PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

Prof. Chris Clifton

24 August 2020

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

PURDUE UNIVERSITY
Department of Computer Science

Web

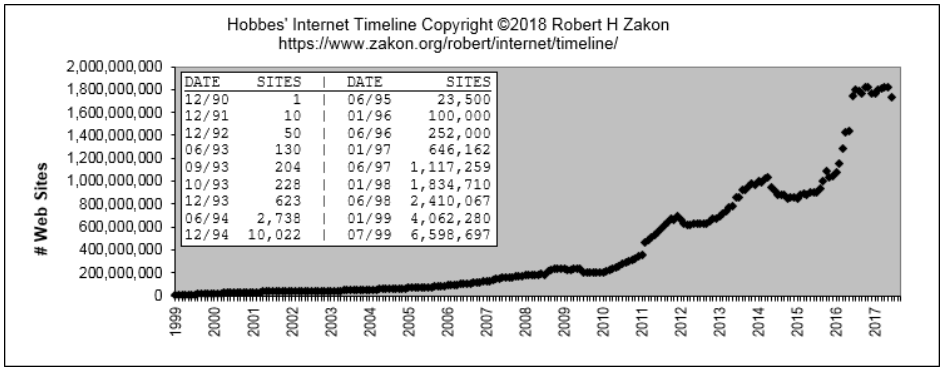
- Web opened the door for many important applications
- Information Retrieval
  - Web Search
  - Information Recommendation by content or by collaborative information
- Web Services
- Semantic Web
- Web 2.0
- XML
- Social Network
- ……..

PURDUE UNIVERSITY
Department of Computer Science

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)



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

PURDUE UNIVERSITY
Department of Computer Science

Why this course?

- Managing Data is one of the primary uses of computers
- Most of this data is NOT contained in structured databases
  - *IDC estimates that by 2025, 80 percent of data will exist as unstructured data - commonly appearing in e-mails, memos, notes from call centers and support operations, news, user groups, chats, reports, … and Web pages.*
  - Text in Web pages or emails; image; audio; video; protein sequences..
  - Therefore, no carefully structured queries
- How do we find this?

Information Retrieval

5

PURDUE UNIVERSITY
Department of Computer Science

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

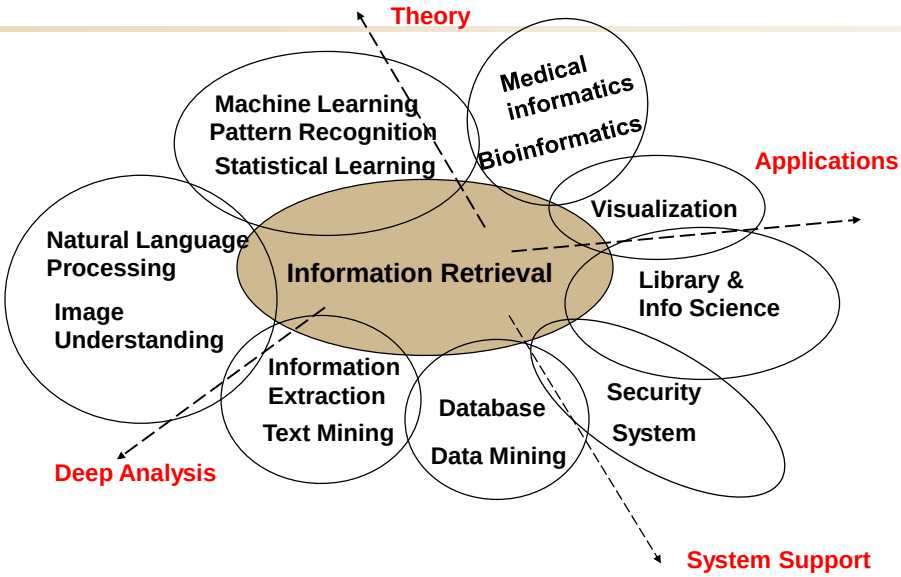
PURDUE UNIVERSITY
Department of Computer Science

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

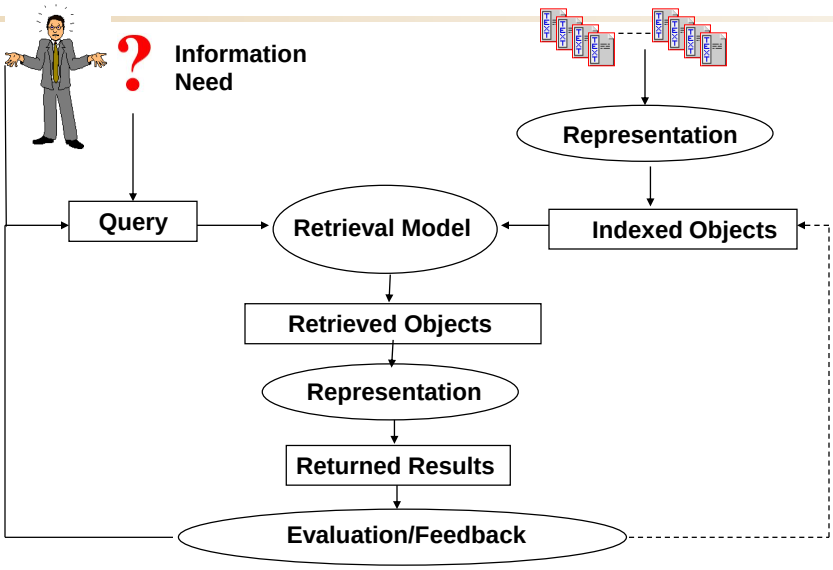
PURDUE UNIVERSITY
Department of Computer Science

