

Introduction à la recherche d'information

Mohand Boughanem

bougha@irit.fr

<http://www.irit.fr/~Mohand.Boughanem>

Université Paul Sabatier de Toulouse

Laboratoire IRIT , UMR5055



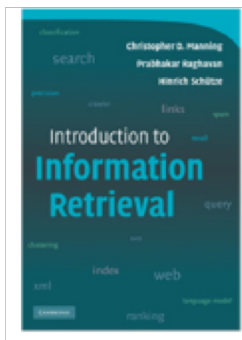
- Fondements de la Recherche d'information (RI)
 - Définition et contours de la RI
 - Problématique de la RI
 - Ouvrir la boîte d'un système de RI
- Quelques thématiques de recherche en RI
- Conclusion

Qu'est ce que la RI ?

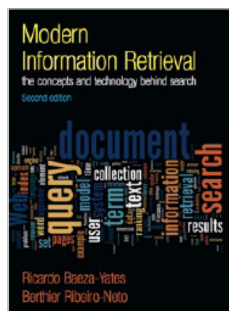
- Recherche d'information (RI) :
 - Ensemble des méthodes et techniques pour l'acquisition, l'organisation, le stockage, la recherche et la **sélection d'information pertinente pour un utilisateur**



Qu'est ce que la RI ?



IR is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).



Information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogs, structured and semi-structured records, multimedia objects. The representation and organization of the information items should be such as to provide the users with easy access to information of their interest.



IR: The techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system.

Qu'est ce que la RI ?



Le premier ouvrage francophone traitant de la recherche d'information



Qu'est ce que la RI ?



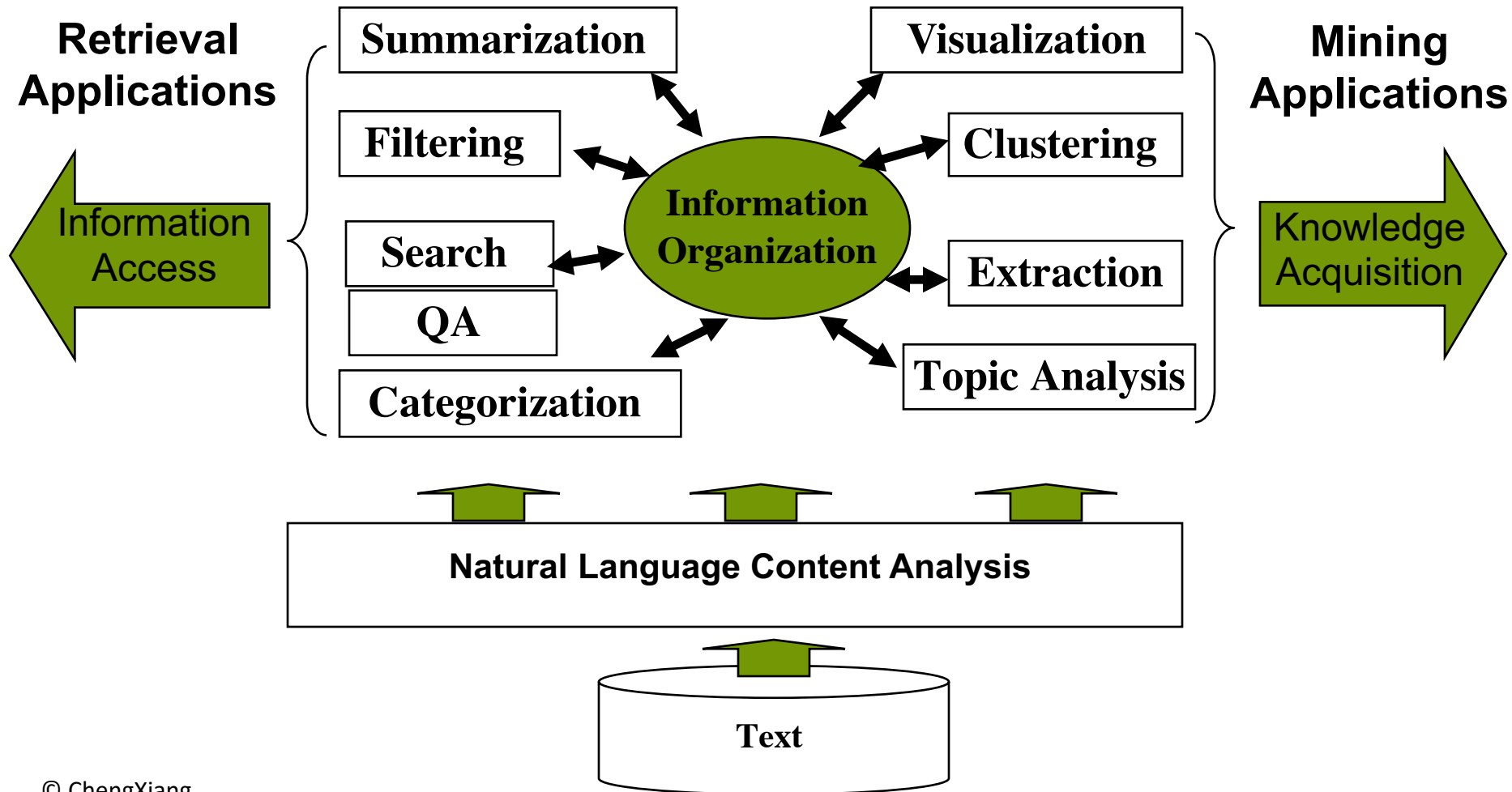
.. Mais pas seulement

- Plusieurs domaines d'application
 - Internet (Web, Forum/Blog search, news)
 - Entreprises (entreprise search)
 - Bibliothèques numériques «digital library»
 - Domaine spécialisé (médecine, droit, littérature, chimie, mathématique, brevets, software, ...)
 - Nos propres PC (Yahoo! Desktop search)

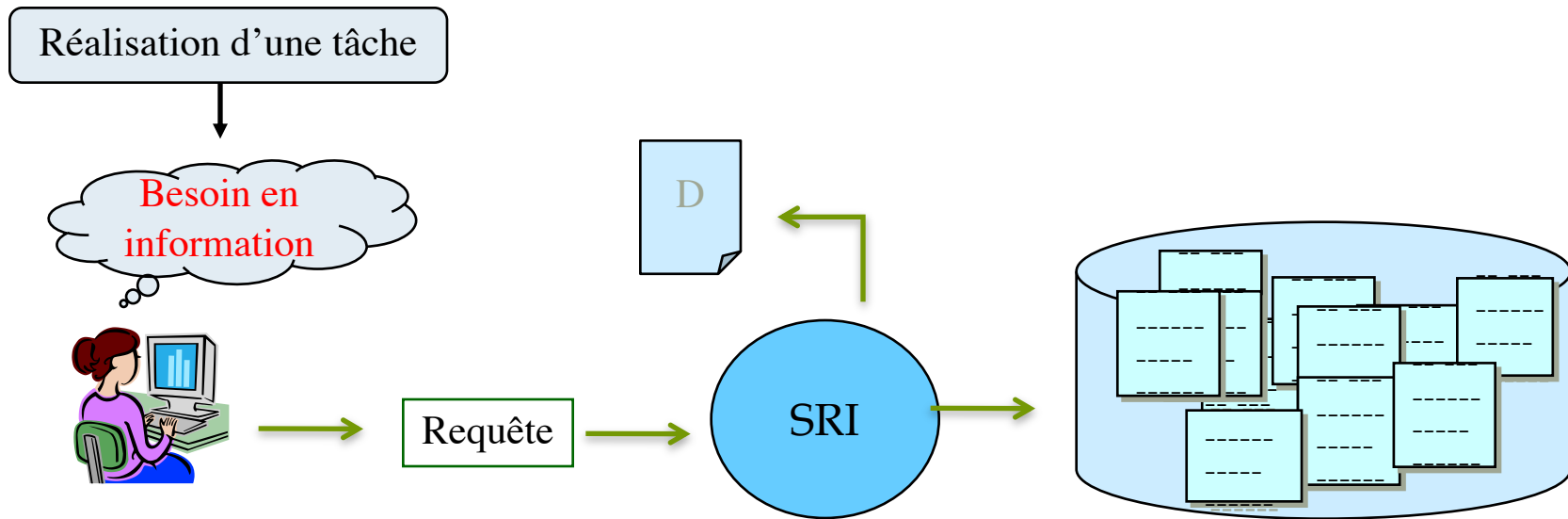
RI et domaines liés

- Recherche d'information s'intéresse à l'accès à des textes (documents textuels)
- Domaines liés
 - Fouille de textes (Text Mining)
 - Détection des « patterns » dans un texte (latent topics, ...)
 - Extraction d'information
 - Représentation des connaissances
 - Traitement automatique de la langue (NLP)
 - Compréhension d'un texte
 - Représentation sémantique d'un texte
 - Bases de données
 - Gestion de données structurées
 - Requêtes précises

Tâches de RI : vue large



Focus sur le « search » : RI adhoc



- Sélectionner dans une collection
 - les informations (items, documents, ..)
 - ... pertinentes répondant aux
 - ... besoins en information des utilisateurs

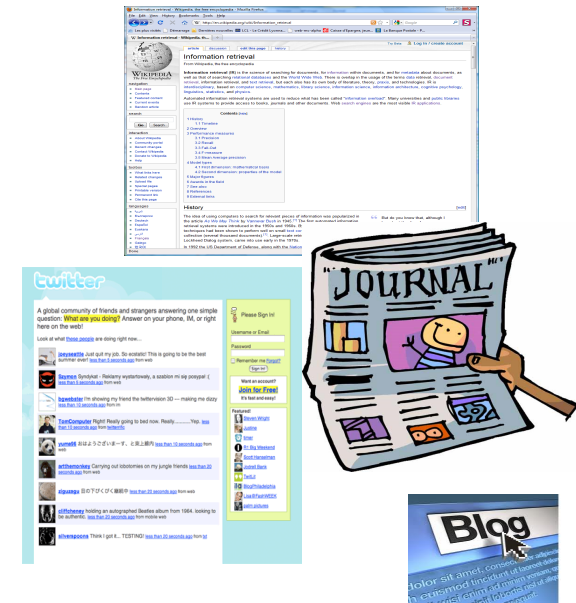
Information – document

● Nature

- **Texte**, images, sons, vidéo, graphiques, etc.
- Exemples texte : web pages, email, livres, journaux, publications, blog, Word™, Powerpoint™, PDF, forum postings, brevets, etc.

