ПРАКТИЧНЕ ЗАВДАННЯ №2
АНАЛІЗ ПОШУКОВИХ ЗАПИТІВ

1. Мета завдання

Метою завдання є формування практичних навичок інформаційного пошуку.

Інформаційний пошук (англ. Information retrieval) – наука про пошук неструктурованої документальної інформації. Особливо це відноситься до пошуку інформації в документах, пошук самих документів, здобуття метаданих з документів, пошуку тексту, зображень, відео та звуку у локальних реляційних базах даних, у гіпертекстових базах даних таких, як Інтернет та локальні інтранет.

2. Зміст завдання

Розробити картки підприємств, перелік та кількість яких визначено для кожного студента (виконавця) у додатку А, за результатами аналізу пошукових запитів в системах Google та Yandex. Примірник картки підприємства у додатку Б (слід зауважити, що картка підприємства може не обмежуватися однією сторінкою та мати 3, 4 та навіть 5 сторінок об'єму).

3. Механізм інформаційного пошуку

Автоматичні системи інформаційного пошуку використовують для зменшення так званого «інформаційного перевантаження». Багато університетів та публічних бібліотек використовують системи інформаційного пошуку для полегшення доступу до книжок, журналів та інших документів. Найвідомішим прикладом систем інформаційного пошуку можна назвати пошукові системи в Інтернеті.

Об'єктом інформаційного пошуку є текстова інформація, зображення, аудіо, відео інформація.

З інформаційним пошуком змикаються проблеми: розсилки інформації (information routing); сортування інформації (information filtering); упорядкування (класифікація) інформації (information categorization); відбір інформації (information extraction).

Для інформаційного пошуку розробляють: алгоритми інформаційного пошуку (retrieval algorithms); підходи інформаційного пошуку(retrieval approaches); стратегії інформаційного пошуку (retrieval strategies).

Для його здійснення створюють: методи інформаційного пошуку (retrieval utilities); засоби інформаційного пошуку (information retrieval systems); комп'ютерні пошукові програми (search engines).

До проблем інформаційного пошуку належать питання: представлення даних, інформації, знань (data, information, knowledge); представлення інформації в сучасних інформаційних сховищах (representation of information); багатомовний інформаційний пошук (cross-language information retrieval); одночасний інформаційний пошук (parallel information retrieval); розподілений інформаційний пошук (distributed information retrieval); суспільний інформаційний пошук (social information retrieval)

Напрям інформаційний пошук відносять до проблем: застосовної (прикладної) лінгвістики (applied linguistics); обробки природної мови (natural language processing);

Завданням інформаційного пошуку є знаходження відповідних (до пошукового запиту) інформаційних об'єктів, або документів серед доступного для пошуку матеріалу. Завдання для інформаційного пошуку задається у вигляді інофрмаційного запиту (query), який може містити слова, фрази чи речення або комбінацію їх. Переважна більшість пошукових систем орієнтована на роботу з пошуковими термінами — словами або словосполученнями, які пошукова система розпізнає як одне ціле. Для здійснення інформаційного пошуку потрібно мати збірку інформаційних об'єктів (бібліотека, комп'ютерні файли) і систему (алгоритм або програму) яка здійснює пошук. Для здійснення інформаційного пошуку користувач (людина або інформаційна система) формує інформаційний запит (information query). Результатом

пошукової роботи є список документів який укладається згідно певного принципу. Такий список називають впорядкованим (ranked list, ranked results).

Пошукова система переглядає всі доступні інформаційні одиниці (документи) зі збірки і відбирає документи відповідні до інформаційного запиту. Оскільки реальні пошукові системи знаходять не всі відповідні документи, говорять про точність пошукових систем (system accuracy). Результатом роботи пошукової системи є список відібраних документів (retrieved documents list), серед яких є відповідні до запиту документи (relevant documents). Для ідеальної пошукової системи список відібраних документів та відповідних документів повинні збігатися. В реальних пошукових системах в списках відібраних документів знаходяться і невідповідні до запиту документи. Тому говорять про ефективність пошукових систем. Ефективність пошукових систем оцінюється двома параметрами: пошукова відповідність (precision) та пошукова якість (recall). Пошукова відповідність визначає частку відповідних документів серед відібраних на запит. Пошукова відповідність визначає якість отриманого результату інформаційного пошуку. Пошукова якість визначає частку отриманих системою відповідних до запиту документів серед загального числа відповідних до запиту документів у збірці. Загальне число відповідних до запиту документів завжди є невідомим і може бути встановлене лише при повному перегляді збірки людиною. Крім того роботу пошукових систем оцінюють швидкодією — часом, за який отримують список відповідних до запиту документів.