IR and other disciplines



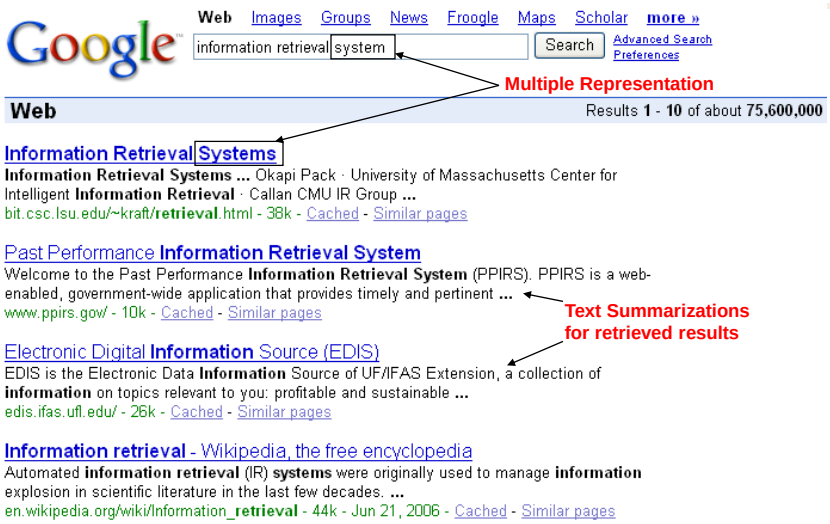
PURDUE UNIVERSITY
Department of Computer Science

Some core concepts of IR



PURDUE UNIVERSITY
Department of Computer Science

Some core concepts of IR



PURDUE UNIVERSITY
Department of Computer Science

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)

PURDUE UNIVERSITY
Department of Computer Science

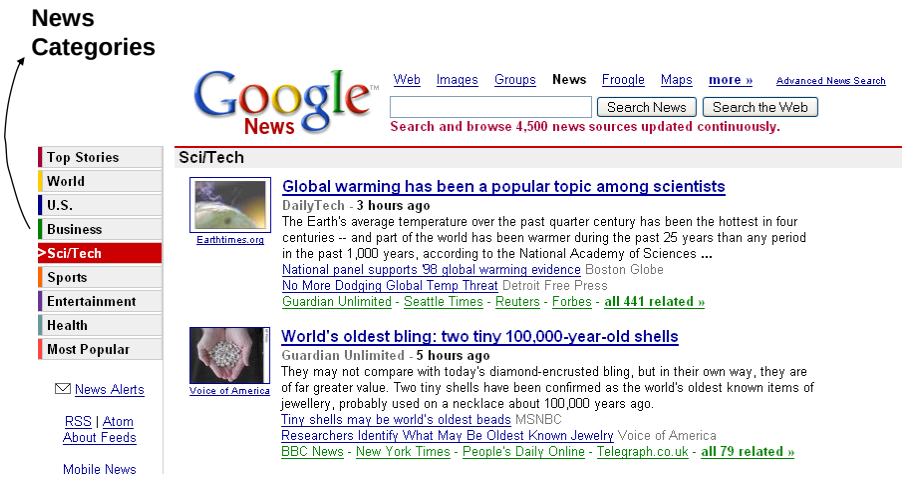
IR Applications

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
- ……..

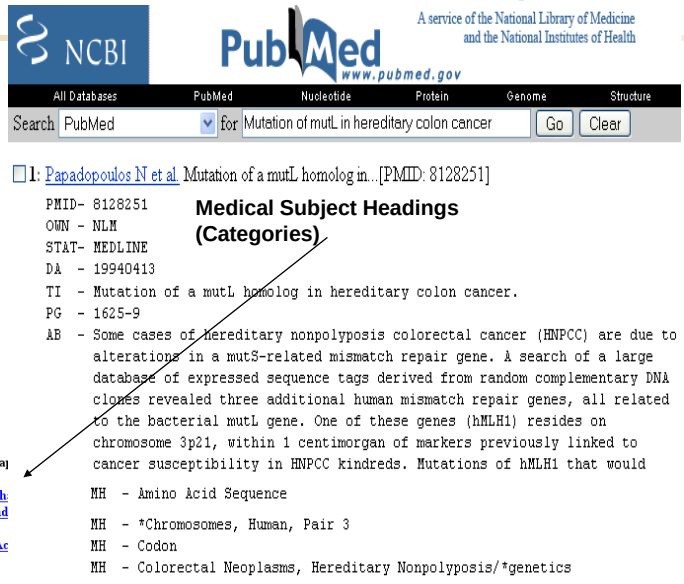
PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Text Categorization



