Information Retrieval and Web Search

What is Information Retrieval (IR)

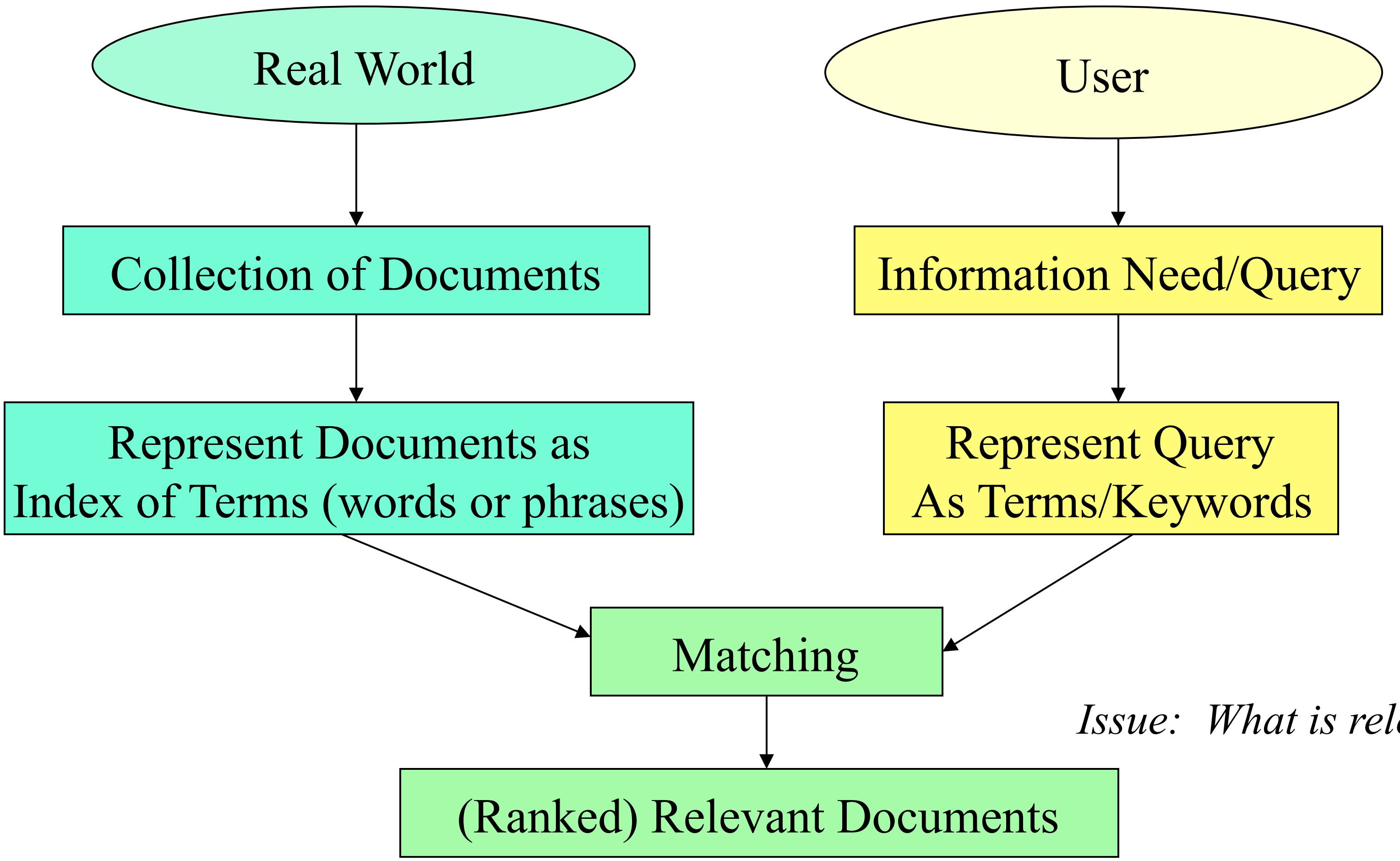
- Gerard Salton, 1968:
  Information retrieval is a field concerned with the structure, analysis, organization, storage, searching and retrieval of information
- Manning, Raghavan and Schutze, 2008:
  Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers)
    - "Document" is the generic term for an information holder (book, chapter, article, webpage, etc)

- Web Search is the branch of IR where the collection of documents are those located on the web

What is tough about IR?

