

Извличане на информация - въведение

проф. д-р инж. Христо Вълчанов
207-4E

<http://cs.tu-varna.bg>

Литература

1. Х. Вълчанов, В.Алексиева. Извличане на информация в Интернет. Ръководство за лаб. упражнения. Варна, 2015.
2. Yi Chang, Hongbo Deng. Query Understanding for Search Engines. Springer, 2020, ISBN 978-3-030-58333-0
3. Nicola Ferro, Carol Peters. Information Retrieval Evaluation in a Changing World. Springer, 2019, ISBN 978-3-030-22947-4
4. Tetsuya Sakai, Douglas W. Oard, Noriko Kando. Evaluating Information Retrieval and Access Tasks. Springer, 2021, ISBN 978-981-15-5553-4
5. Introduction to Information Retrieval. <https://nlp.stanford.edu/IR-book/information-retrieval-book.html> (октомври 2022)
6. Ceri S., A. Bozzon. Web Information Retrieval. Springer-Verlag. 2013.
7. Croft B. Search Engines: Information Retrieval in Practice. <https://ciir.cs.umass.edu/downloads/SEIRiP.pdf> (октомври 2022)
8. Using search engines. <https://edu.gcfglobal.org/en/internetbasics/using-search-engines/1/> (октомври 2022)

Формат на изпита

текущ контрол – две контролни по време на лекция
Контролните се провеждат под формата на тест в
системата за е-обучение на ТУ-Варна.

Оценка

Точките от упражнения (до 60т.) се събират с
точките, получени от контролните (до 40т.)

Технически университет – Варна
Катедра “Компютърни науки и технологии”

Христо Георгиев Вълчанов
Венета Панайотова Алексиева

ИЗВЛИЧАНЕ НА ИНФОРМАЦИЯ В ИНТЕРНЕТ



Ръководство за лабораторни упражнения

Варна
2015

Извличане на информация – Information Retrieval (IR)

- “Извличането на информация е област, занимаваща се със структурата, анализа, организацията, съхраняването, търсенето и извличането на информация” (Gerard Salton, 1968).
- Общата дефиниция може да бъде приложена към много типове информация и приложения за търсене.
- Основният фокус на IR от 1950 е върху текста и документите.

Какво е документ?

- Примери: web страници, е-мейли, книги, текстови съобщения, MS Word, PowerPoint, PDF, постинги, патенти и др.
- Общи свойства:
 - Значително текстово съдържание
 - Определена структура (заглавие, автор, тематика, подател и др.)

Документи и записи в бази данни

- Записите в базите данни (tuples) типично се състоят от добре дефинирани полета (attributes):
 - Банков запис с номера на сметки, баланси, имена, адреси, дата раждане и т.н.
- Лесни за сравняване полета с добре дефинирана семантика за заявките с цел откриване на съвпадение.
- Текстът е много по труден за обработка.

Документи и записи в бази данни - пример

- Пример за запитване към банкова база данни:
 - “Намери записи в баланс > 500000 във фирми, намиращи се в Лондон”;
 - Съвпаденията се намират лесно чрез сравнение със стойностите на полетата в записа.
- Пример за запитване към търсеща машина:
 - “Намери банковите скандали с България”;
 - Този текст трябва да бъде сравнен с текста във всички нови публикувани истории.

Сравняване на текст

- Сравняването на текста от запитването с текста в документа и определянето на добро съвпадение е ключов момент в IR.
- Не е необходимо точно съвпадение на думи:
 - Различни начини за запишем едно и също нещо с естествен език;
 - Някои публикации ще имат по-добро съвпадение от други.

Обхват на IR

- IR е много повече от обикновен текст и много повече от обикновено търсене във Web.
- Потребителите, използващи IR работят с различни медии, различни типове приложения за търсене и различни задачи.

Други медии

- Нови приложения, изискващи нови медии:
 - Видео, музика, изображения, говор.
- Съдържанието, както текста, е трудно за описание и сравнение:
 - Текст може да се използва за представяне на съдържание (тагове).

