# Overview of Information Retrieval (IR)

J. Savoy
Université de Neuchatel

C. Manning, P. Raghavan, H. Schütze: *Introduction to Information Retrieval*. Cambridge University Press, 2008.
M. Boughanem, J. Savoy: Recherche d'information. Hermes, Paris, 2008.
W.B. Croft, H. Turtle: Introduction to Information Retrieval. Spring course, 1997.
J. Allen: Information Retrieval Course. University of Massachusetts at Amherst, 2004.



## IR Domains

- What makes a system like Google or Bing Search tick?
  - How does it gather information? What tricks does it use?
  - Extending beyond the Web
- How can those approaches be made better?
  - Natural language (NL) understanding?
  - User interactions?
- What can we do to make things work quickly?
  - Faster computers (Moore's law)? Caching? Compression?
- How do we decide whether it works well?
  - For all queries? For special types of queries?
  - On every collection of information?
- What else can we do with the same approach?
  - Other media?
  - Other languages?
  - Other tasks?



## Outline

- **What is Information Retrieval (IR)?**
- Basic IR process
- Simple model of IR
- The Web
- Conclusion

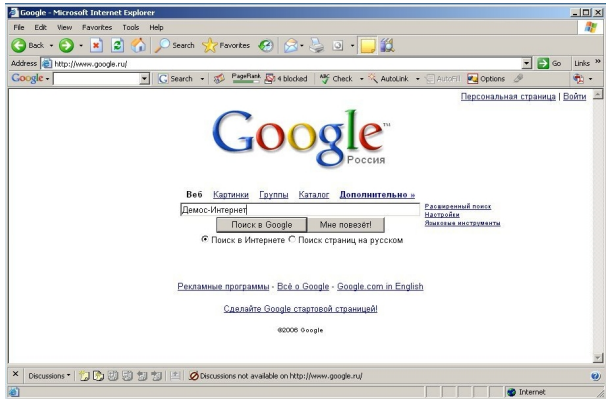


## Definition

Information retrieval deals with the *representation*, *storage*, *organization* of, and *access* to information items. These information items could be references to real documents, documents themselves or even single paragraphs, as well as web pages, spoken documents, images, pictures, music, video, etc.
[Baeza-Yates & Ribiero-Neto, 1999]

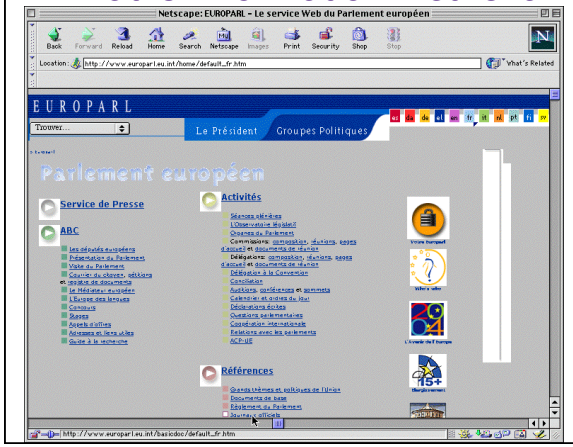
The requests are vague and imprecise description of the user's information need.

## What is Information Retrieval




5

## What is Information Retrieval




6

## Sample Systems

- IR systems
  - Verity, Fulcrum, Excalibur, Eurospider
  - OpenText (Hummingbird)
  - Smart, Lucene, Okapi, Lemur, Inquery
- Database systems
  - Oracle, Informix, Access, MySQL
- Web search and In-house systems
  - Thomson-Reuters (Westlaw), LEXIS/NEXIS, Dialog
  - Google, Yahoo!, Lycos, AltaVista, Northern Light, Teoma,
  - HotBot, Direct Hit, …
  - Ask Jeeves
- And countless others...

7

## Need to manage a huge volume

- 10 MB
  - Papers written by a researcher over a ten years period
- 100 MB
  - All e-mails of a person during 10 years
- 100 GB
  - Text of all books in a small university library
- 40 TB
  - The complete text-only of the Web in 2005
  - The complete Library Of Congress in text format (27 M of items) (see www.loc.gov)
- 167 TB
  - The complete Web in 2002
- 91,850 TB
  - The deep Web in 2002
- 440,606 TB
  - All e-mails around the planet

Lyman P., Varian H. R. *How much information? 2003*, available at the web site www.sims.berkeley.edu /how-much-info/

8

## With Relational Database?

A relation for all books in the library
Other fields not shown

| BookID | AuID | EdsID | Title | Year | Pages |
|---|---|---|---|---|---|
| L1 | A1 | E2 | Language and representation in information retrieval | 1990 | 335 |
| L2 | A2 | E1 | Information retrieval and hypertext | 1996 | 278 |
| L3 | A3 | E5 | Automatic text processing | 1989 | 356 |
| L4 | A4 | E4 | Information retrieval | 1979 | 208 |
| L5 | A5 | E6 | Online information retrieval | 1986 | 256 |

9

## With Database

And the query about the content

```
Select  name, title, year
   from  book, author
   where title = "Information retrieval"
```

```
Name            Title                  Year
-----           -----                  ----
van Rijsbergen  Information retrieval  1989
```

Do we solve the problem?

