Information Retrieval and Web Search

Introduction

Information Retrieval (IR)

- The indexing and retrieval of textual documents.
- Searching for pages on the World Wide Web is the “killer app.”
- Concerned firstly with retrieving relevant documents to a query.
- Concerned secondly with retrieving from large sets of documents efficiently.

Typical IR Task

- Given:
  - A corpus of textual natural-language documents.
  - A user query in the form of a textual string.
- Find:
  - A ranked set of documents that are relevant to the query.
Relevance

- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (information need).

Keyword Search

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order (bag of words).
Problems with Keywords

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
  - “PRC” vs. “China”
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - “bit” (unit of data vs. act of eating)

Beyond Keywords

- We will cover the basics of keyword-based IR, but...
- We will focus on extensions and recent developments that go beyond keywords.
- We will cover the basics of building an efficient IR system, but...
- We will focus on basic capabilities and algorithms rather than systems issues that allow scaling to industrial size databases.

Intelligent IR

- Taking into account the meaning of the words used.
- Taking into account the order of words in the query.
- Adapting to the user based on direct or indirect feedback.
- Taking into account the authority of the source.
IR System Architecture

IR System Components
- **Text Operations** forms index words (tokens).
  - Stopword removal
  - Stemming
- **Indexing** constructs an *inverted index* of word to document pointers.
- **Searching** retrieves documents that contain a given query token from the inverted index.
- **Ranking** scores all retrieved documents according to a relevance metric.

IR System Components (continued)
- **User Interface** manages interaction with the user:
  - Query input and document output.
  - Relevance feedback.
  - Visualization of results.
- **Query Operations** transform the query to improve retrieval:
  - Query expansion using a thesaurus.
  - Query transformation using relevance feedback.
Web Search

- Differences:
  - Must assemble document corpus by spidering the web.
  - Can exploit the structural layout information in HTML (XML).
  - Documents change uncontrollably.
  - Can exploit the link structure of the web.

Web Search System

Other IR-Related Tasks

- Automated document categorization
- Information filtering (spam filtering)
- Information routing
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering
History of IR

• 1960-70’s:
  – Initial exploration of text retrieval systems for “small” corpora of scientific abstracts, and law and business documents.
  – Development of the basic Boolean and vector-space models of retrieval.
  – Prof. Salton and his students at Cornell University are the leading researchers in the area.

IR History Continued

• 1980’s:
  – Large document database systems, many run by companies:
    • Lexis-Nexis
    • Dialog
    • MEDLINE

IR History Continued

• 1990’s:
  – Searching FTPable documents on the Internet
    • Archie
    • WAIS
  – Searching the World Wide Web
    • Lycos
    • Yahoo
    • Altavista
IR History Continued

• 1990’s continued:
  – Organized Competitions
    • NIST TREC
  – Recommender Systems
    • Ringo
    • Amazon
    • NetPerceptions
  – Automated Text Categorization & Clustering

IR History Continued

• 2000’s
  – Link analysis for Web Search
    • Google
  – Automated Information Extraction
  – Parallel Processing
    • Map/Reduce
  – Question Answering
    • TREC Q/A track

IR History Continued

• 2000’s continued:
  – Multimedia IR
    • Image
    • Video
    • Audio and music
  – Cross-Language IR
    • DARPA Tides
  – Document Summarization
  – Learning to Rank
Recent IR History

- 2010’s
  - Intelligent Personal Assistants
    - Siri
    - Cortana
    - Google Now
    - Alexa
  - Complex Question Answering
    - IBM Watson
  - Distributional Semantics
  - Deep Learning

Related Areas

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

Database Management

- Focused on *structured* data stored in relational tables rather than free-form text.
- Focused on efficient processing of well-defined queries in a formal language (SQL).
- Clearer semantics for both data and queries.
- Recent move towards *semi-structured* data (XML) brings it closer to IR.
Library and Information Science

• Focused on the human user aspects of information retrieval (human-computer interaction, user interface, visualization).
• Concerned with effective categorization of human knowledge.
• Concerned with citation analysis and bibliometrics (structure of information).
• Recent work on digital libraries brings it closer to CS & IR.

Artificial Intelligence

• Focused on the representation of knowledge, reasoning, and intelligent action.
• Formalisms for representing knowledge and queries:
  – First-order Predicate Logic
  – Bayesian Networks
• Recent work on web ontologies and intelligent information agents brings it closer to IR.

Natural Language Processing

• Focused on the syntactic, semantic, and pragmatic analysis of natural language text and discourse.
• Ability to analyze syntax (phrase structure) and semantics could allow retrieval based on meaning rather than keywords.
Natural Language Processing: IR Directions

- Methods for determining the sense of an ambiguous word based on context (*word sense disambiguation*).
- Methods for identifying specific pieces of information in a document (*information extraction*).
- Methods for answering specific NL questions from document corpora or structured data like FreeBase or Google’s Knowledge Graph.

Machine Learning

- Focused on the development of computational systems that improve their performance with experience.
- Automated classification of examples based on learning concepts from labeled training examples (*supervised learning*).
- Automated methods for clustering unlabeled examples into meaningful groups (*unsupervised learning*).

Machine Learning: IR Directions

- Text Categorization
  - Automatic hierarchical classification (Yahoo).
  - Adaptive filtering/routing/recommending.
  - Automated spam filtering.
- Text Clustering
  - Clustering of IR query results.
  - Automatic formation of hierarchies (Yahoo).
- Learning for Information Extraction
- Text Mining
- Learning to Rank