УДК 004.8: 681.3

Т. А. Шерепа
Національна бібліотека України імені В.І.Вернадського

Аналіз значущості термінів документів у CDS/ISIS-сумісних базах даних

Викладено запропоновану методику автоматичного індексування інформації у CDS/ISIS-сумісних базах даних з оцінкою значущості термінів і виявлення ключових слів документів для покращення повноти й точності видачі результатів пошуку.

Ключові слова: *автоматичне індексування, пошукові терміни, значущість терміну, розрізнювальна сила, асоційовані терміни, тезауруси, пошукові запити.*

У зв'язку з розвитком документних комунікацій усе більш актуальними постають проблеми ефективного доступу до інформації. Мета процесу індексування в документальних системах подібна до мети каталогізації у бібліотеках: приписати кожній одиниці зберігання деяку множину ідентифікаторів, які б відображали зміст документа. В традиційних бібліотеках у ролі ідентифікаторів змісту виступають відповідні шифри, які визначають предметну класифікацію і місце зберігання документа. З розвитком автоматичної обробки документів звичайний процес каталогізації перетворився на процес індексування, що призначений для надання кожному елементу ідентифікаторів, які також називають індексаційними термінами, ключовими словами, дескрипторами. Усі ці терміни відображають зміст документа і керують пошуком, вибираючи ті документи, терміни яких є найбільш схожими з термінами пошукового запиту.

З наданням пріоритетів формуванню та використанню електронних публікацій, що передбачається «Державною програмою розвитку діяльності Національної бібліотеки України імені В.І.Вернадського на 2005–2010 роки» [1], на бібліотеку покладається завдання формування універсального фонду національних інформаційних ресурсів, ефективне багатоаспектне використання якого потребує якісного індексування електронних документів колекцій бібліотек.

У даній роботі ставиться ціль визначення методики автоматичного індексування бібліографічної інформації у CDS/ISIS-сумісних базах даних для визначення ключових слів документів з метою покращення повноти й точності видачі результатів пошуку в базі даних електронних документів.

Усі нові електронні документи, що надходять до пошукової системи, мають пройти процес індексування: для кожного документа формується його пошуковий образ (профайл), що включає інформацію, яка буде далі використовуватись при пошуку. Ця інформація, у найпростішому випадку ключові слова, зберігається в базі даних.

Пошукова система електронних колекцій бібліотек Національної бібліотеки України ім. В.І.Вернадського (НБУВ) розроблена на базі пакета прикладних програм CDS/ISIS. CDS/ISIS (Computer Documentation System / Integrated System Information Services) є універсальним інструментарієм для створення автоматизованих систем бібліотек, архівів і музеїв, тобто для обробки структурованих нечислових баз даних [2].

Принциповою вимогою до програмних засобів, на основі яких створена пошукова система електронних колекцій, є її відповідність концепції вільного поширення [3], що обумовлено необхідністю тиражування електронних видань на компакт-дисках без будь-яких обмежень за проектом закону України «Про використання Відкритих форматів даних та Вільного програмного забезпечення в державних установах і державному секторі господарства» (від 18 червня 2003 р.), де наголошується на необхідності використання суб'єктами державного сектора «для провадження всіх публічних сервісів та створення інформації» лише вільного програмного забезпечення та програмних засобів з вільною ліцензією [4].

Головною особливістю CDS/ISIS є автоматичне створення й підтримка файлів швидкого доступу («індексних файлів») до кожної бази даних, що забезпечує максимальну швидкість пошуку навіть за великих об'ємів даних. Ці файли називаються словником пошукових термінів, і вміщують усі терміни, які можуть бути використані під час пошуку в базі даних [2].

Автоматичне індексування базується на текстах вихідних документів, або, принаймні, на фрагментах текстів, таких, як заголовки або реферати. Більшість результатів автоматичного індексування не є досконалими, але мають наступні значні переваги перед ручним індексуванням [5]: ефективність пошуку по відношенню до видачі релевантних документів, отриманих автоматичними методами є не менша, а то й вища, ніж при ручному індексуванні цих документів; менша вартість автоматичного індексування; витрачання значно меншого часу висококваліфікованого персоналу.