10

## Comparing IR to Databases

| | Database | IR |
|---|---|---|
| Data | structured | unstructured |
| Fields | Clear semantics (domain) | No fields (other than text) |
| Model | Determinist | Probabilistic |
| Queries | Defined (SQL, relational algebra), complex, complete specification | Free text (NL) flat, Boolean, partial |
| Access | Primary keys | ? |
| Matching | Exact | Best |
| Recoverability | Critical (concurrency control, recovery, atomic operations) | "try again" |

11

## Basic Approach to IR

- Most successful approaches are statistical
  - Directly, or an effort to capture and use probabilities
- Why not natural language understanding?
  - i.e., computer understands documents and query and matches them
  - State of the art is brittle in unrestricted domains
  - Can be highly successful in predictable settings
    - e.g., information extraction on terrorism/takeovers (MUC)
    - Medical or legal settings with restricted vocabulary
- Could use manually assigned headings
  - e.g., Library of Congress headings, Dewey Decimal headings
  - Human agreement is not good
  - Hard to predict what headings are "interesting"
  - Expensive

12

## Relevant Items are Similar

- Most successful approaches are statistical
  - No deeper natural language understanding
- Much of IR depends upon idea that
  similar vocabulary → similar meanings
  similar vocabulary → relevant to same queries
- Usually look for documents matching query words
- "Similar" can be measured in many ways
  - String matching / comparison
  - Same vocabulary used
  - Same meaning of text

13

## Example of NLP

- Polysemy
  Same words → different meanings
  Only one sense in Java?
  (an island, coffee, a dance, a domestic fowl, a computer programming language)
  BSE (Bovine Spongiform Encephalopathy, Bombay Stock Exchange (or Boston, Beirut, Bahrain), Breast Self-Examination, Bachelor of Science in Engineering, Basic Service Element, etc.
- Synonymy / references
  Mr Major arrived in France today. The prime minister will meet the President tomorrow. The Conservative leader will then travel to Moscow where he will meet Mr Gorbachev. Mrs Major will join her husband in Russian, where this son of a circus artist is a relative unknown figure.

14

## Selecting the Right Term

In every case, two people favored the same term with probability < 0.20" [Furnas *et al*. CACM, 1997, p. 964]

Test1: Prob. two persons gives the same term

Test2: Prob. one person gives the most frequently used term

Test3: Prob. one person gives one of the three terms given by another

| #objects | Editor 5 | Editor 25 | Objects 50 | Group 64 |
|---|---|---|---|---|
| Test1 | 0.07 | 0.11 | 0.12 | 0.14 |
| Test2 | 0.15 | 0.21 | 0.45 | 0.52 |
| Test3 | 0.21 | 0.30 | 0.28 | 0.34 |

15

## What is this About?

6 x cubains

5 x nombre, floride, côtes

4 x réfugiés

3 x parvenus

2 x garde, atteint, année, pays

1 x utilisées, unis, gros, années, économie, américaine, américains, tendance, embarcations, éclatement, bateaux, indiqué, responsable, importante, dégradation, légalement, décédés, record, voyage, frêles, jan, mer, illégalement, résidence, agit, pratiquement, cubaine, augmentation, important, titre, fuyant, fui, miami, jamais, furent, whitlock, embarquer, afp, ats, atteignant, bateau, solides, connu, union, er, samedi, américaines, dernière, chris, etats, loi, observateurs, obtenir, passées, exode, présent, soviétique, entraîné, remarqué

