УДК 51.001.57+004.652.4+004.827

Р. Нога, Н.Б. Шаховська,
Національний університет “Львівська політехніка”,
кафедра інформаційних систем та мереж

АНАЛІТИЧНИЙ ОГЛЯД МЕТОДІВ ТА ЗАСОБІВ ОПРАЦЮВАННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ

Проаналізовано проблеми опрацювання розрізнених текстових даних. Виділено задачі опрацювання текстових даних. Показано, що наявне математичне та програмне забезпечення недостатнє для одночасного розв’язання задач опрацювання множинних текстових ресурсів.

Ключові слова: ключове слово, індексування, пошук.

There are analyzed the problems of disparate text data processing. The problems of text data processing are highlighted. It is shown that the existing mathematical software and insufficient for the simultaneous solution of problems handling multiple text resources.

Key words: keyword, indexing, searching.

Вступ

Повноцінне й оперативне забезпечення суспільства новітньою інформацією є необхідною передумовою підвищення ефективності та інноваційної віддачі наукових досліджень.

Вирішення цієї проблеми потребує інтеграції зусиль наукових установ і вищих навчальних закладів для пропагування досягнень української науки та сприяння формуванню у суспільства сучасного наукового світосприйняття, активізації євроінтеграційних процесів у вітчизняному науково-інформаційному просторі.

Впровадження світової інформаційної мережі Інтернет дало можливість будь-якій людині в будь-який час з будь-якої точки світу отримати доступ до будь-якої інформації, що є в цій мережі. Електронна інформація почала мати все більше значення в усіх сферах життя сучасного суспільства. В інформаційних сховищах, розподілених по всьому світу, зібрані террабайти різної інформації. Основну частину цієї інформації складає текстова інформація. Розвиток інформаційних ресурсів Інтернет створив, а з часом і багаторазово підсилив проблему інформаційного перенавантаження. Адже за існуючими оцінками, неструктуровані дані – головним чином текст – складають не менше 90 % інформації. І лише 10 % приходиться на структуровані дані, завантажені в реляційні СУБД [5]. Звичайно такий потужний потік неструктурованої інформації заважає пошуку необхідної інформації. Стара проблема – "де знайти необхідну інформацію" змінилась новою проблемою – "як вибрати необхідну інформацію". Процес відбору необхідної інформації з загального інформаційного потоку почав вимагати великих затрат часу. Перед людством стала проблема створення інтелектуальних систем пошуку необхідної інформації, або як їх ще називають – технологій глибинного аналізу текстів.

Технологія глибинного аналізу текстів – це саме той інструмент, котрий дозволяє аналізувати великі обсяги інформації в пошуках тенденцій, шаблонів і взаємозв’язків, здатних допомогти в виборі необхідної інформації для прийняття стратегічних рішень. Ця технологія – це новий вид пошуку, котрий на відміну від традиційних підходів не тільки знаходить списки документів, формально релевантних запитам, а і допомагає знайти в них необхідну інформацію. Використовуючи аналогію з бібліотекою, ця технологія подібна до відкриття перед читачем книги з підкресленою необхідною інформацією. Порівняйте це з видачею читачеві стосу книг та документів, в котрих десь є інформація, необхідна читачеві, однак знайти її має він сам. Процес осмисленого пошуку являється далеко не тривіальним, адже часто в багатьох документах присутній тільки натяк на необхідну інформацію. Необхідні потужні інтелектуальні можливості, щоб знайти те, що потрібно [3]

Пошукові системи в корпоративній мережі призначені для роботи з масивами текстових документів підприємства, що мають обсяги від декількох гігабайт до декількох десятків гігабайт. Такі програми мають бути реалізовані в мережному варіанті, при якому доступ до бази даних на сервері локальної мережі, здійснюється з робочих станцій співробітників. Джерелами інформації в корпоративних мережах виступають не тільки сайти і сторінки, що наповнюють ці сайти і які створені за єдиними стандартами, але і безліч інших баз даних і різних репозиторіїв структурованих і неструктурованих даних. Різнорідність джерел робить неможливим просте перенесення в корпоративне середовище звичайних пошукових машин. Існує кілька варіантів мережного виконання пошукової програми [8].

Перший і найпростіший – це можливість пошуку в мережному оточенні. Така програма може індексувати файли, що розташовані не тільки на локальному комп'ютері, але і на дисках інших робочих станцій, з'єднаних в локальну мережу. При цьому пошук може здійснюватися тільки з комп'ютера, на якому встановлена система і розташована база даних, включаючи пошуковий індекс. Багато-користувацький режим, при якому користувачі з своїх робочих місць можуть звернутися до бази даних за інформацією, не забезпечується.

Другий варіант – це пошукові системи, що працюють по Інтернет протоколу. В цьому випадку база даних і основна програма розташована на центральному сервері локальної мережі, а всі користувачі мають доступ до інформації з своїх комп'ютерів через стандартний Інтернет браузер. Тобто все відбувається так само, як і при пошуку в глобальному Інтернеті. Користувач, працюючи в Інтранет мережі, для доступу до бази даних вводить адреси баз даних і далі шукає інформацію по стандартній схемі, із стандартним інтерфейсом пошукової системи. Природно, що пошукові програми, створені на основі пошукових Інтернет систем, в основному використовують Інтернет протоколи у випадку роботи у режимі багатокористувацький.