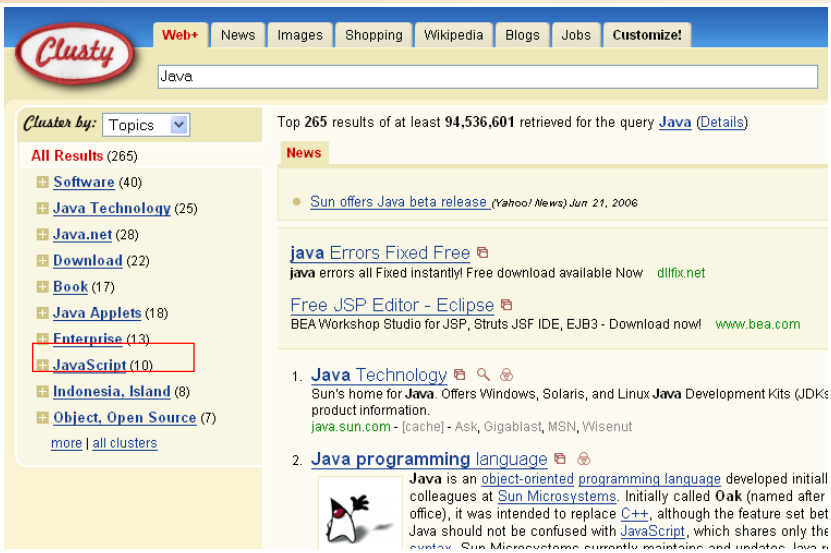
PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Text Categorization



PURDUE UNIVERSITY
Department of Computer Science

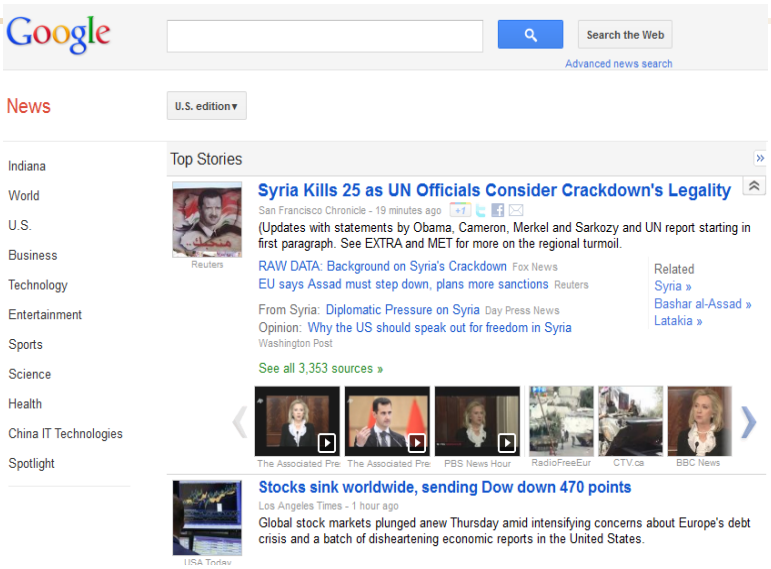
IR Applications: Document Clustering



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Content Based Filtering

Keyword Matching



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Collaborative Filtering

Other Customers with similar tastes

PURDUE UNIVERSITY
Department of Computer Science

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

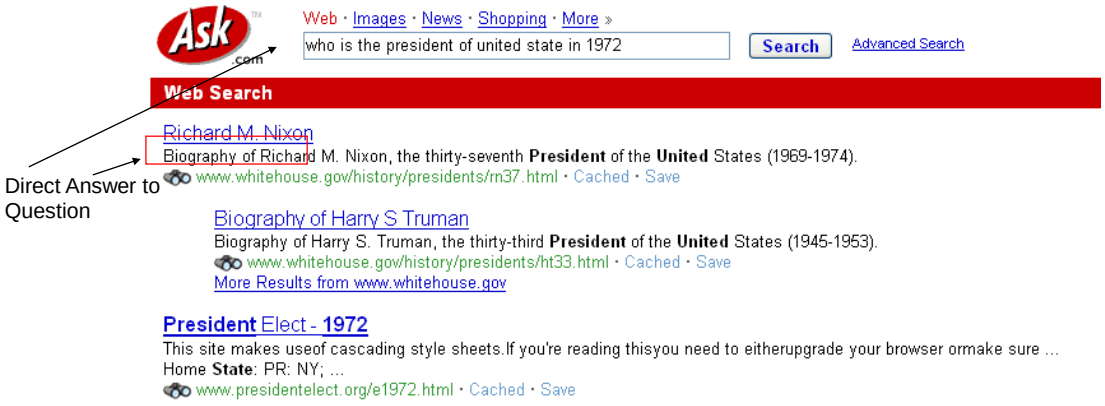
- Inference based on the relationship between entities

Inherited Disease → **Gene** → **Chemical**

Drug discovery

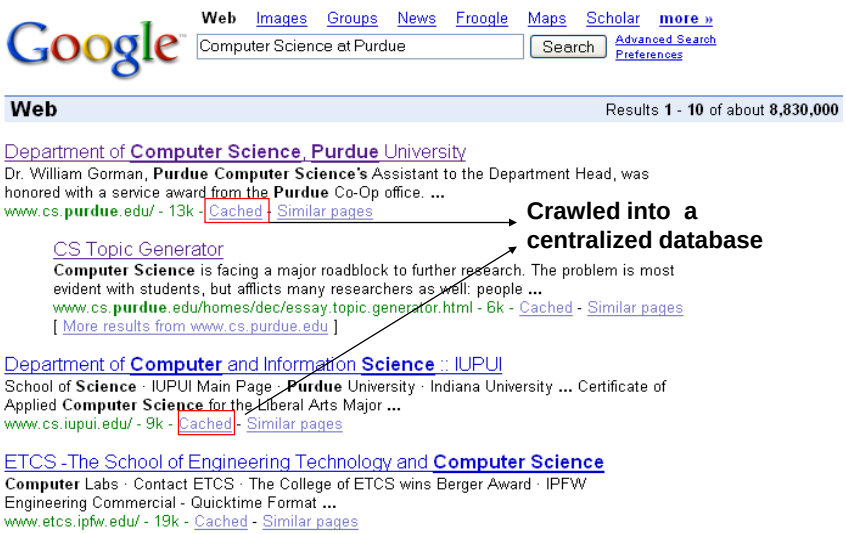
PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Question Answering