4. Стратегії інформаційного пошуку

Стратегії інформаційного пошуку визначають ступінь подібності документів, що розглядаються, до пошукового запиту. Ступінь подібності визначається згідно робочої гіпотези: чим частіше пошуковий термін зустрічається в документі, тим більше «відповідним» є цей документ до пошукового запиту.

Стратегії інформаційного пошуку розробляються не тільки для визначення відповідності, але і для вирішення проблем, які пов’язані з неоднозначністю мови – один і той самий термін може позначати різні концепти (ключ в механіці означає зовсім не те, що в шифруванні), один і той же концепт може позначатись різними термінами (обласний центр Львівської області має назву Львів і Місто Лева).

Стратегія інформаційного пошуку це алгоритм, який, переглядаючи набір документів (Д1, ..., Дn), встановлює їх відповідність до пошукового запиту (ПЗ). Оскільки пошуковий термін зустрічається в документах різну кількість раз, можна говорити про різну ступінь відповідності до пошукового запиту. Цей алгоритм обчислює коефіцієнт відповідності (similarity coefficient) (КВ) для кожного документу КВ(ПЗ, Ді), де 1 ? і ? n.

Існують такі стратегії інформаційного пошуку:

– з використанням векторно-просторового представлення (vector space model);
– пошук імовірності появи пошукового терміну в документі (probabilistic retrieval);
– з побудовою мовної моделі для кожного документу (language models);
– з побудовою мережі припущень, яка використовується для встановлення відповідності документу до пошукового запиту (inference network);
– з Булевим індексуванням, коли кожному пошуковому терміну присвоюється своя «вага», що потім враховується при побудові впорядкованих списків документів (Boolean indexing);
– з використанням не проявленого семантичного індексування (latent semantic indexing);
– з побудовою нейромереж (neural networks);
– з використанням продуктивних алгоритмів, коли початковий пошуковий запит «еволюційно» видозмінюється (genetic algorithms);
– з використанням нечітких множин, коли документу ставиться у відповідність нечітка множина (fuzzy set retrieval).

5. Пошуковий сервіс в Інтернет

Пошукові системи зазвичай мають три компоненти:

– **агент (павук, кроулер або робот)**, який переміщується по мережі і збирає інформацію;

– **база даних**, яка містить інформацію, що зібрано павуками;

– **пошуковий механізм**, який користувачі використовують як інтерфейс для взаємодії з базою даних.

Засоби пошуку типу агентів, павуків, кроулерів і роботів використовуються для збору інформації про документи, які знаходяться в мережі Інтернет. Це спеціальні програми, які займаються пошуком сторінок в мережі, збирають гіпертекстові посилання з цих сторінок і автоматично індексують інформацію, яку вони знаходять для побудови бази даних. Кожний пошуковий механізм має власний набір правил, якими визначається збір документів.

Агенти є найінтелектуальнішими з пошукових засобів. Вони можуть робити більше, ніж просто шукати: вони можуть виконувати транзакції від імені користувача. Вже зараз вони можуть шукати сайти специфічної тематики і повертати списки сайтів, відсортованих за їх відвідуваністю. Агенти можуть обробляти вміст документів, знаходити та індексувати інші види ресурсів, не лише сторінки. Вони можуть бути запрограмовані для витягання інформації з вже існуючих баз даних. Незалежно від інформації, яку агенти індексують, вони передають її назад до бази даних пошукового механізму.

Павуки здійснюють загальний пошук інформації в Інтернет. Павуки повідомляють про зміст знайденого документа, індексують його і добувають підсумкову інформацію. Вони також переглядають заголовки, деякі посилання і відправляють проіндексовану інформацію до бази даних пошукового механізму.

Кроулери переглядають заголовки і повертають тільки перше посилання.

