НТІ

НАУКОВО-ПРАКТИЧНИЙ ІНФОРМАЦІЙНИЙ ЖУРНАЛ

НАУКОВО-ТЕХНІЧНА ІНФОРМАЦІЯ

2(28)'2006

ПЕРЕДПЛАТНИЙ ІНДЕКС 22879

ЗАСНОВНИКИ:

Міністерство освіти і науки України

Український інститут науково-технічної і економічної інформації (УкрІНТЕІ)

Засновано 11 березня 1999 р.
Свідоцтво про реєстрацію Міністерства інформації України: серія КВ, № 3720

Постановою президії ВАК України від 9лютого 2000 р. № 2-02/2 журнал «НТІ» включено до переліку №4 наукових фахових видань України в галузі технічних наук

Постанова президії ВАК України від 13.12.2000 р. № 2-01/10 ухвалила зараховувати статті, опубліковані в журналі «НТІ», як фахові в галузі економічних наук

СЛОВО ГОЛОВНОГО РЕДАКТОРА

Цей номер журналу присвячено питанням теорії і практики науково-інформаційної та інформаційно-аналітичної діяльності, в тому числі використанню науково-технічної і економічної інформації для аналітичної обробки при підготовці проектів управлінських рішень.

Традиційно сторінки журналу надаються для обговорення проблем економіки науково-технологічної та інноваційної діяльності, стратегії технологічного передбачення в інноваційній діяльності, методам прогнозування розвитку економіки та оцінки економічної ефективності проектів.

Вважаємо, що читачам журналу буде цікаво ознайомитися з методом обробки даних космічного моніторингу як окремим специфічним видом наукової інформації, а також запропонованою концепцією анотованого пошуку інформації та принципами побудови інтелектуальних навчальних систем.

З повагою.
Головний редактор В. Пархоменко В. Пархоменко

Рекомендовано до друку
Вченою радою Українського інституту науково-технічної і економічної інформації
та редакційною колегією журналу

ВІД РЕДАКЦІЇ:

Журнал НТІ друкує матеріали, що відповідають профілю видання.

Редакція рецензує наукові статті, листується з читачами на сторінках журналу та на сайті.

Відповідальність за достовірність інформації, що міститься в друкованих матеріалах. несуть автори.

Передплатити
журнал НТІ можна за каталогом Укрпошти (індекс **22879**) у відділі маркетингових досліджень по забезпеченню реалізації інформаційної продукції УкрІНТЕІ, **тел. 528-23-45**.

Адреса редакції:
03680, Київ, вул. Горького, 180, УкрІНТЕІ,
тел. 528-25-57
htpp: //www:uintei.kiev.ua

Відповідальний за випуск
В.І. Воронков

Редактор Н.М. Кучеренко
Технічний редактор Л.М. Басова
Комп'ютерна верстка Б.О. Грабовський
Комп'ютерний набір Н.В. Дудник

Підписано до друку 8.06.2006 р.
Формат 60 ×84 1/8. Друк. арк. 7,5. Обл.-вид. арк. 9,6
Тир. 300 прим. Зам. 177в.
03680, Київ, вул. Горького, 180.
Видавничо-поліграфічне відділення УкрІНТЕІ

КОНЦЕПЦІЯ АНОТОВАНОГО ПОШУКУ


С.М. Брайчевський,
канд. фіз-мат. наук,

Д.В. Ланде,
канд. техн. наук

Вступ. Парадокс щодо розвитку мережних пошукових систем полягає в тому, що їхнє технічне вдосконалювання в рамках традиційної парадигми неминуче призводить до лавиноподібного зростання баз даних, і відповідно, обсягів релевантних вибірок, які кінцевий споживач у підсумку не в змозі обробити [1]. Сучасні технології надають можливість здійснювати витончені операції над даними, але чим ефективніше вони застосовуються, тим менш придатним виявляється результат. Схоже, що технічний прогрес у цьому випадку не поліпшує, а погіршує ситуацію.

Існуючі інформаційно-пошукові системи первісно проектувалися для забезпечення релевантності вибірки в поєднанні з вимогою повноти пошуку, але саме в цьому і полягає їхній головний недолік. Неконтрольований рівень пертинентності вибірки при цьому різко знижує ймовірність одержання користувачем саме тієї інформації, яка йому потрібна.

