УДК 004.738.5

Л. Б. Чирун, В. В. Кучковський, В. А. Висоцька
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

ОСОБЛИВОСТІ МЕТОДІВ КОНТЕНТ-АНАЛІЗУ ТЕКСТОВИХ МАСИВІВ ДАНИХ WEB-РЕСУРСІВ У МЕЖАХ РЕГІОНУ

© Чирун Л. Б., Кучковський В. В., Висоцька В. А., 2015

Описано метод інтегрованого опрацювання неоднорідних інформаційних ресурсів web-систем, який ґрунтується на моделі подання даних як узгодженого поєднання значень, правил їх зображення, правил інтерпретації та структури. Метод передбачає декомпозицію загального процесу на підпроцеси інтеграції значень, синтаксису даних, семантики і структури. Перевагою такого підходу до інтеграційних процесів є можливість їх виконання на рівні метасхем даних, що зменшує кількість звернень до власне даних web-систем, обсяги яких можуть бути значними. У статті 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. У роботі проаналізовано основні проблеми електронної комерції та функціональних сервісів управління контентом. Запропоновано метод управління комерційним контентом як етап життєвого циклу контенту в системах електронної комерції. Метод управління комерційним контентом описує процеси формування інформаційних ресурсів у системах електронної контент-комерції та спрощує технологію управління комерційним контентом. У цій роботі проаналізовано основні проблеми електронної контент-комерції та функціональних сервісів опрацювання комерційного контенту. Запропонований метод дає можливість створити засоби опрацювання інформаційних ресурсів у системах електронної контент-комерції та реалізувати підсистему управління комерційним контентом. Розглянуто питання розроблення методів та програмних засобів опрацювання інформаційних ресурсів у інтернет-системах. Сформульовано новий підхід застосування та впровадження бізнес-процесів для побудови таких систем. Розроблено методи та програмні засоби опрацювання контенту та інформаційного ресурсу.

Ключові слова: контент, інформаційний ресурс, комерційний контент, контент-аналіз, життєвий цикл контенту, контент-моніторинг, контентний пошук, система електронної контент-комерції, web-ресурс, значення даних, інтеграція даних, розподілені системи даних, неоднорідні дані, бізнес-процес, система управління контентом, життєвий цикл контенту.

In the paper the method of integrated processing of heterogeneous information resources web-systems is described. This method is based on the model of data description as a coherent combination of data values, rules of data representation, interpretation rules and data structure. The method involves decomposition of general process into subprocesses of data values integration, data syntax integration, semantics and structure integration. The advantage of this approach is that the integration process can be performed at data metascheme level. It allows to reduce the number of access operation to very large data sets of web-systems. 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. In the paper the main problems of e-commerce and content function management services are analyzed. The method of commercial content

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

management as the content life cycle stage in electronic commerce systems is proposed. The method of commercial content management describes the information resources forming in electronic content commerce systems and automation technology that simplifies the commercial content management. The main problems of electronic content commerce and functional services of commercial content management are analyzed. The proposed method gives an opportunity to create an instrument of information resources processing in electronic commerce systems and to implement the subsystem of commercial content management. The article discusses the development of unified methods and software tools for processing information resources in the Internet systems. A new approach to application and implementation of business processes is formulated for the construction of these systems. The methods and software tools of content and information resource processing are developed.

Key words: content, information resources, commercial content, content analysis, content lifecycle, content monitoring, content search, electronic content commerce systems, web-resource, data value, data integration, distributed data systems, heterogeneous data, business-process, content management system, content lifecycle.

Вступ. Загальна постановка проблеми

Класичні моделі та методи теорії баз даних орієнтовані на організацію зберігання й опрацювання структурованих даних у фактографічних інформаційно-пошукових системах. Найчастіше ці дані є числовими значеннями, що описують характеристики інформаційних об'єктів. Але часто інформація подається у вигляді не структурованих масивів даних, а масиву контенту. На відміну від традиційних баз даних, орієнтованих на повне та точне подання даних достатньо простої змістової структури, системи опрацювання Web-ресурсів орієнтовані на часткове, наближене подання даних зі значно складнішою змістовою структурою, що подається на вході у формі контенту. Основною функцією будь-якої системи опрацювання Web-ресурсів є інформаційне забезпечення користувачів на основі видавання відповідей на їх запити. Видавання системою опрацювання Web-ресурсів необхідних даних реалізується за допомогою головної операції – проведення контентного пошуку. *Контентний пошук* є процедурою підбору контенту, що містить відповідь на поставлені користувачем питання [1–2]. На відміну від фактографічних інформаційно-пошукових системи, які у відповідь на запит користувача видають конкретні відомості (факти), системи опрацювання Web-ресурсів у результаті контентного пошуку надають користувачеві масив контенту, зміст якого відповідає його запиту.

Зв'язок висвітленої проблеми із важливими науковими та практичними завданнями

Контентний пошук у системі опрацювання Web-ресурсів здійснюється на основі запиту, що надійшов від користувача, на підбір необхідного йому контенту. Потреба користувача в актуальній та оперативній інформації в процесі практичної діяльності є *інформаційною потребою* (ІП). Під дією отримуваного контенту ІП користувачів постійно модифікується, змінюється і трансформується. Тому її неможливо однозначно визначити і описати. ІП, яка подається у вигляді деякої послідовності властивих їй значень у фіксовані моменти часу, та висловлена природною мовою, і є *контентним запитом*, з яким користувач звертається до системи опрацювання Web-ресурсів. Якщо запит користувач неправильно сформулював і він не відображає його дійсної ІП у момент звернення до системи опрацювання Web-ресурсів, то під час контентного пошуку в ній фактично розглядається не ІП користувача, а контентний запит, у відповідь на який видається контент. Отже, реакцію системи опрацювання Web-ресурсів необхідно розглядати не лише відносно ІП, але і щодо контентного запиту. Для подання цих відношень у системах опрацювання Web-ресурсів є два фундаментальні поняття: пертинентність і релевантність. *Пертинентністю* є відповідність змісту контенту ІП користувача. Контент, зміст якого задовольняє ІП, є пертинентним. *Релевантність* є відповідністю вмісту контенту контентному запиту в тому вигляді, в якому він сформульований, а контент, вміст якого відповідає запиту користувача, є релевантним. Автоматизація процесу контентного пошуку потребує формалізації подання основного змісту контентного запиту і контенту у вигляді відповідного пошукового розпорядження (ПРК) і пошукових образів контенту

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

(ПОК), для запису яких застосовують спеціальні мови – інформаційні або інформаційно-пошукові. В процесі проведення контентного пошуку в системах опрацювання Web-ресурсів визначається ступінь відповідності змісту контенту і запиту користувача зіставленням ПОК із ПРК. А на основі такого зіставлення приймається рішення про видавання/невидавання релевантного/нерелевантного контенту. Рішення про видавання/невидавання контенту у відповідь на запит приймається на основі деякого набору правил, за яким ця система опрацювання Web-ресурсів визначає міру змістової близькості між ПОК і ПРК. Такий набір правил є *критерієм змістової відповідності* K_c, який задається явно/неявно. Насправді K_c ґрунтується не на раніше введеному понятті релевантності, а на понятті *формальної релевантності* – відповідності вмісту ПОК і ПРК. *Фактичну релевантність*, тобто змістову відповідність контенту щодо запиту, встановлює користувач, намагаючись зрозуміти їх вміст.

Аналіз останніх досліджень та публікацій

Загальна функціональна структура систем опрацювання Web-ресурсів

На рис. 1 подано загальну діаграму варіантів довільного Web-ресурсу.

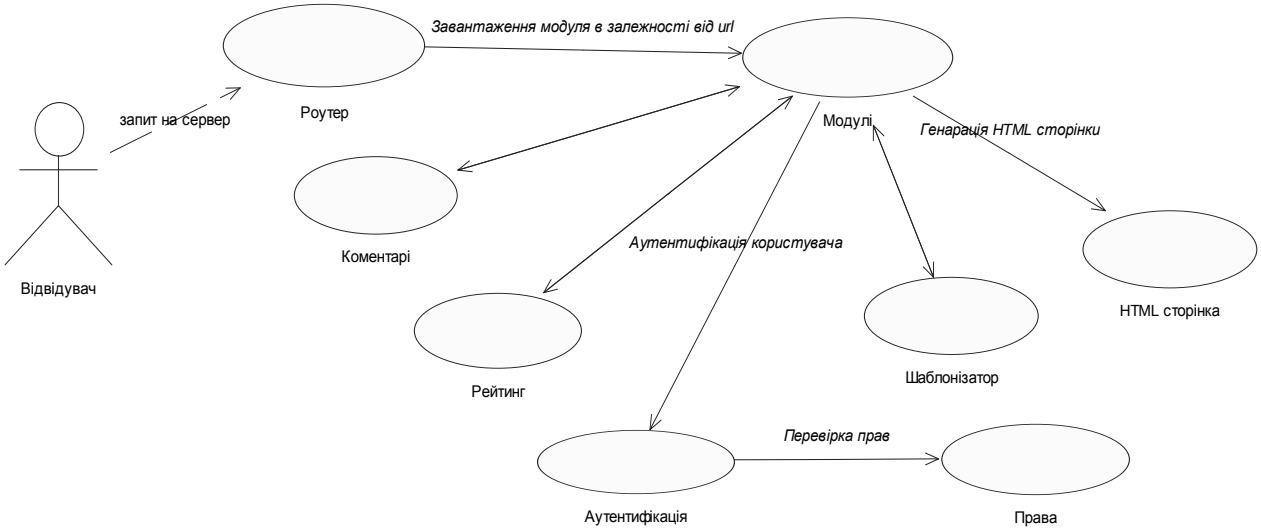

Рис. 1. Загальна діаграма варіантів довільного Web-ресурсу

Обов'язковим елементом системи опрацювання Web-ресурсів є контентно-пошукова підсистема. До складу контентно-пошукової підсистеми системи опрацювання Web-ресурсів входять чотири основні модулі (рис. 2) [1–8]:

1. Модуль реєстрації користувача і введення запиту.
2. Модуль опрацювання контенту.
3. Модуль пошуку контенту.
4. Модуль збереження та подання контенту.

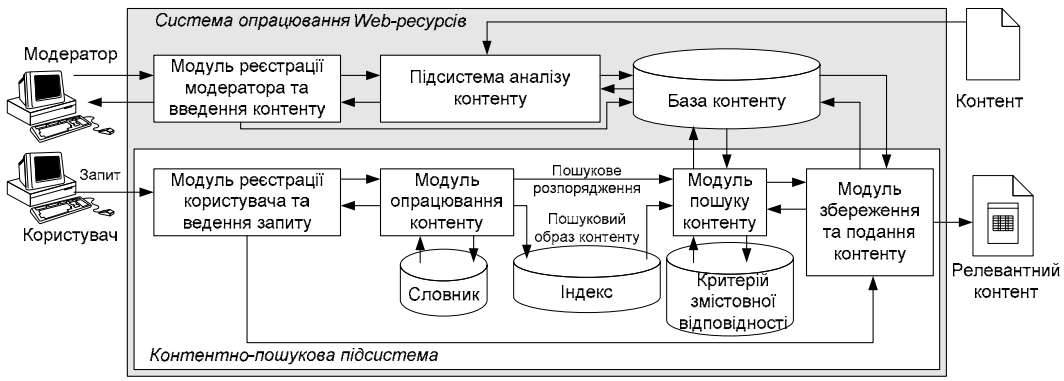

Рис. 2. Загальна функціональна структура системи опрацювання Web-ресурсів

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

Контент, що надходить на вхід системи опрацювання Web-ресурсів, не структурований, не форматований та не формалізований. Тому підсистема введення і реєстрації вирішує такі основні завдання: збереження оригіналу; фільтрування контенту; форматування контенту; рубрикація контенту; виявлення ключових слів; присвоєння контенту унікальних ідентифікаторів (реєстрація), а також ведення таблиці синхронізації імен (за необхідності – збереження колишніх імен); виявлення дублювання контенту; формування дайджестів; збереження контенту в анотованій базі даних; вибіркове поширення контенту між модераторами системи опрацювання Web-ресурсів.

Контент, що надходить, без внесення в нього яких-небудь змін прямує у модуль збереження в базі контенту. База контенту є простою сукупністю файлів, розподіленою за каталогами. Але такий тип подання бази контенту характеризується двома недоліками: неефективним використанням дискового простору; низькою швидкістю доступу за великої кількості файлів. Тому для зберігання контенту застосовують засоби стиснення і швидкого пошуку контенту. В цьому випадку модуль зберігання є сукупністю стандартних або спеціалізованих засобів архівації, СУБД тощо для забезпечення можливості доступу до даних за поданим ідентифікатором.

Контент після його реєстрації та введення надходить на вхід модуля опрацювання, завданням якого є формування для кожного контенту ПОК, в який вноситься інформація, необхідна для подальшого його пошуку. ПОК зберігаються в індексі. Логічно індекс є таблицею, рядки якої відповідають контенту, а стовпці – інформаційним ознакам, на основі яких будується пошуковий образ. Така таблиця є сильно розрідженою, і зберігати всі значення немає сенсу. Тому використовують пряму або інверсну форму зберігання, тобто згортку таблиці за рядками або стовпцями відповідно. Оскільки під час згортання таблиці структура індексу ускладнюється, для його підтримки використовують СУБД. Запит користувача після його реєстрації та введення контентно-пошукова підсистема перетворює на пошукове розпорядження і передає в модуль пошуку, завданням якого є знаходження в індексі ПОК, які задовольняють ПРК відносно K_c. Ідентифікатори релевантного контенту подають з виходу модуля пошуку на вхід модуля зберігання, який і видає користувачеві релевантний контент.

Проблема формального подання змісту контенту

Недоліки подання контенту природною мовою. Природна мова є універсальною знаковою системою, що слугує для обміну інформацією між користувачами. Оскільки запит, що надходить на вхід контентно-пошукової підсистеми, записаний на природній мові, то доцільно її використовувати як основний засіб подання текстового контенту під час всього циклу функціонування системи опрацювання Web-ресурсів. Але в сучасних системах операцію встановлення відповідності між запитом і контентом виконує комп'ютер, що фактично виключає застосування природної мови як основного засобу подання контенту. Це пояснюється істотними недоліками природної мови з погляду технології автоматичного опрацювання контенту.

