Introduction to Information Retrieval

Konstantin Todorov

{firstname.lastname}@lirmm.fr

University of Montpellier

October, 2018

Outline

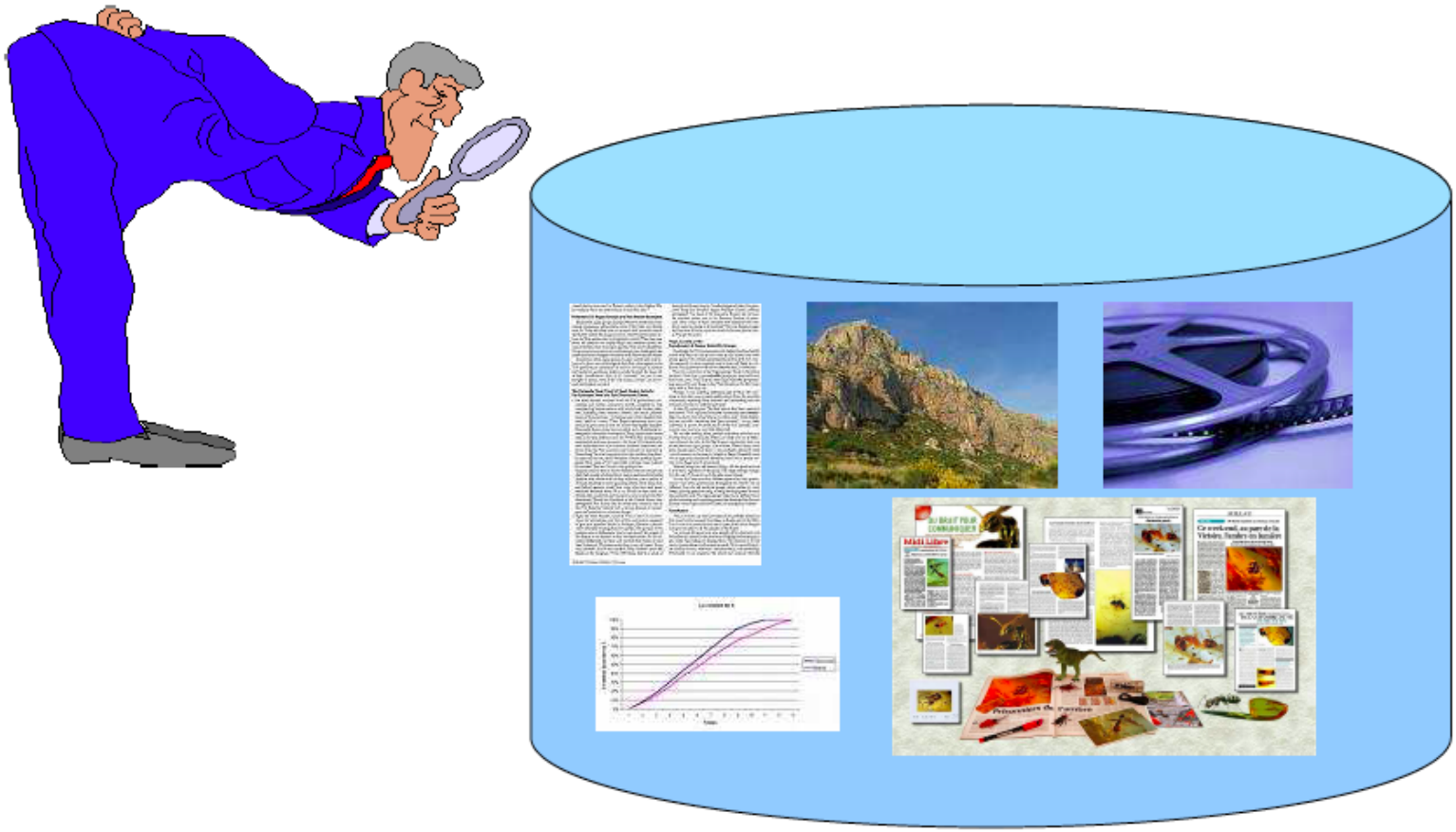
Outline

What is information retrieval?

An information retrieval system (IRS) is a system that allows to discover meaningful and useful information with respect to a query in a big collection of documents.

- Structured: Relational databases (*data retrieval*)
- Unstructured: text, video, sound (*information retrieval*)

What is information retrieval?



From a course by C. Hudelot

Goal of this course

Answer the questions:

- What is the theory behind an IRS?
- How to measure the efficiency of an IRS?
- What is hiding behind a web search engine?
- Where IR ends and where Data Mining begins?
- What are the current challenges?

Introduction

When and where do we have to do with IR?

- Everyday activity of every person who goes on the Web

Important concepts:

- A document: a wikipedia web page, a tweet, an image,...
- A query: an expression of a certain information need of the user

Introduction

Bjork

Web Images Maps Shopping Vidéos Plus Outils de recherche

Environ 3 760 000 résultats (0,20 secondes)

Les cookies assurent le bon fonctionnement de nos services. En utilisant ces derniers, vous acceptez l'utilisation des cookies.

OK En savoir plus

björk.com
bjork.com/ ▾ Traduire cette page
as you may have noticed, things have changed around here. we're exploring new worlds, which some of you might have heard of, starting with biophilia, **björk's** ...

Björk - Wikipédia
fr.wikipedia.org/wiki/**Björk** ▾
Björk Guðmundsdóttir ([ˈpjœrk ˈkvʏðmʏnts,toʊhtɪr]) est une musicienne, chanteuse, compositrice et actrice islandaise, née le 21 novembre 1965 à ...
Goldie - Matthew Barney - Discographie de Björk - Vidéographie de Björk
Vous avez consulté cette page 3 fois. Dernière visite : 28/12/13

bjork.fr - Le site francophone sur **Björk**
www.**bjork**.fr/ ▾
Une communauté francophone dédiée à la chanteuse et aux musiques islandaises propose des forums de discussion, des actualités et des séquences vidéo.

Björk - AlloCiné
www.allocine.fr › Stars › Toutes les stars › Islande ▾
Björk (**Björk** Gudmundsdottir), Actrice, Compositeur. Découvrez sa biographie, sa carrière en détail et toute son actualité.

Björk | Listen and Stream Free Music, Albums, New Releases ...
https://myspace.com/**bjork** ▾ Traduire cette page
Björk's profile including the latest music, albums, songs, music videos and more updates.

Björk - It's Oh So Quiet - YouTube


www.youtube.com/v/htobTBICvUU ▾
1 juil. 2007 - Ajouté par bjorkdotcom

Björk

Suivre

1 645 431 abonnés sur Google+

Björk Guðmundsdóttir est une musicienne, chanteuse, compositrice et actrice islandaise, née le 21 novembre 1965 à Reykjavik. Wikipédia

Conjoint : Matthew Barney
Conjoint : Þór Eldon (m. 1986–1987)
Films : Dancer in the Dark, Drawing Restraint 9, Prêt-à-porter, Plus

Titres

All Is Full of Love	1997	Homogenic
It's Oh So Quiet	1995	Post
Army of Me	1995	Post
Moon	2011	Biophilia
Hidden Place	2001	Vespertine

Posts récents


" i'm not gonna do the same stuff again , because if I did I'd be bored shitless . " source: http://bit.ly/18CQ0nS 20 déc. 2013

Introduction

Important concepts: relevance

A relevant document is that which contains the researched information.

- difficulty: the researched information corresponds to an information need which is expressed by a query
- the relevance can be used to evaluate the quality of an IRS

Important concepts: user feedback

- active user involvement
- user profile

Introduction

jaguar

Web Images Maps Shopping Plus Outils de recherche

Environ 74 000 000 résultats (0,27 secondes)

Les cookies assurent le bon fonctionnement de nos services. En utilisant ces derniers, vous acceptez l'utilisation des cookies.

OK En savoir plus

Jaguar France - Voitures de Sport et Voitures de luxe
www.**jaguar**.fr/
Découvrez les voitures de luxe **Jaguar**. Alliant héritage et technologie, les berlines et voitures de sport **Jaguar** vous feront vivre une expérience de conduite ...

Images correspondant à **jaguar** - Signaler des images inappropriées

Jaguar - Wikipédia
fr.wikipedia.org/wiki/**Jaguar**
Statut de conservation UICN. (NT) NT : Quasi menacé. Statut CITES. Sur l' annexe I de la CITES Annexe I , Rév. du 01-07-1975. Le **jaguar** (Panthera onca) est ...

Jaguar (automobile) - Wikipédia
fr.wikipedia.org/wiki/**Jaguar**_(automobile)
Jaguar, de son nom officiel « **Jaguar** Cars Ltd », est une marque automobile connue pour ses voitures de luxe et ses modèles sportifs. La marque est la ...

Jaguar

Automobile

Jaguar, de son nom officiel « Jaguar Cars Ltd », est une marque automobile connue pour ses voitures de luxe et ses modèles sportifs. La marque est la propriété de l'indien Tata Motors depuis 2008. Wikipédia

Date de fondation : 1922

PDG : Ralf Speth

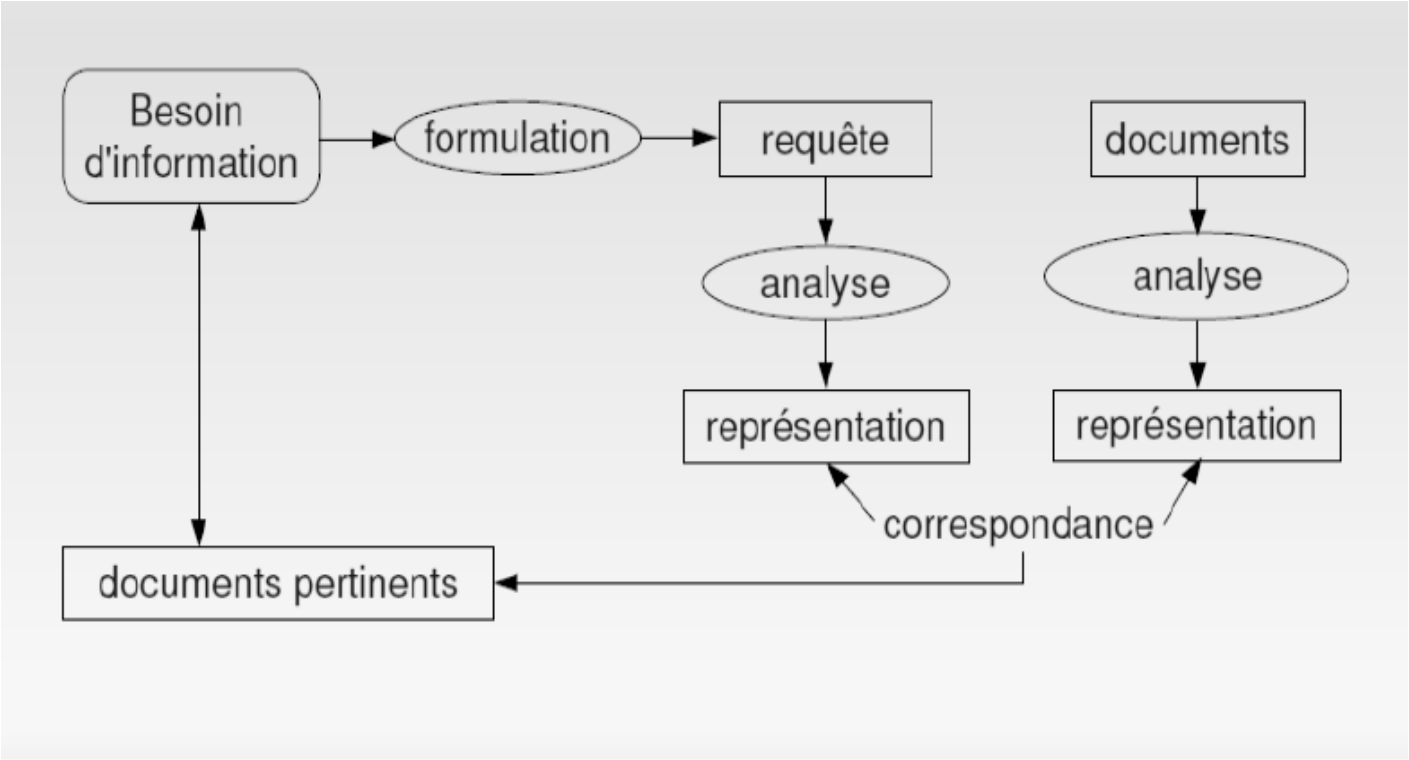
Fondateurs : William Walmsley, William Lyons

Signaler un problème/Plus d'infos

Afficher les résultats pour

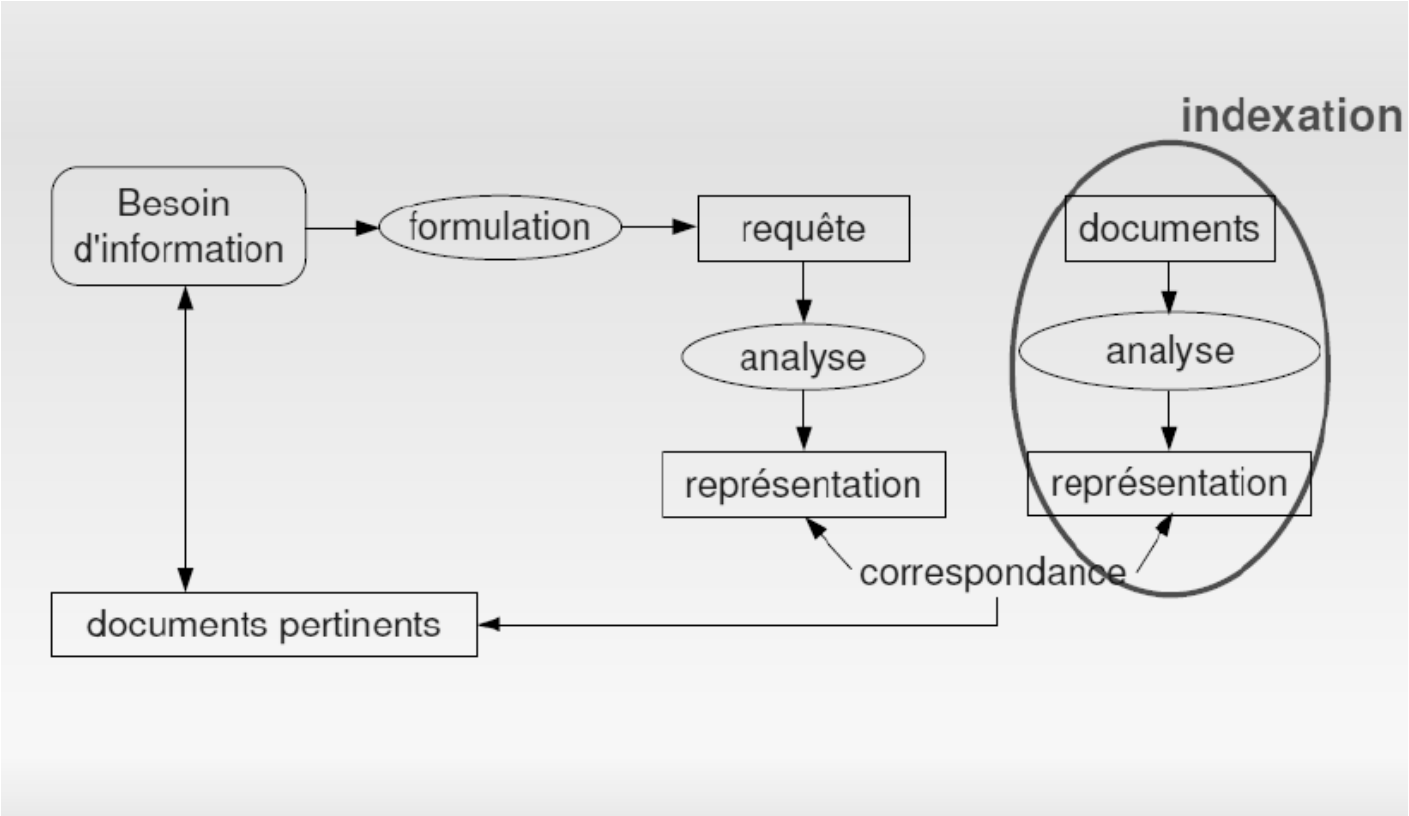
Jaguar (Animal)
Le jaguar est un mammifère carnivore de la famille des félidés. C'est l'un des cinq « grands félins » du ...