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Web Search



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Federated Search

Valuable → Searched by *Federated Search*

PURDUE UNIVERSITY
Department of Computer Science

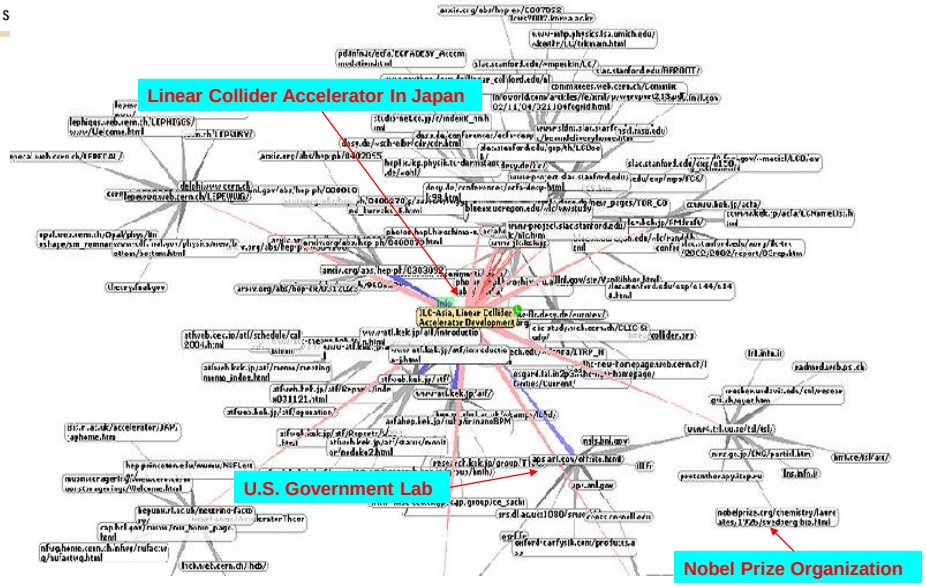
IR Applications: Expertise Search

INDURE: Indiana database of university research expertise

www.indure.org

PURDUE UNIVERSITY
Department of Computer S

IR Applications: Citation/Link Analysis



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Citation/Link Analysis

Citation/Link importance

2457 citations found. **Only retrieving 1000 documents.**

Dempster, A.P., Laird, N.M., & Rubin, D. *Maximum-likelihood from incomplete data via the EM algorithm.* Journal of the Ro
39. di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992). Understanding motor events: A neurophysic
Research, 91, 176--180.

CiteSeer Home/Search Document Not in Database Summary Related Articles Check

This paper is cited in the following contexts:

First 50 documents Next 50

A Probabilistic Model of Gaze Imitation and Shared Attention - Matthew Hoffman David (Correct)

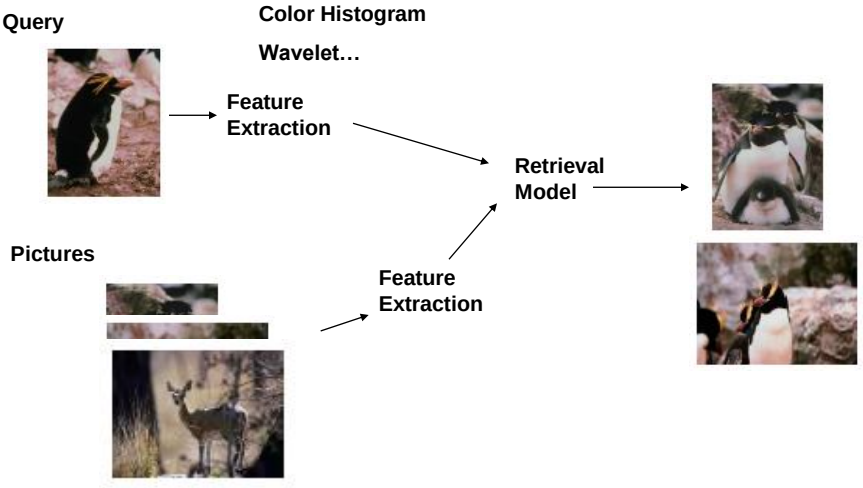
No context found.

Dempster, A.P., Laird, N.M., & Rubin, D. *Maximum-likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statisti
Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolati, G. (1992). Understanding motor events: A neurophysiological study. E
176--180.