- One issue is how to **represent documents** so others might retrieve them
  - Need to match the text of the document with the query
  - In full, free-text systems this is an issue because documents and queries are expressed in language
    - and language is synonymous and polysemous
    - methods for solving the language issue are difficult
  - Sometimes called the **vocabulary gap** or mismatch
- Given the retrieval of some documents how to decide which ones are **most relevant** to the user's query
  - Most often implemented as a **ranking** of the resulting documents

Typical Information Retrieval System



Issue: What is relevance?

Text Retrieval Conference (TREC)

- Co-sponsored by the National Institute of Standards and Technology (NIST) & the Defense Advanced Research Projects Agency (DARPA)
  - Begun in 1992
- Purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.
  - Provides document collections, queries and human judges
  - Main IR track was called the “Ad-Hoc Retrieval Track”
- Has grown in the number of participating systems and the number of tracks each year.
  - Tracks have included cross-language retrieval, filtering, question answering, interactive, web, novelty, video, blog search …

IR System Research

- Traditional IR System research assumes that a user is interested in finding out information on a particular topic
- TREC collections and research experiments
  - build IR systems with different retrieval models
  - test against a standard collection of newswire documents
  - human evaluators judge relevant documents
  - report system evaluations in terms of precision and recall
  - Example type of query:

    I am interested in all documents that discuss oil reserves and current attempts to find new reserves, particularly those that discuss the international financial aspects of the oil production process.

Information Needs

- Other branches of research focus on the user and whether the user's underlying information seeking is satisfied
- Early theories by Belkin, Oddy, etc.
  - Functions of the retrieval system to model the user's information need in an interactive retrieval session:
    - Characterize User
    - Get initial information need
    - Develop need context
    - Formulate information need
    - Conduct search for documents
    - Evaluate results
    - Feedback from user

IR Systems: Constructing the Index

- Process documents and identify terms to be indexed
  - Terms are often just the words
    - Usually stemming is applied and stop words removed
  - Sometimes basic noun phrases are also added, particularly proper names
- Compute weights of terms, depending on model definition
- Build index, a giant dictionary mapping terms to documents
  - For each term,
    - keep a list of documents that it occurs in
    - weights

IR Systems: Models

- Vector Space Models
  - Widely used weights known as TF-IDF (term frequency / inverted document frequency)
    - TF – frequency of the term in the document (normalized by document length): $freq_{td}$ / $length_d$
      - Intuition: more frequently occurring terms are more important
    - IDF – invert the document frequency, the number of documents in the collection that the term occurs in: $\log(N / n_t)$

      where N is number of docs, n_t is number of docs with term t
      - Intuition: terms occurring in all documents are less important to distinguish which ones are relevant to the query
    - TF-IDF = TF * IDF
- Other models include
  - Probabilistic models
  - Language models
  - Boolean models

IR Systems: Queries and matching

- Natural language queries are converted to terms, usually called keywords
  - In web search, typical queries are keywords already
- Query terms are used to retrieve documents from the index
- Model defines how to match query terms to documents, using the weights, and usually resulting in a score for each document
- Documents are returned in order of relevance score

IR Systems: Evaluation

- Human judgments as to whether returned documents are relevant to the query
- Precision and recall can be used to evaluate a set of returned documents

Human judgments -> System:	**Relevant**	**Non-Relevant**
Retrieved	a (true positives)	b (false positives)
Non-Retrieved	c (false negatives)	d (true negatives)

Precision = a / a + b

Recall = a / a + c

IR Systems: Another Evaluation Measure

- The F-measure is a combination of recall and precision, averaged using the harmonic mean
  - Let P be precision and R be recall

$$F = (\beta^2 + 1) PR / (\beta^2) P + R$$

  - Typically, the measure is used for $\beta = 1$, giving equal weight to precision and recall

$$F_{\beta=1} = 2 PR / P + R$$

- Ranked Retrieval Evaluation
  - Given the top k ranked documents, compute precision and recall at every position
  - Mean Average Precision
    - Average the precisions over all positions k in the ranking

