PURDUE UNIVERSITY

CS34800
Information Systems

Information Retrieval

Prof. Chris Clifton

14 November 2016

Indiana Center for Database Systems

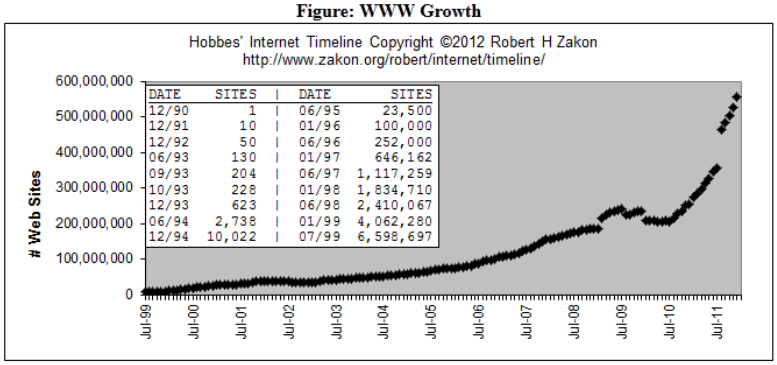
1

Why Information Retrieval:

Information Overload:

"… The world produces between 1 and 2 exabytes (10^{18} bytes) of unique information per year, which is roughly 250 megabytes for every man, woman, and child on earth. …" (Lyman & Hal 03)

Figure: WWW Growth



DATE	SITES	DATE	SITES
12/90	1	06/95	23,500
12/91	10	01/96	100,000
12/92	50	06/96	252,000
06/93	130	01/97	646,162
09/93	204	06/97	1,117,259
10/93	228	01/98	1,834,710
12/93	623	06/98	2,410,067
06/94	2,738	01/99	4,062,280
12/94	10,022	07/99	6,598,697

Why Information Retrieval:

Information Retrieval (IR) mainly studies unstructured data:

Text in Web pages or emails; image; audio; video; protein sequences..

Merrill Lynch estimates that more than 85 percent of all business information exists as unstructured data - commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, … and Web pages.

Unstructured data:

No structure: no primary key as in RDBMS

Semantic meaning unknown: natural language processing systems try to find the meaning in the unstructured text

IR vs. RDBMS

- Relational Database Management Systems (RDBMS):
  - Semantics of each object are well defined
  - Complex query languages (e.g., SQL)
  - Exact retrieval for what you ask
  - Emphasis on efficiency
- Information Retrieval (IR):
  - Semantics of object are subjective, not well defined
  - Usually simple query languages (e.g., natural language query)
  - You should get what you want, even the query is bad
  - Effectiveness is primary issue, although efficiency is important

IR vs. RDBMS

RDBMS and IR get close to each other

RDBMS -> IR

- Combine exact search and inexact text search

  Find an article published between 1999 and 2004 that talks about Oracle and Internet.

IR -> RDBMS

- Use information extraction to convert unstructured data to structured data: extract company names and their headquarter locations from news
- Semi-structured representation: XML data; queries with structured information

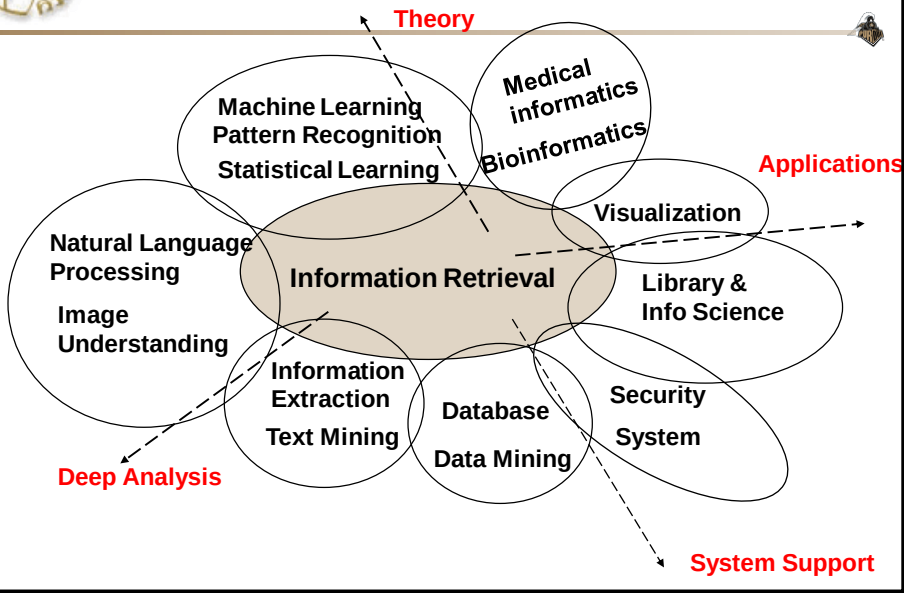
Information Retrieval: Challenges

- Data is unstructured
  - Need to guess what is ~~important~~ relevant
- Query is unstructured
  - Need to guess user intent
- But computers don't guess!