Наступний рівень – це програмні системи, що мають клієнт-серверну архітектуру з власною клієнтською частиною програми. Програма клієнт встановлюється на всіх робочих станціях мережі, а програма сервер забезпечує індексування інформації всієї мережі, створення бази даних на сервері та доступ до неї всіх користувачів. Такі системи складніші в розробці, але мають більше функціональних можливостей, чим системи, які використовують стандартний браузер. Наприклад, для розмежування доступу користувачів до різних видів корпоративної інформації використовуються системні засоби, а інтерфейс призначений для користувача можна зробити більш функціональним і зручним.

1. Постановка задачі та її актуальність

До найактуальніших засобів інтелектуального аналізу текстів належать технології виділення фактографічної інформації про об'єкт пошуку з врахуванням анафоричних посилань на них (посилання на об'єкт, названий в тексті); нечіткий пошук; тематичне і тональне (точне і повне) рубрикування; кластерний аналіз сховищ і добірка документів; виділення ключових тем; побудова анотацій; побудова багатовимірних частотних розподілювачів документів і їх дослідження за допомогою OLAP-технологій; використання методів інтелектуального аналізу тексту для визначення напрямку дослідження великих добірок документів і отримання нової інформації.

Найсучаснішими напрямками отримання інформації з текстів сьогодні є:

- аналітична обробка фактів; ведення досьє;
- отримання та структурування фактографічної інформації;
- пошук інформації по запитах з використанням тезаурусів;
- напрямку пошуку інформації, об'єктів в сховищі документів, в добірці документів;
- анотування документів, побудова дайджестів по об'єктах;
- проведення тематичного аналізу документів (кластеризація та рубрикування);
- побудова та динамічний аналіз семантичної структури текстів;
- виділення ключових тем і інформаційних об'єктів;
- визначення загальної і об'єктної тональності інформації;
- дослідження частотних характеристик текстів.

Впровадження інтелектуальних систем пошуку необхідної інформації в світовій інформаційній системі дає можливість значно спростити такий пошук. Але пошукові системи розроблені для широкого загалу користувачів не завжди зручні в спеціалізованих системах пошуку та аналізу інформації. Тому в даний час в багатьох галузях планується і ведеться розробка методів та програмних продуктів, котрі могли б вести пошук та аналіз інформації в певних параметрах, які задає дана галузь.

Стаття присвячена аналізу сучасного стану математичного та програмного забезпечення опрацювання текстової інформації з різнорідних джерел даних.

2. Основний матеріал

2.1. Аналіз методів опрацювання текстових даних

Метод виділення ключових слів. Використання методу виділення ключових слів з тексту дає можливість пошуку потрібної інформації за короткий проміжок часу.

Ключове слово – слово, або сталий вислів природної мови, яке використовується для вираження деякого аспекта змісту документа (або запиту); слово, яке має істотне смислове навантаження [9]. Воно може служити ключем під час пошуку інформації в інтернеті чи на сторінці сайту. При використанні методу координатного індексування пошукові образи представляють собою множини ключових слів, які, в такому випадку, називають, також унітермами.

Між ключовими словами можуть існувати відношення синонімії, або еквівалентності сенсу, тобто, синонімії з точки зору цієї інформаційно-пошукової системи. Накопичення ключових слів шляхом змістовного аналізу текстів, або алгоритмічно, наприклад, порівнянням слів тексту з фіксованим переліком неключових слів, є важливим етапом при виборі вихідної лексики дескрипторних мов інформаційно-пошукових; відібрані ключові слова об’єднуються, далі, в дескриптори. У дескрипторних словниках (інформаційно-пошукових тезаурусах) даються посилання від ключових слів до відповідних дескрипторів. Ключові слова – це база результатів пошуку. Слід пам’ятати, що ключовим словом може бути не тільки слово, але й словосполучення.

Мистецтво виділення ключових слів, витягання найбільш важливих або характерних фраґментів з одного або багатьох джерел інформації, стало невід’ємною частиною нашого життя. Новини, які пропонуються, – це реферат світових подій дня. Котировки на цінні папери – «сухий залишок» інформації про купівлі-продажі, яку щохвилини породжує ринок. Хоча деякі виробники вже зараз пропонують інструменти для виділення ключових слів, обсяг інформації в мережі росте і оперативно отримувати її коректні зведення стає все складніше. Такі інструменти, як функція Autosummarize в Microsoft Office, системи IBM Intelligent Text Miner, Oracle Context і Inxight Summarizer (компонент пошукового механізму Altavista), безумовно, корисні, але їх можливості обмежені виділенням і вибором оригінальних фраґментів з початкового документу і з’єднанням їх у короткий текст. Підготовка ж короткого викладу має на меті описати основний зміст тексту.