1. *Різноманіття засобів подання змісту контенту.* Основним засобом передавання змісту текстового контенту є лексика природної мови, але функцію передавання змісту виконують інші елементи: контекст; парадигматичні відношення між словами; контентуальні відношення між словами; посилання на слова (словосполучення, фрази тощо), раніше згадувані в тексті контенту.

2. *Семантична неоднозначність.* Контент природною мовою є семантично неоднозначним переважно через синонімію та багатозначність вживаних слів.

3. *Синонімія* є тотожністю або близькістю за значенням слів, що виражають те саме поняття, які відрізняються одне від іншого або відтінками значень, або стилістичним забарвленням, або одночасно обома названими ознаками. Синонімами природної мови є окремі слова або словосполучення.

4. *Багатозначність* характеризує можливість неоднозначного розуміння змісту окремих слів природної мови. Багатозначність слів має два різновиди – полісемію і омонімію, які поширеніші в синтаксичних мовах (наприклад, слов'янських), ніж в аналітичних (наприклад, германських).

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

Таблиця 1

Різновиди багатозначності в природній мові

Назва	Тлумачення	Приклади
Полісемія	Збіг назв різних предметів, що мають будь-які загальні властивості або ознаки. До типових загальних властивостей, що слугують базою полісемії, належать подібність предметів, їх суміжність (просторова, тимчасова тощо), а також однакове функціональне призначення.	*команда* (військовий підрозділ, екіпаж судна, спортивна), *промінь* (сонця, лазера, світла), *соняшник* (рослина, стебло рослини), *дзвони* (багато дзвонів, дзвонити), *дідько* (чорт, погана людина), *марка* (поштова, авто, торгова), *батьківщина* (вітчизна, місцевість батьків)
Омонімія	Збіг назв різних предметів, що не мають загальних властивостей. Омонімічні слова, тотожні за написанням і звучанням, слід відрізняти від омографів – слів, що позначають різні предмети, однакові за написанням, але різні за звучанням, наприклад: “замок” (дверний) – “замок” (палац). Але, оскільки контентно-пошукова підсистема оперує із запитом і контентом на природній мові, внаслідок чого фонетика мови не має вирішального впливу на зміст контенту, омографи можна прирівняти до омонімічних слів.	*диск* (компакт-диск, спортивний диск, кістка, для дискування землі), *порох* (для вогнепальної зброї, пилюка), *кран* (машина, умивальник), *гребінь* (у півня, для волосся, гори, хвилі), *карі* (приправа, очі), *літа* (роки, літо), *крона* (дерева, польські гроші, елемент живлення), *ясен* (дерево, ясен день), *граната* (зброя, фрукт), *колонка* (газова, таблиці, на бензозаправці), *стовпець* (в землі, в таблиці), *пліт* (огорожа, човен, засіб для плавання, покривало), *орел* (птах, бік монети), *гусениця* (комаха, в тракторі або танку), дати *газу* (речовину, пришвидшити рух), сидить козак на тім *боці* (стороні чи частині тіла), *журавель* (птах, біля криниці), *ключ* (птахів, дверний, музичний, в землі), *вуйко* (дядько, ведмідь, місяць), *пекло* (під землею, жарко), *доля* (судьба, частина), *губа* (у людини, в морфлоті заходи – гауптвахта, в морі керченська губа), *край* (кінець чогось, земля), *лист* (лавровий, повідомлення), *став* (ставок, стати, на край обриву або на хибний шлях), *забрало* (дієслово, елемент шолома), *торба* (сумка, одяг), *рукав* (одягу, річки), *рок* (музика, доля), *воля* (свобода, сила волі), *шишка* (від удару, соснова), *дружина* (жінка, військо), *халява* (частина чобота, легко та доступно), *шайба* (хокей, елемент кріплення), *долина* (звук долинає, рівнина), *чайник* (для води, людина)

5. *Еліпсність.* У контенті на природній мові трапляються еліпси або пропуски слів. Еліпсність контенту відіграє негативну роль під час безпосередньої роботи з ним користувача. Але вона негативно вплине і в тому випадку, коли контент природною мовою опрацьовує комп’ютер.

Інформаційно-пошукові мови. Неможливість використання природної мови як основного засобу подання контенту в контентно-пошуковій підсистемі призводить до необхідності застосування штучних мовних засобів. *Інформаційно-пошуковою* мовою (ІПМ) є спеціалізована штучна мова, призначена для опису основного змісту контенту, що надходить в систему, щоб забезпечити можливості подальшого пошуку [1–2]. ІПМ створюється на основі природної мови, але відрізняється від неї компактністю, наявністю чітких граматичних правил і відсутністю семантичної неоднозначності. Існує два основні типи ІПМ: класифікаційні та дескрипторні.

Принципова відмінність між цими типами мов полягає в процедурі побудови речень (фраз) мови. У деяких мовах у їх лексичний склад поряд із словами, що означають прості поняття, заздалегідь вводять також словосполучення і фрази, що відображають складні поняття. Для запису змісту контенту ІПМ використовують лише окремі елементи з цього набору, зокрема і готові складні поняття. Фактично побудова складних синтаксичних конструкцій замінена вибором відповідного складного поняття (у вигляді словосполучення або фрази) із готового набору, наприклад,

{*Політика.Внутрішня.Федеральна*, *Політика.Внутрішня.Регіональна*, ..., *Політика.Зовнішня*...}

За допомогою таких мов (*класифікаційних*) класифікують контент, тобто зараховують його до класів, позначених лексичними одиницями ІПМ. Частковим випадком класифікаційної ІПМ є

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

рубрикатор, лексичними одиницями якого є назви тематичних рубрик. Загалом під *рубрикатором* деякої предметної області (ПО) розуміють орієнтований граф, що складається з незалежних дерев. Листя дерев називатимемо *рубриками* – об'єктами, що інкапсулюють знання про конкретні фрагменти ПО. Всі нелистові вершини є класифікаційними родово-видовими узагальненнями листових вершин і використовуються лише для введення контентного пошуку. Рубрикатор формує група експертів, на підставі їх знань про ПО з урахуванням ІП користувачів. На рис. 3 наведено приклад рубрикатора деякої ПО.

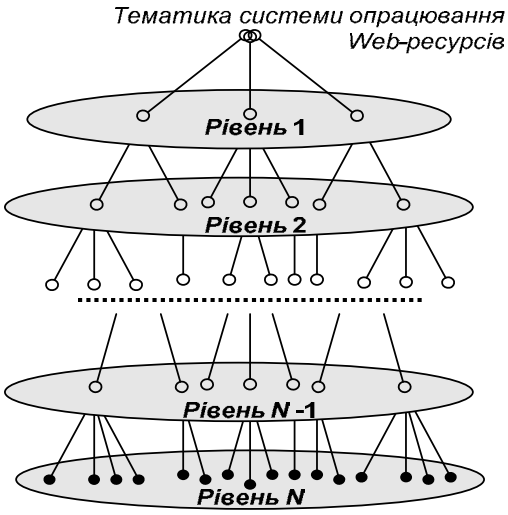

Рис. 3. Приклад рубрикатора

Для рубрикації наукових робіт можна побудувати різні типи онтології предметної області $O =< V, Q, F >$, з певними властивостями, перевагами та недоліками:

I. Модель $O_1 =< V, \{\}, \{\} >$, де V – ключові слова предметної області.

II. Модель $O_2 =< V_1 \cup V_2, \{\}, F >$, де V_1 – всі можливі флексії, а застосування алгоритму стемінгу для відсікання флексій сформує множину основ; V_2 – множина основ ключових слів; F – скінченна множина функцій інтерпретації (аксіоматизація, обмеження), заданих на поняттях чи відношеннях онтології O для рубрикації тексту.

III. Модель $O_3 =< V_1 \cup (V_2' \cup P_2'), \{\}, F >$, де V_1 – всі можливі флексії, а застосування алгоритму стемінгу для відсікання флексій сформує множину основ; V_2 – множина основ ключових слів; $(V_2' \cup P_2')$ – лематизація через визначення частин мови у речені (так званий POS tagging) та застосування до слова правила стемінгу відповідно до частини мови; F – скінченна множина функцій інтерпретації (аксіоматизація, обмеження), заданих на поняттях чи відношеннях онтології O для рубрикації тексту.

IV. Модель $O_4 =< V, \{IS - A\}, \{\} >$, де V – ключові слова предметної області та $\{IS - A\}$ – словник рубрики.

Модель системи онтологій для визначення рубрики наукової роботи:

$$\Sigma^O =< O^{meta}, \{O^{domain}\}, \Xi^{\inf} >,$$

де O^{meta} – онтологія мови (для перекладача); O^{domain} – онтологія предметної області (для рубрикації); $\Xi^{\inf}$ – тематики наукових робіт (для шуканих спільних напрямів наукових досліджень).

Особливість класифікаційних мов. Оскільки складні поняття задаються заздалегідь, до початку процедури запису контенту за допомогою ІПМ, слова, які їх створюють, також наперед зв'язані (скоординовані) певними зв'язками. До іншого типу мов належать дескрипторні ІПМ, в яких лексичні одиниці наперед не зв'язані концептуальними відношеннями. Складні синтаксичні

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

конструкції – речення або фрази – створюються в цих мовах об'єднанням (координацією) лексичних одиниць під час процедури подання змісту контенту системи. Готових речень або фраз у таких мовах немає, тому відсутні обмеження на утворення складних понять. Фактично з невеликої кількості лексичних одиниць ці мови дають змогу будувати речення, що висловлюють будь-який зміст. Розрізняють дескрипторні ІПМ з *граматикою* і *без граматики.* Перші характеризуються наявністю суворих правил формування синтаксичних конструкцій. Наприклад, якщо використовують дескрипторну ІПМ з позиційною граматикою, в якій для опису дій прийнято на першому місці записувати назву дії, далі суб'єкта, а потім об'єкта цієї дії, фраза "Софія тримає іграшку" може виглядати як "тримати Софія іграшка". У дескрипторних ІПМ без граматики таких правил немає, і порядок розташування лексичних одиниць в ПОК або ПРК не відіграє ролі. Тобто наведений вище приклад може бути однаково поданий послідовностями "тримати Софія іграшка", "Софія тримати іграшка" тощо. Існують дескрипторні ІПМ з *контрольованою* і з вільною *лексикою*. *Лексичний* склад перших строго обмежений і зафіксований в словнику ІПМ, тоді як на лексичний склад інших не накладають жодних обмежень, і він може постійно поповнюватися за рахунок додавання нових лексичних одиниць.

Аналіз отриманих наукових результатів

Опрацювання вхідної текстової інформації

Оскільки контент, що надходить на вхід системи опрацювання Web-ресурсів, записаний на природній мові, в ній обов'язково повинна проводитися операція перекладу текстів вхідного контенту з природної мови на ІПМ. Тип використовуваної ІПМ істотно впливає на суть процесів управління контентом у конкретних системах опрацювання Web-ресурсів. У разі використання дескрипторної ІПМ така операція перекладу є *індексуванням,* у разі застосування рубрикатора – *рубрикацією*. Сьогодні серед дескрипторних ІПМ найбільшого поширення в автоматизованих системах опрацювання Web-ресурсів набули мови без граматики і без контролю за словником. У разі використання говорять про *повнотекстове індексування.*

У операції перекладу виділяють два етапи:

1. Аналіз змісту контенту з метою виділення з нього відомостей про відомі системі об'єкти, їхні властивості, а також відношення між ними.
2. Подання цих відомостей на ІПМ, тобто ухвалення рішення про приписування цьому контенту виразів ІПМ (про приєднання відповідних виразів на ІПМ в ПОК).

Етап аналізу змісту контенту пов'язаний з необхідністю використання лінгвістичних і екстралінгвістичних знань. Лінгвістичні знання є загальними для однієї мови і вже доволі добре формалізовані, тоді як екстралінгвістичні істотно залежать від конкретної ПО, а завдання їх формалізації є одним з найскладніших. У зв'язку з цим у сучасних системах опрацювання Web-ресурсів етап аналізу контенту найчастіше зводиться до лінгвістичного аналізу, що проводиться з метою *нормалізації* слів і словосполучень. Під нормалізацією слів розуміємо їх приведення до канонічної форми (наприклад, для іменників – називного відмінку, однини тощо), під нормалізацією словосполучень – нормалізацію складових і запис їх в певній послідовності (наприклад, спочатку записується основне слово, а потім – залежні слова). Нормалізовані слова і словосполучення є *термінами.*

Лінгвістичний аналіз контенту складається з двох етапів: *морфологічний* та *синтаксичний*.

Мета *морфологічного аналізу* полягає в здобутті основ (під основою розуміємо словоформу з відсіченим закінченням) зі значеннями граматичних категорій (наприклад, частина мови, рід, число, відмінок) для кожної зі словоформ. Розрізняють точні та наближені методи морфологічного аналізу. Використання словника словоформ у точних методах дає змогу легко подолати труднощі морфологічного аналізу, пов'язані з такими явищами в українській мові, як, наприклад, чергування голосних і приголосних. За такого підходу завдання отримання основ слів і граматичних ознак зводиться переважно до пошуку в словнику і вибору відповідної інформації (власне ж морфологічний аналіз потрібний лише в тому випадку, якщо словоформа не знайдена в словнику). Якщо словник достатньо повний, швидкість опрацювання матеріалу доволі висока, але об'єм необхідної

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

пам'яті в 2–3 рази більший, ніж у разі використання словника основ. Морфологічний аналіз з використанням словника основ ґрунтується на флективному аналізі, мета якого – правильне виділення основи слова. Основні труднощі використання цього підходу пов'язані з явищем омонімії основ слів. Щоб усунути її, перевіряється сумісність виділеної основи слова і його закінчення. В основу наближених методів морфологічного аналізу покладена гіпотеза, згідно з якою за кінцевими літерами і буквосполученнями можна практично однозначно визначити граматичний клас слова. Внаслідок проведення морфологічного аналізу можуть виникати неоднозначності під час визначення граматичної інформації, які ліквідуються після проведення синтаксичного аналізу.