● Hétérogénéité

- langage (multilingues)
- media (multimédia)

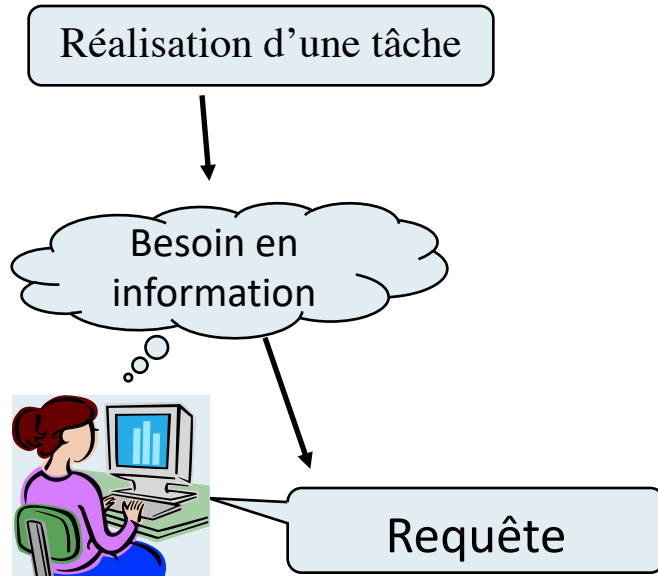


Question



- Information, document, unité/granule/passage
- Comment représenter le contenu du document (caractères, mots simples, phrases, paragraphes)
 - → Ambiguïté du langage naturel (polysémie, synonymie, ...)

Besoin en information (information needs)



- Besoin en information lié à une tâche
 - dépend de la tâche
 - une expression mentale d'un utilisateur
- Requête
 - Ensemble de mots-clés
 - → Une représentation possible du besoin en information



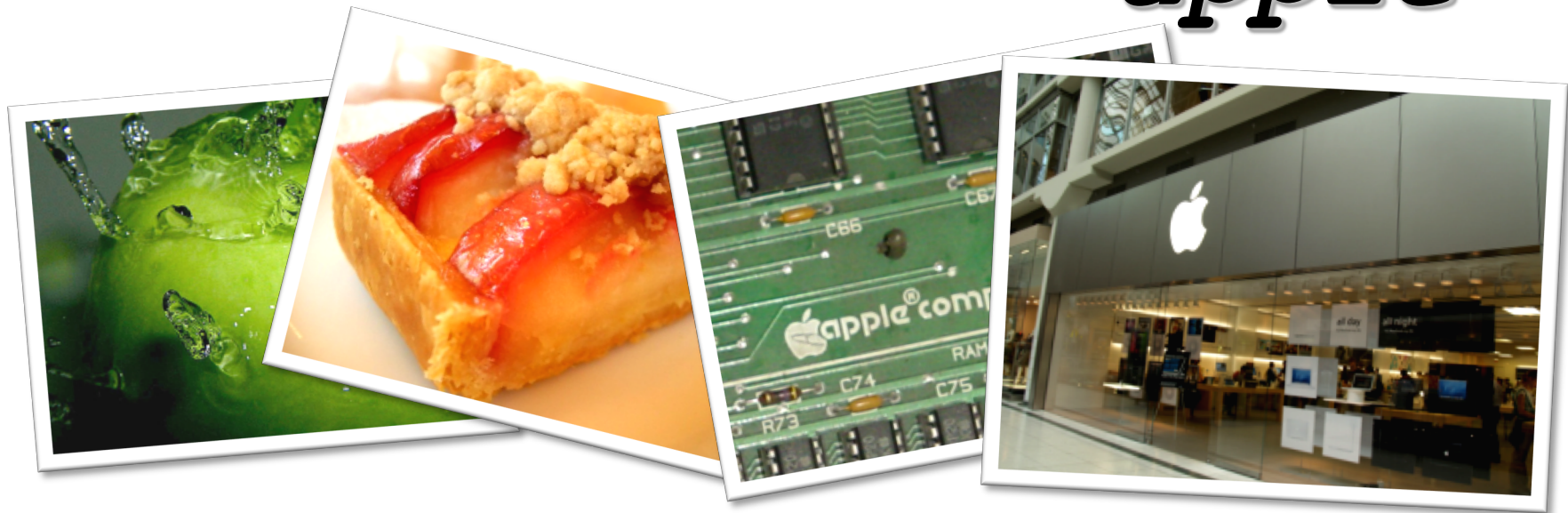
Question

- Comment capturer le besoin de l'utilisateur (l'intention de l'utilisateur)

Besoin en information-Intention de la requête (query intent)

What's in a query?

apple



© David J. Brenes, Daniel Gayo Avello, Kilian Pérez-González

Pertinence - Relevance

- Au cœur de tout système de RI
 - Relation entre le document et ... la requête ou le besoin de l'utilisateur ?
- Plusieurs facteurs influencent la décision de l'utilisateur, tâche, le contexte, nouveauté, style, compréhension, temps, ...
- Pertinence par document

Goffman, 1969: ‘...the relevance of the information from one document depends upon what is already known about the subject, and in turn affects the relevance of other documents subsequently examined.’

Pertinence - relevance

Type of relevance(survey) (Saracevic 2007)

- Plusieurs pertinences
 - **Thématique** (topical): relation entre le sujet exprimé dans la requête et le sujet couvert dans le document.
 - **Contextuelle** (Situation) : relation entre la tâche, le problème posé par l'utilisateur, la situation de l'utilisateur et l'information retrouvée.
 - **Cognitive** : relation entre l'état de la connaissance de l'utilisateur et l'information sélectionnée



Question

- Processus subjectif (humain), dépend de plusieurs facteurs
→ difficile à automatiser

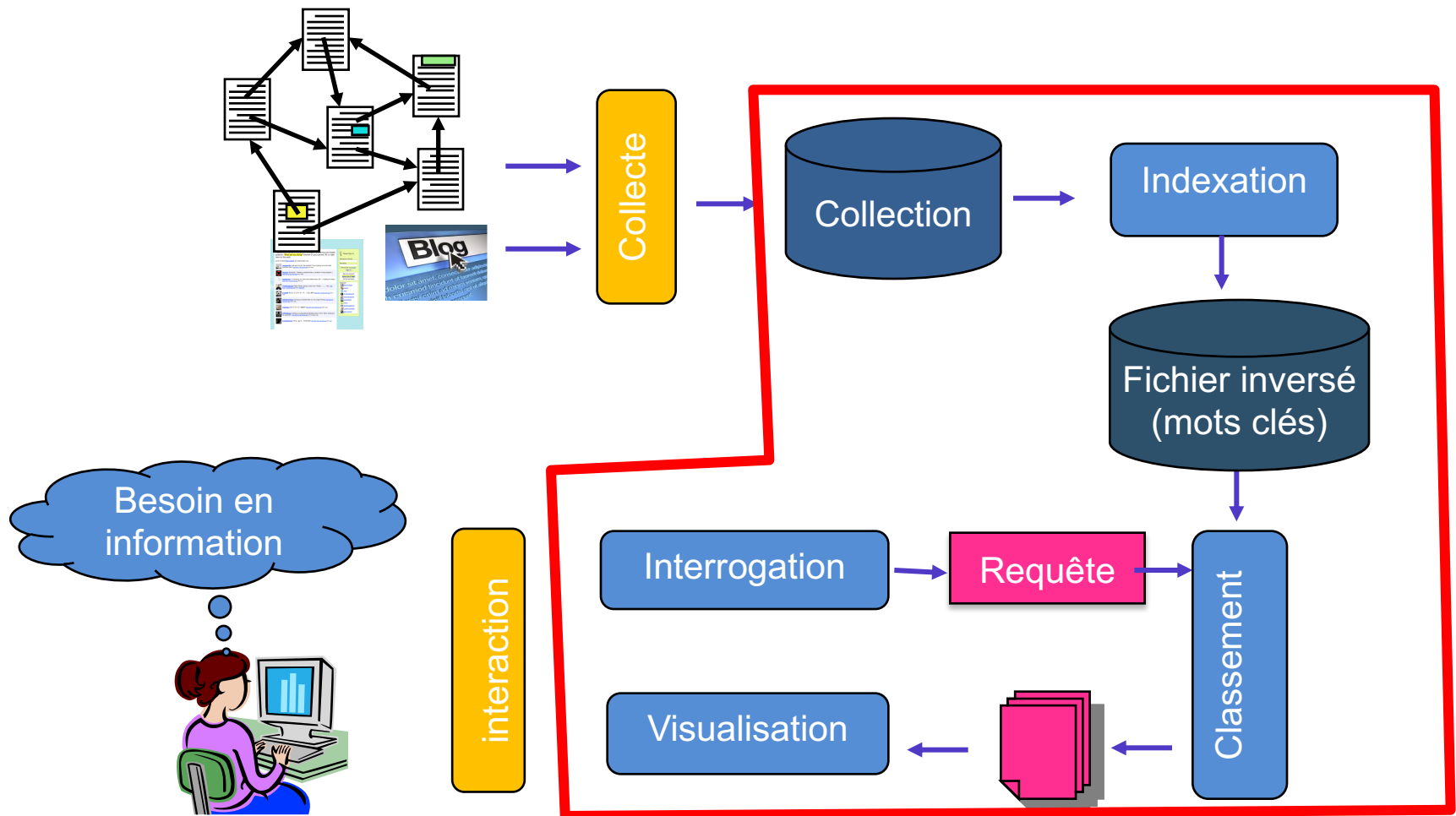
Hypothèses de base

- Besoin = requête
 - Besoin confondu avec la requête utilisateur (une liste de mots clés)
- Document et requête
 - Représentés par des termes (mots simples, groupes de mots, ...) → Sac de mots
- Pertinence
 - Traduite par la similarité de vocabulaire (mots) entre la requête et le document → thématique

Démarche RI

- Interpréter le texte au lieu de le comprendre
- Représenter le contenu comme une liste de mots simples (sac de mots « Bag of words »)
- Exploiter des indicateurs statistiques (comptage de mots) du texte pour évaluer sa pertinence

Architecture d'un SRI

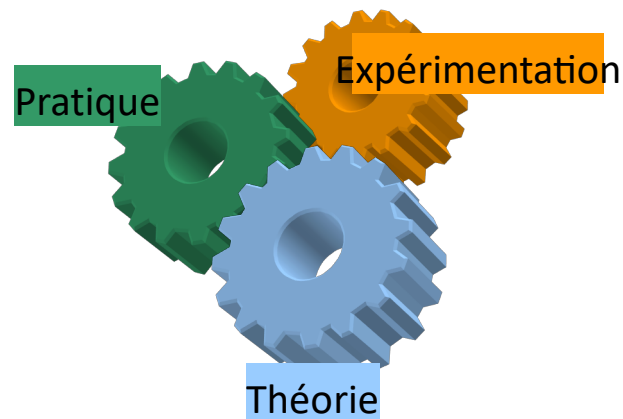


Recap.