16

## The Original Text

<DOCNO> ATS.940101.0004
<KW> etats-unis refugies cubains nombre record
<TI> Nombre record de réfugiés cubains parvenus en Floride en 1993.
<LD> Miami, 1er jan (ats/afp) Plus de 3500 réfugiés cubains sont parvenus sur les côtes de Floride en 1993, un nombre jamais atteint depuis 1980, ont indiqué samedi les garde-côtes américains. L'année dernière, 3656 Cubains ont atteint les côtes de Floride en bateau, soit 43% de plus qu'en 1992, année durant laquelle ils furent au nombre de 2557, selon Chris Whitlock, un responsable des garde-côtes. Le nombre de réfugiés décédés durant le voyage n'est pas connu.
<TX> Il s'agit du plus important exode depuis que 125 000 Cubains étaient parvenus en Floride après avoir fui leur pays par la mer en 1980. Les observateurs en Floride ont remarqué que les réfugiés avaient tendance à présent à s'embarquer sur des bateaux plus gros et plus solides que les frêles embarcations utilisées les années passées.
<TX> Pratiquement tous les Cubains atteignant légalement ou illégalement les côtes américaines peuvent obtenir un titre de résidence aux Etats-Unis, selon la loi américaine. Le nombre de Cubains fuyant leur pays est en augmentation depuis que l'éclatement de l'Union Soviétique a entraîné une importante dégradation de l'économie cubaine.

17

## The Point?

- Basis of most IR is a very simple approach
  - find words in documents
  - compare them to words in a query
  - this approach is very effective!
- Other types of features are often used
  - phrases
  - named entities (people, locations, organizations)
  - special features (chemical names, product names)
    - difficult to do in general; usually require hand building
- Focus of research is on improving accuracy, speed
- …and on extending ideas elsewhere

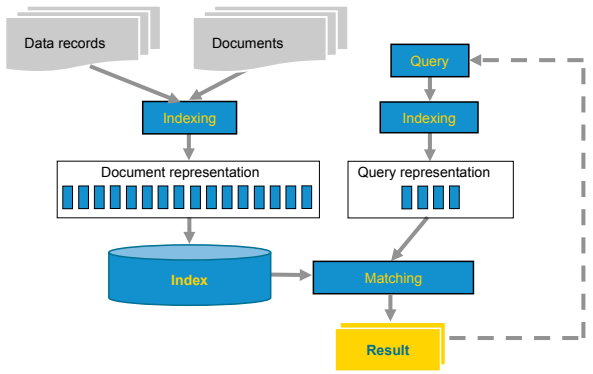
18

## Outline

- What is Information Retrieval (IR)?
- Core idea of IR-related work
- **Basic IR process**
- Simple model of IR
- The Web
- Conclusion

19

## IR "Flow"




20

## Some Issues that Arise in IR

- Text representation (indexing)

  Given a text document, identify the concepts that describe the content and how well they describe it
  - what makes a “good” representation? (surface words, NLP)
  - how is a representation generated from text?
  - what are retrievable objects and how are they organized?
- Representing information needs (query formulation)

  Describe and refine information needs as explicit queries
  - what is an appropriate query language?
  - how can interactive query formulation and refinement be supported? (e.g., interface does not always encourage query acquisition).

21

## Some Issues that Arise in IR

- Comparing representations (retrieval)

  Compare text and information need representations to determine which documents are likely to be relevant
  - what is a “good” model of retrieval?
  - how is uncertainty represented?
- Evaluating effectiveness of retrieval

  Present documents for user evaluation and modify query based on feedback
  - what are good metrics?
  - what constitutes a good experimental test bed?
  - learning schemes

22

## The Retrieval Problem

- Problems:
  - mismatch between document and query due to language ambiguity (synonym, homonym, paraphrasing, metaphor, word forms, typo)
  - mismatch between document and query due to incomplete understanding of problem ("garbage in, garbage out")
  - noisy document collection (OCR)
  - misleading content (spam etc.)
  - authority, source, actuality, copyright
  - conflicting goals: maximizing relevant information vs. minimizing irrelevant information
  - relevance is subjective and context-dependent

23

## Outline

- What is Information Retrieval (IR)?
- Core idea of IR-related work
- Basic IR process
- **Simple model of IR**
- The Web
- Conclusion

24

## Vector-Space Model

- Document can be represented by a set of (weighted) keywords
- Topic can be represented using the same formalism
- Indexing is the process to select / extract the most appropriate keywords
- Automatic indexing:
  - Step 1: Select, format, coding
  - Step 2: Granularity (XML)
  - Step 3: Tokenization (segmentation)
  - Step 4: Stopword removal, normalization
  - Step 5: Stemming

25

## Step 3: Tokenization

- What is a word / token? Sequence of letters?
  - IBM360, IBM-360, ibm 360, …
  - Richard *Brown* vs *brown* paint vs *Brown* is the
  - Database system, data base system, data-base system
- FR: “porte-clefs” (key ring) "chemin de fer" (railway)
- DE: “Bundesbankpräsident” =
  “Bund” + es + “Bank” + “Präsident”
- ZH: 我不是中国人 = 我 (I) 不 (not) 是 (be)中国人 (Chinese)