Звичайний процес індексування складається з однієї або декількох наступних операцій [5]:

— відбору індексаційних термінів, що використовуються для опису змісту документа;

— призначення цим термінам деякої ваги, що відображає значущість термінів;

— відношення кожного з термінів до відповідного типу;

— виявлення відношень між термінами, до яких відносяться, наприклад, синонімічні, ієрархічні, асоціативні та ін.

При автоматичному індексуванні баз даних на основі пакету прикладних програм CDS/ISIS індексний файл містить для кожного терміну список ідентифікаторів записів бази даних (документів), в яких цей термін зустрічається. В цьому списку зберігається також інформація про всі входження терміну в документ у

вигляді послідовності цілих чисел, що задають позиції даного терміну в даному документі. Ця інформація може бути корисною при виконанні пошуку за запитом із накладанням деяких обмежень на близькість термінів запиту в тексті документа. Крім того, структура інвертованого файлу забезпечує його швидку модифікацію при долучені в колекцію нових документів.

Таким чином, автоматичне індексування за допомогою пакета прикладних програм CDS/ISIS забезпечує виділення термінів з масиву документів, відкидаючи слова, що попередньо зазначені у стоп-словнику, зберігає в індексному файлі всю інформацію про зв'язки термінів з документами з точністю до позиції відповідного терміну у відповідному документі і надає доступ до цієї інформації. Але для подальшого аналізу: надання ваги термінам, індексування термінів/документтів за визначеними оцінками значущості (вагою), та дослідження взаємозв'язків термінів та термінів/документів прямих інструментів CDS/ISIS не дає. Однак, за допомогою ISIS_DLL, прикладного програмного інтерфейсу ISIS для операційних систем Windows та Linux, що розроблений та вільно поширюється UNESCO [6], і мови програмування, що припускає використання ISIS_DLL, можна отримати доступ до попередньо сформованого словника пошукових термінів відповідної бази даних CDS/ISIS для подальшого аналізу.

Словник стоп-слів будується з орієнтацією на вилучення другорядних частин мови (сполучників, прийменників та ін.), загальних дієслів, прикметників та прислівників (бути, знати, робити, великий, малий та ін.), займенників та чисельників. Це загальновживані слова, вилучення яких не вплине на якість пошуку, більш того може його покращити.

Комп'ютерна морфологія є необхідною в прикладних системах, які ведуть пошук і аналіз інформації, що представлена природною мовою. Основними функціями, що забезпечуються комп'ютерним морфоаналізом є отримання всіх словоформ слова, постановка слова в задану форму та отримання граматичних характеристик словоформи.

Для англомовних текстів є популярною операція виділення кореня, основи слова (stemming), що дозволяє всі однокореневі слова замінити коренем, і при пошуку їх не розрізняти, наприклад, алгоритм Портера, що широко використовується і дає прийнятні результати, хоча і з деякими похибками. Однак статистичний підхід завжди припускає можливість помилок при порівнянні пошукового запита користувача з документами, і пропонує різні методики їх нейтралізації .

Системи природномовного пошуку, машинного перекладу, безклавіатурного введення мовної інформації до комп'ютерів, автоматичного редагування, реферування та індексування, природномовні інтерфейси до інформаційно-комп'ютерних систем різного призначення все наполегливіше набувають ознак обов'язковості у програмному інструментарії комп'ютерів. Незважаючи на природність даного твердження, досвід науково-технічних здобутків української мови з цього напряму поки що слід оцінювати як досить скромний. Численні приклади з розвитку фахової лексикографії підтверджують, що за репертуаром і кількісними та якісними показниками українська лексикографія поки що відстає, наприклад, від англійської, німецької, французької, російської та не повністю відповідає сучасним потребам. А через відсутність багатьох типів словників досі неможливе і

створення комп'ютерних лексикографічних систем, що програмує відставання в галузі вітчизняної лінгвістичної технології взагалі [7].

