COEN 169

Statistical Properties of Text

Yi Fang

Department of Computer Engineering

Santa Clara University

Statistical Properties of Text

- How is the frequency of different words distributed?
- How fast does vocabulary size grow with the size of a corpus?
- Such factors affect the performance of information retrieval and can be used to select appropriate term weights and other aspects of a search engine.

Word Frequency

- A few words are very common.
  - 2 most frequent words (e.g. “the”, “of”) can account for about 10% of word occurrences
- Most words are very rare.
  - Half the words in a corpus appear only once
- Called a “heavy tailed” distribution, since most of the probability mass is in the “tail”

Sample Word Frequency Data

(from B. Croft, UMass)

Frequent Word	Number of Occurrences	Percentage of Total
the	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

Frequencies from 336,310 documents in the 1GB TREC Volume 3 Corpus
125,720,891 total word occurrences; 508,209 unique words

Zipf's Law

- Rank (*r*): The numerical position of a word in a list sorted by decreasing frequency (*f*).
- Zipf (1949) "discovered" that:

$$f \cdot r = k \ \ (\text{for constant } k)$$

- This is a statistically empirical law, not an exact physical law!

Zipf's Law

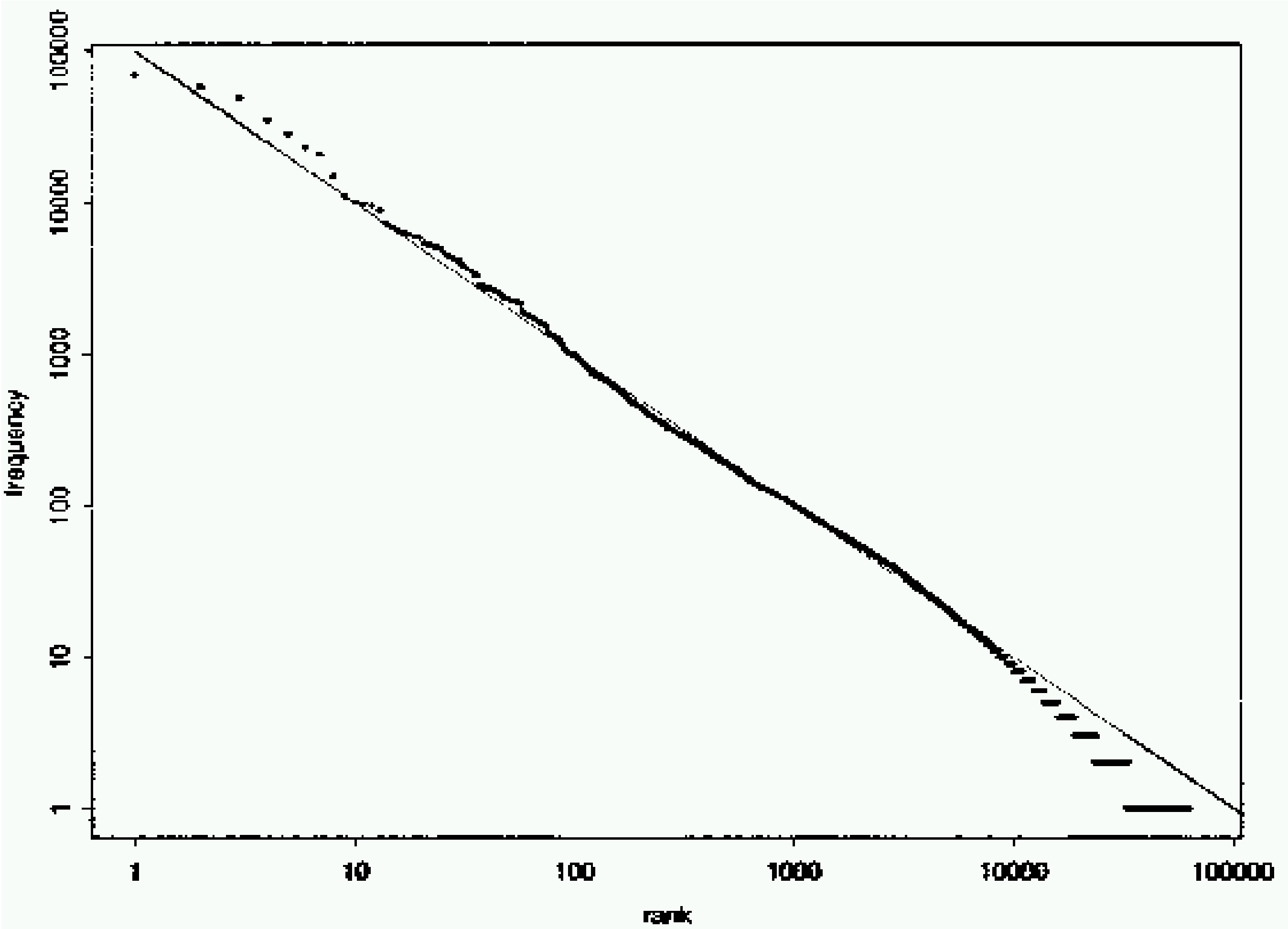
$$f = \frac{k}{r}$$

$$\log(f) = \log(\frac{k}{r})$$

$$\log(f) = \log(k) - \log(r)$$

- If Zipf's law holds true, we should be able to plot log(*f*) vs. log(*r*) and see a straight light with a slope of -1