Головна відмінність між засобами виділення ключових слів полягає в тому, що вони, за суттю, формують короткий виклад або набір цитат з певного матеріалу. Обидва типи викладу мають дві основні цілі: визначити найбільш важливу думку повного тексту та виділити ключові слова.Реферат може бути загальним або орієнтованим на специфічного користувача. Реферати першого типу орієнтуються на широке коло читачів; до них не висуваються жодні спеціальні вимоги, оскільки реферат не призначений для якоїсь однієї групи читачів. Реферати другого типу, навпаки, адресовані конкретному користувачеві або групі користувачів з їх специфічними потребами (наприклад, дітям).

До недавнього часу загальні реферати користувалися більшою популярністю, проте, розповсюдження повнотекстових пошукових механізмів і засобів фільтрації інформації, що адаптуються до вимог конкретних користувачів, приводять до того, що реферати, які налаштовуються, набувають все більшого значення.

Використання реферативних методів дозволяє оптимізувати розміри тексту, проте не забезпечує автоматичного пошуку документів. Можна визначити три найважливіші моменти, які не враховані під час виділення ключових слів з тексту:

а) поділ на частини на ключових словах (по форматуванню);

б) виділення ключових слів (за вагою);

в) формування узагальнюючого документа (у вигляді статистики).

Використання запиту ключового слова для вибірки даних з різнотипних джерел. Запит ключового слова є множиною ключових слів $K_1,...,K_n$. Елемент задовольняє запит ключового слова, якщо виконується одна з таких умов [1]:

1) елемент містить як мінімум одне {K1,...,Kn} значення атрибуту (у цьому випадку його називають релевантним елементом);

2) елемент є зв'язаним (у будь-якому напрямку) із доречним (релевантним) елементом (у цьому випадку його називають зв'язаним елементом).

Запити предикату і запити сусіднього ключового слова відрізняються від традиційних структурованих запитів тим, що користувач може конкретизувати ключові слова, на відміну від фіксованих значень і забезпечення лише приблизної структури інформації. [1]

Індекс, що розглядається, базується на розширенні інвертованих списків. Така техніка широко використовується для пошуку інформації. Інвертований список – це двовимірна матриця, де i-й рядок подає індексоване ключове слово K_i, а j-й стовпець моделює значення I_j. Комірка в i-му рядку та j-му стовпці позначається як (K_i, I_j) і містить відомості про кількість подій з ключовим словом K_i в атрибуті елемента I_j. Якщо комірка (K_i, I_j) не є нульовою, то ми кажемо, що елемент I_j є індексований на K_i. Ключові слова впорядковані в алфавітному порядку, а елементи впорядковані за їх ідентифікаторами.

Запит на основі ключових слів можна застосувавати для пошуку у різнотипних джерелах даних. Такий запит може бути трансформований у запит мовою SQL – для реляційних джерел та XML; пошук – для .xls-файлів та текстових файлів.[2]

Побудова інвертованих індексів для визначення методу доступу до розрізнених даних. Розглянемо атрибут предиката $(A, \{K_1,..., K_n\})$ в запиті предиката. Елементи задовольняють предикат, якщо вони містять деякі з ключових слів $K_1,..., K_n$ в їхньому атрибуті A [4]. Щоб опрацювати предикати атрибуту ефективно, індекс повинен показати, які атрибути містять задане ключове слово. Є декілька шляхів фіксації типів атрибуту в індексації.

Перший спосіб – сформувати індекс для кожного атрибуту, але, як показано у [4], це може привести до істотних витрат у структурі індексу. Іншим способом є конкретизація імені атрибуту в комірках інвертованого списку. Проте, цей метод повинен значно ускладнювати відповідь на запит. Рішення, яке пропонує інвертований список, полягає в збиранні даних про імена атрибутів з індексованими ключовими словами, щоб зберегти і розмір індексу, і час пошуку. Кожного разу, коли ключове слово *k* з'являється у значенні атрибуту, додається рядок до інвертованого списку для *k*//*a*//. Для кожного елементу *I* існує стовпець *I*. У комірці (*k*//*a*//, *I*) записується номер події k в атрибути елементів I. Для того, щоб відповісти на запит предикату з атрибутом предикату $(A, \{K_1,..., K_n\})$, необхідно виконати пошук ключового слова для $\{K_1 // A //,..., K_n // A //\}$. Кожне індексоване ключове слово є конкатенацією ключового слова і атрибуту. Предикат асоціації $(R\{K_1,..., K_n\})$на елементі повертає TRUE, якщо вони мають асоціацію типу R з елементами, які містять деякі з ключових слів $K_1,..., K_n$ ув значеннях атрибуту.

Недоліком інвертованого списку в контексті просторів даних є його значна структура (кількість рядків дорівнює кількості можливих об'єктів простору даних та їх характеристик) та розрідженість, оскільки не між усіма об'єктами простору даних можна встановити асоціацію.