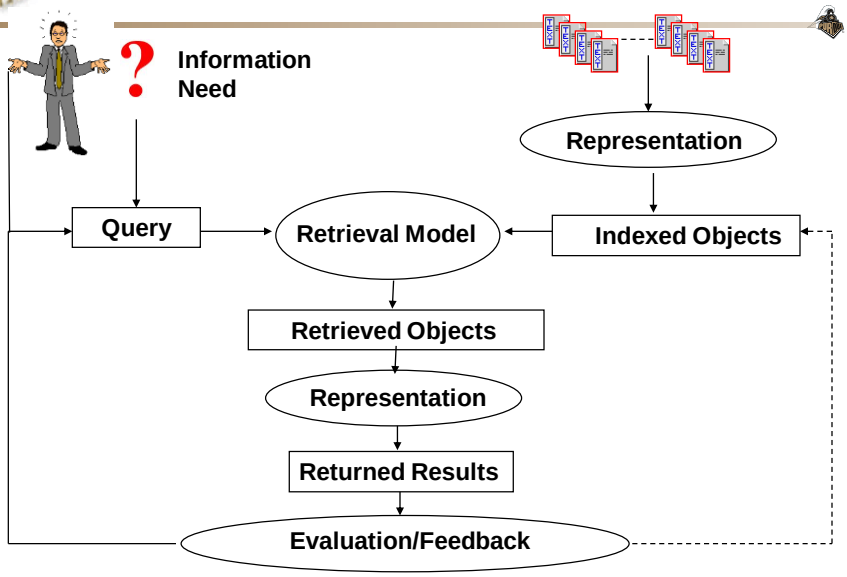
Inferring relevance and intent from data, query is the science of Information Retrieval

6

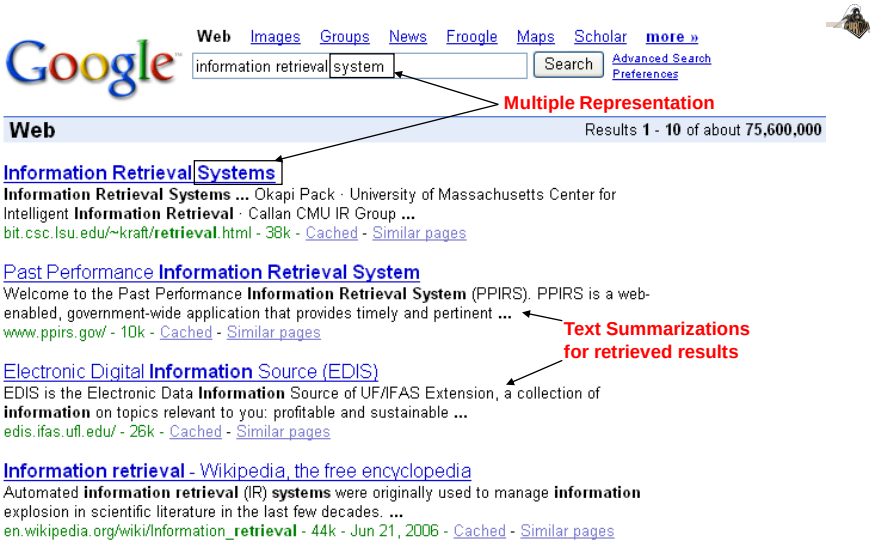
IR and other disciplines



Some core concepts of IR



Some core concepts of IR



Some core concepts of IR

Query Representation:

- Bridge lexical gap: system and systems; create and creating (stemmer)
- Bridge semantic gap: car and automobile (feedback)

Document Representation:

- Internal representation of document contents: a list of documents that contain specific word (inverted document list)
- Representation of document structure: different fields (e.g., title, body)

Retrieval Model:

- Algorithms that best match meaning of user query and available documents. (e.g., vector space model and statistical language modeling)

IR Applications

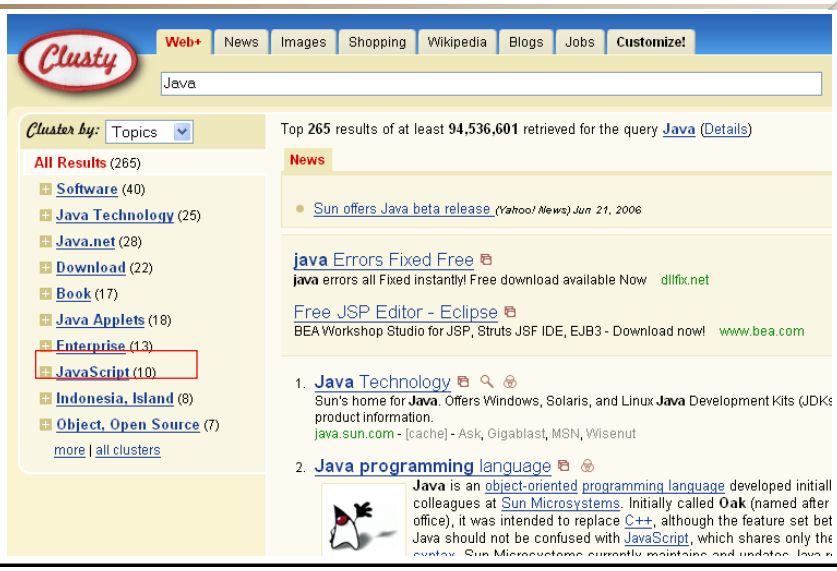
Information Retrieval: a gold mine of applications

- Web Search
- Information Organization: text categorization; document clustering
- Information Recommendation by content or by collaborative information
- Information Extraction: deep analysis of the surface text data
- Question-Answering: find the answer directly
- Federated Search: explore hidden Web
- Multimedia Information Retrieval: image, video
- Information Visualization: Let user understand the results in the best way
- ……..

