geq 0{,}9$ (алг. 2).

Крок 6. Фільтрування контенту, не придатного для побудови дайджестів (при $W \leq l$, де l – значення порогу вилучення контенту за допомогою правил структуризації та модерації контенту із самонавчанням) та статистично змістових дублів.

Крок 7. Вибір $V = q$ контенту із більшою вагою, де $q = const$ і задається модератором.

Етап. 2. Побудова тексту дайджесту з відібраного контенту.

Крок 1. Побудова словника з відібраного контенту (алг. 1).

Крок 2. Застосування контент-аналізу до тексту (табл. 6).

Крок 3. Фільтрування речень, що не відповідають семантичним правилам структуризації та модерації контенту.

Крок 4. Формування гіпертекстового подання дайджесту, його змісту і посилання на вихідні джерела.

Етап. 3. Редагування сформованого тексту дайджесту.

Крок 1. Перевірка обсягу c_i сформованого контенту. Якщо $c_i < C$, то виконання кроку 2, інакше етапу 4.

Крок 2. Видалення з вхідного потоку контенту, який використали для формування дайджесту.

Крок 3. Виконання етапів 1–2.

Крок 4. Дописування до попередньо сформованого дайджесту отриманого та перехід до кроку 1.

Етап. 4. Форматування тексту дайджесту як окремого контенту та збереження в БД із посиланням на джерело.

Таблиця 6

Етапи контент-аналізу текстової інформації

Етап	Характеристика етапу контент-аналізу текстової інформації
1	2
Визначення сукупності джерел або контенту	За допомогою набору заданих критеріїв, яким відповідає кожний контент: заданий тип джерела; один тип контенту; задані сторони, що беруть участь у процесі комунікації; зіставлений розмір повідомлень (мінімальний обсяг/довжина); частота появи повідомлень; спосіб розповсюдження повідомлень; місце розповсюдження повідомлень; час появи повідомлень тощо.
Контент-аналітичний відбір	Формування вибіркової сукупності контенту за критеріями обмеженої вибірки з більшого масиву інформації за допомогою процедури із набором точно визначених дій для опрацювання без будь-яких змін усіх об`єктів дослідження.
Виявлення лінгвістичних одиниць	Дотримання чітких вимог до вибору лінгвістичної одиниці аналізу: достатньо велика для інтерпретації значення; достатньо мала, щоб не інтерпретувати багато значень; легко ідентифікується; кількість одиниць достатньо велика для проведення вибірки. При прийнятті за одиницю аналізу теми її розмір не має виходити за межі абзацу; нова тема виникає при виникненні нових характеристик лінгвістичної одиниці.

Продовження табл. 6

1	2
Виділення одиниць обчислення та формування класифікатора	Одиниці обчислення можуть збігатися із змістовними одиницями або носити специфічний характер. У першому випадку процедура аналізу зводиться до підрахунку частоти вживання виділеної змістовної одиниці, в іншому – дослідник на основі аналізованого матеріалу і цілей дослідження висуває одиниці обчислення (фізична протяжність текстів; площа тексту, заповнена змістовними одиницями; кількість рядків, абзаців, знаків, колонок тексту; розмір/вид файлу; кількість рисунків з певним змістом, сюжетом).
Процедура обчислення	Стандартні прийоми класифікації за виділеними угрупуваннями із формул математичної статистики та теорії ймовірності.
Інтерпретація результатів	Охоплює всі здобуті фрагменти тексту, висновки спираються не на частину результатів, а враховуються всі без винятку. Виявляються і оцінюються характеристики тексту, які дозволяють робити висновки про те, що хотів підкреслити або приховати його автор, або на основі статистичного набору підрахованих коефіцієнтів за певний період часу на визначену категорію прогнозують зміни попиту на контент.