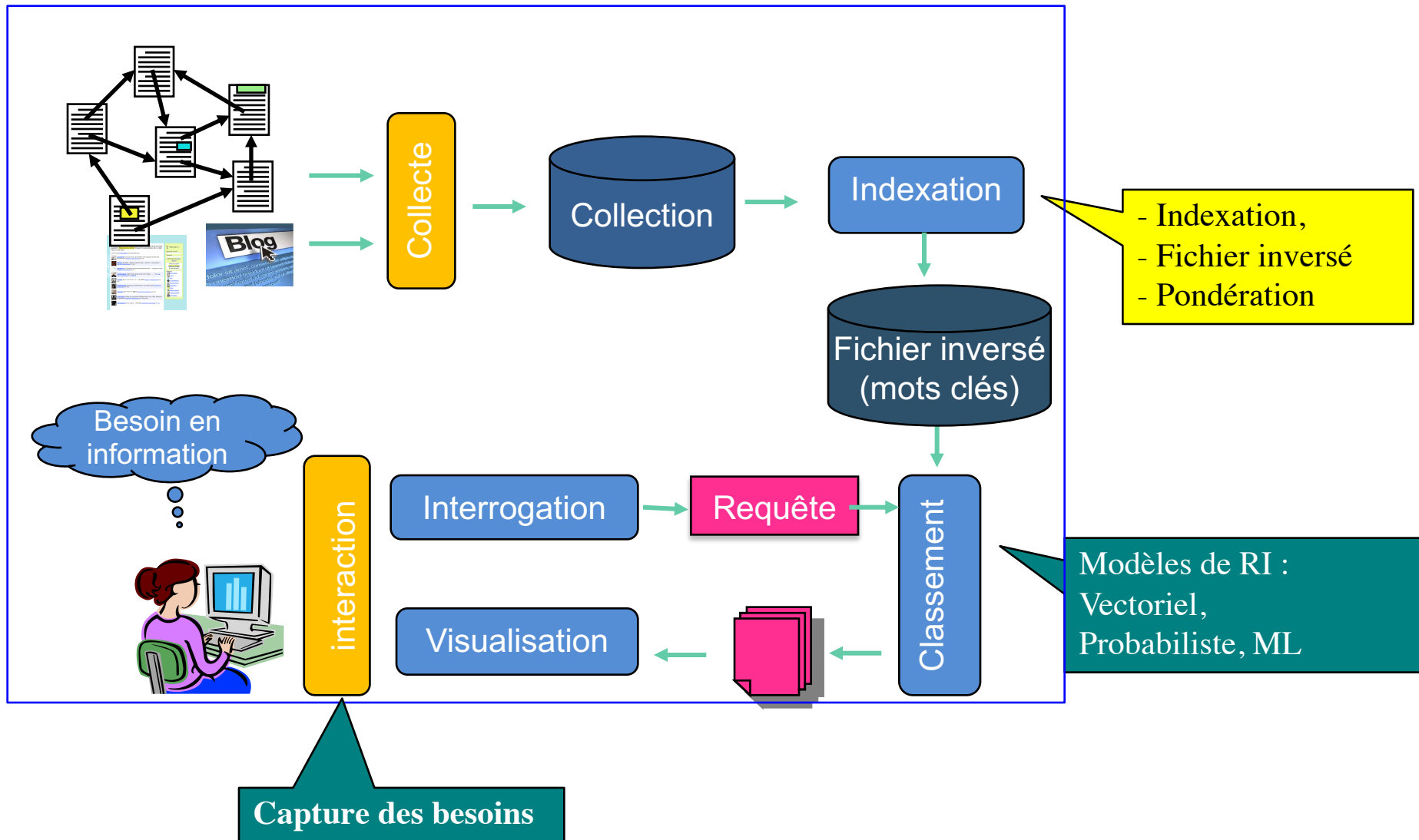
- Représentation (indexation) du document
 - Comment construire une représentation à partir du document ?
 - Quelle organisation physique pour ces index?
- Représentation des besoins
 - Comment capturer le besoin de l'utilisateur ?
 - Comment exprimer le besoin (langage de requêtes) ?
- Comparaison/ranking document-requête (des représentations)
 - Comment mesurer (décider) la pertinence d'un document ?
- Évaluation des performances
 - Comment comparer les SRI ?
 - Quelle démarche (empirique/ analytique) ?
 - Quelles métriques ?

Recap.

- Proposer des solutions
 - modèles, techniques, approches, outils pour répondre à ces problèmes
- ...avec 2 contraintes
 - Quels supports théoriques ?
 - Souvent basés sur des théories mathématiques : Probabilités, statistiques, ensembles, algèbre, logique floue, analyse de données, ...
 - Comment valider ?



Ouvrir la “boîte d’un système de RI”



Définition

- Processus permettant de construire un ensemble d’éléments « clés » permettant de caractériser le contenu d’un document / retrouver ce document en réponse à une requête
- Approches
 - Guidée par un vocabulaire contrôlé vs. Libre
 - Statistique (distribution des mots) et/ou TALN (compréhension du texte)
 - Approche courante est plutôt statistique avec des hypothèses simples
 - Redondance d’un mot marque son importance
 - Cooccurrence des mots marque le sujet d’ un document
- Résultat d’une indexation textuelle
 - Chaque document est représenté par une liste d’index (de mots clés)
 - L’index permet d’accéder (sélectionner) aux documents

Approche générale

- Décomposer le texte (Parsing)
- Segmenter les séquences de caractères en mots (Tokenizing)
- Normaliser
 - Textuelle: ponctuation, dates, case
 - Linguistique : Racinisation (stemming)/lemmatisation
- Supprimer les mots communs (stop word removal)
 - En fonction d'une "short list" "the", "and", "or", ou mots fréquents
- Regrouper les mots

<Title>: Information retrieval
(Corps du texte) : Information retrieval (IR)
is the science of searching of documents.
N.I.S.T launched TREC

Information retrieval Information retrieval
IR is the science of searching of
documents N.I.S.T launched TREC

information retrieval information retrieval
IR is the science, of **search**, document **NIST**
launched TREC

information, retrieval, information,
retrieval, IR, science, **search**, document
NIST **launch** TREC

information 2, retrieval 2, IR 1, science 1,
search 1, document 1, NIST 1, launch 1,
TREC 1

Un sac de
mots (BOW)

Ségmentation (tokenization)

—Pas d’espaces en chinois et en japonais

- Ne garantit pas l’extraction d’un terme de manière unique

Chinese tokenization

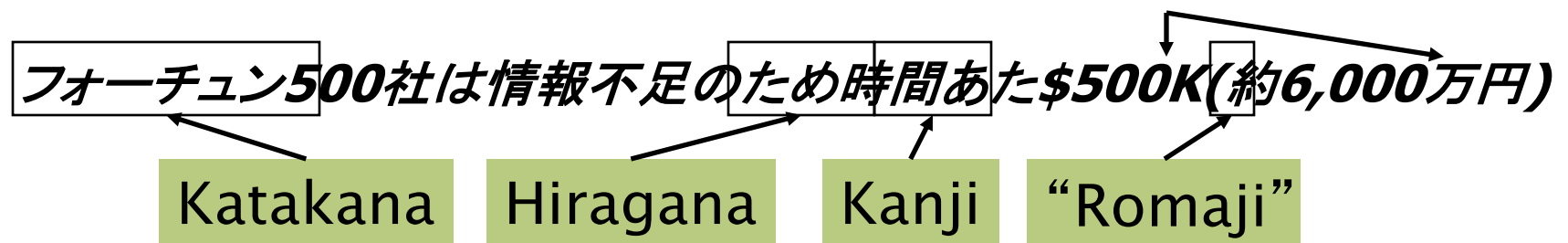
1. Original text

旱灾在中国造成的影响

(the impact of droughts in China)

Ségmentation (tokenization)

- Pas d’espaces en chinois et en japonais
 - Ne garantit pas l’extraction d’un terme de manière unique
- Japonais encore plus compliqué avec différents alphabets



L’ utilisateur peut exprimer sa requête entièrement en Hiragana

Fichier Inversé (Inverted File)

d1:
So let it be
with
Caesar. The
noble
Brutus hath
told you
Caesar was
ambitious

d2:
I did enact
Julius
Caesar I
was killed
i' the
Capitol;
Brutus
killed me.

Traitement
=
Indexation

Term	N docs	Tot Freq	Ptr
ambitious	1	1	1
be	1	1	2
brutus	2	2	3
capitol	1	1	5
caesar	2	3	6
did	1	1	
enact	1	1	
hath	1	1	
I	1	2	
i'	1	1	
it	1	1	
julius	1	1	
killed	1	2	
let	1	1	
me	1	1	
noble	1	1	
so	1	1	
the	2	2	
told	1	1	
you	1	1	
was	2	2	
with	1	1	

Doc #	Freq
2	1
2	1
1	1
2	1
1	1
1	1
2	2
1	1
1	1
2	1
1	2
1	1
2	1
1	1
1	2
2	1
1	1
2	1
2	1
1	1
2	1
2	1
1	1
2	1
2	1
1	1
2	1
2	1

d1:
So let it be with
Caesar. The noble
Brutus hath told you
Caesar was
ambitious

d2:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

d3:
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.
I did enact Julius
Caesar I was killed
i' the Capitol;
Brutus killed me.

