

# Търсещи машини

проф. д-р инж. Христо Вълчанов

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

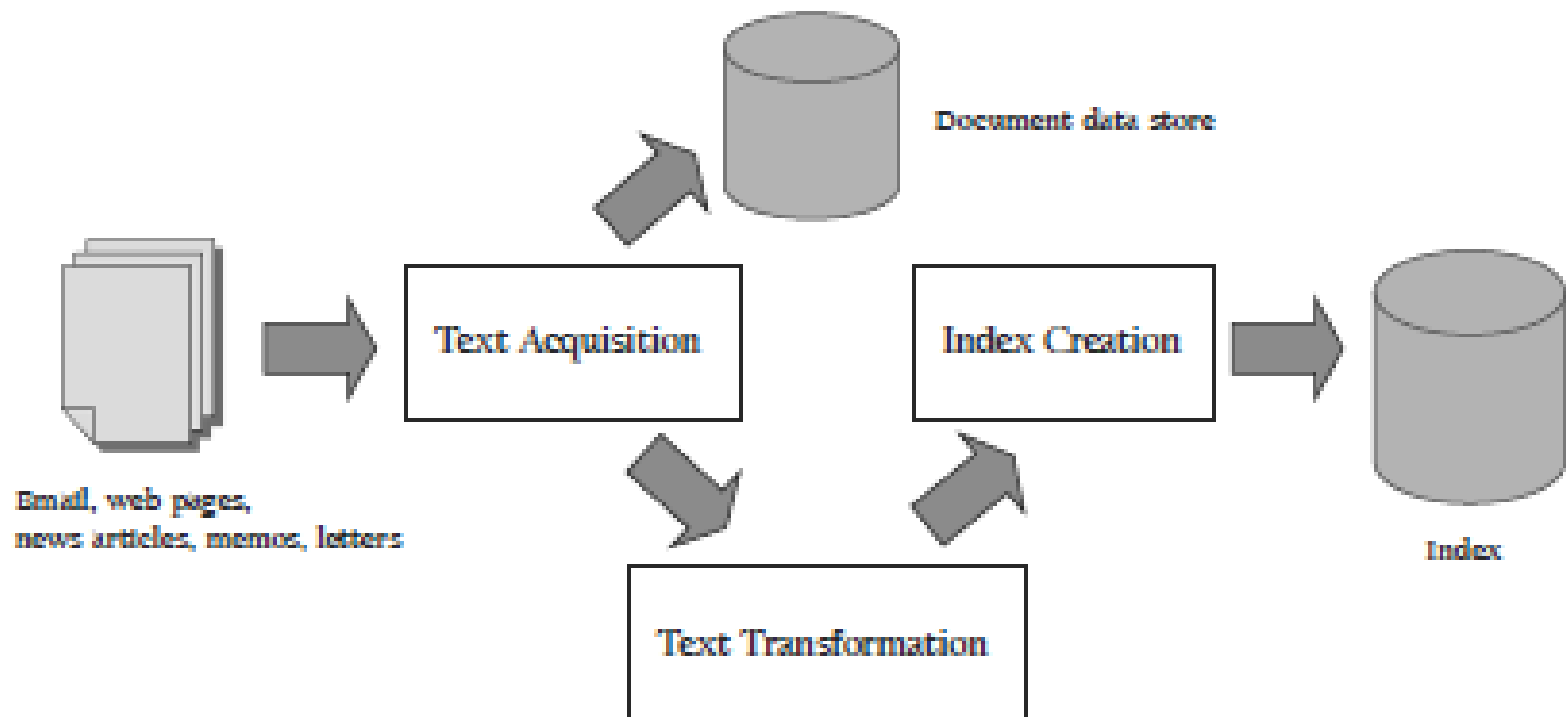
# Архитектура на ТМ

- Софтуерната архитектура съдържа софтуерни компоненти, интерфейси, предоставени от компонентите и отношения между тях.
- Архитектурата се определя от 2 основни изисквания:
  - Ефикасност (време за отговор).
  - Ефективност (качество на резултатите).

# Основни функции на ТМ

- Процес на индексиране.
- Процес на запитване.

# Процес на индексиране



# Изискване на текст

- Crawler:
  - Идентифицира и изисква документи за ТМ;
  - Различни типове – web, enterprise, desktop;
  - Web crawler-ите следват линковете за да откриват документи:
    - Трябва ефикасно да откриват огромен брой страници (coverage) и да ги поддържат актуални (freshness).