Російські (українські) роботи з комп'ютерної лінгвістики для швидкого пошуку словоформи у словнику використовують допоміжні структури: дерево основ (коренів) та закінчень. Більшість слів мови відмінюються стандартним чином, тобто мають однакові закінчення в однакових граматичних формах. Як наслідок цього, всі різні парадигми (моделі відмінювання слів) компактно представляються у вигляді строк у трьох таблицях, а кожна основа слова зберігає посилання на відповідну строку таблиці. Більшість закінчень у парадигмах також є стандартним, і для зберігання так само використовують таблицю закінчень, а в таблицях парадигм зберігаються посилання на відповідні закінчення. Таким чином, кожне слово описується основою й кодом парадигми відмінювання слова.

Алгоритм імовірнісного морфоаналізу відрізняється від точного тим, що замість основи слів використовується дерево суфіксів, сформоване автоматично на етапі компіляції словника. В дерево суфіксів включаються кінцеві частини основ, що зустрічаються не менше 30 разів у словах з однаковою парадигмою відмінювання і мають довжину не більше 4-х символів за наявністю в основі не менше 3-х символів. Емпірично доведено, що ці величини забезпечують найбільшу точність аналізу.

При пошуку кожного закінчення знаходиться найдовше співпадання кінцевої частини слова з одним із суфіксів за умови однакової парадигми і частоти входжень не менше 30. В якості найбільш вірогідної парадигми відмінювання приймається та, при якій сумарна довжина суфікса й закінчення виявляється найбільшою. При наявності декількох кандидатів рівної довжини пріоритет надається парадигмі з більшою частотою входжень до словника [8].

Таким чином, використання основ слів в якості термінів веде за собою значне підвищення ефективності пошуку. Мовознавці дослідили, що загальновживані слова становлять у наукових текстах до 80 % загальної кількості слів. Звичайно, в різних науках — по-різному. Математика, наприклад, їх потребує найменше, інші науки — більше. У будь-якому випадку загальновживані слова дають найбільшу кількість помилок [9].

Автоматична система виявлення ключових слів, як правило, використовує статистичний частотний аналіз (методика В. Пурто). Якщо f — частота, з якою зустрічаються різні терміни в тексті, а u — відносне значення значущості терміну, тоді залежність $f(u)$ може бути апроксимована формулою:

$$f(u) = C\frac{1}{u},$$

тобто добуток частоти використання слів і їх значущості є константою. Подана гіпотеза використовується для виявлення двох границь значень частот. Слова з частотою менше нижньої границі вважаються дуже рідкісними, з частотою більшою за верхню границю — загальними, такими, що не несуть змістовного навантаження. Слова з частотою, що знаходиться між двох границь, найкраще характеризують зміст конкретного документа (використання такої оцінки вперше ввів

Лун). Однак, вибір границь — процедура достатньо суб'єктивна. Ключові слова, що виділяються програмно, аранжуються згідно з частотою їхнього використання.

Помічено, що відповідне значення має не тільки частота вживання слова в конкретному документі, але й кількість документів, в яких це слово зустрічається. За цією теорією найбільш важливими вважаються більш рідкісні, а не часті терміни. В роботах Спарка Джонса експериментально показано, що якщо N — кількість документів і n — кількість документів, в яких зустрічається даний індексний термін (ключове слово), то вага терміну, що визначається за формулою

$$W = \log \frac{N}{n} + 1,$$

приводить до більш ефективних результатів, а саме точності пошуку, ніж без використання оцінки значущості терміну [10].

В якості оцінки, що забезпечила б високі показники і точності, і повноти пошуку, може бути взятий добуток попередніх двох оцінок [5].

Зовсім інший підхід має місце у другій моделі індексування, відомої як метод оцінки розрізнювальної (дискримінаційної) сили терміну. В цій моделі більшу значущість має той термін, що робить документи максимально несхожими один на одний. Тим самим забезпечується максимально можлива віддаленість одного документа від іншого у просторі індексування. І навпаки, меншу значущість має термін, що робить документи більш схожими один на одний, внаслідок чого розрізняти їх стає важче. Чим більше буде розрізнення окремих документів, тобто чим менш будуть схожими відповідні вектори індексаційних термінів, тим легше буде знаходити одні документи, одночасно відкидаючи інші [5].

