УДК 004.9

В. Фольтович[1], М. Коробчинський[2], Л. Чирун[1], В. Висоцька[1]
[1]Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж,
[2]Воєнно-дипломатична академія імені Євгенія Березняка

МЕТОД КОНТЕНТ-АНАЛІЗУ ТЕКСТОВОЇ ІНФОРМАЦІЇ ІНТЕРНЕТ-ГАЗЕТИ

© Фольтович В., Коробчинський М., Чирун Л., Висоцька В., 2017

Зроблено огляд концепції аналізу, видів аналізу, класифікації, проаналізовано методи здійснення та описано сучасну систему аналізу змісту. Для аналізу змісту сайту використано такі методи: аналіз змісту відвідувачів, аналіз необхідної кількості контенту, графічний аналіз контенту. Проаналізовано методи і засоби проектування, використано методи і технології графіків, що відображають логічну й фізичну модель системи. Визначено взаємодію об'єктів у системі, упорядкованих за часом їх виконання. Визначено основні функції системи, мету розвитку системи і використання цієї системи в майбутньому. Розроблено концептуальну модель системи, визначено вхідні, вихідні дані; описано їх вимоги до системи. Інтелектуальна інформаційна система аналізу змісту текстової інформації в електронному видавництві є потужним інструментом в управлінні та успішній діяльності будь-якої газети.

Ключові слова: аналіз, контент-аналіз змісту інформаційних ресурсів, контент-аналіз, рейтингова оцінка, система управління контентом.

Were made a review of the concept of analysis, types of analysis, classification of analysis, methods of its implementation and made a review of modern system of analysis of content. To analyze site content were used the following methods: analysis of the visitor content, analysis of the required amount of content, graphical analysis of content. Were analyzed the methods and design tools. Used methods and technologies of charts that display logical and physical model of the system. Determined the interaction of objects in the system that are ordered by time of their execution. All these charts reflect the structure and functioning of the system. Determined the basic functions of the system, the purpose of system development and usage of this systems in the future. Developed the conceptual model of the system, defined incoming, outgoing data and conducted their description and the description of system requirements. Thus, this intellectual – information system of content – analysis of textual information at the e-publishing house is a powerful tool in the management and successful operation of any newspaper.

Key words: analysis, content analysis content, information resource, content – analysis, rating evaluation, content management system.

Вступ. Загальна постановка проблеми

Останнім часом людство здійснило значний крок у розробленні та впровадженні новітніх інформаційних технологій (ІТ) [1–30]. З розвитком ІТ вирішено багато складних завдань, але й позначилися нові, одним з яких є контент-аналіз текстової інформації. Це кількісно-якісний метод аналізу масивів тексту для подальшої змістовної інтерпретації отриманих кількісно-якісних показників [1–30]. Розвиток інтернет-технологій та його служб надав доступ людству до практично необмеженої кількості контенту, але виникла проблема його достовірності та оперативності [1–30]. Важливою характеристикою контенту є його адекватність, тобто певний рівень відповідності створюваного за допомогою отриманої інформації образу реального об'єкта, процесу чи явища [1–30]. Для оперативності, достовірності та адекватності контенту впроваджують ІТ контент-аналізу.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Це дає змогу отримувати інформаційній системі (ІС) контент за всіма видами її діяльності. Результат контент-аналізу тексту використовують при визначенні його тональності, дублювання, наявності спаму та виявлення нових подій для визначення тематичних сюжетів його потоків. За отриманою інформацією можна оперативно втручатися в діяльність ІС для підвищення рівня її функціональності та популярності серед користувачів [31–37].

Аналіз останніх досліджень та публікацій

ІС контент-аналізу є доволі успішними і не потребують великих витрат і часу на отримання потрібного результату [1–30]. Водночас з використанням цього типу ІС підвищується рівень успішності продукту на 30 % [1–30]. Для ефективності функціонування інтернет-газети необхідно постійно відслідковувати контент про рівень статей, діяльність її користувачів тощо. З використанням таких ІС аналізують контент та здійснюють технічний аналіз сайту. Для аналізу контенту сайту існують такі методи: аналіз контенту з погляду відвідувача, аналіз кількості необхідного контенту, графічний аналіз контенту тощо. Для технічного аналізу існують такі методи: технічний аналіз доступності сайту, технічний аналіз способів управління інформаційним наповненням сайту тощо [1–30].

Вважливим та перспективним контент-аналіз є в управлінні інформаційними ресурсами [1–30]. Без відповідної актуальної інформації зробити це важко. Управління є важливою частиною будь-якого виду аналізу і проводиться на основі поточних даних, а також динаміки змін попередніх даних. Базова система аналізу контенту передбачає такі можливості: швидке поновлення контенту, пошук контенту на певному ресурсі, збирання контенту про постійних та потенційних клієнтів, аналіз цільової та постійної аудиторії, формування та редагування опитувань, аналіз відвідування ресурсу. У цьому напрямку працюють такі провідні світові виробники засобів опрацювання інформаційних ресурсів, як Apple, Google, Intel, Microsoft, Amazon тощо [9]. У разі автоматизації інтернет-газети за допомогою ІС аналізу контенту зменшуються обсяги роботи, час на опрацювання та отримання необхідної інформації, зростає продуктивність роботи системи, що зменшує витрати і час на отримання потрібного результату. Відомим методом аналізу текстової інформації є контент-аналіз – стандартна методика дослідження в галузі суспільних наук (рис. 1) [1–30], предметом якої є аналіз змісту текстових масивів і комунікативної кореспонденції (коментарів, форумів, електронного листування, статей тощо).