Причини надлишковості результатів стандартного інформаційного пошуку можуть бути розділені на дві якісно різні категорії: дублювання інформації та інформаційна невідповідність. Істотним є те, що приналежність документа до числа дублів має цілком об'єктивний характер і може визначатися автоматично на підставі формальних критеріїв. Навпаки, інформаційна невідповідність породжує проблеми суто суб'єктивного характеру, тому що машина не в змозі визначити, чи відповідає зміст даного документа інформаційним потребам даного користувача.

Тому зрозуміло, що пошукові технології повинні бути розширені за рахунок застосування додаткових семантичних засобів, що дають змогу або скоротити розрив між рівнями релевантності й пертинентності, або певним чином його компенсувати.

Модифікація задачі пошуку

Найбільш перспективним з існуючих сьогодні напрямів, безсумнівно, є автоматичне групування результатів пошуку [2], тобто розбивка релевантної вибірки документів на кластери. Разом із тим це не вирішує проблему по суті, оскільки хоча й допомагає орієнтуватися в результатах пошуку, але аж ніяк не сприяє скороченню їхніх обсягів. Головна перевага автоматичного групування полягає в ієрархічній організації результатів пошуку. Це дає змогу на першому етапі мати справу з обмеженим набором кластерів, а потім уже переходити до складу того або іншого кластера. Проте складність полягає в тому, що розбивка вибірки на групи здійснюється на підставі близькості документів, що розуміється формально. Ця обставина, природно, призводить до того, що кінцевий ефект залежить від багатьох, у тому числі й випадкових, факторів і має явно неконтрольований характер.

Особливої актуальності набувають підходи, що дають змогу переформулювати задачу пошуку таким чином, щоб його результати дійсно могли бути без зусиль сприйняті користувачем.

Одним із головних принципів, покладених в основу більш адекватних підходів, на наш погляд, є відмова від вимоги повноти пошуку.

Вартою уваги є постановка задачі попередньої обробки початкової сукупності документів, яка передбачає сформування деякого ефективного набору даних, що відбиває в розумному наближенні її зміст і призначений для подальшого пошуку по ньому.

Сама по собі така постановка задачі не є новою: вона широко й успішно застосовується у сфері автоматичного реферування документальних потоків. Саме продуктивність подібної методики в суміжній сфері й змушує нас уважно придивитися до її можливостей стосовно інформаційного пошуку.

Анотований пошук

У технологічному плані пропонується реалізація принципу попередньої обробки текстового матеріалу за допомогою методик, характерних для іншої сфери інформаційних технологій, а

саме контент-аналізу. Така обробка передбачає автоматичне виділення найбільш значущої інформації і відсівання "сміття", що дасть змогу споживачеві працювати з наборами даних, досить обмеженими за обсягом, і в разі правильної організації може істотно підвищити рівень пертинентності результатів пошуку. Концепція також передбачає свого роду кластеризацію, однак розподілу за групами підлягає не тільки релевантна вибірка, але й вихідний набір документів, в якому ведеться пошук.

У рамках концепції використовуються терміни "анотований пошук" і "анотована база даних", оскільки основні алгоритми пошуку і структура бази даних нагадують ті, які використовуються під час автоматичного реферування.

Центральна ідея пропонованої концепції полягає в тому, що релевантність документа варто визначати стосовно деякого його інформаційного образу. Причому останній має бути побудований саме так, щоб відбивати основний зміст документа. Такий образ документа (або групи документів) у рамках даної концепції називається анотацією.

Структура і форма анотації не мають принципового значення, але в кожному разі вона повинна містити впорядкований набір термінів та/або фраз, що входять до складу відповідного документа і мають певний рівень вагових значень. Вага може характеризувати значимість термінів або фраз у документі і може визначатися різними методами залежно від властивостей предметної сфери та поставленої задачі. Крім того, оскільки споживача в остаточному підсумку цікавлять тексти документів, сукупність анотацій має бути доповнена системою відповідних посилань. Разом вони утворять деякий набір метаданих, що має бути включений у загальну індексну систему бази даних.

На рисунку наведено схему функціонування анотованої бази даних.

Технологічна реалізація анотованого пошуку

Як інформаційно-технологічна основа розглядається база даних традиційної інформаційно-пошукової системи з властивою їй структурою, включаючи, наприклад, індексні, інверсні, словникові таблиці тощо.