Таким чином, значущість терміну m вимірюється його розрізнювальною силою і визначається як різниця між значенням середньої попарної подібності документів, коли термін m є відсутнім у векторах документів, і значенням середньої попарної подібності документів, коли m присутній. Якщо термін є цінним, його присутність повинна робити документи менш схожими один на одний, знижуючи значення оцінки подібності середнього попарного порівняння документів і роблячи вище вказану різницю невід'ємною. Для термінів, що не розрізнюють документи, подана різниця має від'ємне значення.

Оскільки підрахування середніх попарних значень порівнянь документів потребує виконання порядку N^2 операцій для N документів, то більш простим, з точки зору обчислень, є метод визначення наповненості (густини) простору як суми значень оцінки подібності між окремим документом і центроїдом простору документів.

Якщо V_j — набір термінів (вектор термінів) документа j, і V_{ij} — вага (частота) терміну i у документі j, тоді центроїд усіх точок, що зображають документи масиву, визначається як «середній» документ C, де:

$$C_i = \frac{1}{N} \sum_{j=1}^{N} V_{ij} .$$

Якщо подібність документів k і j вимірювати за допомогою деякої функції порівняння векторів $S(V_k, V_j)$, де S змінюється від 1 для повністю співпадаючих документів до 0 для зовсім різних пар документів, то компактність простору документів можна представити як

$$Q = \sum_{j=1}^{N} S(C, V_j),\ 0 \le Q \le N,$$

тобто як суму значень подібності між кожним документом і центроїдом. Великі значення Q вказують на велику компактність простору документів, а відповідно, і більшу схожість між документами.

Вплив окремого терміна m в густину простору можна визначити шляхом обчислення функції $Q_m - Q$, де Q_m — це компактність простору документів, коли термін m вилучений з усіх векторів документів. Якщо термін m є цінним, з точки зору відображення змісту, то $Q_m > Q$, тобто простір документів після вилучення терміну m буде більш густим. Для термінів, що не мають задовільної оцінки розрізнювальної сили $Q_m < Q$. Тобто, значення розрізнювальної сили $(DV)_m$ терміну m визначається як різниця $Q_m - Q$. Таким чином, можна ранжувати всі терміни в порядку зменшення їх розрізнювальної сили. Використання цієї оцінки впливає на точність пошуку.

Забезпечення високих показників точності й повноти пошуку досягається наданням термінам ваги, що дорівнює добутку значень розрізнювальної сили і частотної оцінки [5].

Практичне дослідження ефективності автоматичного індексування проведене на базі пошукової системи електронної колекції документів бази даних НБУВ, що містить автореферати дисертацій, захищених в Україні у 2004 р. Аналіз документів проведений на основі попередньо сформованого словника пошукових термінів бази даних CDS/ISIS колекції документів за допомогою інтерфейсу ISIS_DLL. Дані словника пошукових термінів (терміни назв та рефератів дисертацій) за допомогою мови програмування PHP вигружені до реляційної бази MySQL наступної структури.

Основні статистичні характеристики даної електронної колекції документів, що отримані за допомогою PHP-скриптів та SQL, наведені в таблиці.

Характеристика	**Значення**
Кількість документів електронної колекції	3119
Кількість термінів без повторювань, що містять більше 3-х літер	45490
Застосування наступного алгоритму виявлення словоформ: слова із співпадаючою основою (коренем) слова більше ніж 6 літер; 6-ти літер при закінченні не більше 3-х літер; 5-ти літер — 2-х літер закінчення; 4-х — однієї літери — вважаються однокореневими	20505
Кількість термінів після вилучення з попередньої сукупності термінів із документною частотою використання 1 (що зустрілися в одному документі)	10554
Кількість термінів після вилучення з попередньої сукупності термінів із дуже високими значеннями частоти, що використовуються більш ніж у 25 % документів	10522
Кількість термінів після вилучення з попередньої сукупності термінів із від'ємним значенням розрізнювальної сили	7385

Так як електронна колекція документів бази даних НБУВ, що містить автореферати дисертацій, зберігає і тематики цих документів, то на основі отриманих термінів із задовільною оцінкою значущості, можуть бути побудовані тематичні (предметні) тезауруси. До кожної тематики відносять терміни, ймовірність використання яких у цій тематиці перевищує ймовірність їхнього використання в будь-якій іншій тематиці.