The classical approach to information retrieval



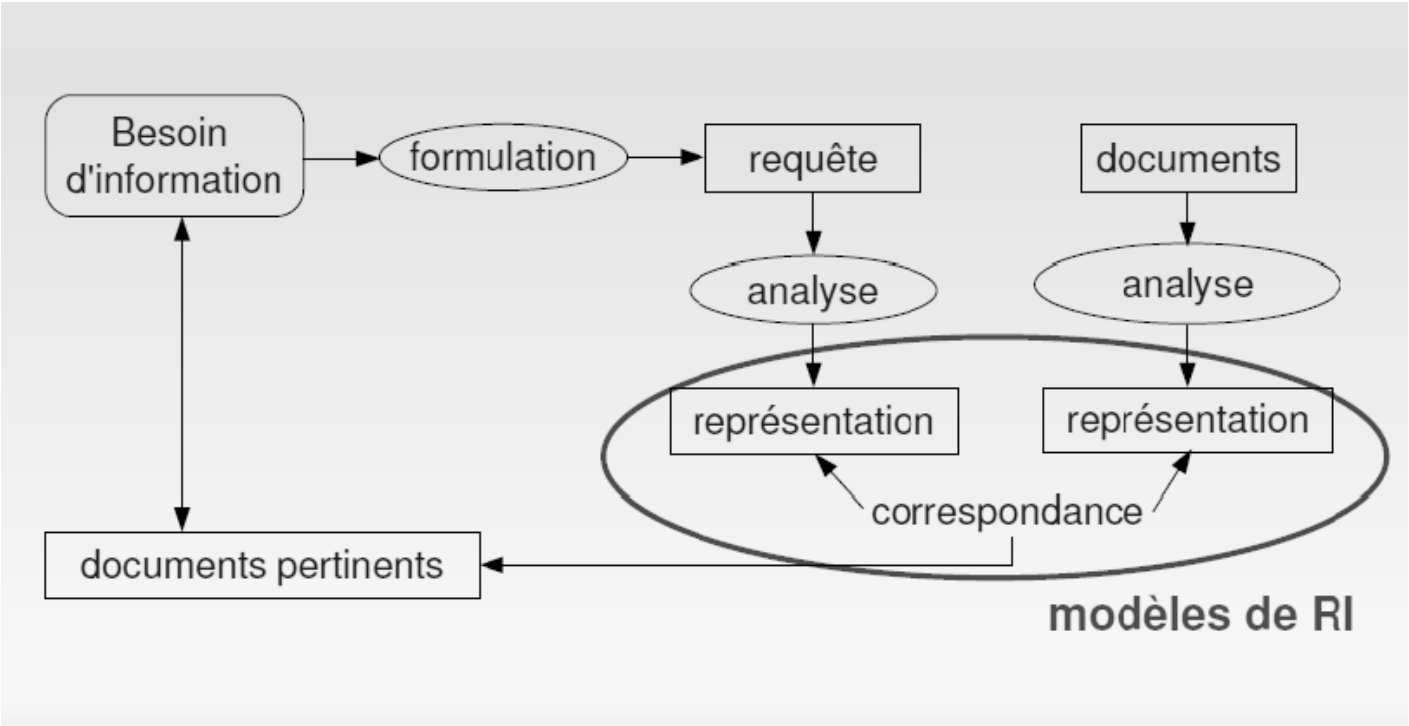
Introduction

The classical approach to information retrieval: indexing



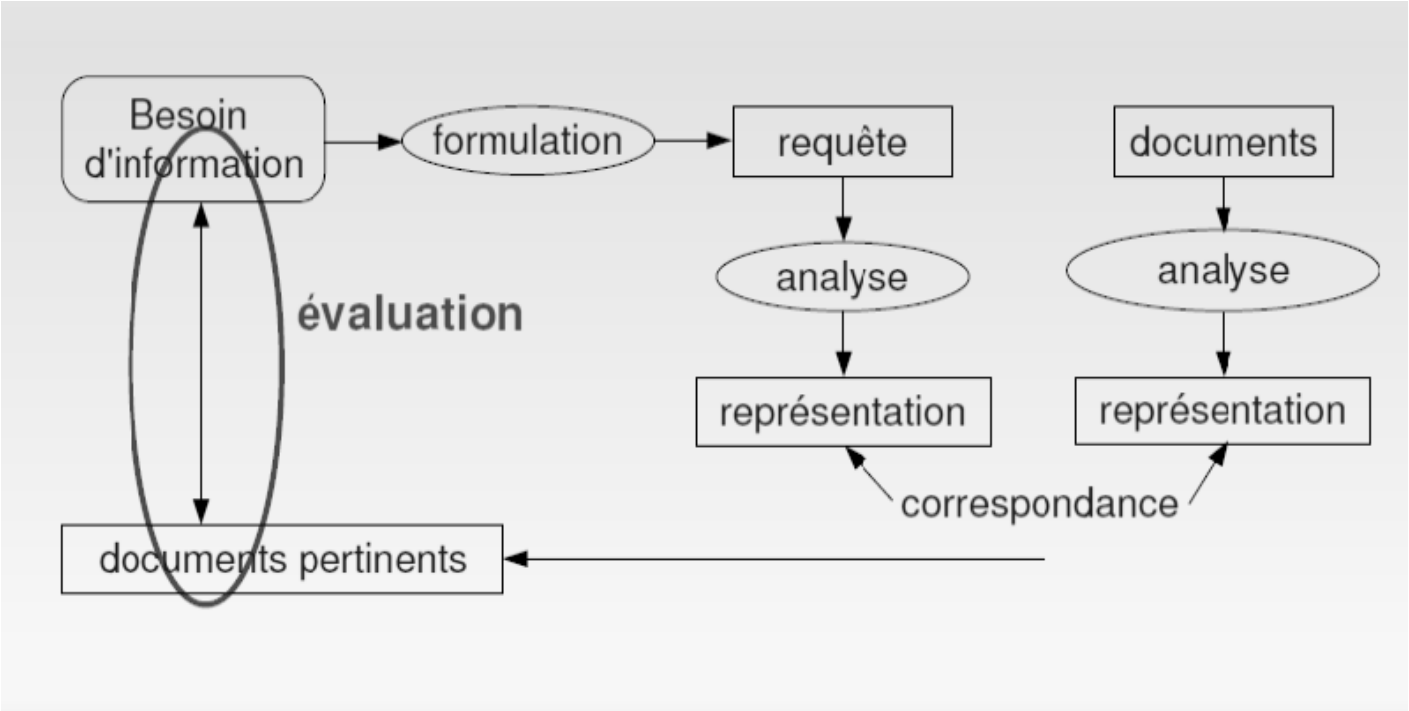
Introduction

The classical approach to information retrieval: models



Introduction

The classical approach to information retrieval: evaluation



Outline

Indexing

- Index: a data structure built from text to speed up the search by providing an appropriate representation of the documents.
- When a text collection is updated, the index has to be updated, too.
- Semi-static collections: collections which are updated at reasonable regular intervals (say, daily)
- Most real text collections, including the Web, are indeed semi-static

  Although the Web changes very fast, the crawls of a search engine are relatively slow

Indexing

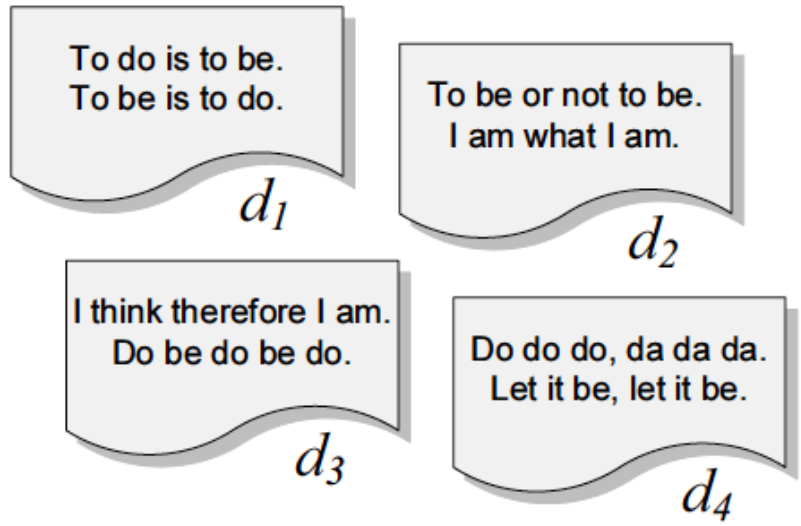
Basic concepts

- Inverted index: a word-oriented mechanism for indexing a text collection to speed up the searching task
- The inverted index structure is composed of two elements: the vocabulary and the occurrences
- The vocabulary is the set of all different words in the text
- For each word in the vocabulary the index stores the documents which contain that word (inverted index)

Indexing

Term-document matrix

Vocabulary	n_i	d_1	d_2	d_3	d_4
to	2	4	2	-	-
do	3	2	-	3	3
is	1	2	-	-	-
be	4	2	2	2	2
or	1	-	1	-	-
not	1	-	1	-	-
I	2	-	2	2	-
am	2	-	2	1	-
what	1	-	1	-	-
think	1	-	-	1	-
therefore	1	-	-	1	-
da	1	-	-	-	3
let	1	-	-	-	2
it	1	-	-	-	2



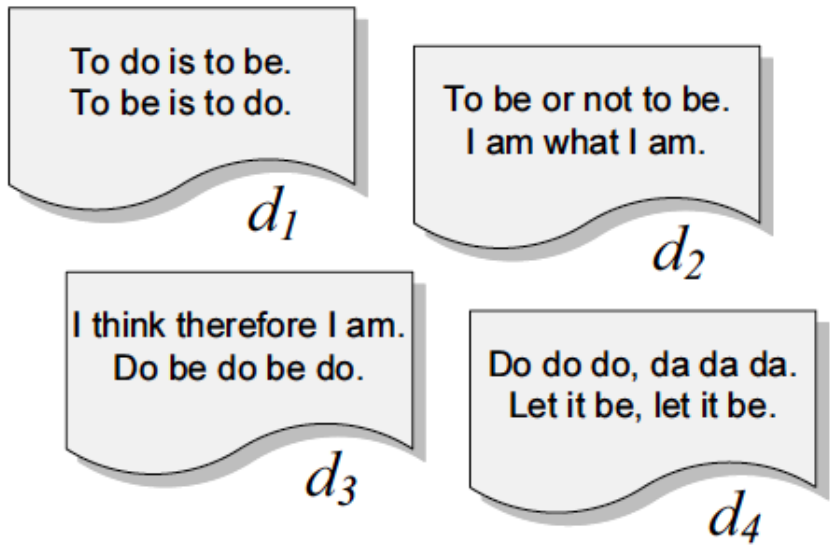
Taken from Baeza-Yates and Ribeiro-Neto.

Indexing

Inverted index (basic form)

Limit the required space for storing

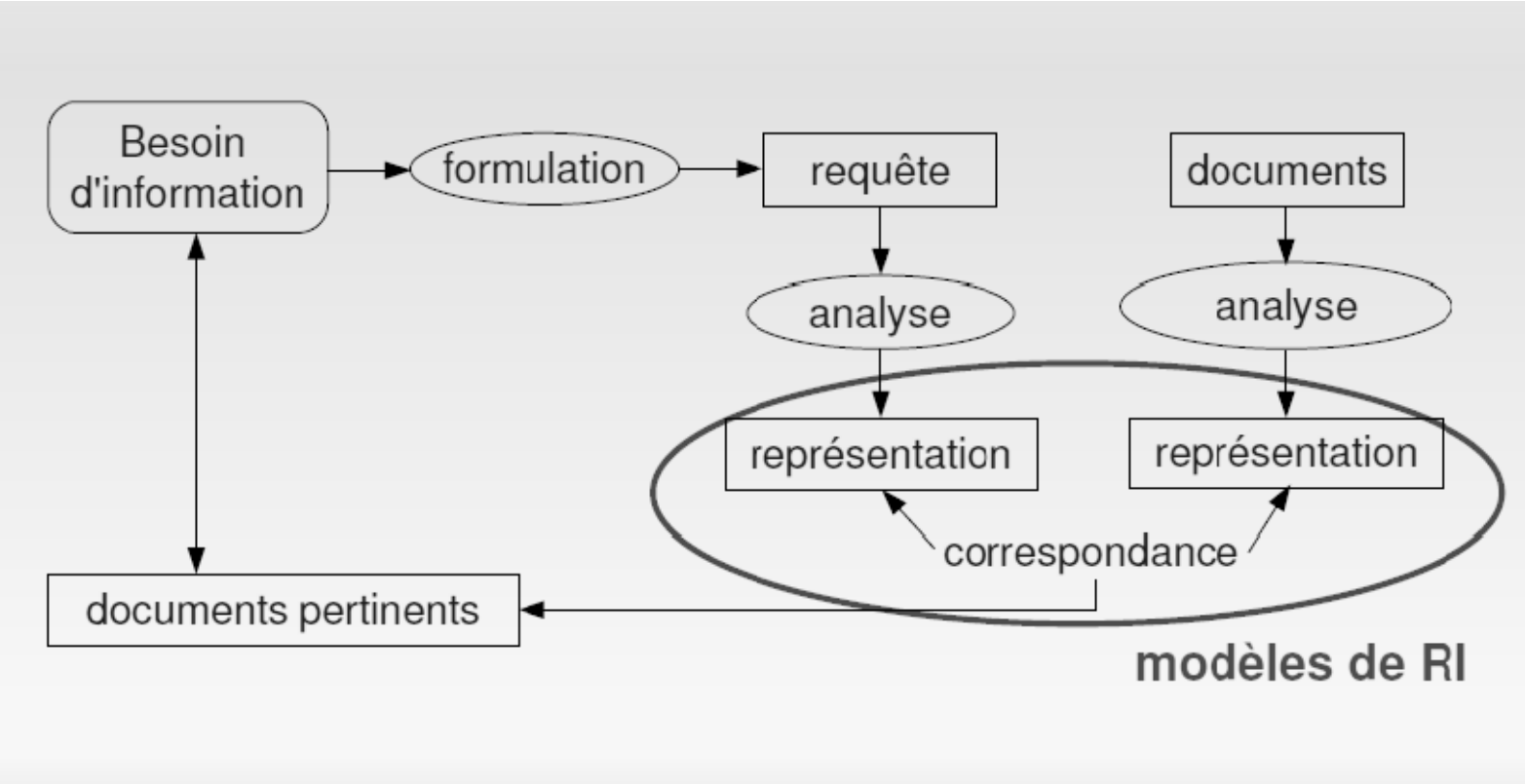
Vocabulary	n_i	Occurrences as inverted lists
to	2	[1,4],[2,2]
do	3	[1,2],[3,3],[4,3]
is	1	[1,2]
be	4	[1,2],[2,2],[3,2],[4,2]
or	1	[2,1]
not	1	[2,1]
I	2	[2,2],[3,2]
am	2	[2,2],[3,1]
what	1	[2,1]
think	1	[3,1]
therefore	1	[3,1]
da	1	[4,3]
let	1	[4,2]
it	1	[4,2]



Taken from Baeza-Yates and Ribeiro-Neto.