Fit to Zipf for Brown Corpus



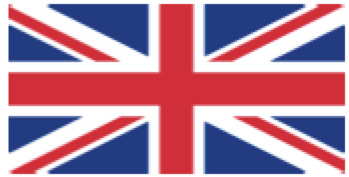
$k = 100{,}000$

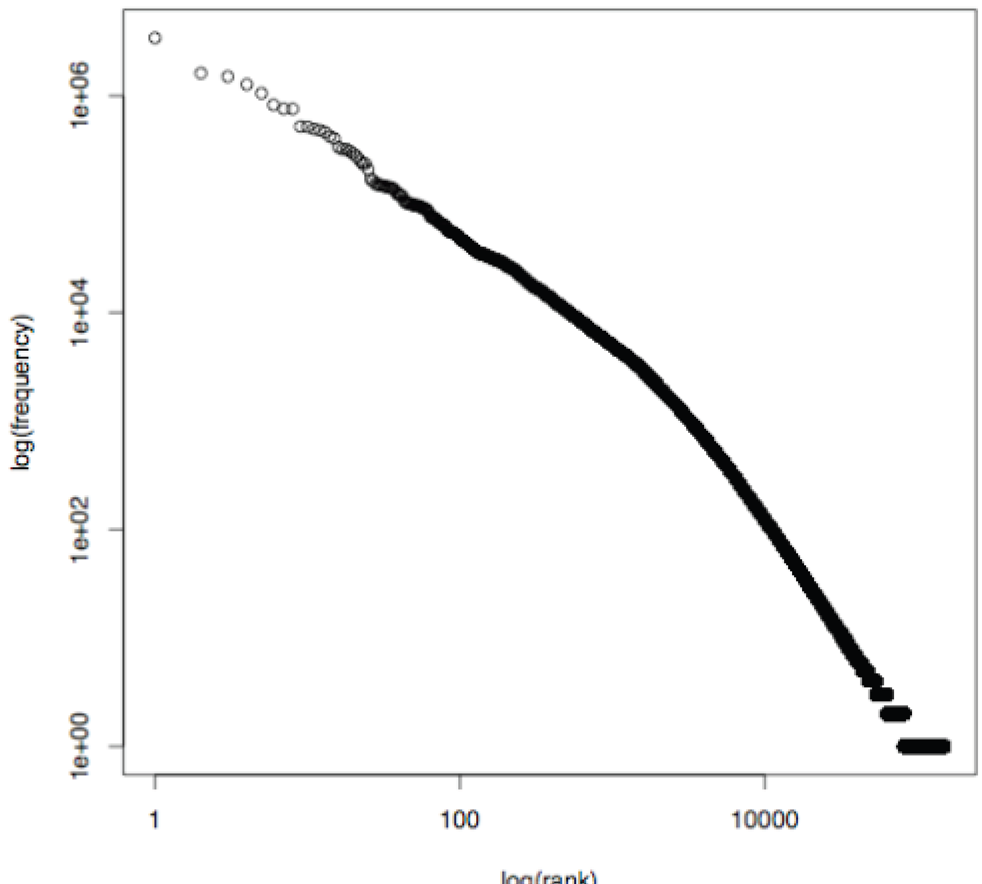
Zipf's Law

- If probability of word of rank *r* is p_r and *N* is the total number of word occurrences in the corpus:

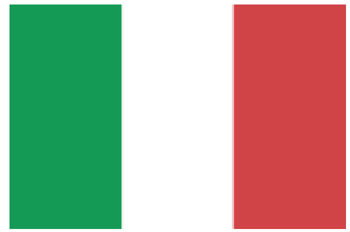
$$p_r = \frac{f}{N} = \frac{A}{r} \quad \text{for corpus independent const } A \approx 0.1$$

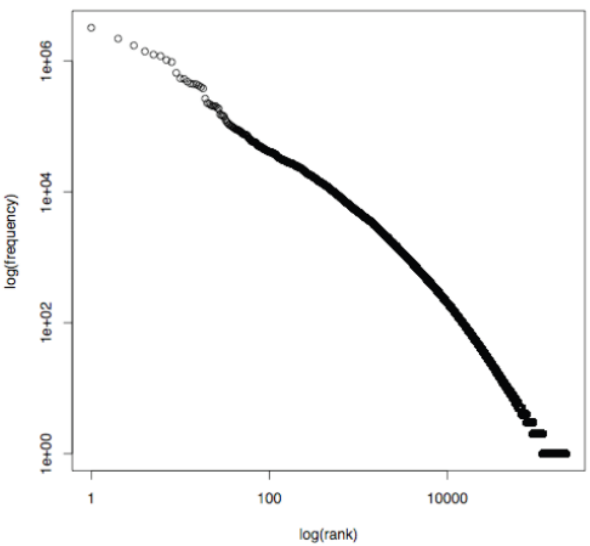
Fit to Zipf for European Parliament Corpus (Koehn ’05)

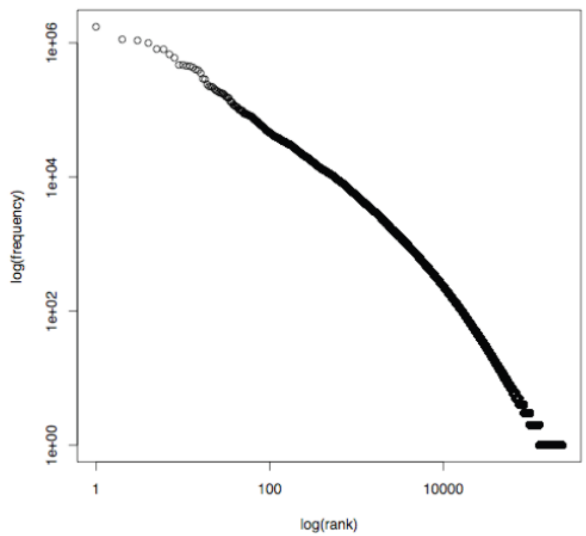


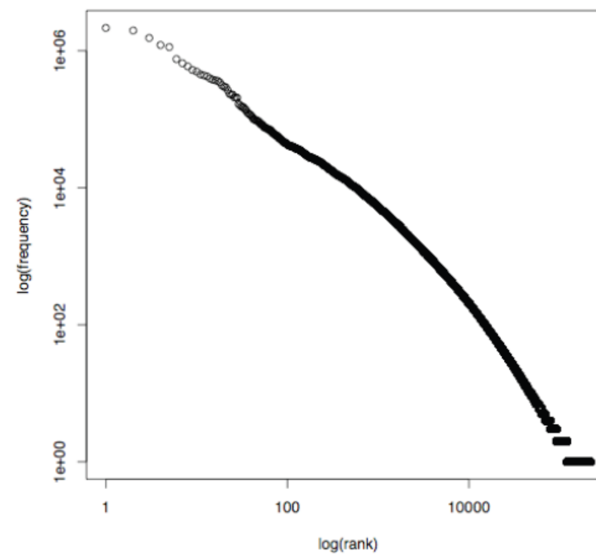


Does Zipf’s Law generalize across languages?