IR Systems: Improving Retrieval

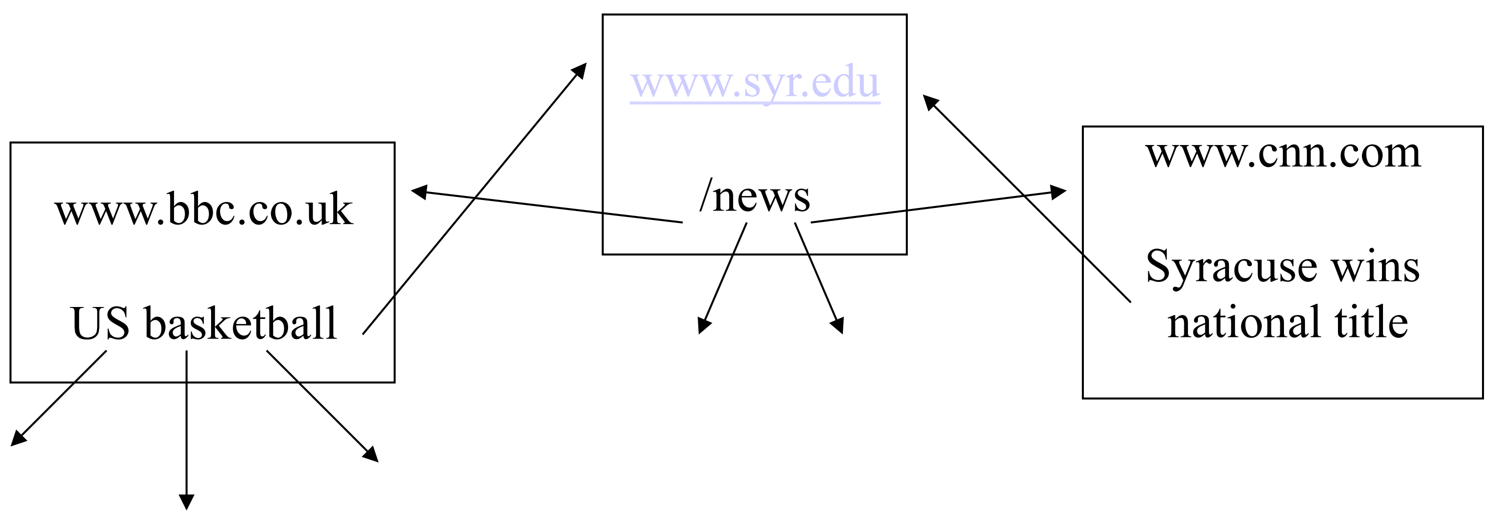
- Query expansion, adding semantically similar words or context words
  - For example, use WordNet to add synonyms to query terms
    - What sense to use? The first?
  - Results are mixed
    - Synonyms added for incorrect sense will throw results off badly
- Relevance Feedback
  - One technique consistently shown to improve retrieval
  - Human relevance feedback – after human has selected a few really relevant documents, add terms from those documents to the query
  - Pseudo-relevance feedback
    - Perform one retrieval and assume that the top n documents are relevant
    - Use those documents to add terms to the query

Web Search

- With the advent of the Web, basic IR was applied to this scenario of linked documents world wide
  - Company like Google keeps a **giant index** of documents for search
- Why/How would IR be different on the Web
  - Compared to a database of documents, the Web
    - Is far larger
    - Is more dynamic: web sites update whenever, links may not be permanent
    - Collection frequencies needed for Inverse Document Frequency (IDF) are so impermanent
  - Quality control of documents on the web is not present
  - No such thing as a complete inverted file for the entire web – many hidden pages (Deep Web)
  - Importance of ranking results
    - Impact of pay for ranking

The Web has structure: Web Graph

- View the collection of static web pages as a graph with “hyperlinks” between them
- Hyperlink in HTML, given by the anchor tag, will give the URL of another web page
  - in-degree is the number of links coming to a page from other pages
  - out-degree is the number of links on the page