Outline

The classical approach to information retrieval: models



Models

A model is a mathematical abstraction of a real process, which helps to study and predict the behaviour of a system.

Modeling in IR consists of two main tasks:

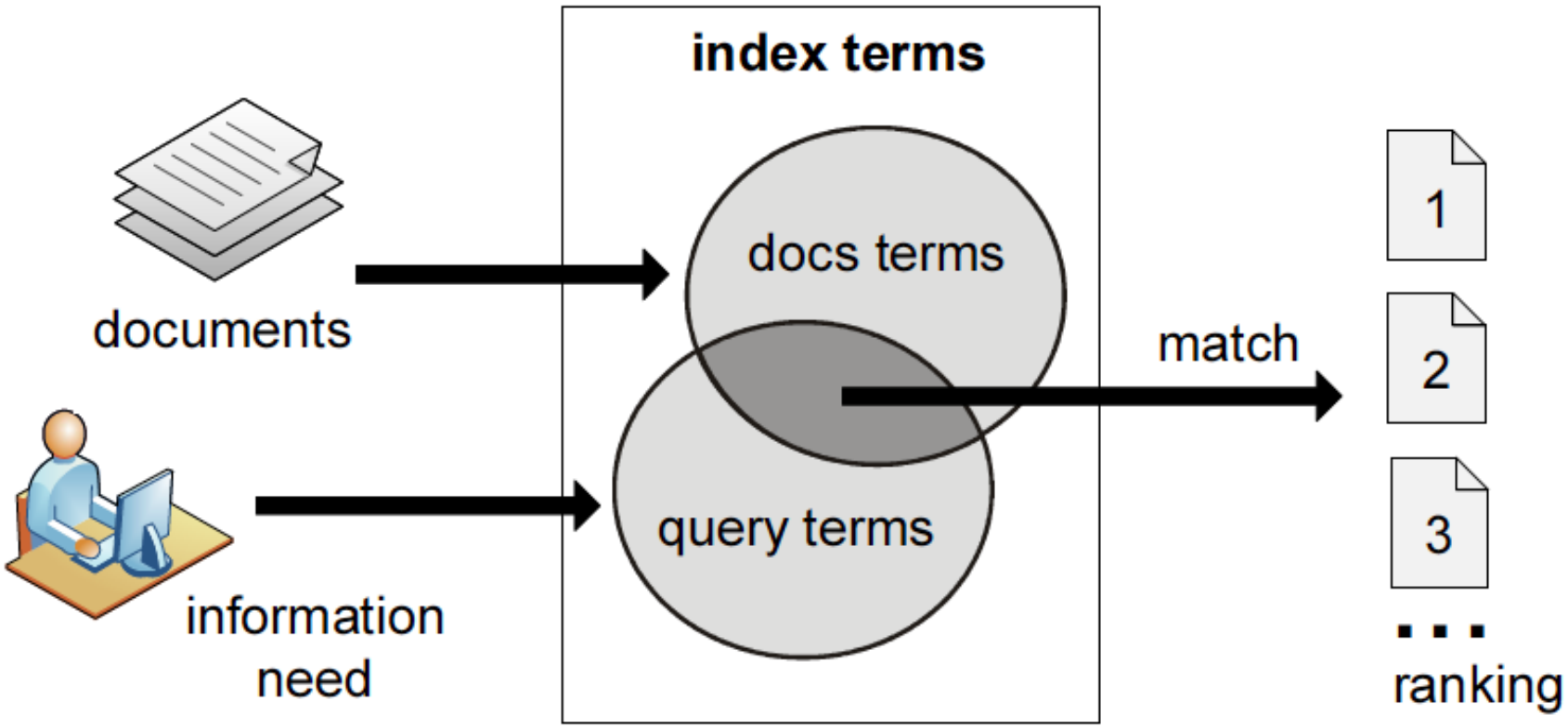
- The conception of a logical framework for representing documents and queries
- The definition of a ranking function that allows quantifying the similarities among documents and queries

Models

IR systems use the *index terms* to retrieve documents

- Index term (terms from the vocabulary of the index):
  - In a restricted sense: it is a keyword that has some meaning on its own; usually plays the role of a noun
  - In a more general form: it is any word that appears in a document
- Retrieval based on index terms can be implemented efficiently (see Indexing section)
- Index terms are simple to refer to in a query

The IR process

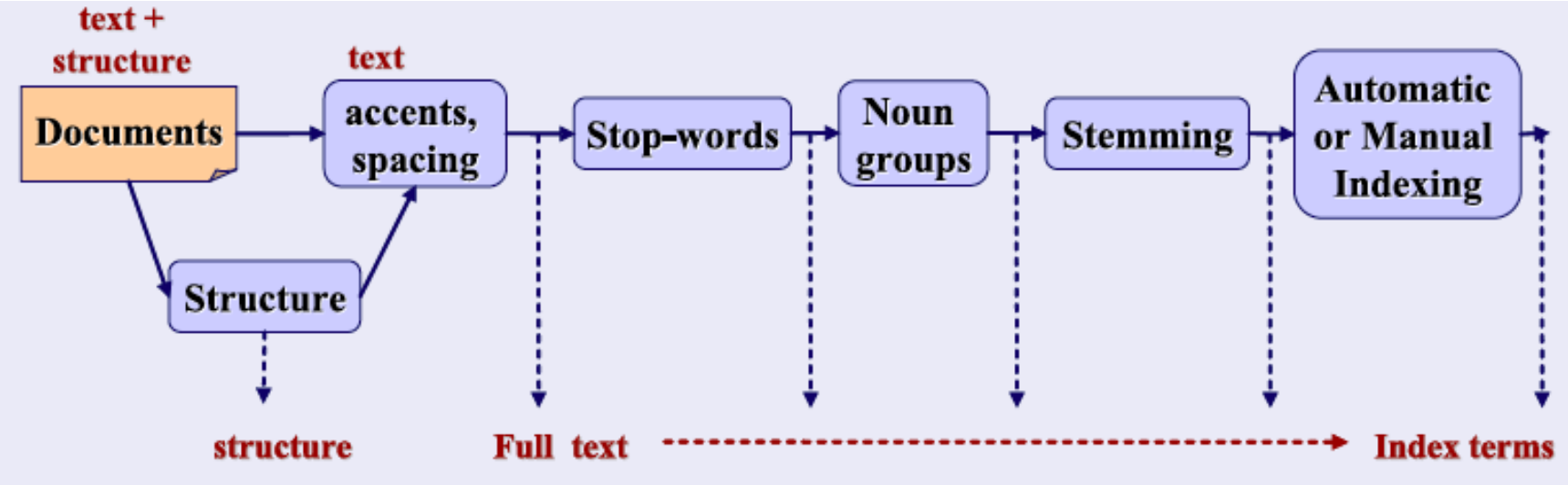


Taken from Baeza-Yates and Ribeiro-Neto.

Outline

Document Processing

Representation of a document: a logical view



From full text ——————-> to a set of index terms

Document Processing

Representation of a document: preprocessing

- Stop words: articles, conjunctions, too frequent (common) words,...
  specific to a language: "the", "a", "and", "her",...
- Noun groups: remove adjectives, verbs, adverbs
- Stemming / Lemmatization: strip words down to their canonical forms
  {"computers, "computing", "computation"} -> "comput"

Outline

Boolean Model

The simplest model based on set theory and exact matches.

- The term-document matrix contains only zeros and ones: 0 if a term is not found in a document, 1 - otherwise
- Documents are returned if they satisfy a given boolean expression: a query is given as a set of index terms connected by logical operators AND, OR, NOT
- The results are non-ordered: a document is either relevant or not (to a query)

Boolean Model

Boolean truth tables

a	b	not ¬b	and a ∧ b	or a ∨ b
0	0	1	0	0
0	1	0	0	1
1	0		0	1
1	1		1	1

For x, y, c, d index terms, a query:

- *x* AND *y* ?
- (*x* AND *y*) OR *c* AND NOT *d* ?

Boolean Model

An example

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Boolean Model

An example

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Information need: The documents that talk about Brutus and Caesar, but not about Calpurnia.

Query: Brutus AND Caesar AND NOT Calpurnia

Boolean Model

An example

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Information need: The documents that talk about Brutus and Caesar, but not about Calpurnia.

Query: Brutus AND Caesar AND NOT Calpurnia

Boolean Model

An example

	Anthony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Anthony	1	1	0	0	0	1
Brutus	1	1	0	1	0	0
Caesar	1	1	0	1	1	1
Calpurnia	0	1	0	0	0	0
Cleopatra	1	0	0	0	0	0
mercy	1	0	1	1	1	1
worser	1	0	1	1	1	0

Query: Brutus AND Caesar AND NOT Calpurnia
Boolean expression: 110100 AND 110111 AND 101111 = 100100

Results for this query...?

Boolean Model

Drawbacks of this model

- No notion of partial matching
- No ranking (absence of a grading scale)
- The query is given as a Boolean expression
  -> the user has to know how to do that!
- The Boolean queries formulated by the users are most often too simplistic
- The model frequently returns either too few or too many documents

Outline

Term Weighting

Motivation

- The terms of a document are *not equally "useful"* for describing the document content
- Some index terms carry more information than others
- There are properties of an index term which are useful for evaluating the importance of the term in a document

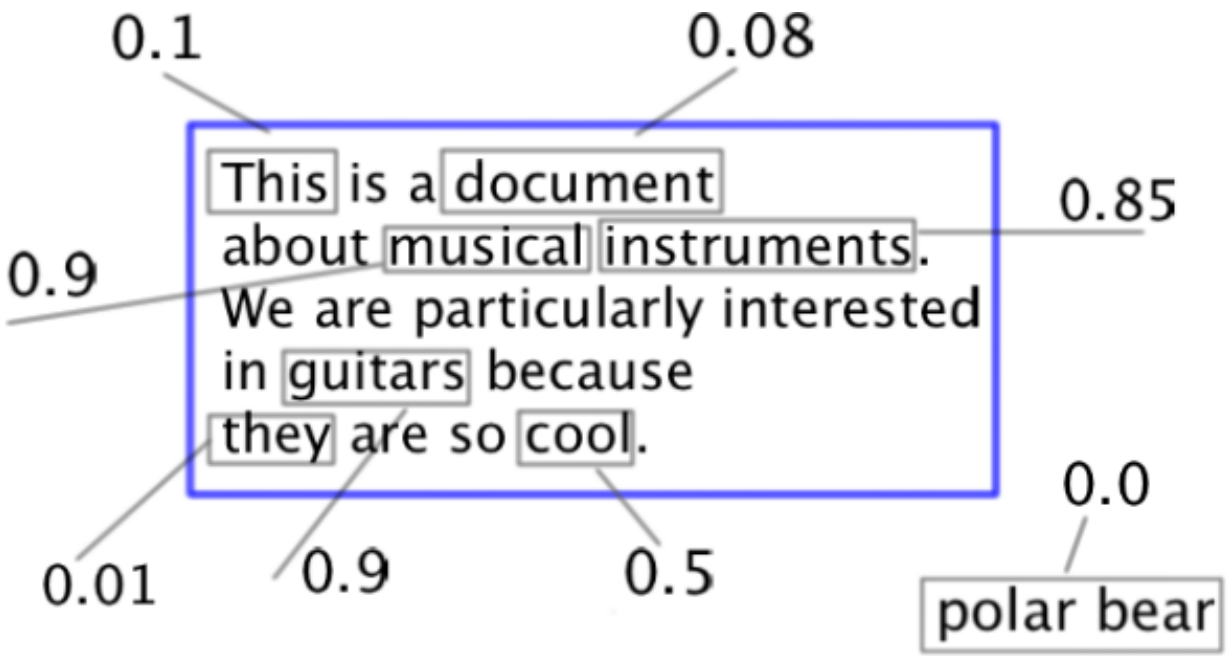
Important or unimportant?

- A word which appears in all documents of a collection?
- A word that appears in only 1 document, several times?

Term Weighting

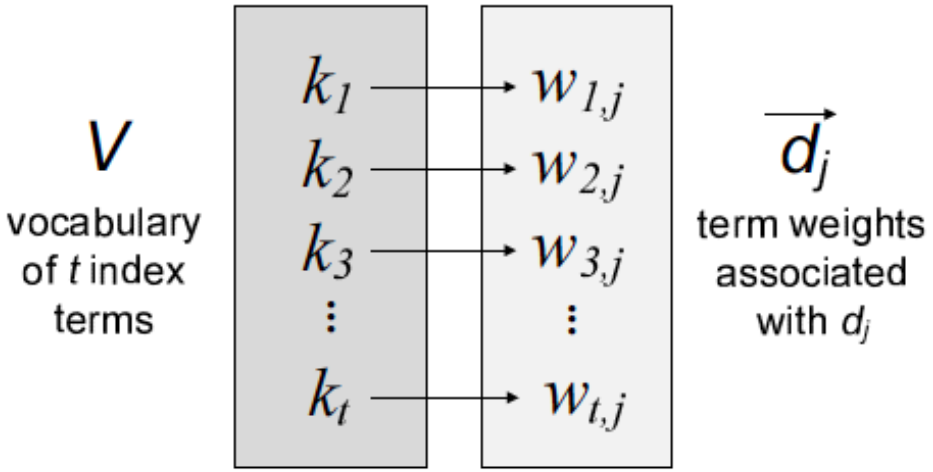
We associate a weight to every term with respect to a document:

- a term k_i
- a document d_j
- a weight $w_{i,j} \geq 0$



Term Weighting

Let $V = (k_1, k_2, \ldots, k_t)$ be the vocabulary (the set of all index terms)



We can express a document as a vector: $\mathbf{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$.

(We will see that in detail in the vector model.)

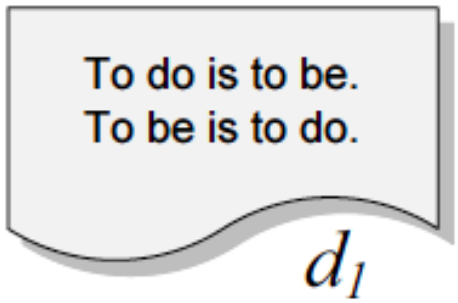
Term Weighting

How to set these weights ?

Term Weighting

- Boolean model: $w_{i,j} \in \{0, 1\}$.

