Информационни операции

Обработка на данните

Теми

- Предпоставки
- Съвременни приложения

Съвременно развитие

- Технологична среда
  - Интернет – информационна вселена без граници
  - мултимедия
  - средства за визуализация
  - изчислителна мощност на компютрите
- Социална среда
  - разширяване на приложните области
  - повишена технологична култура на потребителите
  - повишени изисквания към функционалността на системите

Съвременни приложения

- Извличане на информация
  - Information Retrieval
- Анализ на данни
  - Data Analysis
- Откриване на информация
  - Data Mining
- Откриване на информация в мултимедия
  - Multimedia Mining
- Визуализация на информация
  - Information Visualization
- Други

Извличане на информация

Извличане на информация

- Определения
  - Търсене и извличане на контекстно-зависима информация от данни
  - Събиране, съхраняване, организация и достъп до елементи информация
- Предпоставки за развитие
  - Internet
    - универсално хранилище на човешкото знание и култура
    - безплатен източник на информация
  - Cloud Computing
    - натрупване на голямо количество данни
    - бърз достъп във високо технологична среда

Извличане на информация

- Проблеми
  - нерегламентирани заявки за извличане
  - разнородни информационни източници
  - огромно количество
  - продължително търсене
- Причини за проблемите
  - липса на адекватен единен модел на данните, напр. на тези, съдържащи се в Интернет
  - липса на качествени описания и структури
  - изоставане на софтуерните технологии в сравнение с потребителските фактори

Извличане на информация

- Съвременни изследвания в областта
  - моделиране на данни
  - методи за филтриране, класификация и категоризация на документи
  - потребителски интерфейси
  - езици за формиране на заявки
  - визуализация на информацията
  - системни архитектури
  - и др.

Данни или информация?

- **Извличане на данни**
  - намиране на документи, съдържащи определени ключови думи
  - добре дефинирана семантика
  - нетърпимост към грешки
- **Извличане на информация**
  - информация относно тема
  - недефинирана семантика
  - толеранс на грешките
- **Информационна система IR system**
  - интерпретира съдържанието на намерените документи
  - ранжира намереното по отношение на потребителски критерии
  - оценява приложимостта на намерените документи

Документи

- Основни информационни единици за IR
  - текстови документи
- Съдържание на документите
  - синтаксис
  - семантика
  - структура
- Метаданни за документи
  - дескриптивни
  - семантични

Модели на търсене

- Прост интерактивен модел
  - дълги списъци от намерени документи без оценка на приложимостта
  - статична цел на потребителя
  - итеративно-настройващи се заявки
  - използва се от web search machines
- Самообучение на потребителите
  - използване на хипервръзки и навигация в търсенето (пример: near-miss информационни системи за изследване на възможни бедствия)
  - `berry-picking' model
    - актуализация на целите на потребителя
    - обобщаване на резултати от множество под-заявки
    - главен резултат: акумулираното знание, получено по време на търсенето, не в крайната извадка

Видове търсене

- **разглеждане - browsing**
  - ненасочено изследване на информационни структури
  - следва се от селекция
- **запитване – enquiry**
  - получаване на ново (под) множество от информационни елементи, не събирани заедно досега
  - следва се от разглеждане
- **селектиране - select**
  - избор от организирана информация
  - използва се като резултат или за формулиране на друга операция, напр. запитване
- **сканиране**
  - целенасочено изследване на заглавия, термини, категории и др.
- **навигация**
  - сканиране + селектиране

Формулиране на заявки за търсене

- Методи за дефинирани на заявки
  - Специфициране на думи, фрази, дескриптори за сравнение
  - Избор на ресурсна колекция, метаданни или друго информационно множество, отговарящо на критериите за приложимост на намереното
- Средства за въвеждане на заявки
  - графични средства за манипулация
    - избор от меню
    - директна манипулация с графични обекти по екрана
    - попълване на формуляри с текстови полета и падащи менюта, в които се дефинира заявката QBF
    - описание на пример на намереното QBE
  - вербални средства за описание
    - команден език
    - естествен език

