УДК 004.942:622.32

АВТОМАТИЗАЦІЯ ПРОЦЕСУ КЛАСИФІКАЦІЇ ЗАПИТІВ КОРИСТУВАЧІВ ПО ДЖЕРЕЛАХ ДАНИХ НАФТОГАЗОВОЇ СПРАВИ

Т.Р. Стисло, В.І. Шекета

ІФНТУНГ, 76019, м. Івано-Франківськ, вул. Карпатська,15, тел. (03422) 46067,
e-mail: taras_07@mail.ru

Представлено структурні рішення для пошукових задач в рамках процесу видобування інформації для інтегрованого середовища WEB-орієнтованої цифрової бібліотеки з метою реалізації інтелектуальних функцій підтримки запитів користувачів та логічної структуризації документів на основі джерел даних, знань, метаданих та метазнань.
Ключові слова: пошукові задачі, видобування інформації, цифрова бібліотека, інтелектуальні функції, підтримка запитів.

Предложены структурные решения для поисковых задач в рамках процесса извлечения информации для интегрированной среды WEB-ориентированной цифровой библиотеки с целью реализации интеллектуальных функций поддержки запросов пользователя и логической структуризации документов на основе источников данных, знаний, метаданных и метазнаний.
Ключевые слова: поисковые задачи, извлеченные информации, цифровая библиотека, интеллектуальные функции, поддержка запросов.

The structured solutions for introduced mining routines in the framework of process for information retrieval for integrated environment of WEB-based university digital library with the goal of implementation of intelligent functions for support of user's queries and documents logical structure generation based on data sources, knowledgesources, metadates and metaknowledges is proposed.
Keywords: mining routines, information retrieval, digital library, intelligent functions, queries support.

Вступ. На сьогоднішній день нафтогазова предметна область дозволяє виділити ряд джерел даних та знань щодо опису нафтогазових об'єктів. В бібліотеці НУНГ джерела інформації по даних та знаннях щодо предметної області представлені такими напрямами, як:

1) нафтогазова справа;
2) буріння;
3) видобування нафти і газу;
4) газонафтопроводи і нафтогазосховища;
5) обладнання нафтових і газових промислів [1-2].

У пропонованому дослідженні основним завданням є побудова цифрової бібліотеки НУНГ з введеними автоматизованими інтелектуальними функціями підтримки запитів користувачів, тому у якості класифікаційного джерела знань по предметній області вибрано дослідження [3-5], які закладають основу реалізації саме інтелектуально виділених інтелектуальних функцій шляхом аналізу особливостей нафтогазових об'єктів як джерел інформації шляхом введення основних методів інформаційного моделювання нафтогазових об'єктів з прив'язкою до сучасних інформаційних технологій. Проте виділені в джерелах [2-5] об'єкти та відношення між ними не дозволяють будувати представлення метарівня, що описані у роботах [6-7].

Тому **метою** пропонованого дослідження є виділення складових процесу класифікації запитів користувачів (студентів, викладачів, дослідників) за джерелами даних та знань нафтогазової предметної області з метою автоматизації та інтелектуалізації даного процесу.

Технологія видобування інформації уможливлює вирішення таких проблем, як: представлення, зберігання, організації доступу до інформаційних входжень [8-9]. Представлена організація інформаційних входжень повинна забезпечити користувачу можливість легкого доступу до релевантних даних та знань. Складність даної задачі зумовлюється певною невизначеністю щодо формального представлення інформаційних потреб користувача, тому в фактичних задачах предметної області інформаційна потреба користувача виражається у вигляді запиту, який обробляється пошуковою машиною. На формальному рівні структура такого запиту представляється множиною ключових слів, які виражають інформаційну потребу користувача:

$$UserInformationNeed = \{KeyWords\}^{set} \rightarrow$$
$$\rightarrow UserQuery \rightarrow$$
$$\rightarrow SearchEngine\{LocalSearch, WWW\} \rightarrow \quad (1)$$
$$\xrightarrow{Goal} \xrightarrow[Inrormation\ Retrieval]{}$$
$$\{RelevantEntries\}^{set}$$

Для заданого запиту користувача основною ціллю системи щодо видобування інформації є отримання релевантних входжень до запиту користувача. У даному контексті виділяють три основні задачі: видобування інформації, видобування даних та видобування знань. Видобування даних полягає в визначенні тих документів у яких ключові слова з запиту користувача зустрічаються із заданою частотою:

$$\left[\{Keywords\}^{set} \subset UserQuery\right] \Rightarrow$$
$$\Rightarrow \left[\{KeyWords\}^{set}\, in \{Documents\}^{set}\right] \quad (2)$$

Проте наявність такого документа ще не означатиме, що він задовольнятиме інформаційні потреби користувача. Тому більший інтерес представлятимуть задачі з видобування інформації про певні об'єкти предметної області, аніж задачі видобування даних з метою задоволення заданого запиту. Відповідно функціональність мови видобування даних (програмної або формальної) спрямована на видобування всіх тих об'єктів, які явно задовольняють визначеним умовам, сформульованим у вигляді логічних виразів або у вигляді реляційної алгебри. Таким чином, для систем видобування даних наявність хоча б одного помилкового об'єкта в множині видобутих визначатиме загальну неуспішність результату. При цьому в складних предметних областях об'єкти характеризуються неточністю, неповнотою та неструктуризованістю, що суттєво збільшує імовірність виникнення помилки. Додаткова складність зумовлюється також тим, що, як правило, об'єктами видобування є природно-мовні тексти, що не завжди є достатньо структурованими і чітко визначеними з семантичної точки зору:

$$\{SubjectDomainObjects\}^{set} \Big| =$$
$$\Big| = \{Unstructured, WeaklyStructured, Fuzzy\} = \quad (3)$$
$$\Big| = in \{LanguageTexts\}^{WellFounded}_{Semantics}$$

Тому найпоширенішим випадком є видобування даних з реляційної бази даних, що дозволяє вирішувати задачу видобування в умовах визначеної структури та семантики:

$$RDBS.Mining \rightarrow$$
$$\rightarrow \{Stuctured, WellFoundedSemantic\} \quad (4)$$