Завданням *синтаксичного аналізу* є здійснення граматичного розбору речень на основі інформації, закладеної в словнику. На цьому етапі виділяють підмет, присудок, доповнення тощо, між якими вказуються зв'язки з управління у вигляді дерева залежностей.

Будь-які засоби синтаксичного аналізу складаються з двох частин: бази знань про конкретну мову і власне алгоритму синтаксичного аналізу, тобто набору стандартних операторів, що опрацьовують текст на основі цих знань. Джерелом знань (граматичних) є дані, отримані в результаті морфологічного аналізу, а також різні таблиці, які апріорі заповнені стандартно і є результатом емпіричного опрацювання природомовних текстів людиною з метою виділення певних закономірностей, які необхідні для проведення синтаксичного аналізу. Основою цих таблиць є сукупності конфігурацій або набори валентностей (синтаксичних і семантико-синтаксичних), що є списками лексичних одиниць із вказанням для кожної з них всіх можливих варіантів зв'язків із іншими одиницями виразу на природній мові (тобто потенційних зв'язків). Під час практичної реалізації синтаксичного аналізу прагнуть добиватися повної незалежності правил перероблення даних таблиць від їх вмісту, щоб зміна у разі потреби цього вмісту не спричиняла перебудову самого алгоритму [3].

Автоматичне індексування документів ґрунтується на простих, однослівних або багатослівних складних термінах (фразах). Прості, однослівні терміни далеко не ідеальні для індексування, оскільки зміст слів поза контекстом нерідко буває неоднозначним. Терміни-фрази осмисленіші, у них більша дискримінувальна потужність. Для генерації фраз може використовуватися як синтаксичний аналіз, так і евристичні алгоритми. Наприклад, термін-фраза складається з основи фрази (звичайно це її головна частина) і решти компонентів. Термін з частотою входження у контент, що перевищує встановлений поріг, наприклад $df > 2$, наголошується як основа фрази. Іншими компонентами фрази є терміни із середньою або низькою частотою входження. При цьому враховується їх зв'язок з основою фрази, наприклад, розміщення їх в одному реченні або на деякій заданій відстані один від одного.

Для генерації груп взаємозв'язаних слів за зазначеними закономірностями спільного їх входження у контент застосовують методи групування або кластеризації термінів. Якщо уявляти матрицю контент-термінів у вигляді двовимірного масиву, то вищезазначений метод порівнює один із одним стовпці матриці і робить висновок про те, чи входить та або інша група термінів у множину контенту деякої сукупності. Якщо таке неодноразове входження виявлено, то терміни є зв'язаними і групуються в один клас. Прості та складні терміни, що виконують лише граматичну функцію, заносять в списки винятків і віддаляють. Основою сучасних методів автоматичного індексування є присвоювання вагових коефіцієнтів термінам на основі статистичних характеристик.

Нехай в досліджуваній сукупності є N контенту, а tf_{ij} – частота входження терміна T_j, в контент D_i. Індексування на основі частоти терміна дає змогу досягти лише однієї з цілей індексування – повноти пошуку. Тим часом терміни, сконцентровані в окремому контенті деякої сукупності, можна використовувати для підвищення точності пошуку. Це дасть змогу відокремити контент, де такі терміни трапляються, від тих, де їх немає [1].

Нехай df_j – кількість контенту, в якому трапляється термін T_j. Тоді величина $\log\left(N/df_j\right)$ слугує індикатором того, чи є термін T_j дискримінатором контенту.

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

Частоту терміна й отриману вище величину можна об'єднати в межах єдиної моделі індексування за частотою (де w_{ij} позначає вагу терміна T_j в контенті D_i) [1].

$$w_{ij} = tf_{ij} \log\left(N / df_j \right).$$

Ще один статистичний метод індексування ґрунтується на дискримінації за терміном. Тут кожен контент розглядається як крапка в просторі множини контенту. Що більше подібності в множині термінів двох контентів, то ближче розташовані відповідні крапки в просторі контенту (інакше кажучи, підвищується щільність крапок у просторі контенту), і навпаки.

У межах цієї схеми оцінюють якість терміна як дискримінатора контенту, ґрунтуючись на тому, які зміни відбудуться в просторі контенту після введення терміна в індекс. Для кількісної оцінки такої зміни зручно використовувати збільшення або зменшення відстані між контентом. Термін є дискримінатором, якщо його введення збільшує середню відстань між контентом. Тобто термін із кращими дискримінувальними якостями знижує щільність у просторі контенту. Дискримінувальна характеристика терміна T_j, що позначається як dv_j, обчислюється як різниця між щільністю простору контенту до і після введення терміна T_j. Виявилось, що терміни, що часто вживаються, мають негативні значення дискримінувальних характеристик, терміни з середньою частотою – позитивні, а для термінів, що рідко використовуються, ці значення близькі до нуля. Для спільного обліку частоти терміна і його дискримінувальної характеристики застосовують схему зважування, основану на виразі

$$w_{ij} = tf_{ij} dv_j.$$

Набуті значення ваг термінів використовують, приймаючи рішення про приєднання кожного з термінів у ПОК. Але частіше рішення не приймається, а в ПОК заносяться всі терміни, виявлені в контенті, та їхні ваги.

Автоматична рубрикація

У сучасних дослідженнях з цієї проблеми виділяють два основні підходи [6]: рубрикація, основана на *знаннях*, і рубрикація, основана на *навчанні на прикладах*.

Методи автоматичної рубрикації, основані на знаннях. У системах, що реалізовують цей підхід, використовуються заздалегідь сформовані бази знань, в яких описуються мовні вирази, відповідні тій чи іншій рубриці, правила вибору між рубриками. Процес створення таких систем часто порівнюють зі створенням експертних систем для діагностики і класифікації.

Найбільшого поширення серед цих методів набули дві моделі подання знань: модель семантичної мережі [6] і продукційна модель [5]. У першому випадку знання про ПО описують незалежно від рубрикатора в спеціальному вигляді – *тезаурусі*, який зв'язується з одним або більше рубрикаторами гнучкою системою відношень. Під *тезаурусом* розуміємо ієрархічну мережу понять і відношень між ними. Тезаурус може бути розроблений незалежно від будь-якої системи рубрикації. У ньому накопичують різні варіанти подання в тексті понять ПО (*дескрипторів*). Як варіанти (синоніми або еквіваленти) дескрипторів у тезаурусі містяться іменні та дієслівні групи, окремі іменники, прикметники або дієслова.

Тезаурус може бути розроблений в напівавтоматичному режимі. Наприклад, спочатку опрацьовується сукупність контенту великого обсягу за допомогою програм морфологічного і синтаксичного аналізу з метою виділення терміноподібних груп слів. Потім вибрані групи слів досліджують експерти, приймаючи рішення відносно того:

- чи може ця група входити в тезаурус (в цьому випадку вона стає терміном);
- чи є цей термін дескриптором або синонімом іншого дескриптора;
- як мають бути описані відношення цього терміна.

Крім того, в комплекс знань можуть також входити додаткові бази даних, наприклад: географічна база даних, що містить описи географічних об'єктів, база даних організацій, персоналу тощо. Тезаурус і бази даних мають одну структуру і складаються з таких частин:

1. Дескрипторів, які відповідають поняттям або конкретним об'єктам. Зазвичай дескриптором є іменник або іменна група.

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

2. Кожен дескриптор має сукупність текстових входів або синонімів. Текстовий вхід може бути іменником, прикметником або групою іменника. Одне слово може бути синонімом різних дескрипторів. Змістову неоднозначність усувають під час автоматичного опрацювання контенту.

3. Відношення між дескрипторами всередині кожної бази даних, наприклад: ширший термін (вище); вужчий термін (нижче); зв'язаний термін (асоціація); ціле для терміна (частина); частина для терміна (ціле).

4. Відношення між дескрипторами різних баз даних. У цьому випадку додається відношення – "рівність терміна", яке з'являється, коли бази даних утримують дескриптори, що відповідають одному поняттю або об'єкту.

Дескриптор D_1 міститься в *дескрипторному середовищі* дескриптора D, якщо між D_1 і D існує дескрипторне відношення або *транзитивна залежність*. Дескриптор D є *головним дескриптором* середовища.

Ієрархічна організованість тезауруса і наявність тезаурусних зв'язків дають змогу використовувати поняття середовища дескрипторів і головних дескрипторів (*опорних дескрипторів*) середовища для формування дескрипторних кущів, що використовують для автоматичної рубрикації текстів у цій технології. Загалом, комплекс знань є ієрархічною мережею, повноту і цілісність якої підтримують і відстежують експерти. Існує два типи подання рубрик послідовністю опорних дескрипторів у вигляді булевих нормальних форм [1]:

– диз'юнкція опорних дескрипторів $D_1 \vee D_2 \vee \ldots \vee D_n$;
– кон'юнкція диз'юнкцій опорних дескрипторів

$$(D_{11} \vee D_{12} \vee \ldots \vee D_{1n}) \wedge \ldots \wedge (D_{m1} \vee D_{m2} \vee \ldots \vee D_{mn}).$$

Для кожної рубрики рубрикатора може бути вибраний свій тип подання.

Після того, як для всіх рубрик рубрикатора встановлені зв'язки з відповідними опорними дескрипторами, автоматично визначаються рубрики для всіх дескрипторів тезауруса. Для кожного дескриптора створюється список відповідних рубрик із вказівкою того, в яку з диз'юнкцій рубрики входить цей дескриптор. Кожна рубрика в цій технології фіксує запит користувача, який описується за допомогою дескрипторів тезауруса. При цьому в тезаурусі міститься кущ дескрипторів, що відповідає цій рубриці, і встановлюється зв'язок між рубрикою і найвищим дескриптором (опорний дескриптор рубрики) в ієрархії дескрипторного куща. Одній рубриці може відповідати декілька опорних дескрипторів.

Подальший розвиток цієї технології полягає в наданні користувачеві можливості описувати рубрику природною мовою. Зміст процесу рубрикації згідно з цим підходом полягає у виділенні з контенту опорних дескрипторів і відношень між ними з подальшим зіставленням їх із описами рубрик. Ця технологія автоматичної рубрикації текстів дозволяє класифікувати різні типи текстової інформації, швидко налаштовуватися на різні рубрикатори і типи контенту, але має істотні обмеження у використанні, оскільки трудомісткість розроблення тезауруса достатньо висока, воно потребує великих тимчасових затрат (від декількох місяців до декількох років), крім того, тезаурус формується відповідно до тієї або іншої ПО, що робить неможливим використання одного тезауруса для класифікації текстів із різних ПО.

Основою методів, що використовують *продукційну модель* подання знань, є виділення з контенту концепцій (або понять), які заздалегідь описав експерт. Кожне поняття ПО експерт описує за допомогою особливої конструкції – *означення поняття*, що об'єднує в собі набір характерних для цього поняття слів та фраз. Означення поняття є виразом, записаним спеціальною мовою, що дозволяє об'єднувати ці слова і фрази за допомогою стандартних булевих функцій. У означенні поняття під час запису слів і фраз допускається використання символів-шаблонів (&, * тощо), що дозволяє відмовитися від процедури морфологічного аналізу, яка використовується для нормалізації лексики контенту. Оскільки опис понять експерт виконує вручну, то це не є особливою незручністю, проте значно підвищує продуктивність. На додаток до цих функцій у мові означення понять може бути передбачена можливість введення концептуальних обмежень, що полягає у вказанні дотримання порядку слів у тексті, відстані між словами тощо. Крім того, фразам в

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

означенні поняття можуть бути призначені експертні ваги, які показують, наскільки кожна з фраз характерна для певного поняття. Наведемо приклад визначення поняття золото: (gold (&n (reserve ! medal ! jewelry))).

Алгоритм рубрикації текстового контенту

Етап 1. Виділення понять із контенту з використанням даних із бази означень.

Крок 1. Рішення про наявність поняття в тексті приймають, розрахувавши істинність виразу, що визначає поняття, відносно цього контенту. Якщо вираз істинний, то вважається, що поняття наявне в тексті, тоді перехід до кроку 2.

Крок 2. Якщо в означенні поняття є експертні ваги, то обчислюється вага або ймовірність появи цього поняття в опрацьованому тексті з урахуванням частоти тієї фрази, що вживається в тексті повідомлення. Інакше перехід до кроку 3.

Крок 3. Якщо кінець тексту, то перехід до етапу 2, інакше до кроку 1.

Етап 2. Визначення належності контенту до конкретної рубрики.

Крок 1. Аналіз сформованої з контенту множини понять з можливими вагами (сортування, визначення підмножин з більшими вагами та їх концентрація в тексті).

Крок 2. Рішення приймається на основі *правил рубрикації*, що, як і визначення понять, експерт формулює заздалегідь з використанням мови правил. Вирази, записані мовою правил, подібні до конструкції if-then в алгоритмічних мовах програмування. Мова правил дозволяє основувати рішення на комбінації понять, що з'явилися в тексті. Вона дає змогу врахувати ймовірність появи, а також розміщення кожного поняття у тексті. Існує також можливість визначення довжини контенту.

Сукупність визначень понять і правил рубрикації утворює базу правил (рис. 4).

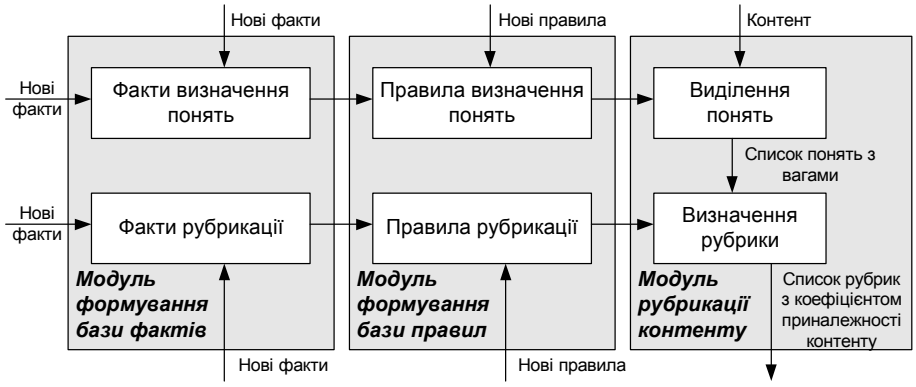

Рис. 4. Продукційна модель процесу рубрикації

Розроблення бази правил є дуже трудомістким процесом, що потребує залучення висококваліфікованих фахівців як із ПО, так і з інженерії знань. Суть цього процесу полягає в опрацюванні більшого масиву рубрикованого контенту, коли для кожної з рубрик виявляються статистичні закономірності, основані на частоті тих слів і фраз, що використовуються, а також на спільній частоті окремих із них. Отримані дані потім використовують експерти для виявлення характерних слів і фраз для опису понять і формування правил рубрикації.