Пошуковий запит вводиться користувачем природною мовою, тобто існує необхідність перевірити кожне слово запиту для вилучення другорядних частин мови та загальновживаних слів, провести аналіз словоформ термінів, що залишились, і співуставити їх із відповідним тезаурусом. Сума оцінок ваги термінів документа може бути обчислена шляхом додавання вагових коефіцієнтів тих термінів документа, які співпадають із термінами пошукового запиту. При нульових результатах пошуку також можуть бути розглянуті документи, що містять терміни, які є асоційованими до термінів пошукового запиту.

Обгрунтуванням цього підходу є припущення того, що якщо термін B завжди використовується з терміном A, то немає значення який з них використовується в пошуковому запиті. Це вказує на абсолютну кореляцію між термінами. Два терміни, які є тісно зв'язаними в асоціативній схемі, мають бути близькими і семантично. Коефіцієнт асоціації може бути обчислений за формулою Дойла [11]:

$$A = \frac{f_{AB}}{f_A + f_B - f_{AB}},$$

де f_{AB} — частота сумісного використання термінів A і B у документах; f_A — частота використання терміну A; f_B — частота використання терміну B.

Також може бути використана оцінка близькості двох термінів за допомогою формули Евклідової відстані:

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (x_{ik} - x_{jk})^2} ,$$

де N — кількість документів; x_{ij} — елементи досліджуваної матриці терм-документ (що містить у собі частоти використання всіх термінів у кожному з документів колекції).

Підрахування цієї та інших оцінок, проведення кластеризації документів електронних колекцій може бути проведене з використанням пакета прикладних програм IDAMS, призначеного для валідації, маніпулювання і статистичного аналізу даних. IDAMS виробляється та вільно поширюється UNESCO. Він включає в себе інструменти маніпулювання й аналізу даних, що є доступними через інтерфейс користувача та командну мову [12]. IDAMS дозволяє підраховувати базові статистичні параметри вибірки — середні, частотні характеристики, кореляції та ін. Основний набір статистичних процедур включає також декілька важливих видів аналізу, таких як кластерний (підтримується шість алгоритмів), дискримінантний, факторний, регресійний та дисперсійний [12].

Після імпортування до пакета IDAMS матриці терм-документ у вигляді текстового файлу з відокремлювачами, на основі отриманих даних необхідно створити словник даних IDAMS, що визначає типи даних та правила їх валідації. На базі словника даних будується файл даних IDAMS, який і буде підлягати обробці і аналізу.

Таким чином, практичне застосування методів індексування колекцій документів електронних бібліотек ставить собі за мету покращення повноти та точності інформаційного пошуку шляхом його інтелектуалізації: уточнення пошукових запитів, ранжування видачі результатів пошуку за оцінкою близькості до пошукового запиту, використання тематичних тезаурусів, використання кластерів документів для звуження масиву пошуку. Також на основі проведення індексації колекцій електронних документів можуть бути розв'язані задачі відстеження змін у часі термінів предметних галузей, авторубрикації та класифікації нових документів та автоматичного реферування документів колекції.

Висновки

1. З розвитком документних комунікацій все більш актуальними постають проблеми ефективного доступу до інформації. Мета процесу індексування в документальних системах: приписати кожній одиниці зберігання деяку множину ідентифікаторів (індексаційних термінів, ключових слів, дескрипторів), що відображають зміст документа і керують пошуком. Автоматичне індексування базується на текстах вихідних документів, тому більшість результатів автоматичного індексування не є досконалими, але мають переваги перед ручним індексуванням, такі як ефективність пошуку по відношенню до видачі релевантних документів, меншу вартість та витрачання меншого часу висококваліфікованого персоналу.