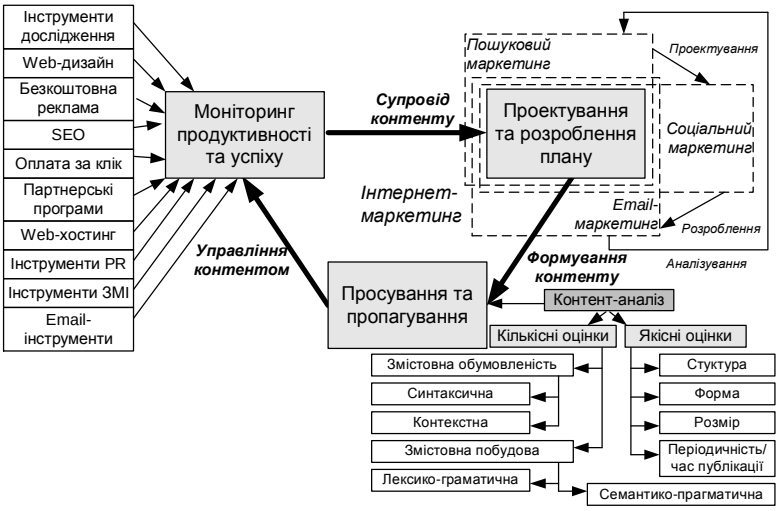


Рис. 1. Інтернет-маркетинг для інтернет-систем

Поняття контент-аналізу не має однозначного визначення, тому ІС, побудовані на основі різних підходів, є несумісними [1–30]. Не існує єдиного та однозначного універсального алгоритму контент-аналізу текстових масивів даних [1–30]. Фахівці з розроблення ІС контент-аналізу текстової інформації, наприклад, для засобів масової інформації або електронної комерції, мусять

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

самі, аналізуючи відомий метод та досліджувану предметну область, адаптувати загальновідомі етапи методу контент-аналізу відповідно до конкретної задачі [1–30]. Актуальність створення ІС контент-аналізу текстової інформації інтернет-газети спричинена зростанням вимог її користувачів та зумовлена такими чинниками: швидкі темпи зростання потреб у достовірному та адекватному контенті, необхідність формування множини оперативного контенту, а також автоматичного фільтрування інформаційного шуму (небажаного контенту та спаму) [31–37].

Виділення проблем

З метою створення ефективної ІСОІР для контент-аналізу розглянемо наявні програмні продукти, системи та сервіси. Сьогодні існує множина засобів та сервісів для контент-аналізу сайтів, серед яких можна виділити такі: google analytics, webmasta.org, cy-pr.com, seogift.ru.

Google Analytics є найпоширенішим, сьогодні його популярність сягає близько 55 % (рис. 2). Сервіс призначений для визначення статистики активності користувачів сайту. Визначає, як часто люди відвідують кожну сторінку, скільки часу проводять на ній і скільки конверсій вона приносить. За допомогою Google Analytics можна дізнатися, що зазвичай шукають користувачі на вашому веб-ресурсі, побудувати маршрут користувачів веб-ресурсу, перевірити швидкість завантаження розділів сайту за допомогою відповідного звіту Google Analytics. Виправте ці проблеми, і сайт буде більш зручним для користувачів. Проаналізуйте, як часто користувачі натискають на Flash- і AJAX-елементи або на посилання для завантаження описів [9].

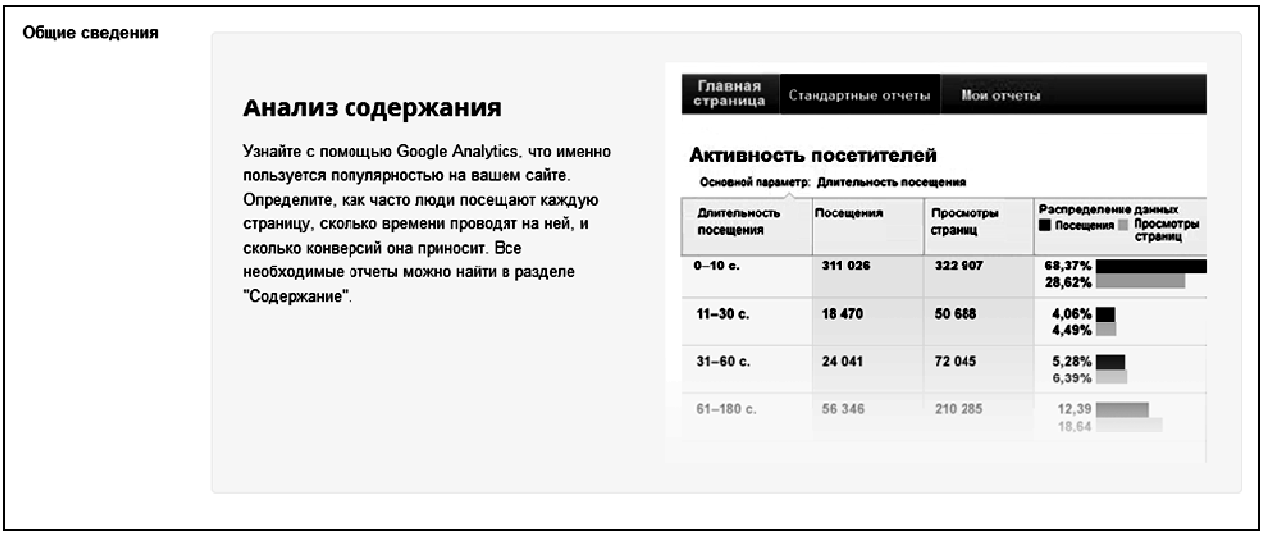


Рис. 2. Сервіс "Google Analytics"

Сервіс "webmasta.org" надає множину інструментів для проведення аналізу вмісту сайту (рис. 3). Аналіз контенту сайту, проведений цим сервісом, дає можливість перевірити повноту ключових слів, а також отримати інформацію, необхідну для ефективної оптимізації текстів. Цей ресурс має інструменти для опрацювання текстів, аналізу тексту, визначення дублікатів у тексті, порівняння текстів. Особливістю цього сервісу є наявність інструментів для аналізу та аудиту сайту, наявність багатьох генераторів. Перевагами сервісу є: швидкість проведення аналізу, доброзичливий інтерфейс, велика кількість інструментів для проведення різних типів аналізу.

Сервіс "cy-pr.com" є інструментом аналізу контенту сайту, що дає змогу якісно оцінити текст, проаналізувати вміст сторінки з погляду сприйняття пошуковими системами (рис. 4). Оцінивши контент, сервіс виявляє недоліки, не помітні на перший погляд, але які значно ускладнюють просування в пошукових системах. За допомогою цього інструменту аналізують: релевантність заголовків сторінки, відсоток входжень в тексті, наявність стоп-слів, щільність слів на сторінці. Перевірка контенту сайту цим сервісом допоможе уникнути помилок і досягти ідеальної оптимізації контенту сайтів під просувний запит. Цей вид аналізу розрахований на професіоналів, він дає вичерпну і дуже точну інформацію, яка допоможе усунути всі недоліки.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Рис. 3. Сервіс "webmasta.org"