Процедури видобування даних в процесі пошуку рішення згідно запиту сформованого користувачем БД вирішують задачі видобування, результати яких не є достатніми для рівня видобування інформації щодо визначених об'єктів. Для досягнення ефективності з точки зору задоволення інформаційних потреб користувача система видобування повинна певним чином інтерпретувати інформаційні входження в множині документів і ранжувати такі входження за ступенем їх релевантності щодо запитів користувача:

$$\left\{\begin{matrix} InformationEntries \\ in\{Documents\}^{set} \end{matrix}\right\} \Big| =_{int\, erpret}$$
$$\left\{Re\, levantEntries_i : RD_i^{UserQueries}\right\}_{i \in N} \quad (5)$$

Така інтерпретація вмістимого документа передбачає виділення синтаксично-семантичної інформації з тексту документа і використання цієї інформації з метою задоволення інформаційних потреб користувача. Складність у даному випадку полягає в:

1) визначенні способу виділення активних синтаксичних і семантичних описів;

2) побудова способу визначення ступеня релевантності документа відповідно до синтаксичних і семантичних конструкцій, тобто введення змінних типу:

$$Document.RD^{Synt.}, Document.RD^{Semant.}$$

$$\left[\begin{matrix} SyntacticDescriptions \cup \\ \cup\, SemanticDescriptions \end{matrix}\right] \Big| =$$
$$\Big| = \left\{\begin{matrix} Document.RD_i^{Synt}, \\ Document.RD_j^{Semant} \end{matrix}\right\}_{i,j \in N} \quad (6)$$

Таким чином, концепція релевантності є центральною в видобуванні інформації. Тому з логічної точки зору основною ціллю системи видобування інформації є видобування всіх документів релевантних запиту користувача і максимально можлива мінімізація кількості результуючих не релевантних документів:

$$Information\, \mathrm{Re}trieval \xrightarrow{\mathrm{Re}levancy} \xrightarrow[Based]{}$$
$$\{\mathrm{Re}\, levantDocument\}^{set}_{\max} \cup \quad (7)$$
$$\cup \{Un\, \mathrm{Re}\, levantDocyments\}^{set}_{\min}$$

Новітній розвиток технологій обробки інформації виходить за рамки початкових цілей щодо індексації тексту і пошуку найбільш корисних документів серед можливих доступних. Новітні рішення полягають у задачах моделювання, класифікації та категоризації документів:

$$\{TextsIndexing \cup Re\, levancySearch\} \Big| =$$
$$\Big| = \left\{\begin{matrix} Models, Classifications, \\ Categorisations \end{matrix}\right\}^{Documents}_{set} \quad (8)$$

Важливими в реалізації задач видобування інформації є також задачі, пов'язані з архітектурами програмних рішень, інтерфейсами користувача, методів візуалізації даних та методів фільтрації даних.

Найбільш перспективними застосуваннями технології видобування інформації є цифрові бібліотеки [8] та задачі конгнітології [6]. Новітні розробки також ведуться в області видобування інформації на основі мультимедійних даних та гіпертекстових технологій в рамках концепції семантичного Web. Особливість Web-технологій полягає в можливості використання їх глобальних, універсальних та уніфікованих баз даних та баз знань.

Успіх Web-технологій зумовлюється концепціями стандартизованого інтерфейсу користувача, який має уніфікований вигляд незалежно від середовища програмного застосування, що дає змогу спростити процес доступу до Web відділивши задачу доступу до інформації від технічних задач, таких як використання комунікаційних протоколів, URL-ресурсів та операційних систем.