Створення анотованої бази даних передбачає створення бази даних пошукових образів первинних документів та їхню кластеризацію, тобто автоматичне формування груп документів із близькими за деякими критеріями пошуковими образами (ПОД).

У разі формування анотованої бази даних найважливішим аспектом є формування бази даних анотацій, тобто пошукових образів кластерів (ПОК), які, власне, і використовуватимуться в процесі пошуку. Ця база даних пов'язана з базою даних кластерів, кожен запис якої відповідає певному кластеру та включає, крім усього іншого, його опис (виконаний методами автоматичного реферування).

Методи автоматичного реферування, а точніше квазіреферування, заснованого на переважному використанні методів статистичного аналізу текстів, використовуються, з одного боку, для створення ПОД, з іншого — і описів, доступних користувачам.

Задача повнотекстового пошуку по надвеликих текстових масивах може виявитися неефективною, наприклад, у романі "Війна і мир" Л.Толстого можна знайти більшість лексем російської мови. Пошук по анотованих текстах у таких випадках вирішує проблему точності. Тобто, замість пошуку по повних текстах виявляється доцільним проводити пошук по анотаціях — пошукових образах документів. Хоча квазіреферат часто для великих текстів виявляється утворенням, що лише віддалено нагадує вихідний текст, який при цьому найчастіше не сприймається людиною, але саме як пошуковий образ документів, що містить зважені ключові слова і фрази, він може приводити до цілком адекватних результатів при повнотекстовому пошуку.

Квазіреферат у більшості відомих систем будується з текстових фрагментів, що мають найбільші вагові значення. Загальна вага текстового блоку на цьому етапі визначається за формулою [3]:

$$Weight = Location + KeyPhrase + StatTerm.$$

Коефіцієнт *Location* визначається розташуванням блоку у вихідному тексті та залежить від того, де з'являється даний фрагмент — на початку, в середині або наприкінці, а також чи використовується він у ключових розділах тексту, наприклад, у висновку.

Ключові фрази (*KeyPhrase*) являють собою конструкції-маркери, що резюмують, типу: "на закінчення", "у даній статті", "відповідно до результатів аналізу" і т.п. Ваговий коефіцієнт ключової фрази може залежати також від оцінного терміна, наприклад, "відмінний".

Статистична вага текстового блоку (*StatTerm*) обчислюється як нормована за довжиною цього блоку сума ваг термінів, що входять у нього — слів і словосполучень. Після виявлення певної, заданої коефіцієнтом необхідного стиснення, кількості текстових блоків з найвищими ваговими коефіцієнтами, вони об'єднуються для побудови квазіреферата.

Слід зазначити, що не тільки анотації у вигляді ПОК, але й описи окремих елементів у базі даних анотацій, доступній на етапі пошуку, створюються на основі засобів автоматичного реферування, які на цьому етапі не враховують інформаційних потреб користувачів, виражених

пошуковими приписаннями (запитами).

У рамках даної концепції передбачається використання методів квазіреферування, перевага яких полягає в простоті реалізації.

У разі звертання користувачів до бази даних передбачається така процедура: запит користувача відпрацьовується за базою даних анотацій, після чого шляхом пошукової процедури виконується формування набору релевантних кластерів, найменування та описи яких, з одного боку, можуть пред'являтися користувачам (якщо їхня кількість не перевищує заданої заздалегідь), а з іншого — якщо кількість результатів пошуку (кластерів) перевищує це значення, то результати пошуку автоматично групуються, утворюючи суперкластери, перелік яких і пред'являється користувачеві.

Отже, в останньому випадку користувачеві пред'являються назви суперкластерів та їхні описи — реферати, складені автоматично вже з урахуванням запитів користувачів. Тобто, вага текстових фрагментів у цьому випадку описується уточненою формулою:

$$Weight = Location + KeyPhrase + StatTerm + UserPref$$

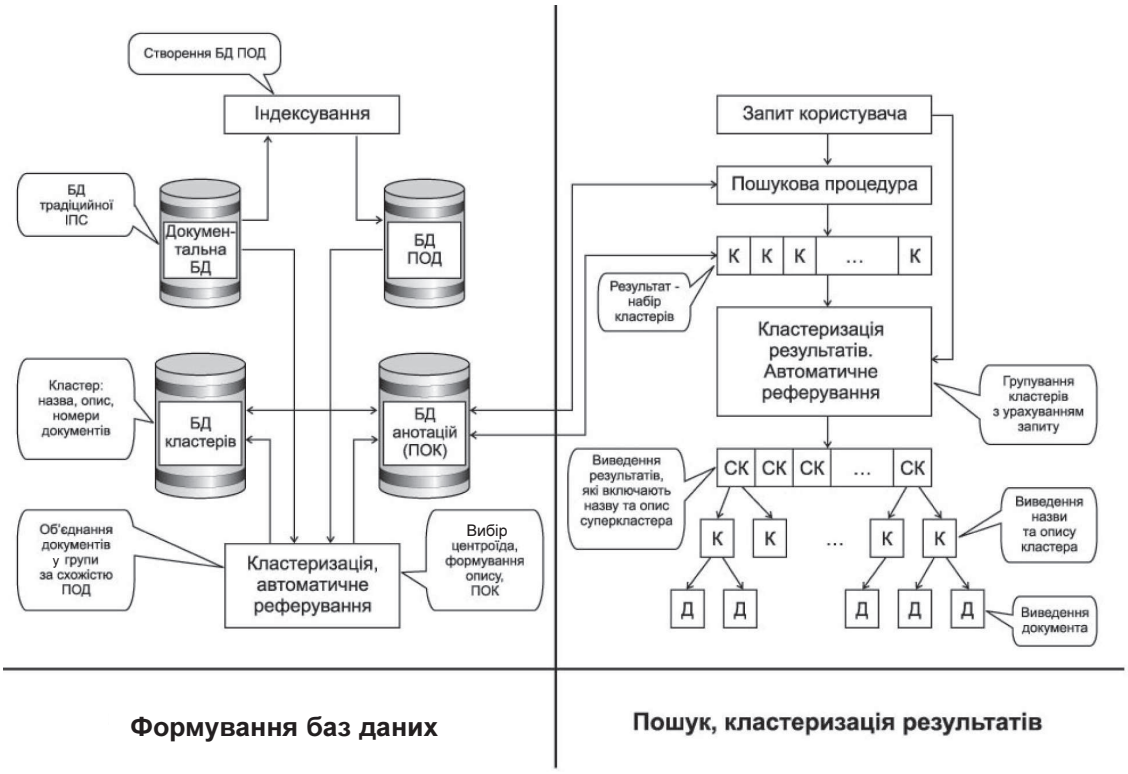
Коефіцієнт *UserPref* — переваги, що надає користувач, залежать від того, наскільки слова і словосполучення, наведені в його запиті, присутні в даному фрагменті.

Надання результатів пошуку може здійснюватися різними способами, залежно від особливостей предметної сфери, структури документальної бази даних, характеру інформаційних потреб користувачів тощо. Як ми зазначили вище, самі анотації є пошуковими образами — внутрішніми елементами системи і користувачеві у вихідному вигляді не пред'являються. Тому припускається, що з метою адекватного відображення результатів пошуку кожен побудований кластер забезпечується описом, що також будується автоматично та видається користувачеві як "етикетка" кластера, яка, на відміну від анотації, являє собою зв'язний текст. За бажанням, користувач може переглянути всі документи, що входять до складу даного кластера.

Припускається, що за такої організації пошуку релевантними виявляться лише ті документи, для яких пошукові терміни запиту користувача є інформаційно-значущими, оскільки самі анотації за своєю природою мають саме таку властивість. Наявність у них виключно термінів або фраз із досить великими ваговими значеннями перешкоджає потраплянню в релевантну вибірку документів, в яких пошукові терміни присутні у вигляді інформаційного шуму.

Висновки

Слід зазначити, що наведена модель у цей час ще не реалізована повністю у вигляді програмно-технологічного забезпечення, однак окремі елементи вже створені й пройшли апробацію. Отже, слід запустити цю модель на реальних надвеликих обсягах даних. До реалізованих елементів належать: традиційні повнотекстові інформаційно-пошукові системи, включаючи авторську розробку — систему InfoRes; алгоритми автоматичного реферування; механізми кластеризації як статичних, так і динамічних масивів інформації, які знаходять уже сьогодні застосування, наприклад, у разі виявлення основних сюжетів у системі контент-моніторингу InfoStream; адаптивні інтерфейси уточнення запитів до



Архітектура і модель функціонування анотованої бази даних

інформаційно-пошукової системи.

