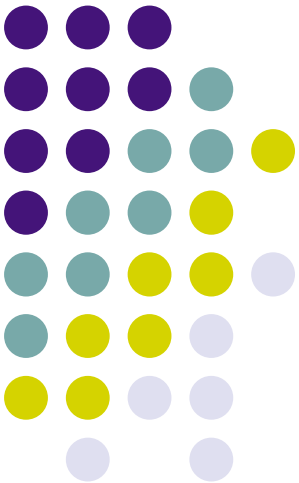
# Search Engines Technology (IR)

J. Savoy
Université de Neuchatel

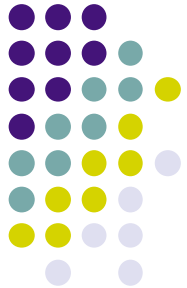


Manning C.D., Raghavan, & P, Schütze, H. *Introduction to information retrieval*. Cambridge University Press, Cambridge (UK), 2008.

W.B. Croft, H. Turtle: Introduction to Information Retrieval. Spring course, 1997.

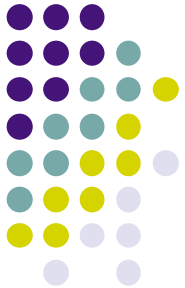
J. Allen: Information Retrieval Course. University of Massachusetts at Amherst, 2004.

# What is Information Retrieval (IR)?



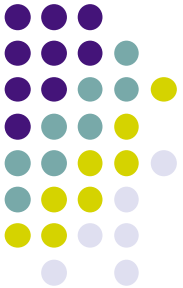
- How to build a search engine?
- How to evaluate IR?
- How to include NLP facets into IR engine?
- How search engines work on the Web?
- … and others answers

# IR domains



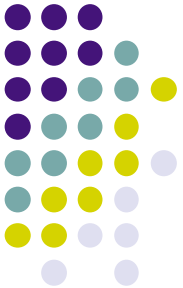
- What makes a system like Google or Yahoo! Search tick?
  - How does it gather information? What tricks does it use?
- How can those approaches be made better?
  - Natural language understanding (NLU)?
  - User interactions?
- How do we decide whether it works well?
  - For all queries? For special 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)?**
- Core idea of IR-related work
- Basic IR process
- Simple model of IR
- The Web
- Conclusion

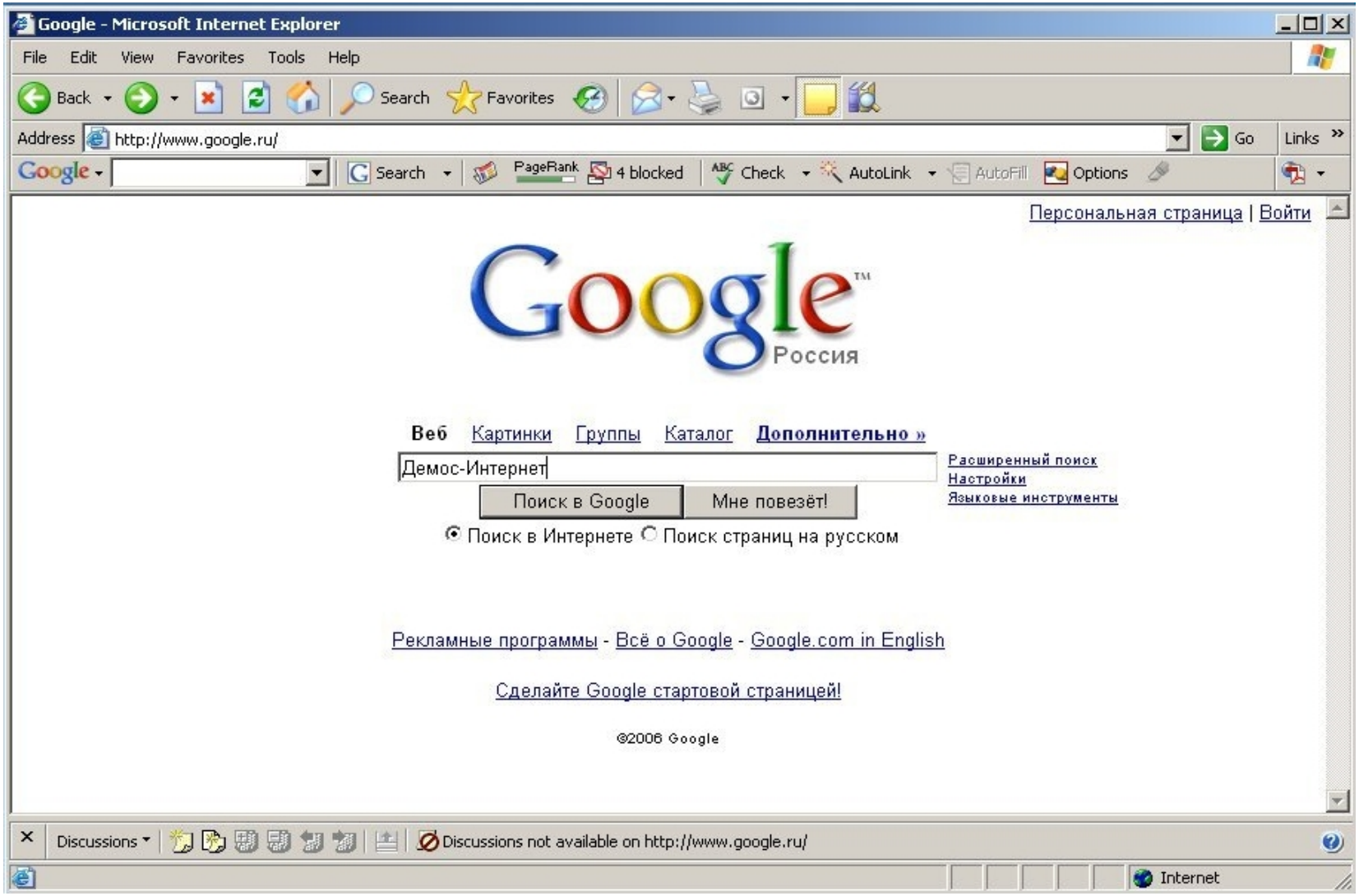
# Definition



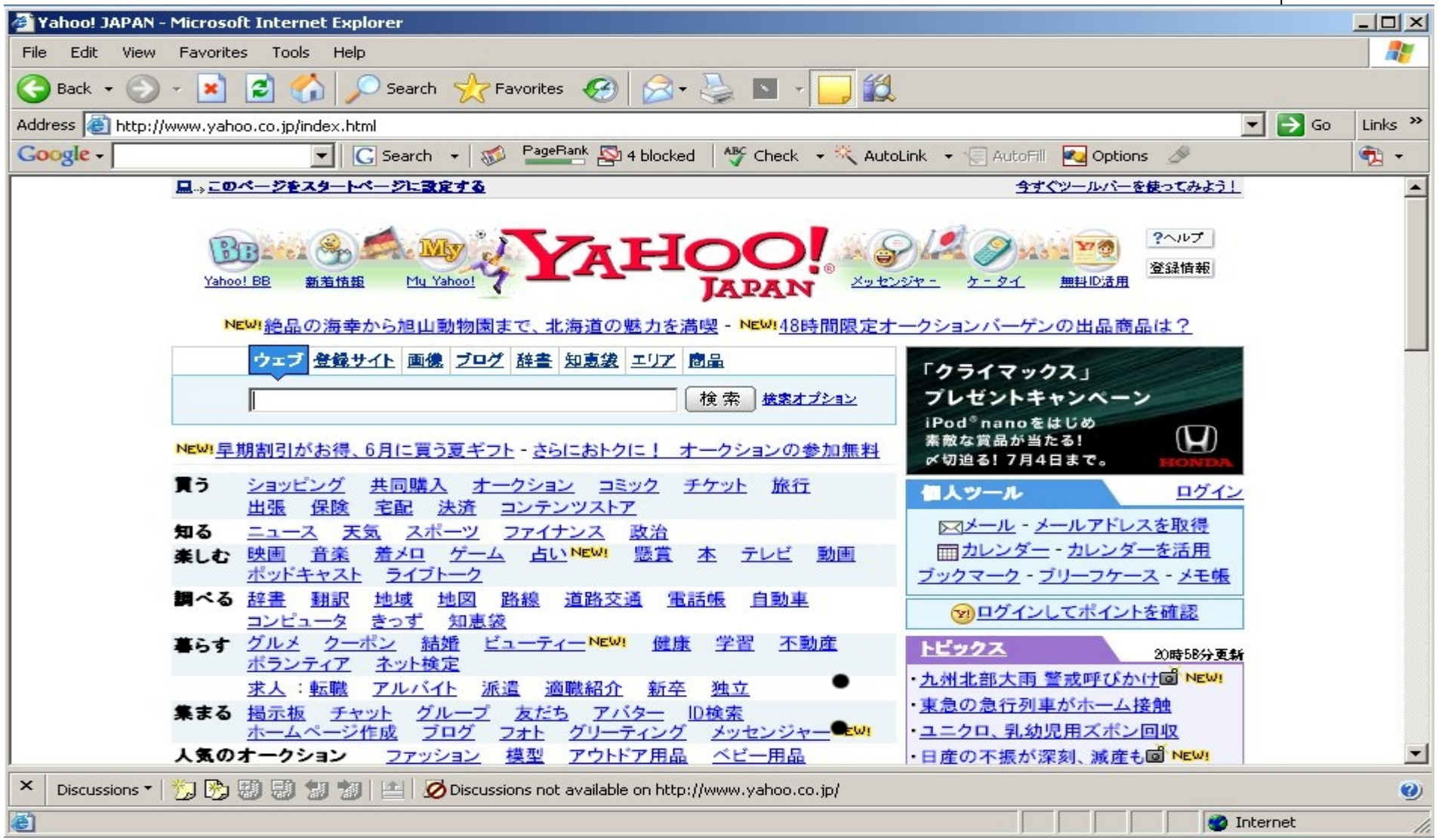
Information retrieval deals with the *representation*, *storage*, *organization* of, and *access* to information items. These information items could be references to real documents, documents themselves or even single paragraphs, as well as web pages, spoken documents, images, pictures, music, video,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