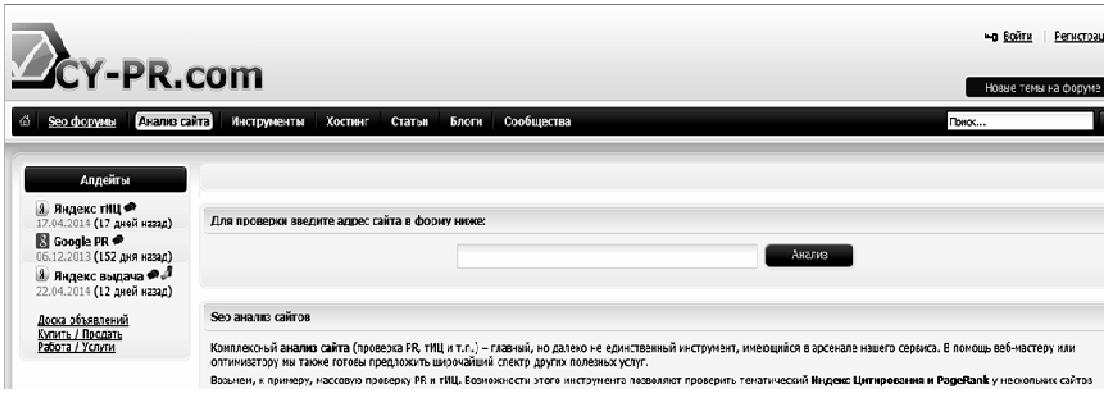


Рис. 4. Сервіс "cy-pr.com"

Сервіс "seogift.ru" використовують для аналізу контенту сайту (рис. 5), який є корисним інструментом для оптимізаторів і веб-майстрів з перевірки щільності ключових слів на сторінці сайту. Сьогодні пошукові системи приділяють пильну увагу контенту сайту. Незамінною деталлю цього сервісу є відображення результату дослідження повного списку всіх слів, що зустрічаються на сторінці сайту, з кількістю їх повторень та процентним співвідношенням. Переваги сервісу: всі інструменти безкоштовні, висока швидкість опрацювання даних, доброзичливий інтерфейс і зручність його використання в роботі, зручна панель для перевірки сайтів, поновлення даних для зареєстрованих користувачів, актуальна інформація про нові технології в пошукових системах.

Рис. 5. Сервіс "seogift.ru"

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Ці ресурси надають велику кількість інструментів для аналізу контенту, здійснюють статистичне оцінювання, дослідження історії, побудову звичайних та спеціальних звітів. Основним недоліком цих програмних продуктів і ресурсів є те, що вони надають множину значень про контент і вміст сайту, але не надають інструментів для відстеження поведінки окремих відвідувачів сайту (таблиця). Цей клас систем зараз активно розвивається і має високу популярність. Найголовнішим у внутрішній оптимізації даних системи є її текстовий вміст, тому контент-аналіз системи подається центральним заходом в ряді всіх дій. У добре збалансованому тексті кожне ключове слово, що зустрічається на сторінці, посилює оперативність і ліквідність всього ресурсу.

Порівняльна характеристика досліджених ресурсів

Назва	Переваги ресурсів	Недоліки ресурсів
Google Analytics	Зручний інтерфейс; велика кількість інструментів для аналізу; статистичне оцінювання відвідування; аналіз трафіку; дослідження історії; велика кількість звичайних і спеціальних звітів	Не надає інструментів для можливості відстежувати поведінку окремих відвідувачів сайту
cy-pr.com	Проведення контент-аналізу ресурсу; визначення щільності слів; визначення відсотку входження слів та кількості однакових слів	Немає спеціальних звітів та інструментів для модифікації контенту, не досліджує історію
webmasta.org	Проведення контент-аналізу ресурсу; перевірка релевантності заголовків; зрозуміла і проста схема звітності	Немає інструментів для модифікації контенту, спеціальних звітів, не досліджує історію
seogift.ru	Всі інструменти безкоштовні; висока швидкість обробки даних; перевірка актуальності інформації; вдалий інтерфейс	Не аналізує трафік та не досліджує історію, не надає інструментів для можливості відстежувати поведінку окремих відвідувачів сайту

Формулювання мети

Метою є аналіз методів контент-аналізу та розроблення ICOIP, яка б здійснювала контент-аналіз текстової інформації інтернет-газети, визначала рейтинг статей, дублювання контенту, здійснювала спостереження за користувачами та оцінювала результати після аналізу. Основні дослідження проводили за допомогою методів контент-аналізу текстової інформації. В результаті освоєно та удосконалено методи контент-аналізу для опрацювання текстових масивів даних, розвинено методи для визначення рейтингу статей. Отримані дані та результати роботи методів враховують під час оновлення інформаційного ресурсу та удосконалення архітектури ICOIP. Практична цінність отриманих результатів полягає у розробленні та апробації нових методів та засобів для систем контент-аналізу текстової інформації, заощадження людських та грошових ресурсів і в побудові зручних для кінцевих користувачів програмних засобів і систем.

Аналіз отриманих наукових результатів

Рейтингове оцінювання в системі проводять двома способами: першим способом проводиться рейтингове оцінювання статей за допомогою трьох критеріїв. Першим критерієм буде кількість звернень або кількість читання окремої статті, який відображатиме кількість відкривань або завантажень статті. Наступним критерієм буде користувацька оцінка – це критерій, який визначає користувач. Загалом це деяка оцінка, яку може поставити користувач певній статті. Наступним критерієм є час читання статті – це критерій, який відображає, скільки часу користувачі витратили на читання статті. При автоматизації за допомогою інформаційної системи аналізу контенту зменшуються обсяги роботи, час на опрацювання та отримання необхідної інформації, зростає продуктивність роботи системи, що, своєю чергою, зменшує затрати коштів і часу на отримання потрібного результату. ICOIP контент-аналізу текстової інформації інтернет-газети повинна виконувати такі завдання:

1. Облік користувачів системи: збереження даних про користувачів системи.
2. Облік контенту системи: збереження даних та статей, які є контентом.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

3. Контент-аналіз текстової інформації Інтернет-газети: аналіз контенту на основі рейтингового оцінювання статей та аналізу статистики.

4. Формування даних, які відображають результати аналізу.

Оскільки для функціонування системи є дуже важливим користувацький аспект, потрібно розробити користувацький інтерфейс, який має бути інтуїтивним і зручним у користуванні, а також має надавати швидкий доступ до основних функцій системи, які призначені для користувача. З цього погляду інтерфейс користувача – це частина інформаційної системи, що взаємодіє з користувачем. Користувач здійснює авторизацію за допомогою об'єкта авторизації.

Після введення статті в систему дані надходять для методу контент-аналізу текстової інформації. Після аналізу результати цієї операції записують до бази даних, після чого адміністратор може проаналізувати дані і виконати одну з дій.