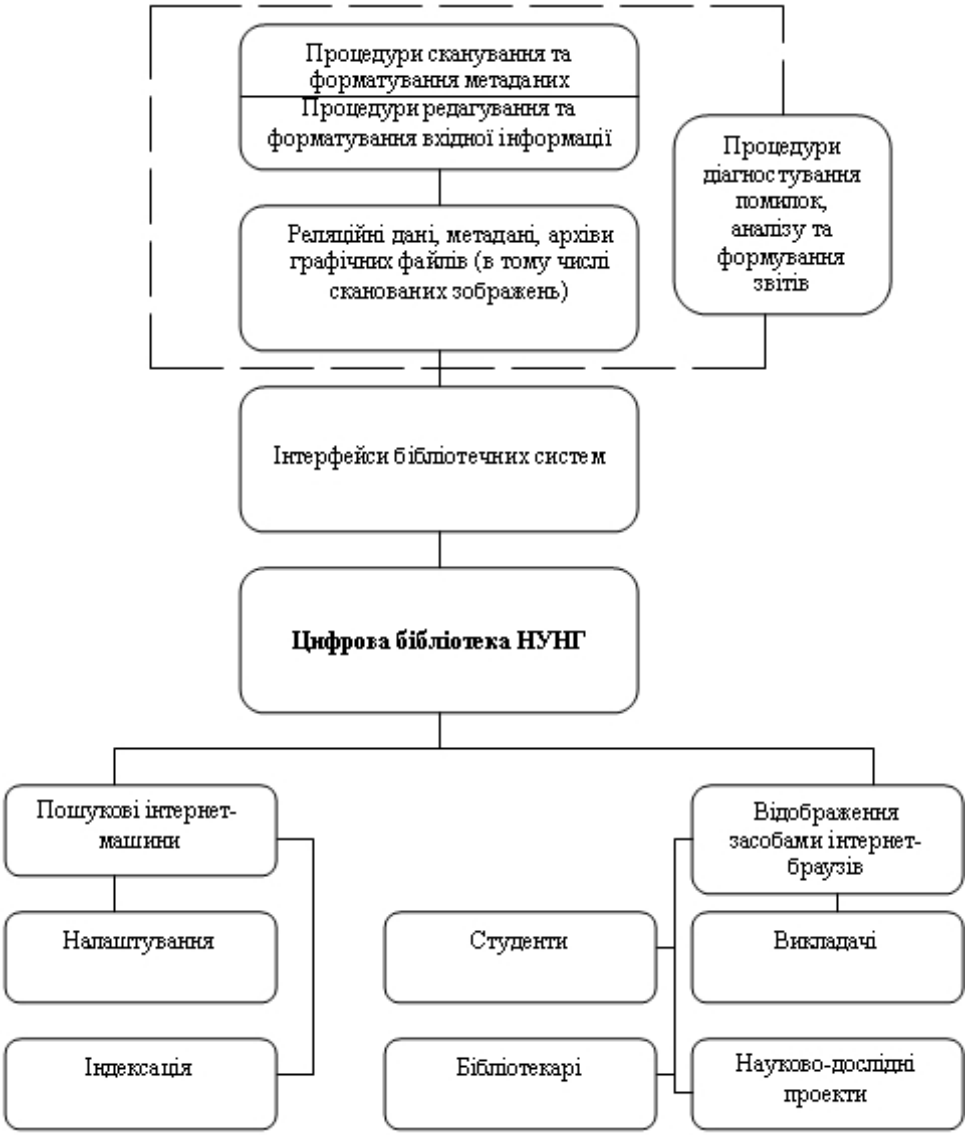

Рисунок 1 – Структурна схема цифрової бібліотеки

Користувачі можуть створювати власні документи, розміщати їх на Web-ресурсах, що дозволяє розглядати Web як єдине середовище для публікування документів. Проте, незважаючи на значне поширення Web-технологій в усіх сферах людської діяльності, проблема пошуку корисної (релевантної) інформації була і залишається актуальною. Процес такого пошуку може бути тривалим і неефективним у зв'язку з масштабністю Web. Неуспішні результати пошуків знижують мотивацію користувача та викликають інформаційну напруженість. Основною причиною даної ситуації є відсутність чітко визначеної базової моделі Web, що відповідно передбачає відсутність визначеної інформаційної структури задовільної якості. Найбільший інтерес для дослідження представляють комп'ютерні технології, техніки, методики, що можуть бути використані для підвищення ефективності для видобування інформації. Одним з можливих ефективних застосувань в області цифрових бібліотек є інтелектуалізація процесу пошуку, що базується на формалізації інтерпретації та процедури видобування інформації людиною, що може бути використана як основа для побудови комплексної процедури обробки інформації.

Ефективне видобування релевантної інформації безпосередньо залежить від завдання (запиту) користувача і від логічної структури (логічного представлення) документа, що може інтерпретуватися системою видобування інформації.

Користувач системи видобування інформації повинен формалізувати опис своїх інформаційних потреб у вигляді запиту засобами мови, що пропонує система. На програмному рівні реалізація системи видобування інформації означає специфікацію множини ключових слів, що передають семантику інформаційної потреби. Відповідно реалізація системи видобування інформації даних передбачає побудову виразу для запиту, що використовується для передачі обмежень, що повинні задовольнятися об'єктами в множині відповідей (результатів). В обох випадках вважається, що пошук користувача щодо корисної інформації формується у вигляді задачі видобування:

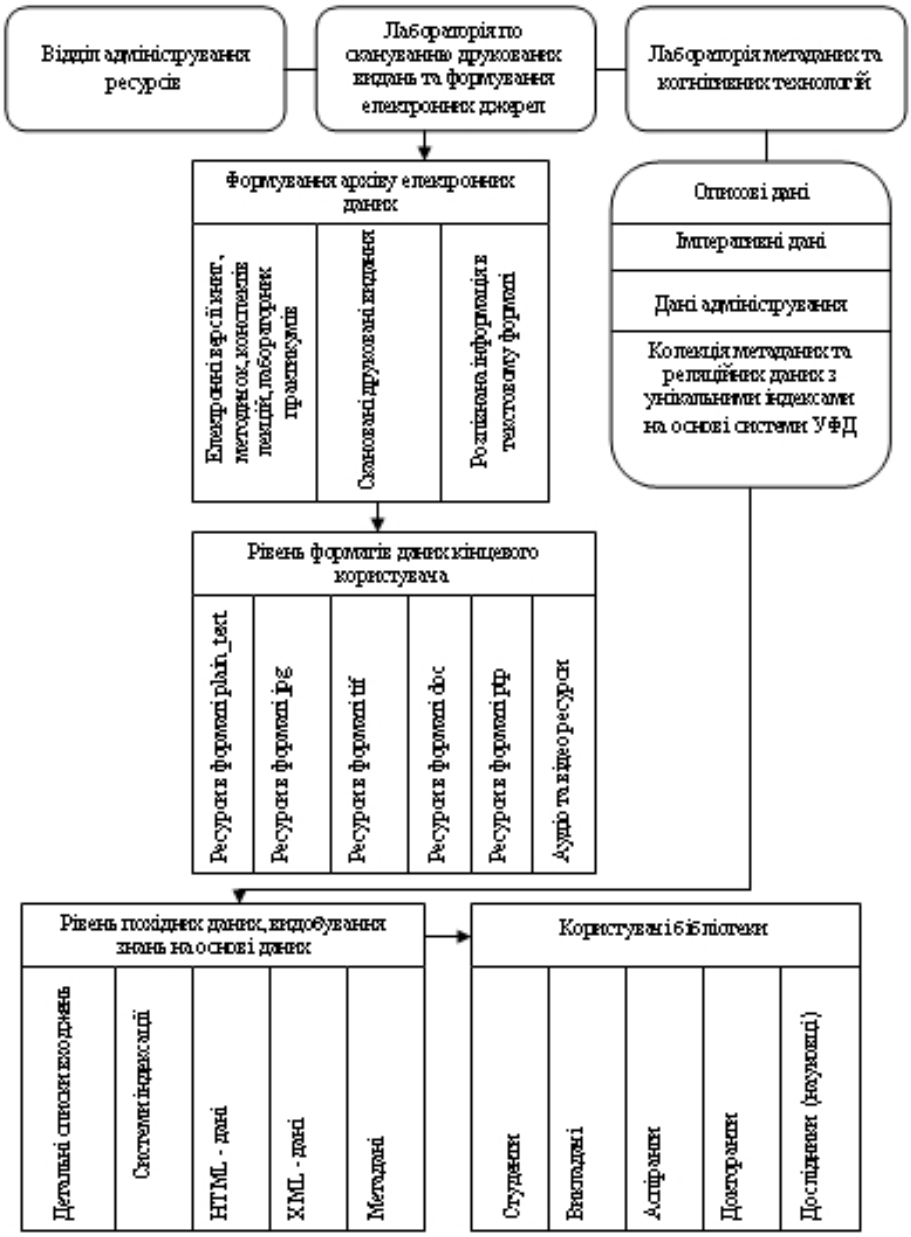

Рисунок 2 – Система формування ресурсів цифрової бібліотеки та доступ до них

$$\{KeyWords\}^{set} \xrightarrow{Semantic} \xrightarrow[InformationNeed]{} \{UserQueries\}_{set}^{Constraints} | = | = \{RetrievalObjects\}^{set} \subset \{Solutions\}^{set} \quad (9)$$