Інтеграційні моделі. Проблема інтеграції колекцій текстових інформаційних ресурсів зводиться, переважно, до інтеграції метаданих їхніх джерел, каталогів, класифікаторів, тезаурусів, онтологій тощо. Як ми вже відзначали, ця проблема набула особливої актуальності у зв'язку з розробками електронних бібліотек. Інтеграція тут розуміється як об'єднання колекцій текстових документів з різних джерел у рамках єдиного джерела. Тут найцікавіші методи, що передбачають матеріалізовану інтеграцію метаданих і віртуальну інтеграцію властиво контенту колекцій текстових документів. Такий підхід використовується, наприклад, у системі Соціонет (http://socionet.ru).

У розробках інтегруючих моделей даних використовується також підхід, заснований на інтеграції моделей даних, що підтримуються різними джерелами. Такі інтегруючі моделі забезпечують одночасно й рішення двоїстої задачі – підтримку множини різних подань тих самих даних. Проекти такого роду розроблені ще до початку 80-х років. Як приклад можна навести спробу інтеграції в єдиній моделі даних можливостей мережної моделі даних CODASYL і реляційної моделі даних. У роботах Л. Калініченка [12] була запропонована методологія синтезу інтегрованої моделі даних.

До цієї ж категорії засобів інтеграції даних відноситься розширення мови SQL – компонента нової версії стандарту мови SQL:200n, що отримала назву SQL/XML [43]. Засоби SQL/XML забезпечують можливості подання схем баз даних SQL і реляційних даних у формі XML-Документів, а також реляційне подання інформаційних ресурсів XML у середовищі баз даних SQL.

Нова технологічна платформа Веб, заснована на стандартах XML, в останні роки привертає увагу багатьох фахівців як ефективний інструмент інтеграції інформаційних ресурсів у багатьох практично важливих випадках. Великий інтерес до середовища XML зв'язаний не тільки з можливостями XML як мови опису даних, але й значною мірою з можливістю використання його для транспорту повідомлень у середовищі Веб.

Конструктивний інтерес до засобів інтеграції інформаційних ресурсів Веб і реляційних баз даних проявляють і розроблювачі нових інформаційних технологій для "Всесвітньої павутини". Розроблений стандарт мови запитів XQuery [13] платформи XML втілює функціональність, властиву інтегруючій моделі даних. Базова модель даних цієї мови підтримує ієрархічні й реляційні структури даних і, отже, забезпечує можливості для інтеграції XML-даних і даних реляційних базах даних. Вона дозволяє явно представляти величезні інформаційні ресурси «схованого» Веб – бази даних SQL, до яких у цей час забезпечується доступ у середовищі Веб за допомогою інтерфейсу HTML-Форм.

Інтеграція на основі Web-сервісів має декілька рівнів:

рівень даних – програмні застосування можуть обмінюватись інформацією. Цей рівень передбачає інтеграцію даних і є найпростішим;

об'єктна взаємодія. Тут йдеться про те, що програмне застосування, розташоване на одному сервері, може запускати програмні процеси на іншому;

інтеграція на рівні стандартної семантики. На цьому рівні сервіси можуть «спілкуватися спільною мовою», обходячи технологічні розбіжності;

На першому рівні інтеграції сервіси будуть потребувати лише стандартизації семантики, тобто, під словами «купівля», «пошук» і «статистика» вони повинні розуміти одне й те ж саме. Якщо семантичних розбіжностей між ними немає, інтеграція не має особливих труднощів. Тобто, використовуючи специфікацію WSDL, програмне застосування може «говорити» системно-незалежною мовою [13].

Недоліком Web-інтеграції є низька якість інтегрованих даних, яка виникає внаслідок використання так званої «стандартної семантики» – фізичного збігу типів даних, розмірностей тощо. У результаті цього не враховуються особливості предметної області, що призводить до наявності повторів даних, неточностей, суперечностей (не встановлюється ступінь довіри до джерела) тощо.

Векторне подання тексту. Сучасні методи векторного подання текстової інформації є розвитком моделей векторних просторів VSM (Vector Space Models), запропонованих в [17]. У цих моделях компоненти текстів, такі як слова, словосполучення, фрагменти текстів, цілі документи, представлені багатомірними векторами. Елементи векторів є значення деякої функції від частоти зустрічі компонентів текстів і їхніх контекстів. Ступінь подібності між компонентами текстів q і d визначається величиною подібності між їхніми векторами q і d. Зазвичай використовується деяка монотонна функція між векторами, наприклад косинус cos(q,d), що для нормованих векторів збігається зі скалярним добутком (q,d).

Далі проаналізуємо програмні засоби опрацювання текстової інформації. Під опрацюваням розумітимемо:

1) пошук,
2) забезпечення релевантності,
3) оптимізація (реферування).

2.2. Аналіз засобів реалізації електронних бібліотек

Під інтеграцією даних в електронних системах розуміють забезпечення єдиного уніфікованого інтерфейсу для доступу користувачів до сукупності автономних джерел, які як правило, мають неоднорідність щодо деяких їх властивостей [14]. Своєрідний клас систем інтеграції представляють системи, в яких за основу прийнято технологію Ініціативи відкритих архівів (Open Archive Initiative – OAI) [15, 16]. У більшості відомих систем цієї категорії їх інформаційні ресурси являють собою колекції текстових документів, передусім наукових публікацій, які автономно формуються у вузлах глобальної мережі, підтримуються та адмініструються їх власниками.

