CS47300: Web Information Search and Management

Prof. Chris Clifton
22 August 2018

Material adapted from course created by Dr. Luo Si, now leading Alibaba research group

Administrivia

- Piazza is active
- Blackboard is now available
  - Don’t expect to find much there
- Office Hours for Prof. Clifton:
  - TBD, but I’ll be available tomorrow 12:30-1:30, LWSN 2142F
- TA office hours (and locations) TBD
  - The department is still working on office assignments…
Basic Concepts of IR: Outline

Basic Concepts of Information Retrieval:
- Task definition of Ad-hoc IR
  - Terminologies and concepts
  - Overview of retrieval models
- Text representation
  - Indexing
  - Text preprocessing
- Evaluation
  - Evaluation methodology
  - Evaluation metrics

Ad-hoc IR: Terminologies

Terminologies:
- Query
  - Representative data of user’s information need: text (default) and other media
- Document
  - Data candidate to satisfy user’s information need: text (default) and other media
- Database|Collection|Corpus
  - A set of documents
- Corpora
  - A set of databases
  - Valuable corpora from TREC (Text Retrieval Evaluation Conference)
Ad-hoc IR: Introduction

- Ad-hoc Information Retrieval:
  - Search a collection of documents to find relevant documents that satisfy different information needs (i.e., queries)
  - Example: Web search

Ad-hoc Information Retrieval:

- Search a collection of documents to find relevant documents that satisfy different information needs (i.e., queries)

Relatively Stable

- Queries are created and used dynamically; change fast
- “Ad-hoc”: formed or used for specific or immediate problems or needs” – Merriam-Webster’s collegiate Dictionary

Ad-hoc IR vs. Filtering

- Filtering: Queries are stable (e.g., Asian High-Tech) while the collection changes (e.g., news)
- More for filtering in later lectures
Content Based Filtering

Information Needs are Stable
System should make a delivery decision on the fly when a document “arrives”

User Profile: Asian High-Tech

AD-hoc IR: Basic Process

Information Need

? Representation

Query

Retrieval Model

Indexed Objects

Retrieved Objects

Evaluation/Feedback
AD-hoc IR: Overview of Retrieval Model

Retrieval Models
- **Boolean**
- **Vector space**
  - Basic vector space
  - Extended Boolean
- **Probabilistic models**
  - Statistical language models
  - Two Poisson model
  - Bayesian inference networks
- **Citation/Link analysis models**
  - Page rank
  - Hub & authorities

Purpose of the Retrieval Model
Determine whether a document is relevant to query

- Relevance is difficult to define
  - Varies by judgers
  - Varies by context (i.e., jointly by a set of documents and queries)
- Different retrieval methods estimate relevance differently
  - Word occurrence of document and query
  - In probabilistic framework, \( P(\text{query|document}) \) or \( P(\text{Relevant|query,document}) \)
  - Estimate semantic consistency between query and document
Types of Retrieval Models

• Exact Match (Document Selection)
  – Example: Boolean Retrieval Method
  – Query defines the exact retrieval criterion
  – Relevance is a binary variable; a document is either relevant (i.e., match query) or irrelevant (i.e., mismatch)
  – Result is a set of documents
    • Documents are unordered
    • Often in reverse-chronological order (e.g., Pubmed)

Types of Retrieval Models

• Best Match (Document Ranking)
  – Example: Most probabilistic models
  – Query describes the desired retrieval criterion
  – Degree of relevance is a continuous/integral variable; each document matches query to some degree
  – Result in a ranked list (top ones match better)
    • Often return a partial list (e.g., rank threshold)
Types of Retrieval Models

• Exact Match (Selection) vs. Best Match (Ranking)
  • Best Match is usually more accurate/effective
    – Do not need precise query; representative query generates good results
    – Users have control to explore the rank list: view more if need every piece; view less if need one or two most relevant
  • Exact Match
    – Hard to define the precise query; too strict (terms are too specific) or too coarse (terms are too general)
    – Users have no control over the returned results
    – Still prevalent in some markets (e.g., legal retrieval)

AD-hoc IR: Basic Process
It never leaves my side, April 6, 2002

Reviewer: "dage456" (Carmichael, CA USA) - See all my reviews

It fits in the palm of your hand and is the size of a deflated wallet (wonder where the money went). I have had my ipod now for 4 months and cannot imagine how I used to get by with my old rio 600 with is 64 megs of ram and.. usb connection. Because of its size this little machine goes with my everywhere and its ten hour battery life means I can listen to stuff all day long.

Pros: size, both physical and capacity. design: It looks beautiful controls: simple and very easy to use connection: FIREWIRE!!

Cons: needs the ability to bookmark. I use my ipod mostly for audiobooks. the ipod needs to include a bookmark feature for those like me.

From Amazon Customer Review of IPod
Iran's government is intensifying a birth control program _ despite opposition from radicals _ because the country's fast-growing population is imposing strains on a struggling economy.

………………

An AP Extra

By ED BLANCHE

Associated Press Writer

NICOSIA, Cyprus (AP)
Text Representation: Indexing

Controlled Vocabulary vs. Full Text

• Controlled Vocabulary Indexing
  – Assign words from a small vocabulary or a node from an ontology
  – Often manually but can be done by learning algorithms

• Full Indexing:
  – Often index with an uncontrolled vocabulary of full text
  – Automatically while good algorithm can generate more representative keywords/key concepts

Mutation of a mutL homolog in hereditary colon cancer.


Johns Hopkins Oncology Center, Baltimore, MD 21231.

Some cases of hereditary nonpolyposis colorectal cancer (HNPCC) are due to alterations in a mutS-related mismatch repair gene. A search of a large database of expressed sequence tags derived from random complementary DNA clones revealed three additional human mismatch repair genes, all related to the bacterial mutL gene. One of these genes (hMLH1) resides on chromosome 3p21, within 1 centimorgan of markers previously linked to cancer susceptibility in HNPCC kindreds. Mutations of hMLH1 that would disrupt the gene product were identified in such kindreds, demonstrating that this gene is responsible for the disease. These results suggest that defects in any of several mismatch repair genes can cause HNPCC.
MeSH Tree Structures

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
   - Bacterial Infections and Mycoses [C01] -
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - Digestive System Diseases [C06] +

4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Physical Sciences [H]

Text Representation: Indexing

PMID: 8128251

TI - Mutation of a mutL homolog in hereditary colon cancer.
MH - *Adenosinetriphosphatase
MH - Amino Acid Sequence
MH - Bacterial Proteins/chemistry/*genetics
MH - Base Sequence
MH - Carrier Proteins
MH - Chromosome Mapping
MH - *Chromosomes, Human, Pair 3
MH - Codon
MH - Colorectal Neoplasms, Hereditary Nonpolyposis/*genetics
MH - *DNA Repair
MH - *DNA-Binding Proteins
Text Representation: Indexing
Controlled Vocabulary

• Pros and cons of controlled vocabulary indexing
• Advantages
  – Many available vocabularies/ontologies (e.g., MeSH, Open Directory, UMLS)
  – Normalization of indexing terms: less vocabulary mismatch, more consistent semantics
  – Easy to use by RDBMS (e.g., semantic Web)
  – Support concept based retrieval and browsing
• Disadvantages
  – Substantial efforts to be assigned manually
  – Inconvenient for users not familiar with the controlled vocabulary
  – Coarse representation of semantic meaning