IR Applications: Text Categorization

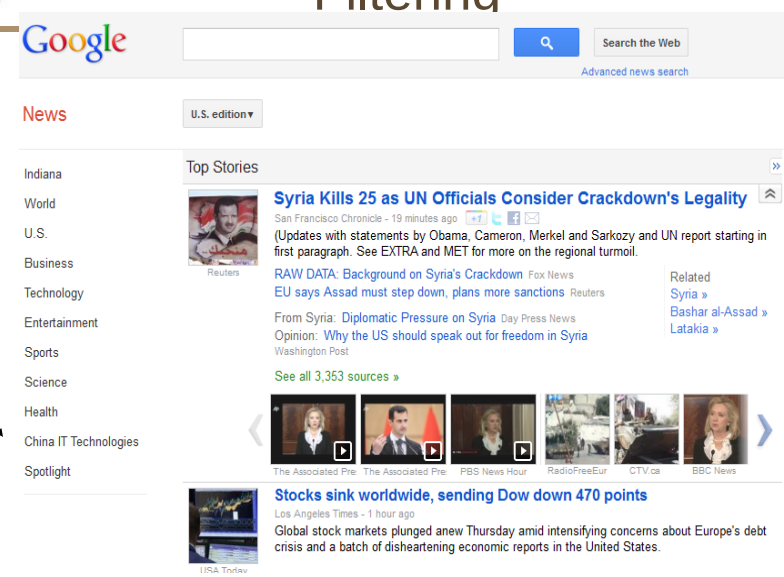
News Categories

IR Applications: Document Clustering

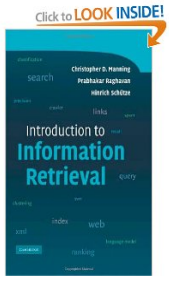


IR Applications: Content Based Filtering

Keyword Matching



IR Applications: Collaborative Filtering



Introduction to Information Retrieval [Hardcover]

Christopher D. Manning (Author), Prabhakar Raghavan (Author), Hinrich Schütze (Author

(14 customer reviews) | Like (6)

List Price: $64.00

Price: $51.50 & this item ships for **FREE with Super Saver Shipping**. Details

You Save: $12.50 (20%)

Special Offers Available

In Stock.

Ships from and sold by **Amazon.com**. Gift-wrap available.

Want it delivered Friday, August 19? Order it in the next 6 hours and 54 minutes, and choos checkout. Details

35 new from $48.00 32 used from $44.00

FREE Two-Day Shipping for Students. Learn more

Customers Who Bought This Item Also Bought

Other Customers with similar tastes

Foundations of Statistical Natural Lang... by Christopher D. Manning

(13)

$56.69

Algorithms of the Intelligent Web by Haralambos Marmanis

(12)

$24.74

The Elements of Statistical Learning: Data Minin... by Trevor Hastie

(45)

$59.09

Speech and Language Processing (2nd Edition) by Daniel Jurafsky

(27)

$101.17

IR Applications: Information Extraction

Bring structure and semantic meaning to text:

- Entity detection

  An 80-year-old woman with diabetes mellitus was treated with gliclazide. Prior to the gliclazide administration, her urinary excretion of albumin, serum urea nitrogen and serum creatinine were normal. After the medication, oliguria, edema and azotemia developed. On the twenty-fourth day when the edema was severe and generalized, gliclazide administration was terminated.

  Diabetes: entity of disease gliclazide: entity of drug

- Recognize Relationship between entities

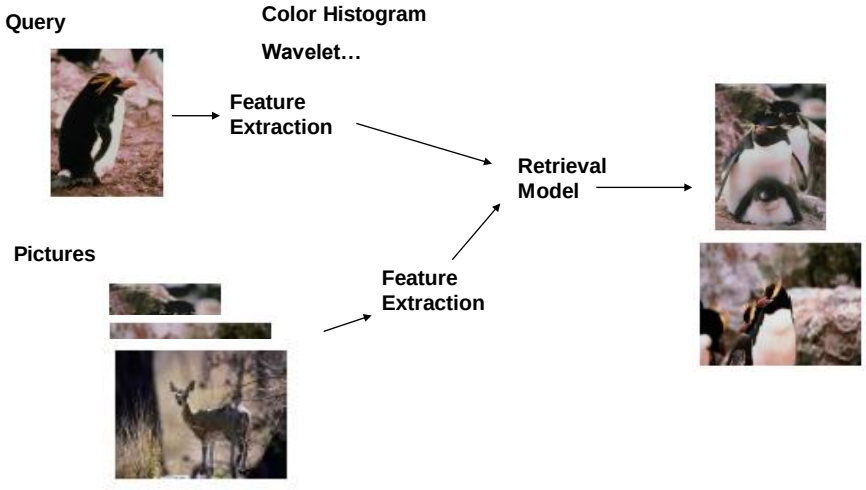
  What type of effect of gliclazide on this patient with diabetes