Роботи можуть бути запрограмовані таким чином, щоб переходити по різним посиланням різної глибини вкладеності, виконувати індексацію і перевіряти посилання в документі. Але, вони можуть застрягати в циклах, адже, проходячи за посиланнями, їм потрібні значні ресурси мережі. Існують методи, що забороняють роботам пошук по сайтах, власники яких не бажають, щоби вони були проіндексовані.

Агенти збирають та індексують різні види інформації. Деякі, наприклад, індексують кожне окреме слово у документі, в той час як інші індексують тільки 100 найбільш важливих слів в кожному документі, індексують розмір документу і кількість слів в ньому, назву, заголовки і підзаголовки і так далі. Вигляд побудованого індексу визначає, який пошук може бути проведений пошуковим механізмом і як отримана інформація буде інтерпретована.

Агенти знаходять інформацію, після чого її розміщують в базі даних пошукового механізму. Адміністратори пошукових систем визначають, які сайти або типи сайтів агенти мають відвідати та проіндексувати. Проіндексована інформація відправляється до бази даних пошукового механізму.

Користувачі можуть розміщувати інформацію прямо в індексі, заповнюючи особливу форму для того розділу, в який вони хотіли б помістити свою інформацію. Ці дані передаються базі даних.

Коли користувач хоче знайти інформацію, доступну в Інтернет, він відвідує сторінку пошукової системи і заповнює форму, що деталізує потрібну йому інформацію. Тут можуть використовуватись ключові слова, дати та інші критерії. Критерії в формі пошуку повинні відповідати критеріям, які використовуються агентами при індексації інформації, яку вони знайшли при переміщені по мережі.

База даних відшукує предмет запиту, що базується на інформації, яка вказана в заповненій формі, і виводить відповідні документи, що підготовані базою даних. Для того, щоб визначити порядок, в якому перелік документів буде показано, база даних застосовує алгоритм ранжування. В ідеальному випадку, розташованими першими в списку будуть документи, що є найбільш релевантними до запиту користувача.

Релевантність – основне поняття при індексації документа в пошукових системах. Релевантність – міра відповідності, тобто це відповідність змісту знайденої сторінки до запиту користувача. Але комп'ютер – не людина, і тому пошукові системи використовують спеціальні алгоритми для визначення релевантности. Теоретичних методів визначення релевантності

більш ніж 20. Але виділяють два основні напрями: лінгвістичне (Рамблер, Яндекс) і статистичне (Google).

Основні російські пошукові системи (зокрема Рамблер) використовують лінгвістичний напрям, тобто пошуковий робот, переглядаючи сторінку, звертає увагу на "літературність" її написання ("чом ти не прийшов" буде більш релевантною, ніж "чом ти не травень прийшов").

Різні пошукові системи використовують різні алгоритми ранжування, однак основними принципами визначення релевантності є наступні:

- **Кількість слів запиту** у текстовому вмісті документу (тобто в html-коді).
- **Теги,** в яких ці слова розташовуються.
- **Місцеположення** шуканих слів у документі.
- **Питома вага слів,** відносно яких визначається релевантність, у загальній кількості слів документу.

Ці принципи застосовуються всіма пошуковими системами. А наведені нижче використовуються деякими, але достатньо відомими (наприклад, AltaVista).

Час – як довго сторінка знаходиться в базі пошукового сервера. Спочатку здається, що це недолугий принцип. Але в Інтернет існує багато сайтів, час життя яких складає близько місяця. Якщо ж сайт існує досить довго, це значить, що його власник є досвідченим за даною темою і користувачу більше підійде сайт, що існує вже кілька років, ніж той, який з'явився тиждень тому за цією ж темою.

Індекс цитованості – як багато посилань на дану сторінку веде з інших сторінок, що зареєстровані у базі пошуковика.

База даних виводить ранжований таким чином перелік документів з HTML і повертає його користувачу, який зробив запит. Різні пошукові механізми вибирають різні способи показу отриманого переліку – деякі відображають лише посилання, інші виводять посилання з декількома першими реченнями документу або заголовок документу разом з посиланням. Коли користувач звертається до посилання на один з документів, цей документ завантажується з сервера, на якому він знаходиться.

Велика частина цільових відвідувачів приходить саме з пошукових систем. Тому важливо знати деякі особливості найбільш популярних з них.