# CS630 Representing and Accessing Digital Information

## Information Retrieval: Basics

Thorsten Joachims
Cornell University

Based on slides from Prof. Jamie Callan and Prof. Claire Cardie

# Information Retrieval

- **Basics**
- **Retrieval Models**
- **Indexing and Preprocessing**
- **Data Structures**

~ 4 lectures

# IR Basics

- **Task definition**
- **Evaluation**
- **Statistical properties of text**

The field of *information retrieval* deals with the

representation,

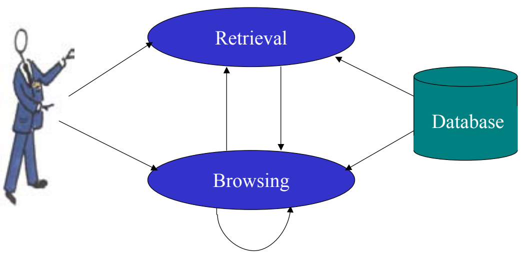
storage,

organization of,

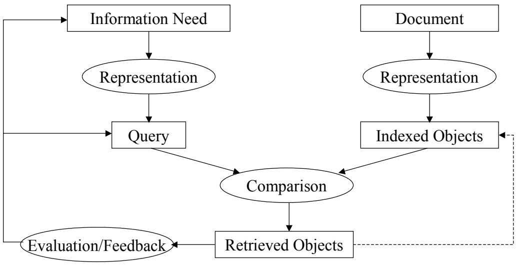
access to

information items.

# User Task

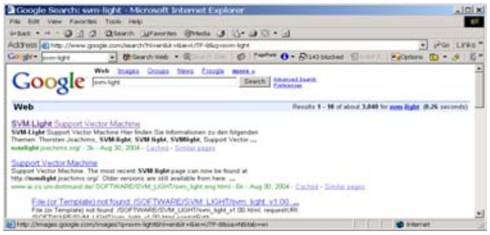


# Basic IR Processes



# Task Definition: Ad-hoc Retrieval

- **Search a large collection of documents to find the ones that satisfy an information need**
  - I.e., find relevant documents
- **Sometimes called “archival” retrieval**
- **Example: Web search engines**


Google Search: svm-light - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address
Go Links
Search Web
143 blocked
Options
Google
Web Images Groups News Froogle more »
Search
Advanced Search
Preferences
Web
Internet

http://www.google.com/search?hl=en&lr=&ie=UTF-8&q=svm-light

svm-light

Results 1 - 10 of about 3,840 for svm light. (0.26 seconds)

SVM-Light Support Vector Machine
SVM-Light Support Vector Machine Hier finden Sie Informationen zu den folgenden Themen: Thorsten Joachims, SVM-light, SVM light, SVMlight, Support Vector ...
svmlight.joachims.org/ - 3k - Aug 30, 2004 - Cached - Similar pages

Support Vector Machine
Support Vector Machine. The most recent SVM light page can now be found at http://svmlight.joachims.org/. Older versions are still available from here. ...
www-ai.cs.uni-dortmund.de/ SOFTWARE/SVM_LIGHT/svm_light.eng.html - 6k - Aug 30, 2004 - Cached - Similar pages

File (or Template) not found /SOFTWARE/SVM_LIGHT/svm_light_v1.00 ...
File (or Template) not found /SOFTWARE/SVM_LIGHT/svm_light_v1.00.html. requestURI

http://images.google.com/images?q=svm-light&hl=en&lr=&ie=UTF-8&sa=N&tab=wi

## Settings for Ad-hoc Retrieval

- **Unranked ad-hoc retrieval**
  - Return an unordered set of documents that satisfies the query
  - Usually used on in Boolean retrieval systems (which you'll hear about soon enough)
  - Disadvantages:
    - Important to create a good query, so that the retrieved set is small
    - Small set may not have enough relevant documents
    - ???
  - Advantages???

## Settings for Ad-hoc Retrieval

- **Ranked ad-hoc retrieval**
  - Return a set of documents that satisfies the query ordered by (presumed) relevance
  - Advantages
    - Large retrieved sets are not a problem
    - Less time spent crafting queries *and* reading documents
  - Disadvantages
    - Good queries are still important
    - ???