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

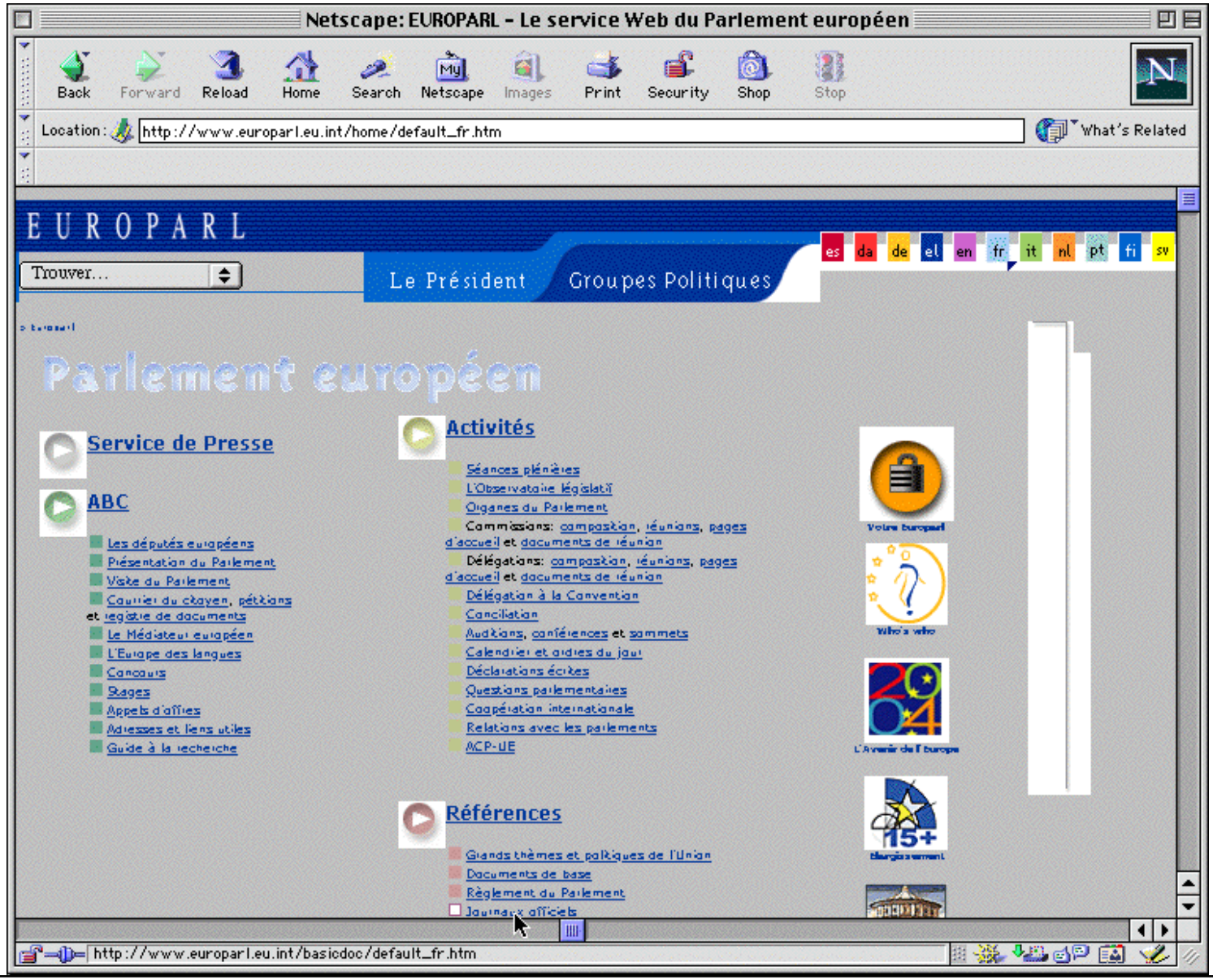
# What is Information Retrieval



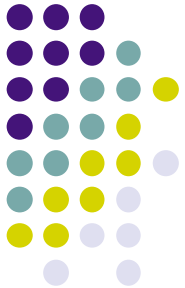
# What is Information Retrieval



# What is Information Retrieval



# What is Information Retrieval

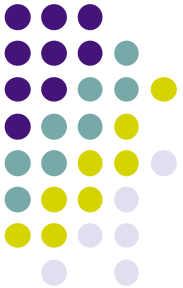






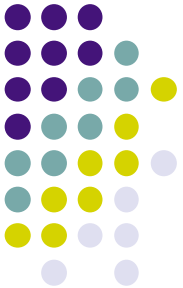
- Quite effective (at some things)
- Highly visible (mostly)
- Commercial successful (some of them, so far)
- What is behind the scene?
  - How do they work?
  - Is there more to it than the Web?

# Sample systems



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

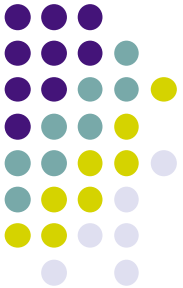
# Evolution



- 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/

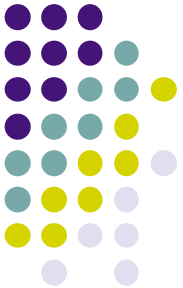
# Searching with databases



- The largest information systems around the world are DB
- Do the same:  use DB
  - Use the relational model
  - “Easy” to define tables
  - Easy to search into tables
  - Effective tool
- Methodologies available, multiple examples

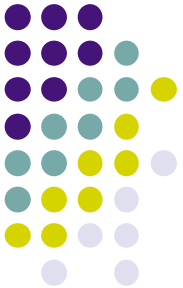
| ID | Name | Address | Book |
|---|---|---|---|
| 1253 | Tintin | Moulinsart 10 | L3 |
| 2345 | Tournesol | Liberty  3 | L5 |
| 345 | Dupont | Central 6a | |
| 674 | Dupond | Central 6b | L13 |

# Searching with databases



| ID | Author | Editor | Title | Year | Pages |
|---|---|---|---|---|---|
| L1 | Blair | Elsevier | Language and representation in information retrieval | 1990 | 335 |
| L2 | Agosti | Kluwer | Information retrieval and hypertext | 1996 | 278 |
| L3 | Salton | Hermes | Automatic text processing | 1989 | 356 |
| L4 | Rijsbergen | Addison | Information retrieval | 1979 | 208 |
| L5 | Harter | Academic | Online information retrieval | 1986 | 256 |

# Searching with databases



And the query about the content

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

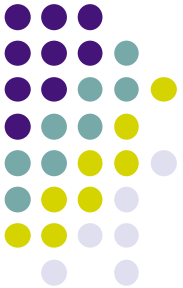
| Name | Title | Year |
|---|---|---|
| Rijsbergen | Information retrieval | 1989 |

Do we solve the problem?

# Comparing IR to databases

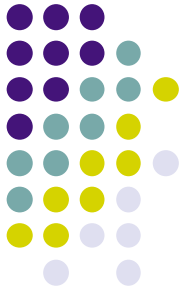
| | Database | IR |
|---|---|---|
| Data | structured | unstructured |
| Fields | Clear semantics (domain) | No fields (other than text) |
| Model | Determinist | Probabilistic |
| Queries | Defined (SQL, relational algebra), complex, complete specification | Free text (NL) flat, Boolean, partial |
|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

# Outline



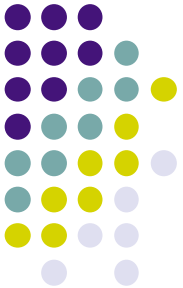
- What is Information Retrieval (IR)?
- **Core idea of IR-related work**
- Basic IR process
- Simple model of IR
- The Web
- Conclusion

# Basic approach to IR



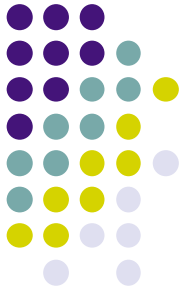
- Most successful approaches are statistical
  - Directly, or an effort to capture and use probabilities
- Why not natural language understanding?
  - i.e., computer understands documents and query and matches them
  - State of the art is brittle in unrestricted domains
  - Can be highly successful in predictable settings
    - e.g., information extraction on terrorism/takeovers (MUC)
    - Medical or legal settings with restricted vocabulary

# Basic approach to IR



- Could use manually assigned headings
  - e.g., Library of Congress (LOC) headings
    Dewey Decimal headings
  - Human agreement is not good
  - Hard to predict what headings are “interesting”
  - Expensive

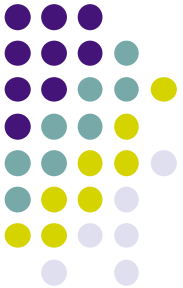
# Relevant items are similar



- 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
  - Probability that documents arise from same model
  - Same meaning of text
- and Natural Language Processing (NLP)?