В загальному випадку після визначення інформаційної потреби формування запиту може бути зведено до формування як слабко визначеної задачі *WeaklyStructuredTask* , так і для переозначеної задачі *OverconstrainedTask* . Найпростішим способом задоволення інформаційної задачі буде перегляд користувачем документів у відповідній множині. Тому процес пошуку зводиться до простого перегляду:

$$Information\ Retrieval \equiv \equiv Review\{Documents\}^{set} \quad (10)$$

Специфікою такого виду видобування інформації буде те, що основні цілі процесу не є чітко визначеними на його початку і можуть змінюватися під час його взаємодії з системою. Таким чином, можна виділити три основних завдання користувача системи видобування інформації:

1) видобування інформації та даних;
2) видобування знань;
3) перегляд (браузінг).

Класичні системи видобування інформації традиційно орієнтовані на задачі видобування даних та інформації, відповідно до гіпертекста системи налаштовуються, як правило, для забезпечення швидкого перегляду. Сучасні цифрові бібліотеки, що використовують Web-інтерфейси, дозволяють поєднувати дані задачі з метою забезпечення покращених властивостей процесу видобування.

Залежно від визначених задач різним буде рівень взаємодії користувача з системою. Таким чином, спільне використання процесу видобування інформації та даних, як правило, використовується в системах, що базується на Web-інтерфейсах. Більше того, такі системи надають також можливість браузінгу, що дозволяє поєднувати процедури видобування інформації та даних з процедурами перегляду (браузінгу).

Процедури видобування інформації браузінгу є ефективними інструментами WWW. Це означає, що користувачі запитують інформацію в інтерактивному режимі. Альтернативним підходом є видобування інформації в автоматичному режимі за допомогою інтелектуальних агентів які виконують збір інформації для користувача. Проте дана опція стане доступною повною мірою тільки з впровадженням стандарту семантичного Web. Інформація корисна для користувача добувається з доступних, як правило гетерогенних, джерел. У даному випадку задача видобування інформації зводиться до фільтрації релевантних входжень, які пізніше сортуються користувачем:

$$Information\ Retrieval \equiv \equiv Sorting\{RelevantEntries\}^{set} \quad (11)$$

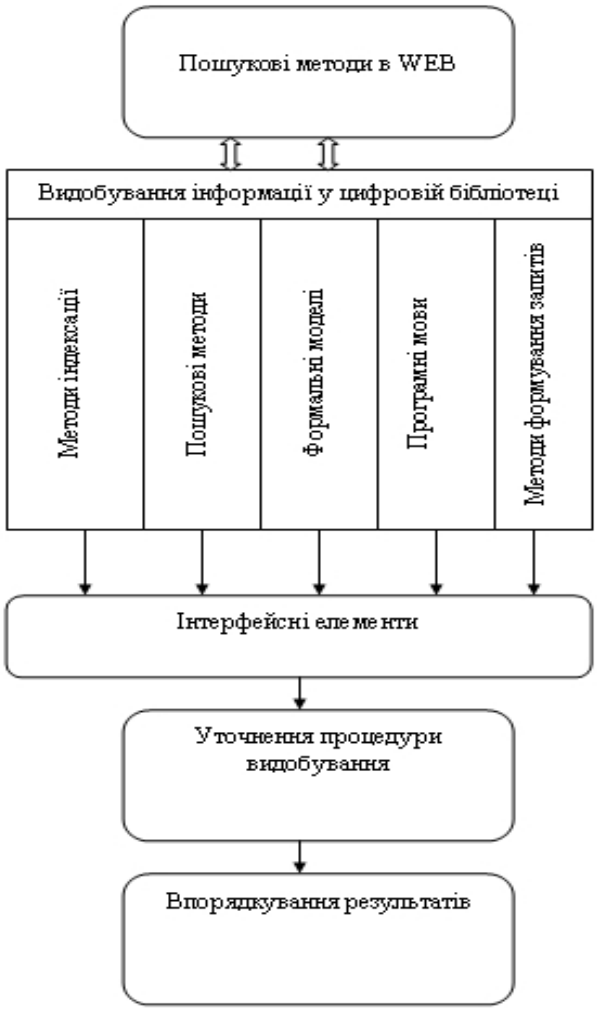

Рисунок 3 – Видобування інформації у цифрових бібліотеках

З точки зору побудови ефективних запитів користувача в інтелектуальних бібліотечних системах важливе значення має дослідження логічної структури документів, що містяться в бібліотеці таких, як: книги, підручники, навчальні посібники, журнали, дисертації, автореферати дисертацій, методичні вказівки, лабораторні практикуми та інше.

Із запуском процедури сканування та оцифрування документів важливого значення набуває розгляд документів в їх сукупності (колекції). Така сукупність ефективно представляється множиною індексованих ключових слів. Такі ключові слова можуть видобуватися безпосередньо з тексту документів або вказуватися користувачами (викладачами, адміністраторами ресурсів). Незалежно від способу генерації (похідні з документів або вказані користувачем) такі ключові слова визначають логічну структуру документа:

$$\begin{aligned} &DocumentCollections = \\ &= \left\{Documents^{set}, \prec_{ord}\right\} = \\ &= \{KeyWords\}^{set} \Big| = \{Indexes\}^{set} \end{aligned} \quad (12)$$

З точки зору методів інформаційних технологій, що визначають задачі роботи з текстовою інформацією, множину всіх документів колекції можна розглядати як множину множини входжень всіх слів в документах. В такому випадку можна стверджувати, що система видобування отримуватиме доступ до логічного представлення документів у формі природноумовних текстів.