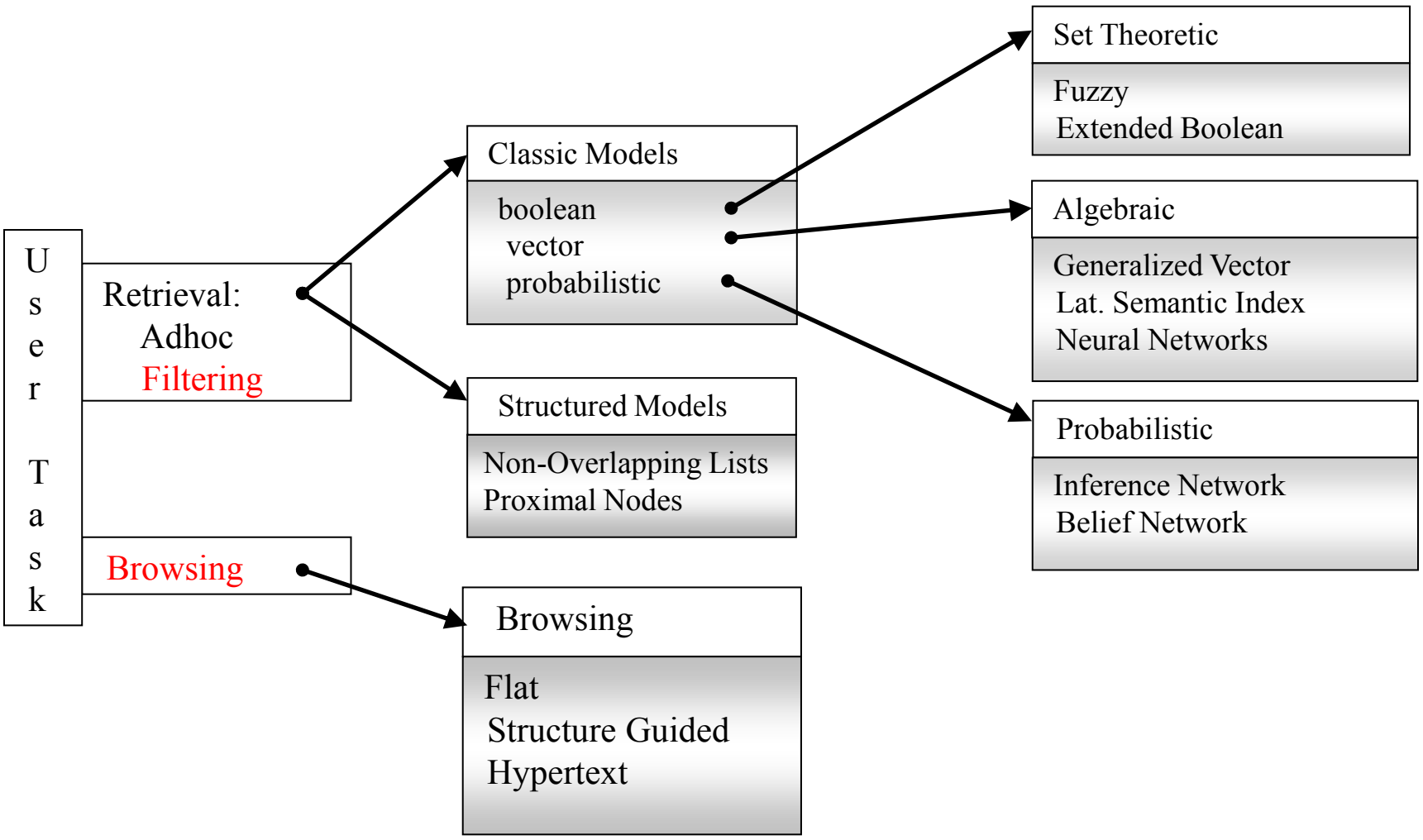
Езици за заявки

- Ключови думи и фрази
  - дизюнкция или конюнкция на отделните думи и фрази?
  - фасети и етикети за улеснение на потребителя
  - приложимостта на резултата зависи от алгоритъма на ранжиране на важността на отделните елементи
    - статистически
      - тегло на важност
      - вероятност за намиране
      - процент на срещане
- Естествен език
  - потребителят контролира приложимостта
  - необходимо е да познава разпознаваемите команди