Exemple issu de C. Manning, P. Raghavan, H. Schütze. Introduction to Information Retrieval. Cambridge University Press, 2008

Fichier Inversé (Inverted File)

Dictionnaire

Mot	Nb Doc	FrqTotal	Ptr
Ambitious	2	6	1
Brutus	2	4	3
capitol	5	15	6

Posting simple

doc	Freq
doc1	3
doc2	2
doc1	1
doc3	7



- Liste triée
- B-Arbre
- Table de hashage (hash-code)
- ...

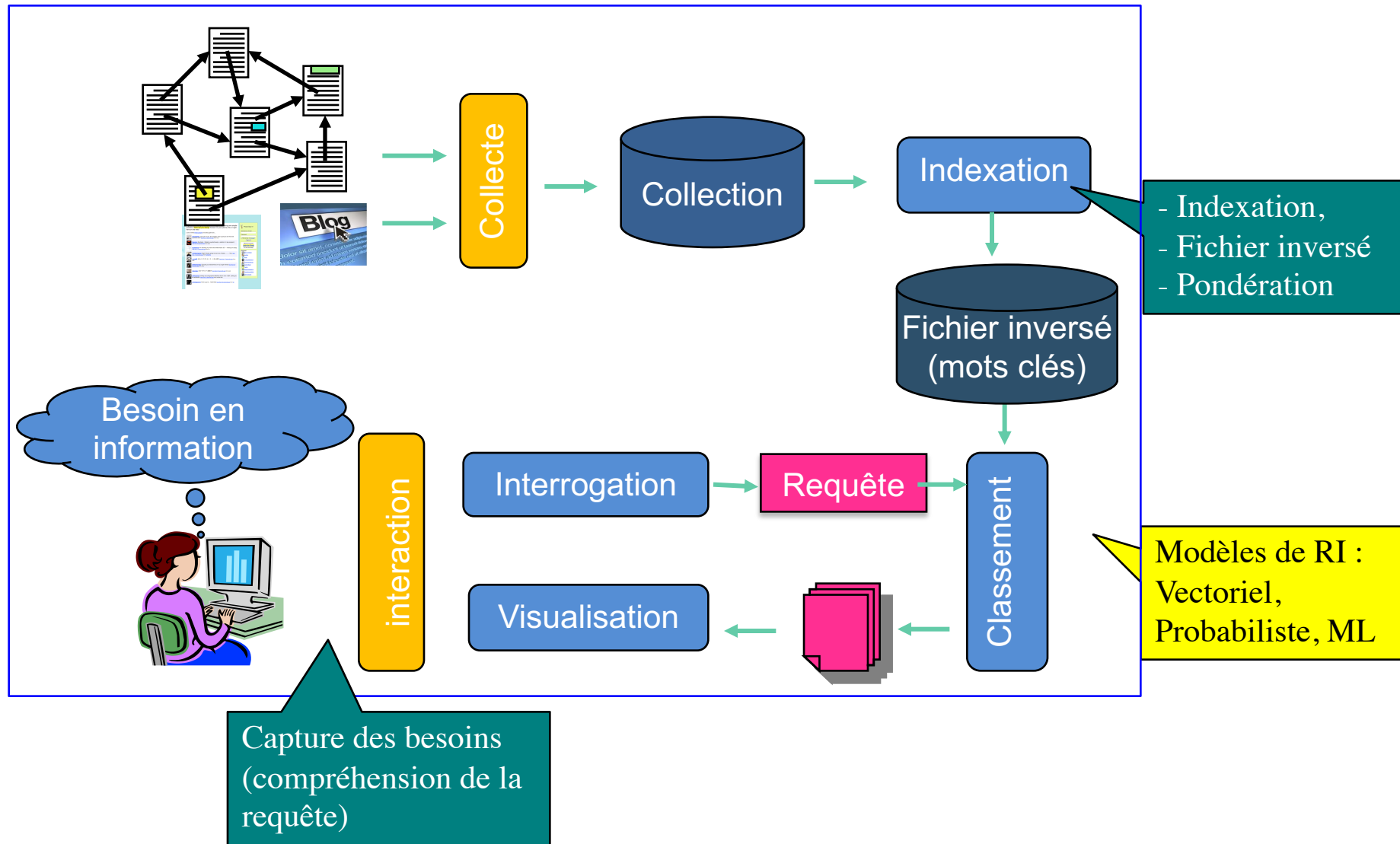
Posting riche

doc	Freq	position	balise
doc1	3	1, 4, 3	1, 5
doc2	2	1	
doc3	2	3	

Position du terme
dans le document
(important pour la
recherche
d'expressions)

Balises (title,
body, anchor,
..)

Ouvrir la boîte d'un SRI



Modèles de RI

- Facteurs utilisés par la majorité des modèles
 - Fréquence du terme dans le document (**tf**), sa fréquence dans la collection (**idf**), sa position dans le texte(*p*), taille du document (**dl**) ...

$$Score(D) = fonction(tf, idf, dl)$$

- Plusieurs modèles théoriques pour formaliser cette fonction
- Elle peut être apprise (apprentissage automatique, approche utilisée par la majorité des moteurs de recherche)

- PIV (vector space model)**

$$\sum_{w \in q \cap d} \frac{1 + \ln(1 + \ln(c(w, d)))}{(1 - s) + s \frac{|d|}{avdl}} \cdot c(w, q) \cdot \ln \frac{N + 1}{df(w)}$$

TF

- DIR (language modeling approach)**

$$\sum_{w \in q \cap d} c(w, q) \times \ln\left(1 + \frac{c(w, d)}{\mu \cdot p(w|C)}\right) + |q| \cdot \ln \frac{\mu}{\mu + |d|}$$

IDF

Length Norm.

- BM25 (classic probabilistic model)**

$$\sum_{w \in q \cap d} \ln \frac{N - df(w) + 0.5}{df(w) + 0.5} \cdot \frac{(k_1 + 1) \times c(w, d)}{k_1((1 - b) + b \frac{|d|}{avdl}) + c(w, d)} \cdot \frac{(k_3 + 1) \times c(w, q)}{k_3 + c(w, q)}$$

- PL2 (divergence from randomness)**

$$\sum_{w \in q \cap d} c(w, q) \cdot \frac{tfn_w^d \cdot \log_2(tfn_w^d) (\lambda_w) + \log_2 e \cdot \left(\frac{1}{\lambda_w} + tfn_w^d\right) + 0.5 \cdot \log_2(2\pi) (tfn_w^d)}{tfn_w^d + 1}$$

$tfn_w^d = c(w, d) \cdot \log_2(1 + c \cdot \frac{avdl}{|d|}), \lambda_w = \frac{N}{c(w, C)}$

Les statistiques au coeur des modèles de RI

Luhn's idea (1958): automatic indexing based on statistical analysis of text



Hans Peter Luhn
(IBM)

“It is here proposed that the frequency of word occurrence in an article furnishes a useful measurement of word significance. It is further proposed that the relative position within a sentence of words having given values of significance furnish a useful measurement for determining the significance of sentences. The significance factor of a sentence will therefore be based on a combination of these two measurements.” (Luhn 58)

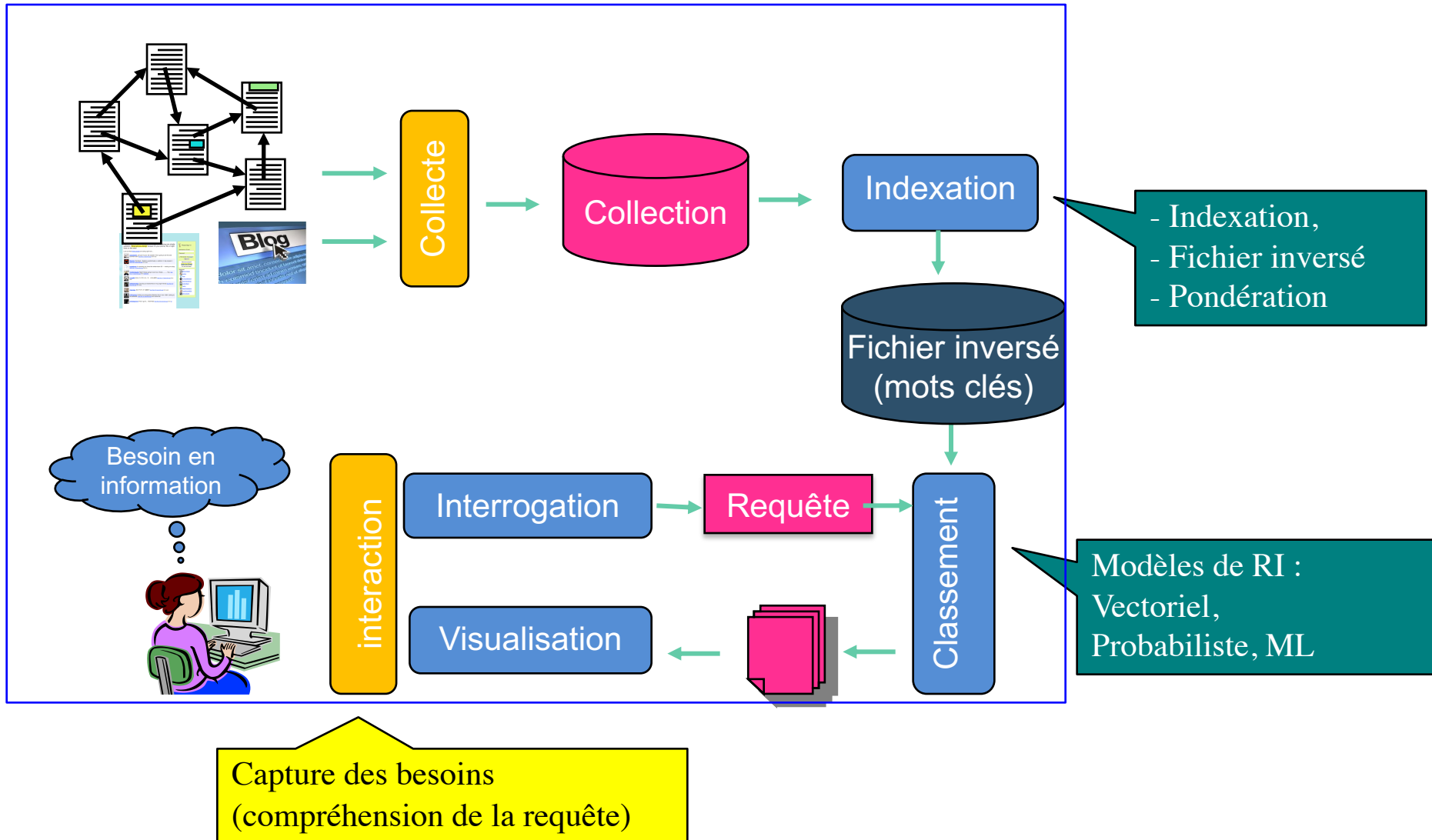
LUHN, H.P., 'A statistical approach to mechanised encoding and searching of library information', *IBM Journal of Research and Development*, 1, 309-317 (1957).