Latent Variable Models for Semantic Orientations of Phrases - Hiroya Takamura Takamura (2006) (Correct)

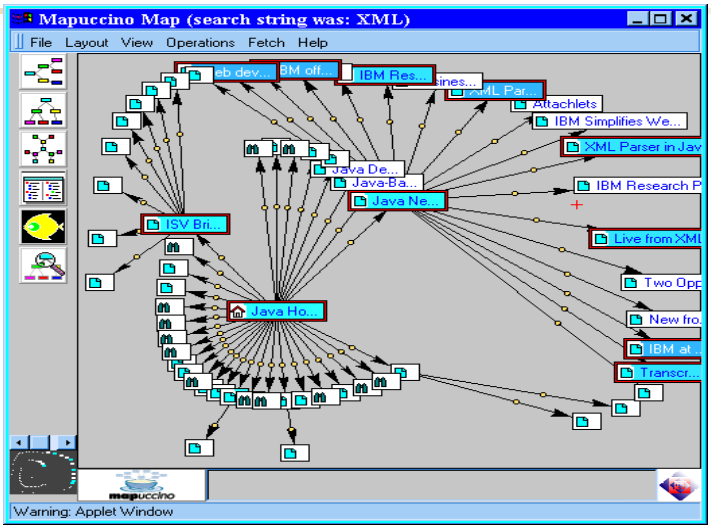
PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Multimedia Retrieval



PURDUE UNIVERSITY
Department of Computer Science

IR Applications: Information Visualization



Partial Structure of pages from a Web subset visualized by Mapuccino

PURDUE UNIVERSITY
Department of Computer Science

Course Goals

- Learn the techniques behind Web search engines, E-commerce recommendation systems, etc.
- Get hands on project experience by developing real-world applications, such as building a small-scale Web search engine, a Web page management system, or a movie recommendation system.
- Learn tools and techniques to do research in the area of information retrieval or text mining.
- Lead to the amazing job opportunities in Search Technology and E-commerce companies such as Google, Microsoft, Yahoo! and Amazon.

PURDUE UNIVERSITY
Department of Computer Science

Logistics

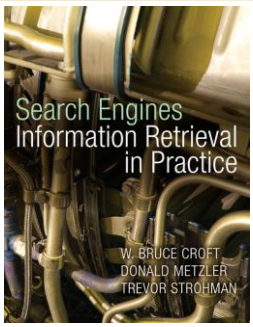
- Time and location: MWF 1:30-2:20, KRAN G016
  - Live remote attendance via WebEx (link found in Brightspace)
  - Lectures recorded and available in Mediaspace (link also found in Brightspace)
- Instructor: Chris Clifton, clifton@cs.purdue.edu
  LWSN 2116E, office hours: TBD
- Teaching assistants:
  - Xiaoni Duan, duan79@purdue.edu
  - Kendal G. Norman, norman17@purdue.edu
  - Md. Masudur Rahman, rahman64@purdue.edu
  - Kaiyuan Zhang, zhan4057@purdue.edu
- Office hours: times TBD, will be on WebEx or Zoom, some also in person
- Webpage: http://www.cs.purdue.edu/~clifton/cs47300/
- Email list: fall-2020-cs-47300-???@lists.purdue.edu
  - Goes to your @purdue.edu email, make sure you get this, as it contains critical announcements.
- Discussion forum: Piazza? Brightspace? Others?
- Prerequisites: CS25100. Having had a Stat course (e.g., STAT 35000) will help.

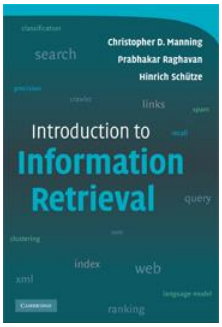
PURDUE UNIVERSITY
Department of Computer Science

Readings

- *Search Engines: Information Retrieval in Practice*, Croft, Metzler, and Strohman, Pearson, 2010. Available as a PDF: ciir.cs.umass.edu/downloads/SEIRiP.pdf

- Recommended:
  Introduction to Information Retrieval, Manning, Raghavan, and Schuze, Cambridge University Press, 2008.





PURDUE UNIVERSITY
Department of Computer Science

Workload

- Homeworks
  - 4-5 written assignments
  - 2-3 more substantial programming projects
  - Late policy: 15% off per day late, maximum of 5 days
  - Five extension days to be used at your discretion
    - No fractional days
    - May not be used to extend submission past last day of class.
    - Late penalties will not be applied until the end of the semester
      - Late days will be applied to your best advantage
- Exams
  - Midterms (2) and final exam

PURDUE UNIVERSITY | Department of Computer Science

CS47300: Web Information Search and Management

Prof. Chris Clifton

26 August 2020

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

Indiana Center for Database Systems™

PURDUE UNIVERSITY
Department of Computer Science

Review of Monday:

- Core concepts of information retrieval
  - Query representation; document representation; retrieval model; evaluation
- Applications of information retrieval
  - Web Search; Text Categorization; Document Clustering; Information Recommendation; Information Extraction; Question Answering…..
- Grade Policy
  - Assignments: 40%; Midterms 24%, Final Exam: 26%; Class attendance: 10%

PURDUE UNIVERSITY
Department of Computer Science

Administrivia

- Brightspace is now available
  - Links to lecture recordings, Piazza, Gradescope~~, and iClicker registration~~
  - And not much else yet
- Piazza is active
  - https://piazza.com/purdue/fall2020/wl202110cs47300le1
  - Or you can access through Brightspace (you may need to access through Brightspace the first time)
- Office Hours for Prof. Clifton:
  - Thursday 8:30-10, WebEx (see course web page for link)
  - In person LWSN 2116E by appointment – send a few times that work for you
- TA office hours TBD, will initially be WebEx and Zoom
  - See course web page for links
  - Department is still working on office assignments, for now all will be online