## Settings for Ad-hoc Retrieval

- **Cross-lingual retrieval (CLIR)**
  - Query in one language (e.g. English)
  - Return documents in other languages (e.g. Korean, Greek, Tamil)
  - Sometimes called "translingual" retrieval

## Settings for Ad-hoc Retrieval

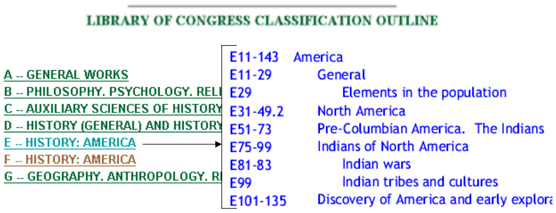
- **Distributed retrieval**
  - Ad-hoc retrieval in a distributed computing environment
    - many text collections
    - reside on different machines
    - possibly different IR system for each machine
  - Issues to address include
    - Database selection
    - Merging results from different databases

## IR Basics

- **Task definition**
- **Evaluation**
  - Issues
  - Test collections
  - Metrics
- **Statistical properties of text**

## Evaluation in IR: History

- **Experimental methodology has been a prominent component of IR research since 1960's**
- **Early work compared manual vs. automatic indexing**



LIBRARY OF CONGRESS CLASSIFICATION OUTLINE

A -- GENERAL WORKS
B -- PHILOSOPHY. PSYCHOLOGY. RELI
C -- AUXILIARY SCIENCES OF HISTORY
D -- HISTORY (GENERAL) AND HISTORY
E -- HISTORY: AMERICA
F -- HISTORY: AMERICA
G -- GEOGRAPHY. ANTHROPOLOGY. RI

| | |
|---|---|
| E11-143 | America |
| E11-29 | General |
| E29 | Elements in the population |
| E31-49.2 | North America |
| E51-73 | Pre-Columbian America. The Indians |
| E75-99 | Indians of North America |
| E81-83 | Indian wars |
| E99 | Indian tribes and cultures |
| E101-135 | Discovery of America and early explora |

## Evaluation in IR: History

- **Manual vs. automatic indexing**
  - Could automatic indexing approach manual quality?
  - Issue: Humans are not as consistent as they think!
- **IR field developed methods of comparing overall system performance**
  - Batch
  - Interactive
- **Until 1990s, problems of scale**

## Types of Evaluation

- **IR components that might be evaluated**
  - Ability to assist formulating queries
  - Speed of retrieval
  - Computing resources required
  - Ability to find relevant documents
- **Evaluation generally comparative**
  - System A vs. system B
  - System A vs. system A’
- **Most common evaluation measure**
  - Retrieval effectiveness

## Ad-hoc Retrieval Example

- **Query:** ***ski areas in New York***
- **Results:**
  - GoSki New York – New York ski areas, snow …
  - NY ski areas on “I Love NY” tourism guide
  - Ski areas in the Adirondack region
  - Press Releases
  - Lake Placid
  - Ski areas in Central NY
  - Ski areas in Cortland County
  - Ski areas in the United States
  - Nordic skiing ski areas wrap up season
  - Greek Peek
  - AYH near ski areas

## Relevance

- **Relevance is difficult to define satisfactorily**
- **A relevant document is one judged useful in the context of a query**
  - Who judges?
  - What is “useful”?
  - Issue of serendipitous utility
  - Humans aren’t consistent in their judgments
  - Judgment depends on more than the document and query
- **With real collections, the full set of relevant documents is never known**
- **All retrieval models include an implicit definition of relevance**

## Test Collections

- **Retrieval performance is compared using a test collection**
  - Set of documents, set of queries, set of relevance judgments
- **To compare two techniques**
  - Each technique is used to evaluate queries
  - Results (set or ranked list) compared using some metric
  - Most common measures: precision, recall
- **Usually use multiple measures to get different perspectives**
- **Usually test with multiple test collections because performance is collection-dependent to some extent**

## Sample Test Collections

| | Cranfield | CACM | ISI | TREC2 |
|---|---|---|---|---|
| Size (documents) | 1,400 | 3,204 | 1,460 | 742,611 |
| Size (MB) | 1.5 | 2.3 | 2.2 | 2,162 |
| Year created | 1968 | 1983 | 1983 | 1991 |
| Word stems | 8,226 | 5.493 | 5,448 | 1,040,415 |
| Stem occurrences | 123,200 | 117,578 | 98,304 | 243,800,000 |
| Avg DocLen (words) | 88 | 37 | 67 | 328 |
| Queries | 225 | 50 | 35 | 100 |

# Finding Relevant Documents

- **For small test collections, can review all documents for a query**
- **Not practical for large collections**
- **Pooling**
  - Retrieve documents using several techniques
  - Judge top n documents for each technique
  - Relevant set is union of relevant documents from each technique
  - Relevant set is a subset of the true relevant set
- **Possible to estimate size of true relevant set by sampling**
- **When testing:**
  - How should unjudged documents be treated?
  - How might this affect the results?