Після того, як користувач здійснив авторизацію, він звертається до статті і за допомогою функції виконує читання статей, отримуючи інформацію у вигляді результату читання статті. Після читання статей дані для оцінювання за допомогою однієї з функцій пересилаються.

Після закінчення операції функції рейтингового оцінювання оновлені дані про популярність статей заносяться до бази даних. Функцію "аналіз статистики" використовують для статистичного оцінювання, для чого цій функції надсилаються дані з бази даних. Адміністратор – єдиний актор у системі, який може додавати або видаляти статті. Після додання нових статей або видалення функції виведення статей передаються дані, які потрібні для виведення останніх та популярних статей. Також цією функцією перевіряють наявність нових статей. Цю систему можна використовувати як інформаційний ресурс, а також аналізувати контент, що допомагає якісніше розміщувати інформацію, яка буде оперативною для користувачів системи. Це підвищує рейтинг інформаційного ресурсу, а отже, збільшує кількість користувачів, що вплине на економічний ефект. Після системного аналізу та обґрунтування проблеми визначено проблеми, притаманні для цієї предметної області, проаналізовано предметну область, а також окреслено та обґрунтовано проблему. Також побудовано дерево цілей, визначено цілі та їх підцілі. Побудовано ієрархію за допомогою методу аналізу ієрархій. Визначено завдання, які має вирішувати інтелектуально-інформаційна система контент-аналізу текстової інформації.

Проаналізовано методи та засоби проектування. За допомогою методів і технологій побудовано діаграми, які відображають логічну та фізичну моделі функціонування системи. Визначено об'єкти та варіанти використанні, наявні в системі. Визначено взаємодію об'єктів, впорядковану за часом їх виконання. Всі ці діаграми повністю відображають структуру та функціонування системи. Визначено основні функції системи, мету розроблення системи та місце застосування системи. Розроблено концептуальну модель системи, визначено вхідні, вихідні дані та описано їх та вимоги до системи.

Відносною мірою кількості семантичної інформації може слугувати коефіцієнт змістовності Z, який визначається як відношення кількості семантичної інформації до її обсягу: $Z = \frac{I_c}{V_д}$ [3].

Отже, обсяг даних $V_д$ у відповідному повідомленні вимірюють за допомогою кількості символів у цьому повідомленні. Тому отримання інформації про яку-небудь систему завжди пов'язане зі зміною ступеня непоінформованості одержувача про стан цієї системи, тобто кількість інформації вимірюється як зміна стану системи. Нижче детальніше розглянемо існуючі кількісні міри інформації. Коефіцієнт інформативності повідомлення визначають за допомогою відношення кількості інформації до відповідного обсягу даних у повідомленні: $Y = \frac{I}{V_д}$ [3].

Із збільшенням коефіцієнта інформативності зменшується й обсяг роботи із перетворення інформації. Тому природним є прагнення до підвищення коефіцієнта інформативності. Кількісно інформацію можна виміряти, враховуючи два початкові поняття: ймовірність випадкової лінгвістичної події і невизначеність, яка є перед виконанням експерименту, результатом якого є вказана подія.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Кожен експеримент завжди має деяку невизначеність результату.

Ентропію визначають за формулою $H = \log S$ [3]. Одиницею виміру ентропії є невизначеність, яку містить дослід з двома рівноймовірними результатами. Це двійкова одиниця або біт. Введення поняття ентропії дає можливість кількісно вимірювати інформацію. Тому в результаті експерименту ми отримаємо нові відомості, тобто деяку нову інформацію. Тому знання того, який повинен бути результат експерименту, зменшує невизначеність. Правильно припустити, що знята в результаті досліду ентропія дорівнює кількості одержаної інформації, тобто $I_0 = \log_2 S$ [3]. Під час опису комбінаторного методу для обчислення кількості інформації та ентропії ми використовували спрощення, за яким всі закінчення досліду вважались рівноймовірними. В реальних дослідженнях такої ситуації практично ніколи не трапляється. Якщо випробування передбачає нерівноймовірні результати, то, очевидно, ентропія такого досліду і отримана від нього кількість інформації відрізнятимуться від аналогічних величин для досліду з рівноймовірними результатами. Для нерівноймовірних результатів ентропія на символ алфавіту $H = \sum_{i=1}^{m} p_i \cdot \log_2 \frac{1}{p_i} = -\sum_{i=1}^{m} p_i \cdot \log_2 p_i$, а кількість інформації в повідомленні, що скадається з k нерівноймовірних символів: $I = -k\sum_{i=1}^{m} p_i \cdot \log_2 p_i$ [3].

Статистичний аналіз проводили в три етапи:

І етап: за допомогою методу масових спостережень збирають первинні статистичні дані. Основний зміст цього етапу полягає в отриманні даних, що характеризують кожну одиницю спостереження;

ІІ етап статистичного аналізу: зібрані дані піддаються первинній обробці, зведенню і групуванню. Метод угруповань дає змогу виділити однорідні сукупності, поділити їх на групи і підгрупи. Результат – це отримання підсумків за сукупністю загалом й за окремими її групами та підгрупами;

ІІІ етап: отримані зведені дані аналізують методом узагальнювальних показників. Основний зміст цього етапу полягає у виявленні взаємозв'язків явищ, визначенні закономірностей їх розвитку та прогнозному оцінюванні [7].

Переваги використання контент-аналізу:

- висока точність аналізу;
- визначення змісту понять "об'єкт" і "ознака" у межах їх частотного висловлювання;
- побудова логіки взаємодії об'єкта і його ознак;
- формування частоти розподілу ознак в об'єктах;
- здійснення формалізованого статистичного аналізу структур тексту.

Дерева рішень є одним з найпопулярніших методів вирішення завдань класифікації та прогнозування. Дерева рішень дають змогу візуально й аналітично оцінити результати вибору різних рішень і використовуються в галузі статистики та аналізу даних для прогнозних моделей. Вперше метод запропонували Ховіленд і Хант (Hoveland, Hunt) наприкінці 50-х років минулого століття [8]. Нехай є множина A з n елементів, m з яких мають деяку властивість S. Тоді ентропія множини A відносно властивості S – це: $H(A,S) = -\frac{m}{n}\log_2\frac{m}{n} - \frac{n-m}{n}\log_2\frac{n-m}{n}$ [8]. Отже, ентропія залежить від пропорції, за якою поділяється множина. У міру зростання цієї пропорції від 0 до 1/2 ентропія теж зростає, а після 1/2 – симетрично спадає. Якщо властивість S не бінарна, а може набувати s різних значень, кожне з яких реалізується в m_i випадках, то ентропія узагальнюється природним чином: $H(A,S) = -\sum_{i=1}^{s}\frac{m_i}{n}\log\frac{m_i}{n}$ [8].

