Колекція Cranfield

Колекція «Cranfield» була сформована під час серії експериментів під назвою «Cranfield experiments» у 1960-х роках. Експерименти були розбиті на два етапи, ціллю яких було оцінити якість систем індексування, що існували на той час. Дослідження проводилися без використання комп'ютерів.

На першому етапі порівнювались чотири системи індексування: універсальна десяткова класифікація (Universal Decimal Classification), алфавітний тематичний каталог (Alphabetical Subject Catalogue), фасетна класифікація (Faceted classification) та система Мортімера Тауба (Mortimer Taube's Uniterm system of co-ordinate indexing).

Спеціалісти з цих систем складали з їх використанням індекси для колекції статей з аеродинаміки. Після чого, автори цих статей складали групи запитів за якими шукали ці статі.

Другий етап відрізнявся від першого способом формування запитів до індексів та оцінкою релевантності документа.

Під час першого етапу кожна система індексування обробляла запити по своєму, при цьому ніхто, окрім спеціаліста з цієї системи, не знав як саме ця обробка виконується. Під час другого експерименту усі ці кроки докладно аналізувались у результаті чого учасники дійшли висновку, що запити краще всього залишати у їх початковій формі.

Також, під час другого експерименту релевантним вважався той документ, який задовольняє інформаційну потребу, а не той на основі змісту якого було згенеровано запит.

Колекцію можна знайти у вигляді архіву за посиланням: http://ir.dcs.gla.ac.uk/resources/test_collections/cran/.

Архів містить саму колекцію з 1400 статей зібраних у одному файлі, файл з запитами, файл з оцінками релевантності документів відносно запитів та файл з поясненнями оцінок релевантності.

Кожна оцінка релевантності документа відносно запиту має вигляд

(Номер запиту, номер документу, оцінка релевантності),

де оцінка релевантності це число від одиниці до п'яти. Значення цих чисел наступне:

a) 1 – документ дає повну відповідь на запитання
b) 2 – документ містить важливу інформацію, відсутність якої б зробила пошук відповіді на запитання неможливим або ж цей пошук вимагав би значних зусиль.
c) 3 – документ містить корисну інформацію яка дає гарне підґрунтя для пошуку відповіді на запитання або ж містить корисні поради щодо того як краще підійти до пошуку відповіді на запитання.
d) 4 – документ містить мінімально корисну інформацію. Наприклад, історичну довідку.
e) 5 – документ не містить корисної інформації

Примітка – Під запитанням мається на увазі запит.

Колекції TREC

«TREC» або ж «Text Retrieval Conference» це конференція метою якої є дослідження різних задач інформаційного пошуку, кожна така задача у рамках конференції маю назву «track». Ця конференція фінансується американським Інститутом стандартів і технологій (NIST) та Агентством з перспективних досліджень розвідувального відомства (IARPA) і проходить з 1992 року.

Під час кожного track(у) розглядається одна конкретна задача, для якої NIST забезпечує необхідні набори даних і дані для тестування.

Наприклад, TREC-1, який відбувався у 1993 році, мав не меті дослідження «ad-hoc information retrieval». У якості набору даних на двох дисках надавалися статі з «Wall Street Journal» (випуски з 1986 по 1992), документи з Федерального регістру (1988 та 1989 років) та Департаменту енергетики і документи від агентства «Newswire». Колекція важила 190 мегабайтів та містила 186.225 унікальних термінів.

У якості тестових даних замість звичайних запитів надавалися «теми» або ж «topics». Використовуючи зміст теми кожен учасник міг самостійно складати запити, замість того щоб користуватись вже готовими.

Кожна тема містила стислий і розгорнутий опис того яким повинен бути релевантний документ, «концепти», тобто набір понять з якими може бути знайомий користувач, що шукає документ, та ще деяку додаткову інформацію яка допомагала складати запити для цієї теми. Таких тем було від 50 до 100.

Для TREC-2 набір даних не змінився, проте тепер був записаний не на 2, а на 3 диски. Кількість термінів та вага колекції майже не змінилась. Кількість тем збільшилась, тепер їх було від 100 до 150.

Інформація про всі інші набори даних які були використані під час інших конференцій, може бути знайдена за посиланням https://trec.nist.gov/data.html. Інформацію про актуальні track(и) можна знайти за іншим посиланням: https://trec.nist.gov/tracks.html.

Також, після закінчення кожного track(у) NIST публікує доповіді, у яких можна почитати про результати тієї чи іншої конференції.

Варто відмітити колекцію документів під назвою «GOV2» яка була зібрана для «Terabyte Track» метою якого було розглянути використання «ad hoc» алгоритмів інформаційного пошуку на колекціях великого розміру.

GOV2 важить 426 гігабайтів та містить 25.205.179 документів приблизно 92% яких це HTML сторінки і 8% яких це PDF документи.

Колекції NTCIR

NTCIR або «National Institute of Japan Test Collections for IR» це набір колекцій для дослідження алгоритмів інформаційного пошуку в умовах міжмовного інформаційного пошуку. Особливістю цих колекцій є те, що вони присвячені східно-азіатським мовам.

Окрім формування колекцій NTCIR, аналогічно до TREC, влаштовує тематичні конференції для вирішення різних задач інформаційного пошуку. Перша така конференція відбулася у 1999 році.

Розглянемо колекції які використовувалися для конференцій, метою яких був «ad hoc information retrieval». Їх усього дві: NTCIR-1 та NTCIR-2.

NTCIR-1 складається з 330.000 наукових документів на японській мові, половина з яких має англійський відповідник, 83-х тем (аналогічні темам TREC) та має вказівки щодо оцінки релевантності документів. Колекцію можна використати для пошуку або документів написаних англійською або для пошуку документів на англійській, що відповідатимуть запиту на японській.

NTCIR-2 складається з 400.000 наукових текстів, написаних японською, 130.000 текстів написаних англійською, 49-ти тем та, як і NTCIR-1, має вказівки щодо оцінки релевантності документів. NTCIR-2 рекомендується використовувати разом з NTCIR-1.

Колекції можна знайти за посиланням:
http://research.nii.ac.jp/ntcir/permission/perm-en.html#ntcir-1

Колекція Reuters-21578 та Reuters-RCV1

Документи з колекції Reuters-21578 з'явилися у стрічці новин Reuters у 1987 році і в тому ж році вони були розподілені по категоріям та зібрані у одну колекцію. Колекцій складається з 21.578 документів. Всі документи розподілені по п'яти великим категоріям: «EXCHANGES», «ORGS», «PEOPLE», «PLACES» і «TOPICS». Кожна з цих категорій складається з під-категорій. Наприклад, до «TOPICS» належать категорії пов'язані з економікою, наприклад: «coconut», «gold», «inventories», «money-supply». До «PLACES» належить категорія «australia», «PEOPLE» належить «parez-de-cuellar», «ORGS» «gatt», «nasdaq» належить до «EXCHANGES».

Колекцію можна знайти за посиланням:
http://www.daviddlewis.com/resources/testcollections/reuters21578/

Колекція Reuters-RCV1 схожа за організацією на попередню проте складається з більш ніж 800.000 документів.

Колекцію, разом з її описом, можна знайти за посиланням:
http://jmlr.csail.mit.edu/papers/volume5/lewis04a/

Колекція 20 Newsgroup

Колекція складається з 18.846 документів без урахування дублікатів, розділених на 20 груп. Фактично, колекція складається з 20 файлів (які називаються «newsgroup»), які містять зміст 18.846 документів.

Колекцію можна знайти за посиланням:
http://qwone.com/~jason/20Newsgroups/