Перевагами цього підходу є належна якість рубрикації і висока швидкодія на тих текстових потоках, для яких вони проектувалися. Основні недоліки таких систем, як і в попередньому випадку: висока трудомісткість і значні затрати, необхідні для розроблення системи; жорстка прив'язка баз знань і алгоритмів до ПО, конкретного рубрикатора, а також розміру і формату рубрикованих текстів. А більшість систем автоматичної рубрикації текстів потребують швидшої та дешевшої побудови.

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

Методи, основані на навчанні на прикладах. Системи автоматичної рубрикації, основані на навчанні на прикладах, розглядають рубрики як поняття, яким потрібно навчитися. Машинне навчання основане на прикладах текстів, які заздалегідь відрубрикували експерти вручну. Можна виділити статистичні й нейромережеві методи рубрикації. Ідея статистичної рубрикації полягає у визначенні міри відповідності термінологічного портрета контенту і термінологічного портрета рубрик на основі статистичних характеристик суб'єктів порівняння. Під термінологічним портретом контенту розуміють сукупність найважливіших термінів, що містяться в тексті контенту. Як показник важливості терміна в контенті найчастіше використовується частота його вживання. Під термінологічним портретом рубрики розуміють набір найхарактерніших для цієї рубрики термінів із їх вагами (термінологічним портретом рубрики часто є множина її характеристичних термінів і частоти їх вживання в рубриці). Отже, семантика рубрики задається однозначно лише її термінологічним портретом.

Зазначимо, що термінологічний портрет розглядають як окремий випадок тезауруса, що має простішу модель і що допускає його автоматичну побудову і коректування.

Формування термінологічних портретів кожної рубрики виконує експерт не вручну, а за допомогою однієї з технологій навчання рубрикатора. Роль експерта зводиться до формування для кожної рубрики *навчальної вибірки* – сукупності максимально коротких фрагментів текстів, що містять повне і мінімально надлишкове лінгвістичне наповнення однієї навчальної рубрики. Виділення характеристичних термінів для рубрики виконується автоматично, на основі їх ваг, які можна отримати в процесі аналізу навчальної вибірки [1]. Наприклад,

$$w_{tr} = \log \frac{N_r}{df_{tr}}, \tag{1}$$

де N_r – кількість контенту в навчальній вибірці, що належить рубриці r, df_{tr} – кількість контенту в навчальній вибірці, що належить рубриці r і що містить термін t [1]. Список характеристичних термінів рубрики впорядкований за зменшенням ваг термінів у ній.

Єдину модель для всіх рубрик одного рубрикатора подають у вигляді двовимірної матриці ваг $\{w_{tr}\}$ [1]. Рубрикація виконується за деяким вирішальним правилом, що враховує як важливість термінів у контенті, так і їх ваги для рубрик. Наприклад, контент належить рубриці r, якщо

$$\sum_t tf_t w_{tr} > k_r, \tag{2}$$

де tf_t – частота вживання терміна t в контенті, k_r – порогові значення для рубрики r. Значення лівої частини вказаного виразу використовують як кількісну оцінку релевантності контенту рубрикам.

Порогові значення для кожної з рубрик визначаються так, щоб, застосувавши вирішальне правило до всієї навчальної вибірки, до цієї рубрики зарахувати максимальну кількість релевантних і мінімальну кількість нерелевантних їй текстів. Обчислення може виконуватися як за допомогою різних математичних методів, так і емпірично. До переваг такого підходу належать:

– простота визначення семантики рубрики, що дає змогу організувати автоматичне навчання рубрик;

– універсальність підходу, що полягає в тому, що в такий спосіб може бути визначена семантика дуже широкого класу рубрик із будь-якої предметної області;

– наявність апарату кількісної оцінки релевантності документів рубрикам;

– висока швидкодія.

Головним недоліком цієї групи методів є нижча порівняно з методами, основаними на знаннях, якість рубрикації.

Основою *нейромережевих методів* рубрикації текстів є використання нейронної мережі (НМ) як навчального класифікатора. Нехай наявна вибірка прикладів текстів, кожен з яких помічений як релевантний або нерелевантний певній рубриці. Завдання НМ, яка навчена на цих прикладах, полягає у визначенні міри релевантності будь-якого нового контенту цій рубриці. Цей підхід передбачає, що семантика рубрики однозначно задається прикладами текстів, що належать їй.

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

Оскільки НМ оперує векторами, для подання контенту використовується одна з векторних моделей, наприклад [1, 7]:

$$(t_1,\ldots,t_D): t_i = \begin{cases} v \Leftrightarrow d_i \in T \\ 0 \Leftrightarrow d_i \notin T \end{cases}, i = \overline{1,D}, v = \frac{1}{\sqrt{N}},$$

де D – потужність словника; d_i – лексична одиниця із словника; T – текст, що розглядається як неврегульована множина лексичних одиниць; N – кількість $d_i \in T$.

Оскільки навчальна вибірка складається з прикладів із заздалегідь відомою приналежністю текстів рубрикам, то є сенс використовувати НМ, в яких реалізована парадигма з вчителем. Так, у [1, 7] пропонується використовувати ймовірнісну нейромережу (ІНМ). НМ має D входів і два виходи, один з яких відображає належність контенту, що подається, до класу релевантних запиту текстів, інший – до класу нерелевантних. На практиці є сенс використовувати лише перший, оскільки сума значень на виходах дорівнює 1. Схема процесу подана на рис. 5 і 6.

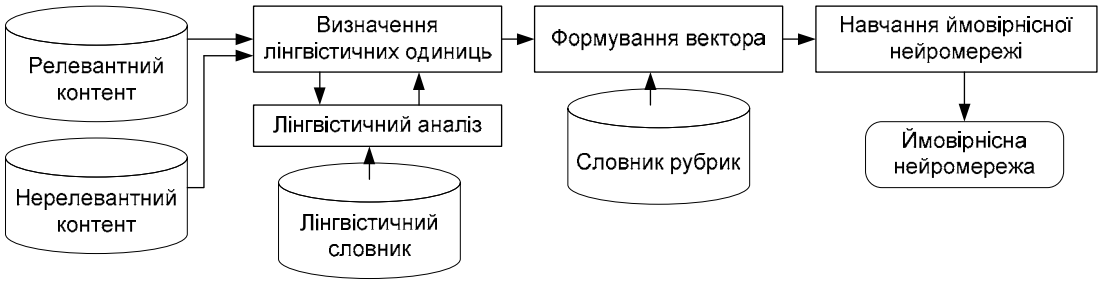

Рис. 5. Навчання процесу рубрикації

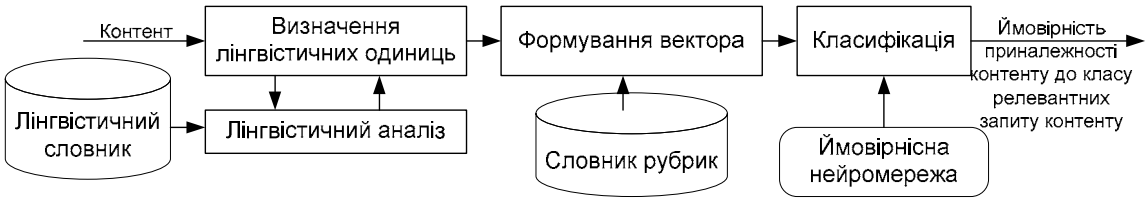

Рис. 6. Визначення ймовірності релевантності контенту рубриці

У словник рубрики можуть входити як прості, так і складні терміни. Він формується, як і в статистичних методах, з тією лише відмінністю, що ваги термінів надалі не використовуються. За якістю рубрикації нейромережеві методи рубрикації посередині між статистичними методами і методами, які основані на знаннях.

Серед основних недоліків нейронних мереж найчастіше називають два факти: експертам незрозуміло, як нейронна мережа працює; на навчання мережі потрібно дуже багато часу.

ІНМ вигідно відрізняється тим, що має: строге математичне обґрунтування (ІНМ – це оптимальний за Байєсом класифікатор); величезну (у тисячі разів більшу) порівняно з іншими нейромережевими парадигмами швидкодію. Характер розв'язаної задачі дає змогу істотно оптимізувати ІНМ, а також усунути залежність обсягу розрахунків від потужності словника [1]. Цей факт дозволяє повністю відмовитися від скорочення словника, небезпечного тим, що можуть бути відкинуті істотні для класифікації терміни. Загалом, вибір цієї нейромережевої парадигми дає змогу звести до мінімуму вказані недоліки.

Пошук текстової інформації

Моделі пошуку текстової інформації характеризуються чотирма параметрами:

- поданням контенту і запитів;
- критерієм змістової відповідності;
- методами ранжирування результатів запиту;
- механізмами зворотного зв'язку, що забезпечують оцінку релевантності користувачем.

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

Розглянемо найпоширеніші моделі пошуку з позиції перших трьох параметрів. *Булева модель* подає контент за допомогою набору термінів, наявних в індексі, кожен з яких розглядається як булева змінна. За наявності терміна в контенті відповідна змінна набуває значення *True*. Присвоєння термінам вагових коефіцієнтів не допускається. Запити формулюються як довільні булеві вирази, що зв'язують терміни за допомогою стандартних логічних операцій: *AND, OR* або *NOT*. Мірою відповідності запиту контенту слугує значення статусу вибірки (*RSV, retrieval status value*). У булевій моделі *RSV* дорівнює або 1, якщо для цього контенту обчислення виразу запиту дає *True*, або 0 – інакше. Всі контенти з *RSV* = 1 вважаються релевантними запиту.

Така модель проста в реалізації, застосовується в багатьох системах опрацювання Web-ресурсів. Вона дозволяє користувачам вводити в свої запити довільні складні вирази. Але ефективність пошуку зазвичай невисока. До того ж сортувати за рейтингом результати неможливо, оскільки всі знайдені контенти мають однакові *RSV*, а термінам не можна присвоїти вагові коефіцієнти. Інколи результати виглядають протиприродно. Наприклад, якщо користувач вказав у запиті десять термінів, зв'язаних логічною операцією *AND*, контент, що містить дев'ять таких термінів, у вибірку не потрапить. Для підвищення ефективності пошуку в ІПС часто застосовують зворотний зв'язок з користувачем. Як правило, система просить користувача вказати релевантність або нерелевантність декількох документів, які на початку списку виведення. Оскільки результати не сортуються за рейтингом, вибір контенту для такої експертної оцінки релевантності ускладнений.

Модель нечітких множин ґрунтується на теорії нечітких множин, що допускає (на відміну від звичайної теорії множин) часткову належність елемента тій чи іншій множині. Тут логічні операції перевизначені так, щоб врахувати можливість неповної належності множині, а опрацювання запитів користувача виконується аналогічно булевій моделі. Але ІПС на основі такої моделі виявляється практично настільки ж нездатною класифікувати отримані результати, як і системи на булевій моделі.

Строга булева модель і модель, що використовує методи теорії нечітких множин, потребують менших обсягів розрахунків (під час індексування та оцінювання відповідності контенту запиту), ніж інші моделі. Вони менш складні алгоритмічно і ставлять не дуже жорсткі вимоги до інших ресурсів, таких як дисковий простір для зберігання подань контенту.

Просторово-векторна модель основана на припущенні, що сукупність контенту можна подати набором векторів у просторі, який визначається базисом, з *n* нормалізованих векторів термінів. Значення першого компонента вектора, що подає контент, відображає вагу терміна в ньому. Запит користувача також подається *n*-вимірним вектором. Показник *RSV*, що визначає відповідність контенту запиту, задається скалярним добутком векторів запиту і контенту. Що більше *RSV*, то вищий, ніж релевантність, контент запиту. Перевага такої моделі в її простоті. Вона дає змогу легко реалізувати зворотний зв'язок для оцінювання релевантності користувачем. Водночас доводиться жертвувати виразністю специфікації запиту, притаманною булевій моделі.

Імовірнісні моделі. У просторово-векторній моделі вважається, що вектори термінів, ортогональні та наявні взаємозв'язки між термінами не повинні братися до уваги. Крім того, в такій моделі не специфікується міра відповідності "запит – контент" і вона оцінюється доволі довільно. Імовірнісна модель враховує всі взаємозалежності та зв'язки термінів, а також визначає такі основні параметри, як ваги термінів запитів і форма відповідності "запит – контент". Ця модель ґрунтується на двох основних параметрах: *Pr(rel)* та *Pr(nonrel*), тобто на ймовірності релевантності та нерелевантності контенту запиту користувача, які розраховуються на основі імовірнісних вагових коефіцієнтів термінів і фактичної наявності термінів у контенті. Релевантність є бінарною властивістю, тому *Pr(rel)*=1–*Pr(nonrel)*. У цій моделі застосовують два вартісні параметри: a_1 і a_2. Вони характеризують відповідно витрати, пов'язані з введенням у результат нерелевантного контенту і пропуском релевантного контенту. Ця модель потребує визначення ймовірності входження терміна в релевантні та нерелевантні частини сукупності контенту, оцінити які доволі складно. Тим часом вона виконує важливу функцію, пояснюючи процес пошуку і пропонуючи теоретичне обґрунтування методів, що застосовувалися раніше емпірично (наприклад, введення деяких систем визначення вагових коефіцієнтів термінів).