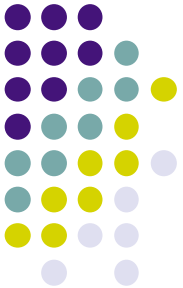
# Example of NLP



- Polysemy
  Same words → different meanings
  Only one sense in Java? Windows?
  He saw a man with a saw. (homographs)

- 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.

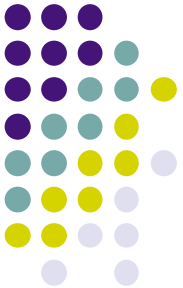
# Example of NLP



- Spelling errors

- Why NLP could be difficult
  - "*Flying* planes could be dangerous"
  - "He saw the girl in the park *with* the telescope"
  - "He eats a fish with a fork"<br>"He eats a fish with a bone"<br>"The ink is in the pen"<br>"The pig is in the pen"
  - "John put the book on the table *in* his pocket" (he put the book or the table?)
  - "He saw her *shaking* hands"

# Selecting the right term



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

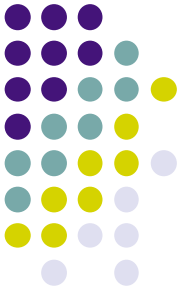
Test1: Prob. two persons gives the same term

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

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

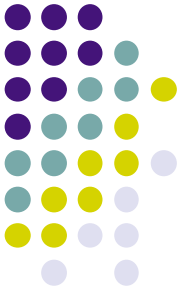
| #objects | Editor 5 | Editor 25 | Objects 50 | Group 64 |
|---|---|---|---|---|
| Test1 | 0.07 | 0.11 | 0.12 | 0.14 |
| Test2 | 0.15 | 0.21 | 0.45 | 0.52 |
| Test3 | 0.21 | 0.30 | 0.28 | 0.34 |

# “Bag-of-Words”



- An effective and popular approach
  {Mary, packet, Montreal, Paris, sent}
- Compares words without regard to *order*
  “Mary sent the packet from Montreal to Paris”
  “Mary sent to Paris from Montreal the packet”
  “Mary sent from Montreal the packet to Paris”

# 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

# The original text

<DOCNO> ATS.940101.0004

<KW>  etats-unis refugies cubains nombre record

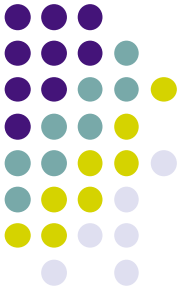
<TI>  Nombre record de réfugiés cubains parvenus en Floride en 1993.

<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.

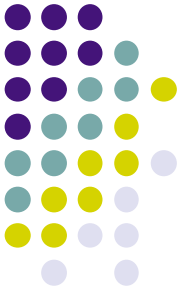
<TX> Pratiquement tous les Cubains atteignant légalement ou illégalement les côtes américaines peuvent obtenir un titre de résidence aux Etats-Unis, selon la loi américaine. Le nombre de Cubains fuyant leur pays est en augmentation depuis que l'éclatement de l'Union Soviétique a entraîné une importante dégradation de l'économie cubaine.

# The point?



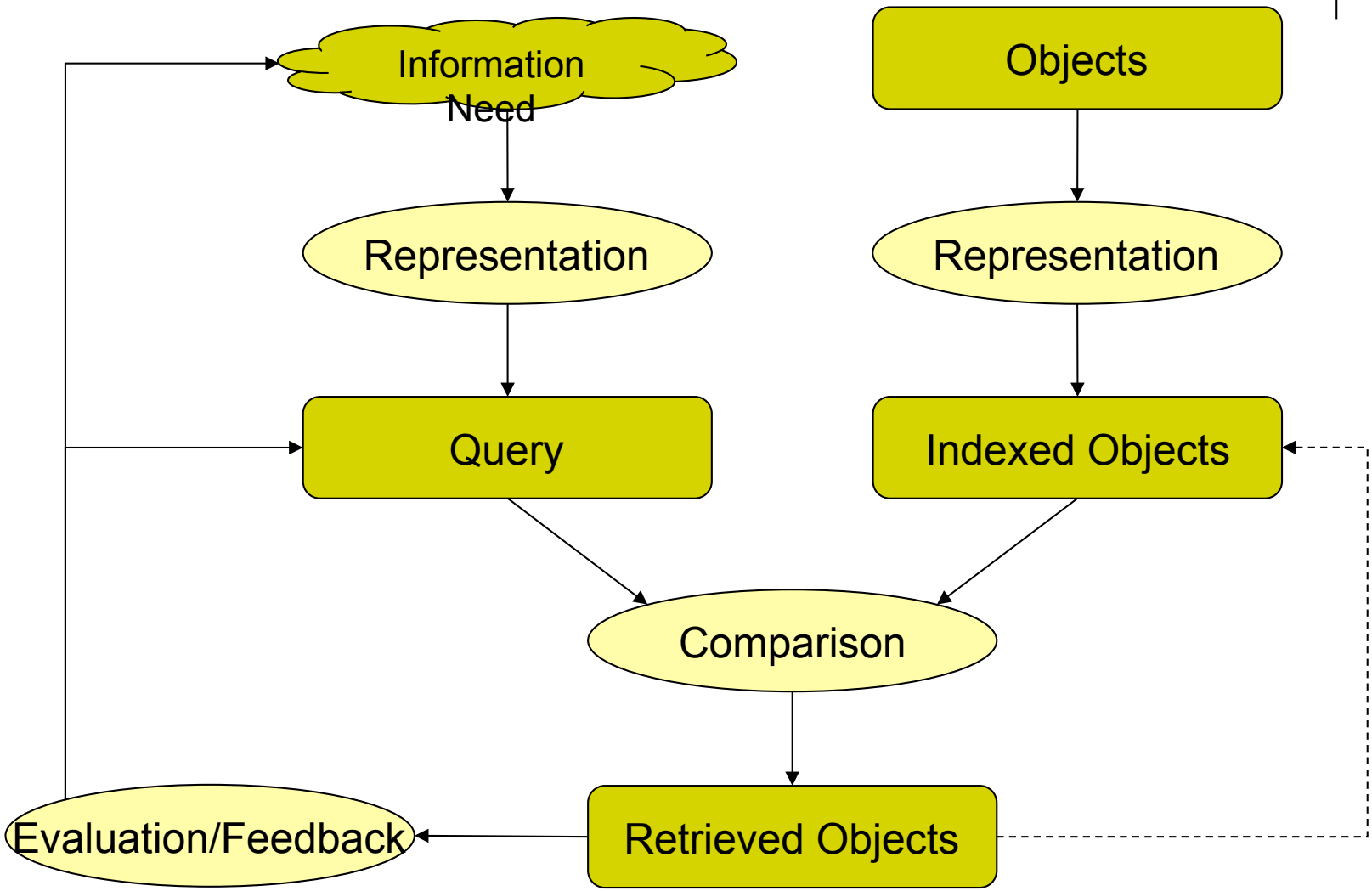
- Basis of most IR is a very simple approach
  - find words in documents
  - compare them to words in a query
  - this approach is very effective!
- Other types of features are often used
  - phrases
  - named entities (people, locations, organizations)
  - special features (chemical names, product names)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

# Outline

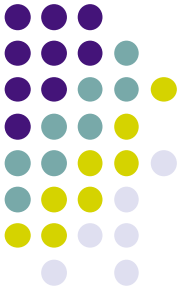


- What is Information Retrieval (IR)?
- Core idea of IR-related work
- **Basic IR process**
- Simple model of IR
- The Web
- Conclusion

# Overview of IR process



# Indexing

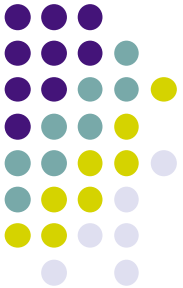


- Text representation (indexing)

  Given a text document, identify the concepts that describe the content and how well they describe it

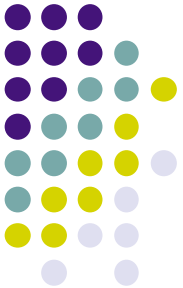
  - what makes a “good” representation? (surface words, NLP, thesaurus)
  - how is a representation generated from text?
- Manual or automatic?
  - controlled vocabulary (e.g., LOC) or free text
  - exhaustivity (all details, main topics)?
  - specificity of the vocabulary (broad terms)?
  - number of terms?