Згідно з технологією OAI, передбачається матеріалізована інтеграція у єдиному репозиторії не самих інформаційних ресурсів, що цікавлять користувачів системи інтеграції, а представлених деяким стандартним чином метаданих, що описують колекції інформаційних ресурсів джерел даного архіву і окремі елементи цих колекцій. Збір таких метаданих для репозиторія здійснюється згідно зі спеціально розробленим протоколом Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH) [16], що забезпечує глобальні послуги доступу та пошуку. Існує кілька підходів до вирішення проблеми створення електронних бібліотек з інтегрованими інформаційними ресурсами. Серед них можна виділити такі два класи таких систем: з інтегрованим веденням ресурсів, з розподіленим веденням ресурсів.

Підхід з інтегрованим веденням ресурсів передбачає збір, збереження й оброблення інформаційних ресурсів у єдиному репозиторії. Такий підхід видається доцільним у випадку, коли інформаційні ресурси організаційно породжуються в одному місці і безпосередньо належать одному власнику. Однак, для централізації всіх ресурсів, такий підхід не зовсім підходить, оскільки потребує вирішення цілого ряду складних організаційно-технічних задач і, насамперед, тих, що спрямовані на збір вихідних інформаційних ресурсів. Крім того, він потребує створення розвинутої структури ведення такої НЕБ.

Підхід з розподіленим веденням ресурсів припускає, що існує багато організацій, які здійснюють самостійне створення і ведення електронних бібліотек і надають можливість доступу до цих ресурсів, включаючи також і організацію пошуку необхідних ресурсів. Крім того, існує "надбудова" над ними, що дозволяє робити пошук за цими ресурсами і, за наявності відповідних умов, надавати доступ до самих ресурсів.

Сьогодні відомо два концептуальних рішення цього підходу. Перше припускає існування механізму перехресного пошуку за багатьма архівами, коли всі ресурси, бібліографічні описи та пошуковий сервіс знаходиться в організації. Пошук здійснюється шляхом безпосереднього звернення до всіх або до вибраних користувачем електронних бібліотек з наступним зведенням одержаних результатів у єдиний список. Друге – пропонує здійснювати збір метаданих, що описують інформаційні ресурси "на місцях" для того, щоб можна було надати централізований пошук в одному місці на основі зібраних метаданих. За суттю це є деякий аналог інтегрованого електронного каталогу.

2.3. Аналіз програмних засобів опрацювання текстів

OBSERVER (http://siul02.si.ehu.es/~jirgbdat/OBSERVER). Ця система пропонує підхід використання вже існуючих онтологій для доступу до гетерогенних, розподілених і незалежно розроблювальних репозиторіях даних [11]. Реалізація такого підходу – ідеологія брокера онтологій предметних областей. Передбачається, що існує множина заздалегідь створених онтологій предметних областей, і користувачу необов’язково “підбудовуватися” під конкретну онтологію. Користувач формулює свій запит на деякій мові, у термінах однієї чи декількох онтологій, і брокер «шукає» релевантні документи, виконуючи транслювання запиту в придатні онтології, а в разі потреби, і сполучення декількох онтологій для більш точної відповіді на запит.

OntoSeek [10]. Ця система розроблена для контекстного отримання інформації з он-лайнових “жовтих сторінок” та каталогів продуктів. Система може працювати як з однорідними, так і з неоднорідними каталогами продуктів. Для точної фіксації контексту може бути застосований інтерактивний підхід, коли користувач поступово уточнює зміст ключових слів, за допомогою лінгвістичної бази даних WordNet. WordNet – це лінгвістична база даних, що складається із сінсетів(synsets) – груп слів, еквівалентних за змістом. WordNet є водночас і лексичним словником

(створеним для декількох європейських мов), і онтологією, що представляє зв'язки між словами у словнику. Опис ресурсу реалізується у вигляді лексичного концептуального графа, де вершини відповідають словам, а іменовані дуги – семантичним відношенням між словами (наприклад, відношення типу "частина", або "підклас", або ін.), назви вершин і дуг також беруть із WordNet, під час створення концептуального графа конкретного ресурсу. Знаходження ресурсів, релевантних до запиту користувача, базується на порівнянні онтологій (лексичних концептуальних графів) цих ресурсів. А саме, при відборі ресурсів, відповідних до запиту користувача, OntoSeek виконує порівняння концептуального графа запиту із існуючими концептуальними графами ресурсів або з частинами цих графів. OntoSeek має централізований сервер, на якому знаходиться база даних лексичних концептуальних графів відомих системі ресурсів, але створення таких графів виконується з боку клієнта. Підхід, використаний в OntoSeek, відрізняється від підходу, який застосовується у моделі W3C Resourse Description Framework (W3C RDF, http://www.w3c.org). У RDF опис структури даних (тобто, схема даних у вигляді <subject, predicate, object>), додається прямо у HTML/XML документ, а не зберігається окремо. Ніяких додаткових умов щодо семантичної узгодженості даних RDF не вимагає.