44

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Ad-hoc IR: Terminologies

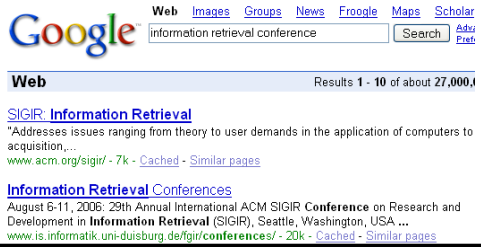
Terminologies:

- 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)

PURDUE UNIVERSITY
Department of Computer Science

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



PURDUE UNIVERSITY
Department of Computer Science

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)

Relatively Stable — **Changes**

  - 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

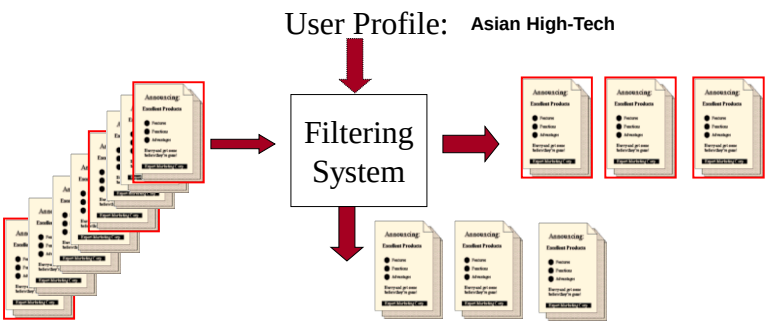
- Filtering: Queries are stable (e.g., Asian High-Tech) while the collection changes (e.g., news)
- More for filtering in later lectures

PURDUE UNIVERSITY
Department of Computer Science

Content Based Filtering

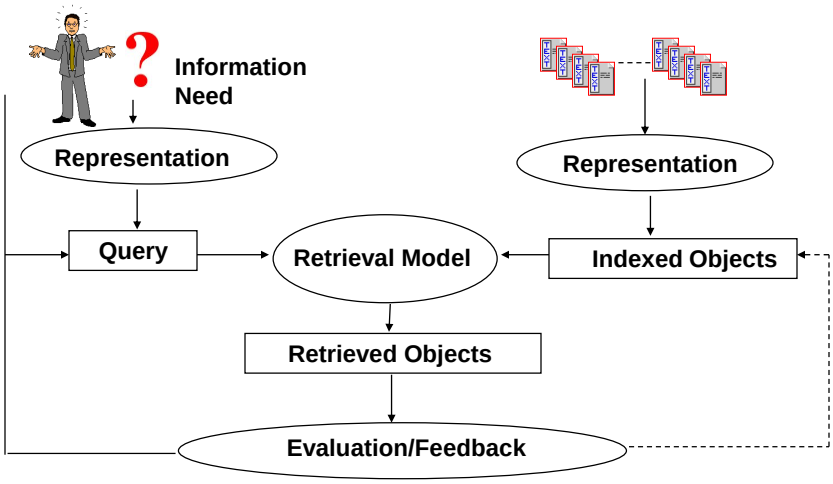
Information Needs are Stable

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



PURDUE UNIVERSITY
Department of Computer Science

AD-hoc IR: Basic Process



PURDUE UNIVERSITY
Department of Computer Science

AD-hoc IR: Overview of Retrieval Models

Retrieval Models

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

PURDUE UNIVERSITY
Department of Computer Science

AD-hoc IR: Overview of Retrieval Models

Purpose of the Retrieval Model

Determine whether a document is relevant to query

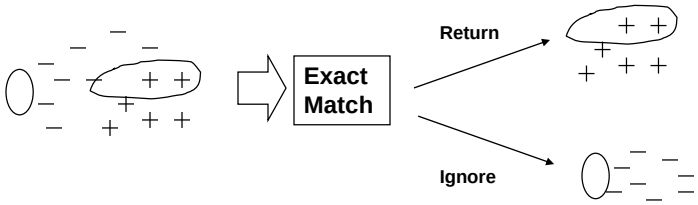
- Relevance is difficult to define
  - Varies by judgers
  - Varies by context (i.e., jointly by a set of documents and queries)
- Different retrieval methods estimate relevance differently
  - Word occurrence of document and query
  - In probabilistic framework, P(query|document) or P(Relevant|query,document)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

52

PURDUE UNIVERSITY
Department of Computer Science

Types of Retrieval Models

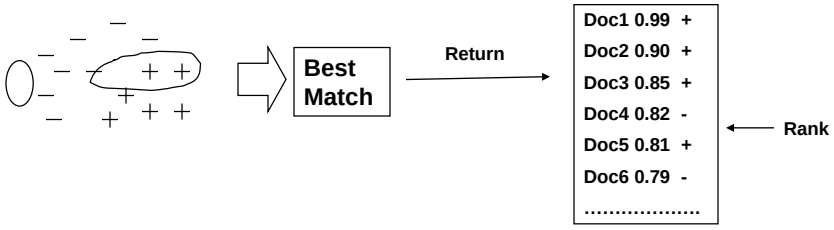
- Exact Match (Document Selection)
  - Example: Boolean Retrieval Method
  - Query defines the exact retrieval criterion
  - Relevance is a binary variable; a document is either relevant (i.e., match query) or irrelevant (i.e., mismatch)
  - Result is a set of documents
    - Documents are unordered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



PURDUE UNIVERSITY
Department of Computer Science

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)