26

## Step 3: Tokenization

- Language independent approach
  *n*-gram indexing [McNamee & Mayfield 2004], [McNamee 2008]
  - different forms possible
    “The White House”
    → “The “, “he W”, “h Wh”, “ Whi”, “Whit”, “hite”, …
    or
    → “the“, “whit”, “hite”, “hous”, “ouse”
  - usually presents an effective approach when facing with new and less known language
  - a classical indexing strategy for JA, ZH or KR
  - trunc-*n*, consider only the first *n* letters
    compute → “compu“

27

## Step 4: Stopword Removal

- Stopword list for the English language
  - No clear and precise decision rule
  - Intelligent matching between query & document terms
  - Reduce the size of the inverted file (30% to 50%)
  - The SMART system suggests 571 words
    (e.g., "a", "all", "are", "back", "your", "yourself", "years"…)
  - The DIALOG system suggests 9 terms
    ("an", "and", "by", "for", "from", "of", "the", "to", "with")
    due to problem with query "vitamin a" or "IT engineer"
  - WIN system (TLR, Thomson Legal & Regulatory, now Thomson Reuters) uses one term ("the")

28

## Step 5: Stemming

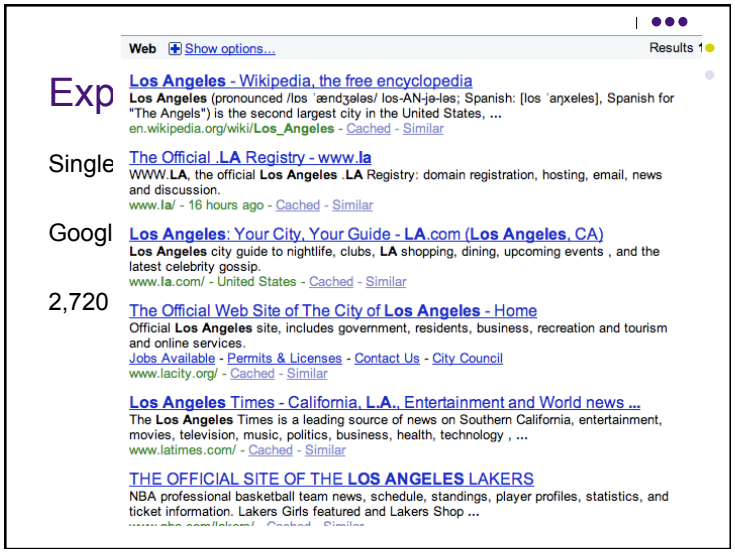
- Stemming
  - matching between documents and queries based on word sense instead of exact match (e.g, "cats" in a document, "cat" in the query)
  - automatic removal of suffixes (stemming)
  - inflectional (number, gender, case)
    "horses" → "horse"
    "actress" → "actor"
    "rosarum" → "rosa"
  - derivational (from one POS to another)
    "establish" → "establishment"



## Experiments with Google

- What is the result of the query "the" sent to Google?
  - or another stopword list (for, be, in, my)?
  - using google.co.uk, google.ca, google.com.au?
- What is the result of the query "bank" vs. "banks" vs. "banking" sent to Google?
  - can you detect the rule used by Google's stemmer?
- What is the result of the query "Bank" vs. "bank" vs. "BANK" sent to Google?
- and if we use another language (French, German)?
  http://members.unine.ch/jacques.savoy/Papers/PageRank.html






## Uppercase vs. Lowercase

| t | Bank | bank | *w* |
|---|---|---|---|
| 35.02 | 1324 | 24 | Gaza |
| 34.03 | 1301 | 36 | Palestinian |
| 33.60 | 1316 | 48 | Israeli |
| 33.18 | 1206 | 26 | Strip |

| t | Bank | bank | *w* |
|---|---|---|---|
| -10.93 | 900 | 1161 | money |
| -10.43 | 624 | 859 | federal |
| -9.59 | 586 | 786 | company |
| -8.47 | 282 | 430 | accounts |

33

## Vector-Space Model

- Indexing weights for term $t_k$ in document $D_i$
  1. frequent terms must a have more weight: $tf_{ik}$
  2. words occurring in less documents (having a greater discrimination power) must have larger weight:
     $idf_k = \log(n/df_k)$ with $n$ = # documents
  3. increase weights for smaller documents
  - the overall formula
    $w_{ik} \approx tf_{ik} \cdot idf_k$
  - many variations possible
    $w_{ik} \approx (\log(tf_{ik})+1) \cdot idf_k$

34

## Example: Small Documents

$D_1$ = "a horse, a horse, my kingdom for a horse".
$D_2$ = "food for cats and dogs".
$D_3$ = "my small horse, but it is a horse".
$D_1$ = {horse 3, kingdom 1}.
$D_2$ = {cat 1, dog 1, food 1}.
$D_3$ = {horse 2, small 1}.
How to store these values (to be effective)?