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

Методи введення зворотного зв'язку з користувачем. На відміну від середовища баз даних, у контентно-пошуковій підсистемі немає чіткого подання контенту і призначених для користувача запитів. Користувачі зазвичай починають з неточного і неповного запиту, а отже – з низької ефективності пошуку, поступово уточнюючи його методом ітерацій. Система підтримує зворотний зв'язок із користувачем, даючи змогу оцінити релевантність контенту, знайденого за первинним запитом. Такий підхід дозволяє підвищити ефективність пошуку. Для спрощення подання зворотного зв'язку використовують просторово-векторну модель пошуку, а користувачеві надана можливість просто відзначити: релевантний контент чи ні.

Множина контенту, що вважається релевантною, формує позитивний зворотний зв'язок, а множина контенту, що розглядається як нерелевантний, – негативний. Існують два основні підходи до використання такого зворотного зв'язку: *модифікація запиту* і *модифікація подання контенту*. Методи, що модифікують подання запиту, впливають лише на поточний сеанс, але ніяк не позначаються на опрацюванні інших запитів. Методи, основані на модифікації подання контенту, впливають і на ефективність пошуку в подальших запитах.

Базове припущення, на яке спирається методологія зворотного зв'язку, полягає в тому, що контент, релевантний деякому призначеному для користувача запиту, подібний до іншого у векторному просторі, тобто відповідні вектори в якомусь сенсі "подібні" один на одного. Використання зворотного зв'язку в механізмах пошуку інформації потребує описовішого і семантично багатшого подання контенту, чогось, що отримують в результаті індексування лише назв або рефератів контенту. Один з можливих способів – індексування всього контенту. Просторово-векторну модель неважко адаптувати до всіх методів пошуку зі зворотним зв'язком, тоді як імовірнісна модель потребує спеціальних розширень.

Модифікація подання запиту. Існують три способи підвищення ефективності пошуку модифікацією подання запиту. Перший – модифікація ваг термінів – передбачає коректування ваг термінів у запиті, що здійснюється об'єднанням вектора запиту і векторів, що подають контент, які отримали позитивну оцінку (позитивний зворотний зв'язок). Разом з цим можливе додаткове коректування за рахунок віднімання векторів, що входять у множину з негативним зворотним зв'язком. Переформований у такий спосіб запит повинен повертати додатковий релевантний контент, аналогічний до того, що потрапив у множину з позитивним зворотним зв'язком. Цей процес є рекурсивним доти, доки якість вибірки та кількість контенту в ній не досягнуть прийнятного рівня. Результати експериментів показують, що позитивний зворотний зв'язок змістовніший та ефективніший. Причина в тому, що контент з множини з позитивним зворотним зв'язком зазвичай однорідніший, ніж той, що формує негативний зворотний зв'язок. Один з ефективних методів використовує весь контент з позитивним зворотним зв'язком, але для віднімання із запиту бере лише ті вектори з негативним зворотним зв'язком, які володіють найбільшим рангом нерелевантності.

Інший метод, названий *методом розширення запиту*, модифікує вихідний запит, додаючи до нього нові терміни. Ці терміни вибирають з контенту з позитивним зворотним зв'язком і сортують на основі їхніх ваг. До запиту додається заздалегідь задана кількість термінів із початку відсортованого списку. Експерименти показують, що останні три методи сортування дають найкращі результати і додавання обмеженої кількості найважливіших термінів переважає врахування всіх термінів. У разі приєднання до запиту більше ніж 20 додаткових термінів ефективність практично не збільшується.

У деяких випадках ці два методи не дають задовільних результатів через неоднорідність контенту із позитивним зворотним зв'язком (оскільки не утворюють компактного кластера в просторі контенту) або через "вкраплення" нерелевантного контенту в множині релевантних. Один зі способів виявлення вказаної ситуації – кластеризація контенту із позитивним зворотним зв'язком та виявлення декількох однорідних кластерів. Такий метод є *розщепленням запиту*. Якщо множину контенту кластеризують, то запит розділяють на підзапити так, щоб кожен підзапит подавав один кластер. Потім можна надбудувати вагові коефіцієнти термінів підзапиту або розширити його за допомогою методів, які описані вище.

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

Модифікація подання документів. Цей підхід передбачає налаштування векторів контенту на основі зворотного зв'язку, та є кластеризацією, орієнтованою на користувача. Суть методу – корекція вагових коефіцієнтів векторів, що потрапили у вибірку, щоб наблизити їх до вектора запиту. Водночас ваги знайденого нерелевантного контенту модифікуються так, щоб віддалити їх від вектора запиту. Але робити це треба обережно – окремі зсуви контенту мають бути невеликі, оскільки оцінювання релевантності користувачем є суб'єктивним. Детальніше моделі пошуку і механізми зворотного зв'язку розглянуто в [1, 4].

Оцінювання якості систем опрацювання Web-ресурсів

Раніше наголошувалося, що в ПОК і ПРК відбивається лише основний зміст повідомлень, що надходять у скороченому вигляді. Тому метод інформаційного пошуку, оснований на подібності ПРК із ПОК, не може повністю забезпечити пошук всієї множини контенту, що відповідає інформаційному запиту. Це призводить до того, що частина контенту, яка відповідає запиту, тобто релевантних йому, залишається невиданою користувачеві. Водночас у множині виданого контенту є такий, які не відповідає запиту, тобто не є релевантним. Фактично у будь-якій реальній контентно-пошуковій підсистемі є два основні типи помилок:

– помилки 1-го роду (або пропуск мети): невидання користувачеві фактично релевантного його запиту контенту;

– помилки 2-го роду (або помилкова тривога, інакше шум): видавання користувачеві нерелевантного контенту, який не відповідає поставленому запиту.

Наявність помилок 1-го і 2-го роду в реальній системі зумовлює розподіл всього масиву контенту системи щодо запиту на чотири підмасиви (табл. 2–3).

Таблиця 2

Розподіл масиву контенту

Масиви	Видані	Невидані
Релевантні	*A* – виданого релевантного контенту	*C* – невиданого релевантного контенту
Нерелевантні	*B* – виданого нерелевантного контенту	*D* – невиданого нерелевантного контенту

Існують такі показники ефективності контентно-пошукової підсистеми, де *a* – кількість виданих релевантних документів; *b* – кількість виданих нерелевантних документів; *c* – кількість не знайдених релевантних документів; *d* – кількість не знайдених нерелевантних документів.

Таблиця 3

Показники ефективності контентно-пошукової підсистеми

№	Коефіцієнт	Характеризує частину	Формула
1	*повноти p*	виданого релевантного контенту у всьому масиві релевантного контенту	$p = \frac{a}{a+c}$
2	*точності* n	виданого релевантного контенту у всьому масиві виданого контенту	$n = \frac{a}{a+b}$
3	*шуму e*	виданого нерелевантного контенту у всьому масиві виданого контенту	$e = \frac{b}{a+b} = 1 - \mathrm{n}$
4	*осаду q*	виданого нерелевантного контенту у всьому масиві нерелевантного контенту	$q = \frac{b}{b+d}$
5	*специфічності k*	не знайденого нерелевантного контенту у всьому масиві нерелевантного еонтенту	$k = \frac{d}{b+d}$

Часто для зручності перераховані показники вимірюють у відсотках, тобто у вказаних формулах з'являється додатковий множник 100 %.

Для оцінювання якості реальних систем найчастіше використовуються лише коефіцієнти повноти і точності. Зрозуміло, що і точність пошуку, і його повнота залежать не лише від

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

властивостей пошукової системи, але і від правильності побудови конкретного запиту, а також від суб'єктивного уявлення користувача про те, що таке потрібна йому інформація. Але за бажанням можна розрахувати і середні значення повноти та точності для конкретної системи, протестувавши її на еталонній базі документів. Ефективна пошукова система має більшу повноту і точність, бажано – 100 %, тобто знаходить весь потрібний контент і нічого зайвого. Але стовідсоткова якість пошуку неможлива, бо на фіксованому рівні потужності пошукового засобу всі спроби поліпшити один з цих параметрів призводять до погіршення іншого (рис. 7).

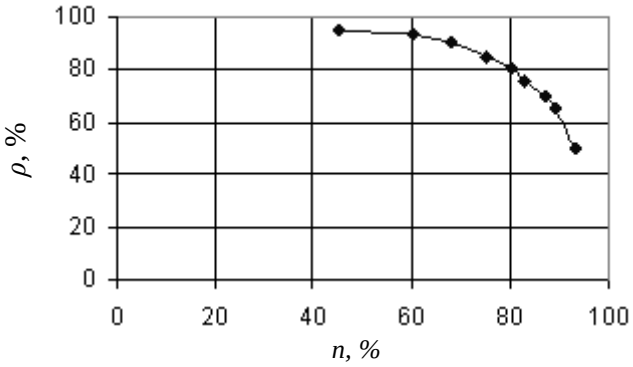

Рис. 7. Приклад залежності між коефіцієнтами повноти і точності

Поряд з перерахованими показниками, які основані на зв'язаності релевантності та видавання, доцільно використовувати також й інші показники ефективності, що зазвичай і роблять на практиці. До основних із них належать:

– швидкодія контентно-пошукової підсистеми (інтервал часу між моментом формулювання запиту й отриманням відповіді на нього);

– пропускна здатність (оцінюється кількістю контенту, що вводяться, і кількістю відповідей за одиницю часу за заданих значень коефіцієнтів повноти і точності);

– продуктивність (оцінюється кількістю користувачів системи і частотою їх звертань);

– надійність роботи (оцінюється ймовірністю того, що система виконуватиме свої функції за заданих умов протягом необхідного часу);

– тип запитів, що обслуговуються системою.

Семантичне моделювання в базах даних

Спочатку в теорії БД основна увага приділялася засобам ефективної організації даних і маніпулювання ними. В результаті виникли три основні моделі даних: ієрархічна, реляційна і мережева. При цьому явно або неявно припускалося, що запропоновані засоби достатньо універсальні для подання знань або інформації про будь-які ПО. Так, і сьогодні прихильники реляційної моделі, що набула найбільшого поширення, часто стверджують, ніби таблична форма подання даних є найзручнішою та інтуїтивно зрозумілою проектувальникові.

Але проектування бази даних у термінах цих моделей часто зводиться до дуже складного і незручного для проектувальника процесу, оскільки ці моделі не містять достатніх засобів подання змісту даних. Семантика реальної ПО повинна незалежним від моделі способом відображатися в свідомості проектувальника. Це призводить до уповільнення процесу розроблення БД і є джерелом потенційних помилок.

З цієї причини останніми роками розвивається такий напрям, як семантичне, або концептуальне, моделювання у базах даних. Його основна мета – організація інтерфейсу проектувальника, а також кінцевого користувача з інформаційною системою на рівні уявлень про ПО, а не на рівні структур даних. Інтерес до цього напряму зріс у зв'язку з розвитком засобів автоматизованого проектування БД на основі CASE-технологій.

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

Сьогодні визначився основний підхід до розв’язання задач семантичного моделювання у базах даних. Він полягає у виділенні двох рівнів моделювання: рівня концептуального моделювання ПО і рівня моделювання власне бази даних [8].

На верхньому рівні здійснюється перехід від неформалізованого опису ПО та інформаційних потреб кінцевого користувача (акторів) до їх формального виразу за допомогою спеціальних мовних засобів. На нижньому – перетворення концептуальної моделі ПО на схему БД і нормалізація схеми БД. Так, у ПО опрацювання Web-ресурсів виділяють таких акторів (рис. 8):

– адміністратор як творець сайта може керувати всіма матеріалами, а також всім, що є на сайті;
– відвідувач: відвідувач сайта, якого цікавить нерухомість.

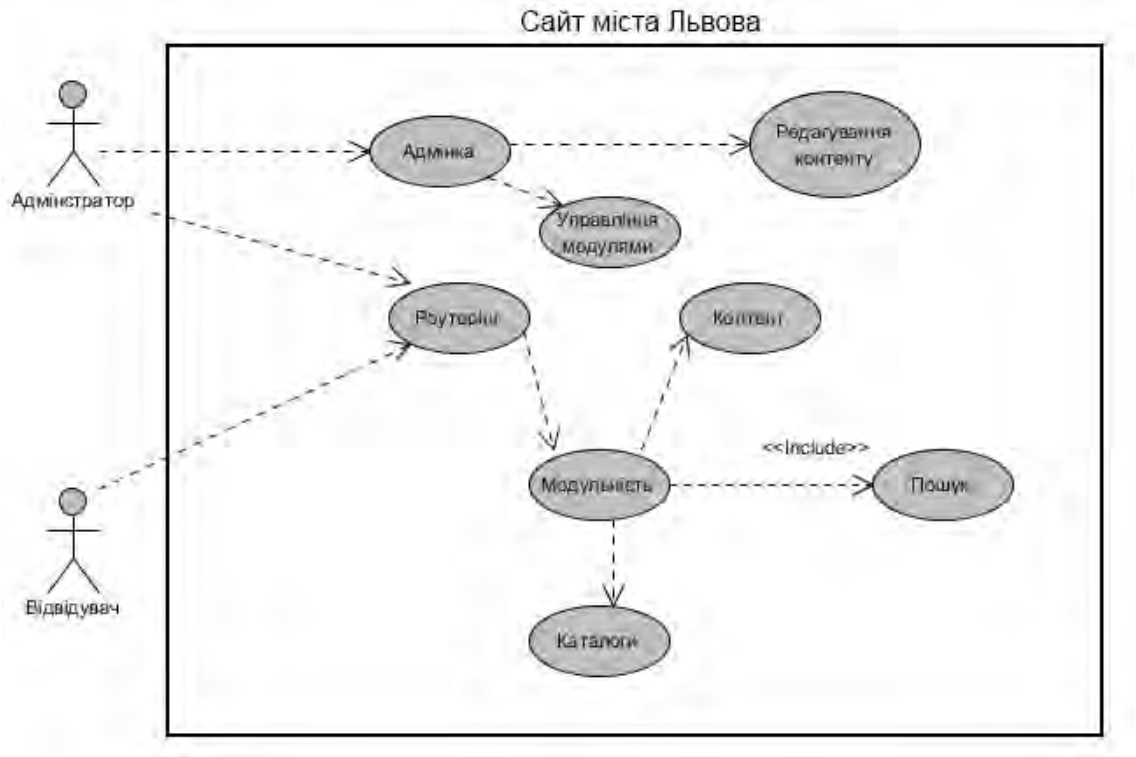

Рис. 8. Діаграма варіантів системи опрацювання Web-ресурсу

Основні вимоги до системи опрацювання Web-ресурсів.