  A term k_i has a weight 1 *in document d_j if it appears (at least once) in d_j.*



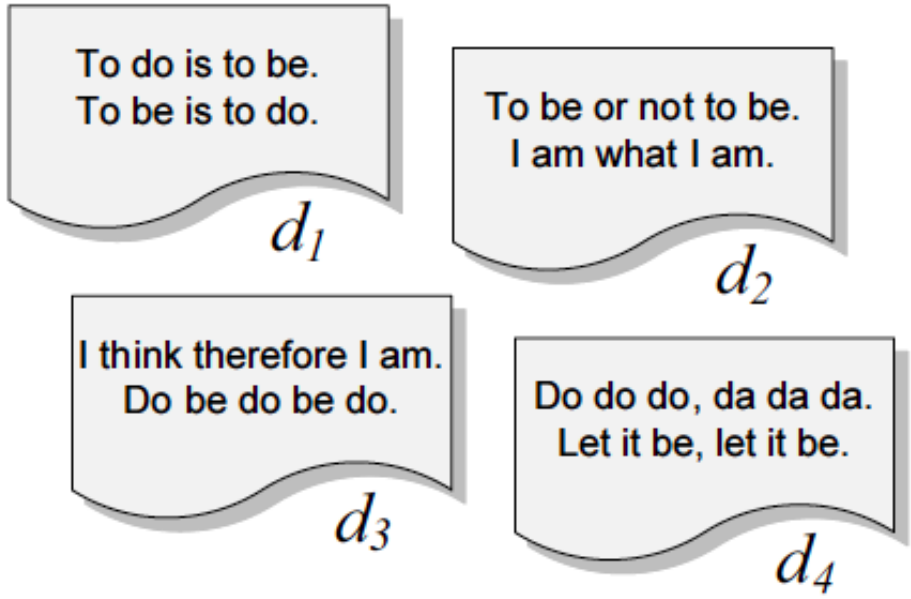
$$w_{do,1} = w_{to,1} = 1$$

Term Weighting

- Term frequency: the number of times a term appears in a document, $f_{i,j}$
- Total term frequency: the number of times a term appears in the whole collection, F_i
- Document frequency: the number of documents a term appears in, n_i

Example: the term *"do"*

$$f(do, d_1) = 2$$
$$f(do, d_2) = 0$$
$$f(do, d_3) = 3$$
$$f(do, d_4) = 3$$
$$F(do) = 8$$
$$n(do) = 3$$



Term Weighting

- The TF-IDF weighting rule
  Allows to take into account the importance of a term within a document, as well as throughout the whole collection.

TF-IDF = Term frequency - Inverse document frequency

For a collection of N documents:

$$IDF(k_i) = \log \frac{N}{DF(k_i)}.$$

IDF diminishes the weight of words that appear too often in the whole collection.

So, we have:

$$\text{TF-IDF}_{i,j} = TF_{i,j} \times IDF_i$$

Document Length Normalization

- Documents are of different sizes
- Longer documents -> more likely to be retrieved!!
- The rank of each document should be penalized (divided) by its length

In the case of vector representation (e.g., with TF-IDF weights) the length of a document is given by the vector norm.

$$\mathbf{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j})$$

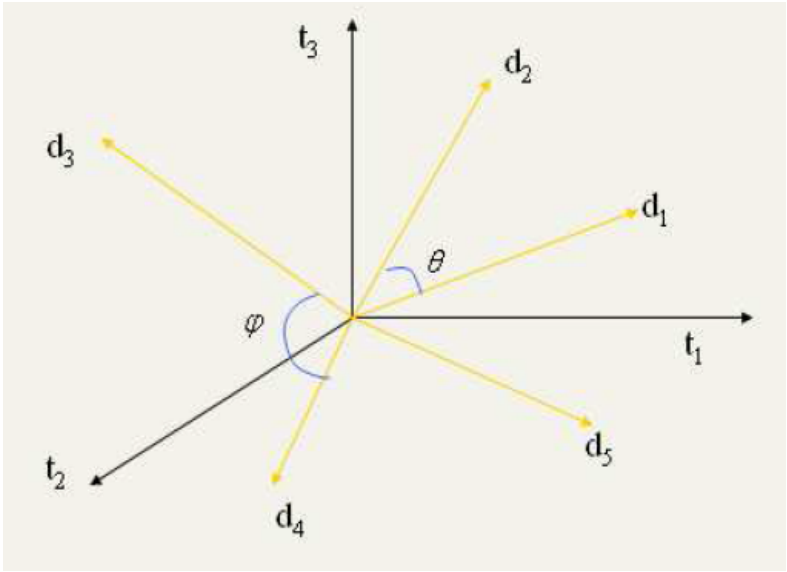
$$|\mathbf{d_j}| = \sqrt{\sum_{k=1}^{t} w_{k,j}^2}$$

Outline

Vector Space Model

Documents are represented as vectors in a common vector space.

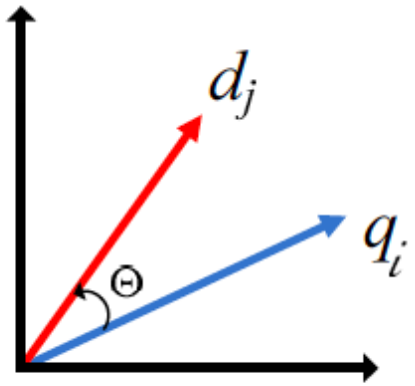
- Index terms = axes (dimensions) of the vector space
- Index terms are assumed to be independent
- The value of a given dimension for a given document corresponds to the weight of the corresponding term that defines that dimension



All documents "live" in the same vector space, same dimension, t.

Vector Space Model

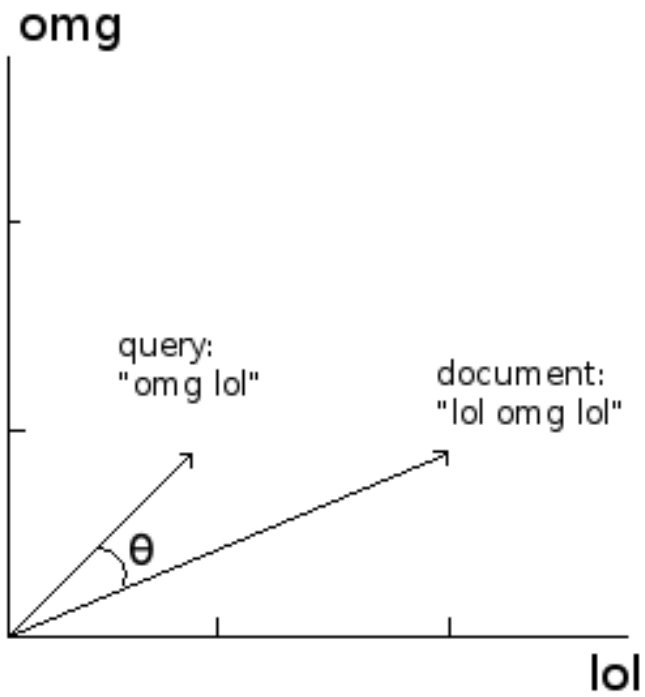
But also: the query is represented as a vector in the same vector space as the documents !



$$\mathbf{d_j} = (w_{1,j}, w_{2,j}, \ldots, w_{t,j}), \mathbf{q_i} = (w_{1,i}, w_{2,i}, \ldots, w_{t,i})$$

Vector Space Model

Example in 2 dimensions:



Now we can measure similarities and distances between documents and between a query and a document.

Vector Space Model

A common similarity measure: the cosine.

$$\cos(\theta) = \frac{\mathbf{d_j} \times \mathbf{q_i}}{\|\mathbf{d_j}\| \|\mathbf{q_i}\|}$$

In a given weighting scheme, we have:

$$sim(\mathbf{d_j}, \mathbf{q_i}) = \frac{\sum_{k=1}^{t} w_{k,j} \times w_{k,i}}{\sqrt{\sum_{k=1}^{t} w_{k,j}^2} \times \sqrt{\sum_{k=1}^{t} w_{k,i}^2}}$$

Vector Space Model

Pros

- Allows to use term weights. This improves the quality of the results
- Retrieve documents which partially correspond to a query
- A natural ranking scheme based on the cosine similarity
- Document length normalization is built-in the ranking function

Cons

- Term independence assumption

Vector Space Model: an Example

Documents:

d_1: "new york times"
d_2: "new york post"
d_3: "los angeles times"

The IDF values of the terms:

$idf(angeles) = log(3/1) = 1.584$
$idf(los) = log(3/1) = 1.584$
$idf(new) = log(3/2) = 0.584$
$idf(post) = log(3/1) = 1.584$
$idf(times) = log(3/2) = 0.584$
$idf(york) = log(3/2) = 0.584$

Vector Space Model: an Example

TF matrix:

	angeles	los	new	post	times	york
d1	0	0	1	0	1	1
d2	0	0	1	1	0	1
d3	1	1	0	0	1	0

TF-IDF matrix:

	angeles	los	new	post	times	york
d1	0	0	0.584	0	0.584	0.584
d2	0	0	0.584	1.584	0	0.584
d3	1.584	1.584	0	0	0.584	0

Vector Space Model: an Example

Query: q: "new new times"

We compute the tf-idf's of every term in the query and then the

Query TF-IDF matrix:

q	0	0	2 *0.584=1,168	0	1 *0.584=0.584	0

Compute the length of the documents and the query:
$|d_1| = \sqrt{(0.584^2 + 0.584^2 + 0.584^2)} = 1.011$
$|d_2| = 1.786$
$|d_3| = 2.316$
$|q| = 0.652$

Vector Space Model: an Example

Finally, the calculate the cosine similarities between the query and each document:

$$\text{cosSim}(d1,q) = (0*0+0*0+0.584*0.584+0*0+0.584*0.292+0.584*0) / (1.011*0.652) = 0.776$$
$$\text{cosSim}(d2,q) = (0*0+0*0+0.584*0.584+1.584*0+0*0.292+0.584*0) / (1.786*0.652) = 0.292$$
$$\text{cosSim}(d3,q) = (1.584*0+1.584*0+0*0.584+0*0+0.584*0.292+0*0) / (2.316*0.652) = 0.112$$

With respect to the similarity values, return the documents in the following order:

d_1
d_2
d_3

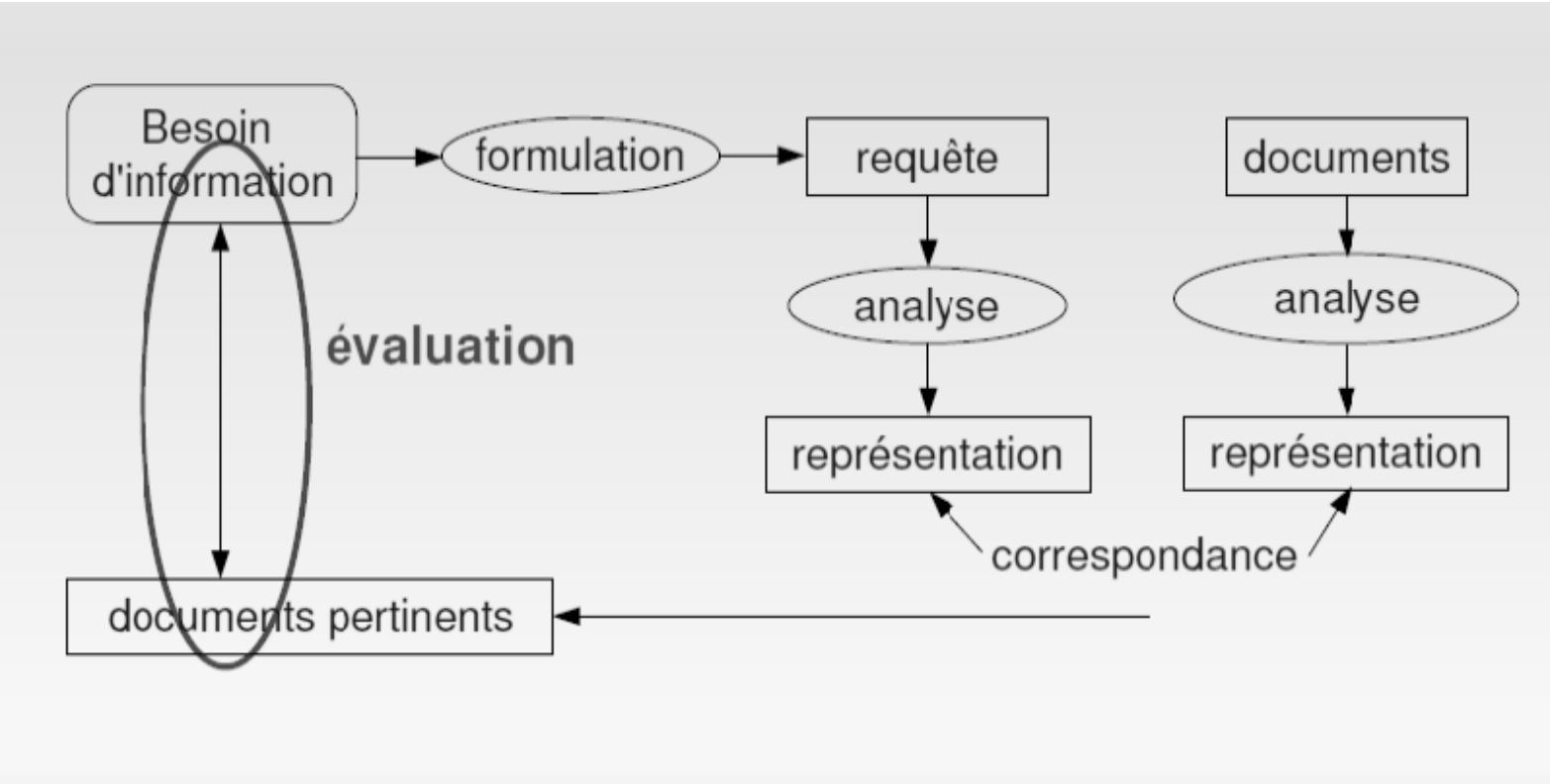
Models

Many more...

- Probabilistic model
- Fuzzy set theoretic model
- ...

Outline

The classical approach to information retrieval: evaluation



Evaluation

The quality of an IRS depends on many factors.

- Many different models, algorithms: which one is the best for my needs?
- Many different parameters within each model (similarity measure, term selection, term-weights): how to set these parameters?

The answers are an outcome of an empirical study.

Evaluation

Relevance of the results: depends on the user's satisfaction wrt. her information need.

Given a set of results: different groups of users => different degrees of satisfaction

Types of users:

- Web
- Entreprise
- Client/Vendor

Types of usage:

- Search to learn: *"The effect of global warming"*
- Search of a precise fact: *"The capital of Brazil"*
- Search of related information: *"Books about Montpellier"*

Evaluation

Exemple

- Besoin d'information : *Est ce que boire du vin rouge est plus efficace que le vin blanc concernant la réduction de risques au niveau du coeur ?*
- Requête : VIN ET ROUGE ET BLANC ET RISQUE ET COEUR ET EFFICACE

Exemple de document

Il s'est lancé dans le coeur de son discours et il a attaqué l'industrie viticole concernant sa sous-évaluation des risques du vin blanc et du vin rouge pour la conduite sous l'influence de drogues

Perfect match between a query and a document. OK ?...

The evaluation has to be done with respect to the information need and not to the query!