Виявлення тематичних сюжетів. Контент, що відображає різні нові тематики, є основою нових груп взаємозалежного контенту при виявленні тематичних сюжетів з такими процедурами:

1) контроль зсередини системи – призначення рівня доступу користувачів до різного контенту;

2) інтеграція контенту – перенесення контенту в нове рішення;

3) підтримка контенту різного типу – зберігання і сортування контенту в центральному репозиторії;

4) детальна документація і контекстно-інтелектуальна довідка;

5) рейтингова система оцінювання статей сайту;

6) шаблонні зміни – загальні зміни форматування контенту частини сайту, відображені на весь сайт;

7) підтримка workflow – створення автоматизованих бізнес-процесів для конкретного контенту;

8) маркування контенту – додавання нових категорій і маркерів до контенту до/після збереження;

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

10) контент-аналіз контентних потоків у системі;

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

12) побудова таблиць взаємозв'язку понять.

Розрахунок рейтингів понять ґрунтується на процедурі підрахунку результату контент-аналізу з врахуванням коефіцієнта c співвідношення позитивних і негативних (щодо обраної позиції) оцінок, думок, аргументів, висвітлених в коментарях користувачів щодо контенту СЕКК. Якщо кількість позитивних оцінок перевищує кількість негативних, то використовується формула $c = \frac{f^2 - f \cdot n}{r \cdot t}$, де f – кількість позитивних оцінок; n – кількість негативних оцінок; r – обсяг аналізованого змісту тексту; t – загальний обсяг тексту. Якщо кількість позитивних оцінок менша за негативну, то використовують формулу $c = \frac{f \cdot n - n^2}{r \cdot t}$.

У табл. 7 подано список наявності етапів життєвого циклу комерційного контенту в розроблених системах. У табл. 6 також подано результати роботи розроблених систем з Google Analytics.

Таблиця 7

Результати функціонування систем електронної контент-комерції

№ з/п	Характеристика	Інформаційний ресурс				
		Фотогалерея Висоцьких	Вголос	Татьяна	Прес-Тайм	AutoChip
1	Підсистема формування контенту	+/–	+	–	+/–	–
2	Підсистема управління контенту	+	+	–	+	+
3	Підсистема супроводу контенту	+/–	+	–	+	+/–
за період часу з 10.2011р. по 11.2011р.						
1	Відвідування	142	199873	43	124653	372
2	Середній час відвідування сайту (хв:с)	2:04	2:48	3:40	2:18	2:49
3	Показник відмовлень (%)	61,27	65,99	60,47	59,89	58,06
4	Досягнута мета	6	0	0	0	54
5	Динаміка (%)	-8,97	39,74	-58,65	23,18	17,72
6	Перегляд сторінок	349	425576	98	245632	1013
7	Кількість сторінок за відвідування	2,46	2,13	2,28	2,09	2,72
8.	Нові відвідування (%)	70,42	36,84	55,81	23,54	75,27
9	Абсолютно унікальні відвідувачі	112	98845	27	45321	290
10	Джерело трафіка – пошукові системи (%)	76,76	41,91	60,47	42,75	54,03
11	Джерело трафіка – прямий трафік (%)	11,27	10,50	39,53	10,50	26,34
12	Джерело трафіка – інші сайти (%)	11,97	47,34	0	46,75	19,62
за період часу з 10.2010р. по 11.2011р.						
1	Відвідування	2033	1813928	186	913929	2423
2	Середній час відвідування сайту(хв:с)	7:03	3:08	6:52	2:58	2:56
3	Показник відмовлень (%)	46,34	62,45	48,68	62,23	48,68
4	Досягнута мета	253	0	0	0	449
5	Динаміка (%)	8,97	39,74	8,65	23,18	17,72
6	Перегляд сторінок	12694	4249331	802	2149567	8423
7	Кількість сторінок за відвідування	6,24	2,34	4,31	2,12	3,48
8.	Нові відвідування (%)	55,53	35,34	22,04	25,65	65,37
9	Абсолютно унікальні відвідувачі	1152	671308	41	334536	1592
10	Джерело трафіка – пошукові системи (%)	49,09	42,77	64,52	40,75	49,94
11	Джерело трафіка – прямий трафік (%)	20,17	15,76	22,04	16,53	29,67
12	Джерело трафіка – інші сайти (%)	30,74	41,21	13,44	42,72	20,39