1. Всі каталоги на сайті перевіряє модератор, додаючи на сайт; якщо інформація не достовірна, не чітка, тоді модератор просто видаляє матеріал.

2. Якщо всі поля заповнені, тоді матеріал успішно додається в базу даних, а також в індекс для пошуку.

3. Якщо не заповнені всі поля для публікації матеріалу, то матеріал просто не додається в БД сайта.

4. Матеріал тоді просто переводиться в архів, забороняється доступ до індексації його пошуковиками, і видаляється з індексу пошуку.

Основні взаємодії між об'єктами нашого проекту опрацювання Web-ресурсів подано на рис. 9 діаграмою комунікації (англ. Communication diagram). Вона відображає організацію інтерфейсу проектувальника, а також кінцевого користувача з системою на рівні уявлень про ПО. На рис. 10 подана діаграма класів системи опрацювання Web-ресурсів, де відображено основні відносини між класами (об’єктами) та їх екземплярами системи опрацювання Web-ресурсів.

На рис. 11 діаграма об'єктів (Object diagram) відображає процес додавання адміністратора. Rights вибирається з ComboBox’а на сайті. А Register заповнюється через тег input, крім параметра Register_id. Він є інкрементом на +1.

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

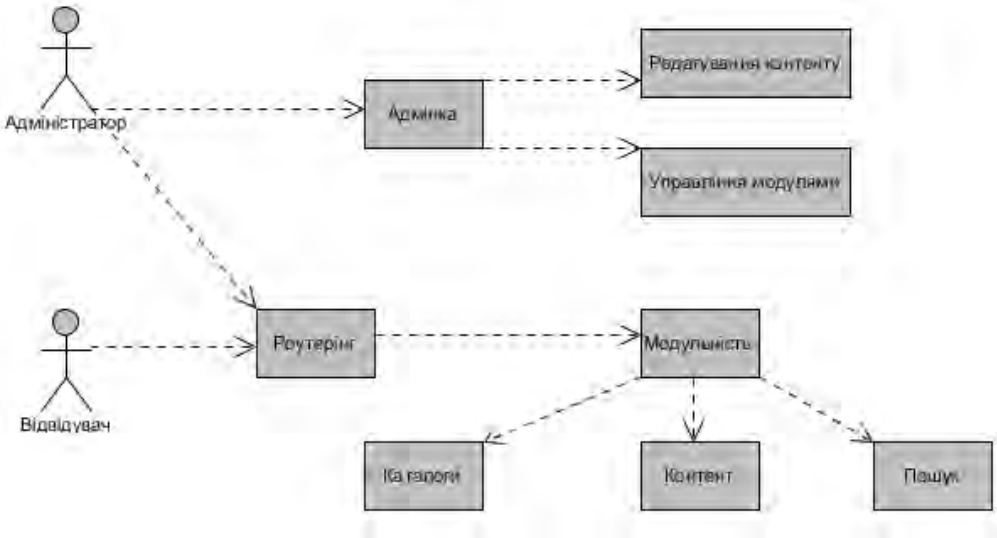

Рис. 9. Діаграма комунікації (communication diagrams)

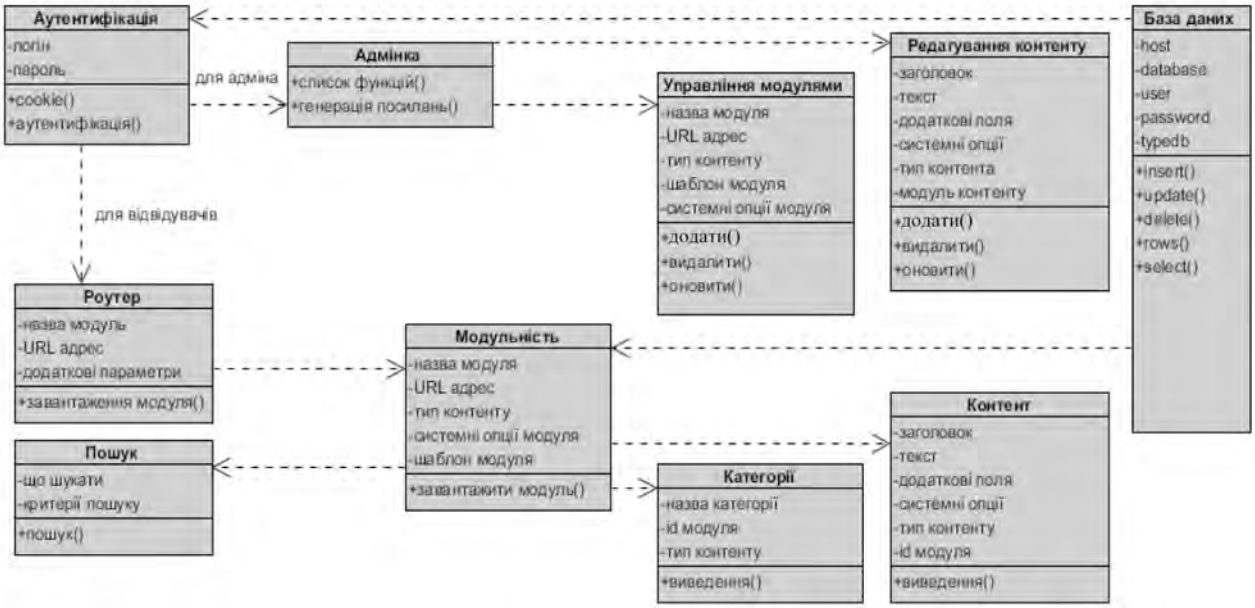

Рис. 10. Діаграма класів (class diagram)

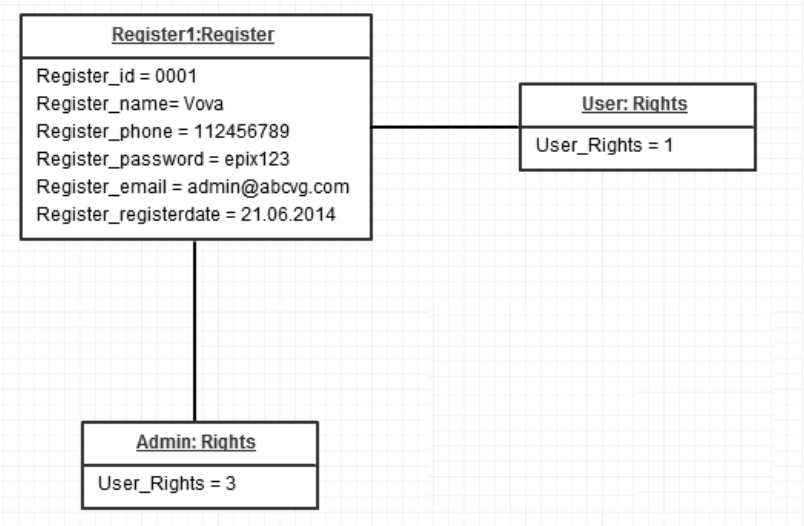

Рис. 11. Діаграма об'єктів (object diagram)

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

Діаграма послідовності процесу опрацювання Web-ресурсів відображає взаємодії основних об'єктів системи, впорядкованих за часом, зокрема, задіяні об'єкти та послідовність відправлених повідомлень між цими об'єктами (рис. 12).

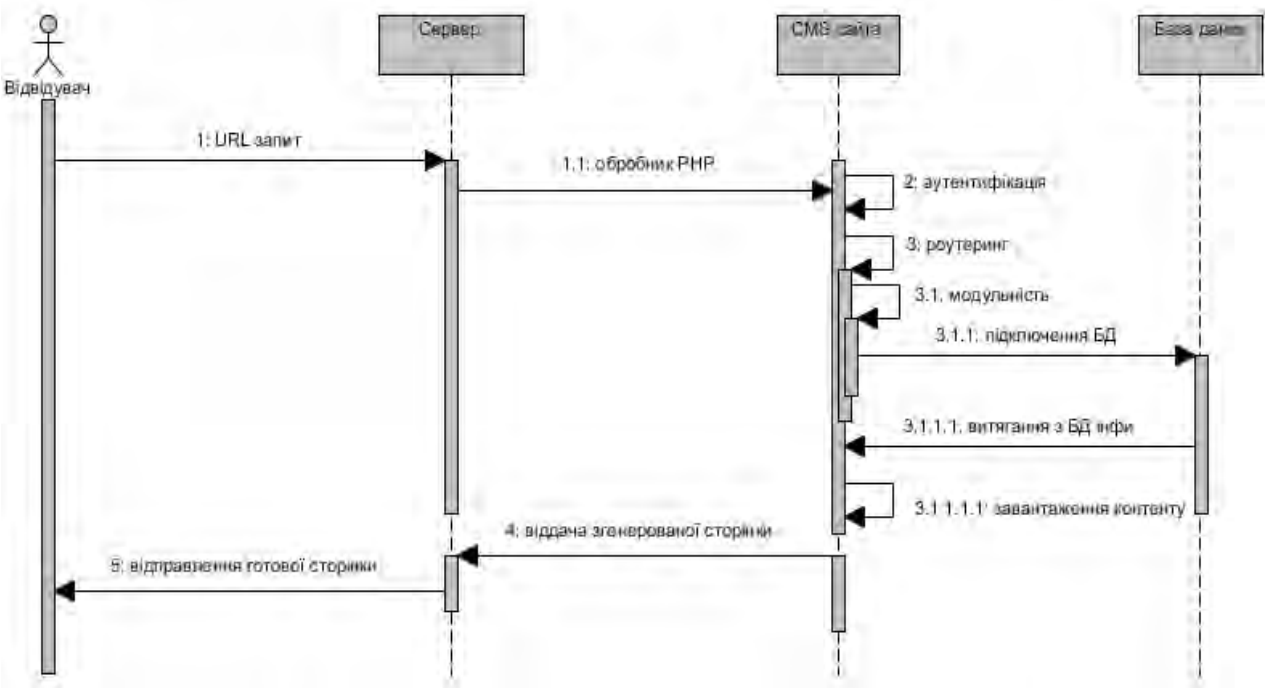

Рис. 12. Діаграма послідовності (sequence diagram)

Діаграма діяльності (рис. 13) розкриває основну діяльність ІС – опрацювання Web-ресурсів.

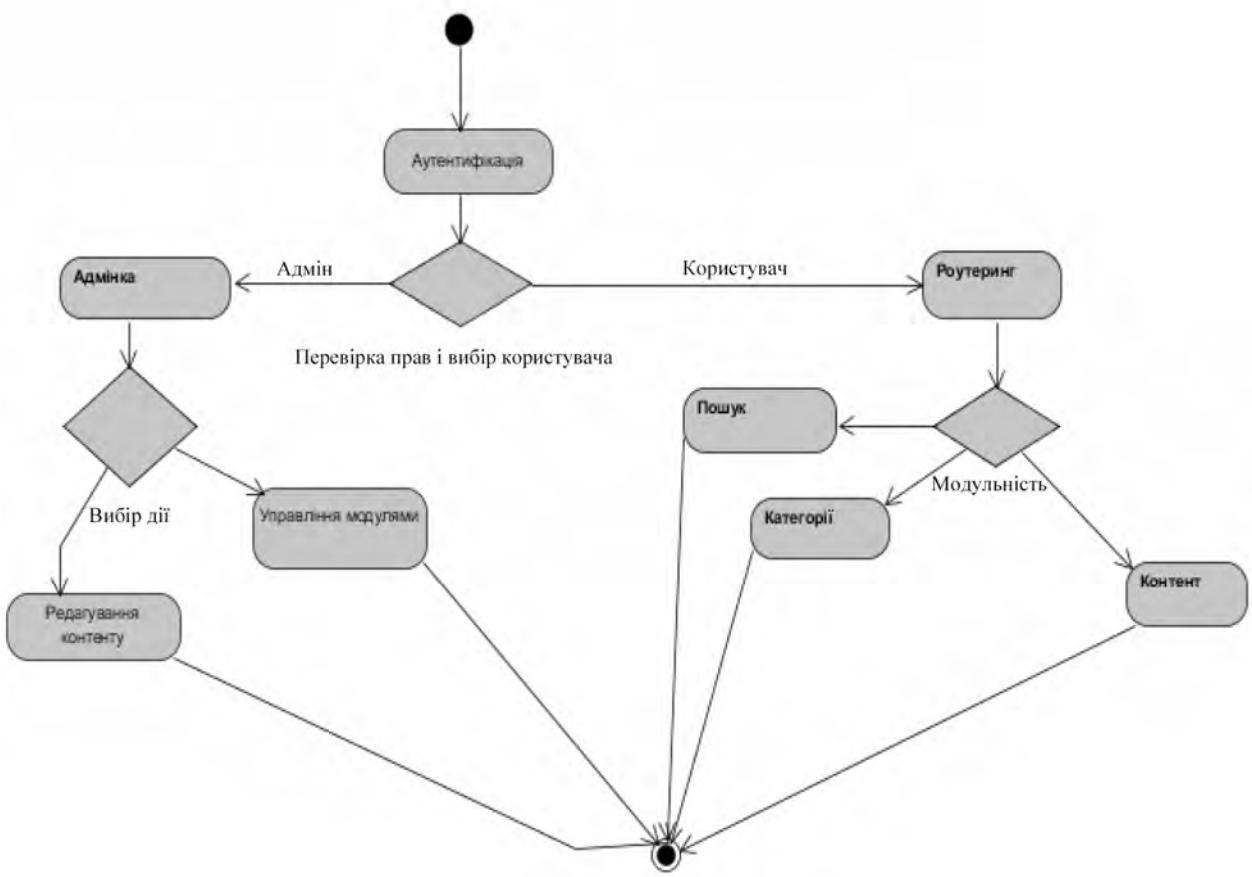

Рис. 13. Діаграма діяльності (Activity diagram)

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

На рис. 14 подана діаграма огляду взаємодії основних об’єктів системи опрацювання Web-ресурсів у разі з’єднання з сервером і завантаження сайта. Діаграма станів на рис. 15 описує основні поведінки окремих найважливіших об'єктів системи опрацювання Web-ресурсів.