# Library of Congress Headings



**A -- GENERAL WORKS**
**B -- PHILOSOPHY. PSYCHOLOGY. RELIGION**
**C -- AUXILIARY SCIENCES OF HISTORY**
**D -- HISTORY: GENERAL AND OLD WORLD**
**E -- HISTORY: AMERICA**
**F -- HISTORY: AMERICA**
**G -- GEOGRAPHY. ANTHROPOLOGY. RECREATION**
**H -- SOCIAL SCIENCES**
**J -- POLITICAL SCIENCE**
**K – LAW**
**L – EDUCATION**
**M -- MUSIC AND BOOKS ON MUSIC**
**N -- FINE ARTS**
**…**

# Library of Congress Headings



**P -- LANGUAGE AND LITERATURE**
**Q – SCIENCE**
**R – MEDICINE**
**S – AGRICULTURE**
**T – TECHNOLOGY**
**U -- MILITARY SCIENCE**
**V -- NAVAL SCIENCE**
**Z -- BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES**

# Thesaurus

772 informatique

N. 1 **Informatique** (*l'informatique*), micro-informatique, mini-informatique ; péri-informatique ; téléinformatique. - Intelligence artificielle ou i.a. - Automation ou automatisation.

2 Automatique, **bureautique**, domotique, novotique, productique, télématique.

3 **Matériel** (*le matériel* ; opposé au *logiciel* 722.11) ; hardware [anglic.]. - **Ordinateur** (ou : calculateur numérique, computer) ; micro-ordinateur ou, fam. micro, pc (*Personal Computer*) ; mini-ordinateur ou, fam., mini ; clone ; machine [fam.], bécane [arg.].

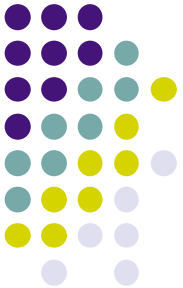
...

11 Logiciel (ou : software, soft) (opposé au matériel 772.3), **programme** ; application, microprogramme, programme enregistré, programme croisé, sous-programme (ou : procédure, routine). - Progiciel (ou : package, produit programme) ; **système d'exploitation** ou os (*Operating System*), système expert. - Menu.

...

23 **Informaticien**, ingénieur système ; analyste, analyste-programmeur, programmeur. - Dactylocodeur, **opératrice de saisie** ; perforateur vérificateur ou perfo-vérif. [anc.]. - Pupitreur. - Bureauticien ; cogniticien.

# Inter-indexer consistency?



**British Library Cataloging in Publication Data**

Gazdar, Gerald

Natural language processing in PROLOG : an introduction to computational linguistics.

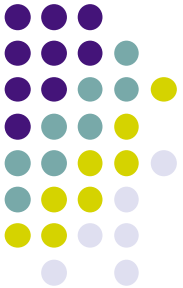
1. Natural language. Analysis. Applications of computer systems. Programming languages. Prolog

I. Title II. Mellish, C. S. (Christopher S.)*1954*--

418

ISBN 0-201-18053-7

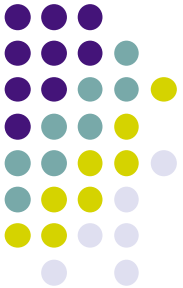
# Inter-indexer consistency?



Gazdar, Gerald.
Natural language processing in PROLOG.
Bibliography : p.
Includes index.
1. Computational linguistics. 2. Prolog (Computer program language) I. Mellish, C. S. (Christopher S.), 1954- . II. Title
P98.G38 1989 410'.28'55133 88-16667
ISBN 0-201-18053-7

- Intersection: {*Prolog*}
- The manual indexing is not the gold standard

# Automatic Indexing



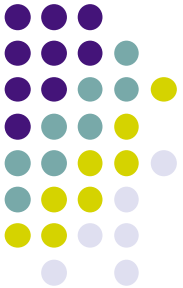
1. Parse documents to recognize structure (fields, paragraphs)
   - e.g. title, date, other fields (clear advantage to XML)
2. Scan for word tokens
   - numbers, special characters, hyphenation, capitalization, etc.
   - languages like Chinese/Japanese need *segmentation*

我不是中国人

我　不　是　中国人

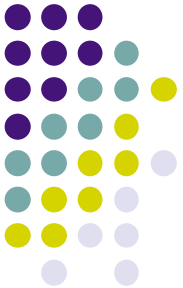
I　not　be　Chinese

# Automatic Indexing (example)



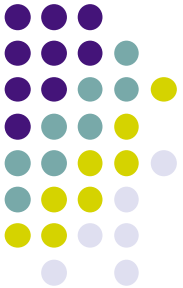
- Lowest level issue is tokenization
  - Does punctuation represent a word break?
  - “bob,alice” → bob alice  “2,103” → 2 103
  - “U.S.” → US  “umass.edu” → umass edu
- Uppercase and lowercase
  - “US President” → us president  but “She gives us”
  - “IT engineer” → it engineer but “It is for you”
- One token?
  - McDonald’s, cann’t, I’ll, you’re, O’Reilly, C|net, Micro$oft, text-based medium, New York-New Haven railroad

# Automatic Indexing



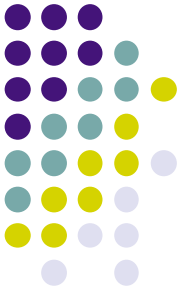
3. Stopword removal
   - based on short list of common words such as “the”, “and”, “a” (Zipf’s law)
   - saves storage overhead of very long indexes
   - can be dangerous (e.g., “The Who”, “and-or gates”, “vitamin A”)
   - available for other languages
     “le”, “des”, “dans”, “mais”, “or”
     “die”, “dem”, “in”, “für”

# Automatic Indexing (example)



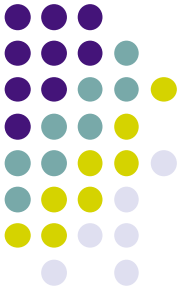
- Diacritics
  - ignore them? “cure” and “curé”, “Apfel”, “Äpfel”
- Spelling variants
  - database vs. data-base vs. data base
- compound construction
  - “Litteraturnobelpreisträger”
    “chemin de fer”
- Phrases can have an impact on both effectiveness and efficiency
  - “information retrieval” and “the retrieval of information”

# Automatic indexing



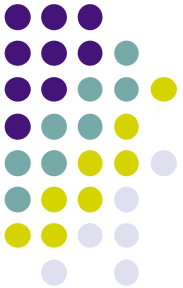
4. Stem words
   - Stemming is commonly used in IR to conflate morphological variants
   - inflections (number, gender, case)
   - derivational suffixes
- It seems reasonable that “dog” in the query match “dogs” in the document
  - can make mistakes but generally preferred
  - not done by most Web search engines (why?)

# Stemming



- Algorithmic stemmer
  - Light stemmer: removing only inflectional suffixes
    the number (sing / plural), horse, horses
    the gender (femi / masc), actress, actor
    verbal form (person, tense), jumping, jumped
    relatively simple in English (‘-s’, ‘-ing’, ‘-ed’)
  - Stemmer: removing also derivational suffixes
    forming new words (changing POS)
    ‘-ably’, ‘-ment ’, ‘-ship’
    admit → {admission, admittance, admittedly}

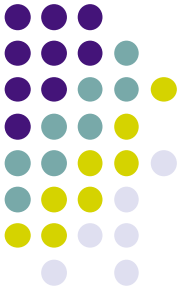
# Stemming



- Typical stemmer consists of collection of rules and/or dictionaries
- Simplest stemmer is “suffix -s” (S-stemmer)
  - If a word ends in «-ies»,
    but not «-eies» or «-aies» then replace «-ies» by «-y»;
  - If a word ends in «-es», but not «-aes», «-ees» or «-oes»
    then replace «-es» by «-e»
  - If a word ends in «-s», but not «-us» or «-ss»
    then remove the «-s»

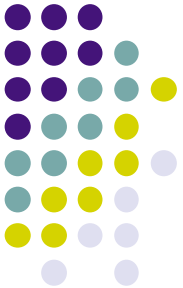
Harman, D. (1991). How effective is suffixing? *JASIS,* 42(1), 7-15

# Stemming



- Example
  - IF (" *-ing ") → remove –ing
    e.g., "king" → "k“, "running" → "runn"
  - IF (" *-ize ") → remove –ize
    e.g., "seize" → "se"
    To correct these rules:
  - IF ((" *-ing ") & (length>3)) → remove –ing
  - IF ((" *-ize ") & (!final(-e))) → remove –ize
- IF (suffix & control) → replace …
  "runn" → "run”
- with exceptions (in all languages)
  box → boxes, child → children
  one walkman → ? (walkmen / walkmans)
  and other problems: "The data is/are …", people