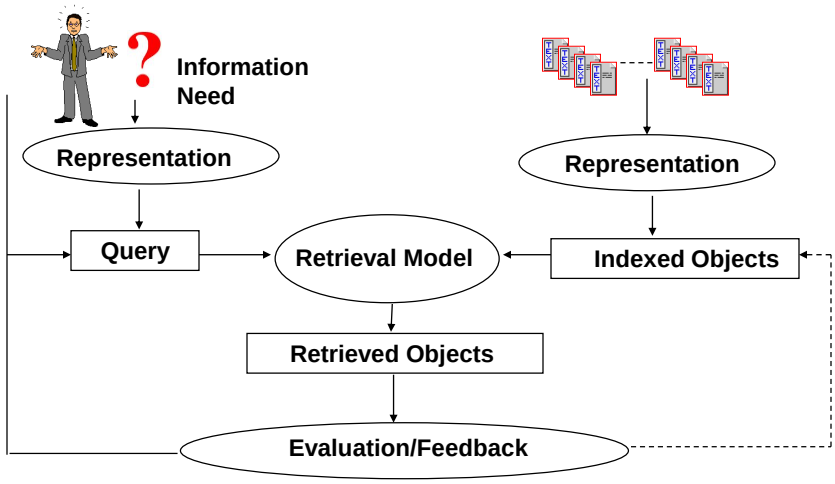
PURDUE UNIVERSITY
Department of Computer Science

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)

PURDUE UNIVERSITY
Department of Computer Science

AD-hoc IR: Basic Process



PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: What computer sees

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

<td><a href="http://www.amazon.com/exec/obidos/tg/cm/member-glance/-
/AJF9GJKJ8UGNX/1/ref=cm_cr_auth/002-1193904-0468830?%5Fencoding=UTF8"><span
style =" font-weight: bold;">"dage456"</span></a> (Carmichael, CA USA) - <a
href="http://www.amazon.com/gp/cdp/member-
reviews/AJF9GJKJ8UGNX/ref=cm_cr_auth/002-1193904-0468830?ie=UTF8“>

See all my reviews</a></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.<br>design: It looks beautiful<br>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.<br /><br />
```

From Amazon Customer Review of IPod

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: TREC Format

```
<DOC>
<DOCNO> AP900101-0001 </DOCNO>
<FILEID>AP-NR-01-01-90 2345EDT</FILEID>
<FIRST>r i PM-Iran-Population Bjt   01-01 0777</FIRST>
<SECOND>PM-Iran-Population, Bjt,0800</SECOND>
<HEAD>Iran Moves To Curb A Baby Boom That Threatens Its Economic
Future</HEAD>
<HEAD>An AP Extra</HEAD>
<BYLINE>By ED BLANCHE</BYLINE>
<BYLINE>Associated Press Writer</BYLINE>
<DATELINE>NICOSIA, Cyprus (AP) </DATELINE>
<TEXT>
  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.
  ……
</TEXT>
</DOC>
```

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Indexing

- Indexing
  - Associate document/query with a set of keys
- Manual or human Indexing
  - Indexers assign keywords or key concepts (e.g., libraries, Medline, Yahoo!); often small vocabulary
  - Significant human efforts, may not be thorough
- Automatic Indexing
  - Index program assigns words, phrases or other features; often large vocabulary
  - No human effort ➔ low cost

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

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.

PURDUE UNIVERSITY
Department of Computer Science

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]

PURDUE UNIVERSITY
Department of Computer Science

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
```

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

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 Index:

- 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)

PURDUE UNIVERSITY
Department of Computer Science

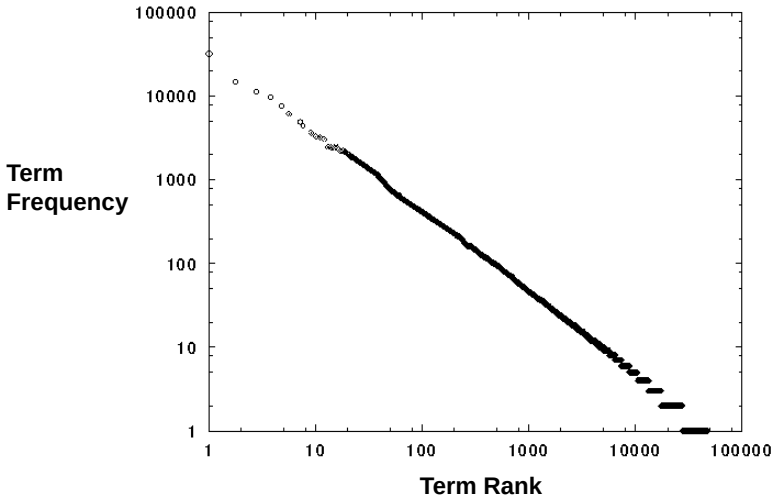
Text Representation: Indexing
Statistical Properties of Text

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

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Indexing
Statistical Properties of Text



PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Indexing
Statistical Properties of Text

- Observations from language/corpus independent features
- A few words occur very frequently (High Peak)
  - Top 2 words: 8%-15% (e.g., words that carry no semantic meanings like “the”, “to”)
- Most words occur rarely (Heavy Tail)
- Representative words often in the middle
  - e.g., market and stock for WSJ
- Rules formally describe word occurrence patterns: Zipf’s law, Heaps’ Law

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Indexing
Statistical Properties of Text