Основні результати реалізовані в Інтернет-проектах “Вголос” (Львів, vgolos.com.ua), “Victana” (Львів, victana.lviv.ua), “Татьяна” (Херсон, tatjana.in.ua), “Прес-Тайм” (Львів, presstime.com.ua), “AutoChip” (Вінниця, autochip.vn.ua), “Фотогалерея Висоцьких” (Львів, fotoghalereja-vysocjkykh.com), “Курси валют” (Львів, kursyvalyut.com), “Добрий ранок” (Львів, dobryjranok.com), “Інформація для бізнесу” (Львів, goodmorningua.com), “Львівська загальноосвітня школа № 3” (Львів, www.зсш3львів.in.ua) та в навчальному процесі Національного університету “Львівська політехніка” під час викладання окремих розділів таких дисциплін: “Технологія програмування та створення програмних продуктів”, “Математична лінгвістика”, “Технологія проектування програмних систем”, “Експертні системи та автоматизовані системи навчання”, “Системний аналіз та проектування комп’ютерних інформаційних технологій”, “Основи автоматизованого проектування складних об’єктів і систем”, “Інформаційні технології”, “Комп’ютерно-технологічні засоби інформаційної діяльності”.

На рис. 4–9 подано результати роботи розроблених систем з Google Analytics у вигляді графіків та діаграм, з яких видно, що за наявності всіх етапів життєвого циклу контенту на сайті суттєво збільшується обсяг відвідувань та унікальних користувачів.

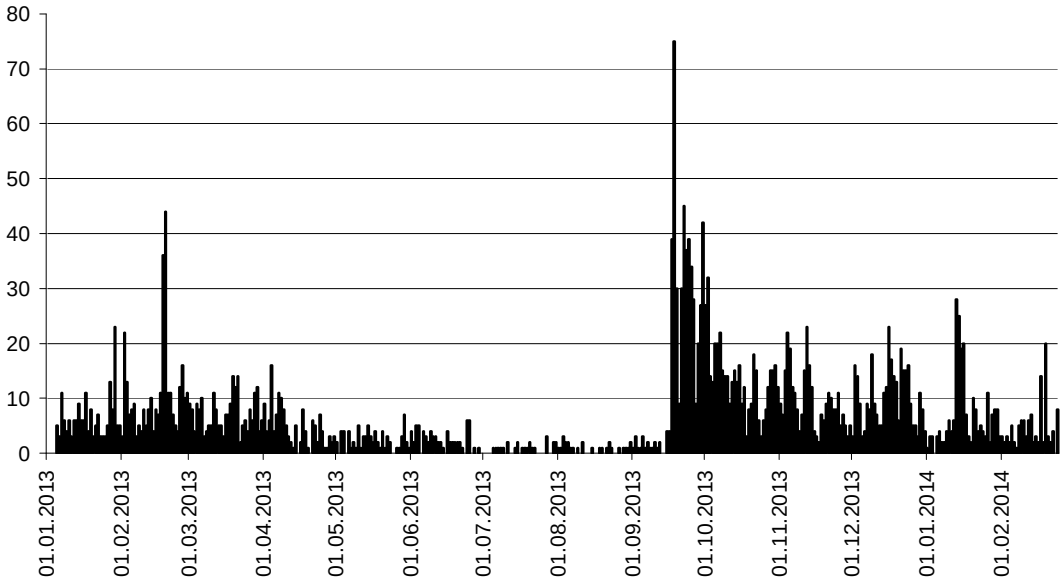

Рис. 4. Розподіл щоденного відвідування інформаційного ресурсу "Victana" (Львів, victana.lviv.ua)

На рис. 4 подано розподіл відвідування інформаційного ресурсу "Victana" (Львів, victana.lviv.ua) за період часу від 01.012013 до 01.01.2014 рр. Спочатку функціонування інформаційного ресурсу "Victana" був активізований лише модуль управління комерційним контентом. Внаслідок специфічності контенту (призначений для студентів) з 01.05.2013 р. до 01.09.2013 р. був спад відвідуваності інформаційного ресурсу. Наприкінці вересня 2013 р. в системі електронної контент-комерції "Victana" було активізовано модулі формування та супроводу комерційного контенту, тому відвідуваність інформаційного ресурсу збільшилась майже вдвічі.

