Секція: **Інформаційні технології**

УДК 025.4

Бурда А.– ст. гр. КА-11
Тернопільський національний технічний університет імені Івана Пулюя

ПРИНЦИП РОБОТИ ПОШУКОВИХ ІНФОРМАЦІЙНИХ СИСТЕМ

Науковий керівник: асист. Федорів П.С.

Пошукова інформаційна система —онлайн-служба, яка надає можливість пошуку інформації на сайтах в інтернеті, а також у групах обговорення та ftp-серверах. Класична ІПС складається з трьох частин: Web павук (робот, агент); індексна база і пошуковий механізм (алгоритм).

Web павук - це програма, яка працює на декількох комп'ютерах підключених до мережі Інтернет, вона слідує по гіперпосиланнях з веб сторінок і викачує всі знайдені файли. Веб павуком може керувати власник сайту. Досить зберегти в кореневу директорію сайту спеціальний файл robots.txt. У цьому файлі на спеціальній мові описані команди для веб павуків. Це необхідно, в першу чергу, для приховування приватної інформації від пошукової системи. Були випадки, коли через ІПС Google при введенні запиту "номера кредитних карток" виводилася приватна інформація. Більш того, такі павуки уміють обходити рекламні трюки по просуванню сайтів, якими користуються власники сайтів для збільшення відвідувальності сайтів. Веб роботи дуже суворо відносяться до таких обманів і не вносять такі сайти до бази. Також веб роботи приймають заявки на індексацію тільки що створеного сайту. На нові веб сайти ніхто не посилається, і прийти рекурсивно по посиланнях інших сайтів не можна.

Інша частина ІПС - це індексатор, завданням якого є обробка "викачаного Інтернету". Це складніша система. Вона витягує всі слова з викачаних документів, і складає в певну індексну базу. Для кожного слова витягується інформація про те, як це слово розташоване на веб сторінці: позиція слова в тексті сторінки, кількість входжень слова в сторінці, колір і шрифт, використаний для оформлення слова.

Витягнуті слова заносяться в спеціальні словники. При занесенні в словник часто відсікають закінчення і суфікси для ефективнішого зберігання інформації. Але це знижує точність пошуку. Словники є частиною індексу і з ідентифікаторами веб сторінок. Будь-якому слову із словника відповідає набір doc_id-документів, в яких це слово зустрічається. Роботою по постійному формуванню інверсного індексу займаються сортувальники. Перед тим, як обробити запит користувача на пошук, пошукова система робить ряд кроків: Перевіряє орфографію запиту. Іноді, в процесі швидкого набору тексту робляться помилки. Новітні системи можуть знаходити помилки в словах і пропонувати ввести свій правильний варіант. Відбувається генерація схожих по сенсу слів і різних відмінкових форм. Наприклад, на запит „купити слона" будуть, також шукатися "продати слона", "продаж слонів". Це істотно розширить межі пошуку. Для цього використовується спеціальні морфоаналізатори. Існує два типи морфоаналізаторів: імовірнісні і імовірно-словарні. Останні якісніші, оскільки оброблене слово додатково перевіряється по словнику. Запит перекладається іншими мовами. Встановлення стоп-слова (займенники, приводи). Останнім часом не використовується. Раніше це робилося для економії обчислювальних ресурсів. І лише після цього виконується запит на пошук.