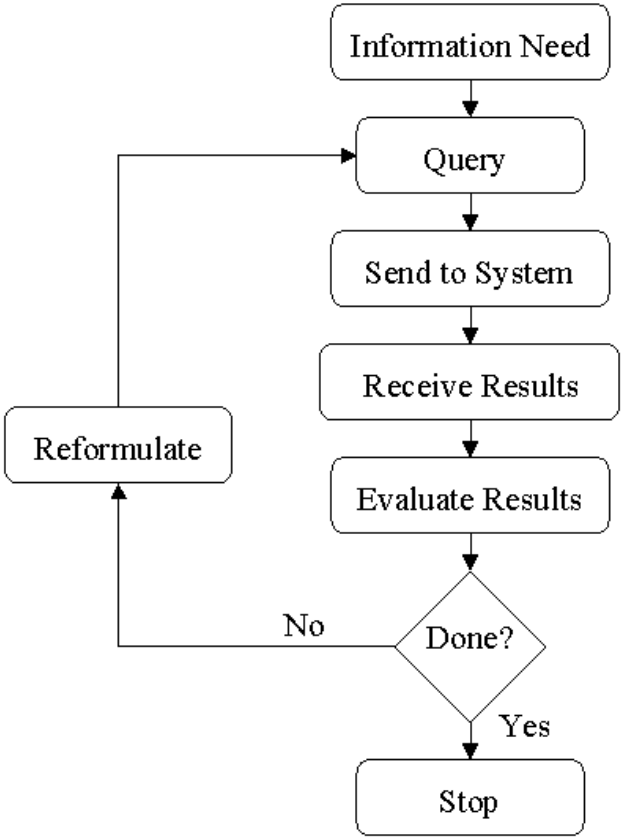
Естествени езици

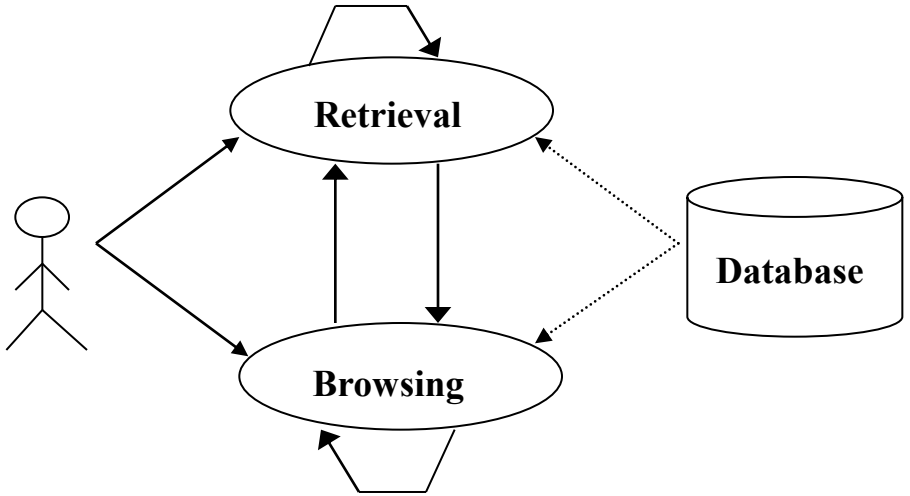
- Въпроси и отговори
  - от синтаксиса на въпроса се извлича информация за свойствата на отговора (напр. каква част от изречението е)
- FAQ finder
  - установяване на съответствие между двойки въпрос и отговор
- Предефинирани типове въпроси
  - идентифициране на типа въпрос на потребителя и последващо перифразиране (доуточняване)
- Свободен текст на въпросите
  - алгоритъм за разделяне на използваните думи и оценяване на тяхното значение