Поняття ентропії тісно пов'язане з теорією інформації. Спрощено, ентропія – це середня кількість бітів, які потрібні, щоб закодувати атрибут S як елемент множини A. Якщо ймовірність

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

появи *S* дорівнює 1/2, то ентропія дорівнює 1, і потрібен повноцінний біт; а якщо *S* з'являється не рівноймовірно, то можна закодувати послідовність елементів *A* ефективніше.

Все це наводить на думку про те, що вибирати атрибут для класифікації треба так, щоб після класифікації ентропія стала якомога меншою (властивість *S* у цьому випадку – значення цільової булевої функції). Ентропія при цьому буде різною в різних нащадках, і загальну суму треба визначати, враховуючи те, скільки результатів залишилося у розгляді за кожним з нащадків. Загальноприйняте в теорії дерев прийняття рішень означення виглядатиме так.

Нехай множина *A* елементів, деякі з яких мають властивість *S*, класифікується за допомогою атрибута *Q*, що має *q* можливих значень. Тоді приріст інформації (*Information gain*) визначається як:

$$Gain(A,Q) = H(A,S) - \sum_{i=1}^{q} \frac{|A_i|}{|A|} H(A_i,S),$$

де A_i – множина елементів *A*, на яких атрибут *Q* набуває значення *i* [8].

На кожному кроці алгоритм повинен вибирати той атрибут, для якого приріст інформації максимальний. Одне з питань, який виникає в алгоритмі дерева рішень – це оптимальний розмір кінцевого дерева. Так, невелике дерево може не охопити тієї чи іншої важливої інформації щодо вибіркового простору. Тим не менше, важко сказати, коли алгоритм повинен зупинитися, тому що неможливо спрогнозувати, додавання якого вузла дасть змогу значно зменшити помилку [8].

Сфери застосування: медицина (діагностика захворювань), молекулярна біологія (аналіз будови сполучень), банківська справа (оцінювання кредитоспроможності), промисловість (контроль за якістю продукції, виявлення дефектів), випробування (якість зварювання).

Переваги використання дерев рішень: швидкий процес навчання; інтуїтивно зрозуміла класифікаційна модель; побудова непараметричних моделей; простота побудови; дерева рішень дають можливість формувати правила з бази даних природною мовою.

Першим етапом реалізації системи є проектування бази даних. Так, найважливішим відношенням у нашій системі буде відношення *Statti* – це інформація про статті, їх вміст тощо. Наступним відношенням буде відношення *User* – це інформація про користувача як фізичну особу, та відношення *Category* – це відношення, яке містить інформацію про категорії статей.

Рис. 6. Головна сторінка сайту

Рис. 7. Виведення статей за категоріями

Відношення *Statti* містить такі атрибути, як *id* – атрибут, в якому зберігається ідентифікатор статті; первинним ключем є тип *int*, *title* – атрибут, в якому записується заголовок статті тип *varchar* (255); *discription* – атрибут, в якому записується короткий текст статті тип *text*; атрибут *text* – значення, в якому зберігається повний текст статті тип text; date – атрибут який використовується для відображення дати написання статті тип date; img_src – атрибут, який використовується для збереження даних про шлях до зображення тип varchar (255), avtor – атрибут, в якому зберігаються дані про автора; тип varchar (255), reed – атрибут, в якому зберігаються дані про кількість звертань

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

до статті; тип unsigned int, атрибут mark – зберігає значення про рейтинг статті; unsigned int, атрибут category – значення про те, до якої категорії належить стаття; тип unsigned int є зовнішнім ключем. Відношення User містить такі атрибути, як id-атрибут, в якому зберігається ідентифікатор користувача, є первинним ключем тип int, name – ім'я користувача тип varchar (255), surname – прізвище користувача тип varchar (255), date – атрибут, в якому зберігається дата народження; тип date; login – атрибут, в якому зберігається логін користувача; тип varchar (255), атрибут password для збереження паролю користувача; тип varchar (255) використовує функцію шифрування MD5. Наступні відношення в базі даних не використовуються для проведення операцій системою. Основним їх призначенням є використання для виведення статей за категоріями та пошуку статей за допомогою відповідного класу системи.

Під час розроблення ICOIP визначено методи та засоби реалізації завдання. Сайт, який відображає програмну частину системи, написано мовою програмування PHP за принципами об'єктно-орієнтованого програмування та засобів, які використовують для створення сайтів.

Абстрактний клас ACore є класом-предком для таких класів: main, category, reg, login, view. Також цей клас містить методи, які виконують певні дії: так, наприклад, метод construct під'єднує до бази даних. Метод get_header використовується для виведення головної частини сайту. Метод get_menu використовується для виведення меню сайту. Метод get_leftmenu виконує виведення категорій статей. Метод get_footer виконує виведення нижньої частини сайту. Ці методи, окрім методу construct, є закритими.

Клас main виконує виведення головної сторінки сайту. Завдяки тому, що він є нащадком абстрактного класу Acore, він наслідує його методи. Клас main має такі методи: get_content, get_galleria. Метод get_galleria виконує виведення галереї, яка розміщується між головною і верхньою частиною сайту. В галереї міститься певна множина зображень, які виводяться одне за одним через певний час. Цей метод призначений для покращення візуальної складової сайту. Метод get_content виконує виведення основної частини сайту, тобто наших статей. Метод використовує два способи виведення статей: це виведення за датою додавання та виведення за рейтинговим оцінюванням. Залежно від вибраних критеріїв метод виконує запит до бази даних, після чого виконує структуроване виведення статей, які зберігаються в базі даних. Клас category є нащадком класу ACore і містить методи, які виконують структуроване виведення статей за категоріями, які містяться в базі даних у відношенні category. Клас reg містить методи, які виконують реєстрування користувача в системі за допомогою відповідної форми, яку заповнює користувач. Форма реєстрації містить такі поля: ім'я користувача, прізвище користувача, дата народження, логін і поле "пароль". Після цього виконується операція реєстрування, коли дані перевіряють на унікальність. Якщо хоча б одне значення є унікальним, виконується реєстрація, якщо жодне значення не є унікальним, то виводиться відповідне повідомлення користувачу.