Проте з досягненням колекціями документів значних розмірів, наприклад в бібліотеці ІФНТУНГ, кількість таких документів перевищує один мільйон входжень, обробка запитів користувача таких колекцій вимагатиме значних ресурсів. Зменшення обчислювальної складності задач даного класу досягається за рахунок зменшення кількості входжень в множині слів, що виконується шляхом формування множин найбільш значущих, незначущих та допоміжних входжень. Дана процедура виконується у кілька етапів:

1) видалення займенників, прислівників тощо;

2) видалення ключових слів, які можна розглядати як похідні від входжень ключових слів кореня;

3) ідентифікація та класифікація груп іменників з метою усунення похідних граматичних форм.

Операції пунктів 1-3 розглядаються як базові операції перетворення текстового контенту. Такі операції дозволяють зменшити представлення документів і перехід від логічного представлення документа у формі природноумовного тексту до представлення у формі індексованих ключових слів. З точки зору ефективності пошуку, логічне представлення у вигляді природно-умовних текстів є більш доцільним, проте відповідно вимагає також більших обчислювальних ресурсів. Тому оптима-

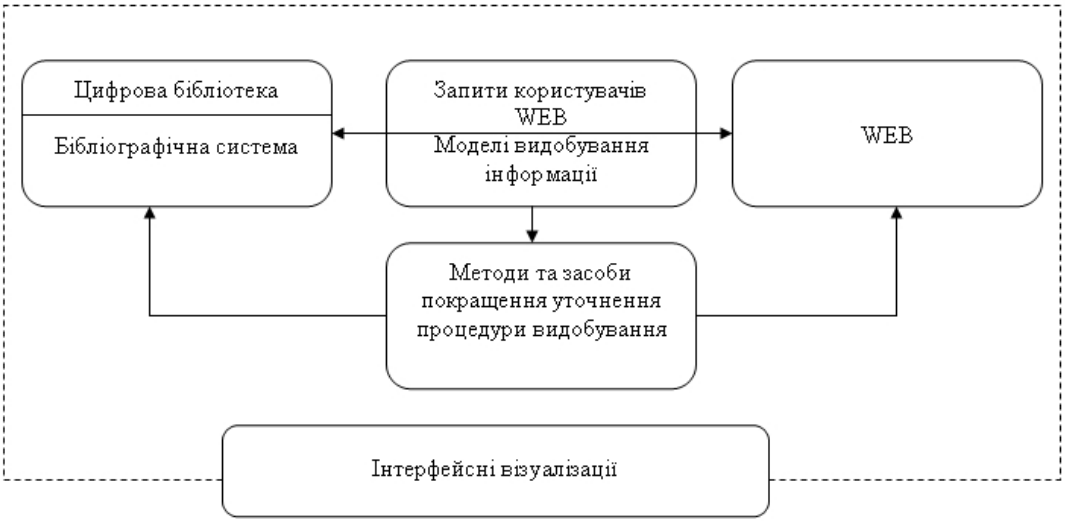

Рисунок 4 – Структура пошукових задач

Рисунок 5 – Опис процесу видобування інформації

льним рішенням є початкова ініціалізація деякої множини категорій визначених користувачем, що забезпечуватиме найбільш стисле представлення документу (документів), але її впровадження призводитиме до отримання процедур видобування низької якості. Використовується також ряд проміжних логічних представлень документа для систем видобування інформації.

Крім використання проміжних стратегій системи видобування інформації можуть розпізнавати також внутрішню структуру, що, як правило, представлена в документі або може бути видобута з нього. В якості такої структури може використовуватися структура розділів, секцій, підсекцій і т. п.

Інформація про структуру документа може бути основною стратегією видобування і ле-

жить в основі моделей видобування інформації на основі структурованого тексту:

$$RetrievalStrategy[DocumentStructure] = \\ |= RetrievalModels \quad (13)$$

Логічне представлення документа є відображенням у якому виконується перехід від природно-умовного до формалізованого представлення,що специфікуються користувачами.

Основне завдання яке вирішувалося на усіх етапах розвитку технологій обробки інформації полягало в побудові ефективних методів організації та накопичення інформації і відповідно її подальшого використання. Контент бібліотечних входжень можна оцінювати на основі його змісту та анотацій:

$$DigitalLibrary|=\{Content\} \cup \{Annotations\} \quad (14)$$

Для забезпечення доступу до вмістимого книг на основі змісту необхідна побудова спеціалізованих структур даних, що забезпечуватимуть більший доступ до збереженої інформації. Класичною і поширеною структурою даних для забезпечення швидкого видобування інформації вважаються колекції виділених ключових слів або концептів з якими асоційовані вказівники на релевантну інформацію у формі документів, засобами індексації:

$$\{KeyWords\}^{set} \cup \{Concepts\}^{set} \rightarrow \\ \rightarrow \{Pointer_i : RelevantEntry_i\}_{i \in N} \quad (15)$$

У відповідних формах представлення індекси є ядром кожної сучасної системи видобування інформації. Вони відповідно забезпечують швидкий доступ до інформації, до даних і дозволять пришвидшувати процедуру обробки запитів:

$$\{UserQueries\}^{set} \leftrightarrow \{Indexes\}^{set} \quad (16)$$

Традиційно в бібліотеках індекси створювались вручну як форма категоризаційних ієрархій:

$$\{CatHierarchies\}^{set}\left|\{Indexes\}^{set}\right| = \\ |=Classify[Documents]^{set} \quad (17)$$

Більшість сучасних нецифрових бібліотек використовують той чи інший вид категоризаційної ієрархії для класифікації своїх документів. Бібліотечні ієрархії створені фахівцями з предметної області (бібліотечна справа). Новітні інформаційні технології дозволяють автоматизувати процес побудови комплексних індексів. Автоматизація побудови індексів забезпечує представлення проблем видобування інформації з акцентом на систему, а не на інформаційні потреби користувача.

З даної точки зору важливим є розрізнення між двома основними підходами до проблем видобування інформації:

1) видобування інформації орієнтоване на комп'ютер *ComputerBased Retrieval* ;

2) видобування інформації орієнтоване на користувача *UsedBased Retrieval* .