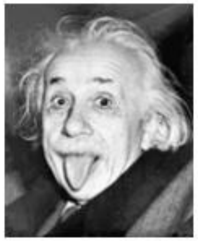




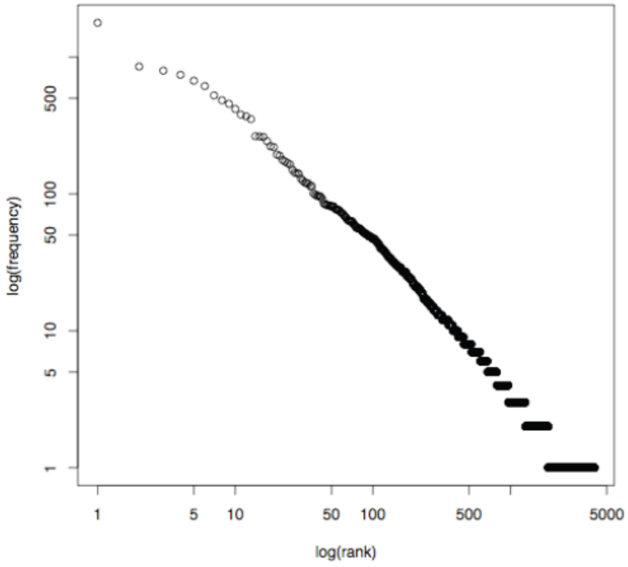
Across different texts

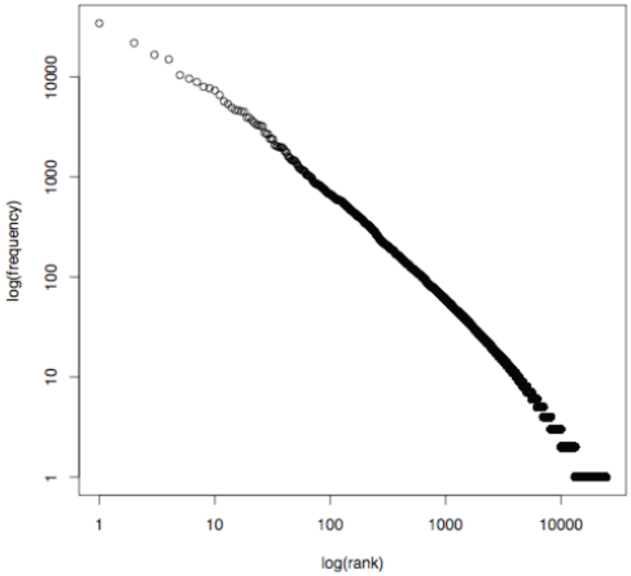
Alice in Wonderland

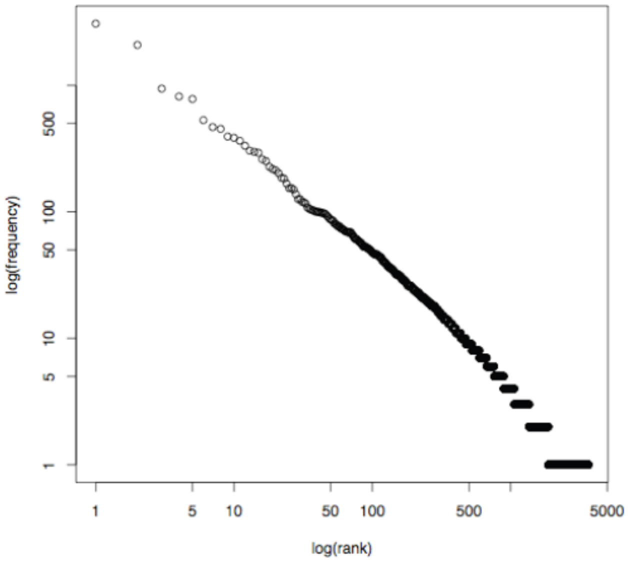
War and Peace



Relativity







Zipf's Law

Zipf's Law holds true for:

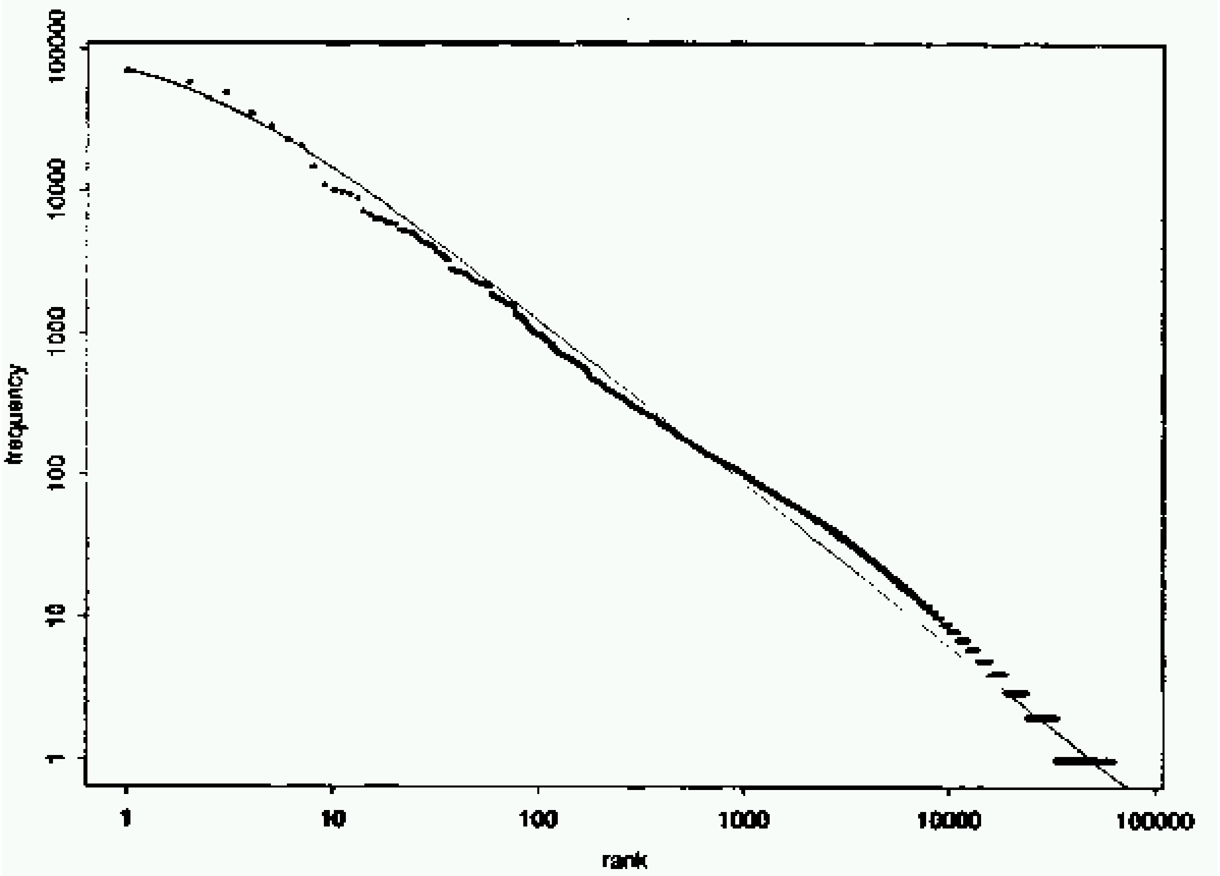
- different languages
- different sizes of text
- different genres
- different topics
- different complexity of content