Обхват на IR - медии

| Content | Applications | Tasks |
|--------------|-------------------|--------------------|
| Text | Web search | Ad hoc search |
| Images | Vertical search | Filtering |
| Video | Enterprise search | Classification |
| Scanned docs | Desktop search | Question answering |
| Audio | Forum search | |
| Music | P2P search | |

Задачи на IR

- *Ad-hoc търсене* – намиране на релевантни документи за произволни текстови заявки
- *Филтриране* – идентифициране на интересите на потребителя за откриване на документи.
- *Класифициране* – използване на множества от етикети (класове) за документи.
- *Отговор на запитвания* – получаване на специфичен отговор на запитвания.

Ключови въпроси в IR - *релевантност*

- Какво е?
- **Дефиниция:** *Релевантния документ съдържа информацията, която потребителят търси, когато изпраща заявка към търсещата машина.*
- Редица фактори оказват влияние на преценката на потребителя кое е релевантно – съдържание, новост, стил и др.
- **Тематична релевантност** (същата тема) и **Потребителска релевантност** (всичко допълнително).

Релевантност

- Моделите на IR дефинират идеята за релевантност.
- **Модел на извличане** - формално представяне на процеса на съвпадане на заявката с документ.

Релевантност

- Алгоритмите за рейтинговане (ranking), използвани в търсещите машини, се базират на моделите на IR.
- Повечето модели описват статистически свойства на текста вместо лингвистични такива:
 - Отброяване на обикновени особености на текста вместо парсване и анализ на изречения.

Ключови въпроси в IR - оценка

- Експериментални процедури и измерване за сравняване на изхода на системата с потребителските очаквания.
- Методите за оценка на IR се използват в много области.
- Типично се използват *тестови колекции* от документи, запитвания и доказателства за релевантност (TREC <http://trec.nist.gov>).
- Използване на информация за потребителските активности (log data, clickthrough data).

Оценки

- ***Precision*** – процентът на извлечените документи, които са релевантни.
- ***Recall*** – процентът на релевантните документи, които са извлечени.

(Cyril Cleverdon, 1960)

Ключови въпроси в IR – информационната нужда на потребителите

- Оценката на търсенето е потребителски-центрирана.
- Текстовите запитвания често се явяват лошо описание на това, което потребителите действително искат.
- Техники за прецизиране на запитванията – разширяване на запитванията, подсказвания, обратна връзка за подобряване на рейтинговането.

IR и търсещи машини

- Търсеща машина (Search Engine) е практическо приложение на техниките на извличане на информация от огромни текстови колекции.
- Типични представители:
 - Web Search Engines;
 - Open Source Search Engines (Lucene, Lemur, Galago).
- Основните ключови въпроси са същите, както при IR, но са налични и нови такива.

IR и търсещи машини

Information Retrieval

Relevance

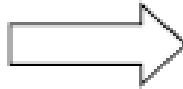
- Effective ranking*

Evaluation

- Testing and measuring*

Information needs

- User interaction*



Search Engines

Performance

- Efficient search and indexing*

Incorporating new data

- Coverage and freshness*

Scalability

- Growing with data and users*

Adaptability

- Tuning for applications*

Specific problems

- e.g. Spam*

Проблеми при търсещите машини - Performance

- Измерване и подобряване продуктивността на търсенето – подобряване времето за отговор, повишаване скоростта на индексирание.
- *Индекси* – даннови структури за повишаване на продуктивността на търсенето.

Проблеми при търсещите машини – Dynamic Data

- Промяна (обновяване, добавяне, изтриване) на колекции данни в които се търси.
- Главна задача – изискване на документи (crawling):
 - Обхват (Coverage) – какъв обем е бил индексирани;
 - Актуалност (Freshness) – колко скоро е бил индексирани.
- Обновяване на индексите докато се обработват запитвания.

Проблеми при търсещите машини – Scalability

- Да се направи така, че всичко да работи всеки ден с милиони потребители и терабайтове информация.

Проблеми при търсещите машини – Adaptability

- Промяна и настройка на компонентите на търсещите машини като алгоритми за рейтинговане, стратегии на индексирание, интерфейси за различни приложения.

Проблеми при търсещите машини – Spam

- Идентифициране и премахване на документи, съдържащи неточни думи.

Въпроси?