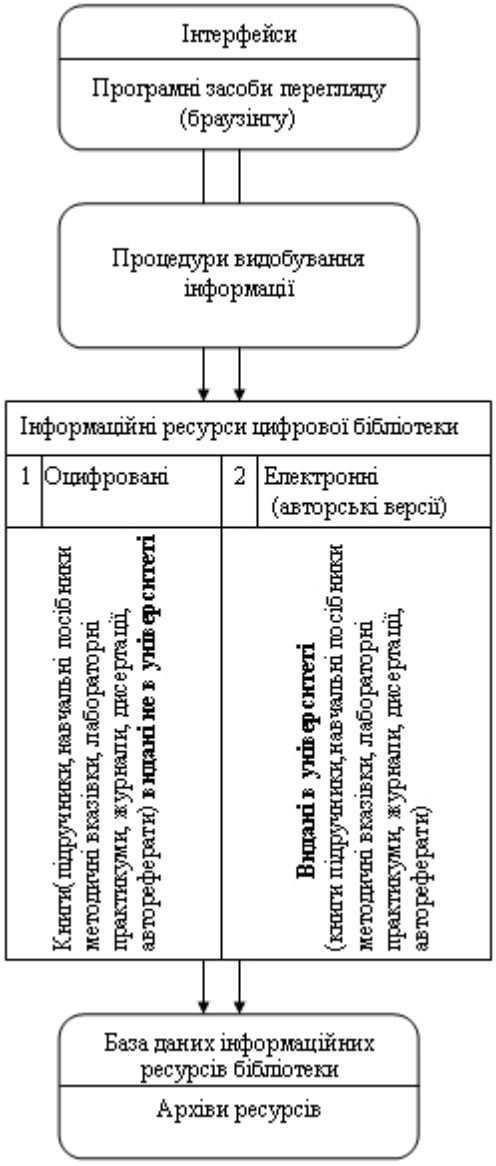

Рисунок 6 – Структура інформаційної взаємодії з ресурсами цифрової бібліотеки

У першому випадку проблема видобування в основному полягатиме в побудові ефективних індексів, обробки запитів користувача з високою ефективністю і побудови ранжованих алгоритмів, що покращують якість множини відповідей:

$$ComputerBased\,Retrieval| = \{ \{Indexes\}^{set}, \\ \{UserQueries\}^{set}, Solutions^{set} | Ranged \} \quad (18)$$

У другому випадку проблема видобування інформації полягатиме в дослідженні поведінки користувача, розуміння його основних потреб і визначення такого факту, яким чином таке розуміння впливає на організацію і функціонування систем видобування:

$$UserBased\,Retrieval| = \begin{Bmatrix} \{UserBehaviour\}^{set}, \\ \{UserNeeds\}^{set}, \\ \{FeedBacks\}^{set} \end{Bmatrix} \quad (19)$$

Відповідно до такого підходу обробка запитів на основі ключових слів не є ефективним рішенням для проблем видобування інформації.

Системи видобування інформації на початкових етапах її функціонування, що використовувалися і використовуються в бібліотечних системах у своїй еволюції пройшли ряд етапів.

В першому поколінні такі системи були автоматизованими інформаційними системами, базованими суто на бібліотечних технологіях, наприклад, веденні каталогу і карток (алфавітного, систематичного, каталогу іноземної літератури, періодичних видань, дисертацій, авторефератів дисертацій, службового, індикаторного) і на базовому рівні уможливлювали пошук за іменем автора чи назвою книжки.

Друге покоління систем дозволило підвищити функціональність пошуку шляхом його виконання за заголовками, ключовими словами і додатковими можливостями побудови комплексних запитів.

Третє покоління систем, що розвиваються у даний час, закцентовано на покращеному графічному інтерфейсі бібліотечних систем, розвинутих електронних форматах, гіпертекстових ресурсах і відкритій програмній архітектурі з можливістю легкої інтеграції нових модулів та програмних рішень, зокрема стосовно інтелектуалізації процесу видобування даних та знань.

Аналіз пошукових машин у середовищі Веб свідчить, що вони базуються на системі індексації входжень, подібної до тієї, що використовується в бібліотеках. Проте розвиток Веб-технологій дає змогу виділити такі новітні рішення в даній області:

– значне здешевлення вартості доступу до інформаційних ресурсів, що суттєво розширює коло можливих та потенційних клієнтів;

– впровадження засобів і методів розширеного доступу, що робить можливим віддалений доступ з обчислювальних та мобільних обчислювальних пристроїв;

– можливість вільного доступу користувачів до медійних ресурсів Веб призвела до появи таких інформаційних явищ та технологій, як: соціальні мережі, блоги, форуми.

Важливою особливістю Веб-технологій та Веб-орієнтованих цифрових бібліотек є високий рівень їх інтерактивності, що дозволяє клієнту обмінюватися повідомленнями, документами, медіа ресурсами, програмним забезпеченням, а також можливістю мережевого спілкування за низькою вартістю. Крім того, клієнти самі обирають час доступу, що визначає нову тенденцію щодо зручності та ефективності мережевого доступу. Таким чином, однією з основних особливостей веб-бібліотек є надвисокий ступінь їх інтерактивності. Проте, незважаючи на це, основна проблема залишається невирішеною – проблема ефективного видобування інформації релевантного до потреб користувача. Таким чином, у динамічному Веб-середовищі, що є основою сучасних цифрових бібліотек, необхідне впровадження технологій інтелектуального пошуку, що є одним з пріоритетів у концепції семантизації Веб. В умовах зростаючого попиту на мережевий доступ актуальним стає питання пропозиції оптимальних способів індексації та зменшення часу обробки запитів. Крім того, якість задачі видобування інформації значною мірою залежить від способу взаємодії користувача з системою. Тому у даному контексті актуальним є питання побудови певних стратегій поведінки користувача і відповідна імплементація таких метаданих в дизайні та архітектурі.