2. Автоматична система виявлення ключових слів, як правило, використовує: статистичний частотний аналіз, аналіз кількості документів, в яких зустрічаються ключові слова, та метод оцінки розрізнювальної (дискримінаційної) сили терміну,

де більшу значущість мають терміни, що роблять документи максимально несхожими один на одний.

3. Проіндексовані терміни можуть бути використані для автоматичного реферування документів електронної колекції та побудови тематичних тезаурусів, де до кожної тематики відносять терміни, що мають високу оцінку значущості, і ймовірність використання яких у деякій тематиці перевищує ймовірність їхнього використання в будь-якій іншій тематиці. На основі тематичних тезаурусів може бути проведена класифікація (авторубрикація) нових документів електронної колекції.

4. Результати видачі пошуку можуть бути ранжовані за сумою оцінок ваги термінів документа, що обчислена шляхом додавання вагових коефіцієнтів тих термінів документа, які співпадають з термінами пошукового запиту. При нульових результатах пошуку можуть бути розглянуті документи, що містять терміни, які є асоційованими до термінів пошукового запиту. При поділенні проіндексованих термінів на кластери, пошуковий запит може бути порівняний із центром кожного кластера, для подальшого звуження масиву пошуку.

5. Автоматичне індексування за допомогою пакета прикладних програм CDS/ISIS забезпечує виділення термінів з масиву документів, відкидаючи стоп-слова та зберігає в індексному файлі всю інформацію про зв'язки термінів з документами. За допомогою прикладного програмного інтерфейсу ISIS_DLL можна отримати доступ до словника пошукових термінів бази даних CDS/ISIS для подальшого аналізу: надання ваги термінам, індексування термінів/документів за оцінками значущості (вагою), та дослідження взаємозв'язків термінів і термінів/документів.

6. Поглиблений аналіз документів електронних колекцій має вивести інформаційні системи бібліотек на якісно новий рівень і сприяти їх трансформації в інтелектуальні системи, що проводитимуть бібліометричні, інформометричні та наукометричні дослідження у великих масивах інформації й дозволять творити нові знання.

1. Про затвердження Державної програми розвитку діяльності Національної бібліотеки України імені В.І. Вернадського на 2005–2010 роки: Постанова Кабінету Міністрів України від 25 серпня 2004 р. № 1085.

2. UNESCO CDS-ISIS databases [Electronic Resource]. — Way of access: URL: http://www.unesco.org/. — Title from the screen.

3. *Шерепа Т.А.* Система галузевих серій електронних видань: основні концептуальні положення // Бібл. вісн. — 2004. — № 1. — С. 26–29.

4. Про використання Відкритих форматів даних та Вільного програмного забезпечення в державних установах і державному секторі господарства: Проект Закону України від 18 червня 2003 р.

5. *Солтон Дж.* Динамические библиотечно-информационные системы. — М.: Мир, 1979. — 558 с.

6. ISIS Application Program Interface ISIS_DLL User's Manual Preliminary Version BIREME, São Paulo, July 2001 [Electronic Resource]. — Way of Access: URL: http://www.bireme.br/.

7. *Широков В.А.* Всеукраїнський лінгвістичний діалог у контексті теорії лексикографічних систем // Мовознавство. — 2003. — № 6. — Way of Access: URL: http://ulif.org.ua/ulp

8. *Ермаков А.Е., Плешко В.В.* Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог'2004. — М.: Наука, 2004 [Электронный ресурс]. — Way of access: URL: http://www.rco.ru/article.asp?ob_no=627

9. Як дібрати С Л О В О ? [Електронний ресурс]. — Way of access: URL: http://dict.linux.org.ua/dict/other/SSR/RE1.html.

10. *Семенов Ю.А.* Современные поисковые системы [Электронный ресурс]. — Way of access: URL: http://www.penza.fio.ru/misc/admin/tcpip/retr4514.htm.

11. *Ланкастер Ф.У.* Информационно-поисковые системы. — М.: Мир, 1972. — 308 с.

12. IDAMS Statistical Software [Electronic Resource]. — Way of Access: URL: http://www.unesco.org/webworld/idams. — Title from the screen.

Надійшла до редакції 19.12.2005