Клас login містить методи, які виконують авторизацію користувача в системі за допомогою відповідної форми, яку заповнює користувач. Форма авторизації містить такі поля: логін "поле", до якого користувач вписує логін, і поле "пароль", куди користувач вписує особистий пароль. Після цього виконується операція авторизації, за якою перевіряють дані в полі "пароль/логін" на відповідність у відношення user бази даних: якщо хоча б одне значення не збігається, виводиться відповідне повідомлення користувачу для зміни пароля/логіна і виконується повторна авторизація.

Клас view містить методи, які використовуються для виведення окремої статті, коли користувач натискає кнопку детальніше. Завдяки тому, що він є нащадком абстрактного класу Acore, він наслідує його методи. Клас view має метод get_content. Метод get_content виконує структуроване виведення статті, яку вибирає користувач для читання. Також при використанні методів цього класу збільшується поле кількості звернень до статті та поле "час читання", яке міститься у відношенні statti і відношенні danis і використовується для рейтингового оцінювання статей.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

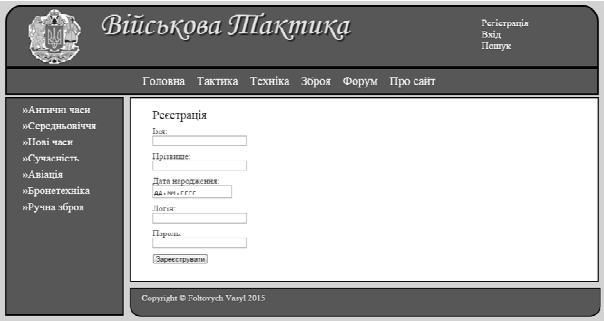


Рис. 8. Web – форма реєстрації користувача

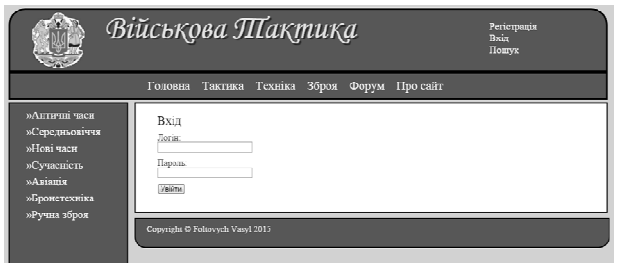


Рис. 9. Web – форма входу в систему

Рис. 10. Виведення окремої статті

Рис. 11. Статистика рейтингів статей

Клас admin використовується для адміністрування сайту. Містить методи, які виконують адміністрування сайту. Також для того, щоби адміністратор зміг здійснювати адміністрування, він повинен пройти авторизацію. Клас add_statti містить методи, які виконують додавання статті. Додавання статті виконується за допомогою відповідної форми, яка містить такі поля: тема статті, короткий опис, текст, зображення, автор, прочитано, оцінка, категорія. Поле “тема статті” використовується для введення теми статті. Поле “короткий опис” слугує для введення стислого опису статті. Поле “текст” слугує для введення тексту статті. Поле “зображення” слугує для збереження шляху до зображення, яке прикріплюється за кожною статтею. Поле “автор” використовується для введення даних про автора статті. Поля “прочитано” та “оцінка” використовують для рейтингового оцінювання. Поле “категорія” використовують для введення категорії статті.

Клас updete_statti містить методи, за якими редагують статтю за допомогою відповідної форми, яка містить такі поля: тема статті, короткий опис, текст, зображення, автор, прочитано, оцінка, категорія. Поле “тема статті” використовується для введення теми статті. Поле “короткий опис” слугує для введення стислого опису статті. Поле “текст” слугує для введення текст статті. Поле “зображення” слугує для збереження шляху до зображення, яке прикріплюється за кожною статтею. Поле “автор” використовується для введення автора статті. Поля “прочитано” та “оцінка” використовуються для рейтингового оцінювання. Поле “категорія” використовується для введення категорії статті.

Клас delete_statti містить методи, які виконують видалення статті з бази даних за допомогою відповідних запитів. Крім того, у системі містяться також відповідні класи для редагування, видалення та додавання нових категорій: add_category, updete_category, delete_category. А також містяться відповідні класи для додавання та інших операцій над пунктами меню.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

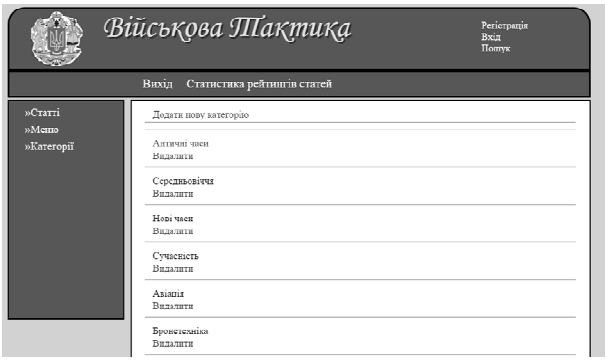


Рис. 12. Адміністрування сайту

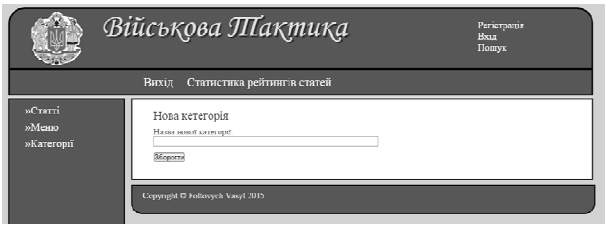


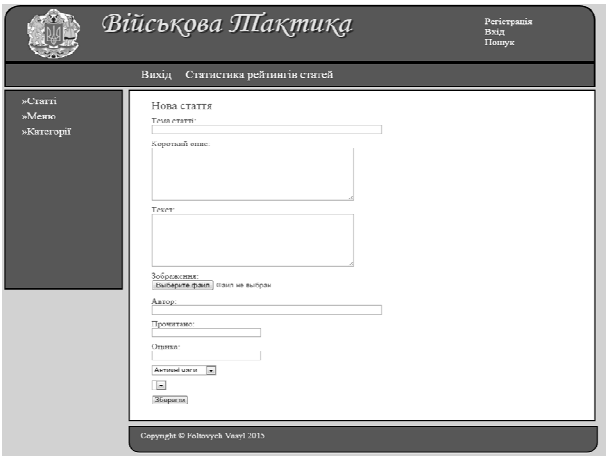
Рис. 13. Web-форма додавання нової категорії