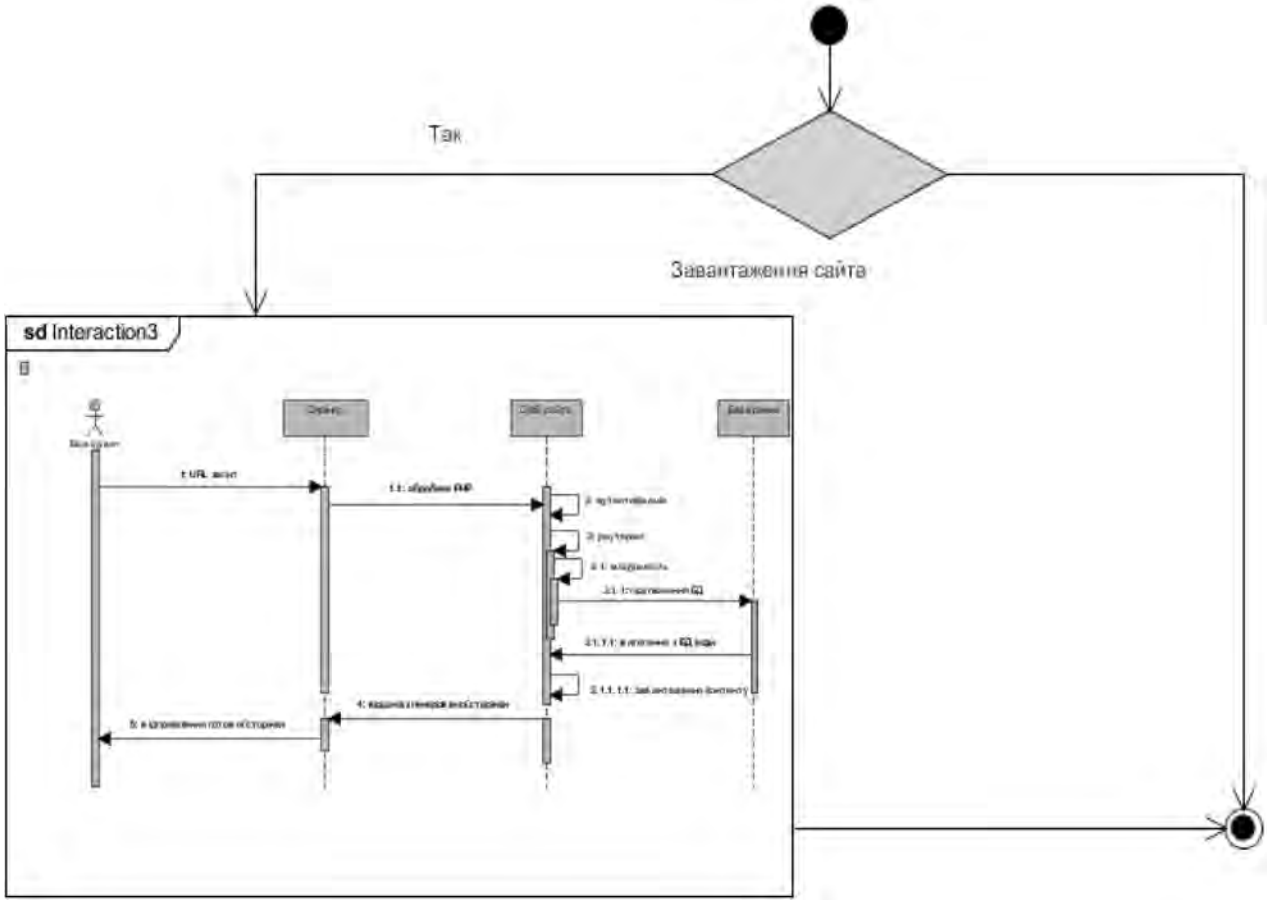

Рис. 14. Діаграма огляду взаємодії (interaction overview diagram)

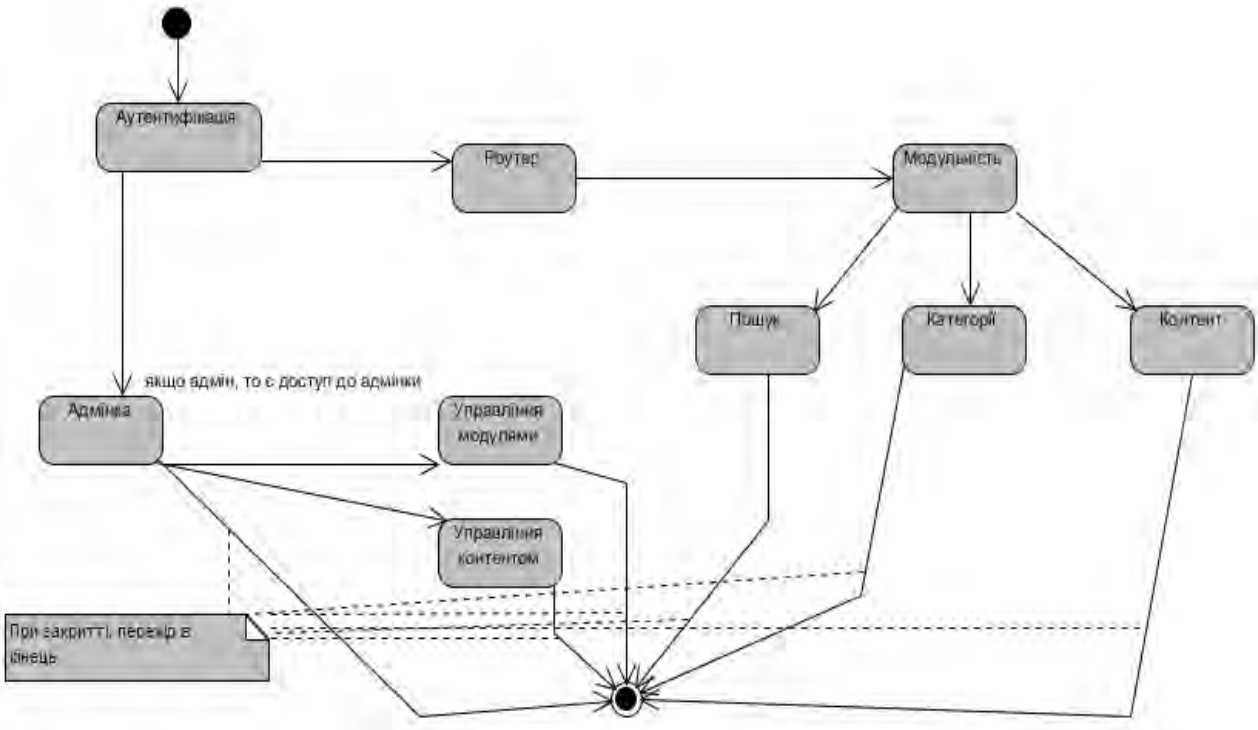

Рис. 15. Діаграма станів (state machine diagram)

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

Діаграма розгортання застосовується для подання загальної конфігурації та топології розподіленої програмної системи опрацювання Web-ресурсів і містить зображення розміщення компонентів у окремих вузлах системи (рис. 16). Діаграма розгортання показує наявність фізичних сполук – маршрутів передавання інформації між апаратними пристроями, які задіяні в реалізації системи. Діаграма компонентів на рис. 17 відображає залежності між компонентами програмного забезпечення опрацювання Web-ресурсів та структури сервера, враховуючи компоненти вихідних кодів, бінарні компоненти, та компоненти, що можуть виконуватись. Модуль програмного забезпечення поданий як компонента. Деякі компоненти існують під час компіляції, деякі – під час компонування, а деякі під час роботи програми.

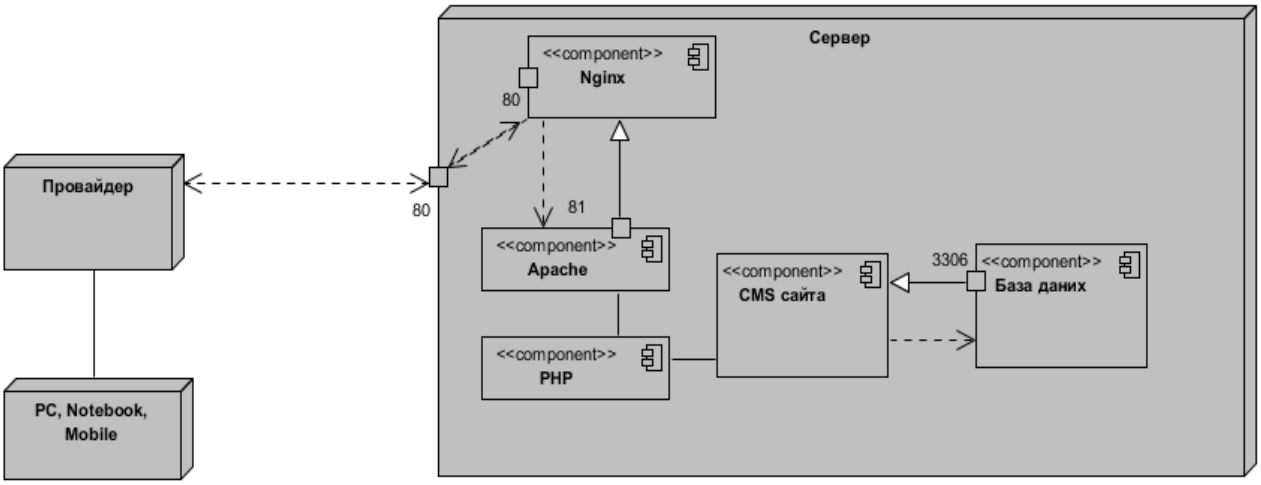

Рис. 16. Діаграма розгортання (deployment diagram)

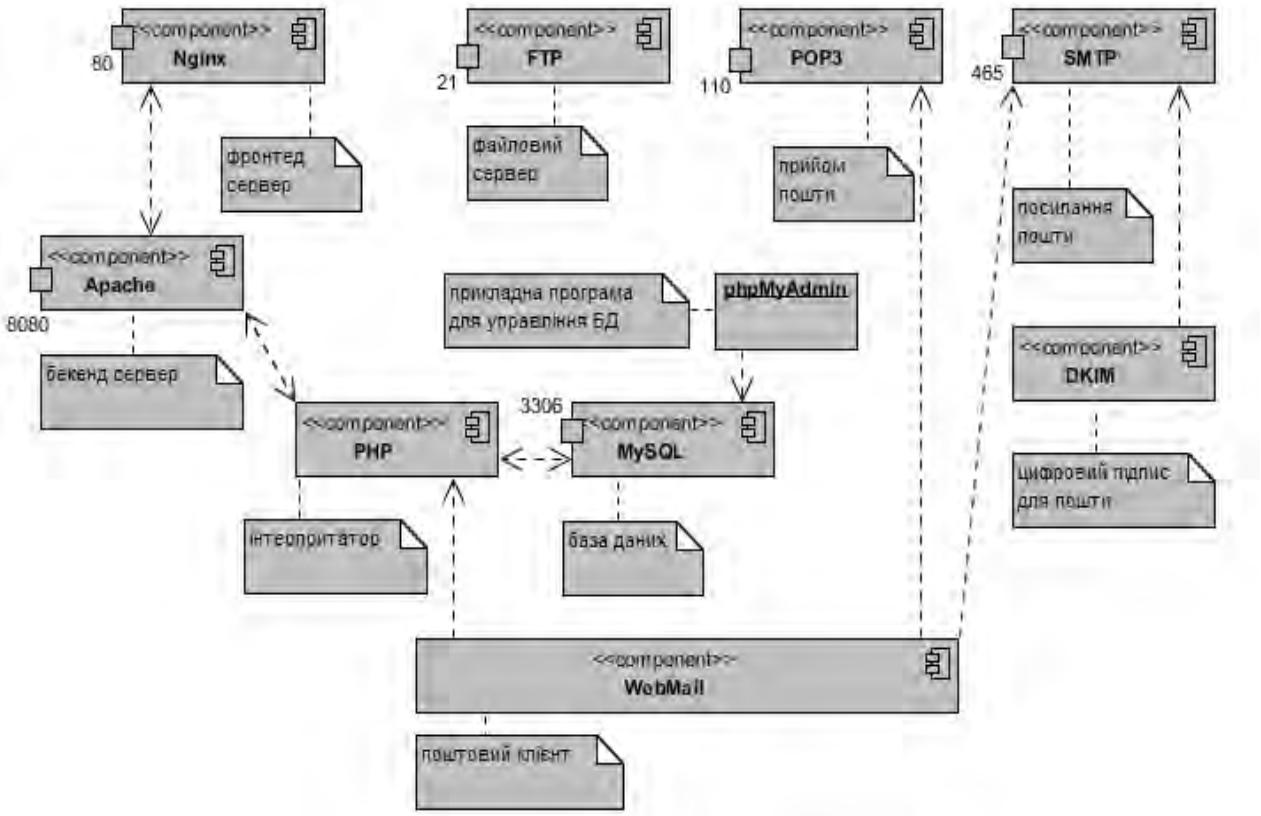

Рис. 17. Діаграма компонентів (component diagram)

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

На рис. 18 діаграми пакетів відображають організацію елементів у групи з якою-небудь ознакою з метою спрощення структури та організації роботи з моделлю системи опрацювання Web-ресурсів.