Оскільки однією з основних тенденцій розвитку Веб є електронна комерція, то тенденції розвитку комерціалізації поширились також на цифрові бібліотеки. Оптимальним рішенням в даному випадку є видавання кожному користувачеві бібліотеки електронної картки на зразок банківської кредитної картки, яка одночасно виконує роль читацького квитка у звичайному розумінні і містить такі дані: назва ВУЗу, номер читацького квитка, прізвище, ім'я, по батькові користувача, назва групи, факультету, дата видачі, персональний штрих-код та фотографія користувача, печатка бібліотеки, підпис директора, а також містить кредитні кошти для надання платних послуг, передбачених статутом бібліотеки. Відповідно користувач картки може отримувати платні послуги як безпосередньо у бібліотеці, так і віддалено. Виконання віддалено доступу до платних послуг вимагає високого рівня питання безпеки таких операцій, що забезпечується на основі процедур шифрування. Додатковою проблемою є питання забезпечення приватності та конфіденційності доступу, тобто читацькі інтереси користувачів повинні бути захищенні в рамках їх профілю. Таким чином, впровадження Веб-орієнтованої цифрової бібліотеки вимагає ефективного вирішення питань конфіденційності та безпеки доступу. Також впровадження такого стандарту бібліотеки актуалізує питання авторських, патентних та суміжних прав, що очевидно недостатнім чином описані та захищені відповідними законодавчими актами. Дане питання є важливим, оскільки воно є також одним з ключових питань при побудові та підтримці цифрових бібліотек великої розмірності. В даному контексті новітні Веб-базові платформи для blog-ресурсів, wikies – ресурсів є серйозним конкурентом для класичних джерел друкованої продукції (видавництв), які на сьогоднішній день також використовують електронне представлення видань з метою підвищення комерційної ефективності їх роботи. Тому співвідношення двох даних напрямів також містить у собі суттєві колізії в питаннях дотримання авторських, патентних та суміжних прав. Додаткові проблеми виникають також з впровадженням процедур сканування, розпізнавання та побудови запитів на видобування інформації з використанням різних природних мов користувачів, що особливо актуально в умовах збільшення кількості іноземних студентів.

Для практичної реалізації необхідний детальний аналіз процесу видобування інформації. Даний процес інтерпретується в термінах

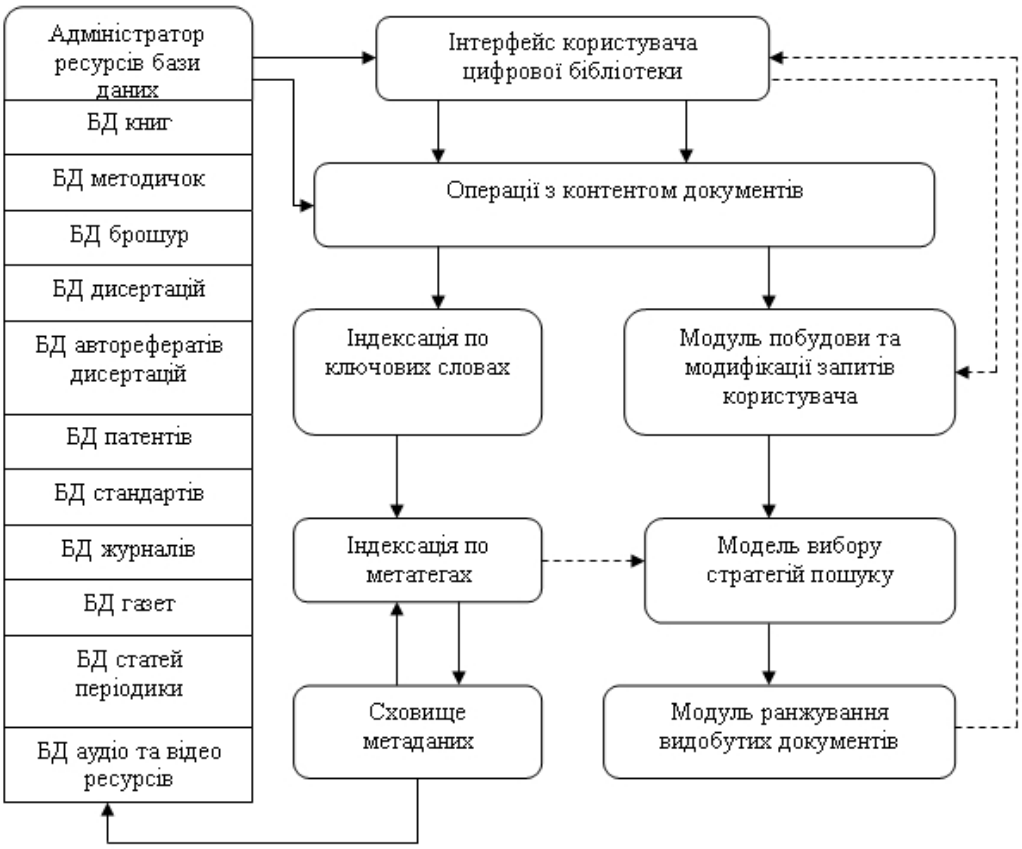

Рисунок 7 – Структуризація процесу видобування інформації

компонентних під процесів. Для опису процесу видобування інформації використовується підхід зображений на рисунку 7.

З даного рисунку видно, що з початком ініціалізації процедури необхідно визначити базу даних текстового контенту. В проекті цифрової бібліотеки НУНГ дану задачу виконує адміністратор системи бібліотечних БД. При цьому він вирішує такі завдання:

– виділяє документи для використання *SelectedDocuments* ;

– визначає операції, що будуть виконуватись над текстовим контентом *ContentOperations* ;

– визначає структуру текстового контенту: елементи контенту, що будуть основою процесу видобування *ContentStructure* .

Таким чином, множина операцій, що виконуються над контентом, дає змогу генерувати його у логічне представлення.

$$LogicalStructure\left\{\begin{matrix} ContentOperations^{set}, \\ DefinedContentStructure \end{matrix}\right\} \quad (20)$$

Після того як визначено логічне представлення документа адміністратор БД виконує побудову систем індексації текстового контенту.

Таким чином, представлення індексів є вирішальним фактором вибору структури даних, оскільки дозволяє швидкий пошук у великих об'ємах даних. В якості основних ресурсів виступає час і доступний дисковий простір необхідний на визначення текстового контенту БД і відповідно процедура побудови індексів базується на процедурі побудови запитів в процесі видобування.