TextAnalyst. TextAnalyst розроблений як інструмент для аналізу змісту текстів, смислового пошуку інформації, формування електронних архівів, і надає користувачеві наступні основні можливості:

1) аналіз змісту тексту з автоматичним формуванням семантичної мережі з гіперпосиланнями – отримання смислового портрета тексту в термінах основних понять і їх смислових зв'язків;

2) аналіз змісту тексту з автоматичним формуванням тематичного дерева з гіперпосиланнями – виявлення семантичної структури тексту у вигляді ієрархії тем і підтем;

3) смисловий пошук з урахуванням прихованих смислових зв'язків слів запиту зі словами тексту;

4) автоматичне реферування тексту – формування його смислового портрета в термінах найбільш інформативних фраз;

5) кластеризація інформації – аналіз розподілу матеріалу текстів за тематичними класами;

6) автоматична індексація тексту з перетворенням в гіпертекст;

7) ранжування всіх видів інформації про семантику тексту за «ступенями значимості» з можливістю варіювання детальності її досліджень.

Забезпечення релевантності. Основною задачею перелічених систем є пошук інформації у великих повнотекстових масивах [8]. В базі даних таких систем можуть закачуватися будь-які текстові джерела інформації, у тому числі великого обсягу: енциклопедії, довідники, архіви періодичних видань, цілі бібліотеки спеціальної літератури, архіви документів корпорацій, спеціалізовані архіви типу історичних, патентних, судових, розшифровки розмов, протоколи і багато що інше. У відповідь на конкретний запит система видає множину посилань. Далі система має обробити кожне посилання і видати всі відповідні тексти, тобто система має шукати не просто документи, а інформацію, що міститься в них.

Існуючі технології пошуку недостатньо ефективні, щоб знайти у великій кількості різнорідній інформації глобальної мережі відомості, що відповідають запиту. Тому розвивається процес дворівневого пошуку: перший – тематичний пошук і відбір даних в повнотекстову базу даних, другий – пошук в повнотекстовій базі даних. Всі пошукові системи WWW побудовані за принципом Single Shot Relevancy, тобто «релевантність з першого попадання»: ми робимо один запит і одержуємо потрібні результати. Ми можемо якось змінити запит, але спеціальних механізмів, що забезпечують зворотний зв'язок, не передбачається.

Підхід Single Shot Relevancy відповідає ідеї пошуку, вираженій терміном search, що припускає приблизність, він забезпечує достовірність, а в корпоративних умовах не менше важлива точність, тут важливо виявлення саме того документа, що потрібен, і про існування якого користувач знає наперед. Для того, щоб підвищити точність пошуку по запитах (drill down analysis – «аналіз з підвищеним рівнем деталізації») необхідно зробити пошук «інтерактивним». Для цього користувач одержує можливість, використовуючи зворотний зв'язок, коректувати запити і поступово добиватися необхідної йому точності.

Для ефективності такої процедури потрібні нові методи представлення результатів пошуку, простого списку сторінок з короткою анотацією, як це роблять звичайні пошукові машини, недостатньо. Результати можуть бути одержані в текстовій формі, з вказівкою зв'язків між окремими компонентами. Існують великі перспективи у графічних формах представлення результатів пошуку. Особливо слід виділити роботу з мультимедійними даними, жодна пошукова система не працює з такими даними. Розв'язання цієї задачі потребує якісно нових підходів. Сучасні пошукові машини побудовані на декількох основних принципах: дані зберігаються у вигляді окремих документів, тому для підвищення продуктивності можна використовувати розподілену архітектуру, використовуючи метадані, можна якимось чином визначати значення документа

і, спираючись на моделі, добиватися необхідної релевантності.

Застосування у пошукових машинах. Серед багатьох популярних пошукових систем слід визначити найбільш відомі системи: dtSearch, Greenstone, Google Search Appliance, Google Custom Search, Google Desktop, Autonomy, RetrievalWare, Моделі і засоби систем баз даних і знань Яndex.ServerStandard 3.0, PolyAnalyst, МЕТАТЕКА. Далі розглянемо їх основні можливості.

Пошукова система dtSearch. Основне призначення програми dtSearch 7.0 – пошук інформації в локальному і мережному оточенні. Вона містить індексатор документів, менеджер бібліотек індексів, індексатор даних, що знаходяться на CD. Система забезпечує пошук інформації різних типів, і на різних мовах, включаючи zip, rtf, pdf, html, xml, документи Microsoft Office (Word, Excel, PowerPoint) і WordPerfect. Підтримується кодування Unicode. Допускаються декілька видів пошуку, а саме морфологічний, фонетичний пошук, а також пошук синонімів і пошук в словах з орфографічними помилками. Для лінгвістичного опрацювання текстів dtSearch використовує засоби WordNet. Розроблені додаткові засоби підтримки української мови.