- Inference based on the relationship between entities

Inherited Disease → Gene → Chemical

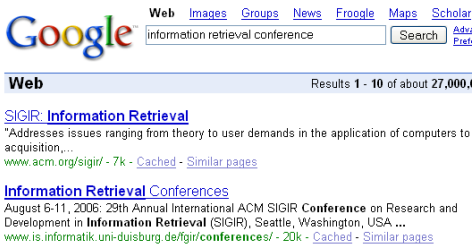
Drug discovery

IR Applications: Multimedia Retrieval



Ad-hoc Information Retrieval

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



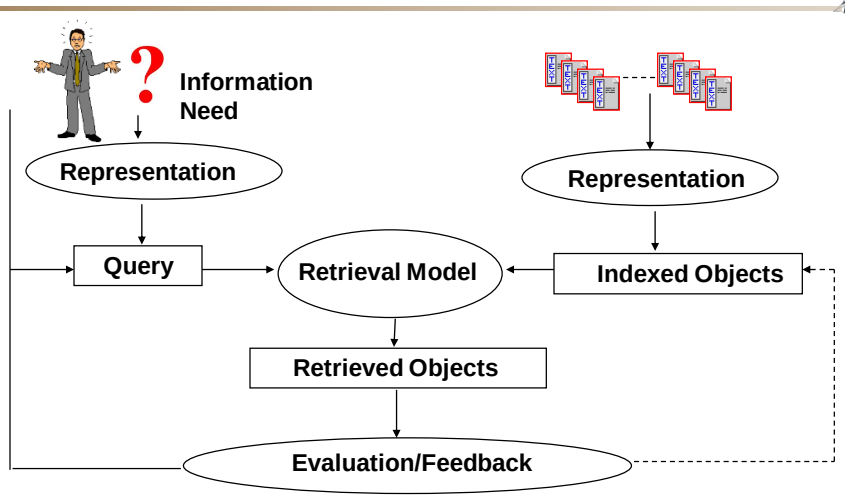
Ad-Hoc IR: Outline

- 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

- 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: Basic Process



AD-hoc IR: Overview of Retrieval Model

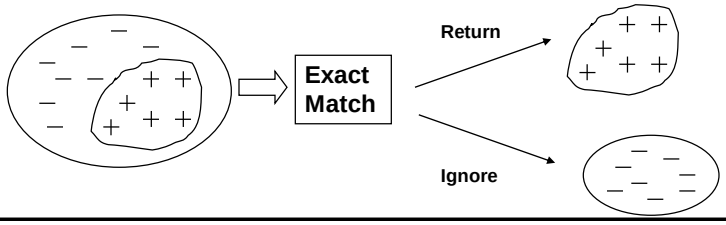
Retrieval Model

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(query|document) or P(Relevance|query,document)
  - Estimate semantic consistency between query and document

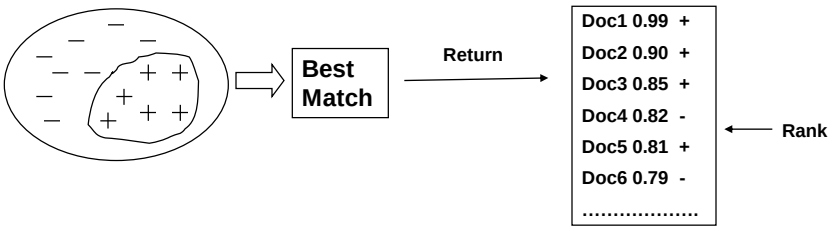
Types of Retrieval Models

- 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)

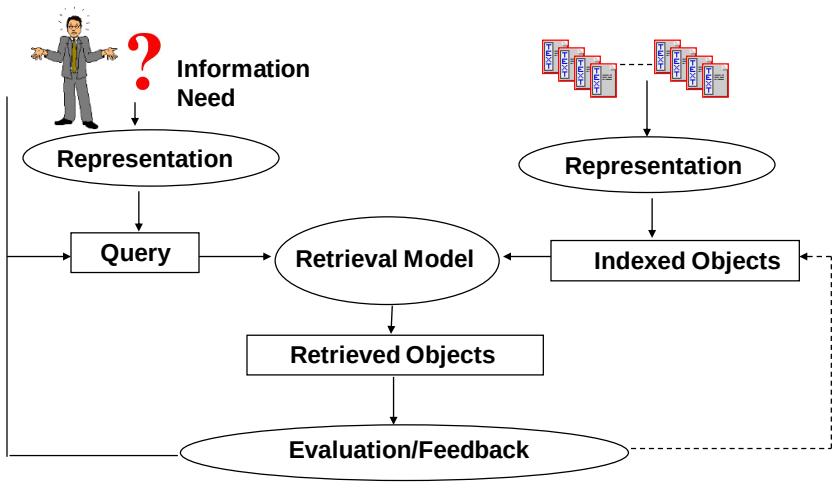
- Best Match is usually more accurate/effective
  - Do not need precise query; representative query generates good results
  - Users have control to explore the rank list: view more if need every piece; view less if need one or two most relevant
- Exact Match
  - Hard to define the precise query; too 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: Example Retrieval Models