Mandelbrot (1954) Correction

- The following more general form gives a bit better fit:

$$f = P(r + \rho)^{-B} \qquad \text{For constants } P, B, \rho$$

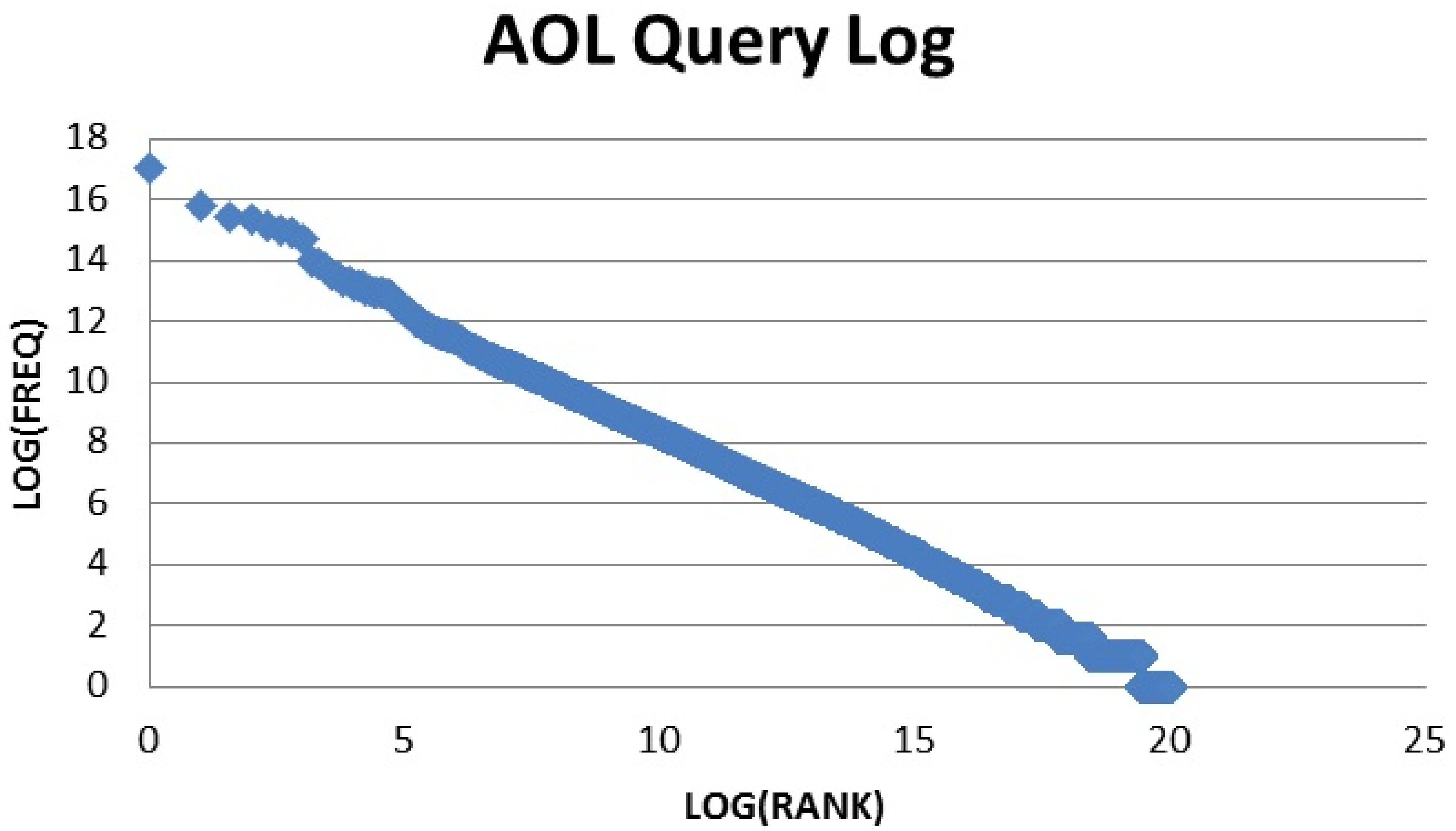
Mandelbrot Fit



Mandelbrot's function on Brown corpus

$P = 10^{5.4}$, $B = 1.15$, $\rho = 100$

Zipf's Law in search queries



Zipf’s Law in Web search queries

- Same trend: a few queries occur very frequently, while most occur very infrequently
- In Web search, half the queries issued on a given day are unique
- Search engines are usually tweaked to do well on those queries it is likely to “see” again and again

Zipf's Law Impact on IR

- Good News: Stopwords will account for a large fraction of text so eliminating them greatly reduces inverted-index storage costs

- Bad News: For most words, gathering sufficient data for meaningful statistical analysis (e.g. for correlation analysis for query expansion) is difficult since they are extremely rare.

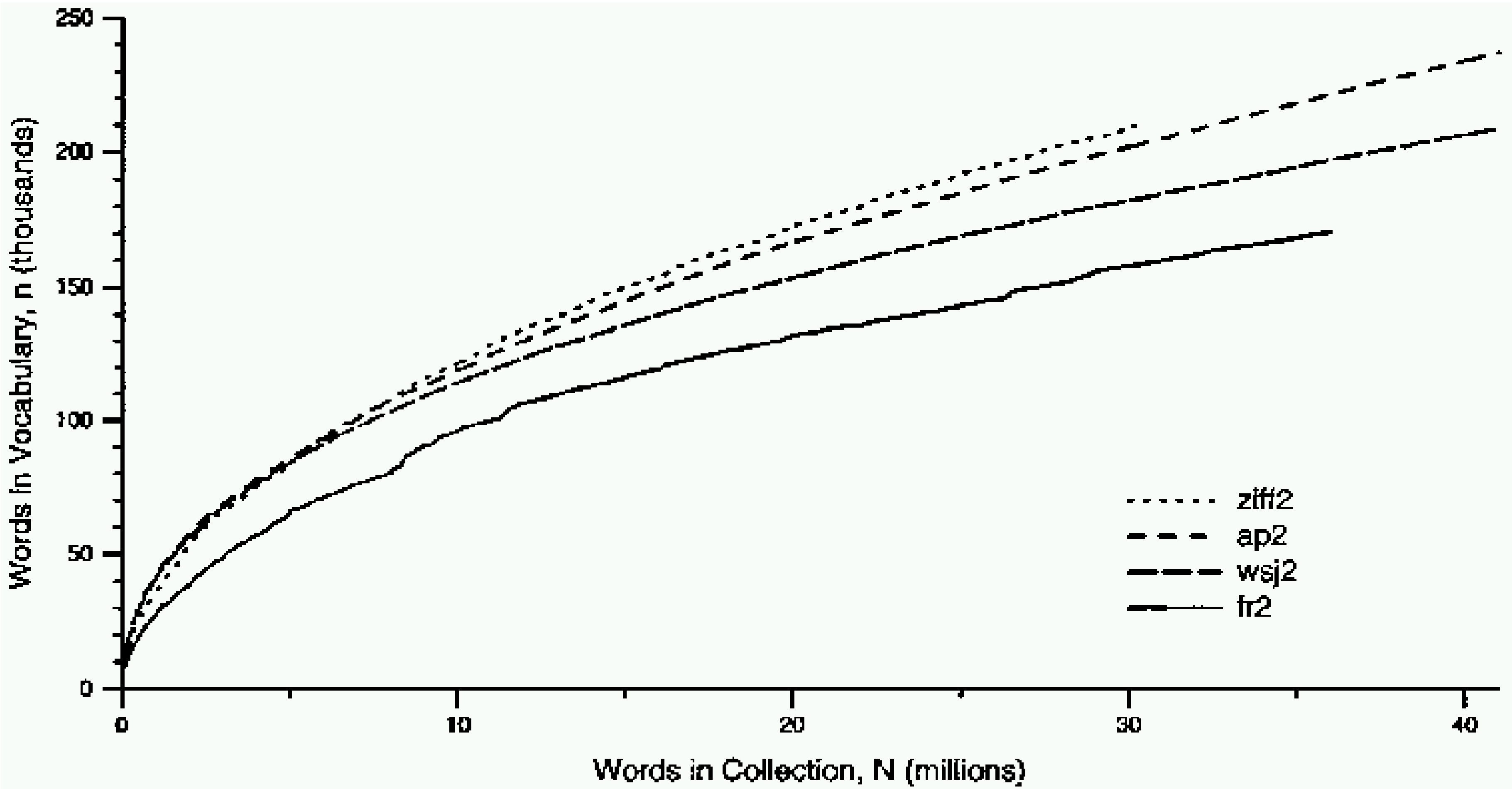
Heaps' Law

- If V is the size of the vocabulary and the n is the length of the corpus in words:

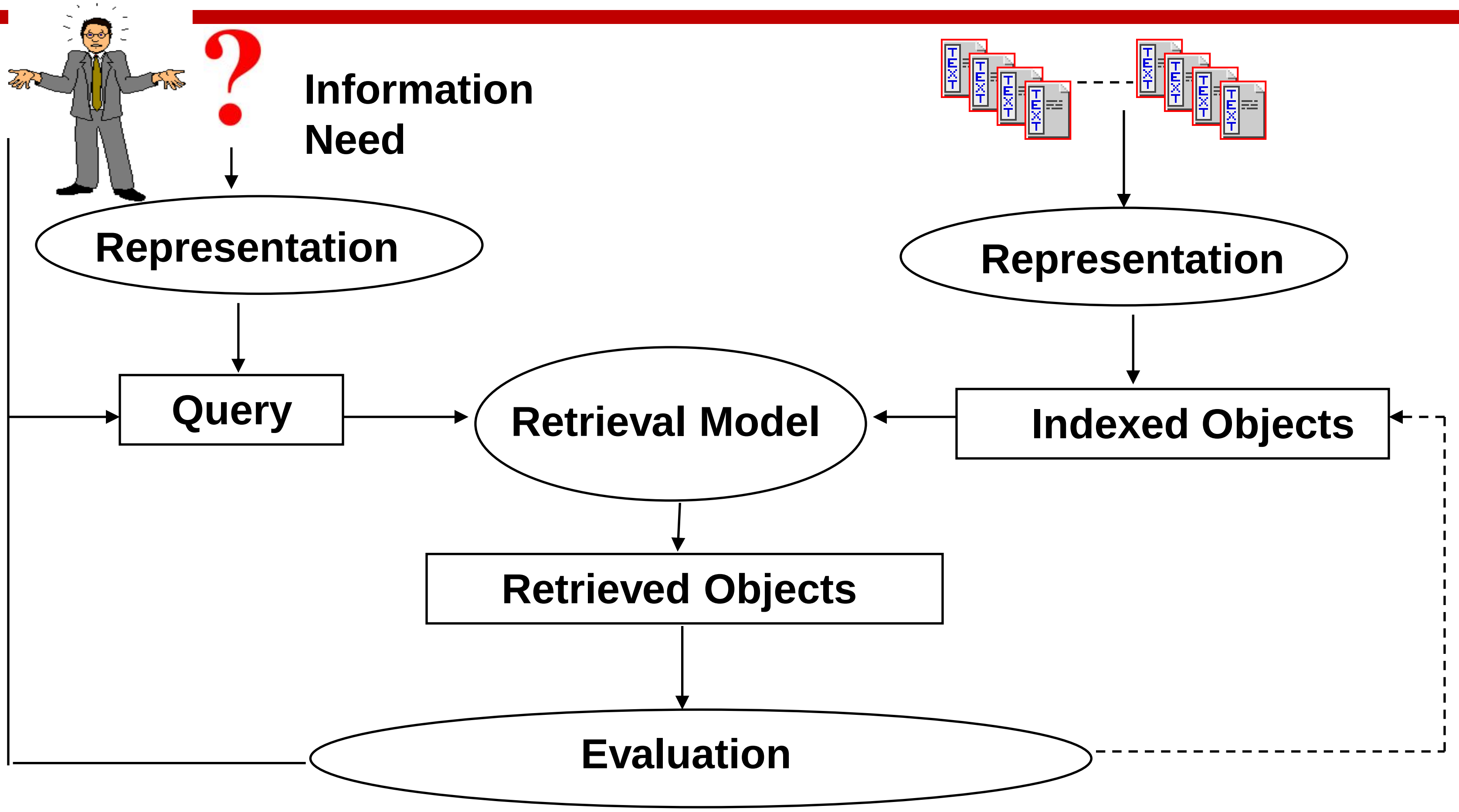
$$V = Kn^{\beta} \quad \text{with constants } K,\ 0 < \beta < 1$$

- Typical constants:
  - $K \approx 10–100$
  - $\beta \approx 0.4–0.6$ (approx. square-root)

Heaps' Law Data



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?

Why System Evaluation?

- There are many retrieval models/ algorithms/ systems, which one is the best?
- What is the best component for:
  - Term selection (stopword removal, stemming…)
  - Term weighting (TF, TF-IDF,…)
  - Ranking function (dot-product, cosine, …)
- How far down the ranked list will a user need to look to find some/all relevant documents?

Why System Evaluations?

- From all the ranking schemes that are possible with given weighting/ranking schemes, which one has the best performance?
- For a fair comparison:
  - Should be all evaluated on the same collection of documents
  - Should be all evaluated on the same set of questions/queries
  - Should be all evaluated using the same measures

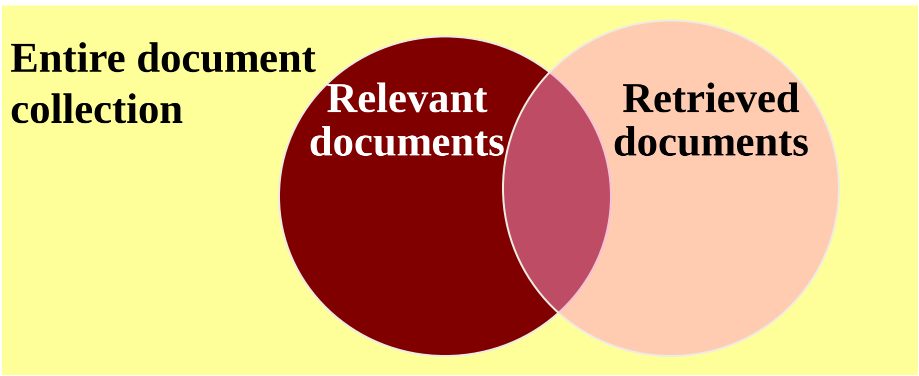
Difficulties in Evaluating IR Systems

- Effectiveness is related to the ***relevancy*** of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
  - Subjective: Depends upon a specific user's judgment.
  - Situational: Relates to user's current needs.
  - Cognitive: Depends on human perception and behavior.
  - Dynamic: Changes over time.