Evaluation

Measure the performance of an IRS:

we need a **quantitative metric** associated to the results produced by an IR system

In a reply to a given query, an IRS returns an ordered list of documents

- The order of the returned documents has to reflect their degree of relevance to the information need
- How to measure relevance (different models) -> how to evaluate the quality of the returned results

Evaluation

Classical ground truth methodology:

- A set of documents
- A set of queries
- A relevance judgement: a measure for every pair (document, query)

A bit of history: the Cranfield initiative (1950s):

- Experiments on manually collected documents
- Study of 2 quantities: the fraction of relevant documents retrieved, and the fraction of retrieved documents that are relevant
- First insights, first test **reference collection**

Reference collections, which are based on the foundations established by the Cranfield experiments, constitute the most used evaluation method in IR

A reference collection is composed of:

- A set $\mathcal{D}$ of pre-selected documents
- A set I of information need descriptions used for testing
- A set of relevance judgements associated with each pair $[i_m, d_j]$, $i_m \in \mathcal{I}$ and $d_j \in \mathcal{D}$

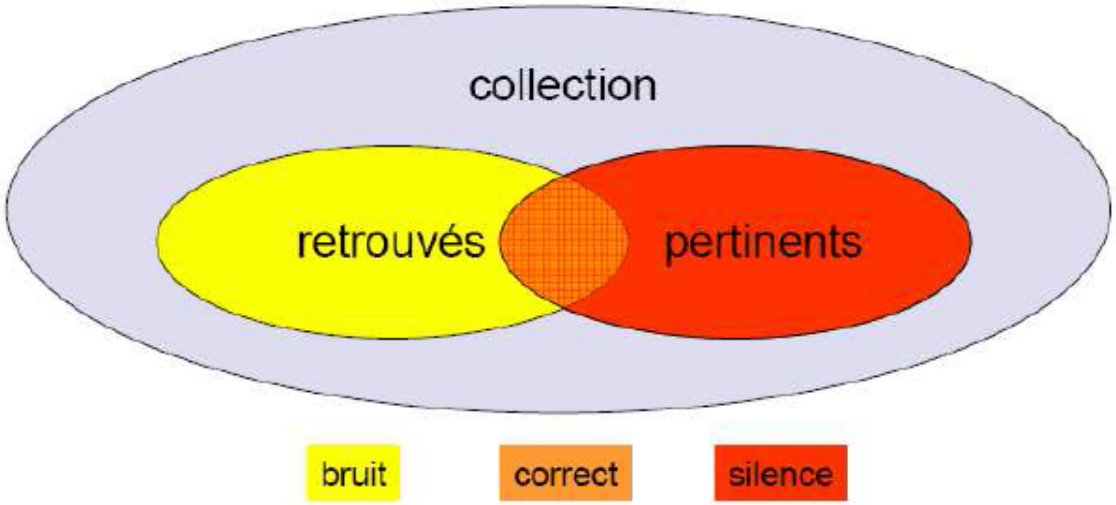
The relevance judgement has a value of 0 if document d_j is non-relevant to i_m, and 1 otherwise

These judgements are produced by human specialists

We can partition the documents in 4 categories:

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

tp = true positives, **fp** = false positives **tn** = true negatives, **fn** = false negatives



Precision and Recall

$$P = \frac{relevant\_found}{all\_found}, \qquad R = \frac{relevant\_found}{all\_relevant}.$$

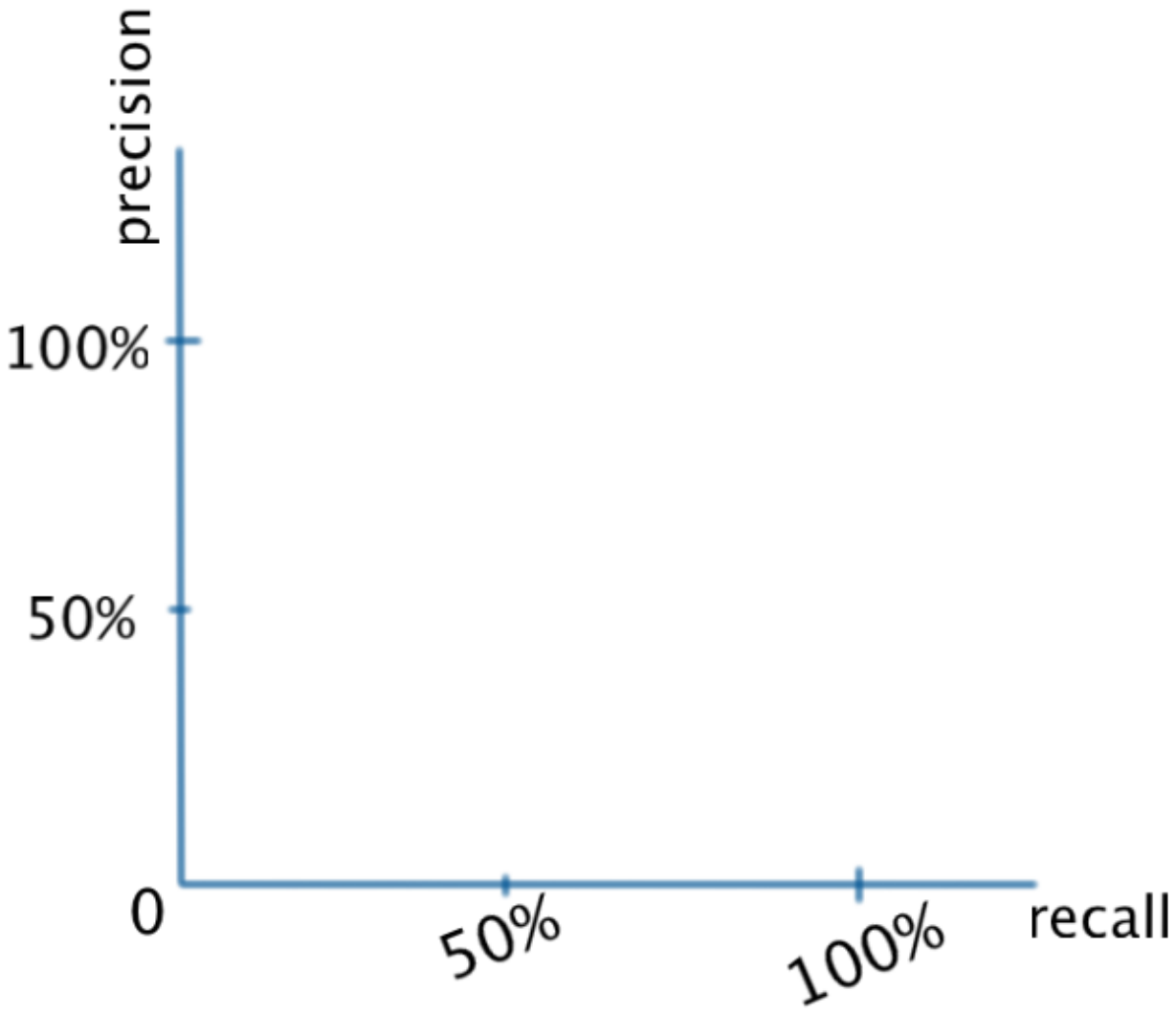
Some observations

- The definitions of precision and recall are based on the whole set of found documents
- The user is usually not presented with the whole set of found documents
- User goes through a ranked list of documents, examining them one by one, starting from the top
- Precision and recall change as the user advances in the list of found documents

Therefore, it is convenient to plot a curve showing the evolution of P and R

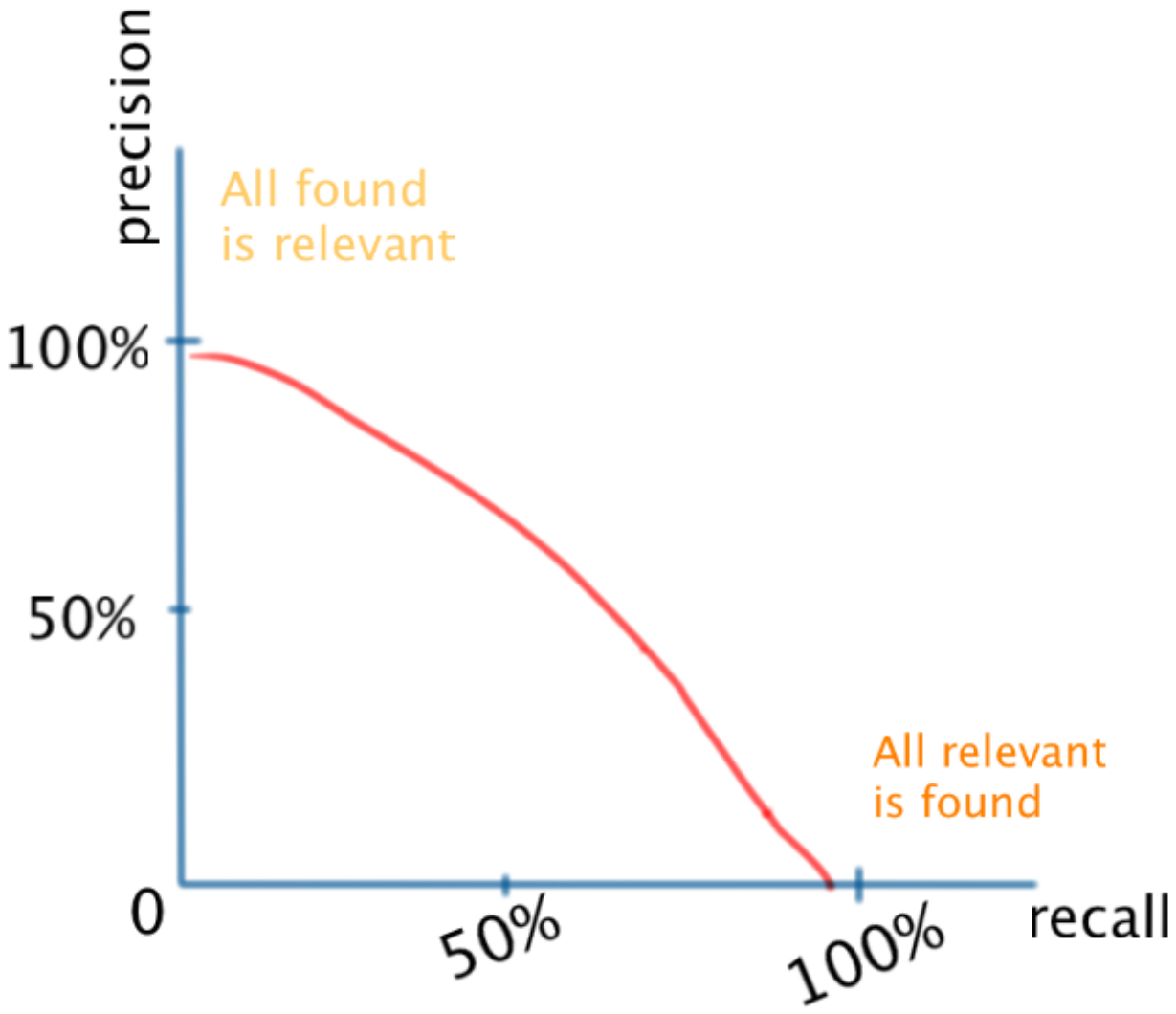
Evaluation

Precision and Recall



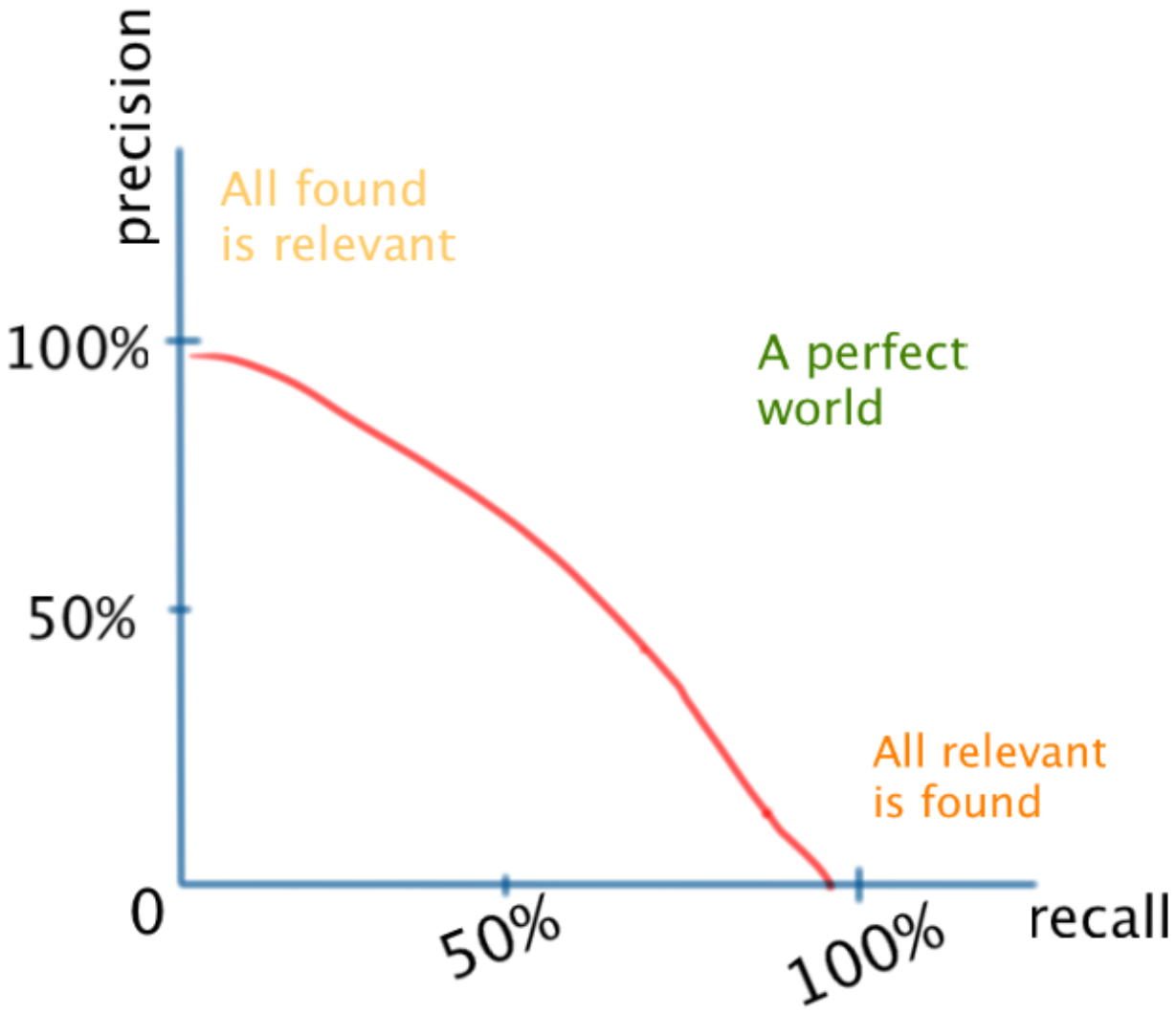
Evaluation

Precision and Recall



Evaluation

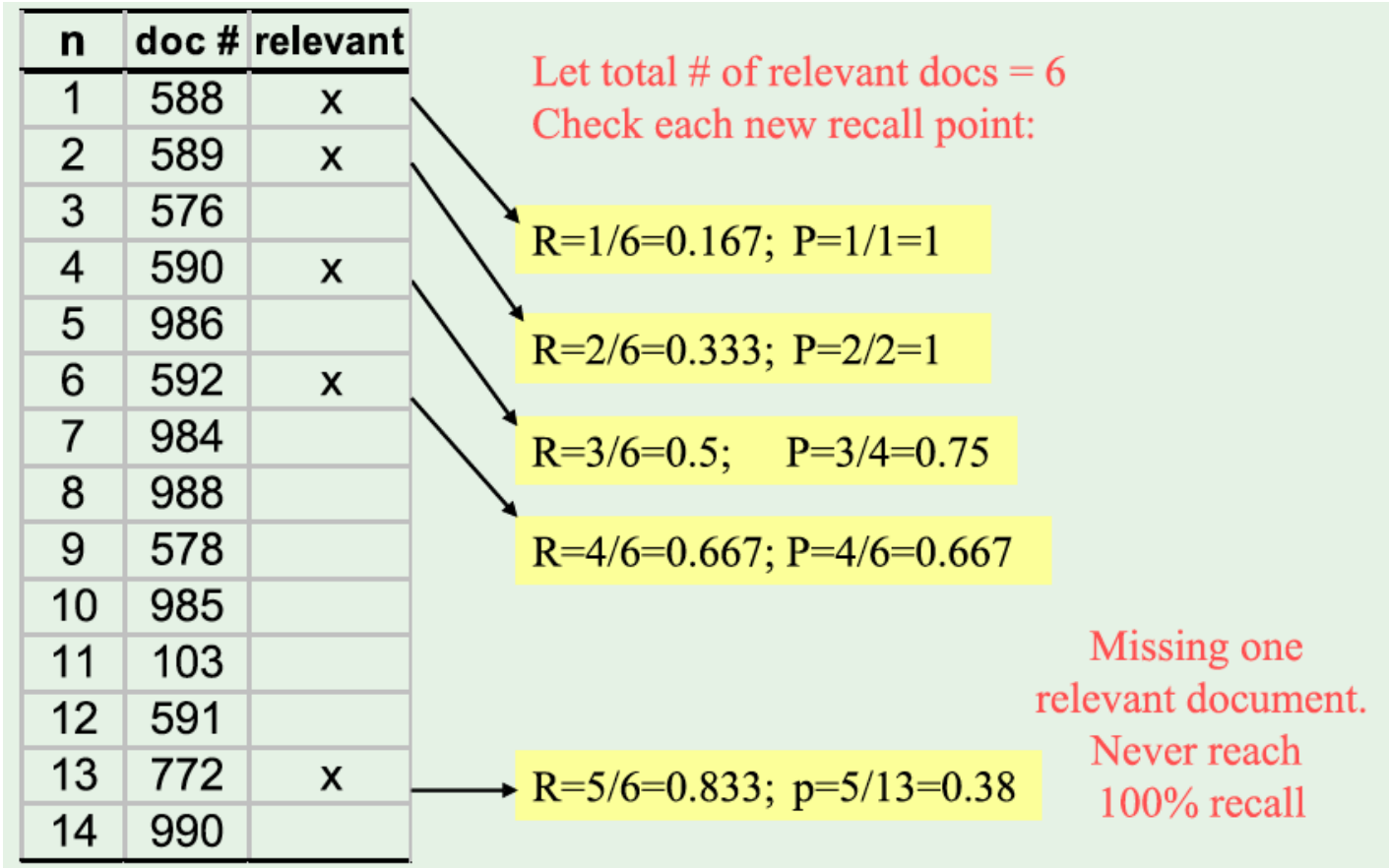
Precision and Recall



Evaluation

Precision and Recall

An example:



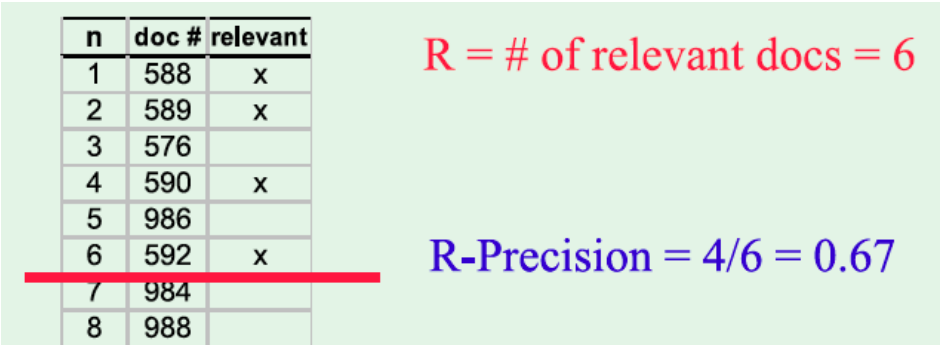
We suppose we know the number of relevant documents.

Evaluation

Single Value Measures

Often, we'd prefer one single value to express the performance of the IRS.

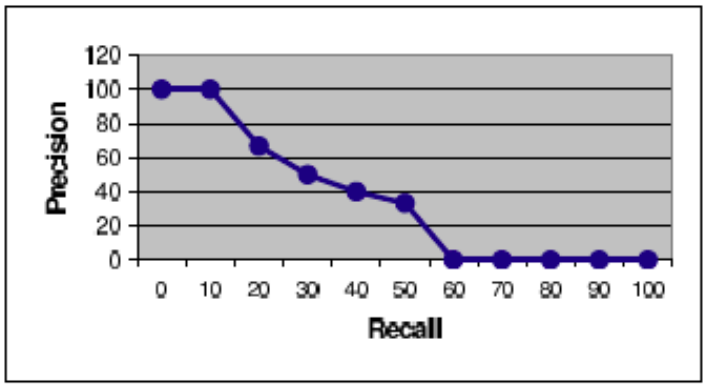
- Precision over the top *n* found documents:
  - e.g., *P*@5, *P*@10 (5 or 10 documents have been seen)
  - good precision on the top of the list -> positive user impression
  - web search
- R-Precision:
  - *R* - the total number of relevant docs for a given query
  - compute precision at the *R*th position of the ranking
  - e.g., $R = 6$, but 4 relevant in the top 6 => R-Precision = 4/6



Evaluation

Single Value Measures

- MAP: mean average precision over the first n documents
  - e.g., for 10 documents: $MAP = \frac{P1+P2+...+P10}{10} = \frac{1+1+0.6+0.5+0.4+0.33+0+0+0+0}{10}$



Recall	Precision
0	100
10	100
20	66.6
30	50
40	40
50	33.3
60	0
70	0
80	0
90	0
100	0

- F-measure (the harmonic mean of precision and recall at the j-th position of the ranking)
  - $F(j) = \frac{2P(j)R(j)}{P(j)+R(j)} \in [0,1]$
  - high F <= both P and R high
  - good compromise between P and R

Evaluation

Accuracy?

	Relevant	Nonrelevant
Retrieved	tp	fp
Not Retrieved	fn	tn

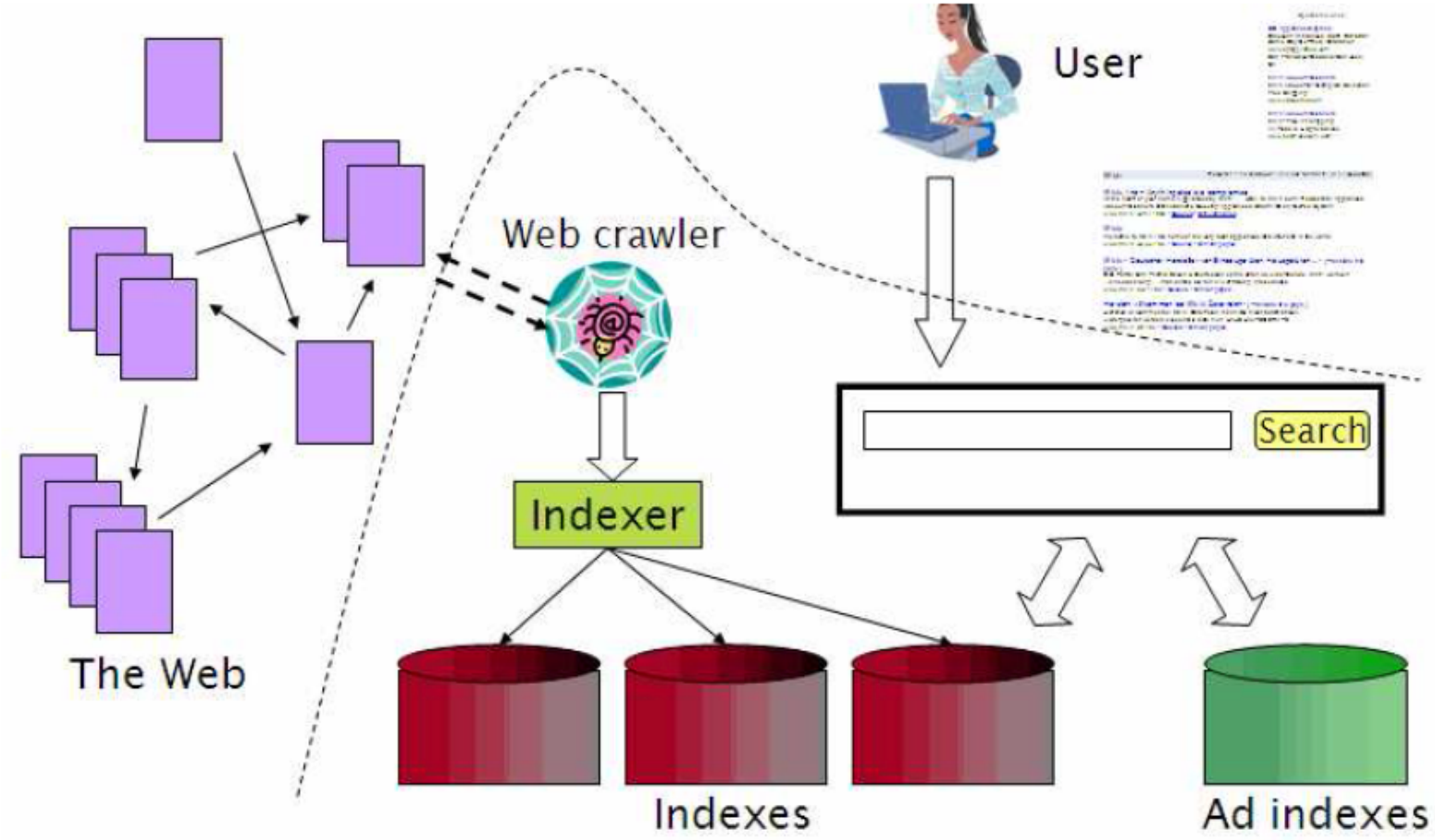
$$Accuracy = \frac{tp+tn}{all}$$

- A measure often used in classification (see following lecture)
- Not appropriate for IR because...

- Data are biased: about 99.9% of the documents are non relevant
- We would like to tolerate error: an IR system is expected to return something!

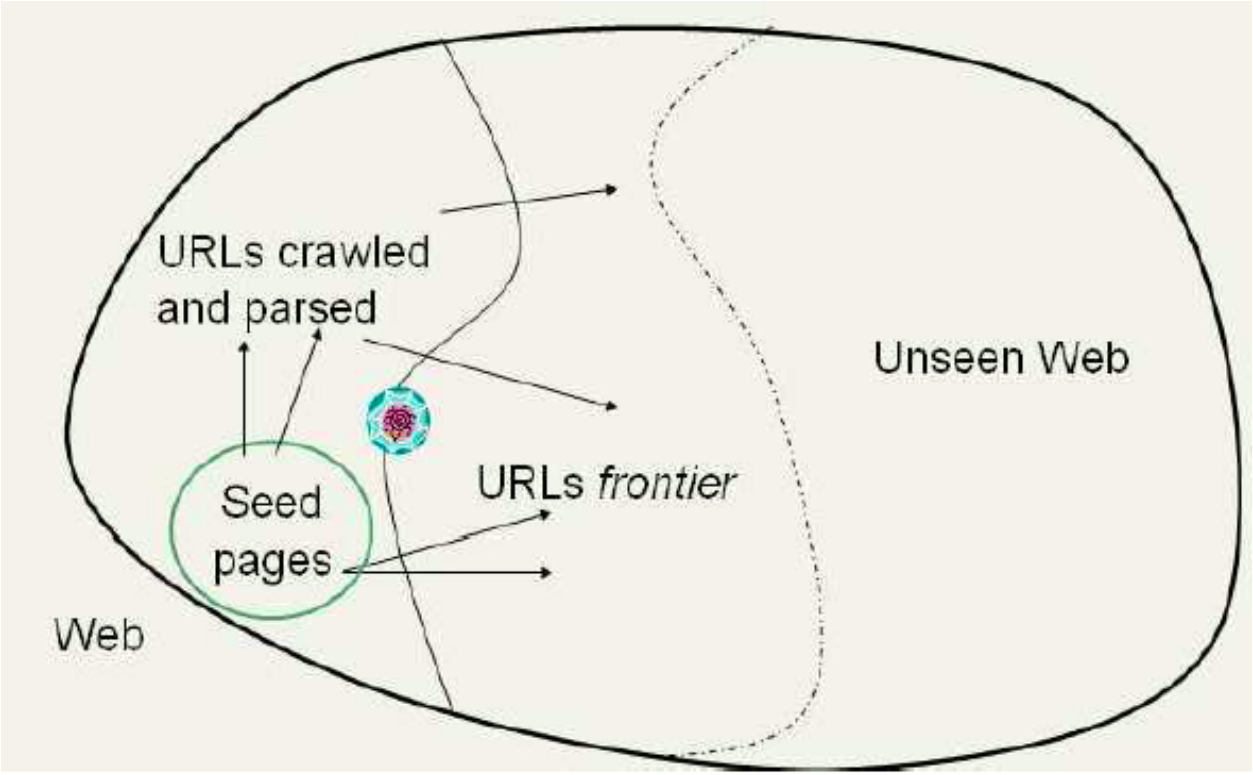
Outline

Web Search



From H. Schuetze

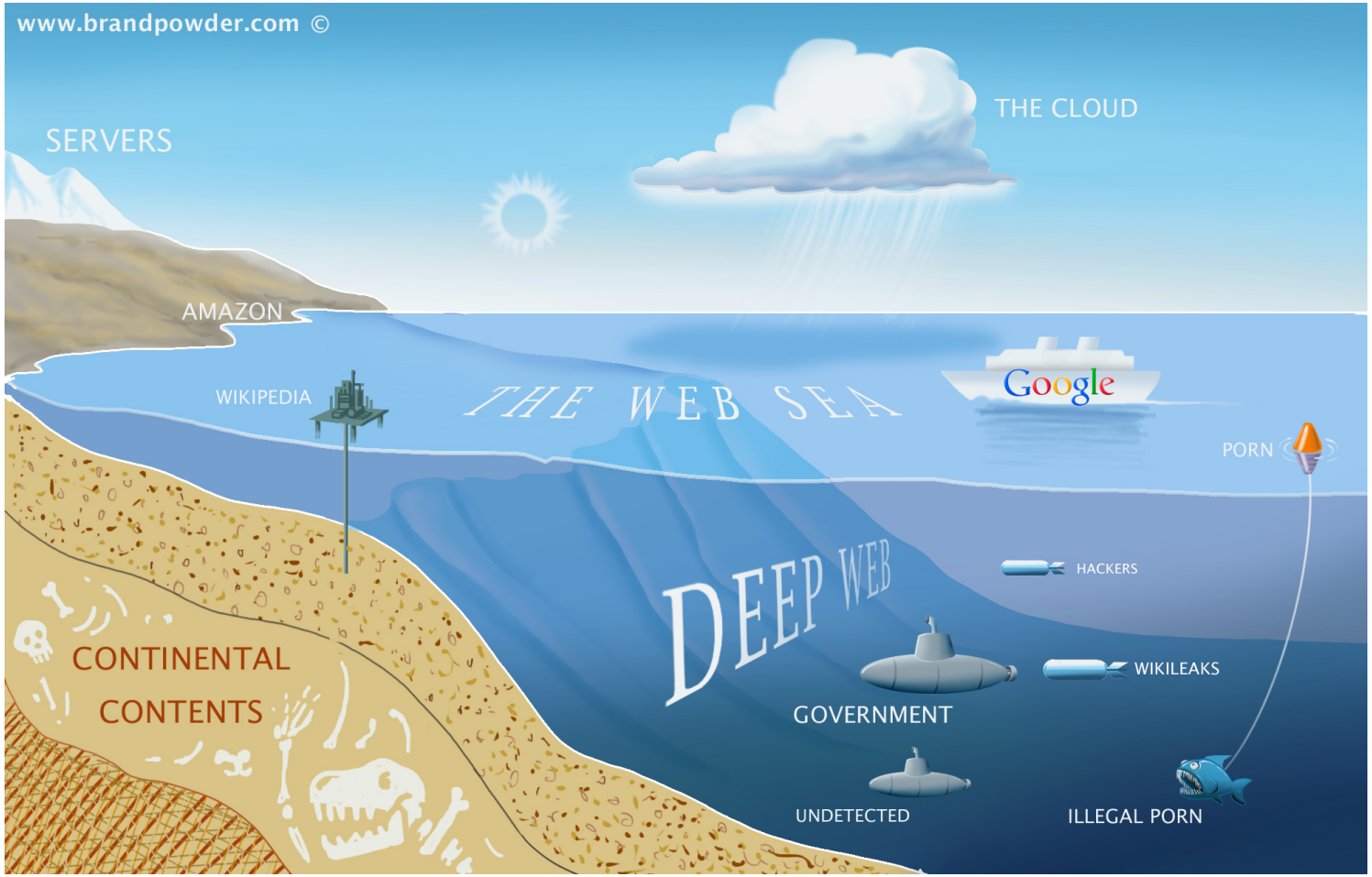
An idea about crawling...