# Изискване на текст

- Feeds
  - Потоци документи в реално време (блогове, видео, радио, телевизия).
  - RSS reader е общ стандарт.
- Конвертиране
  - Конвертират се документи в консистентен текстов формат с метаданни (HTML, Word -> XML)
  - Конвертира се текстовото кодиране за различни езици.

# Изискване на текст

- Хранилища на документи
  - Съхраняват текст, метаданни и друга съответстваща информация за документите
    - Метаданни – информация за документа (дата на създаване);
    - Друга информация (линкове, котви).
  - Предоставят бърз достъп до съдържанието на документите за всички компоненти на ТМ.
  - Могат да се използват релационни бази данни.

# Трансформация на текст

- **Парсер**

- Обработване на последователност от текстови лексеми (tokens) за разпознаване на структурни елементи (заглавия, линкове и др.).
- Думите се разпознават от *Tokenizer*.
- Таговите езици (HTML, XML) често се използват за описание на структурата.



# Трансформация на текст

- ***Stopping***
  - Премахване на общи думи.
  - Влияе на ефективността и ефикасността.
  - Може да е проблем при някои запитвания.

# Трансформация на текст

- ***Stemming***

- Групира думи, производни на общ корен.
- Обикновено е ефективно, но не при всички запитвания.
- Предимствата зависят от различните езици.

# Трансформация на текст

- **Анализ на връзки (*Link analysis*)**
  - Използва линковете и котвите в WEB страниците.
  - Анализът определя *популярността* (PageRank).
  - Текстът в котвите може значително да разшири обхвата на страниците, сочени от линковете.
  - Има значително влияние при търсенето в WEB.

# Трансформация на текст

- ***Извличане на информацията***
  - Идентифицира класове на индексни термини, които са важни за определени приложения (класове като хора, компании, дати)

# Трансформация на текст

- **Класификатор**
  - Идентифицира класово-базирани метаданни за документите (присвоява етикети към документи)
  - Зависи от конкретното приложение.

# Създаване на индекс

- **Статистика за документите**
  - Събира се брой и позиция на думите и други свойства
  - Използва се в алгоритмите за рейтинговане.

# Създаване на индекс

- **Задаване на тегла (*Weighting*)**
  - Изчислява тегла за индексните термини.
  - Използва се в алгоритмите за рейтинговане
  - Пример: ***tf.idf***

*Комбинация от честотата на появяване на индексния термин в документа (term frequency) и честотата на появяване на индексния термин в цялата колекция документи (inverse document frequency).*