Система Greenstone є Open Source-рішенням для створення "цифрових бібліотек". Включає пошук з попередним індексуванням документів різних форматів, і перш за все doc і pdf, які можуть бути представлені і у вигляді архівних файлів. Система створює каталог документів, конвертує їх в html-формат, а потім забезпечує віддалений доступ до бібліотек та інших ресурсів за допомогою браузера.

Програмно-апаратний комплекс *Google Search Appliance* забезпечує пошук документів в рамках корпоративної мережі. Пошуковий механізм комплексу забезпечує роботу більш ніж з 200-ми типами файлів.

Google Desktop Search (GDE). Безкоштовна локальна версія відомої пошукової системи Google. На жаль, як сам Google Desktop Search, так і ряд інших безкоштовних зарубіжних пошукових систем поки малопридатні для текстових масивів на українській мові. Вони не працюють з українською морфологією, погано індексують українськомовні текстові масиви.

Технологія компанії *Autonomy* для корпоративних систем є інструментарієм для автоматизованого керування інформаційними потоками. Основні наукові принципи Autonomy базуються на інформаційній теорії Клода Шеннона, байесовских ймовірностей і нейронних мережах. Концепція адаптивного моделювання ймовірності дозволяє системі Autonomy ідентифікувати шаблони в тексті документа і автоматично визначати подібні шаблони в масиві інших документів.

Обробляючи шаблони рядків у документах, система Autonomy визначає кореляцію образів і виявляє закономірності серед великих масивів документів. При цьому не враховуються ніякі специфічні правила (зокрема і лінгвістичні). Оскільки система не базується на визначених раніше ключових словах, вона може працювати з будь-якими мовами.

Інформаційно-пошукова система *RetrievalWare* є засобом повнотекстового і атрибутивного пошуку. До документів, з якими RetrievalWare здатна працювати, належать тексти в різних форматах і кодуваннях, електронні таблиці, бази даних, поштові повідомлення і т. д. Система володіє додатковим інструментарієм, який дозволяє налаштуватися на підтримку документів специфічних форматів. В основі системи покладена технологія адаптивного розпізнавання образів, яка базується на нейронних мережах для обробки інформації і діє як система, яка виділяє з масиву збережену інформацію і індексує бінарні образи, що самоорганізовуються. До переваг застосування цієї технології для пошуку текстової інформації можна зарахувати здійснення нечіткого пошуку, мовну незалежність, малі обсяги індексних файлів.

Основою технології *семантичного пошуку* є використання семантичних мереж, які описують значення слів природної мови і зв'язки між поняттями, що визначаються ними. Нема підтримки української морфології. Семантична мережа словника цієї мови включає близько 40 тисяч семантичних груп в базовому варіанті. Це дозволяє користувачу вводити запит природною мовою і система сама шукає всі документи, контекст яких збігається з контекстом запиту. Застосування семантики дозволяє враховувати загальний контекст документа.

Система *Яndex.ServerStandard* 3.0 є системним сервісом для організації повнотекстового пошуку інформації у заданій колекції документів. Складається з двох основних логічних частин: індексатора і пошукового серверу. Індексатор аналізує документи, серед яких має здійснюватися пошук, і зберігає інформацію про них у спеціальних індексних файлах. Пошуковий сервер після запуску знаходиться в постійному очікуванні запитів, які можуть бути представлені на природній мові. Пошук може здійснюватися з урахуванням морфології мови, в одній або декількох колекціях документів. Яndex.Server 3.0 підтримує формати html, xml, rtf, pdf, doc, mp3 і багато інших. Вміст документів, що індексуються, також може бути отриманий при зверненні до довільної бази даних, зокрема, MySQL і MS SQL. Система надає можливість кластеризації результатів пошуку (групує знайдені документи відповідно до зовнішніх атрибутів).

Ядром механізму обробки контента *InfoStream* є повнотекстова інформаційно-пошукова система InfoReS. Технологія InfoStream дозволяє створювати повнотекстові бази даних і здійснювати пошук інформації, формувати тематичні інформаційні канали, автоматично створювати рубрики інформації, формувати дайджести, таблиці взаємозв'язків понять (як вони зустрічаються в мережних публікаціях), гістограми розподілу вагових значень окремих понять, а також динаміки їх зустрічі за часом. За допомогою InfoStream можна обробляти дані у форматах Microsoft WORD, rtf, pdf, і всіх текстових форматах (простий текст, html, xml). Технології InfoStream дозволяють створити комплекс підтримки документального інформаційного сховища, в якому реалізується інтегроване інформаційно-пошукове середовище на основі веб-рішень.

Засоби реферування. Далі проаналізуємо засоби автоматичного реферування. Результати порівняння подано у таблиці 1.

Як бачимо, жодна з розглянутих систем не дозволяє одночасно забезпечити розв'язання задач поділу на частини, виділення ключових слів та опрацювання множини документів.