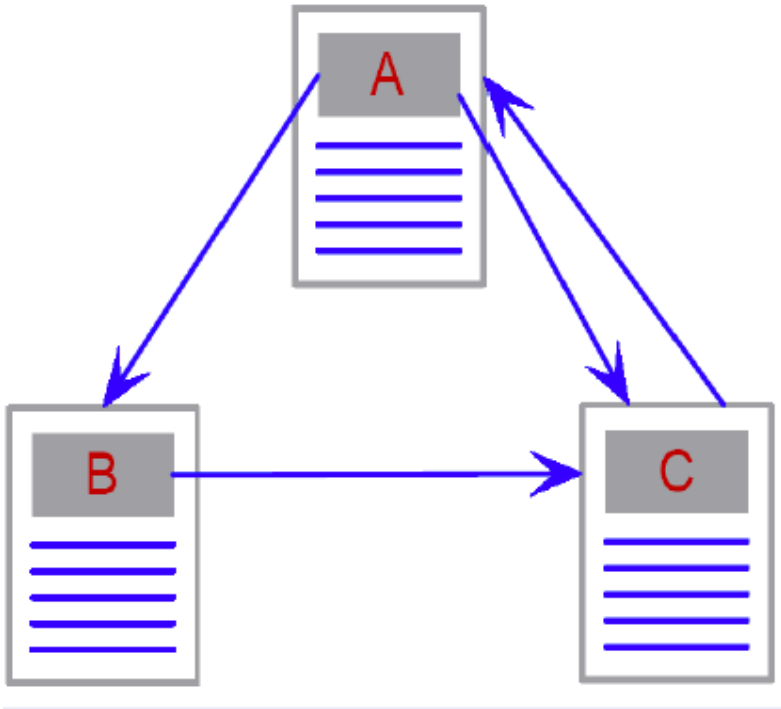
Web Search

The Deep Web?

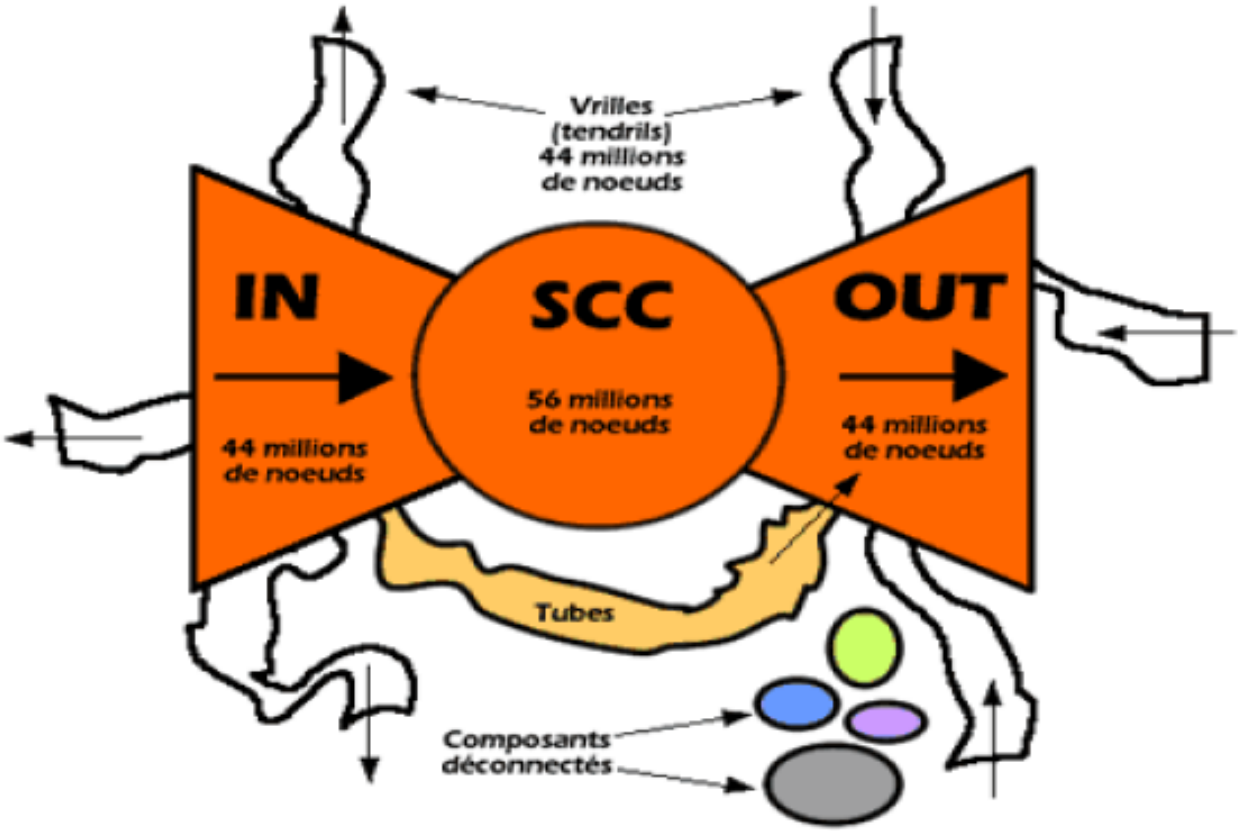
Search engines index about 6% of the entire web content.



The static web: a graph representation induced by the hyperlinks.



The Bow-tie model of the web:



Three main page categories: IN, OUT, SCC (Strongly Connected Component)
From Broder et al.

Web Search

Link Analysis

- Google's PageRank (Brin and Page, 1998)
- Many different variants of this algorithm

Main idea:

- The Web is an oriented graph
- The links carry important information
  - A link between two pages indicates a relevance relation between them
  - The text of the link describes the target page: important for indexing !

    ...well, not always:
    "You can find useful information about New Delhi's economy here."

Web Search

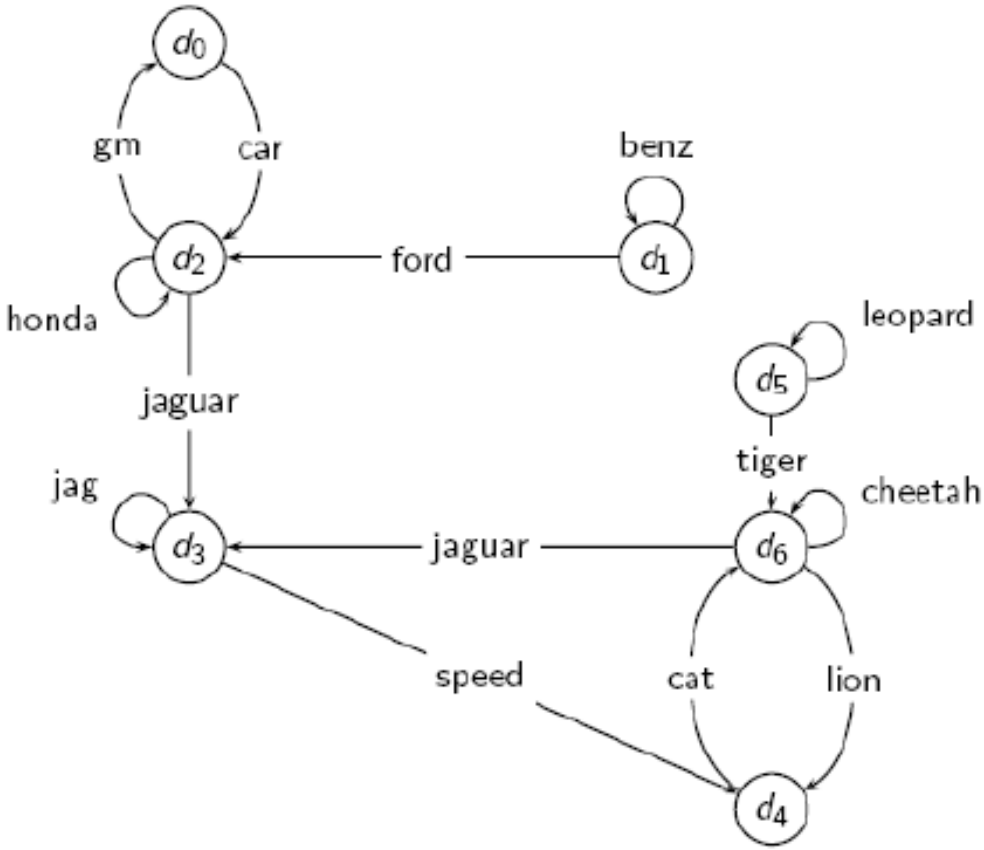
The algorithm in a nutshell:

- A random walk through the Web graph (modeled as a Markov chain)
- At each step, the walker leaves a page by following one of its links (with an equal probability)
- If a page has no out-links: jump on to an arbitrary page with a given probably p
- The PageRank is the probability that the walker is at page d, after a given time, t

–> "The likelihood that a person randomly clicking on links will arrive at a certain page, after a time t."

–> A rough estimate of how important a website is.

Example:

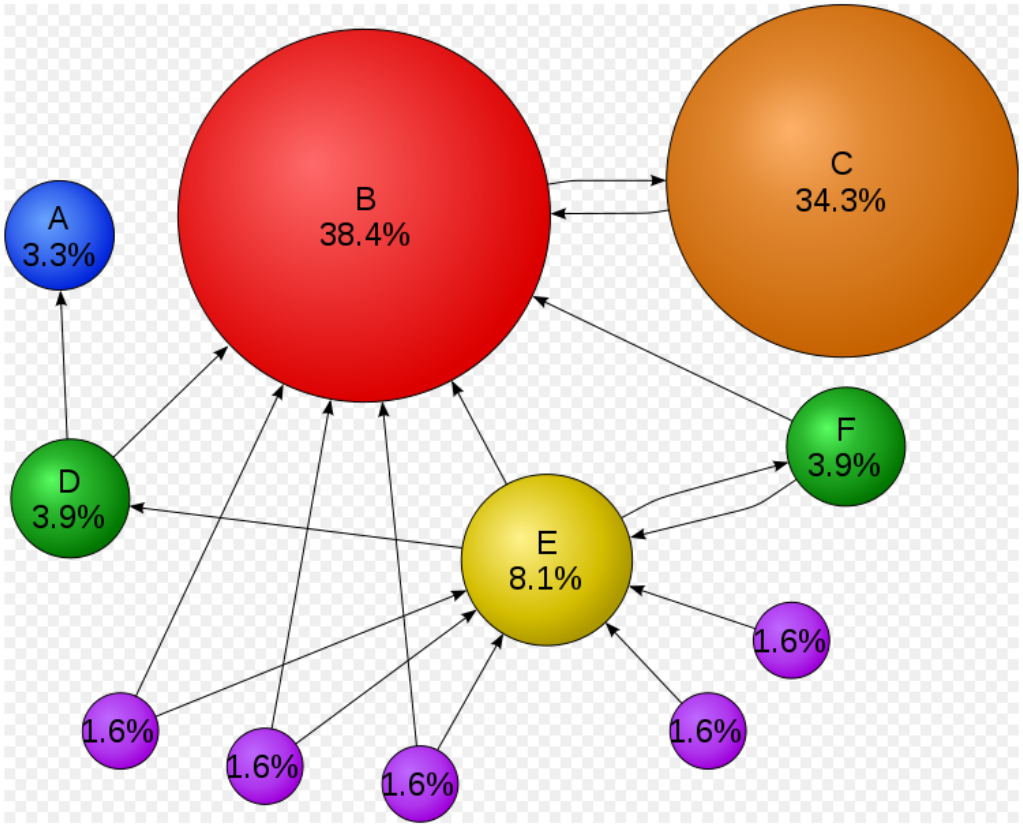

d_0
gm
car
benz
d_2
ford
d_1
honda
leopard
d_5
jaguar
tiger
jag
cheetah
d_3
jaguar
d_6
speed
cat
lion
d_4

Example: the transition matrix.

	d_0	d_1	d_2	d_3	d_4	d_5	d_6
d_0	0.00	0.00	1.00	0.00	0.00	0.00	0.00
d_1	0.00	0.50	0.50	0.00	0.00	0.00	0.00
d_2	0.33	0.00	0.33	0.33	0.00	0.00	0.00
d_3	0.00	0.00	0.00	0.50	0.50	0.00	0.00
d_4	0.00	0.00	0.00	0.00	0.00	0.00	1.00
d_5	0.00	0.00	0.00	0.00	0.00	0.50	0.50
d_6	0.00	0.00	0.00	0.33	0.33	0.00	0.33

Web Search

The importance of a page depends not only on the number of links.



source: Wikipedia

So, let's start.

1. $PR(p) = \sum_{p_i -> p} 1$

OK, but if p_i has many out-links, the probability of going to p becomes smaller!

2. $PR(p) = \sum_{p_i -> p} \frac{1}{out(p_i)}$

That's better. But if p_i is a very important page?

3. $PR(p) = \sum_{p_i -> p} \frac{PR(p_i)}{out(p_i)}$

And what if I don't follow the links, but decide to jump to a random page?

4. $PR(p) = \frac{(1-d)}{N} + d \times \sum_{p_i -> p} \frac{PR(p_i)}{out(p_i)}$

Web Search

When a query comes:

- Retrieve the pages that satisfy the query (see standard models)
- Rank the pages by using PageRank
- Return to the user the ranked list of pages

Outline

Multimedia IR

What about images, video, sound...?