Методи за извличане на информация

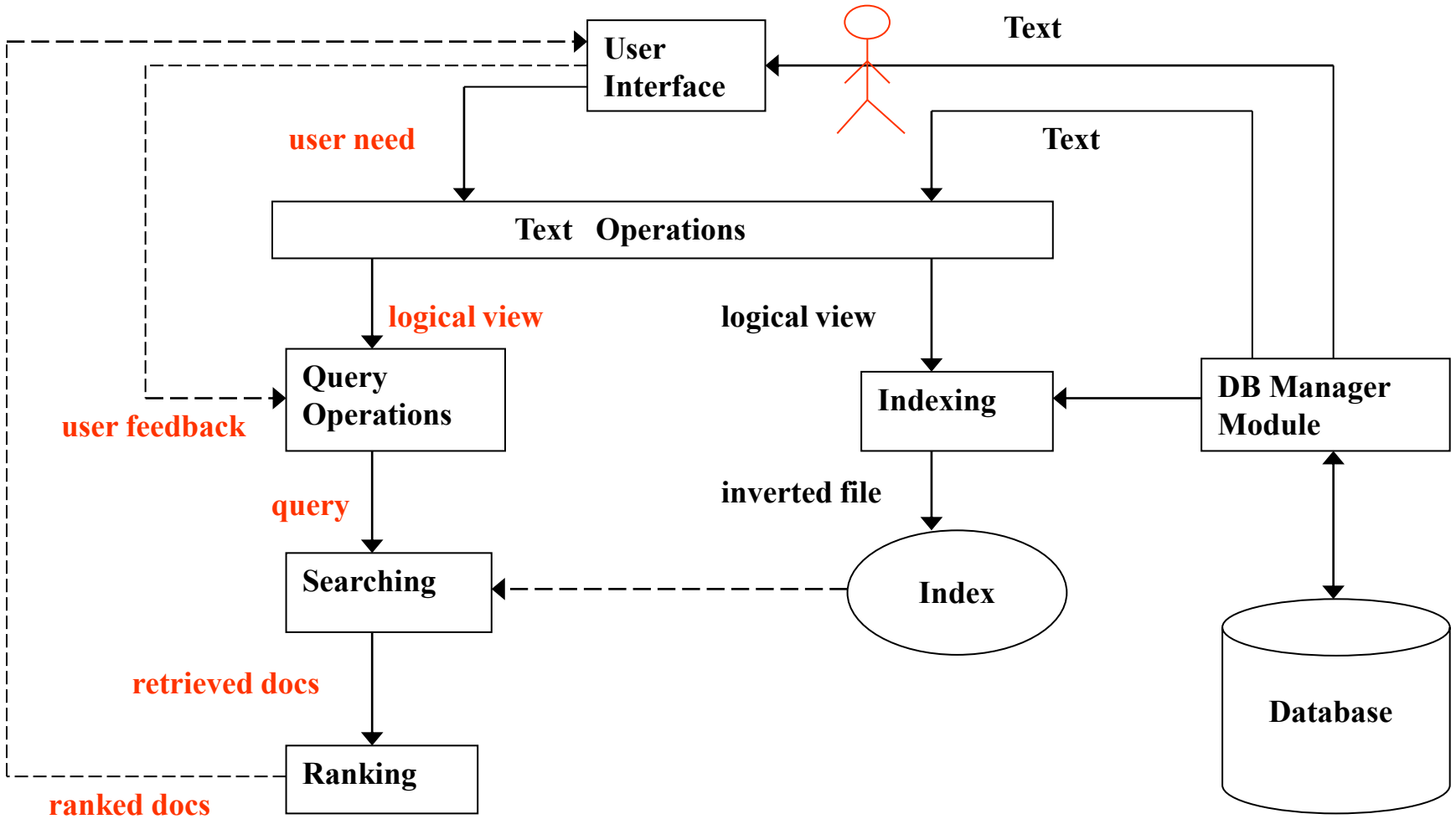


Алгоритми за търсене



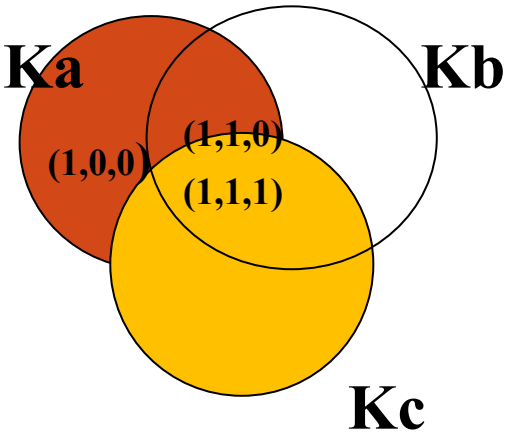


Процес на извличане на информация

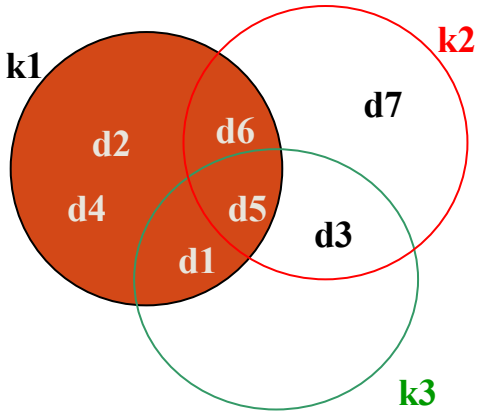


Примери

- Индекси на документите
  - степен на приложимост на документа
  - критерии за приложимост
  - тежест на критериите
- Заявки за намиране на съвпадение на индекси
- Булево търсене в множество => има или няма
  - функция на принадлежност $\{0,1\}$
- Размито множество
  - функция на принадлежност $[0,1]$



Примери



- Векторен модел
  - степен на приложимост на документа по критерии с различна тежест

	k1	k2	k3	q • dj
d1	2	0	1	5
d2	1	0	0	1
d3	0	1	3	11
d4	2	0	0	2
d5	1	2	4	17
d6	1	2	0	5
d7	0	5	0	10
q	1	2	3	

Специализирани приложения

- Digital Libraries
  - Многоезични документи
    - речници
  - Мултимедийни документи
    - синхронизация на потоци информация
    - намиране по метаданни
    - визуални езици за заявки
    - *Query By Image Content*
  - Структурирани документи
    - приложени една или повече структури на данните
    - описания чрез SGML
  - Разпределени документи
    - във физическо или логическо пространство
- Разпределено търсене Federated search
  - изпращане на една заявка към различни сървъри за търсене и последващо обобщаване на резултатите в единен формат

Стандарти за системи IR

- Комуникационни протоколи
  - Z39.50 - протокол за търсене, извличане,сортиране и преглеждане на информация от отдалечена база данни; използват се атрибути: use, relation, position, structure, truncation, completeness.
  - WAIS - Wide Area Information Servers – протокол за търсене на текст в индексирани отдалечени бази от данни
  - Dienst – протикол за форматиране на документи при клиент-сървър комуникации

Стандарти за системи IR

- Стандарти за данни
  - Resource Description Framework (RDF)
    - разделяне на обект и описание
  - Text Encoding Initiative (TEI)
    - комбиниране на данни и метаданни
- Стандарти за метаданни
  - MARC
  - Dublin Core - 15 основни елемента за описание на цифров обект по отношение на:
    - content (Title, Subject, Description, Source, Language, Relation, Coverage)
    - intellectual property issues (Creator, Publisher, Contributor, Rights)
    - digital objects (Data Type, Format, Identifier)
  - Warwick Framework
    - пакети и връзки между тях