A "topic": Q = "Food for horses"
Q = {horse 1, food 1}.

35

## Inverted File Organization

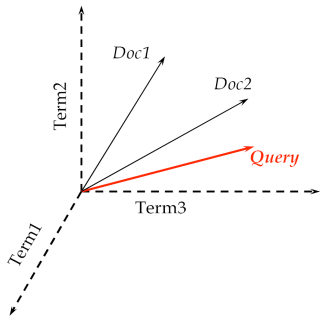
| | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|
| horse | 3 | | 2 |
| cat | | 1 | |
| kingdom | 1 | | |
| dog | | 1 | |
| small | | | 1 |
| food | | 1 | |

Q =
horse ={$D_1$, 3; $D_3$, 2}
food = {$D_2$, 1}

36

## Vector-Space Model

In general, we can view documents and the query as vector in a *t* dimensional space (*t* = # indexing terms)




## Comparison

- Documents are vectors
- Topic is represented by a vector
- Compare item by item and when the same item is present both in the document and in the query, increase the similarity between the corresponding document and the query (intern product, with $w_{ij}$ = term $t_k$ and document $d_j$ and $w_{qk}$ = weight of term $t_k$ in the query)

$$sim(Q, D_i) = \sum_{k=1}^{t} w_{ij} \cdot w_{qj}$$

38

## Inverted File Organization

Inverted file
(index of a book)

| | |
|---|---|
| horse | {$D_1$, 3; $D_3$, 2} |
| cat | {$D_2$, 1} |
| kingdom | {$D_1$, 1} |
| dog | {$D_2$, 1} |
| small | {$D_3$, 1} |
| food | {$D_2$, 1} |

Q = “Food for horses”

horse = {$D_1$, 3; $D_3$, 2}
food = {$D_2$, 1}

Similarity

$D_1 = 3 \cdot 1 = 3$
$D_2 = 1 \cdot 1 = 1$
$D_3 = 2 \cdot 1 = 2$

39

## Comparison

- Or compute the cosine of the angle between the document vector and the query vector or used another similarity measure

Cosine

$$sim(Q, D_i) = \frac{|D_i \cap Q|}{|D_i|^{0.5} \; |Q|^{0.5}} = \frac{\sum_{k=1}^{t} w_{ik} \cdot w_{qk}}{\sqrt{\sum_{k=1}^{t} w_{ik}^2} \cdot \sqrt{\sum_{k=1}^{t} w_{qk}^2}}$$

Dice

$$sim(Q, D_i) = \frac{|D_i \cap Q|}{|D_i \cup Q|} = \frac{2 \cdot \sum_{k=1}^{t} w_{ik} \cdot w_{qk}}{\sum_{k=1}^{t} w_{ik}^2 + \sum_{k=1}^{t} w_{qk}^2}$$

40

## Vector-Space Model

- Problem
  - unigram approach: the fact that a given term occur does not imply that another term has more (or less) chance to co-occur (e.g, "algorithm" and "computer")
  - not clear how to define/weight noun phrase ("sort algorithm", "operating system")
  - various similarity measures
  - baseline system, not the most effective (Boolean, probabilistic, language model, logic-based, …)
  - knowing some relevant document may help the system

41

## Empirical Evidence

- Test-collection (TREC, CLEF, NTCIR, INEX, FIRE)
  - a set of "documents" (article, image, interview, video)
  - a set of topics
  - the relevance information for each topic
- Various subjects / several languages
- Measure by
  - precision (# relevant items / # retrieved items)
  - recall (# relevant items / # relevant items)
  - precision at 10 docs (P@10): precision after retrieving the first 10 docs.
- User interface is important (essential?)

42

## Average Precision (One Query)

| Rank | System A | | System B | |
|---|---|---|---|---|
| 1 | R | 1/1 | nR | |
| 2 | R | 2/2 | R | 1/2 |
| 3 | nR | | R | 2/3 |
| … | nR | | nR | |
| 35 | nR | | R | 3/35 |
| … | nR | | nR | |
| 108 | R | 3/108 | nR | |
| | AP = | 0.6759 | AP = | 0.4175 |
| | | | | -38.2% |

43

For both systems, P10 = 2/10 = 0.2

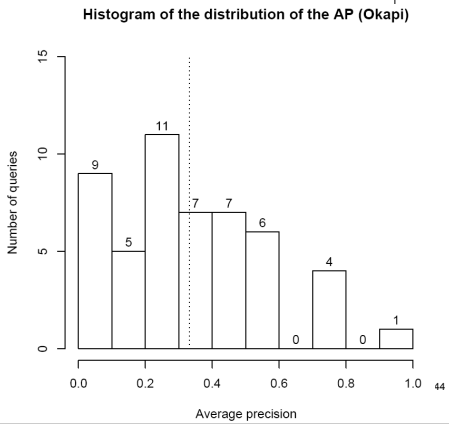
## Mean Average Precision (MAP)

A single value
MAP: 0.3321
or an histogram?