Zipf’s law: relate a term’s frequency to its rank

- Rank all terms with their frequencies in descending order, for a term at a specific rank (e.g., *r*) collects and calculates

f_r : **term frequency**

$p_r = \frac{f_r}{N}$: **relative term frequency**

Total number of words

- Zipf’s law (by observation):

$$p_r = A/r \quad A \approx 0.1$$

So $p_r = \frac{f_r}{N} = \frac{A}{r} \Rightarrow rf_r = AN \Rightarrow \log(r) = -\log(f_r) + \log(AN)$

So Rank X Frequency = Constant

PURDUE UNIVERSITY
Department of Computer Science

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*

PURDUE UNIVERSITY
Department of Computer Science

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

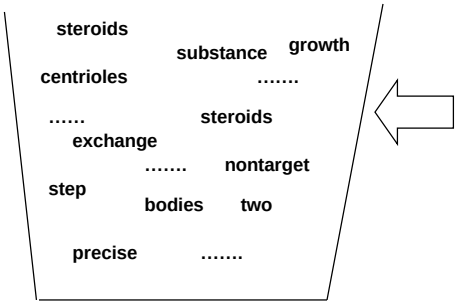
24 stopwords out of total 61 words

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Bag of Words

The simplest text representation: “bag of words”

- Query/document: a bag that contains words in it
- Order among words is ignored



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	

PURDUE UNIVERSITY
Department of Computer Science

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”

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

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

77

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Word Stemming

K Stemmer: based on morphological rules

- If word occurs in a dictionary, do not stem it

- For all other words
  - Remove inflectional endings: plurals to singular; past tense to present tense; remove “ing”
  - Remove derivational endings by a sequence of rules: may make mistake when suffixes indicate different meanings like “sign” to “signify”

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Word Stemming

Problems with Rule-based Stemming

- Rule-based stemming may be too aggressive
  - e.g., execute/executive, university/universe
- Rule-based stemming may be too conservative
  - e.g., European/Europe, matrices/matrix
- Difficult to understand the meaning the stems
  - e.g., Iteration/iter, general/gener

PURDUE UNIVERSITY
Department of Computer Science

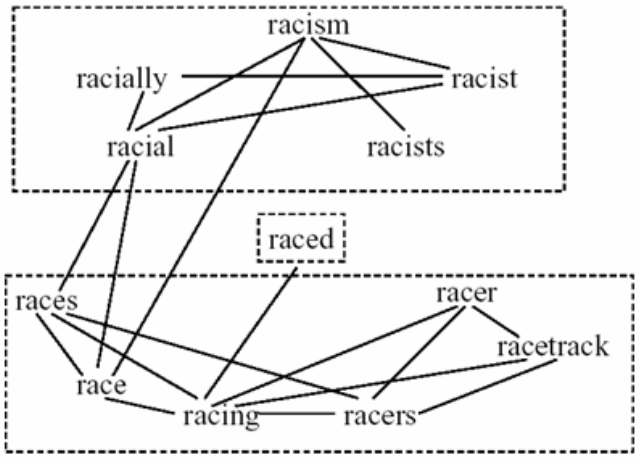
Text Representation: Word Stemming

Corpus-Based Stemming

- Hypothesis: Word variants that should be considered equally often co-occur in documents (passages or text windows) in the corpus
  - Collect the statistics of co-occurrence of words in the corpus and form the connected graph
  - Cut the graph by different methods and find the connected subgraphs to form equivalence classes

PURDUE UNIVERSITY
Department of Computer Science

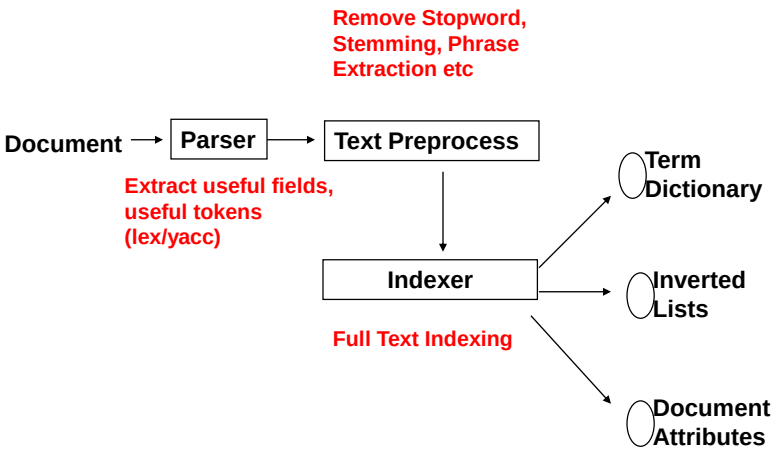
Text Representation: Word Stemming



(Xu & Croft, 1998)

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Process of Indexing



PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Inverted Lists

Inverted lists are one of the most common indexing techniques

- Source file: collection organized by documents
- Inverted list file: collection organized by term
  - one record per term, the lists of documents that contain the specific term
- Possible actions with inverted lists
  - OR: the union of lists
  - And: the intersection of lists

PURDUE UNIVERSITY
Department of Computer Science

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

PURDUE UNIVERSITY
Department of Computer Science

Text Representation: Inverted Lists

Many engineering details

- Update inverted lists: delete/insert a term or document
- Compression: trade off between I/O time and CPU time
- Add more information such as position information
- ……