LUHN, H.P., 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2, 159-165 (1958).

Plusieurs cadres théoriques

- Théorie des ensembles :
 - Boolean model (± 1950)
- Algèbre
 - Vector space model (± 1970)
 - Spreading activation model (± 1989)
 - LSI (Latent semantic Indexing)(± 1994)
- Probabilité
 - Probabilistic model (± 1976)
 - Inference network model (± 1992)
 - Language model (± 1998)
 - DFR (Divergence from Randomness model) (± 2002)
- Apprentissage automatique (Machine learning)
 - Apprendre à ordonnancer (Learning to rank) (2000) (Cours E. Gaussier)
 - RI neuronale : Neural IR, deep IR (Thème de EARIA)

Ouvrir la boîte d'un SRI



Capture des besoins-compréhension de la requête

- Requête idéale difficile à construire (l'utilisateur ne sait pas décrire ce qu'il recherche)
- Mieux capturer les besoins (l'intention)
 - Suggestion de requête, correction, ...
 - Identification de l'intention de la requête
 - Aller au delà d'une requête unique
 - Profilage de l'utilisateur

Intention de la requête

- Plusieurs types d'intention
 - But de recherche (search goal)
 - Navigational, informational, transactional
 - Tâche (Search task)
 - « purchase computer », « job-finding query »
 - Intention verticale (Vertical-oriented intents)
 - Image, video, apps, . . .
 - Sensible au temps (Time-sensitivity)
 - News-sensitive queries
 - Sensible à la localisation (Location-sensitivity)
 - “Restaurant”

Au delà de la requête unique

La RI est un processus interactif et itératif (A/R entre l’utilisateur et le système)

- Prise en compte des anciennes requêtes et de l’historique des recherches

 - ➔ **Implicit feedback** ([Shen et al. 05]), **personalized search** ([Teevan et al. 10])

- Modéliser l’utilisateur

 - ➔ **intent modeling** ([Shen et al. 06]), **task inference** ([Wang et al. 13])

- Prise en compte de la session de recherche

 - ➔ **Active feedback** ([Shen & Zhai 05]), **exploration-exploitation tradeoff** ([Agarwal et al. 09], [Karimzadehgan & Zhai 13])

 - ➔ **Session search** (e.g., [Luo et al. 14])

Keynote SIGIR’2015 ChengXiang Zhai (Towards a Game-Theoretic Framework for Information Retrieval)

Evaluation :
Quelle est la meilleure approche?

- Identifier les critères (Cleverdon 66)
 - Facilité d'utilisation du système
 - Présentation des résultats
 - Efficience (Efficiency) (Coût accès/stockage)
 - Efficacité (Effectiveness) : capacité d'un système à sélectionner des documents pertinents.

Rappel : capacité d'un système à sélectionner tous les documents pertinents de la collection ($R = \text{Nb pert sel} / \text{total pert}$)

Précision : capacité d'un système à sélectionner que des documents pertinents ($P = \text{Nb pert sel} / \text{total selec.}$)

R-Précision, MAP, P@X, RR (Reciprocal Rank) NDGC, BPREF, E-mesure, Coverage, Novelty,

Evaluation en RI est comparative

Démarche d'évaluation

- Démarche Analytique (formelle) :
 - Difficile pour les SRI, car plusieurs facteurs : pertinence, distribution des termes, etc. sont difficiles à formaliser mathématiquement
- Démarche Expérimentale (lab-based evaluation) (Cranfield Paradigm)
 - « **benchmarking** ».
 - Evaluation effectuée sur des collections de tests
 - Collection de test : un ensemble de documents, un ensemble de requêtes et des pertinences (réponses positives pour chaque requête)
- User studies evaluation
 - RI interactive, comportement de l'utilisateur

Campagnes d'évaluation

- TREC - Text REtrieval Conference
 - Évaluation des approches RI (beaucoup de tâches sont évaluées dans cette campagne)
- CLEF - Cross Language Evaluation Forum
 - Évaluation des approches de croisement de langues (multilinguisme)
- INEX - Initiative for the Evaluation of XML Retrieval
 - Évaluation de la RI sur des documents de type XML
- NTCIR- NII Testbeds and community for information access Research

Retour sur la représentation des informations

- Vector representation
 - $V = (w_1, w_2, \dots, w_{|V|})$, the whole vocabulary a million (or a billion) of words
 - Text (document) is represented
 - Document vector : $(0, \textcolor{red}{1}, 0, 0, \textcolor{red}{1}, 0, 0, \textcolor{red}{1}, \dots, \textcolor{red}{1}, \dots, 0)_{|V|}$
 - Term vector
 - term vector : $(0, 0, 0, 0, 0, 0, 0, 0, \dots, \textcolor{red}{1}, \dots, 0)_{|V|}$
 - Document vector (Weighted terms)
 - $(0, \textcolor{red}{2}, 0, 0, \textcolor{red}{3}, 0, 0, \dots, \textcolor{red}{1}, \dots, \textcolor{red}{5}, \dots, 0)_{|V|}$
 - $(0, \textcolor{red}{0.25}, 0, 0, \textcolor{red}{0.29}, 0, 0, \dots, \textcolor{red}{0.02}, \dots, \textcolor{red}{0.6}, \dots)_{|V|}$

$$\begin{cases} \text{count}(w, T) \\ 1 + \log(\text{count}(w, T)) \end{cases}$$

Term weighting is very important in many applications

Beyond bag of words

- Limits of bag of words (word representation system)
 - Synonymy, different word with the same meaning “Car” and “auto”,
 - Polysemy, same word but different meaning “salsa (dance)”, “salsa (sauce)”
- Beyond bag of words → Semantic representation (represent the meaning of a word)
 - Knowledge-based techniques (Ontology-based, ...)
 - Unsupervised approaches: Distributional similarity based representations

“You shall know a word by the company it keeps”



Firth, J. R. (1957).

- How to make neighbors represent words?
 - Instead of using the whole $V = (w_1, w_2, \dots, w_{|V|})$ (e.g. $|V|=10^6$)
 - Represent a word in a dense (low dimensional) vector (25 - 1000) built (learnt)
 - **LSA** based on Matrix Factorization (SVD: Singular Value Decomposition) [S. Deerwester et al 1988]
 - Neural Word Embedding: **Word2Vec** (CBOW, Ski-gram) [T. Mikolov et al 2013], Glove, FastText

Lyon : $[0.015, 0.001, -0.020, 0.015, 0.001]_{300}$

Lyon : $[0, 0, 0, \dots, 1, 0, 0, 0, \dots, \dots, \dots]_{|V|}$

Test for linear relationships, examined by Mikolov et al.

$a:b :: c:?$



$$d = \arg \max_x \frac{(w_b - w_a + w_c)^T w_x}{\|w_b - w_a + w_c\|}$$

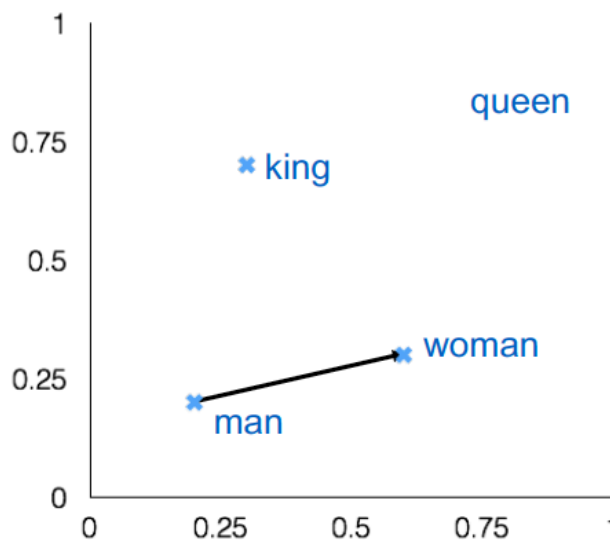
man:woman :: king:?