Here, for one query, the perfect answer

For 9 queries, Okapi "fails"

(ZH, NTCIR-5, indexing unigram & bigram)




44

## Why IR System May Fail

- Spelling error
  «Innondationeurs en Hollande et en Allemagne»
- Stopword list ("ai" in French)
  «AI en Amérique latine» or «IT engineer»
- Stemming ("parlement" ≠ "parlementaires")
  «Elections parlementaires européennes»
- Missing specificity
  «World Soccer Championship»
- Cannot discriminate between relevant and non-relevant
  «Chinese currency devaluation»
- Language use
  «telephone portable» but "natel", "cellulaire"

45

## Outline

- What is Information Retrieval (IR)?
- Core idea of IR-related work
- Basic IR process
- Simple model of IR
- **The Web**
- Conclusion

46

## The Web

- Information explosion
- Magnetic memory is larger than paper
  - 327 TB for paper vs. 3,416,230 TB for magnetic
- These values are increasing
  - The surface web is 17x larger than the Library of Congress
- New phenomena
  - blog (blogcount.com)
  - - P2P (peer to peer file sharing, 5,000 TB (mainly video (59%) and audio (33%)) with 3 M of active users)
- A real challenge for CS and other fields!

47

## The Web

**Market share**
(July 2005, Nielsen//NetRating)

| | |
|---|---|
| Google | 46.2% |
| Yahoo | 22.5% |
| MSN | 12.6% |
| AOL | 5.4% |
| MyWay | 2.2% |
| Ask | 1.6% |
| NetScape | 1.6% |
| Others | 7.9% |

March 2007
http://www.comscore.com/press/release.asp?press=1219)

| | |
|---|---|
| Google | 48.3% |
| Yahoo | 27.5% |
| MSN | 10.9% |
| Ask | 5.2% |
| AOL | 5.0% |

48

## The Web

- Various task-specific search engines
  - General
  - News
  - Shopping
  - For Kids
  - Specialty (medical, gov, legal, QA, travel)
    - Images/ audio / video
    - Metasearch (metacrawler)
    - Country-specific
    - Specific SE for your web site (product)
    - Enterprise search (web + emails + memos + …)

49

## The Web: Query Type

- Informational – want to learn about something (~40%)
  e.g. “low hemoglobin”
- Navigational – want to go to that page (~25%)
  e.g. ”CFF”
- Transactional – want to do something (web-mediated) (~35%)
  Access a service e.g., “Geneva weather”
  Downloads e.g., “Mars surface images”
  Shop e.g., “iTunes”
- Gray areas
  Find a good hub e.g., “car rental seattle”
  Exploratory search “see what’s there”

50

## The Web

Examples
- What is the melting point of lead?
- Origins of conflict in Palestine
- Barack Obama
- SIGIR’2010 online registration
- INRT journal author instructions
- Computer Science Department
- Vincent Cerf
- Official information about abortion
- Sharks Attacks in Australia

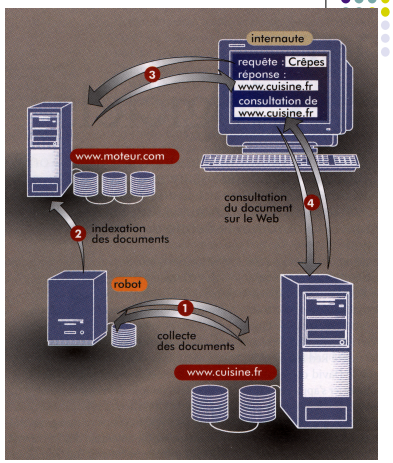
Type
- Q/A (fact)
- Topic relevance
- News search
- Online service
- Known item search
- HomePage finding
- Recall-oriented
- Restricted doc. type
- Geo IR

51

## The Web

A search engine on the Web is not only a IR system (may be this is the smallest part)

1. spider
2. indexer
3. query processor



## The Web

1. Spider (crawler or robot) -- builds the corpus
   - Collects the data recursively
     - For each known URL, fetch the page, parse it, and extract new URLs
     - Repeat
   - Additional data from direct submissions & various other sources
   - Various search engines have different policies -- little correlation among corpora