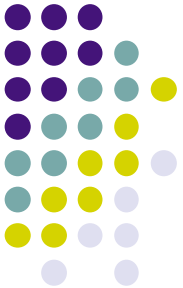
# Stemming



More complex for Germanic languages

- Various forms indicate the plural (+ add diacritics)
  "Motor", "Motoren"; "Jahr", "Jahre";
  "Apfel", "Äpfel"; "Haus", "Häuser"
- Grammatical cases imply various suffixes
  (e.g., genitive with '-es' "Staates", "Mannes")
  and also after the adjectives
  ("einen guten Mann")
- Compound construction
  ("Lebensversicherungsgesellschaftsangestellter"
  = life + insurance + company + employee)

# Stemming

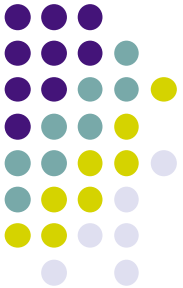


Finno-Hungarian family owns numerous cases (18 in HU, 15 FI)

| | |
|---|---|
| ház | nominative (house) |
| házat | accusative singular |
| házakat | accusative plural |
| házzal | “with” (instrumental) |
| házon | “over” (superessive) |
| házamat | my + accusative sing. |
| házamait | my + accusative + plur. |

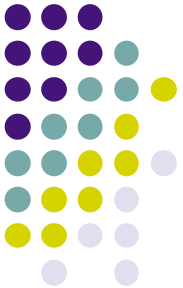
- In FI, the stem may change (e.g., “matto”, “maton”, “mattoja” (carpet))
  It seems that a deeper morphological analyzer is useful for FI
- + Compound construction (“internetfüggök”, “rakkauskirje”)

# Automatic indexing



5. Weight words
   - could be limited to a set of words (Boolean indexing), but not very effective
   - want more “important” words to have higher weight
   - using term frequency in documents (*tf*) and
   - using term frequency in the corpus (*df*)
   - frequency data independent of retrieval model
   - More on this in the next section

# Indexing

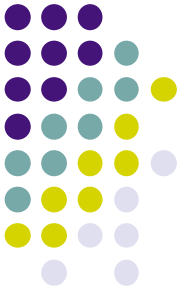


- Representing information needs (query formulation)

  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)
  - selecting the most appropriate search term

# Some issues that arise in IR



- Comparing representations (retrieval)

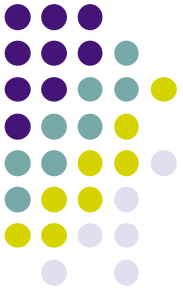
  Compare text and information need representations to determine which documents are likely to be relevant

  - 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

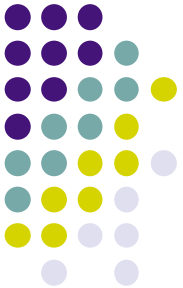
  - what are good metrics?
  - what constitutes a good experimental test bed?
  - learning schemes

# Requests examples



- Homepage searching
  - Doctor Donovan-Peluso
  - Worldnet Africa
  - HKUST Computer Science Dept.
- Ad hoc
  - Death of Kim Il Sung
  - Russian intervention in Chechnya
  - AI in Latin America

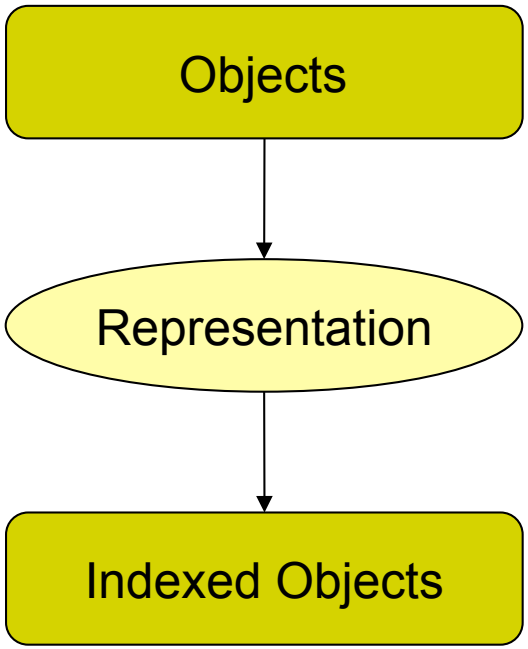
# Outline



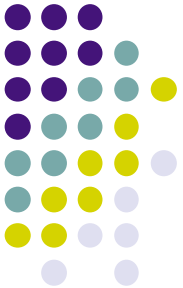
- What is Information Retrieval (IR)?
- Core idea of IR-related work
- Basic IR process
- **Simple model of IR**
- The Web
- Conclusion

# 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:
  - Ignore very frequent (e.g, "a", "the", "was", "you")
  - apply stemming
- Can be done manually

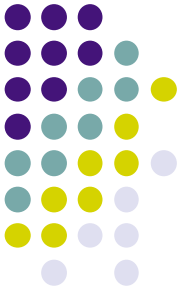


# Vector Space model



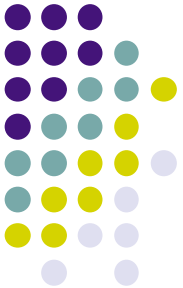
- Stemming
  - matching between documents and queries based on word sense instead of exact match (e.g, "cats" in a document, "cat" in the query)
  - automatic removal of suffixes (stemming)
  - inflectional (number, gender, case)
    "horses" → "horse"
    "actress" → "actor"
    "rosarum" → "rosa"
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

# 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
    $w_{ik} \approx (\log(tf_{ik})+1) \cdot idf_k$

# Example: small document



$D_1$ = "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}.

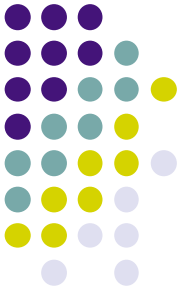
$D_2$ = {cat 1, dog 1, food 1}.

$D_3$ = {horse 2, small 1}.

How to store these values (to be effective)?

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

# Inverted file organization

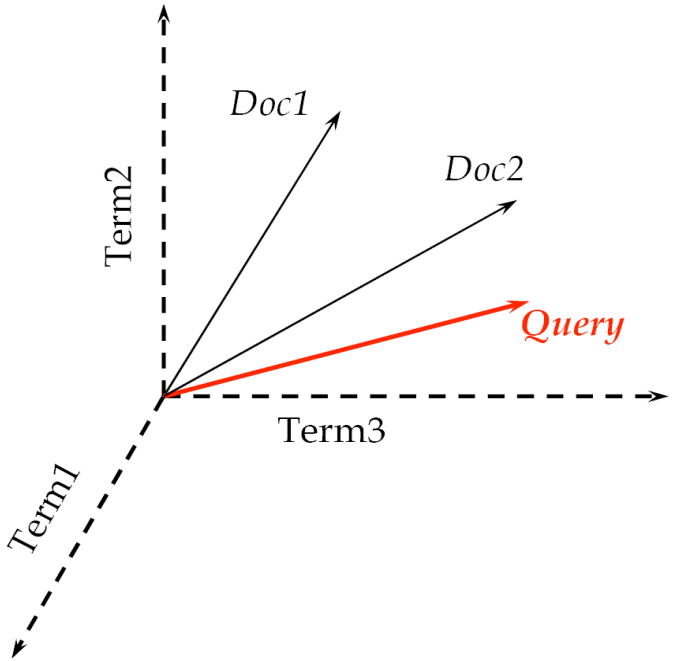


| | $D_1$ | $D_2$ | $D_3$ |
|---|---|---|---|
| horse | 3 | | 2 |
| cat | | 1 | |
| kingdom | 1 | | |
| dog | | 1 | |
| small | | | 1 |
| food | | 1 | |

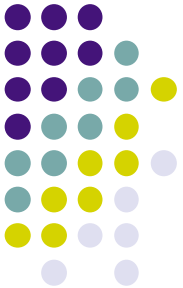
Q =
horse ={$D_1$, 3; $D_3$, 2}
food = {$D_2$, 1}

# Vector Space model

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



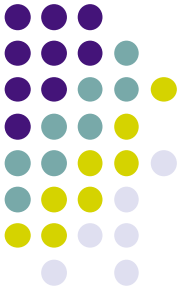
# Comparison



- 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 (inner product, with $w_{ij}$ = term $t_k$ and document $d_j$ and $w_{qk}$ = weight of term $t_k$ in the query)

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

# Inverted file organization



Inverted file

| | |
|---|---|
| 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

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

# Comparison

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

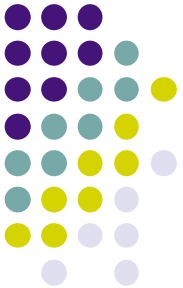
Cosine

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

Dice

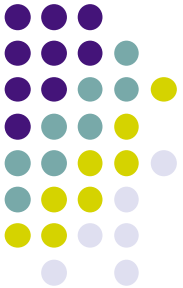
$$sim(Q, D_i) = \frac{|D_i \cap Q|}{|D_i \cup Q|} = \frac{2 \cdot \sum_{k=1}^{t} w_{ik} \cdot w_{qk}}{\sum_{k=1}^{t} w_{ik}^2 + \sum_{k=1}^{t} w_{qk}^2}$$

# Vector Space model



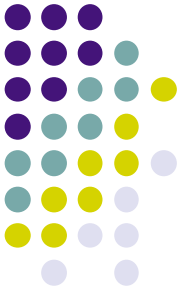
- Problem
  - unigram approach:  the fact that a given term occur does not imply that another term has more (or less) chance to co-occur (e.g, "algorithm" and "computer")
  - not clear how to define/weight noun phrase ("sort algorithm", "operating system")
  - various similarity measures
  - knowing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

# Empirical evidence



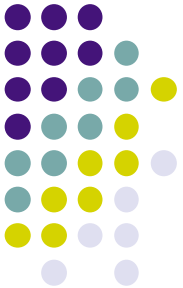
- Test-collection
  - a set of "documents" (article, image, interview, video)
  - a set of topics
  - the relevance information for each topic
- Subject / several languages
- Measure by
  - precision (# relevant items / # retrieved items)
  - recall (# relevant items / # relevant items)
- User interface is important (essential?)