Retrieval Models

- Boolean
- Vector space
  - Basic vector space — SMART, LUCENE
  - Extended Boolean
- Probabilistic models
  - Statistical language models — Lemur
  - Two Possion model — Okapi
  - Bayesian inference networks — Inquery
- Citation/Link analysis models
  - Page rank — Google
  - Hub & authorities — Clever

AD-hoc IR: Basic Process



Text Representation: What you see

It never leaves my side, April 6, 2002

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

Text Representation: What computer sees

<table><tr><td valign="top">
Reviewer:</td>

<td>"dage456" (Carmichael, CA USA) -

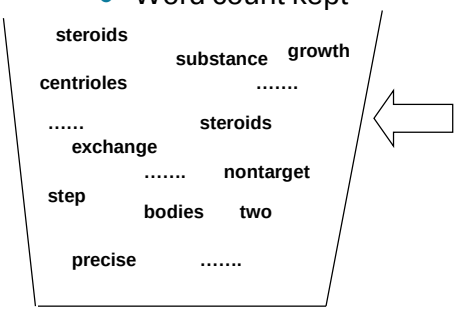
See all my reviews</td></tr></table>It fits in the palm of your hand and is the size of a deflated wallet (wonder where the money went). <p>I have had my ipod now for 4 months and cannot imagine how I used to get by with my old rio 600 with is 64 megs of ram and.. usb connection. Because of its size this little machine goes with my everywhere and its ten hour battery life means I can listen to stuff all day long.<p>Pros: size, both physical and capacity.
design: It looks beautiful
controls: simple and very easy to use<p>connection: FIREWIRE!!<p>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

Bag of Words (aka Vector Space Model)

The simplest text representation: "bag of words"

- Query/document: a bag that contains words
- Order among words is ignored
  - Word count kept



3 steroids	1 cilia-bearing	1 precise	1 two
2 centrioles	1 different	1 receptor	1 unexpected
1 affect	1 exchange	1 regularly	1 vitally
1 already	1 exogenous	1 reveal	1 way
1 Although	1 fluorescent	1 Specific	
1 antibodies	1 growth	1 step	
1 basal	1 identity	1 substance	
1 bodies	1 level	1 suggests	
1 cell	1 localization	1 target	
1 cells	1 nontarget	1 technique	

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

Text Representation: Indexing

Controlled Vocabulary

Mutation of a mutL homolog in hereditary colon cancer.

Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD, et al.

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.

Text Representation: Indexing

Controlled Vocabulary

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] +

Neoplasms by Site
Digestive System Neoplasms
Gastrointestinal Neoplasms
Intestinal Neoplasms
Colorectal Neoplasms
Colorectal Neoplasms, Hereditary Nonpolyposis

