دانشگاه صنعتی شاهرود
دانشکده مهندسی کامپیوتر

مبانی بازیابی اطلاعات و جستجوی وب

هدی مشایخی

۱- مقدمه

Definition of *information retrieval*

Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

Retrieving relevant documents to a query (while retrieving as few non-relevant documents as possible)

- especially from large sets of documents efficiently.
- Deals with the representation, storage, organization of information items, and access to them

تعاریف اولیه

- **Document**: a unit decided to build a retrieval system over
  - textual: a sequence of words, punctuation, etc that express ideas about some topic in a natural language.
    - یک متن کوتاه، یک فصل کتاب، ...
- **Information need**: information required by the user about some topics
- **Query**: formulation of the information need

یک سیستم بازیابی اطلاعات متداول

- ورودی: مجموعه اسناد (corpus) و پرس و جوی کاربر (query)
- خروجی: مجموعه رتبه بندی شده از اسناد مرتبط با پرس و جو

سه مقیاس بازیابی اطلاعات

- Web search
- Personal information retrieval
- Enterprise, institutional, and domain specific search

بازیابی داده، بازیابی اطلاعات

- **بازیابی داده**: کدام آیتم ها شامل یک سری کلمات کلیدی هستند؟ یا پرس و جوی کاربر را ارضا می کنند؟
  - ساختار و معنا تعریف شده است.
  - حتی یک آیتم اشتباه نشان دهنده شکست سیستم است.
- **بازیابی اطلاعات**: زبان طبیعی خوش ساختار نیست و ممکن است مبهم باشد، بنابراین معنا قابل انعطاف است
  - اطلاعات در مورد یک موضوع
  - اشتباهات کوچک قابل پذیرش هستند.

اسناد دارای ساختار، بدون ساختار

- Unstructured text (free text): a continuous sequence of tokens
- Structured text (fielded text): text is broken into fields that are distinguished by tags or other markup
- Semi-structured text: e.g. web page

مفاهیم بیشتر

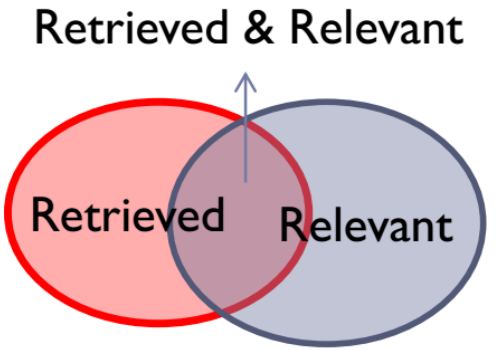
- **Ad hoc retrieval:** اسنادی که مرتبط با یک نیاز اطلاعاتی دلخواه کاربر هستند
- **Relevant:** یک سند مرتبط است اگر از نظر کاربر شامل اطلاعات با ارزش در باب نیاز اطلاعاتی وی باشد. ممکن است دقیقا حاوی کلمات پرس و جو نباشد.
- **Efficiency:** کارایی سیستم بازیابی
  - Precision
  - Recall

کارایی سیستم بازیابی اطلاعات

- **Precision**: Fraction of retrieved docs that are relevant to user's information need
- Precision = relevant retrieved / total retrieved

  = |Retrieved ∩ Relevant | / |Retrieved|
- **Recall**: Fraction of relevant docs that are retrieved
- Recall = relevant retrieved / relevant exist