# Създаване на индекс

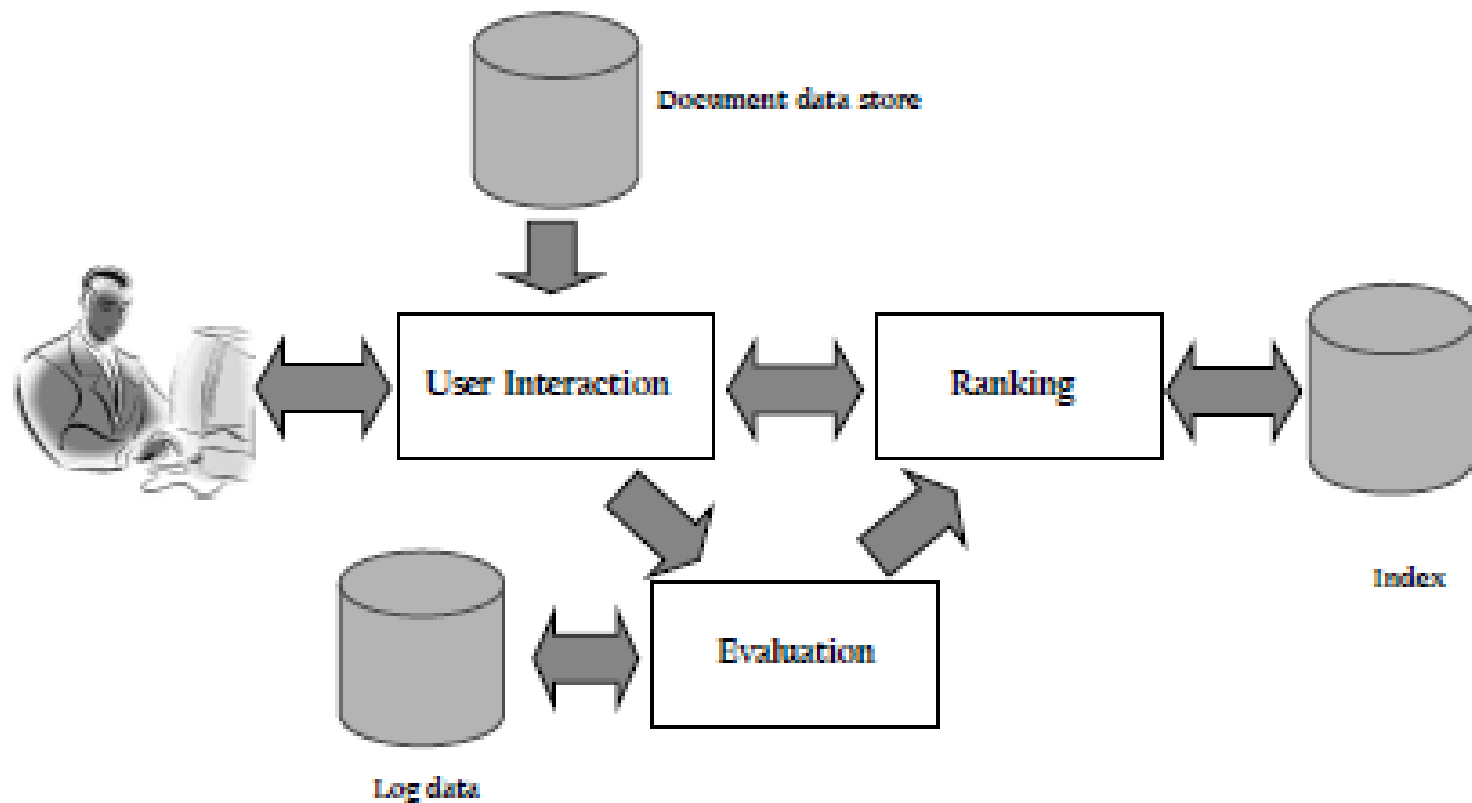
- **Инвертиране**
  - Същината на процеса на индексирание.
  - Преобразува информацията от вида документи-термини във вида термини-документи.
  - Форматът на инвертния файл е проектиран за бърза обработка на запитвания.



# Създаване на индекс

- ***Разпределяне на индекса***
  - Разпределяне на индексите между множество компютри и/или сайтове.
  - Изключително важност за бърза обработка на запитвания с огромен брой документи.
  - Вариации (разпределян на документи, на термини, репликации).

# Процес на запитване



# Взаимодействие с потребителя

- ***Въвеждане на запитвания***
  - Предоставя интерфейс и парсер на езика за запитвания.
  - Повечето web запитвания са много опростени (кавички).
  - Някои приложения могат да използват форми.
  - За описание на сложни запитвания се използват езици за запитвания (query languages).

# Взаимодействие с потребителя

- **Трансформация на запитванията**
  - Подобрява началното запитване (и преди и след първоначалното търсене).
  - Използват се техники за трансформиране на текст, както при документите.
  - *Spell checking, query suggestion* предоставят алтернатива на оригиналните запитвания.
  - *Query expansion, relevance feedback* модифицират оригиналното запитване с допълнителни термини.

# Взаимодействие с потребителя

- **Извеждане на резултати**
  - Формиране на извеждане на рейтинговани документи за запитването.
  - Генериране на *snippets* за показване как запитванията съответстват на документите.
  - Подчертаване на важни думи и пасажии.
  - Получаване на подходящи реклами в много приложения.

# Рейтинговане (Ranking)

- **Формиране на оценка**
  - Изчисляват се оценки за документите на базата на алгоритми за рейтинговане.
  - Ключов компонент на всяка ТМ.
  - Основен вид на оценката е

$\sum q_i d_i$  — теглата на термините на запитването и на документа за термин  $i$ .

# Оценяване

- ***Logging***
  - Съхраняване на потребителските запитвания и взаимодействие (*clickthrough data, dwell time*).
- ***Анализ на рейтинговането***
  - Измерване и настройване на ефективността на рейтинговането.
- ***Анализ на производителността***
  - Измерване и настройване на ефективността на система за търсене (тестови колекции).

# Въпроси?