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

Text Representation: Indexing

Controlled Vocabulary

```
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

Text Representation: Indexing

Full Text Indexing

Full text Indexing: index all text with uncontrolled vocabulary

- Advantages
  - (Possibly) Keep all the information within the text
  - Often no human efforts; easy to build
- Disadvantages
  - Difficult to cross vocabulary gap (e.g., “cancer” in query, “neoplasm” in document)
  - Large storage space

How to build full text Indexing:

- What are the candidates in the word vocabulary? Are they effective to represent semantic meanings
- How to bridge small vocabulary gap (e.g., car and cars)

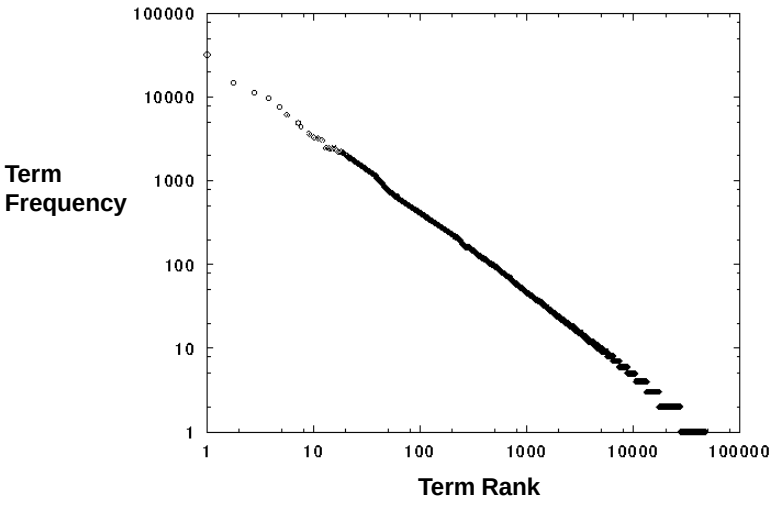
Text Representation: Indexing Statistical Properties of Text

Word	Frequency	Word	Frequency
the	1130021	market	52110
of	547311	bank	47940
to	516636	stock	47401
a	464736	trade	47310
in	390819	…	…
and	387703	…	…
….	…	…	…

Statistics collected from Wall Street Journal (WSJ), 1987

Text Representation: Indexing

Statistical Properties of Text



Text Representation: Text Preprocessing

Text Preprocessing: extract representative index terms

- Parse query/document for useful structure
  - E.g., title, anchor text, link, tag in xml…..
- Tokenization
  - For most western languages, words separated by spaces; deal with punctuation, capitalization, hyphenation
  - For Chinese, Japanese: more complex word segmentation…
- Remove stopwords: (remove "the", "is",..., existing standard list)
- Morphological analysis (e.g., stemming):
  - Stemming: determine stem form of given inflected forms
- Other: extract phrases; decompounding for some European languages "*rörelseuppskattningssökningsintervallsinställningar*"

Text Representation: Text Preprocessing

4 the	1 at	1 different	1 may	1 step
3 and	1 basal	1 exchange	1 nontarget	1 substance
3 by	1 be	1 exogenous	1 not	1 suggests
3 steroids	1 been	1 fluorescent	1 may	1 target
2 centrioles	1 bodies	1 from	1 of	1 technique
2 in	1 can	1 growth	1 precise	1 two
1 affect	1 at	1 has	1 receptor	1 unexpected
1 already	1 cell	1 identity	1 regularly	1 vitally
1 Although	1 cells	1 level	1 reveal	1 way
1 antibodies	1 cilia-bearing	1 localization	1 Specific	1 with

24 stopwords out of total 61 words

Text Representation: Phrases

- Single word/stem indexing may not be sufficient
  e.g., “hit a home run yesterday”
- More complicated indexing includes phrases (thesaurus classes)
- How to automatically identify phrases
  - Dictionary
  - Find the most common N word phrases by corpus statistics (be careful of stopwords)
  - Syntactic analysis, noun phrases
  - More sophisticated segmentation algorithm like “Hidden Markov Model”

Text Representation: Word Stemming

Word Stemming

- Associate morphological variants of words into a single form
  - E.g., plurals, adverbs, inflected word forms
  - May lose the precise meaning of a word
- Different types of stemming algorithms
  - Rule-based systems: Porter Stemmer, Krovetz Stemmer
    Porter Stemmer Example: describe/describes -> describ
  - Statistical method: Corpus-based stemming

Text Representation: Word Stemming

Porter Stemmer

- Based on a pattern of vowel-consonant sequence
  - [C](VC)^m[V], m is an integer
- Rules are divided into steps and examined in sequence
  - Step 1a: ies → i; s →; …..
    cares → care
  - Step 1b: if m>0 eed ee
    agreed → agree
  - ….. Step 5a, Step 5b
- Pretty aggressive:
  - nativity → native

Text Representation: Word Stemming

Examples of Stemming:

- Original Text:

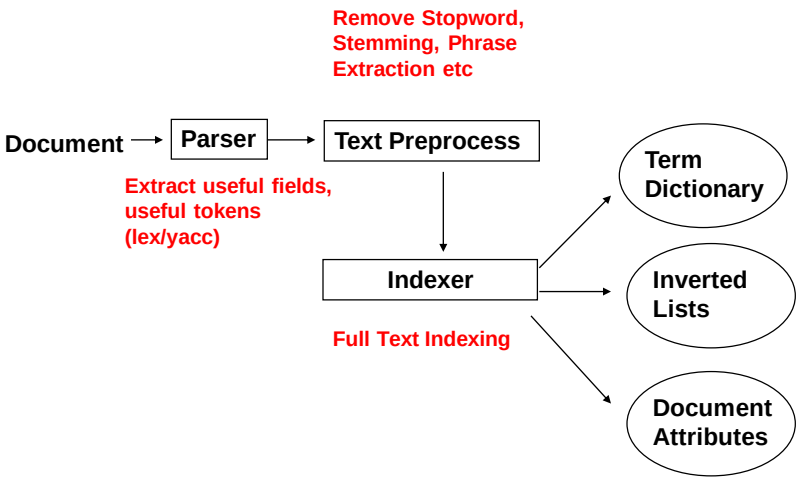