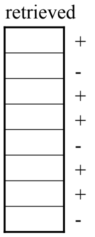
# Evaluation Metrics: Precision and Recall

- **Recall**
  - Percentage of all relevant documents that are found by a search

$$R = \frac{\#\ of\ relevant\ items\ retrieved}{\#\ of\ relevant\ items\ in\ collection}$$

- **Precision**
  - Percentage of retrieved documents that are relevant

$$P = \frac{\#\ of\ relevant\ items\ retrieved}{\#\ of\ items\ retrieved}$$



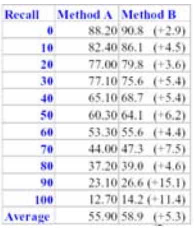
$R$ = 5/10 = 50%

$P$ = 5/8 = 62.5%

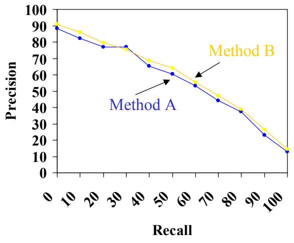
# Evaluation Metrics: Precision and Recall

- **Precision and recall are well-defined for unranked retrieval**
  - Unranked retrieval produces a set of documents
- **For ranked retrieval**
  - The entire collection is ranked (in theory)
    - Compute P at fixed recall points (e.g. precision at 20% recall)
    - Compute P at fixed rank cutoffs (e.g. precision at rank 20)

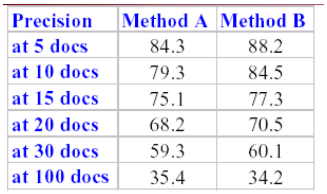
# Recall Precision Tables



| Recall | Method A | Method B |
|---|---|---|
| 0 | 88.20 | 90.8 (+2.9) |
| 10 | 82.40 | 86.1 (+4.5) |
| 20 | 77.00 | 79.8 (+3.6) |
| 30 | 77.10 | 75.6 (+5.4) |
| 40 | 65.10 | 68.7 (+5.4) |
| 50 | 60.30 | 64.1 (+6.2) |
| 60 | 53.30 | 55.6 (+4.4) |
| 70 | 44.00 | 47.3 (+7.5) |
| 80 | 37.20 | 39.0 (+4.6) |
| 90 | 23.10 | 26.6 (+15.1) |
| 100 | 12.70 | 14.2 (+11.4) |
| Average | 55.90 | 58.9 (+5.3) |



# Precision at Fixed Rank Cutoffs



| Precision | Method A | Method B |
|---|---|---|
| at 5 docs | 84.3 | 88.2 |
| at 10 docs | 79.3 | 84.5 |
| at 15 docs | 75.1 | 77.3 |
| at 20 docs | 68.2 | 70.5 |
| at 30 docs | 59.3 | 60.1 |
| at 100 docs | 35.4 | 34.2 |

# F-measure

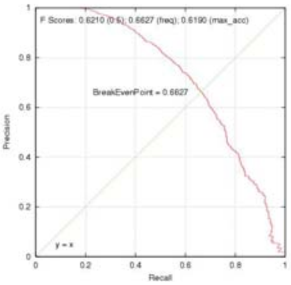
harmonic average of precision and recall

$$F = \frac{2 * (PRECISION \times RECALL)}{(PRECISION + RECALL)}$$

- **rewards results that keep recall and precision close together**
  - R=40, P=60. R/P average = 50. F-measure= 48
  - R=45, P=55. R/P average = 50. F-measure= 49.5

## BreakEvenPoint

- **break even point is the point at which recall equals precision**



## IR Basics

- Task definition
- Evaluation
- **Statistical properties of text**
  - Zipf's Law
  - Collocations and Co-occurrences

## Statistical Properties of Text

- **There are stable, language-independent patterns in how people use natural language**
  - A few words occur very frequently; most occur rarely
  - In general
    - Top 2 words ~ 10-15% of all word occurrences
    - Top 6 words ~ 20% of all word occurrences
    - Top 50 words ~ 50% of all word occurrences

most common words from *Tom Sawyer*

| | | |
|---|---|---|
| 1 | **The** | **3332** |
| | **And** | **2972** |
| | **A** | **1775** |
| | **To** | **1725** |
| | **Of** | **1440** |
| | **…** | |
| 14 | **Tom** | **679** |
| | **With** | **preposition** |

## Statistical Properties of Text

- The most frequent words in one corpus may be rare words in another corpus
  - Example: "computer" in CACM vs. National Geographic
- Each corpus has a different, fairly small "working vocabulary"

These properties hold in a wide range of languages

## Zipf's Law

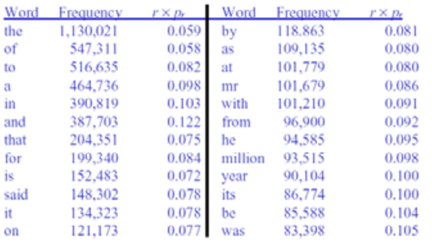
- **Zipf's Law relates a term's frequency to its rank**
  - frequency ∝ 1/rank
  - There is a constant *k* such that *frequency* * *rank* = *k*
  - Rank the terms in a vocabulary by frequency, in descending order