# Empirical evidence



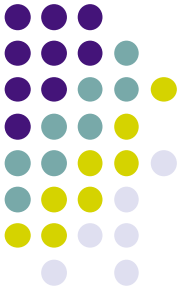
| Rank | System A | | System B | |
|---|---|---|---|---|
| 1 | R | 1/1 | nR | |
| 2 | R | 2/2 | R | 1/2 |
| 3 | nR | | R | 2/3 |
| … | nR | | nR | |
| 35 | nR | | R | 3/35 |
| … | nR | | nR | |
| 108 | R | 3/108 | nR | |
| | Prec@10 | 2/10 | Prec@10 | 2/10 |
| | P@2 | 2/2 | P@2 | 1/2 |

# Why IR system may fail



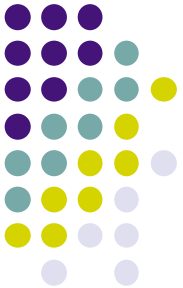
- Spelling error
  «Innondationeurs en Hollande et en Allemagne»
- Stopword list ("ai" in French)
  «AI en Amérique latine» or «IT engineer»
- Stemming ("parlement" ≠ "parlementaires")
  «Elections parlementaires européennes»
- Missing specificity
  «World Soccer Championship»
- Cannot discriminate between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

# Outline



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

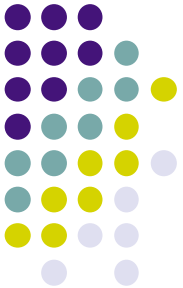
# The Web



- 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!

# The Web



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

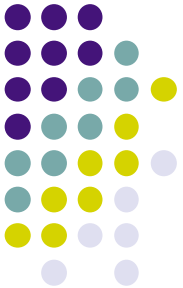
| | |
|---|---|
| Google | 46.2% |
| Yahoo | 22.5% |
| MSN | 12.6% |
| AOL | 5.4% |
| MyWay | 2.2% |
| Ask | 1.6% |
| NetScape | 1.6% |
| Others | 7.9% |

## March 2007

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

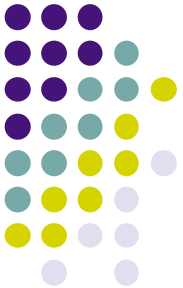
| | |
|---|---|
| Google | 48.3% |
| Yahoo | 27.5% |
| MSN | 10.9% |
| AskOL | 5.2% |
| AOL | 5.0% |

# The Web



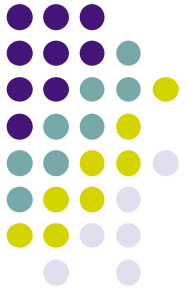
- Various task-specific search engines
  - General
  - News
  - Shopping
  - For Kids
  - Specialty (medical, gov, legal, QA, travel)
  - Images/ audio / video
  - Metasearch (metacrawler)
  - Country-specific
  - Specific SE for your web site (product)
  - Enterprise search (web + emails + memos + …)

# 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”

# The Web



Examples

- What is the melting point of lead?
- Origins of conflict in Palestine
- George Bush
- SIGIR'06 online registration
- INRT journal author instructions
- Computer Science Department
- Andrei Broder
- Official information about abortion
- Sharks Attacks in CA

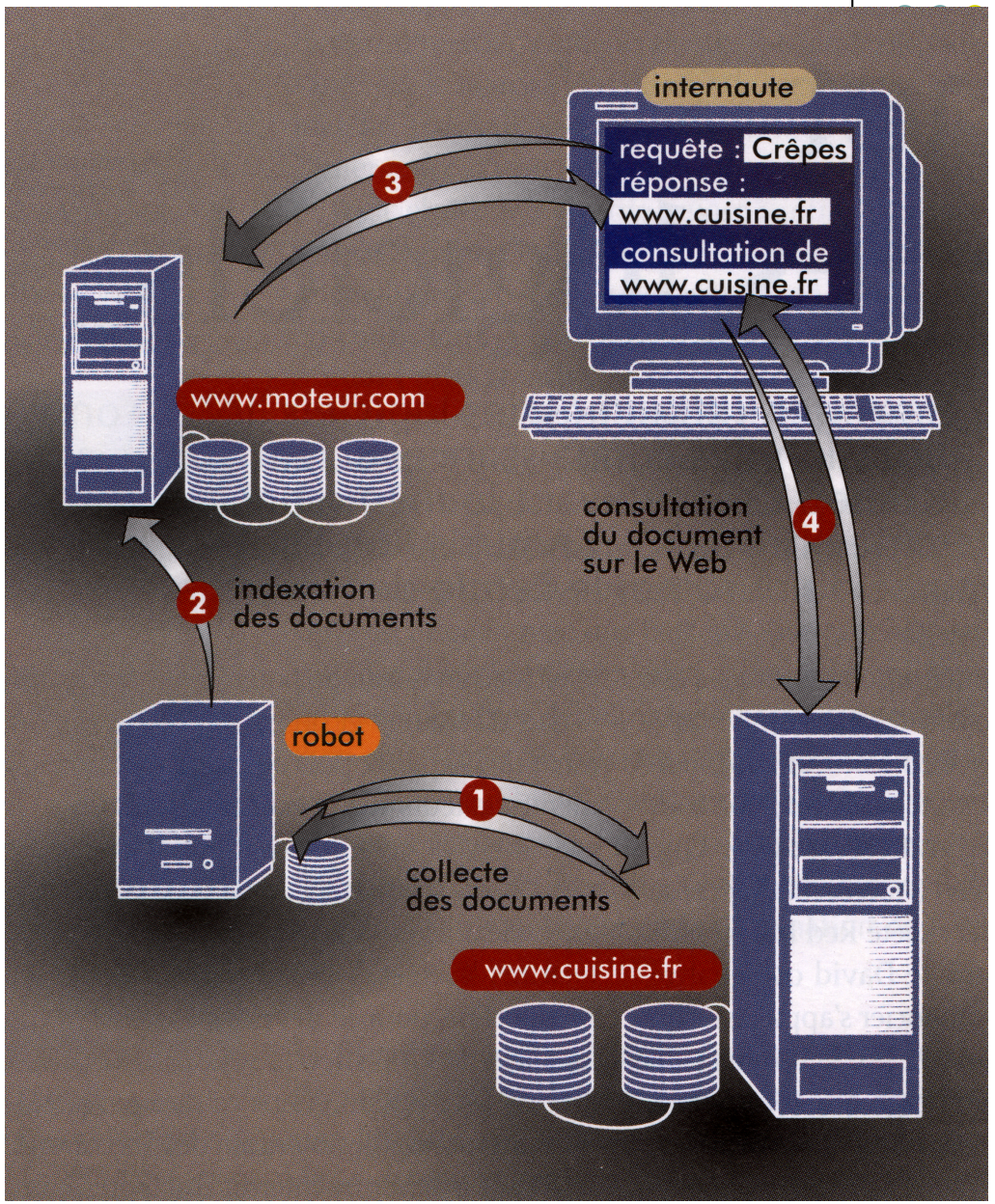
Type

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

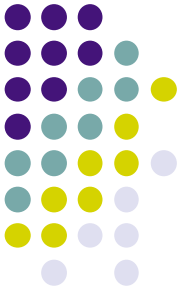
# The Web

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

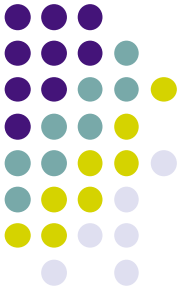


# The Web



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)

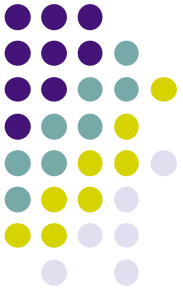
   Various policies wrt which words are indexed, capitalization, support for Unicode, stemming, support for phrases, etc