  = |Retrieved ∩ Relevant | / |Relevant|



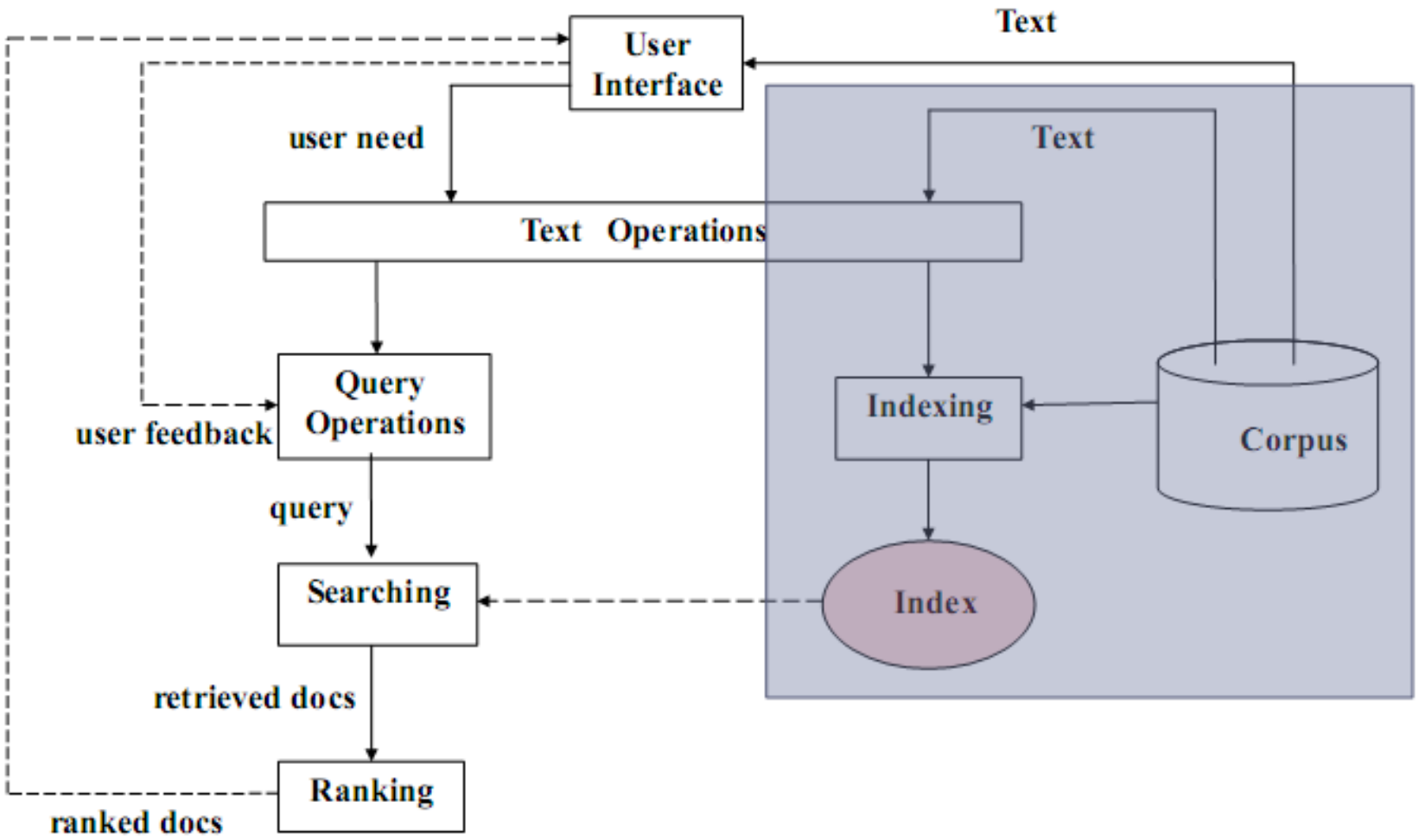
کمینه سازی سربار جستجو

- Search overhead: زمان سپری شده در تمامی قدم ها که در نهایت منتهی به مشاهده آیتم های شامل نیاز اطلاعاتی کاربر می شود.
  - Steps: query generation, query execution, scanning results, reading non-relevant items, etc.

متراکم نمودن داده (Indexing)

- avoid linearly scanning the texts for each query:
  - index the INDEX documents in advance.
- استفاده از ایندکس به جای اسکن خطی اسناد که برای مجموعه های بزرگ از نظر محاسباتی پرهزینه است
  - Indexing depends on the query language and IR model
- **Term** (index unit): A word, phrase, and other groups of symbols used for retrieval

Typical IR system architecture



اجزاء معماری سیستم بازیابی اطلاعات

- **Text Operations:**

  فرمت دهی و استخراج index term

- **Indexing:**

  ایجاد ایندکس برای مجموعه اسناد

- **Query Operations:**

  تغییر پرس و جو برای بهبود بازیابی

- **Searching:**

  بازیابی اسنادی که مرتبط به پرس و جو هستند

اجزاء معماری سیستم بازیابی اطلاعات (ادامه)

- **Ranking:**

رتبه بندی اسناد بازیابی شده با توجه به میزان مرتبط بودن آنها

- **User Interface:**

مدیریت ارتباط با کاربر برای ورود پرس و جو، نمایش پاسخ ها و دریافت بازخورد

جستجوی وب

- کاربرد IR بر روی World Wide Web (اسناد html)
- Web IR:
  - جمع آوری مجموعه اسناد با خزش (crawl) وب
  - استفاده از طرح دارای ساختار صفحات وب
  - به جز term استفاده از ساختار لینک
    - تحلیل لینک، جریان کلیک و غیره

چالش های web IR

- جمع آوری توزیع شده داده ها
- اندازه مجموعه اسناد و حجم پرس و جوهای کاربران
- تخمین میزان ارتباط
- تغییر غیر قابل کنترل اسناد
- اسناد و پرس و جوهای نامعمول و ناهمگن

عناوین اصلی درس

- Introduction
- Indexing & text operations
- IR vector space model
- Evaluation of IR systems
- Machine Learning in IR: Classification, clustering, and ranking
- Web IR

عناوین دیگر در IR

- Cross-language IR
- Image and Multimedia IR, Speech IR, Music IR
- User interfaces for IR
- Parallel and Peer-to-Peer IR
- Digital libraries
- Information science perspective
- Logic-based approaches to IR
- Natural Language Processing techniques
- Recommender systems
- Question answering

مباحث مرتبط

- Database Management
- Artificial Intelligence
- Natural Language Processing
- Machine Learning
- Library and Information Science

منابع

- An introduction to information retrieval فصل اول کتاب

Some slides are extracted from the IR course by Dr Soleymanian