+ king [0.30 0.70]

- man [0.20 0.20]

+ woman [0.60 0.30]

queen [0.70 0.80]



Test word analogies [Mikolov 2014]

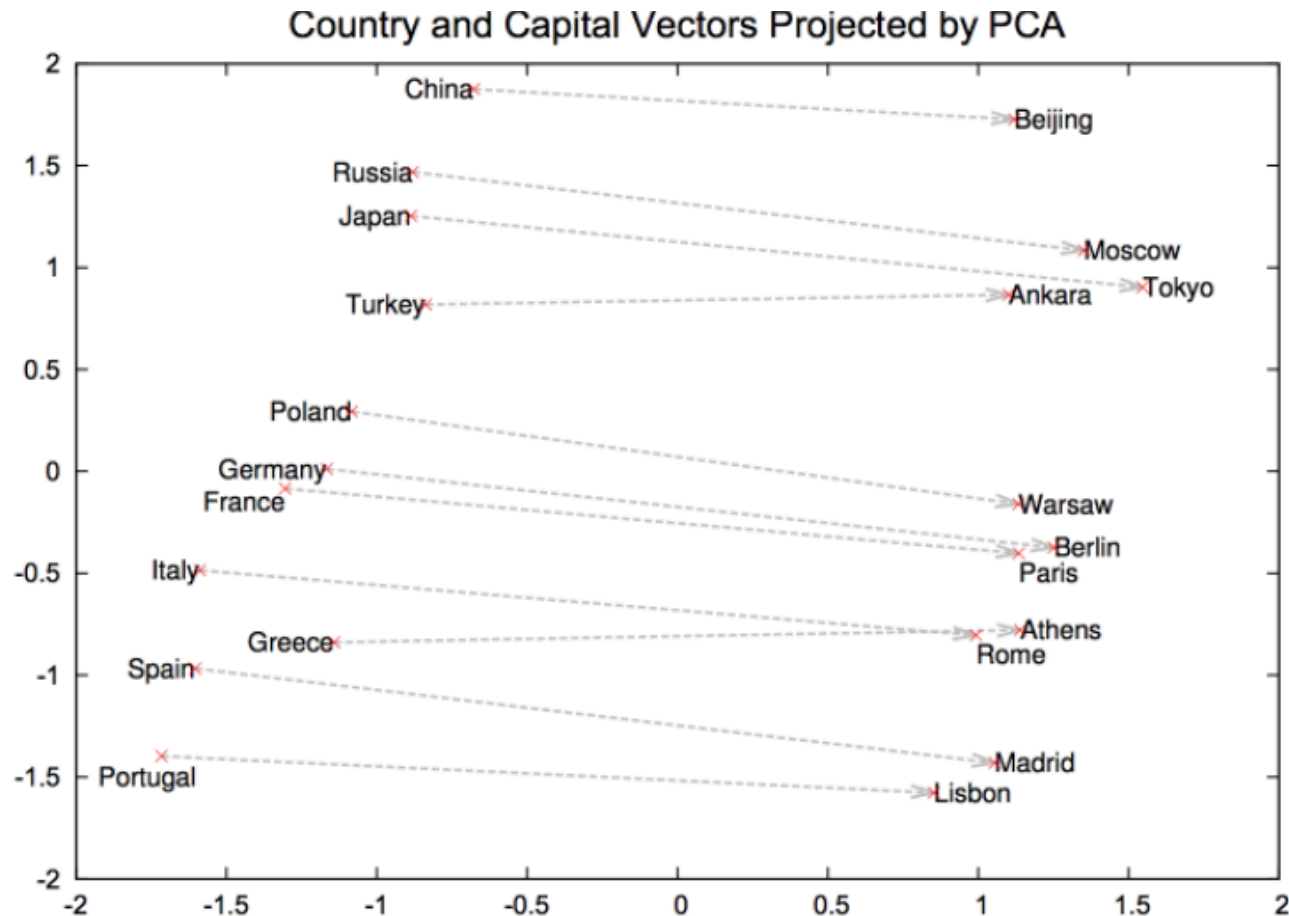
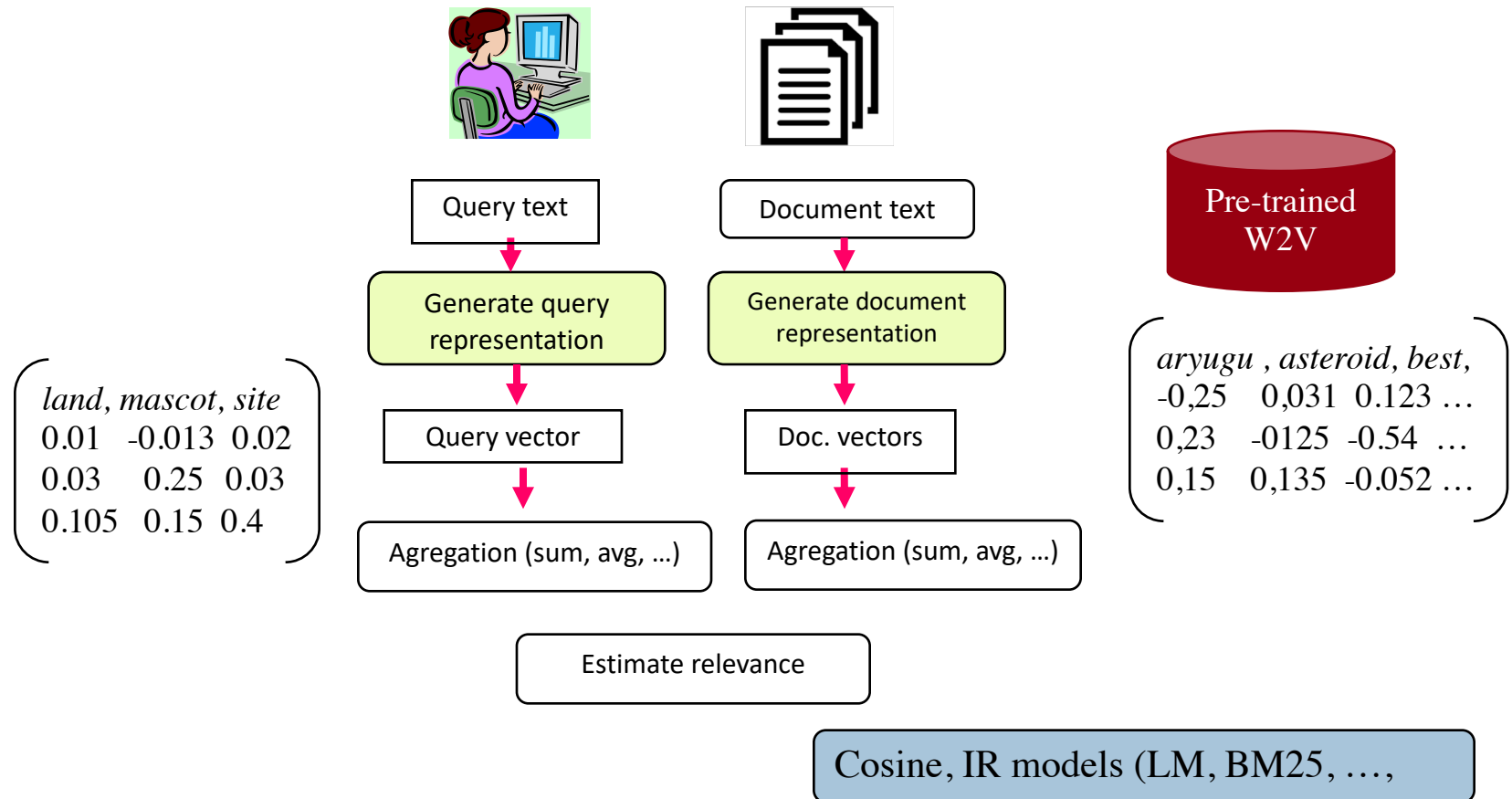


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

- Extend traditional IR models
 - Term weighting, language model smoothing, translation of vocab
- IR models that work in the embedding space
 - Centroid distance, word mover's distance

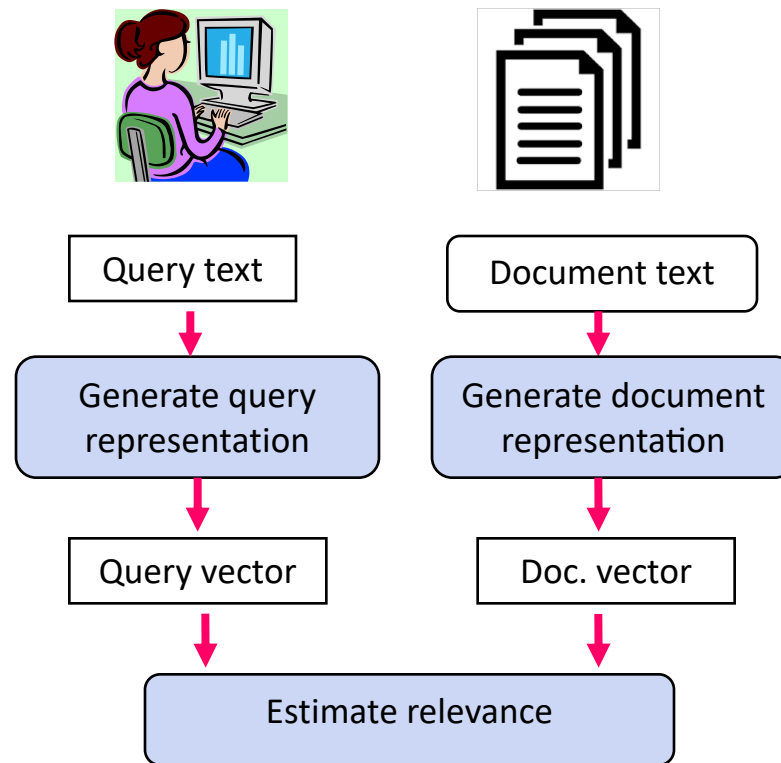


Neural Information retrieval

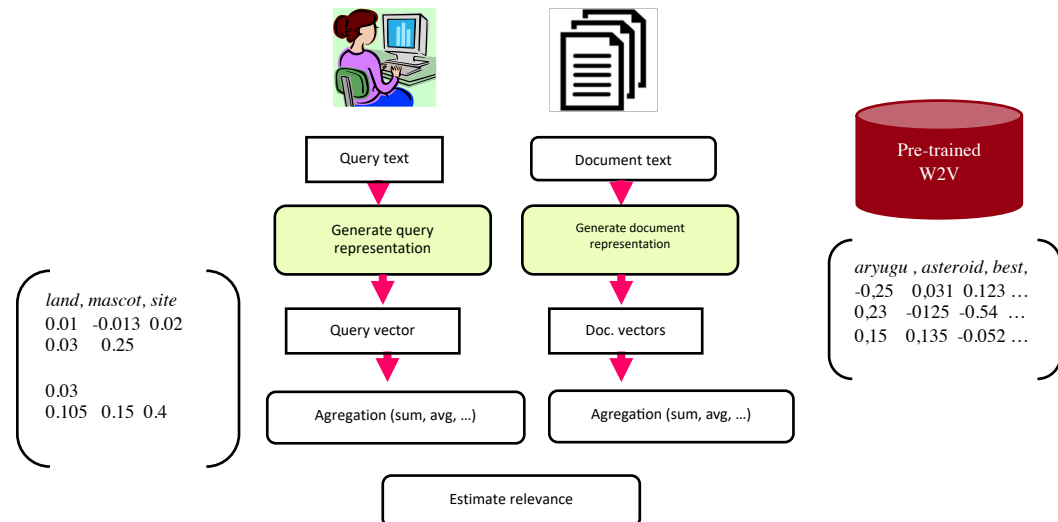
Deep learning et RI

- Information retrieval (IR) in three simple steps:
 - Generate query representation
 - Generate doc. representation
 - Estimate relevance