3. Query processor -- accepts queries and returns answers

   Front end -- does query reformulation -- word stemming, capitalization rules, optimization of Booleans, compounds, etc

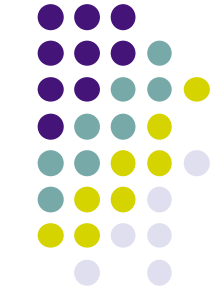
   Back end -- finds matching documents and ranks them

# 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



…
dysphasie & dyslexie
sont

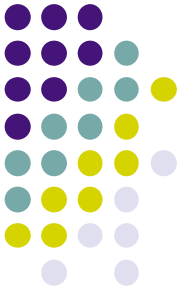
…
dépistage précoce
est essentiel

Dysphasie.be
- en Suisse
- en France

**Dysphasie en Suisse**

Troubles du langage et de la communication, les enfants souffrant de dysphasie sont pris en charge par l'AI, assurance fédérale,

…
maladie génétique
comme les récentes …



# PageRank

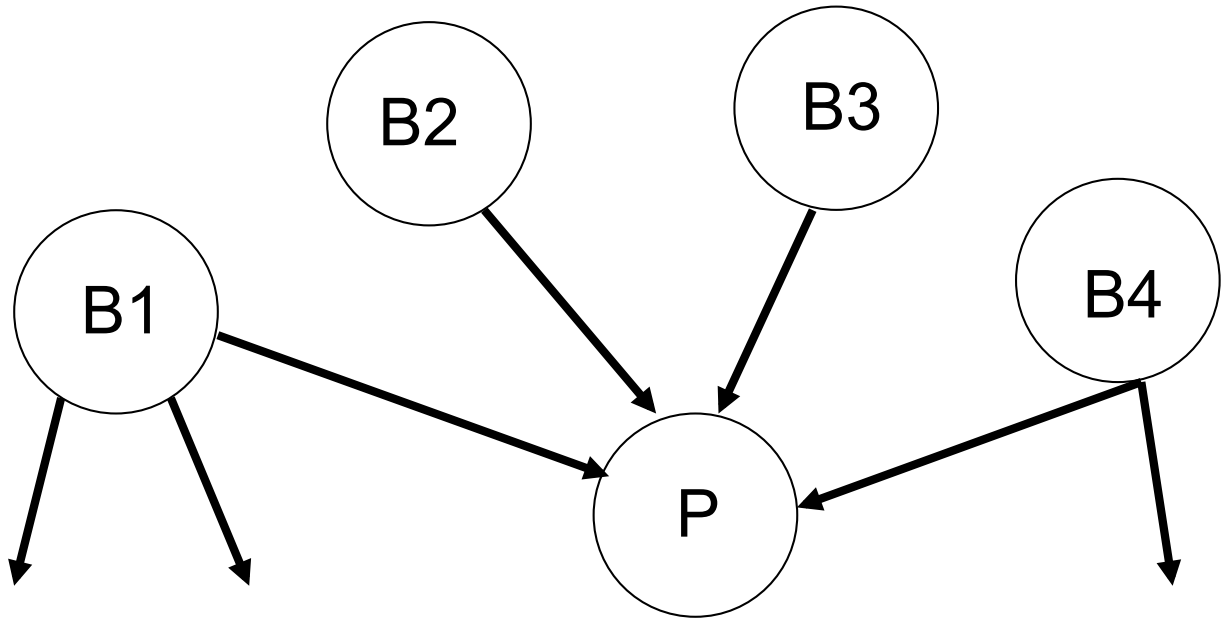
- Initially the surfer is at a random page
  - At each step, the surfer proceeds to a randomly chosen web page with probability *d* (e.g., probability of a random jump = 0.15)
  - or to a randomly chosen successor of the current page with probability 1-*d*
    (e.g., probability of following a random 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.

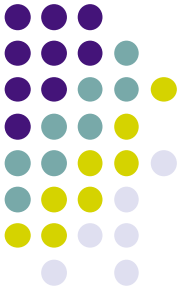
# PageRank

- Extend inductively:

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



# Random Surfer Model



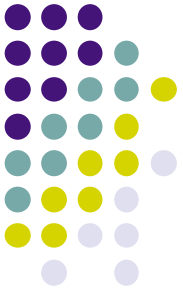
- Formally

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

  $PR^c(D_i)$: PageRank value of page $D_i$ after c cycles
  $C(D_i)$: number of outlinks for page $D_i$ (outdegree)

- But to compute $PR^c(D_i)$, we need $PR^{c-1}(D_j)$
  We do it iteratively (usually 5 iterations is enough)

# Random Surfer Model

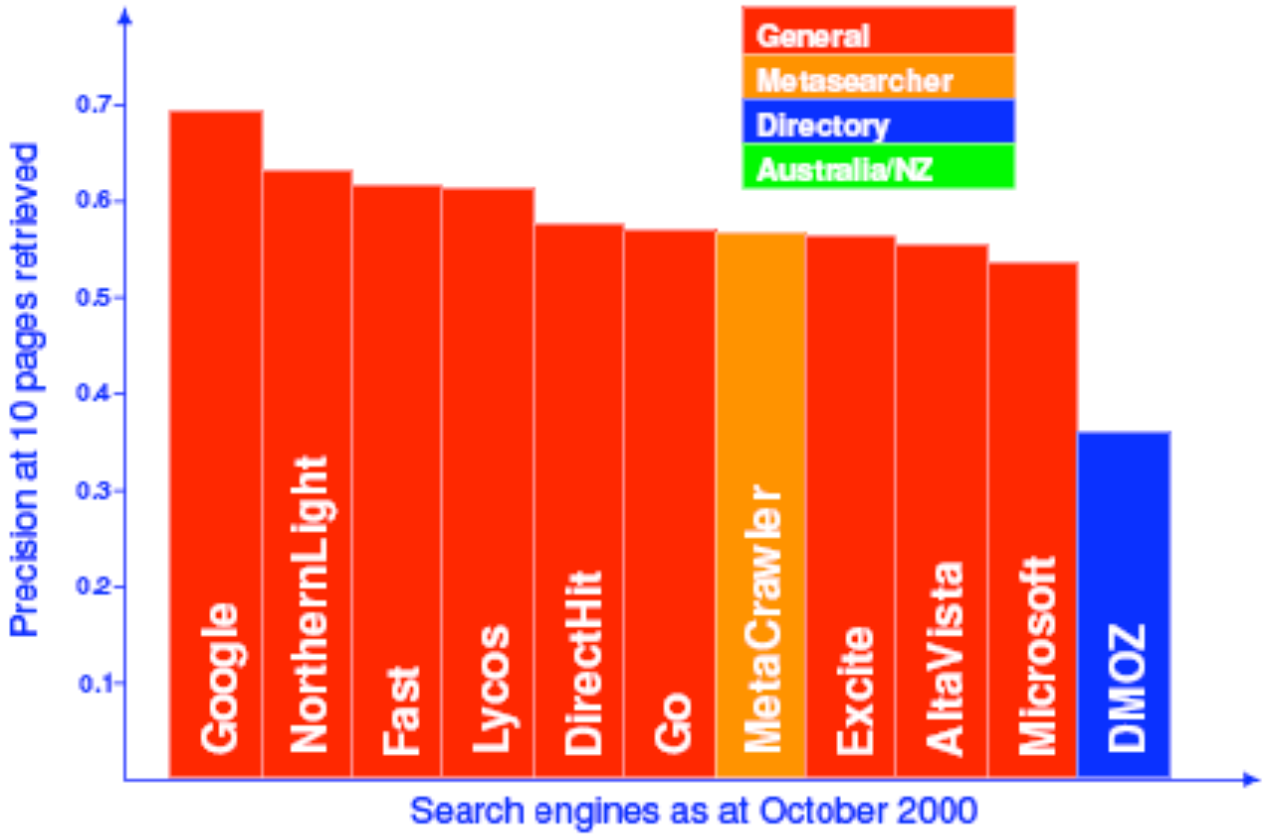


- How realistic is the random surfer model?
  - What if we modeled the back button? [Fagi00]
  - Surfer behavior sharply skewed towards short paths [Hube98]
  - Search engines, bookmarks & directories make jumps nonrandom.
- Biased Surfer Models
  - Weight edge traversal probabilities based on match with topic/query (non-uniform edge selection)
  - Bias jumps to pages on topic (e.g., based on personal bookmarks & categories of interest)