Human Labeled Corpora (Gold Standard)

- Start with a corpus of documents.
- Collect a set of queries for this corpus.
- Have one or more human experts exhaustively label the relevant documents for each query.
- Typically assumes binary relevance judgments.
- Requires considerable human effort for large document/query corpora.

Precision and Recall



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

$$precision = \frac{Number\ \ of\ \ relevant\ documents\ \ retrieved}{Total\ number\ of\ \ documents\ \ retrieved}$$

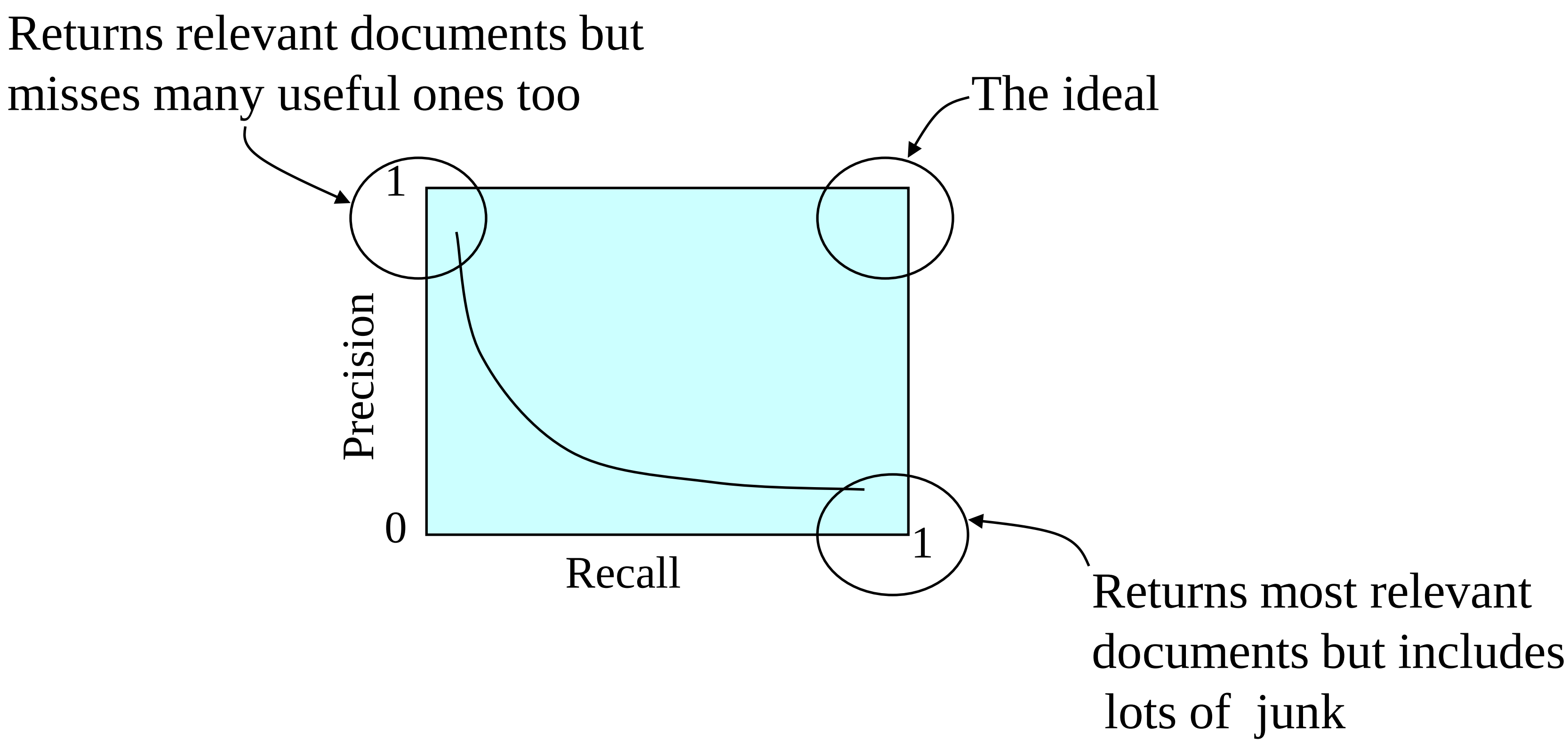
$$recall = \frac{Number\ of\ \ relevant\ documents\ retrieved}{Total\ number\ of\ \ relevant\ documents}$$

Determining Recall is Difficult

- Precision vs. Recall:
  - Precision = The ability to retrieve top-ranked documents that are mostly relevant.
  - Recall = The ability of the search to find ***all*** of the relevant items in the corpus.

- Total number of relevant items is sometimes not available:
  - Sample across the database and perform relevance judgment on these items.
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set.

Trade-off between Recall and Precision



Precision and Recall are inverse proportional

Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; P=5/13=0.38

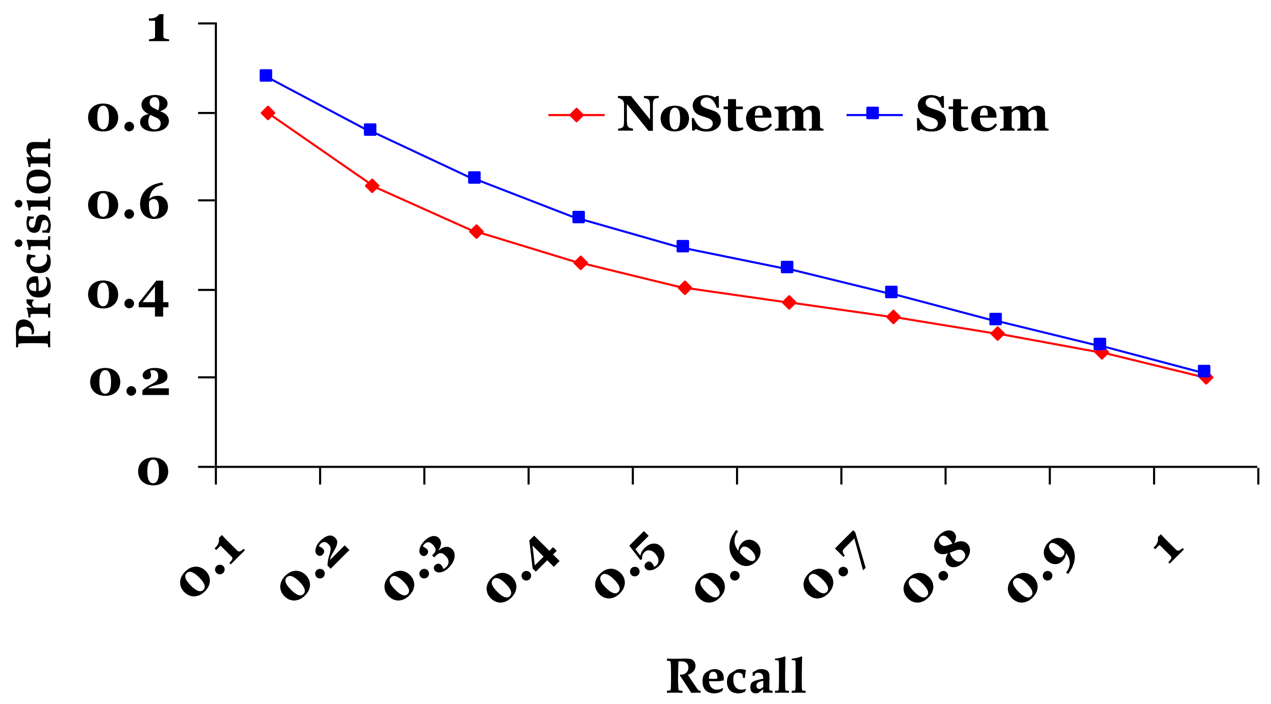
Missing one relevant document.
Never reach 100% recall