Надана модель орієнтована на практичну реалізацію і в явному вигляді містить ряд технологічних обмежень, головне з яких пов'язане з тим, що на етапі індексування пошукові образи документів створюються без урахування запитів користувачів. Оскільки ПОД не є повною копією документів, то заздалегідь не можуть бути враховані всі нюанси інформаційних потреб користувачів, а це може позначитися не тільки на повноті, але й на релевантності. Вирішити цю проблему можуть лише витончені інтелектуальні алгоритми автоматичного реферування.

Разом з тим пропонована організація пошуку надасть можливість вирішити такі важливі задачі:

— автоматичне групування документів і тим самим скорочення реального обсягу простору пошуку;

— пред'явлення користувачеві винятково інформаційно-значущих документів;

— у разі необхідності — виключення дублів з результатів пошуку за збереження їх у самій базі даних.

Згадаємо, що середня довжина запиту до пошукової системи в Інтернет не перевищує двох-трьох слів, можливо в тому числі і через це основні проблеми користувача зводяться до вирішення проблеми релевантності-повноти, і зрештою— пертинентності видачі. Очевидно, пропонована система організації пошуку надасть можливість істотно підвищити його привабливість із погляду користувача.

Через зростаючі обсяги інформації пошукові системи вже сьогодні не в змозі надати користувачеві все те, що йому потрібно з наявного в Інтернеті. Тому завдяки реалізації даної концепції навіть на першому етапі пошуку користувач зможе отримувати відносно невелику і змістовну вибірку.

ЛІТЕРАТУРА

1.Современные информационные потоки: Актуальная проблематика / Брайчевский С.М., Ландэ Д.В. // "Научно-техническая информация", серия 1, № 11. — 2005. — С. 21—33

2. *Григорьев А.Н., Ландэ Д.В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Труды Международного семинара "Диалог'2005". — 2005. — С. 109—111.

3.*Ландэ Д.В.* Поиск знаний в Internet. Профессиональная работа. — М.: "Вильямс", 2005. — 272 с.

ЕТАПИ ТА МЕТОДИ ОБРОБКИ ДАНИХ КОСМІЧНОГО МОНІТОРИНГУ ЗЕМНОЇ ПОВЕРХНІ



О.Ф. Дубина, канд. техн. наук,

С.О.Кондратенко, канд. техн. наук,

Р.М.Осадчук

Нині існує досить багато систем, призначених для отримання інформації про земну поверхню і об'єкти, розташовані на ній. Ця інформація досить різноманітна і залежить від багатьох чинників (задач, що вирішуються, характеристик і параметрів складових системи, стану навколишнього середовища тощо). Однією з основних таких систем є система дистанційного зондування Землі (ДЗЗ). Дані, що отримують за допомогою ДЗЗ, використовуються в картографії, землекористуванні, природознавстві, геодезії, метеорології, військовій сфері діяльності й ін.

Для отримання необхідних даних за допомогою системи ДЗЗ первинна інформація має пройти певні, визначені рівні обробки. Обробка даних ДЗЗ (*табл.* 1), відповідно до світової практики передбачає кілька рівнів обробки [1]

У загальному випадку обробка даних дистанційного зондування включає в себе три етапи: попередню, первинну, вторинну (тематичну) обробки.

Розгорнутий перелік рівнів обробки даних ДЗЗ і склад операцій, що виконуються при цьому, надано в *табл. 2* [1].

На *першому* етапі (рівні ROW, 0), після прийому супутникових даних, записування їх на магнітний носій і виконання необхідних декодувальних і коригувальних операцій відбувається перетворення даних (з урахуванням калібрувань), переданих з космічного апарата, безпосередньо в космічний знімок (наприклад, синтез радіолокаційних зображень з радіоголограм, переданих по радіолінії), а також перетворення їх у формати, зручні для наступних видів обробки. На цьому етапі може проводитися попередня прив'язка знімків до карти необхідного масштабу за орбітальними даними, які передаються разом із цільовою інформацією.

На *другому* етапі проводяться радіометричні і геометричні перетворення (корекція) для виправлення радіометричних і геометричних перекручувань, викликаних нестабільністю роботи космічного апарата і датчика, а