Neural networks can be useful for one or more of these steps



- IR in three simple steps:
 - Generate query representation
 - Generate doc. representation
 - Estimate relevance



Learning low-dimensional representations of words
→ Semantic representation

- IR in three simple steps:
 - Generate query representation
 - Generate doc. representation
 - Estimate relevance

Posterior probability
computed by softmax

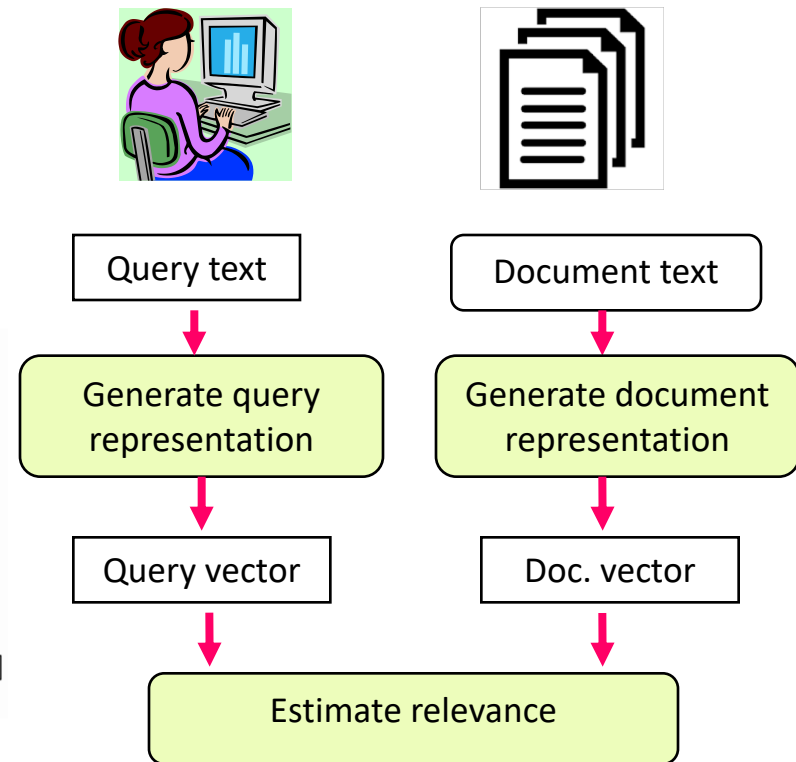
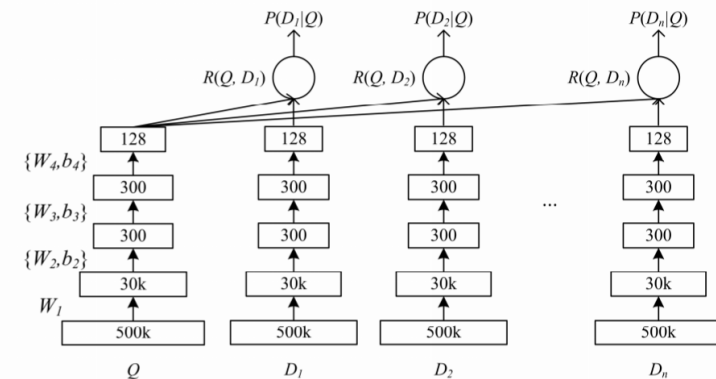
Relevance measured
by cosine similarity

Semantic feature y

Multi-layer non-
linear projection

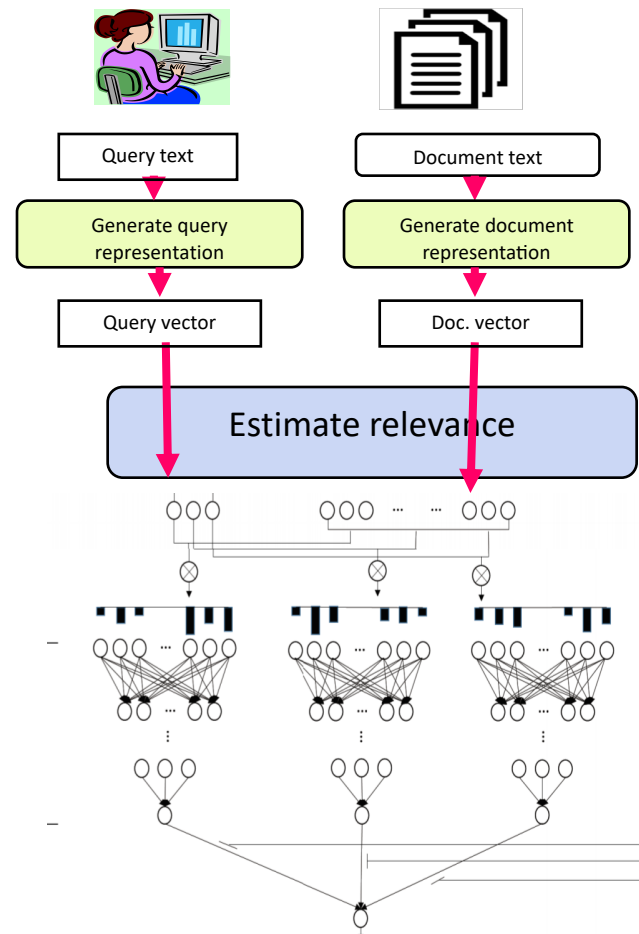
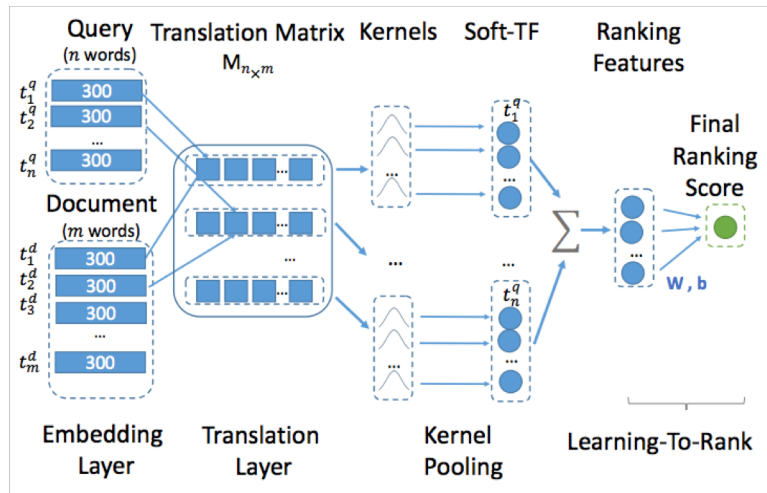
Word Hashing l_1

Term Vector x

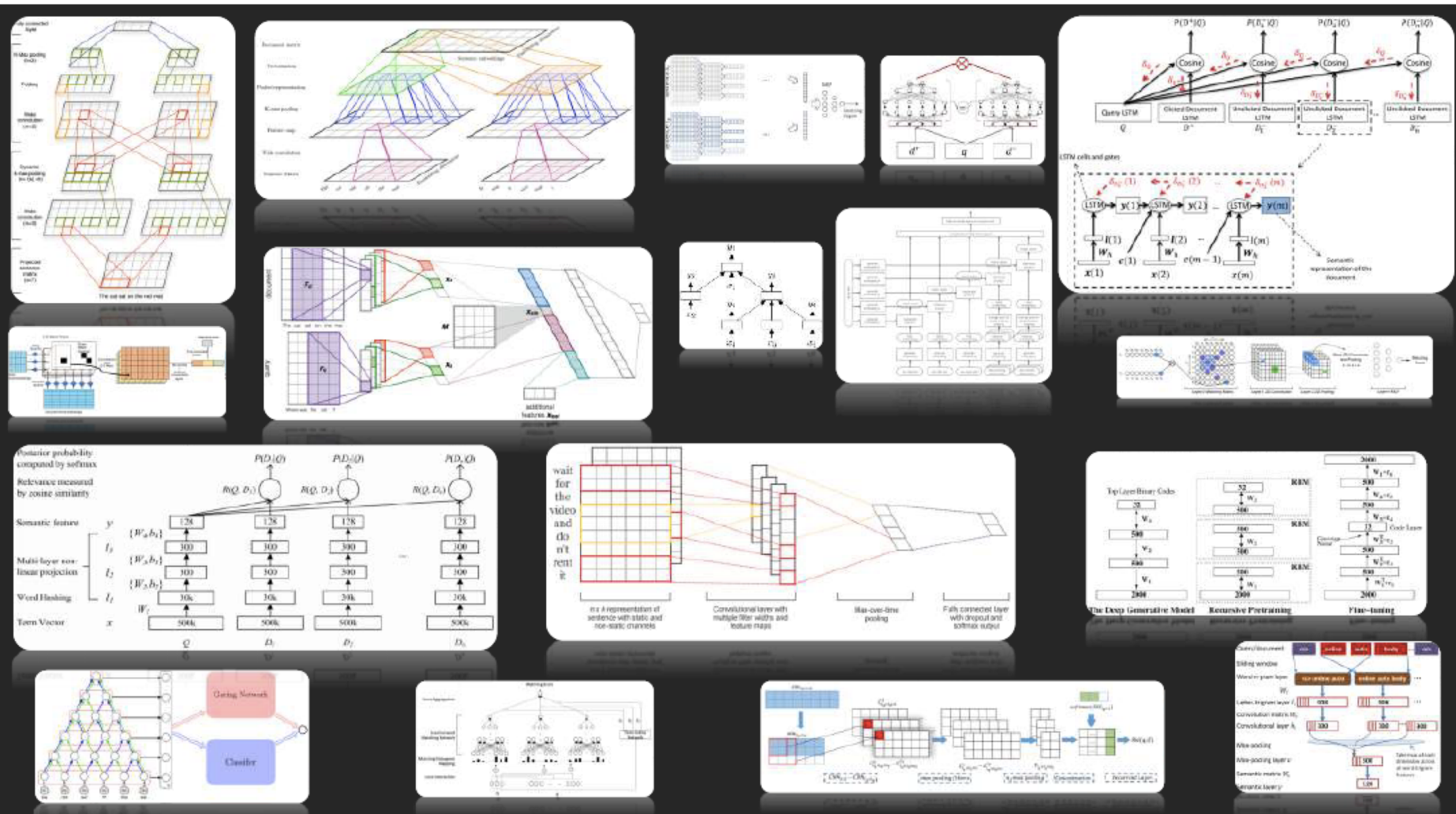


Representation-based models construct a fixed-dimensional vector representation for each text separately and then perform matching within the latent space

- IR in three simple steps:
 - Generate query representation
 - Generate doc. representation
 - Estimate relevance



Interaction-based models compute the interaction between each individual term of both query and documents.