  Information retrieval deals with the representation, storage, organization of, and access to information items

- Porter Stemmer (Stopwords removed):

  Online example:
  http://facweb.cs.depaul.edu/mobasher/classes/csc575/porter.html

  Inform retrieve deal represent storag organ access inform item

Text Representation: Process of Indexing



Text Representation: Inverted Lists

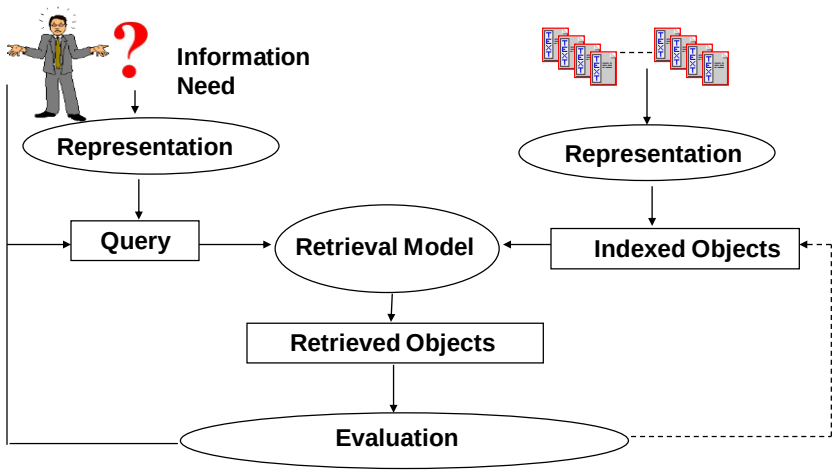
Documents

Doc ID	Text
1	kids question noting in 1960s
2	young man question everything in 1970s
3	kids question questions in 1980s
4	young man question nothing in 2000s

Inverted Lists

Term ID	Term	Documents
1	kids	1,3
2	question	1,2,3,4
3	nothing	1,4
4	in	1,2,3,4
5	19060s	1
6	young	2,4
7	man	2,4
8	everything	2
9	1970s	2
10	questions	3
11	1980s	3
11	2000s	4

AD-hoc IR: Basic Process



Evaluation

Evaluation criteria

- Effectiveness
  - How to define effectiveness? Where can we find the correct answers?
- Efficiency
  - What about retrieval speed? What about the storage space? Particularly important for large-scale real-world system
- Usability
  - What is the most important factor for real user? Is user interface important?

Evaluation Criteria

- Effectiveness
  - Favor returned document ranked lists with more relevant documents at the top
  - Objective measures
    - Recall and Precision
    - Mean-average precision
    - Rank based precision

For documents in a subset of a ranked lists, if we know the truth

	Retrieved	Not retrieved
Relevant	Relevant docs retrieved	Relevant docs not retrieved
Irrelevant	Irrelevant docs retrieved	Irrelevant docs not retrieved

$$\text{Precision} = \frac{\text{Relevant docs retrieved}}{\text{Retrieved docs}}$$

$$\text{Recall} = \frac{\text{Relevant docs retrieved}}{\text{Relevant docs}}$$

Evaluation

Evaluate a ranked list

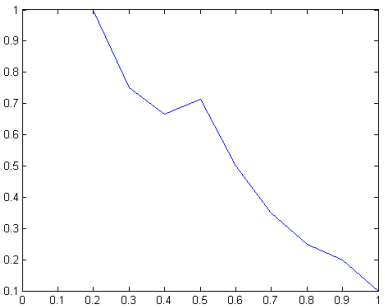
Precision at Recall

- Evaluate at every relevant document

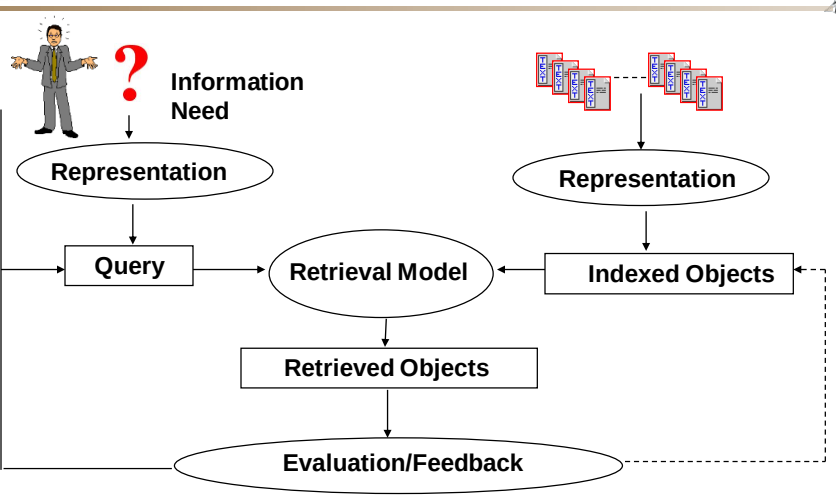
+
+
-
+
+
-
+

Not Retrieved: +++++

Precision	Recall
1	0.1
1	0.2
0.667	0.2
0.75	0.3
0.8	0.4
0.667	0.4
0.714	0.5



Retrieval Models



Retrieval Models: Unranked Boolean

- Unranked Boolean: Exact match method
- Selection Model
  - Retrieve a document iff it matches the precise query
  - Often return unranked documents (or with chronological order)
- Operators
  - Logical Operators: AND OR, NOT
  - Proximity operators: #1(white house) (i.e., within one word distance, phrase) #sen(Iraq weapon) (i.e., within a sentence)
  - String matching operators: Wildcard (e.g., ind* for india and indonesia)
  - Field operators: title(information and retrieval)…

Retrieval Models: Unranked Boolean

Unranked Boolean: Exact match method

- A query example
  (#2(distributed information retrieval) OR
  (#1 (federated search)) AND
  author(#1(Jamie Callan) AND NOT (Steve))

Retrieval Models: Unranked Boolean

WestLaw system: Commercial Legal/Health/Finance Information Retrieval System

- Logical operators
- Proximity operators: Phrase, word proximity, same sentence/paragraph
- String matching operator: wildcard (e.g., ind*)
- Field operator: title(#1(“legal retrieval”)) date(2000)
- Citations: Cite (Salton)

Retrieval Models: Unranked Boolean

Advantages:

- Work well if user knows exactly what to retrieve
- Predicable; easy to explain
- Very efficient

Disadvantages:

- It is difficult to design the query; high recall and low precision for loose query; low recall and high precision for strict query
- Results are unordered; hard to find useful ones
- Users may be too optimistic for strict queries. A few very relevant but a lot more are missing

Retrieval Models: Ranked Boolean