## The Web

2. The indexer -- processes the data & represents it (inverted files)
   - Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc
3. Query processor -- accepts queries and returns answers
   - Front end -- does query reformulation -- word stemming, capitalization rules, optimization of Booleans, compounds, etc
   - Back end -- finds matching documents and ranks them

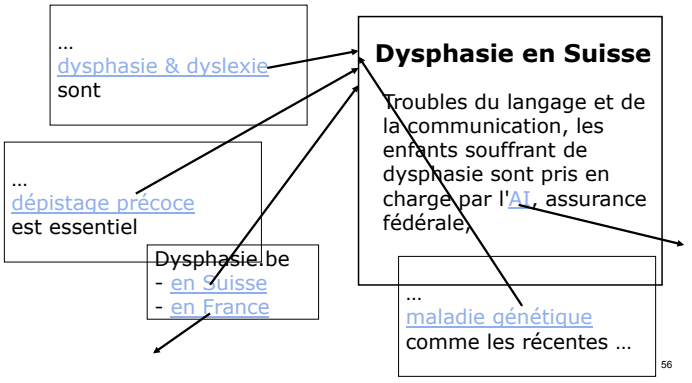


## The Web

- First generation -- use only "on page", text data
  - Word frequency, language
  - AltaVista, Lycos, Excite
- Second generation -- use off-page, web-specific data
  - - Link (or connectivity) analysis
  - - Click-through data (What results people click on)
  - Anchor-text (How people refer to this page)
  - Google (1998) with PageRank
- Third generation -- answer "the need behind the query" (still experimental)



## The Web



## PageRank (Google)

- How can we use hyperlink in IR?
- Initially the surfer is at a random page
  - At each step, the surfer proceeds to a randomly chosen web page with probability *(1-d)* (e.g., probability of a random jump = 0.15)
  - or to a randomly chosen successor of the current page with probability *d* (e.g., probability of following a random outlink = 0.85)
- PageRank of a page = Probability that the surfer is at the page on a given time step

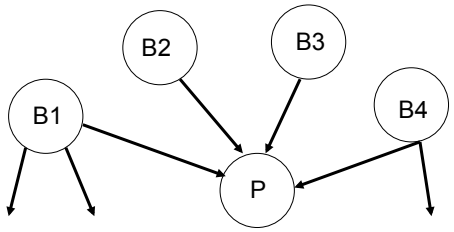
Brin S., Page L., The anatomy of a large-scale hypertextual web search engine, *Proceedings of the WWW7*, Amsterdam, Elsevier, 107-117, 1998.

57

## PageRank (Google)

- Extend inductively:

  Quality of P: Q(P) = Q(B1)/3 + Q(B2) + Q(B3) + Q(B4)/2




58

## Random Surfer Model

- Formally

$$PR^{c+1}(D_i) = (1-d)\frac{1}{n} + d\left[\frac{PR^c(D_1)}{C(D_1)} + \ldots + \frac{PR^c(D_m)}{C(D_m)}\right]$$

  $PR^c(D_i)$: PageRank value of page $D_i$ after c cycles
  $C(D_i)$: number of outlinks for page $D_i$ (outdegree)
- But to compute $PR^{c+1}(D_i)$, we need $PR^c(D_j)$
  We do it iteratively (usually 5 iterations is enough)
- In the first step, the IR system retrieves the web pages (with the corresponding document score). Use PR to rank retrieved page

59

## PageRank (PR)

- Use a web site to compute the PagRank (PR) of various web sites.
- Which firms owns the highest PR values (defined between 0 and 10)?
- Do you think that PR is biased?
- TREC experiments have shown that PR alone does not produce a high MAP (compared to classical IR models). PR is simply *one* component in Google's ranking function (but not the only one).

60

## Evaluation (WT2g)

WT2g (100 queries in TREC-8 and 9)

| IR system | MAP |
|---|---|
| Okapi | 0.2668 |
| *tf idf* | 0.1385 |
| Okapi + links | 0.0874 |
| *tf idf* + links | 0.0682 |

J. Savoy, J. Picard: Retrieval effectiveness on the Web. *Information Processing & Management,* 2001, 37(4), 543-569



## Bookstores

- Does the website design have an impact on retrieval ranking?
- Study:
  - 38 bookstores (online) from directories (*Yahoo*! & *Google*)
  - 206 unique books (best-sellers New York Times, Sept 2002)
  - 4 search engines (*Google*, *Fast*, *MSN*, *AltaVista*)
- Correct answer: a transactional page with the book ISBN

Upstill T., Craswell N., Hawking D., Buying bestsellers online: A case study in search & searchability, *Proceedings of 7th Australasian Document Computing Symponium (ADCS)*, Sydney, 2002.