Висновки

Отже, під час опрацювання текстової інформації з множини розрізнених інформаційних ресурсів необхідно виділити такі задачі:

1) виділення заголовку, авторів, ключових слів та побудова концептуальної моделі тексту;
2) інтеграція у повнотекстову базу даних;
3) пошук у повнотекстових базах даних;
4) забезпечення релевантності запиту;
5) зменшення обсягів текстової інформації та узагальнення тексту з кількох джерел (побудова множинного реферату).

Порівняльна характеристика систем автоматичного реферування

Системи автоматичного реферування\задачі	Поділ на частини	Ключові фрази	Формування узагальнюючого документу
Auto Sumarizer (MS Word)	+		
CONTEXT			+
Data Hammer		+	
DimSum			+
Extractor		+	
GE Summarizer	+		+
Intelligent Miner	+		
IntellScope	+		+

Продовження таблиці

Системи автоматичного реферування\задачі	Поділ на частини	Ключові фрази	Формування узагальнюючого документу
InText			+
InXihtSummarizerPlus	+	+	
ProSum	+		
Search'2005 Developer Kit	+	+	
SMART			+
SUMMARIST			+
TexNet32			+

На основі проведеного аналізу найпопулярніших систем сьогодні мало справляються з виділеними задачами. Зокрема, немає методів та засобів, які б дозволяли розв'язувати усі задачі одночасно. Також проблемою сучасних систем є те, що вони зазвичай орієнтуються на тексти англійською мовою. Для текстів українською мовою розроблено лише неповні онтології у деяких предметних областях. Областю, яка має достатньо багато відомих авторів термінів і інформація, яка є доступна, є область інформаційних технологій. Такі проблеми ускладнюють задачі опрацювання тексту та роботу з ними. Пошук потрібної інформації стає важчим, враховуючи те, що з кожним днем обсяг інформації збільшується. Особливо це актуально для наукових установ та бібліотек. Подальші дослідження будуть спрямовані на розроблення методів та засобів опрацювання наукових статей з метою виділення наукових шкіл та прогнозування їх розвитку.

1. Xin Dong . Indexing Dataspaces // Xin Dong, Alon Halevy. - SIGMOD'07, June 11–14, 2007, Beijing, China. 2. S. – Y. Chien, Z. Vagena, D. Zhang, V. J. Tsotras, and C Zaniolo. Efficient structural joins on indexed XML documents. In Proc. of VLDB, 2002. 3. Розенталь М. Краткий философский словарь / Розенталь М., Юдин П. М.: Политиздат, 1951. – 450 с. 4. Gruninger M., Fox M. Methodology for the Design and Evaluation of Ontologies // Proceedings of IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing. – 1995. – P. 231–238. 5. Гаврилова Т.А. Базы знаний интеллектуальных систем / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб: Питер, 2001. – 384 с. 6. Дубинский А.Г. Некоторые вопросы применения векторной модели представления документов в информационном поиске // Управляющие системы и машины. – 2001. – №4. – С. 77 – 83. 7. Дрейфус Х. Чего не могут вычислительные машины: Критика искусственного разума / Х. Дрейфус. – М.: Прогресс, 1978. – 333 с. 8. Андон П.І., Дерецкий В.А. Процесори пошуку та аналізу природномовної текстової інформації в аналітичних системах – 2001.– № 3 – 4. – С. 144–163. 9. Енциклопедія кібернетики, т. 1, с. 457. 10. Guarino N., Masolo C., Vetere G. Content-Based Access to the Web. IEEE Intelligent Systems, May/June 1999, p.70 – 80. 11. Mena E., Kashyap V., Sheth A., Illaramendi A. OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-Existing Ontologies. In Proceedings of the First IFCIS International Conference on Cooperative Information Systems (CoopIS'96), Brussels (Belgium), June. IEEE Computer Society Press, 1996. 12. Калиниченко Л.А. СИНТЕЗ – язык определения, проектирования и программирования интероперабельных сред неоднородных информационных ресурсов, ИПИ РАН, 1993, 115 с. 13. Grinev, M., Kuznetsov S.: Towards an Exhaustive Set of Rewriting Rules for XQuery Optimization: BizQuery Experience, 6th East-European Conference on Advances in Databases and Information Systems (ADBIS), LNCS 2435 (2002) 340 – 345. 14. Когаловский М.Р. Тенденции развития технологий управления информационными ресурсами в электронных библиотеках // Тр. VIII Всероссийской научн. конф. Электронные библиотеки: перспективные методы и технологи. – Суздаль, Россия. – 2006. – С. 46–55. 15. Лагозе К., Ван де Зомпель Г. Инициатива «Открытые архивы»: создание среды с высокой степенью интероперабельности. Электронные библиотеки. – 2001. – Т. 4. Вып. 6. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2001/part6/LS. 16. The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14. http://www.openarchives.org/ /OAI/2.0/openarchivesprotocol.htm. 17. Salton G. Automatic Text Processing: The Transformation, Analisys, and Retrieval of Information by Computer. – Addison-Wesley, Reading, MA. – 1989. – 530 p.