Ranked Boolean: Exact match

- Similar as unranked Boolean but documents are ordered by some criterion

Retrieve docs from Wall Street Journal Collection

Query: (Thailand AND stock AND market)

Reflect importance of document by its words

Which word is more important?

Many "stock" and "market", but fewer "Thailand". Fewer may be more indicative

Term Frequency (TF): Number of occurrence in query/doc; larger number means more important

Inversed Document Frequency (IDF): Larger means more important

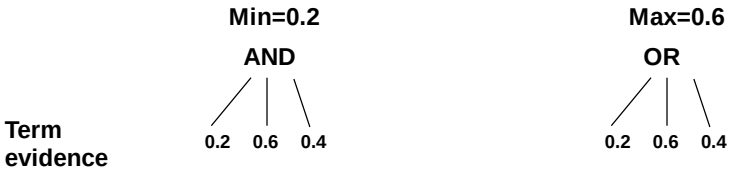
$$\frac{\text{Total number of docs}}{\text{Number of docs contain a term}}$$

There are many variants of TF, IDF: e.g., consider document length

Retrieval Models: Ranked Boolean

Ranked Boolean: Calculate doc score

- Term evidence: Evidence from term i occurred in doc j: (tfij) and (tfij*idfi)
- AND weight: minimum of argument weights
- OR weight: maximum of argument weights



Query: (Thailand AND stock AND market)

Retrieval Models: Ranked Boolean

Advantages:

- All advantages from unranked Boolean algorithm
  - Works well when query is precise; predictive; efficient
- Results in a ranked list (not a full list); easier to browse and find the most relevant ones than Boolean
- Rank criterion is flexible: e.g., different variants of term evidence

Disadvantages:

- Still an exact match (document selection) model: inverse correlation for recall and precision of strict and loose queries
- Predictability makes user overestimate retrieval quality

Retrieval Models: Vector Space Model

- Vector space model
- Any text object can be represented by a term vector
  - Documents, queries, passages, sentences
  - A query can be seen as a short document
- Similarity is determined by distance in the vector space
  - Example: cosine of the angle between two vectors
- The SMART system
  - Developed at Cornell University: 1960-1999
  - Still quite popular
- The Lucene system
  - Open source information retrieval library; (Based on Java)
  - Work with Hadoop (Map/Reduce) in large scale app (e.g., Amazon Book)

Retrieval Models: Vector Space Model

Vector space model vs. Boolean model

- Boolean models
  - Query: a Boolean expression that a document must satisfy
  - Retrieval: Deductive inference
- Vector space model
  - Query: viewed as a short document in a vector space
  - Retrieval: Find similar vectors/objects

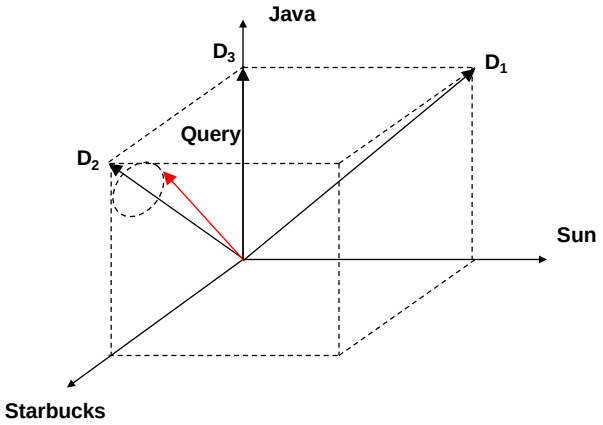
Retrieval Models: Vector Space Model

- Vector representation

	Java	Sun	Starbucks
D1	1	1	0
D2	1	0	1
D3	1	0	0
Query	1	0.2	1

Retrieval Models: Vector Space Model

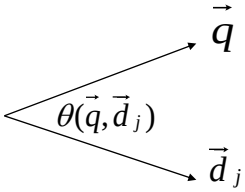
Vector representation



Retrieval Models: Vector Space Model

Give two vectors of query and document

- query as $\vec{q} = (q_1, q_2, ..., q_n)$
- document as $\vec{d}_j = (d_{j1}, d_{j2}, ..., d_{jn})$
- calculate the similarity



Cosine similarity: Angle between vectors

$$sim(\vec{q}, \vec{d}_j) = \cos(\theta(\vec{q}, \vec{d}_j))$$

$$\cos(\theta(\vec{q}, \vec{d}_j)) = \frac{\vec{q} \cdot \vec{d}_j}{\|\vec{q}\|\|\vec{d}\|} = \frac{q_1 d_{j,1} + q_2 d_{j,2} + ... + q_j d_{j,n}}{\|\vec{q}\|\|\vec{d}\|} = \frac{q_1 d_{j,1} + q_2 d_{j,2} + ... + q_j d_{j,n}}{\sqrt{q_1^2 + ... + q_n^2}\sqrt{d_{j1}^2 + ... + d_{jn}^2}}$$

Retrieval Models: Vector Space Model

Vector representation

	Java	Sun	Starbucks
D1	1	1	0
D2	1	0	1
D3	1	0	0
Query	1	0.2	1

Similarity Score	D1	D2	D3
Query	0.59	0.99	0.70

IR review

- 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

Content Based Filtering Filtering

Information Needs are Stable

System should make a delivery decision on the fly when a document "arrives"