На рис. 5 подано розподіл щоденного відвідування інформаційного ресурсу "Фотогалерея Висоцьких" (Львів, fotoghalereja-vysocjkykh.com) за період часу 01.012013-01.01.2014 рр. Від початку функціонування інформаційного ресурсу "Фотогалерея Висоцьких" було активізовано модулі формування, управління та супроводу комерційного контенту. З 01.06.2012 р. було відімкнено модуль супроводу комерційного контенту – відвідуваність зменшилась майже вдвічі.

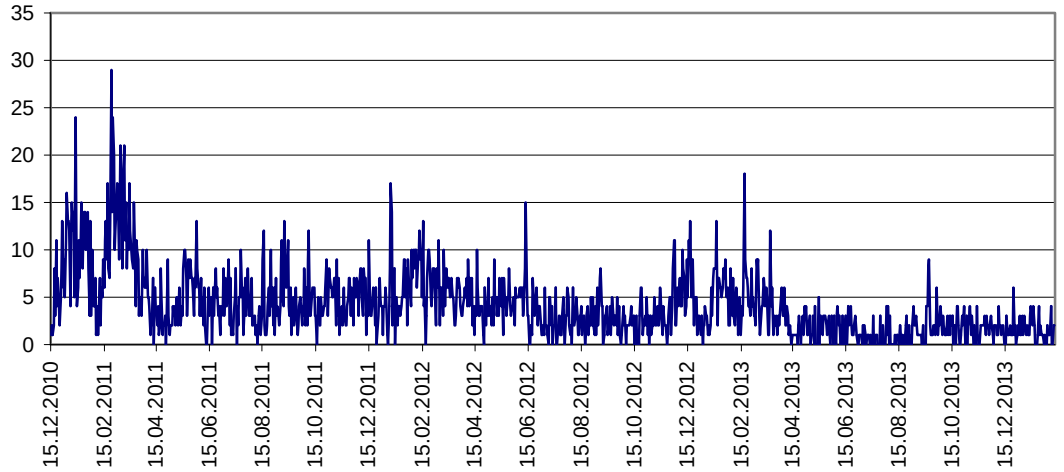

Рис. 5. Розподіл щоденного відвідування інформаційного ресурсу "Фотогалерея Висоцьких"

На рис. 6 подано розподіл щотижневого відвідування інформаційного ресурсу “Фотогалерея Висоцьких” (Львів, fotoghalereja-vysocjkykh.com) за період часу від 01.012013 до 01.01.2014 рр. Від 01.12.2012 р. був знову активізований модуль супроводу комерційного контенту. Відвідуваність збільшилась майже вдвічі. Від 01.04.2013 р. було відімкнено модулі формування, управління та супроводу комерційного контенту. Відвідуваність зменшилась майже втричі.

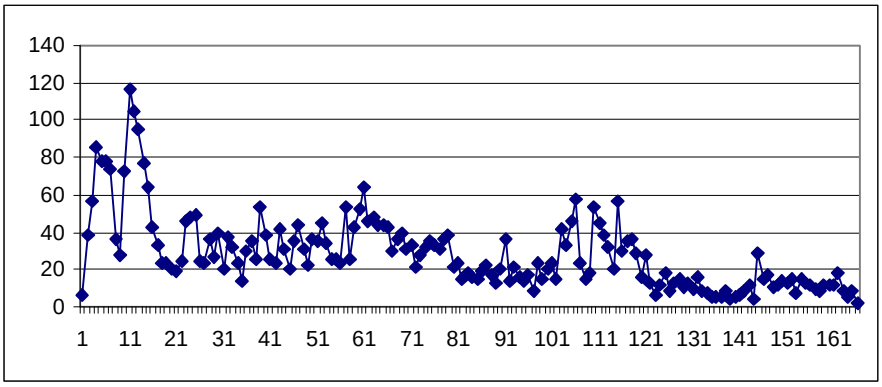

Рис. 6. Розподіл щотижневого відвідування інформаційного ресурсу “Фотогалерея Висоцьких”