Building Search Engines: Web Crawling

- In order to build an index of documents for web search, the web crawler, or spider, has to locate documents
- Required Features:
  - Robustness – it must not get stuck in dead ends or loops
  - Politeness – it must not overwhelm any web server with too fast or too many requests
    - web servers set politeness policies
- Desired Features
  - Quality – should try to give "useful" pages priority
  - Freshness – should obtain updated pages so that the web index has a fairly current version of the web page
  - Performance and efficiency, scalability, operate in a distributed fashion

Building Search Engines: Web Document Processing

- Find content and process into tokens for traditional use in IR indexing
  - content may be text in-between tags
  - image tags may have text attributes to describe the image
  - may discard javascript and other computational elements
  - may even try to discard "noisy" text in the form of web site navigation, standard copyright notices, etc.
    - one technique is to observe that real content text has fewer tags per token than non-content text
- Keywords may be added to the document that don't appear directly in the content
  - metadata tags may have keywords
  - special weights may be added for tokens appearing in header tags
  - anchor text from other pages (see next slide)

Building Search Engines: Anchor Text

- Sometimes the text content of a web page does not contain generally descriptive words for that page
  - home page for IBM did not contain the word “computer”
  - home page for Yahoo did not contain the word “portal”
- Generally descriptive words may be found in anchor text of links, or even near it, that occur in other pages

   Big Blue

  example of a large computing firm

   Big Blue

  an example of a large computing firm is

   here

  - typically, we disregard anchor text words such as “click” and “here”

Building Search Engines: Link Analysis

- Link analysis can be viewed as a development of citation analysis for the web
  - Bibliographic citation analysis used book and article references
  - Bibliometric analysis of bibliographic citation links
    - Web examples: Web of Science from ISI / Citeseer
- The intuition behind link analysis is that a hyperlink from page A to page B represents an endorsement of page B, by the creator of page A.
  - not true for some links, such as links to adminstrative notices on corporate websites - "internal" links are typically discounted.
- Two major algorithms, PageRank and HITS, that give scoring weights for web pages
  - PageRank is from Sergei and Brin, founders of Google
  - such weights are combined with other weights from content tokens and many other ranking criteria

Additional Criteria for Ranking

- Popularity – what are the current topics of the day?
  - Collected from blogs and previous queries
- Click-through results – statistics about which pages users click-on after getting ranked results can inform ranking algorithms to improve later rankings
- Context – keep track of the user's interests, location, situation
  - What do other users like this one like?
- Learning to rank – use machine learning on ranked relevance results to improve rankings
  - Importance of getting relevant documents in the top 10 list
  - Search engine companies have large amounts of data, including relevance judgments in terms of what documents users click on after a query

Evaluating Ranked Retrieval Results

- Evaluation measure: Discounted Cumulative Gain (DCG)
  - Measures relevance at each ranked position
  - Penalizes highly relevant documents that are lower down in the ranks
  - nDCG normalizes over queries (of different lengths)
- Experiments for search engines
  - User judgments are good, but are necessarily small in scope
  - A/B testing
    - Deploy an experimental search engine to some users (group B) while other users get "normal" search engine (group A)
    - Use "click-through" judgments as to which results the users thought would be relevant
    - Evaluate which relevant results are highest ranked

Where is the NLP in IR?

- IR is thought of as a field in its own right, but
- NLP is used at the lower levels in building indexes
  - Stemming, stopwords
  - Named entities and other noun phrases
- Web Search engines incorporating natural language and NLP into queries
  - Not just keywords anymore
  - Query processing to find commonly used patterns and tailor the searches and search results
    - Conversion of units
    - Queries with "how do I . . ." or other patterns
    - Other query processing techniques similar to those in question/answering

Search Engine Company Data Centers

- Google designs data centers specifically for web search
  - http://www.google.com/about/datacenters/
  - Uses lots of low-cost computers networked together
    - Design network and data algorithms for fast performance
  - Acknowledges 15 data centers around the world (probably 2 dozen more)
    - Each data center has up to 10,000 computers
    - Data center at The Dalles, Oregon