Interpolating a Recall/Precision Curve

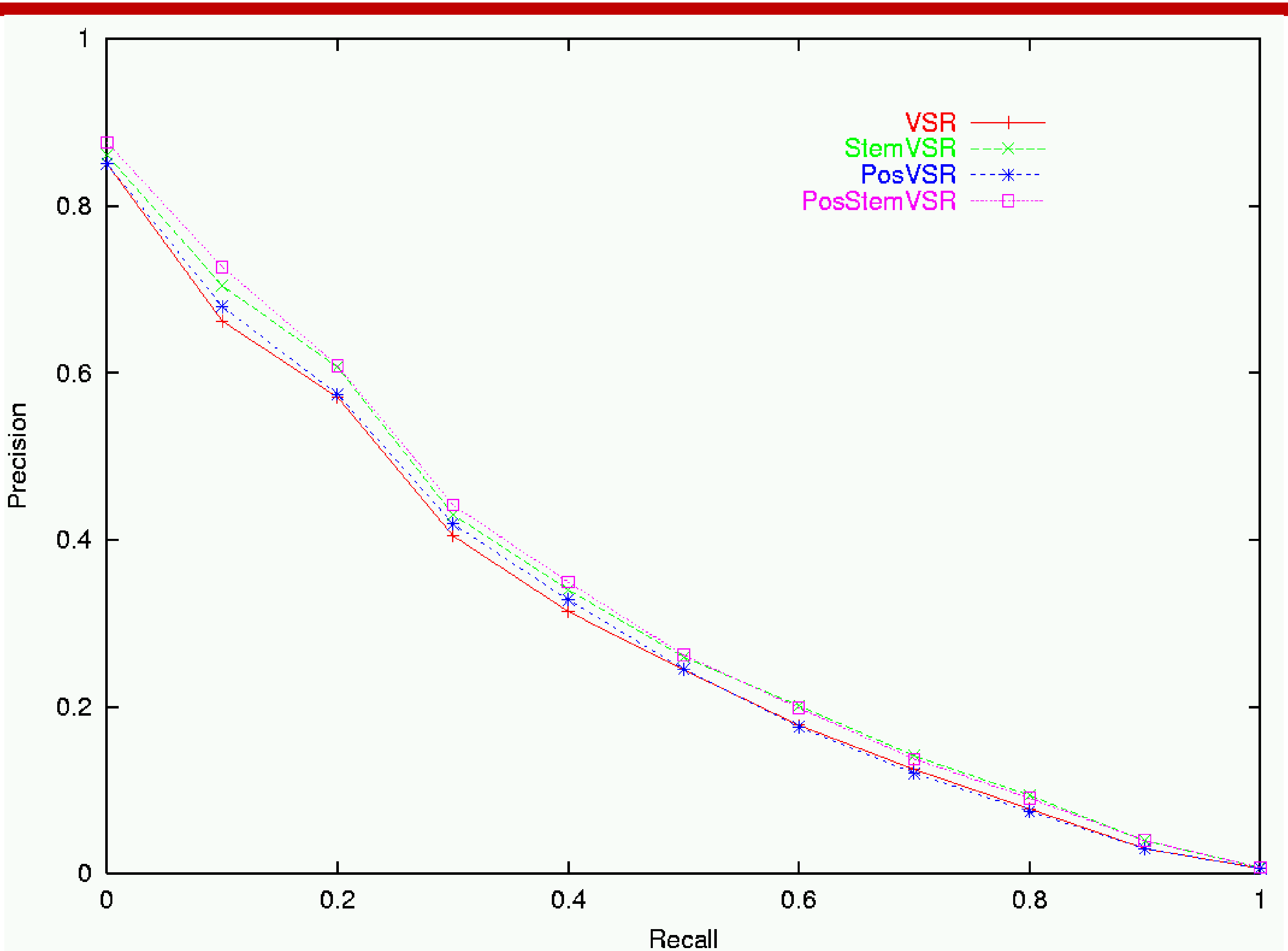
- Interpolate a precision value for each *standard recall level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0$, $r_1 = 0.1$, …, $r_{10} = 1.0$
- The interpolated precision at the *j*-th standard recall level is the highest precision found for any recall level $r >= r_j$

$$P(r_j) = \max_{r >= r_j} P(r)$$

Which system is better?



Sample Recall/Precision Curve



Average Precision

- Average Precision (AP)
  - Average of precision at each relevant document retrieved
  - Precision of an unretrieved relevant document = 0

- Mean Average Precision
  - The mean AP over all queries

F-Measure

- One measure of performance that takes into account both recall and precision.
- Introduced by van Rijbergen, 1979
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

NDCG

- Normalized Discounted Cumulative Gain
- Popular measure for evaluating web search
- Two assumptions:
  - Highly relevant documents are more useful than marginally relevant document
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks
- Typical discount is 1/*log (rank)*
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain

- *DCG* is the total gain accumulated at a particular rank *p*:

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

  - used by Web search companies
  - emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
- discounted gain:

  3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

  = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
- DCG:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Normalized DCG

- DCG numbers are averaged across a set of queries at specific rank values
  - e.g., DCG at rank 5 is 6.89 and at rank 10 is 9.61
- DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking*
  - makes averaging easier for queries with different numbers of relevant documents

NDCG Example

- Perfect ranking:

  3, 3, 3, 2, 2, 2, 1, 0, 0, 0

- ideal DCG values:

  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88

- NDCG values (divide actual by ideal):

  1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88

  – NDCG $\leq$ 1 at any rank position

What is the Right Measure?