На рис. 7 подано для порівняння розподіли щомісячного відвідування інформаційного ресурсу “Фотогалерея Висоцьких” (Львів, fotoghalereja-vysocjkykh.com), щомісячний пошуковий трафік комерційного контенту та щомісячні відвідування постійних користувачів, що повернулися на цей інформаційний ресурс за період часу від 01.012013 до 01.01.2014 рр. Як видно з діаграми на рис. 7, коливання розподілів пропорційне до періодів активацій та відмикання відповідних модулів формування, управління та супроводу комерційного контенту.

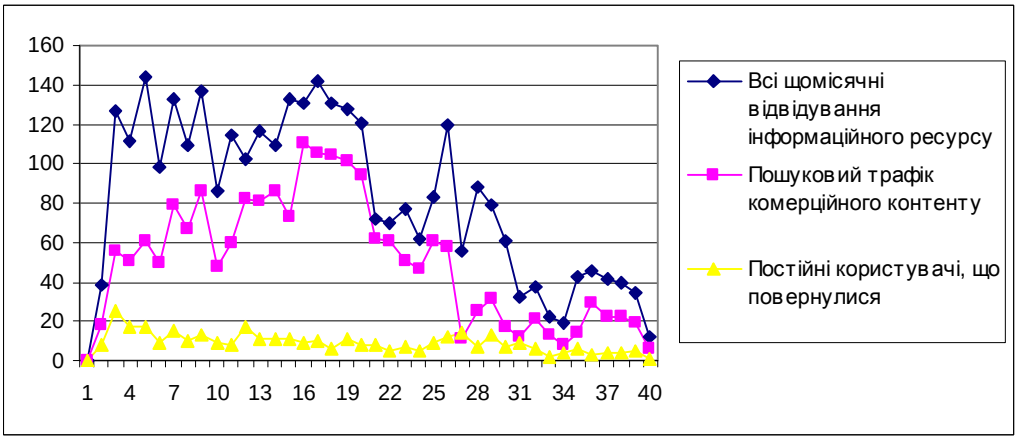

Рис. 7. Порівняння розподілів щомісячного відвідування інформаційного ресурсу “Фотогалерея Висоцьких”

На рис. 8 подано розподіл відвідування інформаційного ресурсу “Victana” (Львів, victana.lviv.ua) за період часу від 01.012013 до 01.01.2014 рр. порівняно з розподілами щоденного відвідування унікальних відвідувачів та відвідувачів, які замовили комерційний контенту на інформаційному ресурсі. Підвищення ефективності систем електронної контент-комерції внаслідок збільшення обсягів реалізації комерційного контенту прямо пропорційно залежить від коливання розподілу відвідування інформаційного ресурсу та пропорційне до періодів активації та відімкнення відповідних модулів формування, управління та супроводу комерційного контенту.