Indexing: text, manual

Users are responsible for providing textual information about an image:

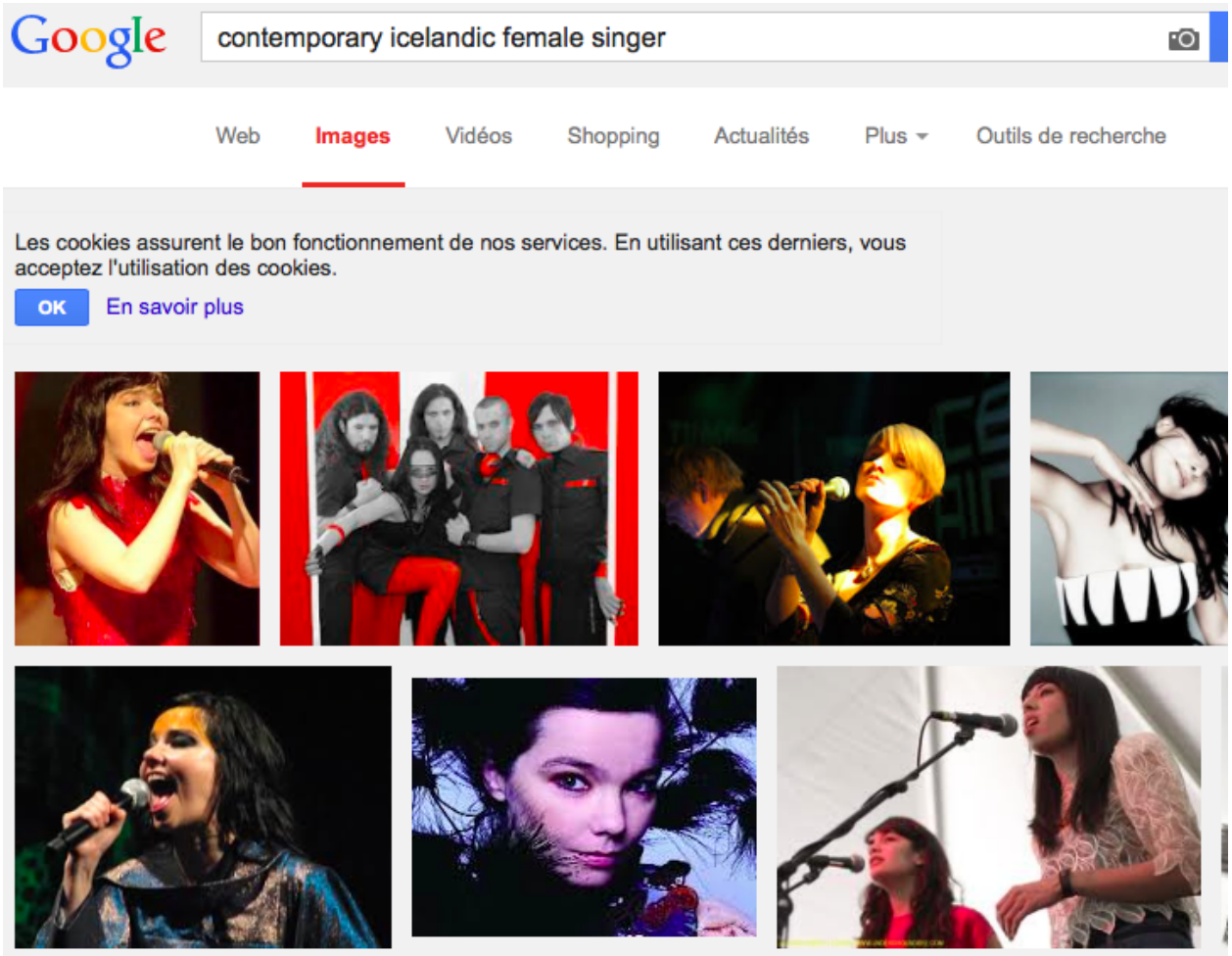
bjork, singer, songwriter, Iceland, female, dancer in the dark, ...

- expensive, depends on the user background, contextual, multilinguism

Indexing: text, automatic

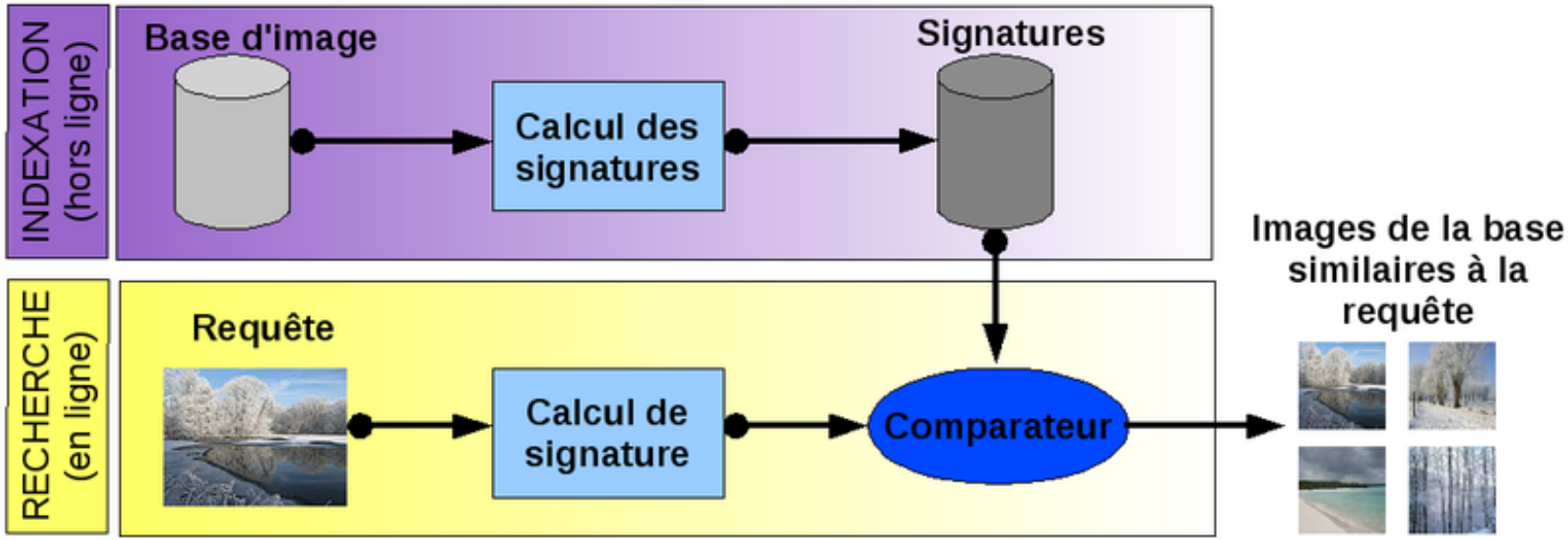
Multimedia IR

If an image is indexed by using textual information,
the query is a set of key words (see previous sections).



Indexing: visual content

- Automatic extraction of numerical descriptors (features) from an image
- Field: content-based image retrieval (CBIR)

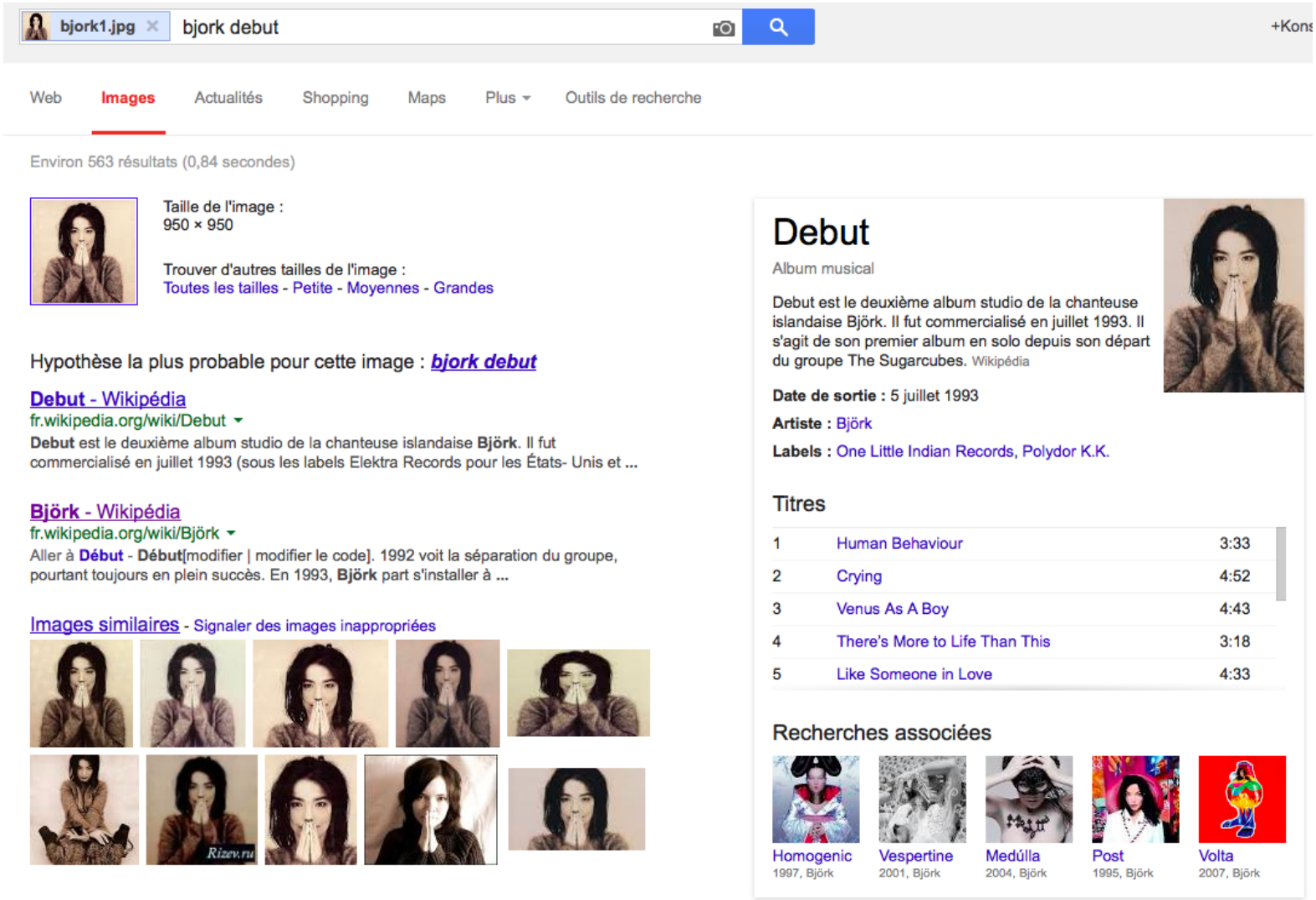


Source: Wikipedia

- Open research problem in CBIR: the semantic gap.

Multimedia IR

If an image is indexed by using its content, the query is an image.



Often a hybrid search...

Outline

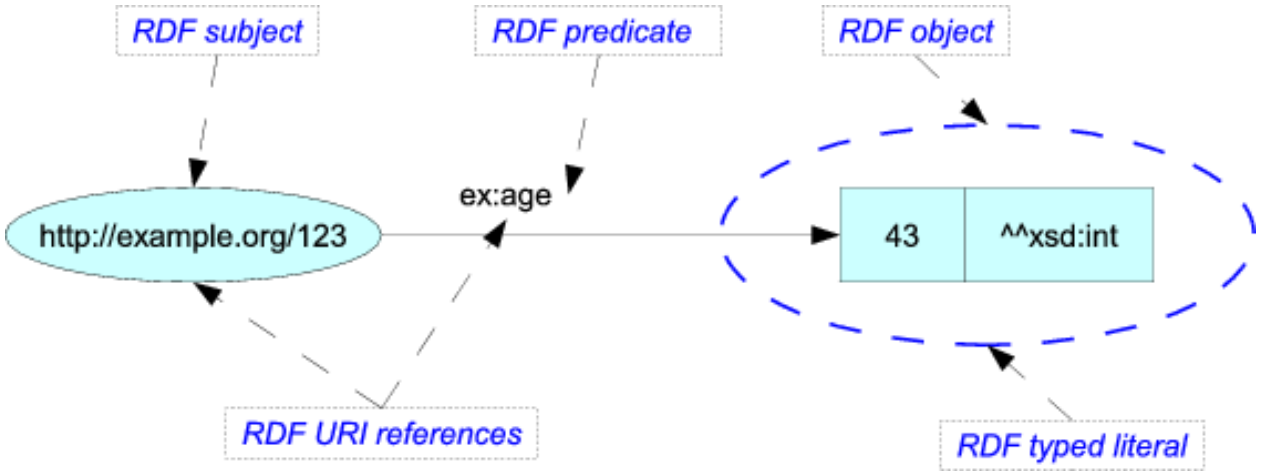
The Web of Data

The web of data, or the semantic web - a new vision of the web.

- From a web of documents -> to a web of data
- Not about how documents are connected, but about how things are connected
- A global network of information
- Discover connected and relevant pieces of information that are not found on one single page, nor on pages linked by hyperlinks
- Using semantics and strict rules of publishing data: we will come back to that in a further lecture.

The Web of Data

The web of data, or linked open data - a new vision of the web.
Publish data in the form of a structured dataset using the RDF formalism.



IR: more like querying a database by using an adapted query language (SPARQL).

The Web of Data

- Today: answers to complex queries are embedded (or buried) in many result pages that the human has to read in order to extract the desired information
- Structuring the web –> ongoing research
- Imagine a data base on the scale of Wikipedia.
- Bridging data base methodology to web search
- Ongoing tentatives: DBpedia, Freebase, Google Knowledge Graph, etc...

Weikum, G., Theobald, M. (2010). From information to knowledge: harvesting entities and relationships from web sources. In Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 65-76). ACM.

The Web of Data

Environ 4 430 000 000 résultats (0,24 secondes)

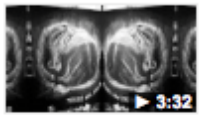
Les cookies assurent le bon fonctionnement de nos services. En utilisant ces derniers, vous acceptez l'utilisation des cookies.

OK En savoir plus

Conseil : Recherchez des résultats uniquement en **français**. Vous pouvez indiquer votre langue de recherche sur la page Préférences.

The National
americanmary.com/ ▾ Traduire cette page
Join Our Mailing List. **The National** - Trouble Will Find Me. New Album Out Now. iTunes / Amazon / Vinyl. NEWS. We'd like to hear your cover of "I Need My Girl.
Tour Archive - Trouble Will Find Me - Think You Can Wait / Exile Vilify

The National (groupe) - Wikipédia
fr.wikipedia.org/wiki/**The_National**_(groupe) ▾
The National en concert à la Brooklyn Academy of Music en 2010. Informations générales. Pays d'origine, Cincinnati, Drapeau des États-Unis États-Unis.
Histoire du groupe - Membres du groupe - Discographie - Bandes originales

The National - "Demons" - YouTube


www.youtube.com/watch?v=N527oBKIPMc ▾
8 avr. 2013 - Ajouté par thenationalofficial
From the album Trouble Will Find Me, out now: http://bit.ly/10NQPA1 Illustration and video by Azar Kazimir ...

Latest and breaking news | **thenational**.ae - **The National**
www.**thenational**.ae/ ▾ Traduire cette page
Get the latest news, business, sport, lifestyle, arts, culture and more from **The National**, the leading English-language voice in the Middle East - **thenational**.ae.

Pourquoi je déteste **The National** | Slate
www.slate.fr/story/75063/pourquoi-je-deteste-**the-national** ▾
15 juil. 2013 - **The National** me donne l'impression que le rock –comme une grande

The National

Compositeur

The National est un groupe américain de rock indépendant formé à Cincinnati en 1999 et basé à Brooklyn. Il se compose de l'auteur-compositeur-interprète Matt Berninger et des quatre frères Dessner, Aaron et Bryce, et Devendorf, Bryan et Scott. Wikipédia

Membres : Matt Berninger, Bryce Dessner, Aaron Dessner, Bryan Devendorf, Scott Devendorf

Labels : Beggars Banquet Records, 4AD, Brassland Records

Origine : Brooklyn, État de New York, États-Unis, Cincinnati, Ohio, États-Unis

Sélections : Grammy Award du meilleur album de musique alternative, plus…

Événements à venir

4 févr. mar.	The National Auckland
6 févr. jeu.	The National Torrensville
7 févr. ven.	The National Sydney

Semantics: The National is a band and they give a concert on 4th February in Auckland.

Retrieval vs. mining?

IR and Data Mining

Information retrieval vs. data mining

- IR is about finding what is already there: as fast as possible
- DM is about finding something new in your data that was not known before: as new as possible

Combine both

- group together the documents with respect to their content (see following lecture on clustering and classification)
- disambiguate the different "meanings" of a query and present the results in categories to the user
- diversify
- user profiling and recommendation: find patterns in user search
- query expansion and reformulation
- ...

Sources and further reading

This course uses references from

- a course by Céline Hudelot given at Ecole Centrale Paris. This is Céline's homepage: http://perso.ecp.fr/ hudelotc/
- a course by R. Baeza-Yates and B. Ribeiro-Neto. Can be found here: http://www.mir2ed.org

Here are two very useful books:

- C. Manning, P. Raghavan, H. Schuetze. Introduction to Information Retrieval, Cambridge University Press, 2008.

  Available online here: http://nlp.stanford.edu/IR-book/information-retrieval-book.html
- R. Baeza-Yates, B. Ribeiro-Neto. Modern Infromation Retrieval. Addison Wesley

Questions?...