Plan

- Fondements de la Recherche d'information (RI)
 - Introduction : définition, contours de la RI
 - Problématique de la RI
 - Ouvrir la boîte d'un SRI
- Quelques thématiques de recherche en RI
- Conclusion

- **Search and Ranking** : Queries and query analysis, Web search, including link analysis, sponsored search, search advertising, adversarial search and spam, and vertical search, **Retrieval models and ranking**, including diversity and aggregated search, Efficiency and scalability, **Theoretical models** and foundations of information retrieval and access
- **Future Directions**: Novel approaches to IR, **Ethics**, economics, and politics, Applications of search to social good, IR with new devices, including wearable computing, neuroinformatics, sensors, Internet-of-Things, vehicles
- **Domain-Specific Applications**: Social search, Search in structured data including email search and entity search, Multimedia search, Education, Legal, Health, including genomics and bioinformatics, Other domains such as digital libraries, enterprise, news search, app search, archival search
- **Content Analysis, Recommendation and Classification**: Filtering and recommender systems, Document representation, Content analysis and information extraction, including summarization, text representation, readability, sentiment analysis, and opinion mining, Cross- and multilingual search, Clustering, classification, and topic models

- **Artificial Intelligence, Semantics, and Dialog:** Question answering, **Conversational systems and retrieval**, including spoken language interfaces, dialog management systems, and intelligent chat systems, Semantics and knowledge graphs, **Deep learning for IR, embeddings**, and agents
- **Human Factors and Interfaces:** **Mining and modeling search activity**, including user and task models, click models, log analysis, behavioral analysis, and attention modeling, Interactive and personalized search, Collaborative search, social tagging and crowdsourcing, Information privacy and security
- **Evaluation:** User-centered evaluation methods, including measures of user experience and performance, user engagement and search task design, Test collections and evaluation metrics, including the development of new test collections, Eye-tracking and physiological approaches, such as fMRI, Evaluation of novel information access tasks and systems such as multi-turn information access, Statistical methods and reproducibility issues in information retrieval evaluation

- Conversational Information Seeking
- Fairness, Accountability, Confidentiality and Transparency in Information Retrieval
- IR for Supporting Knowledge Goals and Decision-Making
- Evaluation
- Machine Learning in Information Retrieval (Learnable IR)
- Generated Information Objects
- Efficiency Challenges
- Personal Information Access

ARTICLE

Research Frontiers in Information Retrieval
Report from the Third Strategic Workshop
on Information Retrieval in Lorne
(SWIRL 2018)

Editors

J. Shane Culpepper, Fernando Diaz, and Mark D. Smucker

Plan

- Concepts de base de la Recherche d'information (RI)
 - Définition et contours
 - Problématique de la RI
 - Ouvrir la boîte noire d'un SRI
- Quelques thématiques de recherche en RI
- Conclusions et repères historiques

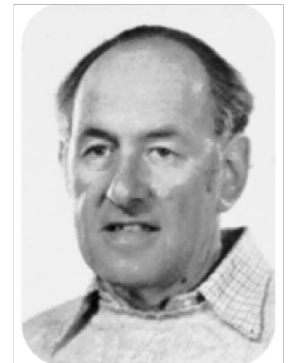
- 1952 Calvin N. Mooers invente le mot « IR »
- 1959 Luhn (RI-statistique)
- 1960 Cranfield I (démarche de validation)
- 1960 Maron and Kuhns (modèle probabiliste)
- 1961 (-1965) Smart (le modèle vectoriel)
- 1968 Premier livre de Salton
- 1975 Livre C van Rijsbergen (accessible sur le web, ver. 1979)
- 1977 Modèle probabiliste (PRP) S. Robertson
- 1978 Première conférence SIGIR
- 1983 Début d' Okapi (modèle probabiliste)

- 1985 RIAO-1 Grenoble
- 1986 Modèle logique («Keith» van Rijsbergen)
- 1990 (tout début du) Learning to rank (développement dans les années 2000)
- 1990 Modèle LSI (Dumais, Deerwester ...),
- 1992 TREC-1
- 1998 Modèle de langue
- 1998 Google
- 2000 CLEF
- 2002 INEX
- 2004 CORIA (Conférence francophone en recherche d'information)
- 2006 EARIA
- 2013 Deep learning (representation learning)

- Plusieurs avancées :
 - Plusieurs Modèles de RI (Luhn, salton, ... BM25, modèle de langue, théorie de l'info, Learning to rank)
 - Évaluation des performances (métriques, benchmark)
 - Traitement de très gros volume de textes (Architecture physique (inverted file))
 - Accès à différents type de textes (journaux, bibliothèques numériques, blogs, microblogs, ...)



Hans-Peter Luhn



Gerard Salton

Constat : les techniques statistiques sont efficaces en RI

- Les modèles statistiques (BOW) s'appuient sur deux hypothèses (
 - Les termes importants sont redondants → LN est redondant
 - La cooccurrence entre les mots indique le sujet du document
 - L'analyse de textes en surface est capable de capturer ces phénomènes
 - Les statistiques « savent » mesurer ces phénomènes
- Deep learning (learning representation, ...)

- Plusieurs programmes d'évaluation :
 - TREC (Text REtrieval Conference)- DARPA,NIST
 - NTCIR (Evaluation campaign on Asian documents)
 - CLEF (Cross Language Evaluation Forum European Language)
 - INEX (INitiative for the Evaluation of XML Retrieval)

 - Collections volumineuses (topics, documents, relevance judgments)
 - Plusieurs tâches
- Comparaisons riches

- Systèmes de RI open source
 - Smart (Cornell)
 - MG (RMIT & Melbourne, Australia; Waikato, New Zealand),
 - Lemur (CMU/Univ. of Massachusetts)
 - Lucene (Nutch)
 - Terrier (Univ Glasgow)
- Permettent d'indexer et d'interroger différents types documents textuels (texte libre, html, xml, pdf, ...). Plusieurs modèles de RI sont programmés (vectoriel, probabiliste (BM25), modèle de langue)

Conclusion

- IR responsable- FACT IR : *Fairness, Accountability, Confidentiality and Transparency*
 - *Fair/juste*
 - Pas Discrimination
 - Pas Dominance
 - Groupes dominants ont des comptes (fréquence d'occurrence) plus importants que d'autres
 - *Accountable /responsable*
 - Justifier ses actions
 - *Confidentiality (Respect vie privée)*
 - Justifier ses actions
 - *Transparence*
 - Expliquer les résultats

transgenders are

transgenders are **mentally ill**

transgenders are **sick**

transgenders are **annoying**

transgenders are **freaks**

Fairness Track at TREC 2019

Conclusion

- Très forte tradition d'expérimentation
- Plusieurs problèmes et verrous à investir
- Domaine difficile : les méthodes intuitives ne sont pas forcément celles qui sont efficaces dans la pratique

Références bibliographiques

- Ouvrages en ligne
 - Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval. 2008 <http://nlp.stanford.edu/IR-book/information-retrieval.html>)
 - Baeza-Yates, R. and Ribeiro-Neto, B. (2011). Modern Information Retrieval - the concepts and technology behind search.
 - Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999
 - Van Rijsbergen (1977) Information Retrieval, Butterworths
 - Frakes and Baeza-Yates, eds. (1992) Information Retrieval: Data Structures & Algorithms, Prentice Hall
 - Witten, Moffat and Bell (1994) Managing Gigabytes plus software, Van Nostrand-Reinhold
 - Baeza-Yates and Ribeiro-Neto, eds. (1999) Modern Information Retrieval Addison-Wesley ([site miroir](#))
 - Recherche d'information : état des lieux et perspectives (M. Boughanem et J. Savoy)
 - Recherche d'information Applications, modèles et algorithmes - Data mining, décisionnel et big data (2013 et 2017)

● Conférences

- ACM SIGIR : Special Interest Group on Information Retrieval
- CIKM: Conference on Information and Knowledge Management
- ECIR: European Conference on Information Retrieval Research
- WSDM: International conference on Web Search and Data Mining
- ICTIR: International Conference on the Theory of Information Retrieval
- CHIIR: Conference on Human Information Interaction and Retrieval
- WWW: TheWebConf
- CORIA : Conférence Francophone en Recherche d'Information et Applications

● Journaux

- JASIST : Journal of the American Society for Information Science and Technology
- IP&M : Information Processing & Management
- IJODL : International Journal on Digital Libraries
- JDOC : Journal of Documentation
- JIR : Journal of Information Retrieval
- ACM-TOIS : Transactions on Information Systems

- Frontiers, challenges, and opportunities for information retrieval:
Report from SWIRL 2018
- <http://sigir.org/resources/>



Merci