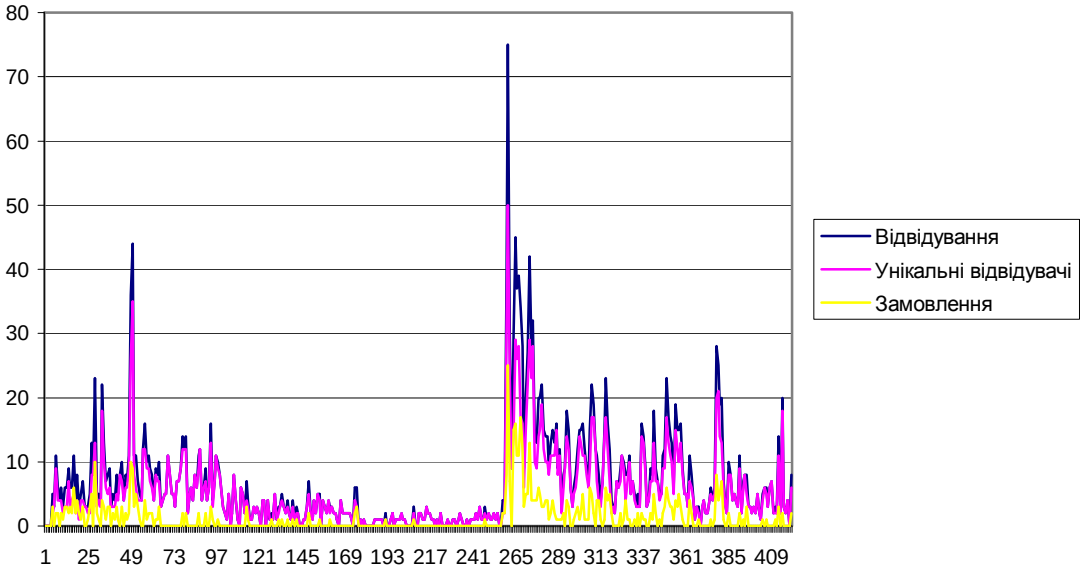

Рис. 8. Порівняння розподілів щомісячного відвідування інформаційного ресурсу "Victana" (Львів, victana.lviv.ua)

На рис. 9 подано розподіл відвідування інформаційного ресурсу "Курси валют" (Львів, kursyvalyut.com) порівняно із постійними користувачами інформаційного ресурсу за період часу від 01.012013 до 01.02.2014 рр.

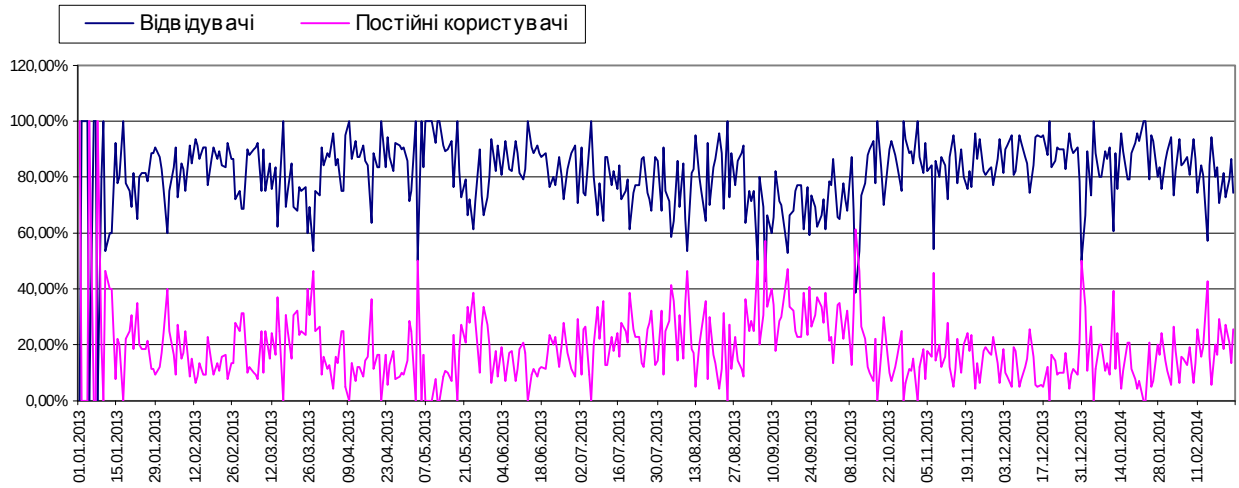

Рис. 9. Порівняння розподілів відвідування інформаційного ресурсу "Курси валют" (Львів, kursyvalyut.com)