## Bookstores

- In the search results
  - Only 14 bookstores (over 38) returns any correct answer within the top 1000
  - Limited to the top 10, only 4 bookstores remain
  - *Amazon* was the most searchable
  - Only *Amazon* had correct results returned by every search engine
  - *Barnes & Noble* performed well on *Google* and *Fast*
  - *Walmart* performed well on *MSN*
  - Only *Fast* returns results from many bookstores



## Bookstores

| Bookstore | S@1 | S@5 | S@10 | S@100 | S@1000 | S@1000 breakdown (AV:FA:GO:MS) | Hostname Results |
|---|---|---|---|---|---|---|---|
| Amazon | 0.124 | 0.325 | 0.402 | 0.492 | 0.584 | 104:83:162:132 | 3903 |
| Barnes and Noble | 0.028 | 0.096 | 0.140 | 0.225 | 0.316 | 0:87:170:3 | 3603 |
| Walmart | 0.010 | 0.030 | 0.045 | 0.070 | 0.075 | 2:0:0:60 | 277 |
| BookSite | 0.000 | 0.004 | 0.005 | 0.013 | 0.013 | 0:0:0:11 | 52 |
| ecampus | 0.0 | 0.0 | 0.0 | 0.005 | 0.012 | 0:7:0:3 | 290 |
| AllDirect | 0.0 | 0.0 | 0.0 | 0.002 | 0.005 | 0:4:0:0 | 52 |
| NetstoreUSA | 0.0 | 0.0 | 0.0 | 0.001 | 0.010 | 0:8:0:0 | 261 |
| Sam Weller's Books | 0.0 | 0.0 | 0.0 | 0.001 | 0.006 | 0:5:0:0 | 22 |
| Books-A-Million | 0.0 | 0.0 | 0.0 | 0.0 | 0.008 | 0:4:0:3 | 775 |
| 1BookStreet | 0.0 | 0.0 | 0.0 | 0.0 | 0.006 | 0:5:0:0 | 17 |
| Wordsworth.com | 0.0 | 0.0 | 0.0 | 0.0 | 0.004 | 1:0:1:1 | 92 |
| TextbookX.com | 0.0 | 0.0 | 0.0 | 0.0 | 0.002 | 0:2:0:0 | 22 |
| CodysBooks.com | 0.0 | 0.0 | 0.0 | 0.0 | 0.002 | 0:2:0:0 | 78 |
| Arthurs Books | 0.0 | 0.0 | 0.0 | 0.0 | 0.003 | 0:1:0:0 | 3 |
| Powells Bookstore | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0:0:0:0 | 1031 |

## Bookstore

- Search engine comparison
- The best SE is not always the same

| Search engine | P@1 | P@5 | P@10 | P@100 |
|---|---|---|---|---|
| AltaVista | 0.14 | 0.39 | 0.45 | 0.50 |
| Fast | 0.00 | 0.02 | 0.05 | 0.18 |
| Google | 0.15 | 0.56 | **0.67** | **0.83** |
| MSN | **0.36** | **0.57** | 0.65 | 0.73 |

65

## Bookstores

- The coverage of bookstore by SE
  - To be retrieved, a page must be found by the crawler!
  - *Amazon* had a larger coverage by all SEs
  - The coverage of *Barnes & Noble* varies widely
  - A large number of pages of *Walmart* were covered by *MSN*
  - *Fast* did not have a large coverage of any one bookstore
  - *AltaVista* had a large coverage only of *Amazon*
- The link coverage is important (for the crawler and during ranking). Based on the number of links to an entire domain, *Amazon* appears in the first position

66

## Conclusion

- Information Retrieval?
  - Indexing, retrieving, and organizing text by probabilistic or statistical techniques that reflect semantics without actually understanding
- Core idea
  - Bag of words captures much of the "meaning"
  - Objects that use vocabulary the same way are related
- Vector-Space model
  - Documents and queries are vectors
  - Various similarity measures
- Web
  - Huge, less structured, various media/languages
  - Link analysis help

67

## Evaluation

Using TREC-2003 (WebTrack), 50 queries

| Model | Prec@5 | Prec@10 |
|---|---|---|
| IR Model | 16.00 | 11.60 |
| HITS, *hub*, $\sigma$=50 | 3.60 | 2.60 |
| HITS, *authority*, $\sigma$=50 | 0.80 | 0.60 |
| PageRank, d=0.85 | 2.00 | 1.60 |

J. Savoy, Y. Rasolofo: Hyperliens et recherche d'information sur le Web. Proceedings JADT 2004, 1000-1007, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/jadt2004-th.htm

68