УДК 004.89
О.Маєвський, О.Назаревич, Г.Шимчук
(Тернопільський державний технічний університет імені Івана Пулюя)

ДОСЛІДЖЕННЯ ІНФОРМАЦІЙНО-ПОШУКОВИХ СИСТЕМ В МЕРЕЖІ ІНТЕРНЕТ

Щорічне споживання різного виду інформації зростає в геометричній прогресії, а спроможність людини опрацювати певний об'єм інформації величиною сталою. Тому проблема автоматизації пошуку інформації є актуальною на сьогодні, що економить людині час на опрацювання більшого її об'єму. Актуальною є проблема створення інформаційно-пошукових систем (ІПС). Розвиток сучасних систем пошуку у великих колекціях даних відбувається одночасно в декількох напрямах. З самих відомих можна назвати системи інформаційного пошуку (IR-системи, Information Retrieval), в числі яких такі, як Google, Яndex, AltaVista, рубрикатори і класифікатори (наприклад, Yahoo або List.Ru), альтернативні системи інтуїтивного пошуку і перегляду (як приклад можна назвати Kartoon), системи відповідей на питання (QA systems, які в літературі прийнято називати системами фактографічного пошуку), наприклад, AnSel, Mulder, AskMSR, системи систематизації та катологізації найдених ресусів інтернету, створення електронної бібліотеки знань (Zotero, del.icio.us).

Предметом даної роботи є дослідження моделей існуючих інформаційних пошукових систем і вивчення підходів до «прозорої інтеграції» пошуку даних в призначений для користувача інтерфейс, тобто реалізації повнотекстового контекстного пошуку різних типів web-документів.

Математично розглядаємо векторний простір множини документів та її атрибутів. Основною ідеєю моделі векторного простору є задача створити кожному документу колекції його образ – вектор в деякому евклідовому просторі так, щоб образи близьких документів були близькі. Отримавши запит користувача, пошукова система, заснована на векторній моделі, будує вектор цього запиту в тому ж просторі і видає список документів, ранжований по ступеню близькості векторів документа і запиту.

Нехай C – колекція документів, T – словник колекції, T(d) – множина всіх термів документа d в колекції C, tf (d,t) – число входжень терма t в документ d, df (t) – число документів колекції C, що містять терм t. Позначимо вагу терма t в документі d через w(d, t), тоді,

$$\omega(d,t) = \frac{tf(d,t)\log(\frac{|C|}{df(t)})}{\sqrt{\sum_{t'} tf(d,t')\log(\frac{|C|}{df(t)})}} \qquad (1)$$

Ця формула заснована на природному статистичному спостереженні, що чим більше локальна частота терма в документі (tf) і більше «рідкість» (тобто зворотна зустрічаємість в документах) терміну в колекції (idf), тим вище маса даного документа по відношенню до терма. Тобто, в даному випадку робиться спроба привласнити більшу масу тим термінам, які «відрізняють» даний документ від решти документів колекції C. Як міра подібності документів звичайно розглядається скалярний добуток між їх векторами. Описаний метод зважування термів називається tf*idf

Оцінка якості пошуку ІПС проводилася методом побудови співвідношення recall-precision на самостійно підготовленому наборі даних.