5. Висновки і перспективи подальших наукових розвідок

У роботі розв'язано актуальну наукову задачу дослідження і розроблення методів та засобів опрацювання інформаційних ресурсів СЕКК із використанням розробленої класифікації, математичного та програмного забезпечення та узагальненої архітектури СЕКК. Проаналізовано термінологію та класифіковано системи електронної контент-комерції для визначення їхніх характерних закономірностей, тенденцій, процесів проектування та моделювання. Розроблено формальні моделі СЕКК для визначення недоліків існуючих методів та засобів опрацювання ресурсів. Розроблено уніфіковані методи опрацювання інформаційних ресурсів СЕКК. Розроблено архітектури модулів СЕКК для реалізації етапів життєвого циклу комерційного контенту.

Розроблено програмні засоби опрацювання інформаційних ресурсів систем електронної контент-комерції. З позиції системного підходу застосовано принципи опрацювання інформаційних ресурсів у СЕКК, що дало змогу розробити методи формування, управління та реалізації комерційного контенту. Реалізовано комплексний метод супроводу контенту, що дає можливість розробити модуль реалізації комерційного контенту.

1. Берко А. Системи електронної контент-комерції / А. Берко, В. Висоцька, В. Пасічник. – Л.: Вид-во Львівської політехніки, 2009. – 612 с. 2. Большакова Е. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Е. Большакова, Д. Ландэ, А. Носков, Э. Клышинский, О. Пескова, Е. Ягунова. – М: МИЭМ, 2011. –272 с. 3. Брайчевский С. Современные информационные потоки / С. Брайчевский, Д. Ландэ // Научно-техническая информация. – 2005. – № 11. – С. 21–33. 4. Клифтон Б. Google Analytics: профессиональный анализ посещаемости веб-сайтов / Б. Клифтон. – М: Вильямс, 2009. – 400 с. 5. Корнеев В. Базы данных. Интеллектуальная обработка информации / В. Корнеев, А. Гареев, С. Васютин, В. Райх. – М: Нолидж, 2000. – 352 с. 6. Ландэ Д. Основы моделирования и оценки електронных информационных потоков / Д. Ландэ, В. Фурашев, С. Брайчевский, О. Григорьев. – К: Інжиніринг, 2006. – 348 с. 7. Ландэ Д. Основы интеграции информационных потоков: монография / Д. Ландэ. – К: Інжиніринг, 2006. – 240 с. 8. Пасічник В. Математична лінгвістика / В. Висоцька, В. Пасічник, Ю. Щербина, Т. Шестакевич. – Л: "Новий Світ – 2000", 2012. – 359 с. 9. Советов Б. Моделирование систем / Б. Советов, С. Яковлев. – М: ВШ, 1998. 10. Федорчук А. Контент-мониторинг информационных потоков / А. Федорчук. – К., 2005. – № 3.

УДК 004.43

О.В. Годич, Ю.О. Прокопів
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ВІЗУАЛЬНА ПРЕДМЕТНО-ОРІЄНТОВАНА МОВА ЗАПИТІВ

Подано результати створення візуальної предметно-орієнтованої мови запитів, яка є візуальним засобом взаємодії з даними і побудови правил моніторингу стану інформаційних систем. Використання розробленої мови надає експертам предметної області потрібні засоби для розширення системи ad-hoc у процесі її використання з метою вчасного реагування на виробничі потреби.

Ключові слова: предметно-орієнтована мова, інформаційні запити, проекційне редагування, HCI.

This article discusses a visual domain-specific query language that supports data interaction and composition of business rules as part of a software system, which can be used directly by domain experts as part of the information system itself. Such approach provides domain experts with necessary tools to enhance the live system in order to meet dynamically changing real-life requirements without the tedious and often complex development/deployment cycles currently used in the software industry.

Key words: domain-specific language, data querying, projectional editing, HCI.

Вступ. Постановка проблеми.

Мова є невід'ємною характеристикою людини, і можливості мов (природних і формальних) є визначальними щодо здатності мислити й успішно вирішувати складні проблеми. Наприклад,