    $f_r$ : frequency of term at rank $r$

    $N$ : total number of word occurrences

    $p_r = f_r / N$ and $\sum_{r=1}^{V} p_r = 1$
  - Empirical observation: $p_r = A / r, \quad A \approx 0.1$
  - Hence: $p_r = \frac{f_r}{N} = \frac{A}{r} \rightarrow rf_r = AN$
  - $k \approx N/10$ for English

## Zipf's Law



| Word | Frequency | $r \times p_r$ | Word | Frequency | $r \times p_r$ |
|---|---|---|---|---|---|
| the | 1,130,021 | 0.059 | by | 118.863 | 0.081 |
| of | 547,311 | 0.058 | as | 109,135 | 0.080 |
| to | 516,635 | 0.082 | at | 101,779 | 0.080 |
| a | 464,736 | 0.098 | mr | 101,679 | 0.086 |
| in | 390,819 | 0.103 | with | 101,210 | 0.091 |
| and | 387,703 | 0.122 | from | 96,900 | 0.092 |
| that | 204,351 | 0.075 | he | 94,585 | 0.095 |
| for | 199,340 | 0.084 | million | 93,515 | 0.098 |
| is | 152,483 | 0.072 | year | 90,104 | 0.100 |
| said | 148,302 | 0.078 | its | 86,774 | 0.100 |
| it | 134,323 | 0.078 | be | 85,588 | 0.104 |
| on | 121,173 | 0.077 | was | 83,398 | 0.105 |

WSJ87 collection (46,449 articles, 19 million term occurrences, 132 MB)

## Predicting Occurrences of Frequencies

- **A word that occurs *n* times has rank** $r_n = \frac{AN}{n}$
  - Example: n=50, A=0.1, N=100,000

    $$r_n = 0.1 * 100{,}000/50 = 200$$

  - Several words may occur *n* times; assume rank $r_n$ applies to last word that occurs *n* times
  - $r_n$ words occur more than *n* times
  - $r_{n+1}$ words occur more than *n+1* times

## Predicting Occurrences of Frequencies

- The number of words that occur exactly *n* times is:

  $$I_n = r_n - r_{n+1} = AN/n - AN/(n+1) = AN/(n(n+1))$$

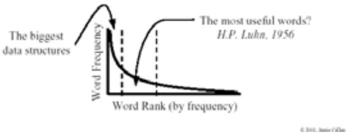
- Highest ranking term occurs once and has rank $r_{max}=AN/1$
- Proportion of words with frequency 1 is:

  $$I_n / r_{\max} = 1/(n(n+1))$$ (independent of text length and A)

- Proportion of words occurring once is 1/2

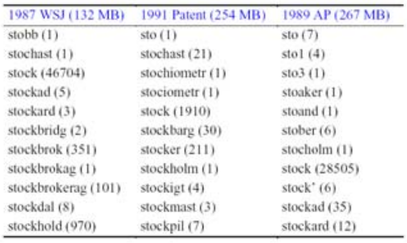
## Statistical Properties of Text

- **Summary:**
  - Term usage is highly skewed, but in a *predictable* pattern
- **Why is it important to know the characteristics of text?**
  - Optimization of data structures
  - Statistical retrieval algorithms depend on them



## Statistical Profiles

- **Can act as a summarization device**
  - Indicate what a document is about
  - Indicate what a collection is about



| 1987 WSJ (132 MB) | 1991 Patent (254 MB) | 1989 AP (267 MB) |
|---|---|---|
| stobb (1) | sto (1) | sto (7) |
| stochast (1) | stochast (21) | sto1 (4) |
| stock (46704) | stochiometr (1) | sto3 (1) |
| stockad (5) | stociometr (1) | stoaker (1) |
| stockard (3) | stock (1910) | stoand (1) |
| stockbridg (2) | stockbarg (30) | stober (6) |
| stockbrok (351) | stocker (211) | stocholm (1) |
| stockbrokag (1) | stockholm (1) | stock (28505) |
| stockbrokerag (101) | stockigt (4) | stock' (6) |
| stockdal (8) | stockmast (3) | stockad (35) |
| stockhold (970) | stockpil (7) | stockard (12) |

## Collocations and Co-occurrences

- **A collocation is an expression consisting of two or more words that occur in a particular order and correspond to some conventional way of saying things**
  - Noun phrases (e.g. *a stiff breeze*, *weapons of mass destruction*)
  - Phrasal verbs (e.g. *to make up*)
  - Stock phrases (e.g. *the rich and famous*, *vim and vigor*)
- **Two words co-occur if they appear in the same context (in general) or the same text (in IR)**
  - Co-occurrence patterns
    - *doctor* with *nurse*, *honorary*, *dentist*, *treat*, *examined*, *bills*, etc.
    - people and companies
      - *Ted Turner* with *Turner Broadcasting*, *Atlanta Braves*, etc.