# The Web

Evaluation:
Precision at 10

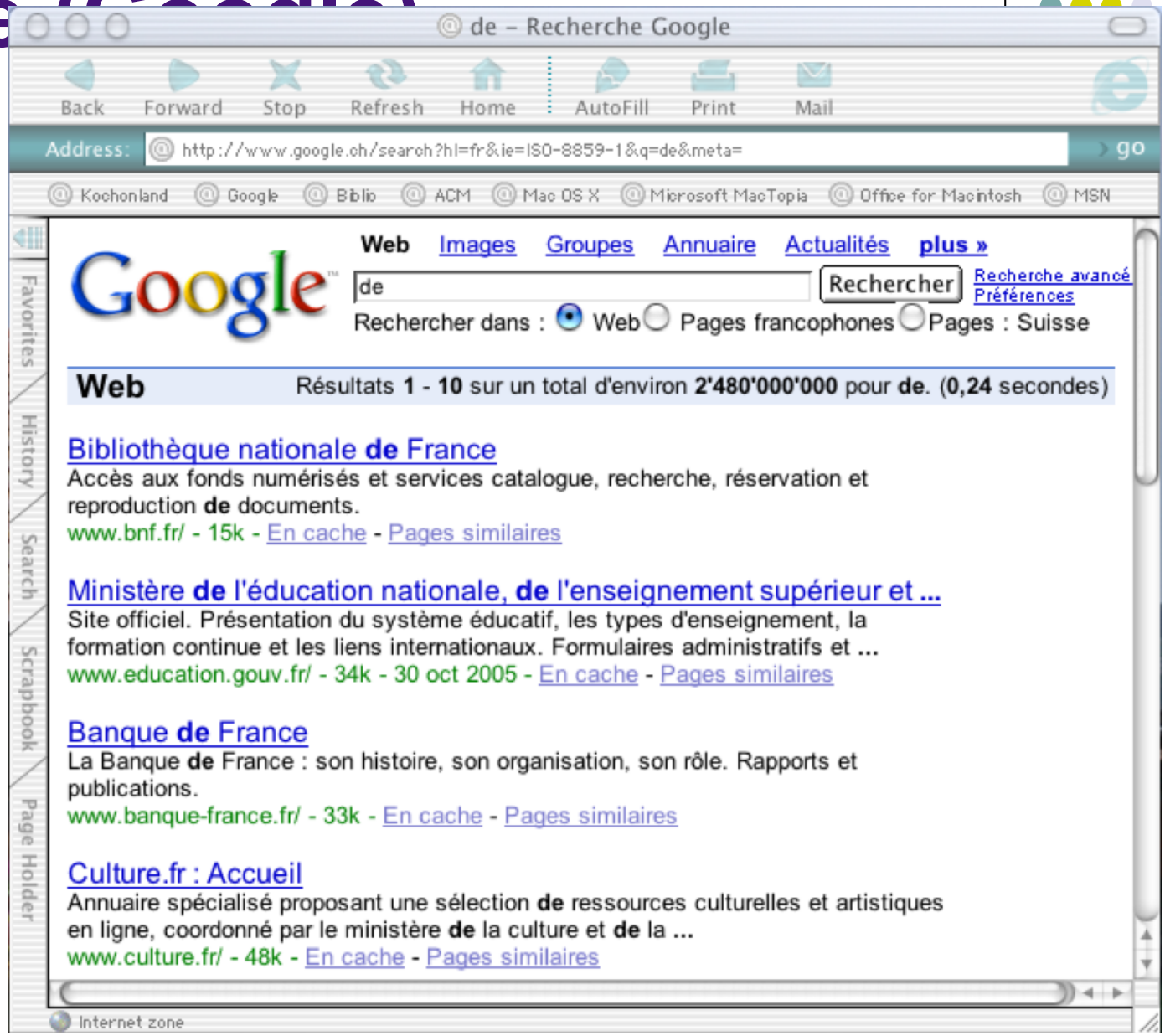
Queries: 106 on-line service queries from AV and EM logs



# Example (Google)

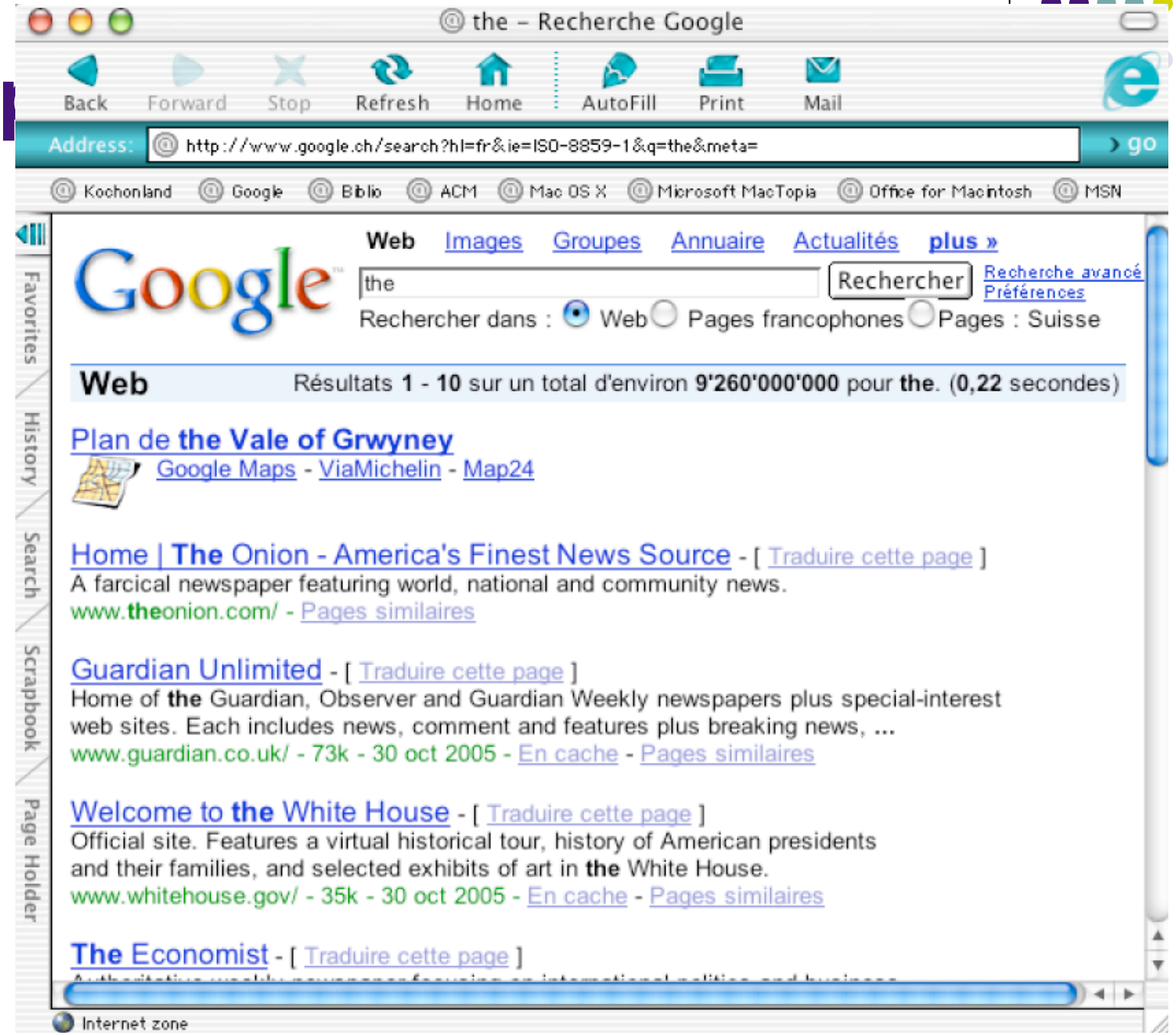
Query “de”

in French

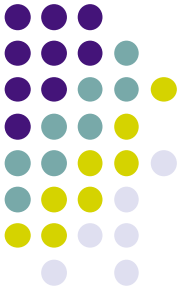


# Examp

Q="the"


the – Recherche Google
Back
Forward
Stop
Refresh
Home
AutoFill
Print
Mail
Address:
http://www.google.ch/search?hl=fr&ie=ISO-8859-1&q=the&meta=
go
Kochonland
Google
Biblio
ACM
Mac OS X
Microsoft MacTopia
Office for Macintosh
MSN
Favorites
History
Search
Scrapbook
Page Holder
Google
Web
Images
Groupes
Annuaire
Actualités
plus »
the
Rechercher
Recherche avancé
Préférences
Rechercher dans : Web Pages francophones Pages : Suisse
Web
Résultats 1 - 10 sur un total d'environ 9'260'000'000 pour the. (0,22 secondes)
Plan de the Vale of Grwyney
Google Maps - ViaMichelin - Map24
Home | The Onion - America's Finest News Source - [ Traduire cette page ]
A farcical newspaper featuring world, national and community news.
www.theonion.com/ - Pages similaires
Guardian Unlimited - [ Traduire cette page ]
Home of the Guardian, Observer and Guardian Weekly newspapers plus special-interest web sites. Each includes news, comment and features plus breaking news, ...
www.guardian.co.uk/ - 73k - 30 oct 2005 - En cache - Pages similaires
Welcome to the White House - [ Traduire cette page ]
Official site. Features a virtual historical tour, history of American presidents and their families, and selected exhibits of art in the White House.
www.whitehouse.gov/ - 35k - 30 oct 2005 - En cache - Pages similaires
The Economist - [ Traduire cette page ]
Internet zone

# 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