Інтерфейс сайту не є завершеним і призначений для наповнення БД даними і тестування роботи системи. Вся інформація, наведена на скриншотах, є вигаданою, а всі збіги випадковими. Основним реалізованим методом аналізу контенту є метод, який ґрунтується на використанні рейтингового оцінювання статей за декількома критеріями. Критеріями методів аналізу контенту з використанням рейтингового оцінювання є кількість звернень, час читання та користувацька оцінка. Також реалізований аналіз статистики. На рис. 11 продемонстровано проведення аналізу рейтингів статей. Також адміністратор має змогу переглянути статистику відвідування окремої статті. За допомогою проведення аналізу визначається й оцінка статті на основі рейтингового оцінювання за допомогою критеріїв: час читання, кількість звернень, користувацька оцінка. Статистичний аналіз проходить в три етапи:

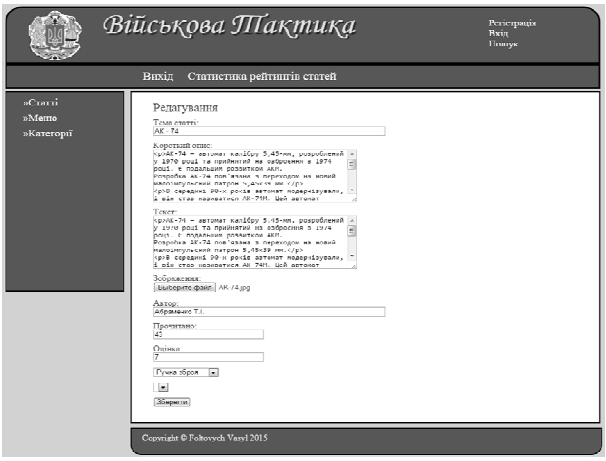
1) на першому етапі за допомогою методу масових спостережень збирають первинні статистичні дані. Основний зміст цього етапу полягає в отриманні даних, що характеризують кожну одиницю спостереження;

2) на другому етапі статистичного аналізу зібрані дані піддаються первинній обробці, зведенню і групуванню. Метод груповань дає змогу виділити однорідні сукупності, поділити їх на групи і підгрупи. Підсумок – це отримання підсумків загалом й за окремими її групами та підгрупами. Результати групування і зведення викладають у вигляді статистичних таблиць. Основний зміст цього етапу полягає в переході від характеристик кожної одиниці спостереження до зведених характеристиках сукупності загалом або її груп;

3) на третьому етапі отримані зведені дані аналізують методом узагальнювальних показників. Основний зміст цього етапу полягає у виявленні взаємозв'язків явищ, визначенні закономірностей їх розвитку та здійсненні прогнозних оцінок.



a



b

Рис. 14. Web-форма додання та редагування нової статті

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

Під час аналізу статистики відвідування окремої сторінки визначають також загальний час читання статті в секундах. Після цього визначають оцінку статті, яку розраховують за трьома критеріями: час читання, кількістьі звернень та користувацька оцінка. Аналіз статистики здійснюється класом statustuk. Для цього використовують дані, які збираються та зберігаються в відношенні бази даних timeread. Саме визначення статистики відвідування статті, рейтингового оцінювання статей та статистики рейтингів статей і є суттєвою відмінністю сайту від інших аналогічних сайтів та інформаційних ресурсів. Крім того, цей метод дає змогу оперативно знаходити інформацію про діяльність сайту та робити відповідні висновки. Основним критерієм цих методів є ефективність і простота використання. Також перевагами цих методів є: точність і достовірність результатів; простота математичних розрахунків; широке поле застосування; можливість отримання динамічних характеристик.

Висновки та перспективи подальших наукових розвідок

В статті спроектовано ІСОІР контент-аналізу текстової інформації інтернет-газети. Основною метою цієї системи є покращення якості та оперативності контенту інтернет-газети. Актуальність створення системи контент-аналізу текстової інформації інтернет-газети спричинена ростом вимог користувачів цих систем та зумовлена такими чинниками: швидкими темпами зростання потреб у достовірній та адекватній інформації, необхідністю формування множини оперативної інформації, а також автоматичного фільтрування небажаної інформації та спаму. Проаналізовано методи та засоби проектування. За допомогою методів і технологій побудовано діаграми, які відображають логічну та фізичну моделі функціонування системи. Визначено об'єкти та варіанти використання, які наявні в системі, основні функції системи, мету розроблення системи та місце застосування системи. Розроблено концептуальну модель системи, визначено вхідні, вихідні дані та наведено їх опис, описано вимоги до системи. Спроектована ІСОІР система контент-аналізу текстової інформації інтернет-газети під час роботи використовує аналіз статистики, рейтингове оцінювання та контент-аналіз, що робить її потужним засобом проведення аналізу. Основними перевагами цих методів є: точність і достовірність результатів, простота математичних розрахунків, широке поле застосування. Створено програмну реалізацію системи, базу даних, описано реалізацію завдання та проаналізовано отримані результати. Інтерфейс системи у відповідних формах призначений для наповнення бази даних даними і тестування роботи системи. Також було проведено верифікацію та валідацію і виправлено всі виявлені відхилення і дефекти у функціонуванні системи. Отже, ця ІСОІР контент-аналізу текстової інформації інтернет-газети є потужним інструментом в управлінні та успішному функціонуванні будь-якої газети.

1. 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. 2. Зміст методики "контент-аналіз" [Електронний ресурс] / Т. Хорошилова // Прикладна лінгвістика. – Режим доступу: http://studentstpl.ucoz.ru/publ/teorija_vozdejstvija/metodika_kontent_ analiza/ soderzhanie_ metodiki_kontent_analiz/12-1-0-116. 3. Математична лінгвістика. [Книга 1. Квантитативна лінгвістика]: навч. посібник / [В. В. Пасічник, Ю. М. Щербина, В. А. Висоцька, Т. В. Шестакевич] // Серія "Комп'ютинг". – Львів: "Новий світ -2000", 2012. – 359 с. 4. Григорьев С. Проведение контент-анализа [Електронний ресурс]: навч. посібник / С. Григорьев. – Режим доступу: http://www.psyfactor.org/lib/k-a2.htm. 5. Контент-аналіз як метод дослідження [Електронний ресурс] / О. Т. Манаєв // Псі-фактор. – Режим доступу: http://psyfactor.org/lib/content-analysis3.htm. 6. Береза А. Електронна комерція / А. Береза, І. Козак, Ф. Левченко. – К: КНЕУ, 2002. – 326 с. 7. Клифтон Б. Google Analytics: профессиональный анализ посещаемости веб-сайтов / Б. Клифтон. – М.: Вильямс, 2009. – 400 с. 8. Деревья решений: общие принципы работы [Електронний ресурс] / Акобир Шахиди. – Доступ: http://www.basegroup.ru/ library/ analysis/tree/description/. 9. Аналіз статистики: Методи багатовимірної статистики. – Режим доступу: http://christsocio.info/content/view/492/102/ – Назва з титул. екрану. 10. Іванов В. Ф. Контент-аналіз: Методологія і методика дослідження ЗМК: навч. посібник / В. Ф. Іванов;[наук. ред. А. З. Москаленко]. – К., 1994. – 112 с. 11. Основы моделирования и оценки електронных информационных потоков / [Д. Ландэ, В. Фурашев, С. Брайчевский,