Враховуючи, що БД документів є індексованою можна ініціалізувати процедуру видобування інформації шляхом виконання наступних кроків:

Крок 1: визначення та специфікація потреб користувача (*UserInformationNeed*);

Крок 2: виконання парсінгу і трансформації визначеної інформаційної потреби на основі текстового контенту;

Крок 3: застосування операції перетворення запиту в процесі генерації логічного представлення ініцілізаційної потреби користувача;

$$UserQueries^{modif} \Big| = LogicalStructure[\\ [UserInformationNeed] \quad (21)$$

Крок 4: виконання обробки запиту з метою отримання видобутих документів.

$$\Big[UserQueries^{process.}\Big] = \\ \Big| = \big[RetrievedDocuments\big]^{set} \quad (22)$$

Швидкість обробки запитів визначається вибраною структурою індексації. Перед відправкою користувачу видобуті документи ранжуються відповідно до ймовірної релевантності.

$$\Big\{\big[RetrievedDocuments\big]^{set}, \prec_{Relev.}^{prob.}\Big\} \quad (23)$$

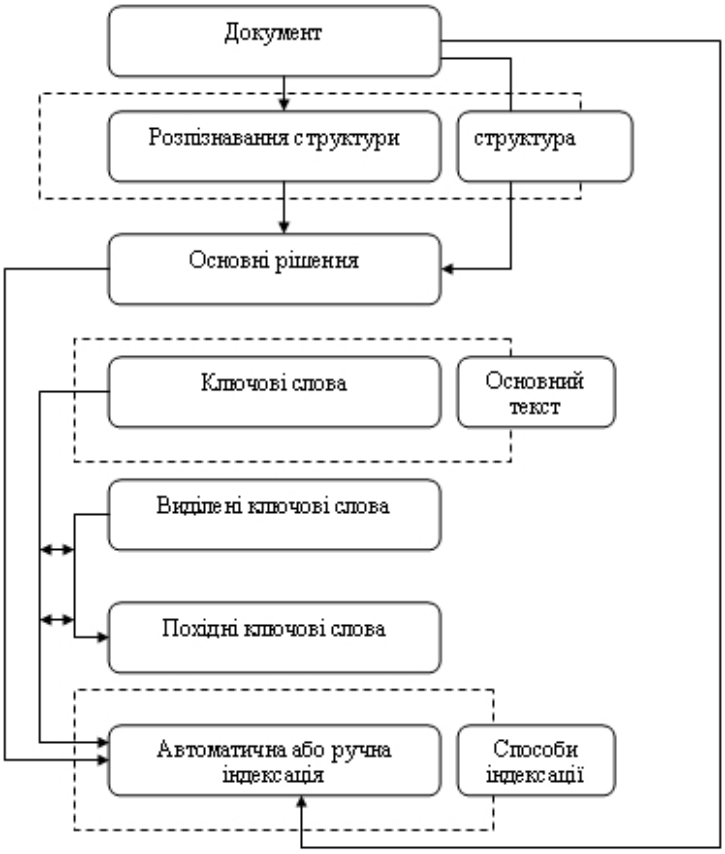

Рисунок 8 – Виділення логічної структури документа

Представлення інтерфейсу користувача в новітніх системах видобування інформації, як правило зводиться до пошукових машин-Веб та технологій Web-браузерів. Основною складністю аналізу ефективності інтерфейсу є те, що користувач як правило не представляє системі свої інформаційні потреби у вигляді формалізованих запитів на отримання ранжованих документів. Проте у кожний момент часу користувач може виділити деяку підмножину документів, що є релевантними, і тим самим змінити налаштування всього процесу видобування:

$$\left[RelevantDocuments\right]^{SubSet}\Big| = \Big| = RetrievalProcess^{Settings} \tag{24}$$

В процесі видобування система використовує документи, вибрані користувачем для побудови послідовних модифікованих запитів:

$$UserQueries^{modif.}\Big| = \left[SelectedDocuments\right]^{SubSet} \tag{25}$$

Імовірно, що такий модифікований запит буде кращим в плані релевантності отриманих результатів.

Висновок

У даній статті представлено структурні рішення щодо формування ресурсів цифрової бібліотеки шляхом оцифровування даних, імпортування даних та формування комплексних інтерфейсів доступу до них. Основним результатом даного дослідження стало представлення структури пошукових задач для процесу видобування інформації в WEB-базованих цифрових бібліотеках шляхом комплексної взаємодії інформаційних ресурсів джерел даних та знань нафтогазової предметної області. В процесі розв'язання даної задачі виділено структуру видобування інформації нафтогазової предметної області та описано логічну структуру поточного виділеного документа.

Подальші дослідження даного напряму будуть спрямовані на програмну реалізацію виділених структурних елементів.

Література

1 Бойко В.С. Довідник з нафтогазової справи / В.С. Бойко, Р.М. Кондрат, Р.С. Яремійчук. – Львів, 1996. – 620 с.

2 Положення про систему каталогів і картотек науково-технічної бібліотеки ІФНТУНГ. // Наказ по університету від 14.09.2010р. № 105.

3 Юрчишин В.М. Інформаційне моделювання нафтогазових об'єктів: монографія / В.М. Юрчишин, В.І. Шекета, О.В. Юрчишин. – Івано-Франківськ: Видавництво Івано-Франківського національного технічного університету нафти і газу, 2010. – 192 с.

4 Юрчишин В.М. Наукові основи застосування інформаційних технологій при управлінні процесами розробки нафтогазових родовищ. Дис...д-ра техн. наук / Юрчишин В.М. – Івано-Франківськ, 2006. – 310 с.

5 Шекета В.І. Інформаційна система для прогнозування нафтогазоносних покладів. Дис... канд. техн. Наук / Шекета В.І. – Херсон. – 1999. – 140 с.

6 Стисло Т.Р. Концептуалізація запитів користувача по базах знань метаданих про бібліотечні ресурси / Т.Р.Стисло, В.І. Шекета, Р.М. Федорак // Вісник Хмельницького національного університету. Технічні науки. – 2009. – Том 4. – С. 132-138.

7 Стило Т.Р. Структуризація інформаційного простору цифрової бібліотеки з допомогою орієнтованих графів. // Матеріали міжнародної конференції «Теоретичні та прикладні аспекти побудови програмних систем TAAPSD'2009». – 8-10 грудня 2009. – Київ, 2009. – С. 108-109.

8 Arms W. Digital Libraries. – MIT Press. – Cambridge, MA. – 2000. – 291p.

9 K. Sparek Jones, P. Willet. Readings in Information Retrieval. Morgan Kaufmann Publishers, Inc. – 1997. – 593 p.

Стаття надійшла до редакційної колегії 09.11.10
Рекомендована до друку професором ***Д. Ф. Тимківим***