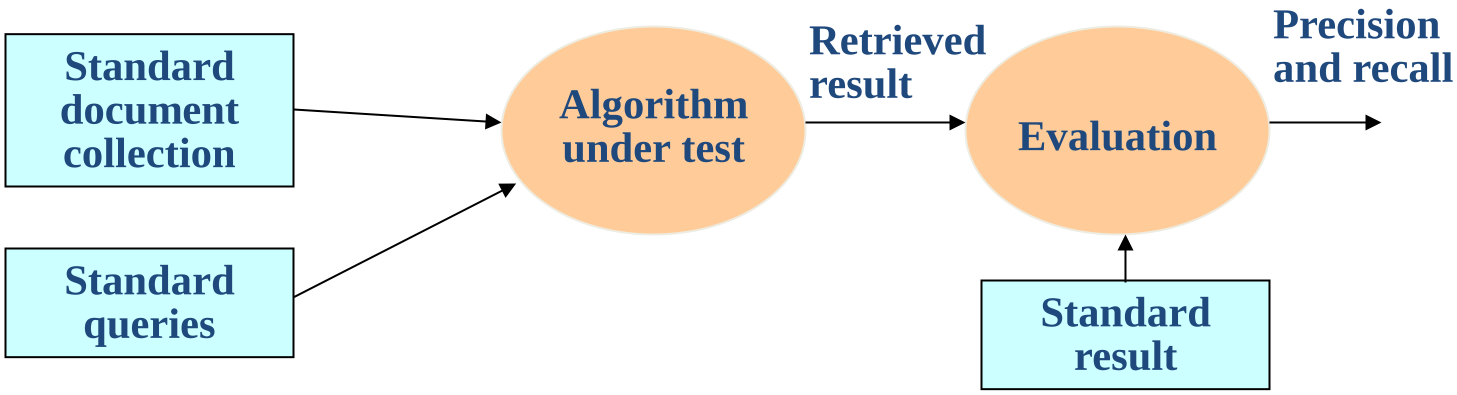
- Precision: “*I'm feeling lucky*”
- Recall: maximizing coverage of topic
- F: explore P/R tradeoff

Benchmarking

- ***Analytical*** performance evaluation is difficult for document retrieval systems because many characteristics such as relevance, distribution of words, etc., are difficult to describe with mathematical precision.
- Performance is measured by ***benchmarking***. That is, the retrieval effectiveness of a system is evaluated on a *given set of documents*, *queries*, and *relevance judgments*.
- Performance data is valid only for the environment under which the system is evaluated.

Benchmarks

- A benchmark collection contains:
  - A set of standard documents and queries/topics.
  - A list of relevant documents for each query.
- Standard collections for traditional IR:
  - Smart collection: ftp://ftp.cs.cornell.edu/pub/smart
  - TREC: http://trec.nist.gov/



Early Test Collections

- Previous experiments were based on the SMART collection which is fairly small. (ftp://ftp.cs.cornell.edu/pub/smart)

Collection Name	Number Of Documents	Number Of Queries	Raw Size (Mbytes)
CACM	3,204	64	1.5
CISI	1,460	112	1.3
CRAN	1,400	225	1.6
MED	1,033	30	1.1
TIME	425	83	1.5

- Most collections available from http://www.sigir.org

The TREC Benchmark

- TREC: Text REtrieval Conference (http://trec.nist.gov/) originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).
- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.

The TREC Benchmark

- Participants are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing.

- Participants submit the P/R values for the final document and query corpus and present their results at the conference.

The TREC Objectives

- Provide a common ground for comparing different IR techniques.
  - Same set of documents and queries, and same evaluation method.
- Sharing of resources and experiences in developing the benchmark.
  - With major sponsorship from government to develop large benchmark collections.
- Encourage participation from industry and academia.
- Development of new evaluation techniques, particularly for new applications.
  - Retrieval, routing/filtering, non-English collection, web-based collection, question answering.

TREC Advantages

- Large scale (compared to a few MB in the SMART Collection).
- Relevance judgments provided.
- Under continuous development with support from the U.S. Government.
- Wide participation:
  - TREC 1: 28 papers 360 pages.
  - TREC 4: 37 papers 560 pages.
  - TREC 7: 61 papers 600 pages.
- Diverse tasks:
  - Information retrieval
  - Text classification
  - Domain-specific retrieval:
    - Blogs, medical domain
  - Cross-language information retrieval

Current TREC Tasks

- Chemical
- Crowdsourcing Track
- Entity Track
- Legal Track
- Medical Records Track
- Microblog Track
- Session Track
- Web Track

New TREC testbed

- Testbed: Clueweb09
  Size (count): 1.04 billion web pages
  Size (TB): 25 Terabytes
  Crawl period: January & February, 2009

Sample TREC Document

<DOC>

<DOCNO> WSJ870324-0001 </DOCNO>

<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>

<DD> 03/24/87</DD>

<SO> WALL STREET JOURNAL (J) </SO>

<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM) MARKETING, ADVERTISING (MKT) TELECOMMUNICATIONS, BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>

<DATELINE> NEW YORK </DATELINE>

<TEXT>

John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said. Industry sources put the value of the proposed acquisition at more than $100 million. ...

</TEXT>

</DOC>

TREC Properties

- Both documents and queries contain many different kinds of information (fields).
- Generation of the formal queries (Boolean, Vector Space, etc.) is the responsibility of the system.
  - A system may be very good at querying and ranking, but if it generates poor queries from the topic, its final P/R would be poor.

Evaluation at TREC

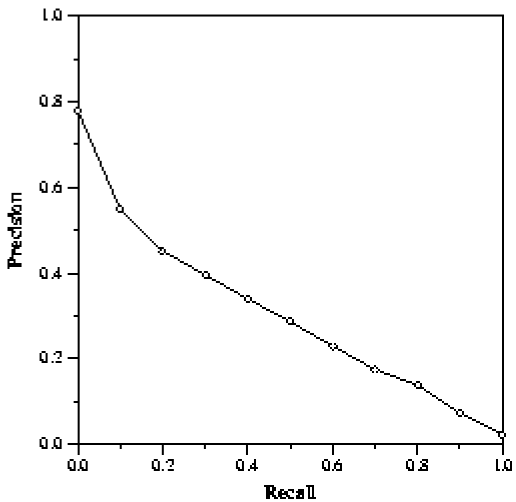
- Summary table statistics: Number of topics, number of documents retrieved, number of relevant documents.
- Recall-precision average: Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- Document level average: Average precision when 5, 10, .., 100, … 1000 documents are retrieved.
- Average precision histogram: Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

Ad hoc results — Fujitsu Laboratories. Ltd.

Summary Statistics	
Run Number	Flab8atd2
Run Description	Automatic, title + desc
Number of Topics	50
Total number of documents over all topics	
Retrieved:	50000
Relevant:	4728
Rel ret:	2990

Recall Level Precision Averages	
Recall	Precision
0.00	0.7796
0.10	0.5490
0.20	0.4517
0.30	0.3954
0.40	0.3397
0.50	0.2863
0.60	0.2291
0.70	0.1745
0.80	0.1381
0.90	0.0720
1.00	0.0224
Average precision over all relevant docs	
non interpolated	0.2930

Document Level Averages	
	Precision
At 5 docs	0.5480
At 10 docs	0.4880
At 15 docs	0.4587
At 20 docs	0.4200
At 30 docs	0.3887
At 100 docs	0.2490
At 200 docs	0.1777
At 500 docs	0.1011
At 1000 docs	0.0598
R Precision (precision after R docs retrieved (where R is the number of relevant documents))	
Exact	0.3203



Recall-Precision Curve

Lecture(s) 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