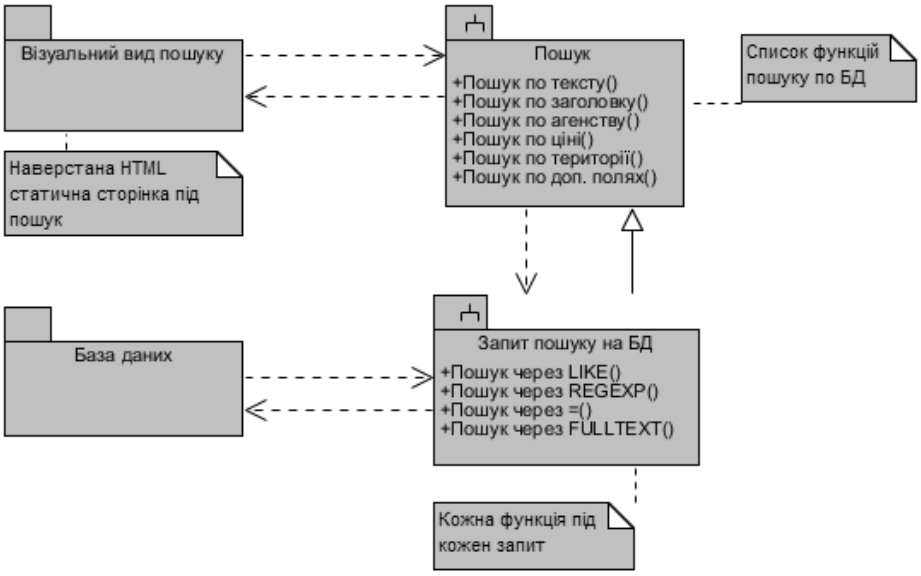

Рис. 18. Діаграма пакетів (package diagram)

Розроблена діаграма пакетів, що відображає, як програмно відбуватиметься пошук по сайту та БД. На діаграмі синхронізації (рис. 19) наведено альтернативне подання діаграми послідовності, що явно показує зміни стану на лінії життя із заданою шкалою часу, тобто зображено швидкість роботи сервера. Корисна в додатках реального часу.

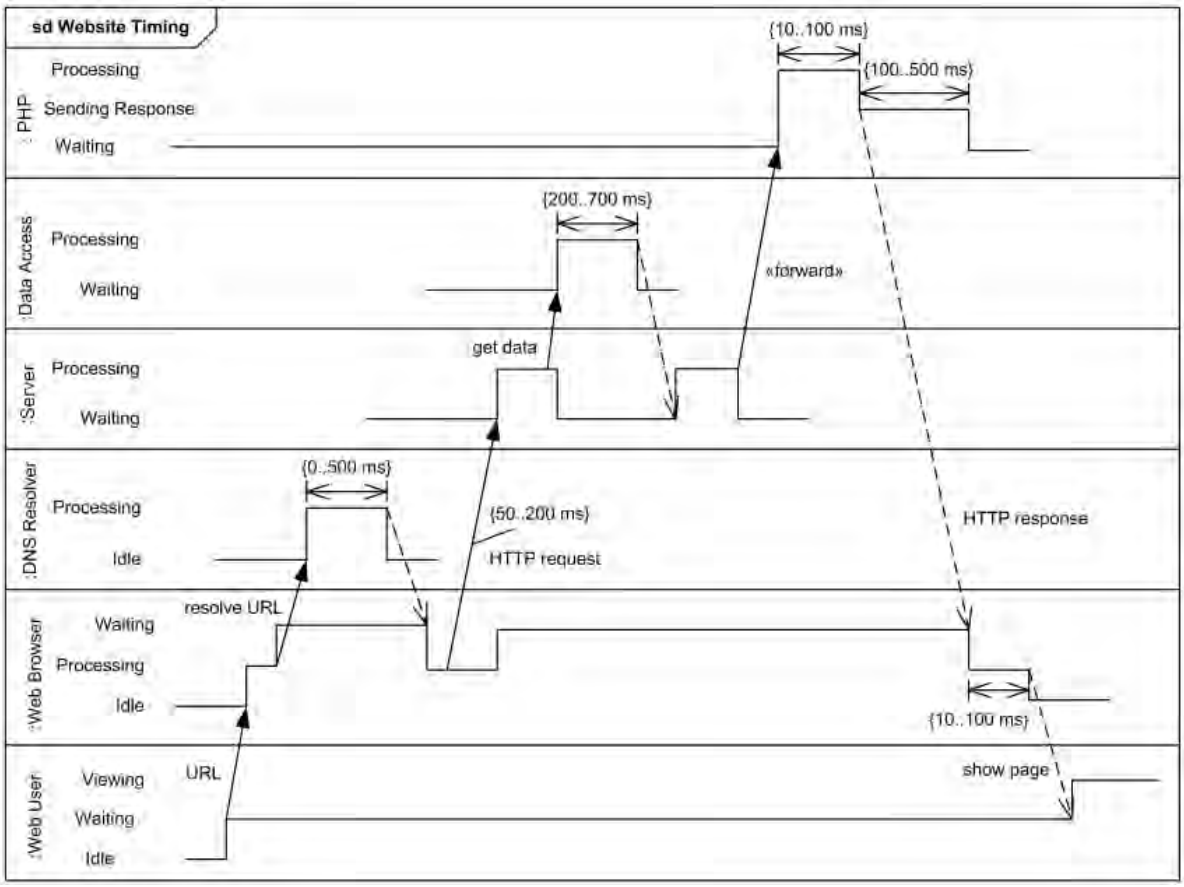

Рис. 19. Діаграма синхронізації (timing diagram)

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

Діаграма композитної/складової структури (рис. 20) демонструє внутрішню структуру класів і взаємодію елементів (частин) внутрішньої структури класу. Діаграми композитної структури використовують спільно з діаграмами класів.

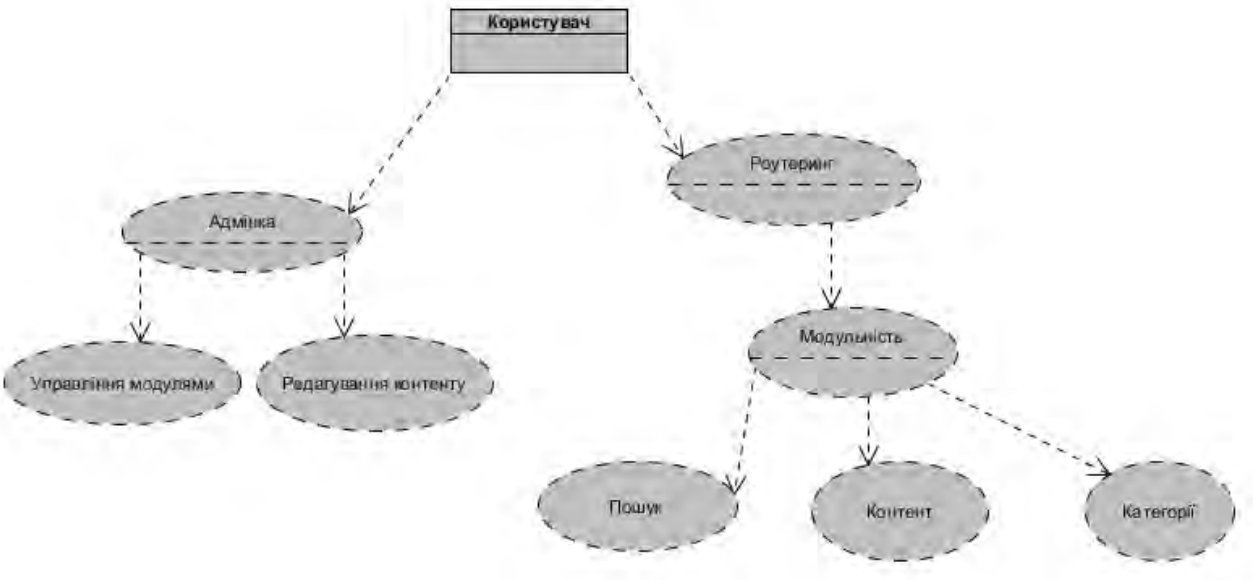

Рис. 20. Діаграма композитної/складової структури (composite structure diagram)

Зазначимо, що за допомогою відомих підходів визначення кількості інформації розв'язують задачі з позиції аналізу загальних властивостей об'єкта і не враховують позиції особи, що приймає рішення (ОПР, в нашому варіанті це користувач Web-ресурсу) стосовно досліджуваного об'єкта (Web-ресурсу). Однак позиції різних ОПР можуть принципово різнитися. Наприклад, один із них аналізуватиме множину ситуацій ризику, дотримуючись крайньої обережності, а інший – буде прихильником раціонального ризику. Тому виникла потреба визначення кількості та якості інформації не тільки для опису наявних властивостей і особливостей досліджуваного об'єкта, але й з погляду формування і досягнення цілей ОПР на основі його бачення як необхідних властивостей та особливостей досліджуваного об'єкта, так і шляхів та засобів їх реалізації. Для цього ОПР повинен мати певний рівень інформованості про об'єкт (табл. 4).

Таблиця 4

Властивості та особливості Web-ресурсу

№	Назва	Повнота	Достовірність	Своєчасність
1	2	3	4	5
1	*Новини*	{ заголовок, текст, картинка, дата новини}	{новини будуть рерайтитись з джерел, в яких рейтинг понад 10}	{інформація переважно актуальна на сьогодні та ще два дні вперед}
2	*Оголошення*	{картинка, заголовок, телефони, електронні пошти, опис оголошення, тип оголошення (якщо нерухомість, то такі пункти: операція, район, адреса, тип, стан, кімнат, поверх, загальна площа, житлова площа, площа кухні, кімнат, гостинка, комуналка, кімната, малосімейка, телефон, електронна пошта, веб-сайт, ціна; якщо робота, то такі: ТОВ вакансії, зарплатня, зайнятість, досвід роботи, освіта, вік, стать)}	{подання ПІБ автора оголошення, підтвердження пошти під час подання оголошення}	{інформація актуальна 2–3 місяці}
3	*Афіша*	{жанр, країна, рік, актори, режисер, час перегляду, картинка, вартість перегляду, анонси фільмів, список доступних кінотеатрів, де проходить перегляд}	{інформація буде братись з офіційних сайтів, кінотеатрів через API}	{актуально не більше ніж на тиждень}

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

Продовження табл. 4

1	2	3	4	5
4	*Погода*	{список з часом (вранці, ввечері, вдень, вночі), кількість градусів за Цельсієм і Фаренгейтом, опис події (ясно, без опадів, вітер, рт. ст., град, дощ, вологість, туман)}	{погода буде братись з сайта метеослужби}	{один день}
5	*Карта міста*	{картинка карти міста, з масштабом}	{карта буде братись з Google Maps}	{актуально на 2–3 місяці}
6	*Довідкова*	{список закладів, телефони та час, коли можна дзвонити}	{буде перевірено кожен телефон, і раз на місяць обдзвін всіх номерів у каталозі}	{може бути актуальна декілька років}
7	*Фотогалерея*	{список альбомів, у альбомі містяться фотографії, їх опис, exif дані, назва, дата завантаження, теги}	{завжди мають бути фотографії певного міста, а не інших}	{немає обмежень щодо дати, оскільки це є склад фотографій}
8	*Транспорт*	{карта, а на ній нанесено всі маршути громадського транспорту в місті}	{тільки ті маршутні таксі, які є в місті}	{актуально, доки маршут не поміняють}
9	*Курс валют*	{список банків, назва валюти, ціна продажу, ціна купівлі}	{інформація буде збиратись з сайта міжбанка}	{актуально на 1–2 години}
10	*Історія*	{хронологія міста, опис важливих подій для міста, і їх дати}	{історія/хронологія має бути написана спеціально для певного міста, а не інших}	{актуально може бути протягом років}

У разі неповноти та нечіткості інформації інформацію редагує до належного стану модератор чи адміністратор сайта.

Максимальний час вирішення проблеми – до 30 хв (пошук джерел, порівняння інформації, редагування матеріалу). Якщо проблеми є на серверному рівні: DDoS атаки, просідання сервера, погана віддача, висока відвідуваність, тоді проблеми вирішують так:

1. DDoS атаки – визначається маска підмережі в мережі, IP адреса, з якої ведуть DDoS атаку, через файрвол ставиться редирект на внутрішній IP того, хто атакує: 127.0.0.1 чи 192.168.0.1.
2. Просідання сервера – тільки його заміною, бо проблема в жорсткому диску є.
3. Погана віддача – підключення сервера до іншого провайдера.
4. Купівля або оренда нового потужного сервера під проект.

Якщо проблеми є на системному рівні: bugs (помилки в коді програми) системи, проблеми з кодуванням, проблеми з пошуком, проблеми з авторизацією користувачів, тоді адміністратор шукає помилку в login, де вона виникає, та виправляє. Цей процес триває від 5 хв до 1 доби. Також є критичним, те що, наприклад, інформація може бути неактуальна, погано оформлена, не читабельна, це все виправляє модератор у дуже короткий термін (30 хв – 1 год).

Висновки та перспективи подальших наукових розвідок

Розглянуто питання розроблення методів та програмних засобів опрацювання інформаційних ресурсів у інтернет-системах. Сформульовано новий підхід застосування та впровадження бізнес-процесів для побудови таких систем. Розроблено методи та програмні засоби опрацювання контенту та Web- ресурсу.

1. Корнеев В. В. Базы данных. Интеллектуальная обработка информации / В. В. Корнеев, А. Ф. Гареев, С. В. Васютин, В. В. Райх. – М.: Нолидж, 2000. – 352 с. 2. Белоногов Г. Г. Автоматизированные информационные системы / Г. Г. Белоногов, В. И. Богатырев. – М.: Сов. радио. – 1973. 3. Попов Э. В. Общение с ЭВМ на естественном языке / Э. В. Попов. – М.: Наука, 1982. 4. Gudivada V. N. Поиск информации в World Wide Web / V. N. Gudivada // ComputerWeek. – № 35_97. – С. 19–21, 26, 27. 5. Hayes P. J. Construe/TIS: A system for content-based indexing of a database of news stories / P. J. Hayes, S. P. Weinstein // In Innovative Applications of Artificial Intelligence 2. – The AAAI Press/The MIT Press, Cambridge, MA. – 1991. – P. 49–64. 6. Лукашевич Н. В. Автоматическое рубрицирование потоков текстов по общественно-политической тематике / Н. В. Лукашевич // НТИ. Информационные процессы и системы. – 1996. – Серия 2. – № 10. – С. 22–30. 7. Гареев А. Решение проблемы размерности словаря при использовании вероятностной нейронной сети для задач информационного поиска / А. Гареев // Нейрокомпьютеры: разработка, применение. – 2000. – № 1. – С. 60–63. 8. Цаленко М. Ш. Моделирование семантики в базах данных / М. Ш. Цаленко. – М.: Наука, 1989.

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