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

О. Григорьев]. – К.: Інжиніринг, 2006. – 348 с. 12. Ландэ Д. Основы интеграции информационных потоков: монография / Д. Ландэ. – К. : Інжиніринг, 2006. – 240 с. 13. Поспелов Д. Ситуационное управление: теория и практика / Д. Поспелов. – М.: Наука. – 1986. – 288 с. 14. CM Lifecycle Poster / Content Management Professionals. – Retrieved 20 July 2010. – Access: http://www.cmprosold.org/ resources/poster/. – Title from the screen. 15. EMC. Content Management Interoperability Services. Appendices. Version 0.5 / EMC, IBM, Microsoft. – Hopkinton: EMC, 2008. – 17 p. 16. EMC. Content Management Interoperability Services. Part I. Version 0.5 / EMC, IBM, Microsoft. – Hopkinton: EMC, 2008. – 76 p. 17. EMC. Content Management Interoperability Services. Part II – REST protocol binding. Version 0.5 / EMC, IBM, Microsoft. – Hopkinton: EMC, 2008. – 79 p. 18. EMC. Content Management Interoperability Services. Part II – SOAP protocol binding. Version 0.5 / EMC, IBM, Microsoft. – Hopkinton: EMC, 2008. – 37 p. 19. Hackos J. Content Management for Dynamic Web Delivery / J. Hackos. – Hoboken: Wiley, 2002. – 432 p. 20. Halvorson K. Content Strategy for the Web / K. Halvorson. – Reading: New Riders Press, 2009. – 192 p. 21. McGovern G. Content Critical / G. McGovern, R. Norton. – Upper Saddle River: FT Press, 2001. – 256 p. 22. McKeever S. Understanding Web content management systems: evolution, lifecycle and market / S. McKeever // Industrial Management & Data Systems (MCB UP), 2003. – № 103 (9). – P. 686–692. 23. Nakano R. Web content management: a collaborative approach / R. Nakano. – Boston: Addison Wesley Professional, 2002. – 222 p. 24. Osgood C. The nature and measurement of meaning / C. Osgood // Psychological Bulletin, 49 (1952). – P. 197-237. 25. Papka R. On-line News Event Detection, Clustering, and Tracking: thesis for the degree doctor of philosophy / R. Papka. – Amherst: Massachusetts University, 1999. – 154 p. 26. Woods R. Defining a Model for Content / R. Woods. – 2010. – Access: http://www. contentmanager.net/magazine/article_785_defining_a_model_for_ content_governance.html. – Title from the screen. 27. Rockley A. Managing Enterprise Content: A Unified Content Strategy / A. Rockley. – Reading: New Riders Press, 2002. – 592 p. 28. Stone W. R. Plagiarism, Duplicate Publication and Duplicate Submission: They Are All Wrong! / W.R. Stone // IEEE Antennas and Propagation, 2003. – Vol. 45. – № 4. – P. 47–49. 29. Sullivan D. Invisible Web Gets Deeper / D. Sullivan // Search Engine Report. – 2002. – Access: http://searchenginewatch.com/sereport/article.php/2162871. – Title from the screen. 30. The Content Management Possibilities Poster [Electronic resource] / Metatorial Services, Inc. – Retrieved 20 July 2010. – Access mode: http://metatorial.com/pagea.asp? id=poster. 31. Methods based on ontologies for information resources processing : Monograph / [Vasyl Lytvyn, Victoria Vysotska, Lyubomyr Chyrun, Dmytro Dosyn] // LAP Lambert Academic Publishing. Saarbrücken, Germany. – ISBN-13: 978-3-659-89905-8, ISBN-10: 3659899054, EAN: 9783659899058. – 2016. – 324 с. 32. Висоцька В. А. Методи і засоби опрацювання інформаційних ресурсів в системах електронної контент-комерції: автореферат дисертації на здобуття наукового ступеня кандидата технічних наук: 05.13.06 – інформаційні технології / Вікторія Анатоліївна Висоцька; Національний університет "Львівська політехніка". – Львів, 2014. – 27 с. 33. Berko A. Features of information resources processing in electronic content commerce / Andriy Berko, Victoria Vysotska, Lyubomyr Chyrun // Applied Computer Science. ACS journal. – Vol. 10, Number 2. – Poland, 2014. – ISSN 2353-6977 (Online), ISSN 1895-3735 (Print). – P. 5–19 [Online]. 34. Vysotska Victoria. Web Content Processing Method for Electronic Business Systems / Victoria Vysotska, Lyubomyr Chyrun // International Journal of Computers & Technology. – Vol 12, No 2. – December 2013. – P. 3211–3220. – ISSN 2277-3061.– [Online] http://cirworld.org/journals/index.php/ijct/article/view/3299. Impacr factor 1,532. (Index Copernicus, NASA ADS, DOAJ, Google Scholar, Eyesource, EBSCO, CiteSeer, UlrichWeb, ScientificCommons, ProQuest CSA Technology Research Database). 35. Висоцька В. А. Моделювання етапів життєвого циклу комерційного web-контенту / В. А. Висоцька, Л.Б Чирун, Л. В. Чирун // Інформаційні системи та мережі. Вісник Нац. ун-ту "Львівська політехніка". – Львів, 2011. – № 715. – С. 69–87. 36. Висоцька, В. А. Особливості проектування та впровадження систем електронної комерції / В. А. Висоцька // Інформаційні системи та мережі. Вісник Національного університету "Львівська політехніка". – Львів, 2008. – № 631. – С. 55–84. 37. Vysotska V. Analysis and evaluation of risks in electronic commerce / V. Vysotska, I. Rishnyak, L. Chyrun // CAD Systems in Microelectronics, CADSM '07, 9th International Conference. – The Experience of Designing and Applications of CAD Systems in Microelectronics. – Lviv, 24 February 2007. – P.332–333.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua