Recherche d'information
de la RI traditionnelle à la RI neuronale

Ecole d'été ETAL 2023 - Marseille

Recherche d’information, quésaco ?

Recherche d'information : définition

Definition

"Information retrieval (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)." C. Manning

Application la plus courante : les moteurs de recherche

Google YAHOO! Yandex Aol. Ask.com Baidu 百度 Bing

Mais aussi dans les entreprises, les bibliothèques numériques, nos ordinateurs
Avec des domaines d'application spécialités (médecine, droit, …)

Recherche d'information : définition

Definition

"Information retrieval (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)." C. Manning

Science interdisciplinaire

- Basée sur l'informatique, les mathématiques, les statistiques, la science de l'information, la psychologie cognitive, la linguistique, …
- Utilisée pour réduire la « surcharge informationnelle »

L'information est partout

SCIENCES SORBONNE UNIVERSITÉ

- En 2000, 18 millions de recherches Google par jour. En 2020 ? Plus de 20 milliards !
- Une personne effectue en moyenne 3 à 4 recherches par jour (1 200 recherches par mois).
- Il y a plus de 80 000 requêtes Google par seconde.
- 92% des utilisateurs d'Internet ont utilisé des moteurs de recherche au moins une fois. Sujets ?
  - 83% sur la santé ou les loisirs,
  - 81% pour le bulletin météo,
  - 78% pour des informations sur les nouveaux produits,
  - 76% pour lire les nouvelles,
  - 72% pour le divertissement
  - 71% pour les achats en ligne.
- Temps passé par mois : 1 heure, 47 minutes et 42 secondes.

Sources:
https://seotribunal.com/blog/google- stats- and- facts/
https://www.twinword.com/blog/how- people- spend- their- time- on- the- internet- infographic/

L’information est partout

- L’information est disponible
- MAIS : le problème réside dans la sélection de l’information
- Trouver la bonne information pour le bon utilisateur au bon moment

Le paradigme « One size fits all »? (un même système pour tous

- Longtemps abordé avec une unique vision :
  une requête → une liste de documents

- Désormais :
  - Personnalisation : profiling, contextualisation
  - Interaction : clarification

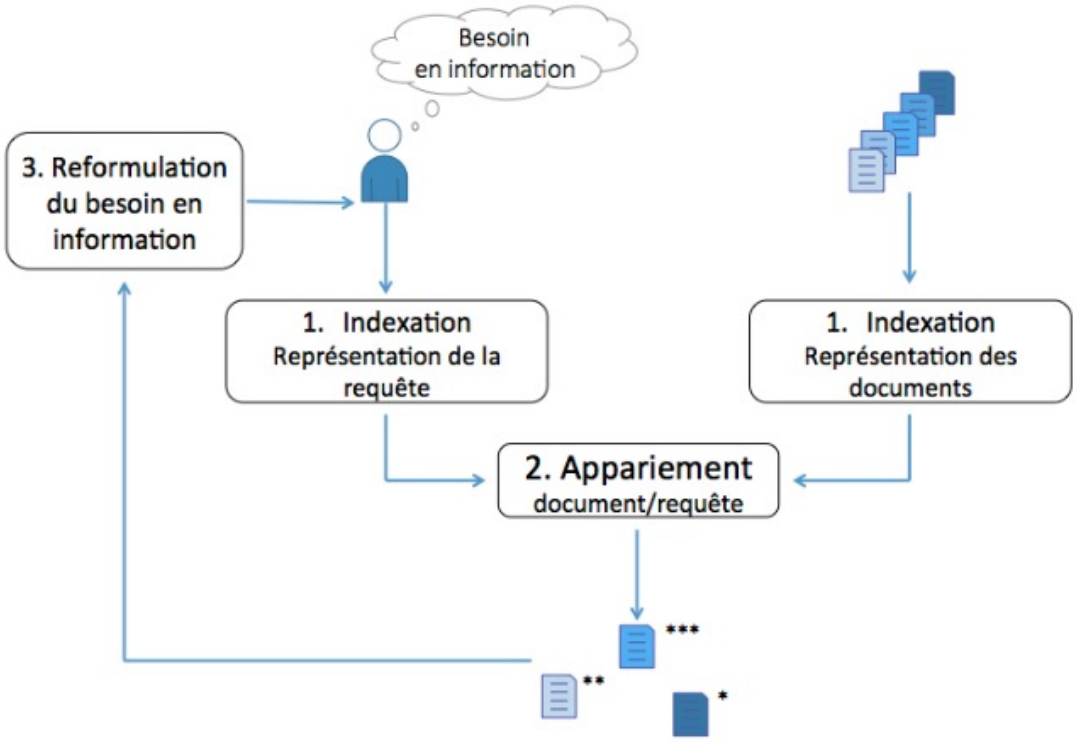
Exemples de tâches en RI

SCIENCES SORBONNE UNIVERSITÉ

- RI ad-hoc : Trouver parmi un ensemble d'articles ceux qui concernent un sujet spécifique : pertinence d'un document ? identification d'experts ?
- Trouver dans un document les passages pertinents, les informations pertinentes concernant un sujet (mots - phrases)
- Faire un résumé du contenu d'un document ou d'un ensemble de documents (éventuellement sur un sujet)
- Rassembler différentes opinions pour un sujet
- Structuration (classification) automatique d'un ensemble de documents / présentation des résultats de recherche
- Suivre dans une collection d'articles l'évolution d'un sujet, Changements de sujets
- Guetter l'arrivée d'informations (appels d'offre, CFP, nouveaux produits, ...)
- Dialoguer avec les clients (e.g. Hot Line, réclamations, ...)
- Filtrage collaboratif : recommander des produits (e.g. Amazone)

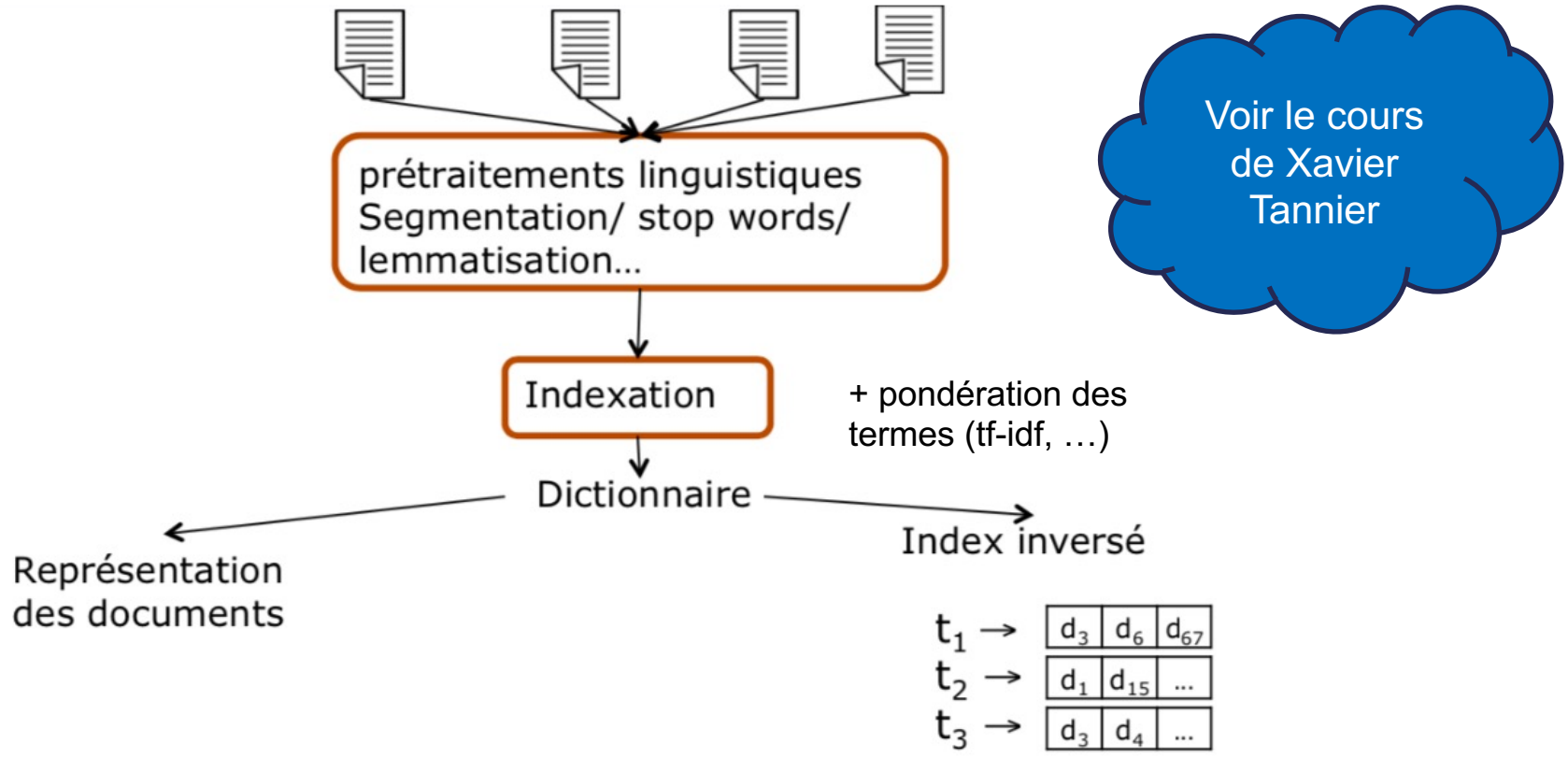
Et comment cela fonctionne ?

Schéma général : Processus en U



- **Indexation :** extraire le contenu d'un document dans un index et représenter la requête → même processus sur les requêtes et les documents pour permettre de faire la correspondance
- **Appariement :** mettre en relation la collection de documents, indexée au préalable, avec la requête, également pré-traitée, afin d'identifier les documents pertinents.
- **Reformulation du besoin en information :** redéfinir le besoin de l'utilisateur au fur et à mesure de la session de recherche.

Indexation

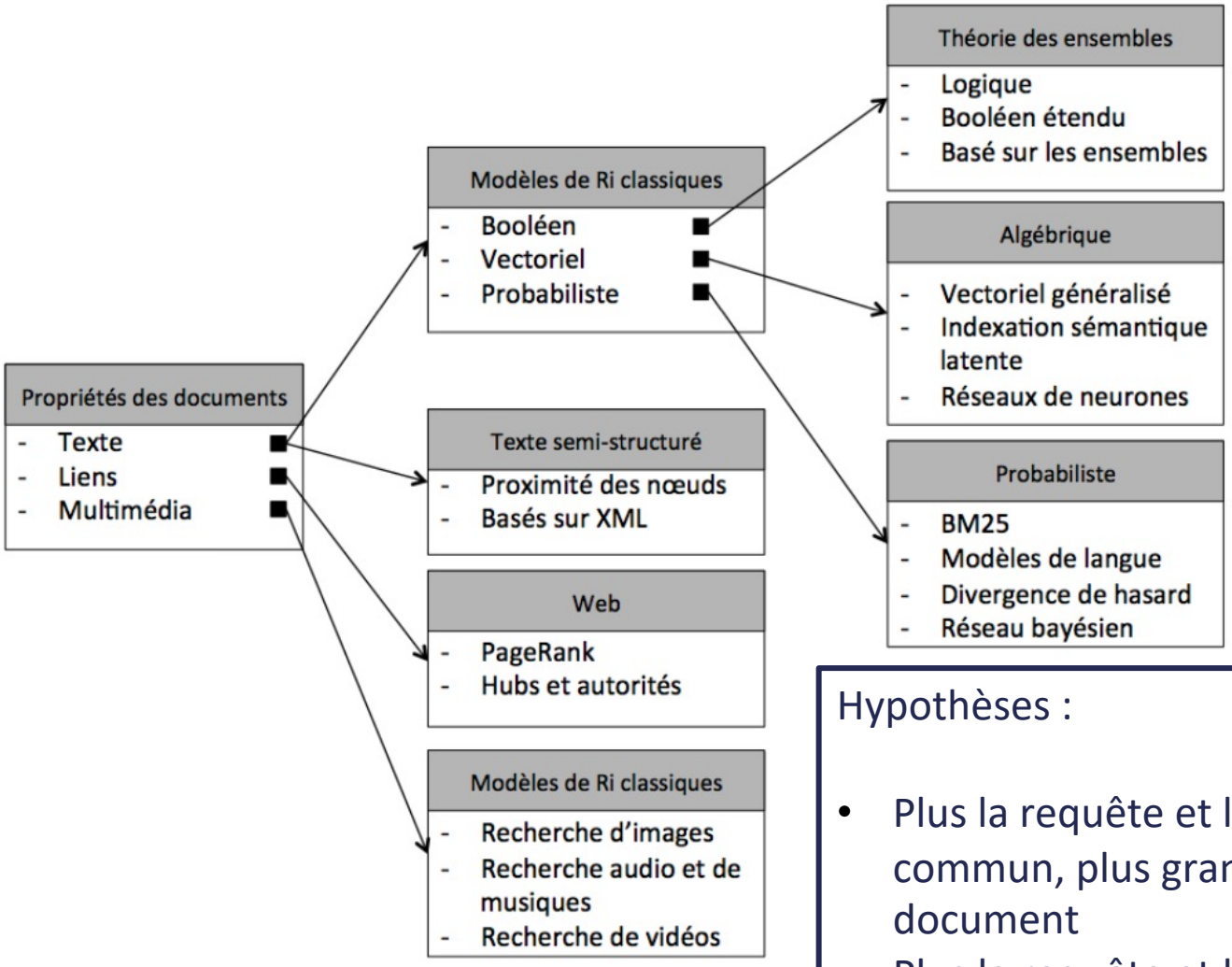


Voir le cours de Xavier Tannier

Deux types d'index

- Index direct : représentation directe des documents
- Index inversé : représentation avec les termes pour point d'entrée.

Appariement

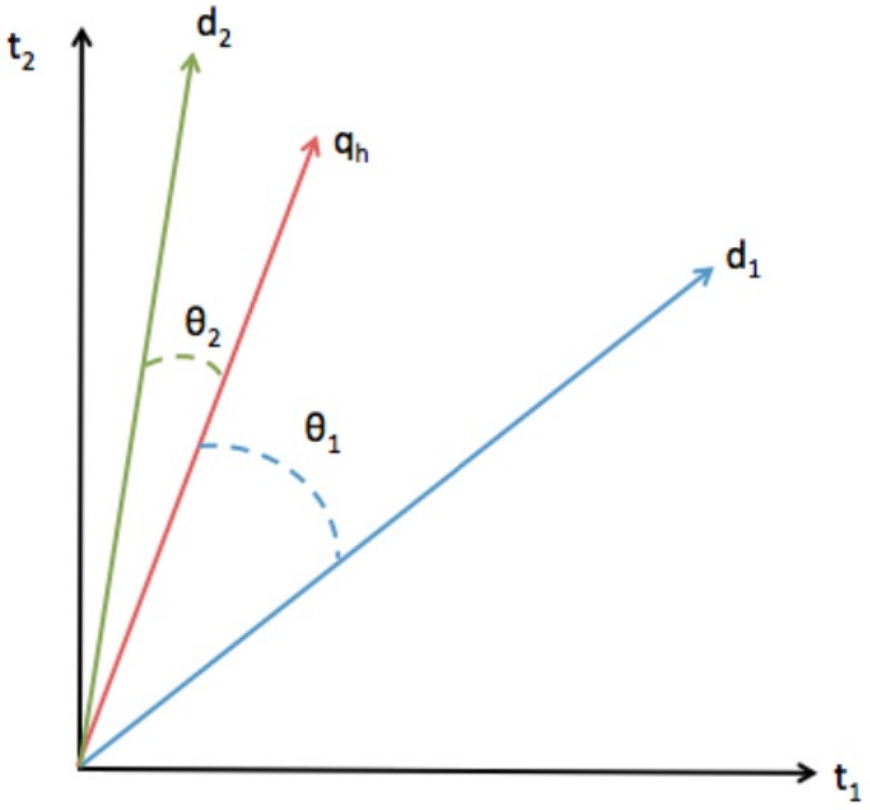


Hypothèses :

- Plus la requête et le document ont de mots en commun, plus grande sera la pertinence du document
- Plus la requête et le document ont une distribution de termes similaire, plus grande sera la pertinence du document

Modèle vectoriel (Salton et al., 1975)

- Espace de caractéristiques $t_i, i = 1...n$, i.e. termes sélectionnés pré-traités
- Représentation des documents - requêtes : vecteur de poids dans l'espace des caractéristiques :
  - document : $d = (x_0, ..., x_{n-1})$
  - requête : $q = (y_0, ..., y_{n-1})$
- Mesures de similarité : cosinus, ...



Modèle Okapi-BM25 (Robertson et al, 1994)

SCIENCES SORBONNE UNIVERSITÉ

- Un des modèles les plus connus et robuste
- Formule générale : BM25 (Robertson et Walker, 1994)

$$s(d,q) = \sum_{i;y_i=1} IDF(y_i).\frac{tf(y_i,d)}{tf(y_i,d) + k_1(1-b+b.\frac{|D]}{avgdl}}$$

avec $|D|$: longueur du document, $avgdl$: longueur moyenne des documents. k_1 et b constantes (e.g., resp 1.2 et 0.75)

- Deux modifications à la pondération tf-idf:
  - Verbosité : on normalise le TF (occurrence du terme) par rapport à la longueur du document pour éviter de privilégier les documents longs.
  - coefficient de saturation de la loi 2-poisson qui modélise la distribution des termes dans les documents (élites, non élites)

Reformulation de requêtes

- Intuition
  - Difficile de formuler les requêtes qui correspondent aux documents de la collection
    * On ne sait pas forcément exprimer ce que l'on cherche
    * On ne sait pas forcément à quoi ressemble le document

- Feedback explicite : utilisateur clique/regarde les documents
  → on peut utiliser ce signal pour reformuler la requête

- Feedback implicite : considérer que le SRI a établi un premier ordre et que les top/flop documents donnent un premier signal de pertinence
  → on peut les utiliser pour reformuler la requête

Principe général

- A partir de la liste ordonnée des r meilleurs documents

$$D_r(q) = d_1, ..., d_r$$

- On peut demander à l'utilisateur de partitionner ces r documents en pertinents (rel)/non pertinents (non-rel). On peut également partitionner à partir des clics des utilisateurs.

$$D_r(q) = \{D_r^{rel}(q) \cup D_r^{non-rel}(q)\}$$

- Principe du relevance feedback : reformuler la requête q pour obtenir une nouvelle requête q' en fonction des documents jugés pertinents ou non pertinents :

$$q' = f(q, D_r^{rel}(q), D_r^{non-rel}(q))$$

Modèle de Rocchio (1971)

SCIENCES SORBONNE UNIVERSITÉ

- Modèle de base de l'expansion/reformulation de requêtes :
  - On considère que la requête et les documents sont modélisés par un vecteur où chaque élément correspond à un terme du vocabulaire qui a un poids, par exemple tf-idf.
  - La nouvelle requête q' correspond globalement à la combinaison des vecteurs documents requêtes : 1) on ajoute la moyenne des vecteurs des documents pertinents et 2) on enlève la moyenne des vecteurs des documents non pertinents.

$$\vec{Q} = (a.\vec{Q_0}) + (b.\frac{1}{|D_{rel}|}\sum_{d+\in D_{rel}} \vec{d+}) - (c.\frac{1}{|D_{non-rel}|}\sum_{d-\in D_{non-rel}} \vec{d-})$$

- Améliorations allant de 20% à 80% par rapport à sans RF
- Différentes variantes :
  - considérer seulement les documents pertinents / que les non-pertinents
  - optimiser a, b, c
  - optimiser le nombre de documents du feedback

Comment on évalue un moteur de recherche ?

Enjeux de l'évaluation en RI

- Quel modèle de RI est le plus efficace ?

- Efficace ? Problème difficile, pas de mesure absolue
  - Critères de qualité d'un système de RI
  - Facilité d'utilisation du système
  - Coût accès/stockage
  - Présentation des résultats
  - Efficacité de la recherche
  - Possibilités de formuler des requêtes riches

Vers un protocole d'évaluation standardisé : Evaluation Cranfield

- Objectif : évaluer la capacité d'un système à retourner des documents pertinents
- De quoi a-t-on besoin ?

Paradigme de Cranfield - première expérimentation "laboratoire" en RI

Évaluation basées sur des collections de test composées de :
- Corpus de documents
- Requêtes
- Jugements de pertinence

- Avantages
  - Peu coûteux
  - Facilite les analyses d'erreurs
  - Répétables
- Inconvénients
  - Jugements de pertinence peuvent être incomplets
  - Quelles hypothèses pour la pertinence ?

Jeux de données

SCIENCES SORBONNE UNIVERSITÉ

→ Campagnes d'évaluation nombreuses (TREC, CLEF, NTCIR, …)

- Ad hoc Test Collections
- Web Test Collections
- Blog Track
- Chemical IR Track
- Clinical Decision Support Track
- Common Core Track
- Confusion Track
- Contextual Suggestion Track
- Interactive Track
- Knowledge Base Acceleration Track
- Legal Track
- Medical Track
- Microblog Track
- Million Query Track
- Novelty Track
- Query Track
- Question Answering Track
- Precision Medicine Track
- Real-time Summarization Track
- Relevance Feedback Track
- Robust Track

→ Jeux de données issus d'articles scientifiques

Le nerf de la guerre : les requêtes…

SCIENCES SORBONNE UNIVERSITÉ

- ➜ Des requêtes réalistes
  - ➜ Souvent issues de logs de recherche
  - ➜ Format : mots-clés ? langage naturel ?
  - ➜ Quelle intention derrière une requête ?

<top>

<num> Number: 501
<title> deduction and induction in English?

<desc> Description:
What is the difference between deduction and induction in the process of reasoning?

<narr> Narrative:
A relevant document will contrast inductive and deductive reasoning.
A document that discusses only one or the other is not relevant.

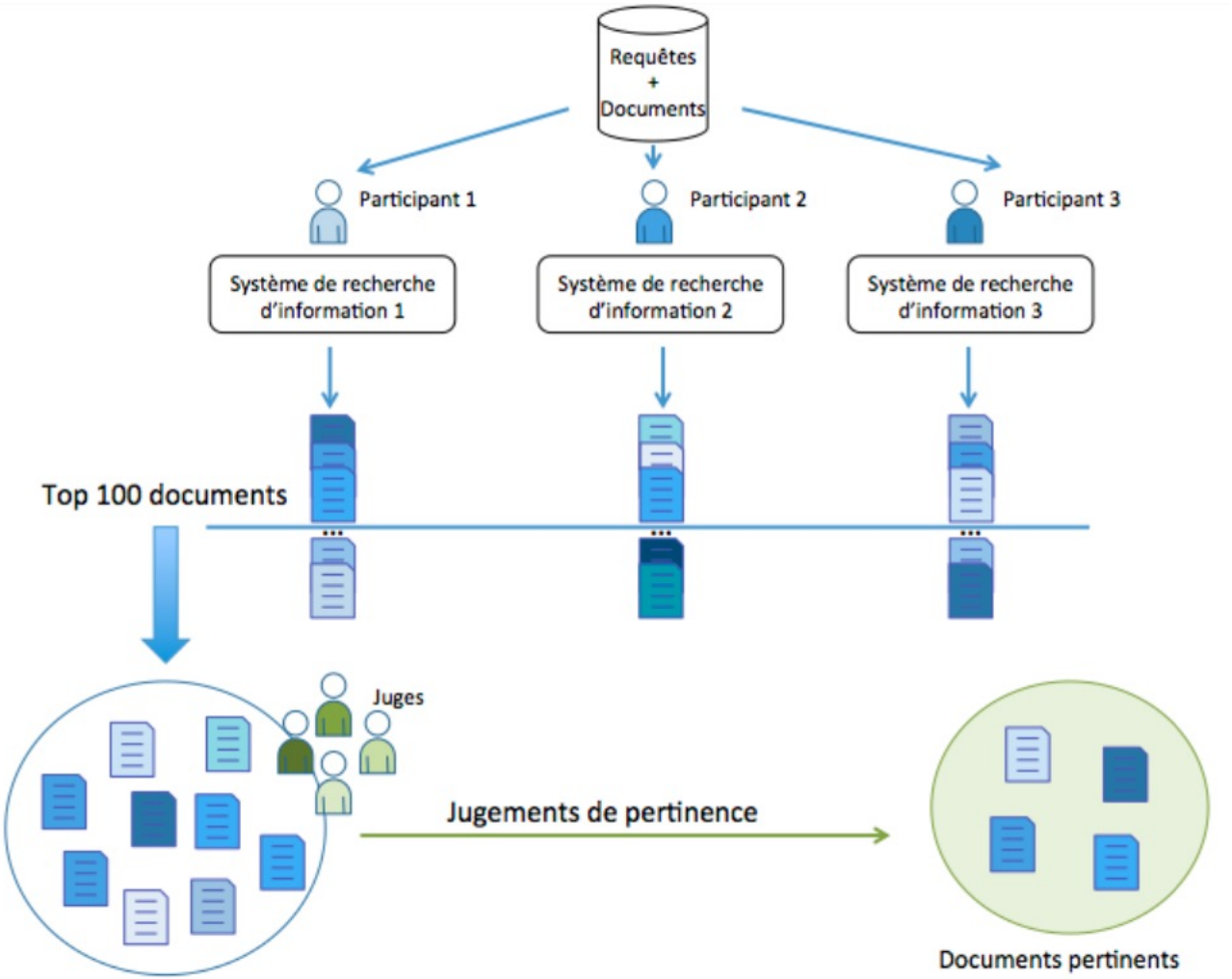
</top>

Le nerf de la guerre : …et les jugements de pertinence

- Pour chaque requête, il faut identifier les documents pertinents
  - Pertinence binaire (0/1) ou graduelle (de 0 à 5)

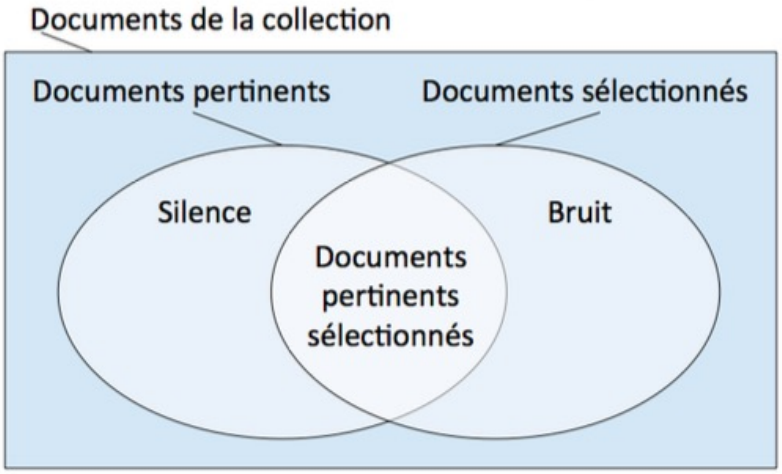
- Juger toute la collection de documents pour chaque requête est trop lourd…

Le nerf de la guerre : …et les jugements de pertinence



Métriques d'évaluation

- Différents types de mesure
  - Précision : nombre de documents pertinents renvoyés / nombre de documents renvoyés
  - Rappel :nombre de documents pertinents renvoyés / nombre de documents pertinents
  - Orientées rang : rang inverse des documents pertinents
  - Ndcg : gain normalisé (pondération de la pertinence en fonction du rang)



Autre protocoles d'évaluation : un pas vers les vrais utilisateurs

- Evaluation basée sur les logs utilisateurs
  - Permet d'appliquer a posteriori des modèles sur des données utilisateurs

- Avantages
  - Besoin en information généré par un utilisateur réel
  - Automatisation de l'évaluation
- Inconvénients
  - "Artificiel"

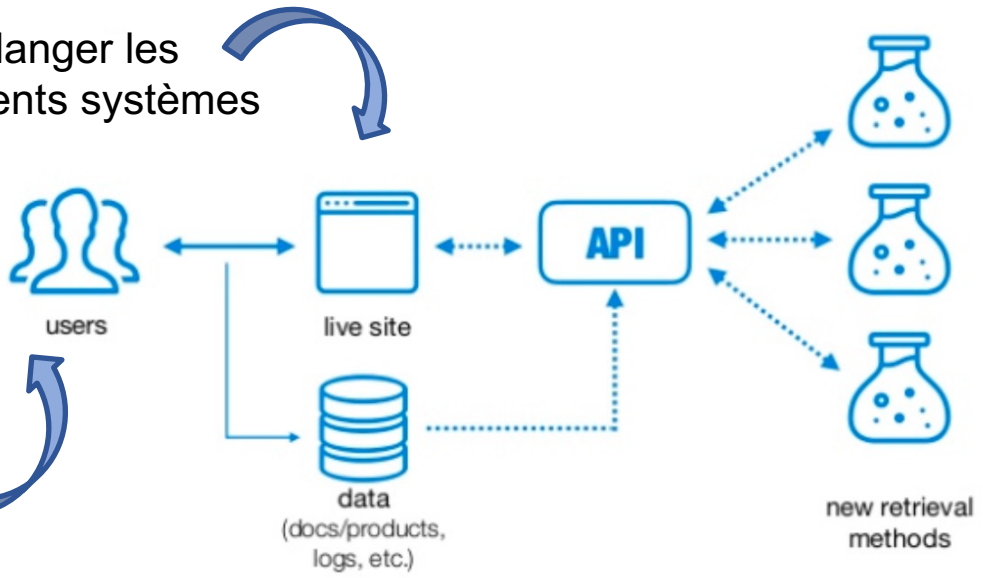
Des évaluation carrément avec les utilisateurs

- Expérimentations utilisateurs :
  - Faire tester le système en direct avec des utilisateurs
  - Dans la vraie vie ou en laboratoire (environnement contrôlé)

- Et plus fun encore : les living labs

Interleaving : mélanger les résultats de différents systèmes

A/B test : mesurer quels systèmes est le plus pertinent à partir des clics utilisateurs



K. Balog, L. Kelly, and A. Schuth. **Head First: Living Labs for Ad-hoc Search Evaluation.** *CIKM'14*

Des évaluation carrément avec les utilisateurs

- User-study
  - Utilisateur interagit en temps réel avec le système

- Avantages
  - Evaluation directe du système
  - Au plus proche de l'utilisateur
- Inconvénients
  - Collecte fastidieuse, coûteuse, ...
  - Difficile d'évaluer toutes les variantes d'un modèle (paramétrage, etc. . .)
  - Evaluation de plusieurs modèles ?

Et si la RI était un problème d'apprentissage automatique ?

Formulation du problème d'apprentissage

SCIENCES SORBONNE UNIVERSITÉ

➜ Calculer un score de pertinence revient à apprendre f(q,d)

→ Plusieurs questions se posent :

→ Comment modéliser la requête et le document ?

Formulation du problème d'apprentissage

SCIENCES SORBONNE UNIVERSITÉ

→ Calculer un score de pertinence revient à apprendre f(q,d)

→ Plusieurs questions se posent :

→ Comment modéliser la requête et le document ?

ID	Feature Description	Category
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in body	Q-D
2	$\sum_{q_i \in q \cap d} c(q_i, d)$ in anchor	Q-D
3	$\sum_{q_i \in q \cap d} c(q_i, d)$ in title	Q-D
4	$\sum_{q_i \in q \cap d} c(q_i, d)$ in URL	Q-D
5	$\sum_{q_i \in q \cap d} c(q_i, d)$ in whole document	Q-D
6	$\sum_{q_i \in q} idf(q_i)$ in body	Q
7	$\sum_{q_i \in q} idf(q_i)$ in anchor	Q
8	$\sum_{q_i \in q} idf(q_i)$ in title	Q
9	$\sum_{q_i \in q} idf(q_i)$ in URL	Q
10	$\sum_{q_i \in q} idf(q_i)$ in whole document	Q
11	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in body	Q-D
12	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in anchor	Q-D
13	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in title	Q-D
14	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in URL	Q-D
15	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in whole document	Q-D
16	$\|d\|$ of body	D
17	$\|d\|$ of anchor	D
18	$\|d\|$ of title	D
19	$\|d\|$ of URL	D
20	$\|d\|$ of whole document	D
21	BM25 of body	Q-D
22	BM25 of anchor	Q-D
23	BM25 of title	Q-D
24	BM25 of URL	Q-D
25	BM25 of whole document	Q-D
26	LMIR.ABS of body	Q-D
27	LMIR.ABS of anchor	Q-D
28	LMIR.ABS of title	Q-D
29	LMIR.ABS of URL	Q-D
30	LMIR.ABS of whole document	Q-D
31	LMIR.DIR of body	Q-D
32	LMIR.DIR of anchor	Q-D
33	LMIR.DIR of title	Q-D
34	LMIR.DIR of URL	Q-D
35	LMIR.DIR of whole document	Q-D
36	LMIR.JM of body	Q-D
37	LMIR.JM of anchor	Q-D
38	LMIR.JM of title	Q-D
39	LMIR.JM of URL	Q-D
40	LMIR.JM of whole document	Q-D
41	Sitemap based term propagation	Q-D
42	Sitemap based score propagation	Q-D
43	Hyperlink based score propagation: weighted in-link	Q-D
44	Hyperlink based score propagation: weighted out-link	Q-D
45	Hyperlink based score propagation: uniform out-link	Q-D
46	Hyperlink based propagation: weighted in-link	Q-D
47	Hyperlink based feature propagation: weighted out-link	Q-D
48	Hyperlink based feature propaga-	Q-D

Liu, 2011. Learning to rank for information retrieval.

Formulation du problème d'apprentissage

SCIENCES SORBONNE UNIVERSITÉ

→ Calculer un score de pertinence revient à apprendre f(q,d)

→ Plusieurs questions se posent :

→ Comment modéliser la requête et le document ?

ID	Feature Description	Category
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in body	Q-D
2	$\sum_{q_i \in q \cap d} c(q_i, d)$ in anchor	Q-D
3	$\sum_{q_i \in q \cap d} c(q_i, d)$ in title	Q-D
4	$\sum_{q_i \in q \cap d} c(q_i, d)$ in URL	Q-D
5	$\sum_{q_i \in q \cap d} c(q_i, d)$ in whole document	Q-D
6	$\sum_{q_i \in q} idf(q_i)$ in body	Q
7	$\sum_{q_i \in q} idf(q_i)$ in anchor	Q
8	$\sum_{q_i \in q} idf(q_i)$ in title	Q
9	$\sum_{q_i \in q} idf(q_i)$ in URL	Q
10	$\sum_{q_i \in q} idf(q_i)$ in whole document	Q
11	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in body	Q-D
12	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in anchor	Q-D
13	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in title	Q-D
14	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in URL	Q-D
15	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in whole document	Q-D
16	$\|d\|$ of body	D
17	$\|d\|$ of anchor	D
18	$\|d\|$ of title	D
19	$\|d\|$ of URL	D
20	$\|d\|$ of whole document	D
21	BM25 of body	Q-D
22	BM25 of anchor	Q-D
23	BM25 of title	Q-D
24	BM25 of URL	Q-D
25	BM25 of whole document	Q-D
26	LMIR.ABS of body	Q-D
27	LMIR.ABS of anchor	Q-D
28	LMIR.ABS of title	Q-D
29	LMIR.ABS of URL	Q-D
30	LMIR.ABS of whole document	Q-D
31	LMIR.DIR of body	Q-D
32	LMIR.DIR of anchor	Q-D
33	LMIR.DIR of title	Q-D
34	LMIR.DIR of URL	Q-D
35	LMIR.DIR of whole document	Q-D
36	LMIR.JM of body	Q-D
37	LMIR.JM of anchor	Q-D
38	LMIR.JM of title	Q-D
39	LMIR.JM of URL	Q-D
40	LMIR.JM of whole document	Q-D
41	Sitemap based term propagation	Q-D
42	Sitemap based score propagation	Q-D
43	Hyperlink based score propagation: weighted in-link	Q-D
44	Hyperlink based score propagation: weighted out-link	Q-D
45	Hyperlink based score propagation: uniform out-link	Q-D
46	Hyperlink based propagation: weighted in-link	Q-D
47	Hyperlink based feature propagation: weighted out-link	Q-D
48	Hyperlink based feature propaga-	Q-D

→ Que doit on prédire ?

→ Un score ?

→ Une classe ?

Formulation du problème d'apprentissage

SCIENCES SORBONNE UNIVERSITÉ

→ Calculer un score de pertinence revient à apprendre f(q,d)

→ Plusieurs questions se posent :

→ Comment modéliser la requête et le document ?

ID	Feature Description	Category
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in body	Q-D
2	$\sum_{q_i \in q \cap d} c(q_i, d)$ in anchor	Q-D
3	$\sum_{q_i \in q \cap d} c(q_i, d)$ in title	Q-D
4	$\sum_{q_i \in q \cap d} c(q_i, d)$ in URL	Q-D
5	$\sum_{q_i \in q \cap d} c(q_i, d)$ in whole document	Q-D
6	$\sum_{q_i \in q} idf(q_i)$ in body	Q
7	$\sum_{q_i \in q} idf(q_i)$ in anchor	Q
8	$\sum_{q_i \in q} idf(q_i)$ in title	Q
9	$\sum_{q_i \in q} idf(q_i)$ in URL	Q
10	$\sum_{q_i \in q} idf(q_i)$ in whole document	Q
11	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in body	Q-D
12	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in anchor	Q-D
13	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in title	Q-D
14	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in URL	Q-D
15	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot idf(q_i)$ in whole document	Q-D
16	$\|d\|$ of body	D
17	$\|d\|$ of anchor	D
18	$\|d\|$ of title	D
19	$\|d\|$ of URL	D
20	$\|d\|$ of whole document	D
21	BM25 of body	Q-D
22	BM25 of anchor	Q-D
23	BM25 of title	Q-D
24	BM25 of URL	Q-D
25	BM25 of whole document	Q-D
26	LMIR.ABS of body	Q-D
27	LMIR.ABS of anchor	Q-D
28	LMIR.ABS of title	Q-D
29	LMIR.ABS of URL	Q-D
30	LMIR.ABS of whole document	Q-D
31	LMIR.DIR of body	Q-D
32	LMIR.DIR of anchor	Q-D
33	LMIR.DIR of title	Q-D
34	LMIR.DIR of URL	Q-D
35	LMIR.DIR of whole document	Q-D
36	LMIR.JM of body	Q-D
37	LMIR.JM of anchor	Q-D
38	LMIR.JM of title	Q-D
39	LMIR.JM of URL	Q-D
40	LMIR.JM of whole document	Q-D
41	Sitemap based term propagation	Q-D
42	Sitemap based score propagation	Q-D
43	Hyperlink based score propagation: weighted in-link	Q-D
44	Hyperlink based score propagation: weighted out-link	Q-D
45	Hyperlink based score propagation: uniform out-link	Q-D
46	Hyperlink based propagation: weighted in-link	Q-D
47	Hyperlink based feature propagation: weighted out-link	Q-D
48	Hyperlink based feature propaga-	Q-D

→ Que doit on prédire ? Un score ou une classe ?

Mais…

Pourquoi n'est-ce pas adapté à un problème de RI ?

Formulation du problème d'apprentissage

SCIENCES SORBONNE UNIVERSITÉ

- Apprendre un ordre total sur un ensemble d'articles X induit que :
  - Cet ordre permettra de comparer tous les couples d'articles dans X
  - Compte tenu de cet ordre total, n'importe quel sous-ensemble de X peut être ordonné.

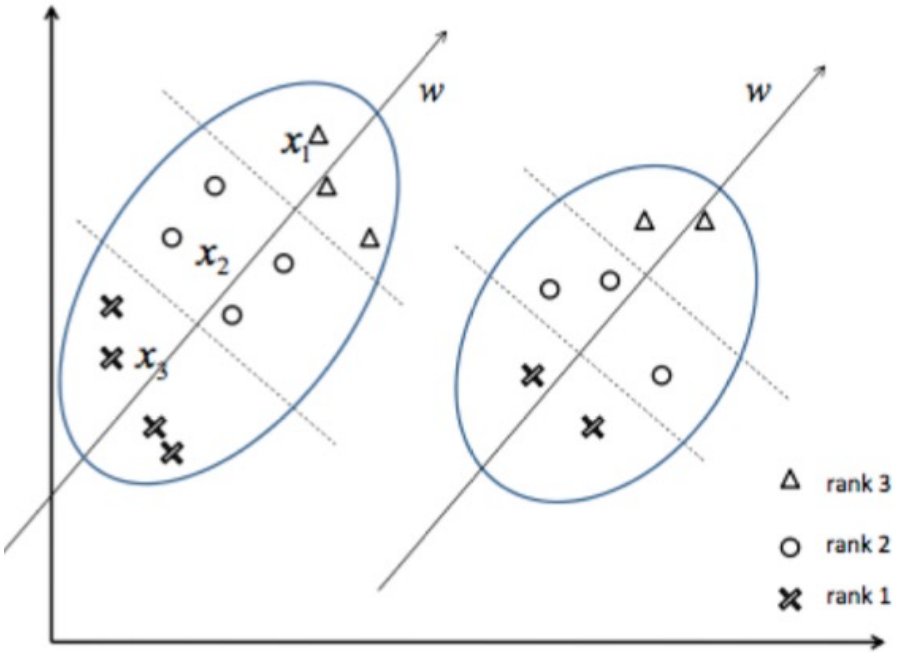
Méthode

- Apprendre à ordonner des paires d'exemples : Une erreur se produit lorsque deux éléments sont mal ordonnés.
  → Seuls les scores relatifs d'une paire sont importants.
- Il n'est pas nécessaire d'ordonner toutes les paires : seul un petit sous-ensemble des exemples est nécessaire.
  → Ceci fournira un ordre partiel sur les éléments de X
- Une fonction d'ordonnancement f sera apprise sur cet ensemble partiellement ordonné.
  → Elle permettra ensuite d'étendre l'ordre partiel à un ordre total sur tous les éléments

Comment apprendre à ordonner ?

SCIENCES SORBONNE UNIVERSITÉ

→ Première approche :
apprendre à détecter le rang
d'un document
(un rang = une classe)

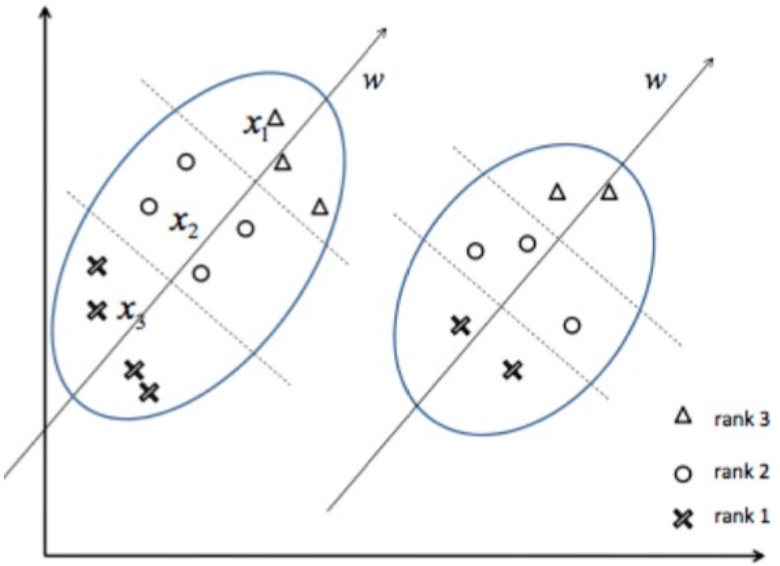


Manning et al, 2008. Introduction to information retrieval

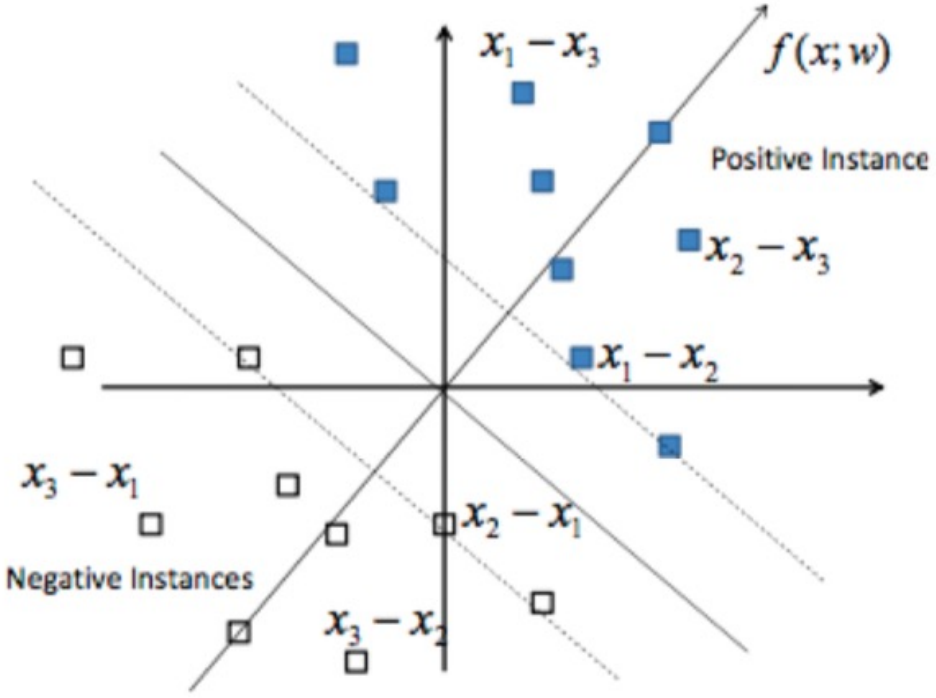
Comment apprendre à ordonner ?

SCIENCES SORBONNE UNIVERSITÉ

→ Première approche : apprendre à détecter le rang d'un document (un rang = une classe)



→ Deuxième approche : apprendre à comparer les documents



Manning et al, 2008. Introduction to information retrieval

Catégorisation des méthodes d'apprentissage d'ordonnancement

- Pointwise
  - Estimer une note ou un label (pertinent ou non) pour un seul document
  - Les documents sont indépendants (classification, ...)

PCE

$$\Delta_{PCE}(p_\theta(q,d,y)) = y \log p_\theta(q,d) + (1-y) \log 1 - p_\theta(q,d)$$

- Facile à mettre en place
- Peut être instable lors de l'apprentissage

Catégorisation des méthodes d'apprentissage d'ordonnancement

- Paiwise
  - Estimation d'une préférence d'ordre partiel dans les paires de documents
  - Documents dépendants au sein de la paire
  - Les paires de documents sont indépendantes

Max-Margin

max-margin loss

$$\Delta(f_\theta(q, d_+), f_\theta(q, d_-)) = \max\{0, 1 - (f_\theta(q, d_+) - f_\theta(q, d_-))\}$$

Le minimum 0 est atteint si

$$f_\theta(q, d_+) > f_\theta(q, d_-) + 1$$

Catégorisation des méthodes d'apprentissage d'ordonnancement

SCIENCES SORBONNE UNIVERSITÉ

- Listwise
  - Estimer l'ordre total dans une liste de documents
  - Les scores des documents dépendent les uns des autres

InfoNCE

Coût défini pour une liste de documents ≈ mesure de RI
La plus utilisée actuellement en LETOR / RI

$$\Delta(f, q, \tilde{D}) = \sum_{d_+ \in \tilde{D} \cap D_q^+} R(q, d_j) \log \frac{\exp(f(q, d_+))}{\sum_{d \in \tilde{D}} \exp(f(q, d))}$$

où R est 1 si d_j est pertinent pour q (0 sinon)

- Très utilisé en pairwise ($\tilde{D}$ avec deux documents) pour le ré-ordonnancement
- Pour l'ordonnancement, on utilise l'ensemble / un sous-ensemble des documents du batch

Enjeu = il faut trouver le meilleur $\tilde{D}$

Lindgren, E. et al. 2021. Efficient Training of Retrieval Models using Negative Cache.

- Essaie de trouver l'ensemble $\tilde{D}$ tel que le gradient soit le plus proche possible du cas où $\tilde{D} = D$

Catégorisation des méthodes d'apprentissage d'ordonnancement

SCIENCES SORBONNE UNIVERSITÉ

Category	Algorithms
Pointwise Approach	**Regression**: Least Square Retrieval Function (TOIS 1989), Regression Tree for Ordinal Class Prediction (Fundamenta Informaticae, 2000), Subset Ranking using Regression (COLT 2006), ... **Classification**: Discriminative model for IR (SIGIR 2004), McRank (NIPS 2007), ... **Ordinal regression**: Pranking (NIPS 2002), OAP-BPM (EMCL 2003), Ranking with Large Margin Principles (NIPS 2002), Constraint Ordinal Regression (ICML 2005), ...
Pairwise Approach	Learning to Retrieve Information (SCC 1995), Learning to Order Things (NIPS 1998), Ranking SVM (ICANN 1999), RankBoost (JMLR 2003), LDM (SIGIR 2005), RankNet (ICML 2005), Frank (SIGIR 2007), MHR(SIGIR 2007), GBRank (SIGIR 2007), QBRank (NIPS 2007), MPRank (ICML 2007), IRSVM (SIGIR 2006), LambdaRank (NIPS 2006),...
Listwise Approach	**Non-measure specific**: ListNet (ICML 2007), ListMLE (ICML 2008), BoltzRank (ICML 2009) ... **Measure-specific**: AdaRank (SIGIR 2007), SVM-MAP (SIGIR 2007), SoftRank (LR4IR 2007), RankGP (LR4IR 2007), ...

Figure 1: source : https://www.programmersought.com/article/188249556/

Aparté sur le temps de calcul

SCIENCES SORBONNE UNIVERSITÉ

→ En entraînement : on définit des paires
(souvent à partir des jugements de pertinence)

→ En inférence : impossible de calculer pour pour tous les documents de la collection
(trop lent)

Ordonnancement en deux étapes (Two-Stage Ranking)

1. On cherche le top-K (ex. $K = 1000$) avec BM25 ou un autre modèle **rapide**
2. On ré-ordonne les K document avec φ

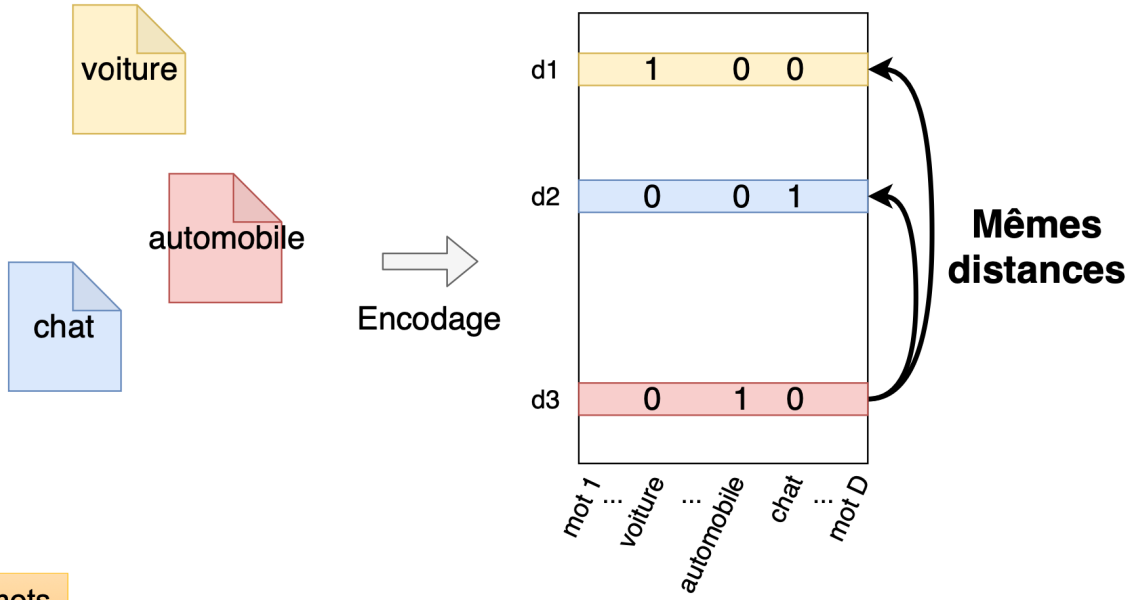
RI neuronale

De l'appariement exact vers l'appariement sémantique

SCIENCES SORBONNE UNIVERSITÉ

→ Appariement exact n'est pas suffisant (surtout du fait du fossé sémantique entre les termes)

Par défaut, la distance entre tous les mots est identique

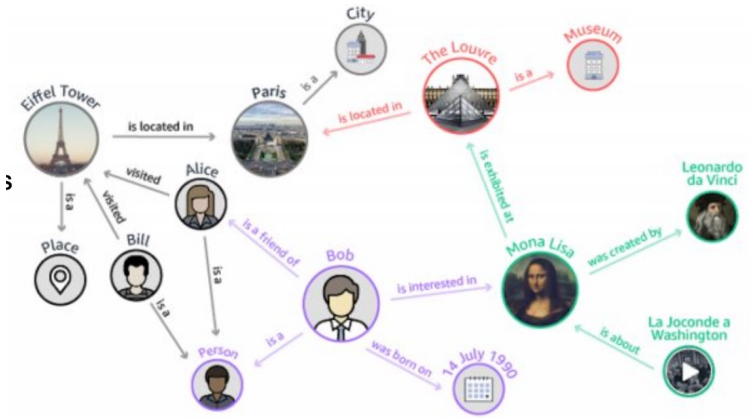


Sémantique = distance entre mots

De l'appariement exact vers l'appariement sémantique

SCIENCES SORBONNE UNIVERSITÉ

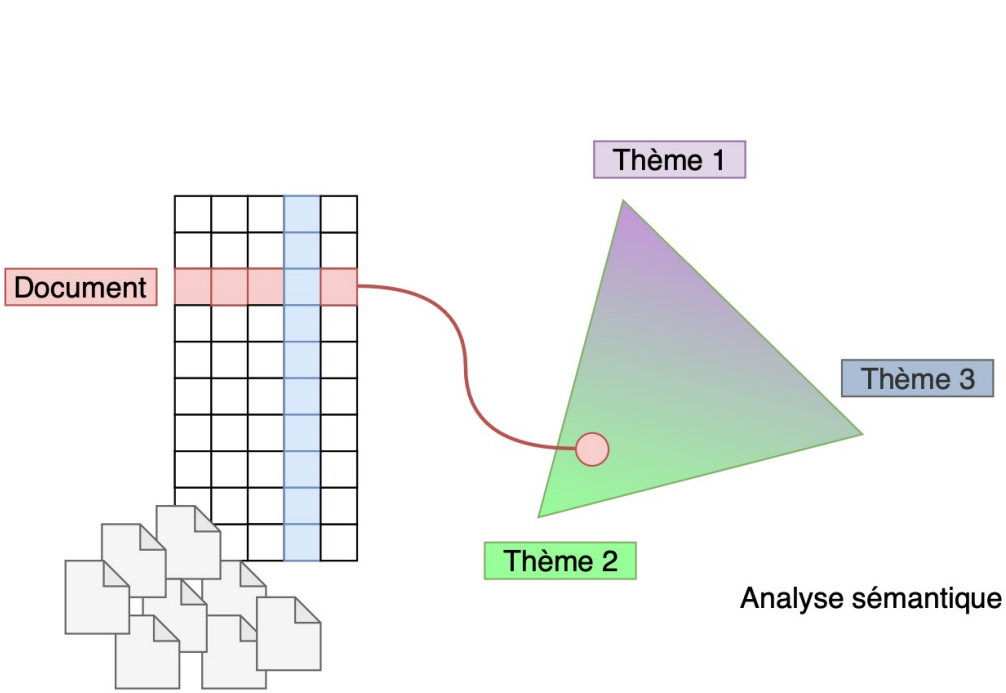
- ➜ Nombreuses approches orientées bases de connaissances.
- ➜ Mais demande des connaissances expertes



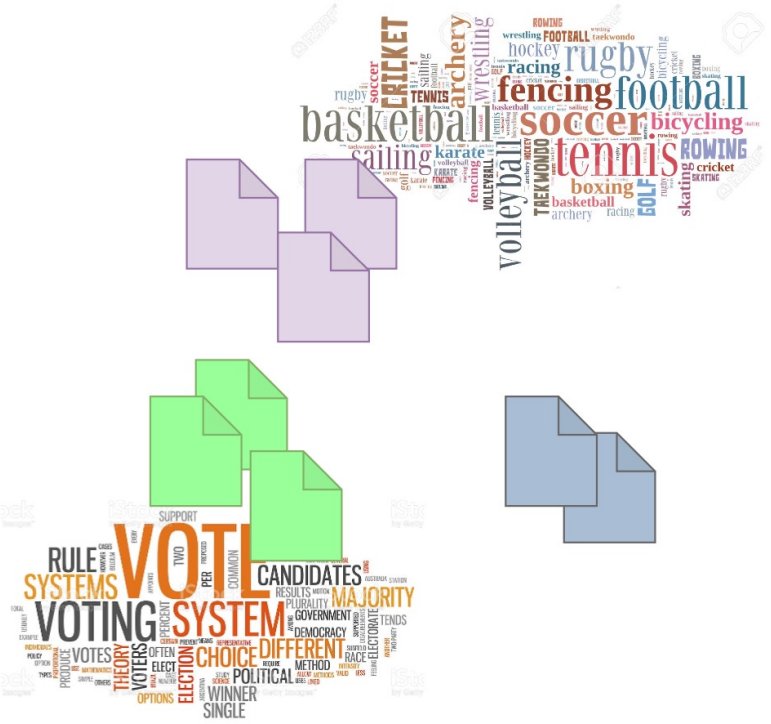
- Reformulation de la requête
- Augmentation du processus d'indexation
- Appariement augmenté
- Profils sémantique des utilisateurs

De l'appariement exact vers l'appariement sémantique

- ➜ Les représentation latentes permettent de dépasser cette contrainte…
  - ➜ Premiers topic modèles : analyse de la distribution des termes pour construire des groupes de termes.



Analyse sémantique



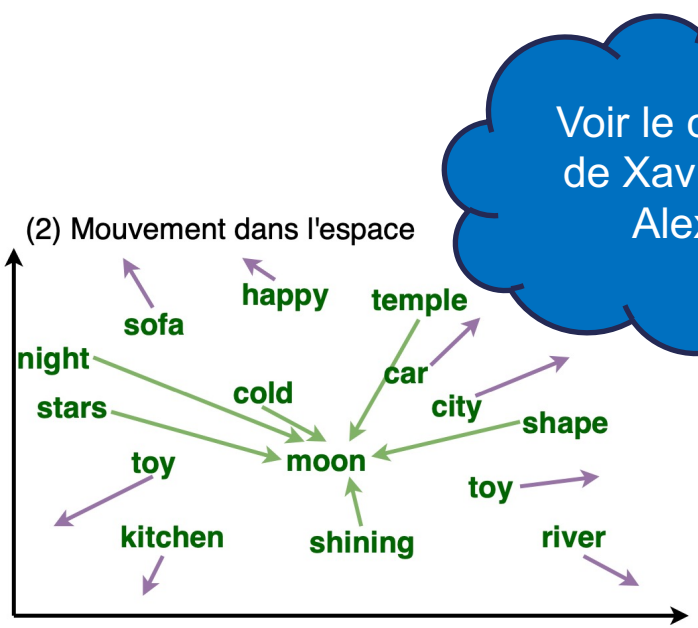
Analyse sémantique latente: LSA, pLSA, LDA

De l'appariement exact vers l'appariement sémantique

SCIENCES SORBONNE UNIVERSITÉ

- ➜ Les représentation latentes permettent de dépasser cette contrainte…
  - ➜ Récents modèles axés sur l'apprentissage profond / embeddings
  - ➜ Et s'avèrent très puissantes….

```
he curtains open and the moon shining in on the barely
ars and the cold , close moon " . And neither of the w
rough the night with the moon shining so brightly , it
made in the light of the moon . It all boils down , wr
 surely under a crescent moon , thrilled by ice-white
sun , the seasons of the moon ? Home , alone , Jay pla
m is dazzling snow , the moon has risen full and cold
un and the temple of the moon , driving out of the hug
 in the dark and now the moon rises , full and amber a
bird on the shape of the moon over the trees in front
 But I could n't see the moon or the stars , only the
rning , with a sliver of moon hanging among the stars
 they love the sun , the moon and the stars . None of
the light of an enormous moon . The plash of flowing w
man 's first step on the moon ; various exhibits , aer
 the inevitable piece of moon rock . Housing The Airsh
oud obscured part of the moon . The Allied guns behind
```



Voir le cours de Xavier & Alex

Comprendre = mesurer une distance entre les mots

2012

Apprentissage de représentation, Word2Vec, FastText, …

Mais est-ce vraiment binaire ?

SCIENCES SORBONNE UNIVERSITÉ

Query: united states president

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

The President of the United States of America (POTUS) is the elected head of state and head of government of the United States. The president leads the executive branch of the federal government and is the commander in chief of the United States Armed Forces. Barack Hussein Obama II (born August 4, 1961) is an American politician who is the 44th and current President of the United States. He is the first African American to hold the office and the first president born outside the continental United States.

Traditional IR models estimate relevance based on lexical matches of query terms in document

Representation learning based models garner evidence of relevance from all document terms based on semantic matches with query

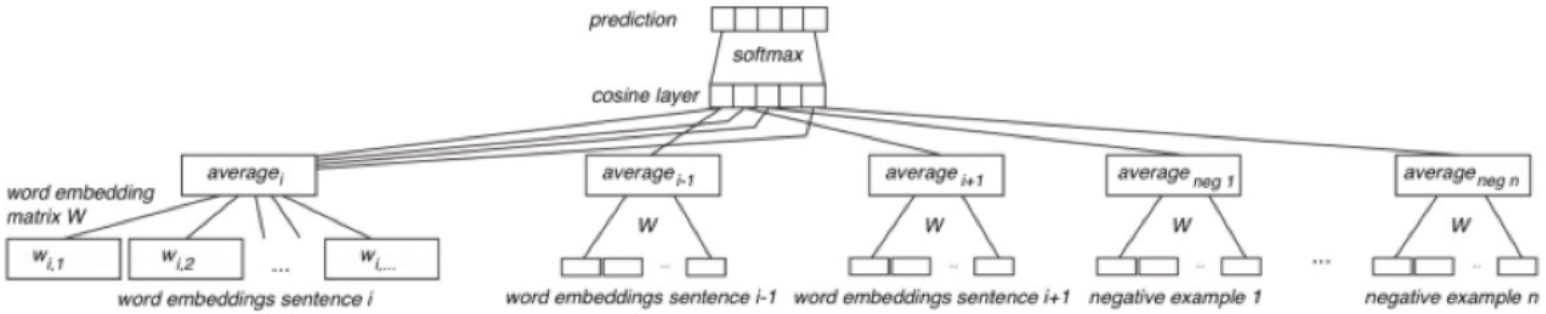
Both lexical and semantic matching are important and can be modelled with **neural networks**

© Kenter et al, 2018. Tutorial – Neural networks for Information Retrieval

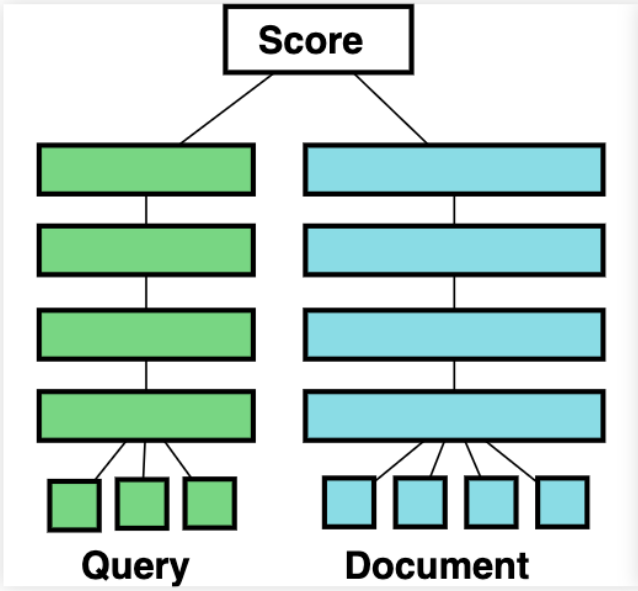
Premiers travaux en RI

- Principe de compositionalité
  - Un document / une requête : moyenne des représentations des mots
  - Eventuellement pondérées par TF-IDF

- Calcul du score :
  - Cosinus (retour aux modèles vectoriels)
  - Réseaux de neurones

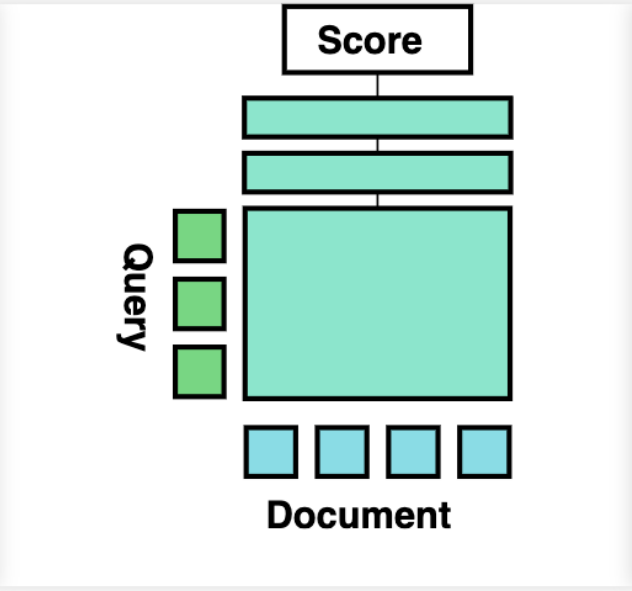


Vers des modèles end-to-end



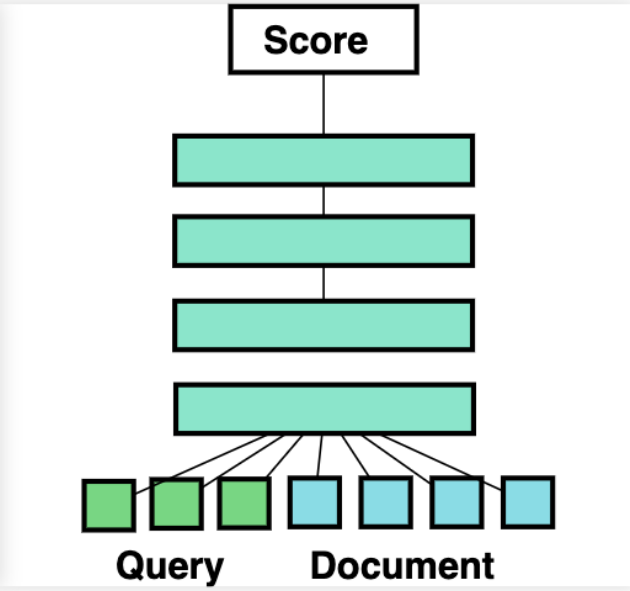
Modèle de représentation

⊞ Rapide
⊟ Moins robuste
👍 ANCE, TAS-Balanced, SparTerm, SPLADE, ...



Modèle d'interaction faible

⊞ Plus robuste (à partir de ColBERT)
⊟ Un peu plus lent
👍 DRMM, ColBERT

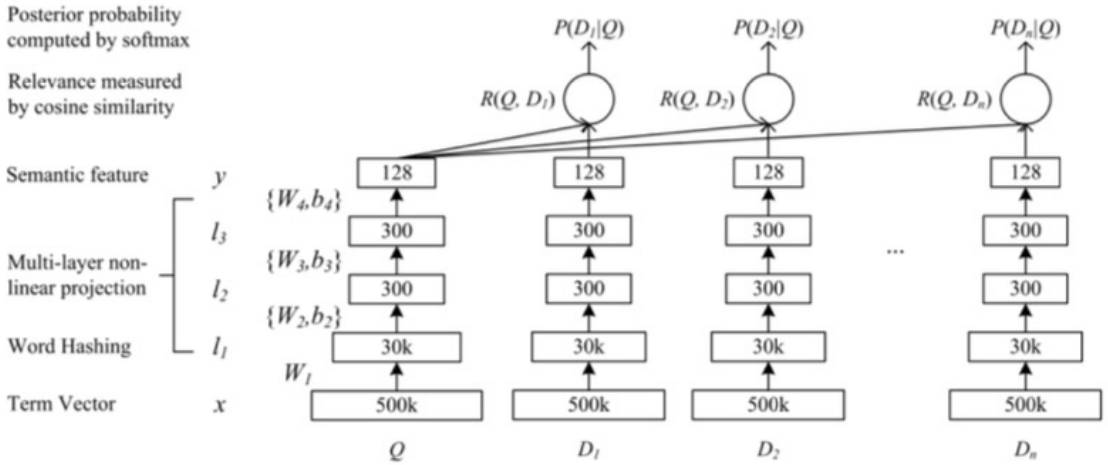


Modèle d'interaction forte

Modèle d'interaction forte
⊞ Meilleures performances
⊟ Très lent
👍 monoBERT

Modèles denses : un des premiers modèles neuronaux

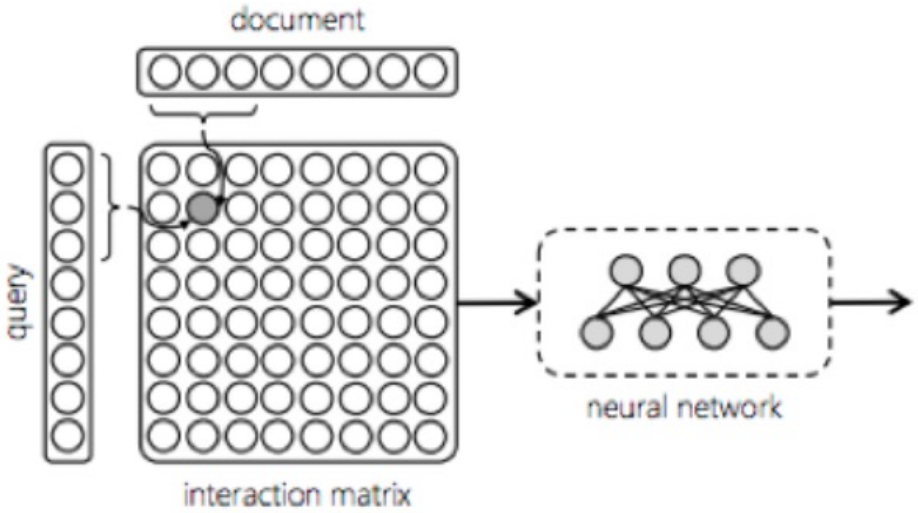
- Deep Semantic Matching Model (Huang et al, 2014)



- On va calculer le score pour plusieurs paires de requête-document (la même requête):
  - 1 paire document pertinent-requête
  - N-1 paires de documents non pertinents-requête
- On va faire en sorte que le score du document pertinent soit supérieur à celui des documents non pertinents.
- Pour cela, les paires ont un facteur commun : la requête.
- Cela permet de capture les facteurs de similarité/pertinence en opposant des exemples négatifs.

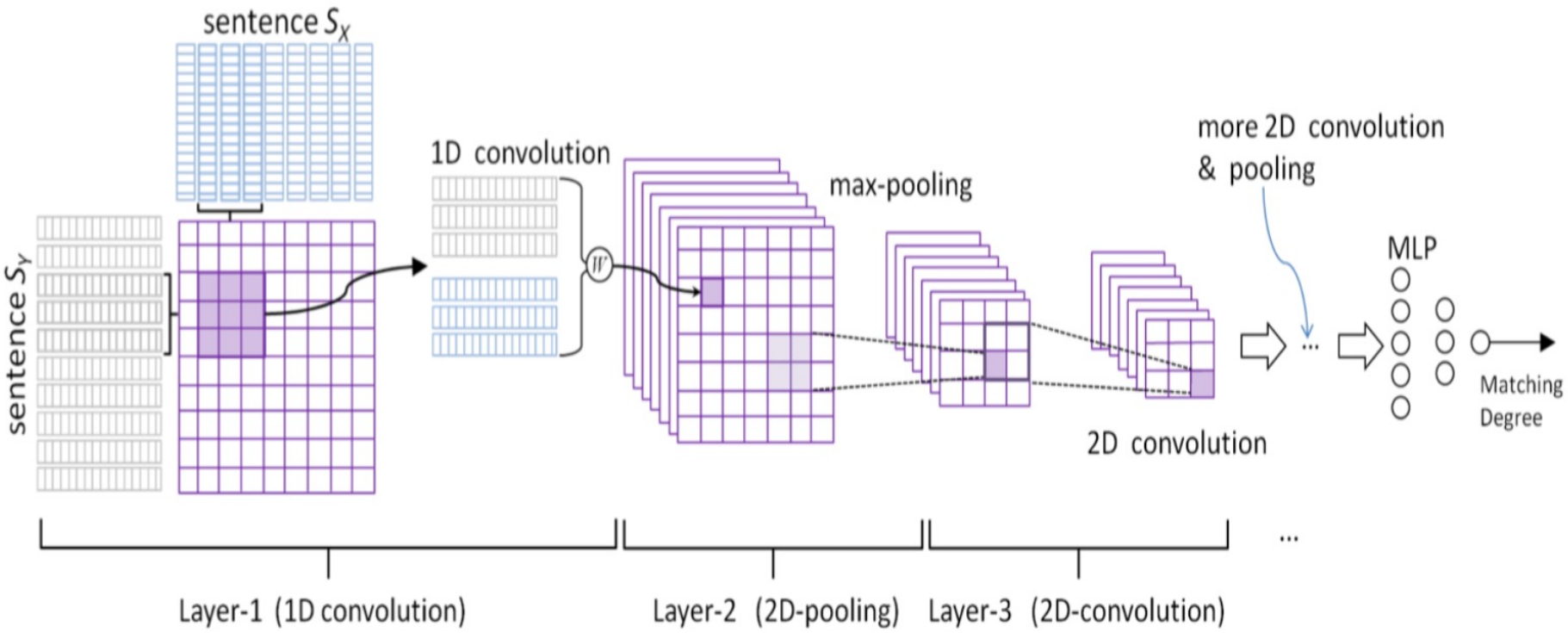
Modèles d'interaction

- Modèles basés sur l'interaction (à droite) : on pré-calcule une matrice d'interaction ayant pour but de calculer la similarités des termes de la requête et des documents (paires à paires) et de donner cette matrice en entrée au réseau de neurones pour apprendre le score de similarité.

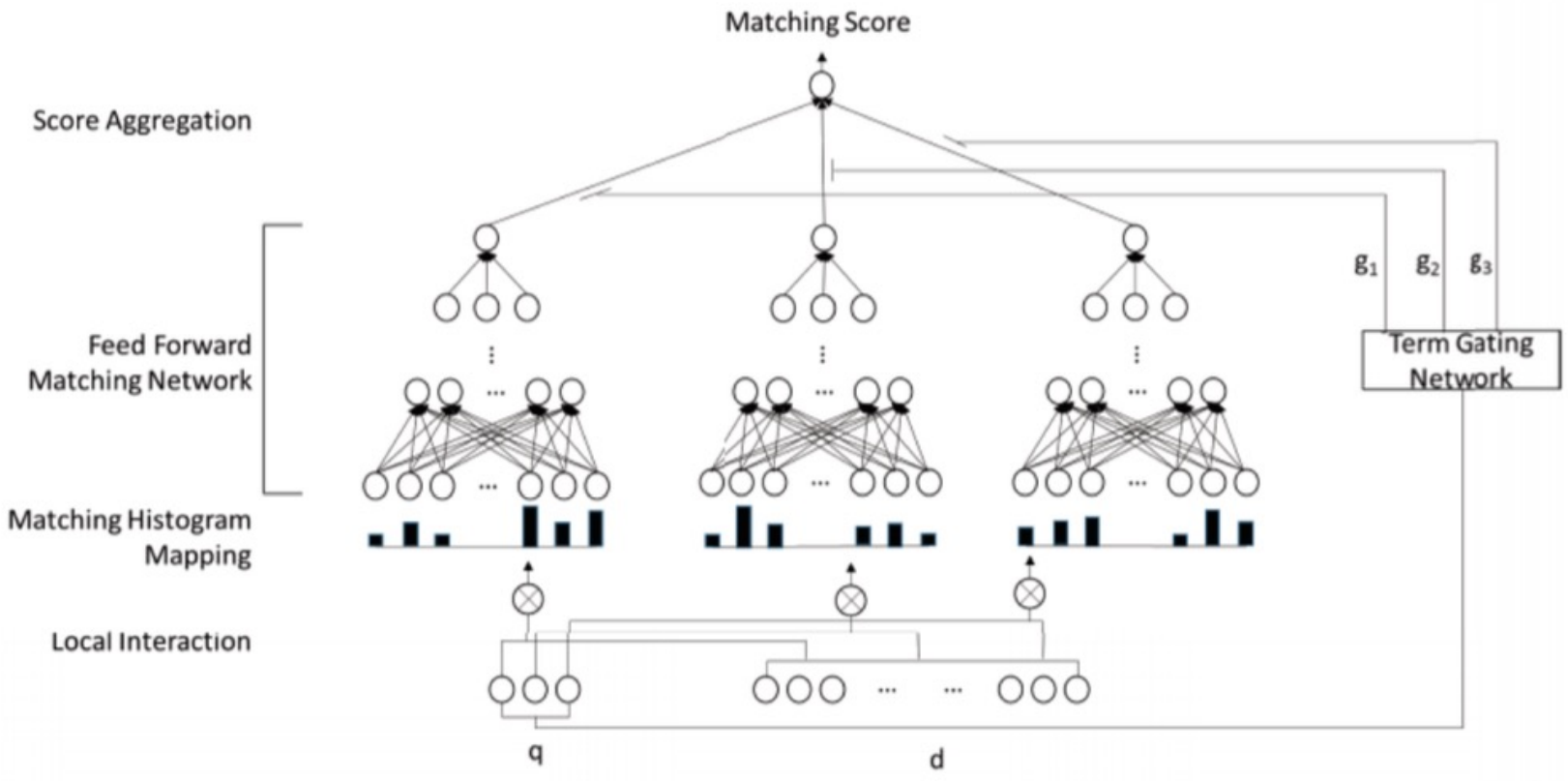


© Kenter et al, 2018. Tutorial – Neural networks for Information Retrieval

ARC-II [Hu et al]



Deep Relevance Matching model [Guo et al. 2016]

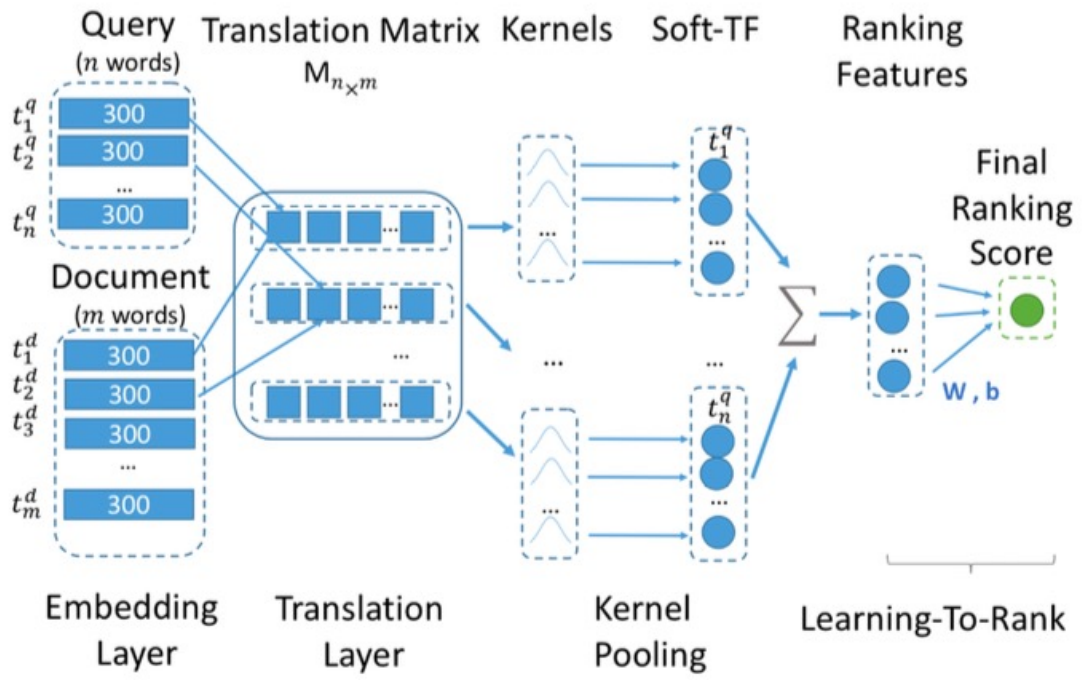


- On transforme la matrice d'interaction en histogramme
- Les histogrammes correspondent à des vecteurs données en entrée du réseau

KNRM – Kernel Neural Relevance Model

SCIENCES SORBONNE UNIVERSITÉ

Un des meilleurs modèles d'interaction (avant BERT...)



Généralisation du DRMM

$$K_k(\underbrace{q_i \hat{d}}_{I_i}) = \exp\left(\frac{(I_{ij} - \mu_k)^2}{2\sigma_k^2}\right)$$

avec $\hat{d} \in \mathbb{R}^{n \times \ell}$ et $q_i \in \mathbb{R}^n$

Xiong, C. et al. 2017. End-to-End Neura
Ad-Hoc Ranking with Kernel Pooling.

Xiong et al, 2017. End-to-end neural ad-hoc ranking with kernel pooling.

Et si on prenait un peu de recul….

- Résultats en RI

Data Set / Model	Robust04		$GOV2_{MQ2007}$		Sogou-Log	
	MAP	P@20	MAP	P@10	NDCG@1	NDCG@10
BM25[46] (1994)^{2,4}	0.255	0.370	0.450	0.366	0.142	0.287
QL[116] (1998)^{1,4}	0.253	0.369	\	\	0.126	0.282
RM3[117](2001)^{4}	0.287	0.377	\	\	\	\
RankSVM[118] (2002)^{2}	\	\	0.464	0.381	0.146	0.309
LambdaMart[97] (2010)^{2,4}	\	\	0.468	0.384	\	\
DSSM[13] (2013)$^{1,2}_{S/R/G}$	0.095	0.171	0.409	0.352	\	\
CDSSM[47] (2014)$^{1,2}_{S/R/G}$	0.067	0.125	0.364	0.291	0.144	0.333
ARC-I[17] (2014)$^{1,2}_{S/R/G}$	0.041	0.065	0.417	0.364	\	\
ARC-II[17] (2014)$^{1,2}_{S/I/G}$	0.067	0.128	0.421	0.366	\	\
MP[18] (2016)$^{1,2}_{S/I/G}$	0.189	0.290	0.434	0.371	0.218	0.380
Match-SRNN[69] (2016)$^{2}_{S/H/G}$	\	\	0.456	0.384	\	\
DRMM[21] (2016)$^{1,2,4}_{A/I/G}$	0.279	0.382	0.467	0.388	0.137	0.315
Duet[23] (2017)$^{3}_{A/H/G}$	\	\	0.474	0.398	\	\
DeepRank[33] (2017)$^{2}_{A/I/G}$	\	\	0.497	0.412	\	\
K-NRM[85] (2017)$^{4}_{A/I/G}$	\	\	\	\	0.264	0.428
SNRM[28] (2018)$^{5}_{S/R/G}$	0.286	0.377	\	\	\	\
SNRM+PRF[28] (2018)$^{5}_{S/R/G}$	0.297	0.395	\	\	\	\
CONV-KNRM[84] (2018)$^{4}_{A/I/M}$	\	\	\	\	0.336	0.481
HiNT[34] (2018)$^{3}_{A/I/G}$	\	\	0.502	0.418	\	\

- BM25 fonctionne bien!
- Plus de données = c'est mieux
- On travaille beaucoup sur les textes courts (titres)

Yang et al., 2019. Critically examining the « neural hype »

Modèles denses modernes

➜ Ils sont tous basés sur le même processus : modèle vectoriel sur des représentations Transformer

$$\hat{q} = BERT_{\texttt{CLS}}(\texttt{[CLS]}\ q_1 \dots q_n\ \texttt{[SEP]})$$

et

$$\hat{d} = BERT_{\texttt{CLS}}(\texttt{[CLS]}\ d_1 \dots q_m\ \texttt{[SEP]})$$

Le score est donné par

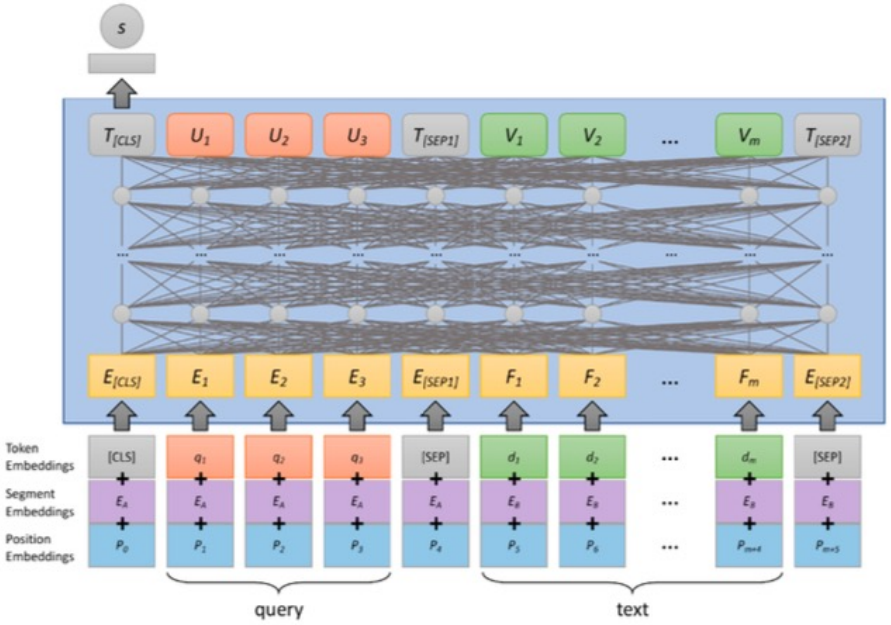
$$s(q,d) = \hat{q} \cdot \hat{d}$$

Modèles cross-encoders : MonoBERT (2019)

SCIENCES SORBONNE UNIVERSITÉ

On utilise BERT en concaténant question et document cible

$$rsv(q, d) = x^{[\mathrm{CLS}]}([\mathrm{CLS}]\ \mathrm{q}\ [\mathrm{SEP}]\ \mathrm{d})$$



Nogueira, R. and Cho, K. 2019. Passage Re-ranking with BERT.

- De très bons résultats
- Travaux récents permettent de réduire le temps de traitement (modèle EPIC de Macavaney et al. 2020).

Et les résultats ?

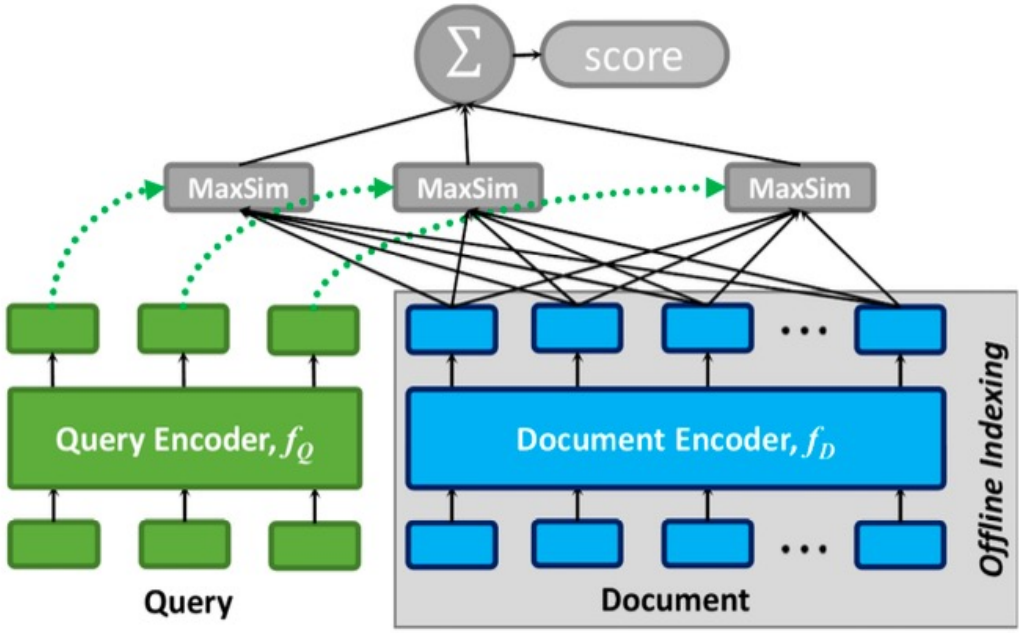
SCIENCES SORBONNE UNIVERSITÉ

	Method	TREC 2019 DL Passage		
		nDCG@10	MAP	Recall@1k
(3a)	BM25 (Anserini, $k = 1000$)	0.5058	0.3773	0.7389
(3b)	+ monoBERT$_{\text{Large}}$	0.7383	0.5058	0.7389
(4a)	BM25 + RM3 (Anserini, $k = 1000$)	0.5180	0.4270	0.7882
(4b)	+ monoBERT$_{\text{Large}}$	0.7421	0.5291	0.7882

ColBERT (2020)

ColBERT

- Encore plus simpole = Modèle d'interaction basé sur une aggrégation (maximum)
- permet de construire un index !



Khattab, O. and Zaharia, M. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT.

Modèles parcimonieux pour éviter le pré-ordonnancement basé sur BM25

SCIENCES SORBONNE UNIVERSITÉ

- ➜ Etape de pré-ordonnancement réalisée avec BM25 (orienté rappel)
  - ➜ Identification rapide des documents pertinents
  - ➜ Représentations continues pas adaptées pour l'indexation
- ➜ Réordonnancement avec un modèle neuronal (orienté précision)
  - ➜ Affinage de la pertinence : moins efficient mais plus fin en termes de sémantique
- ➜ Comment combiner index parcimonieux et représentations denses ?

Modèles parcimonieux : SNRM (Zamani et al, 2018)

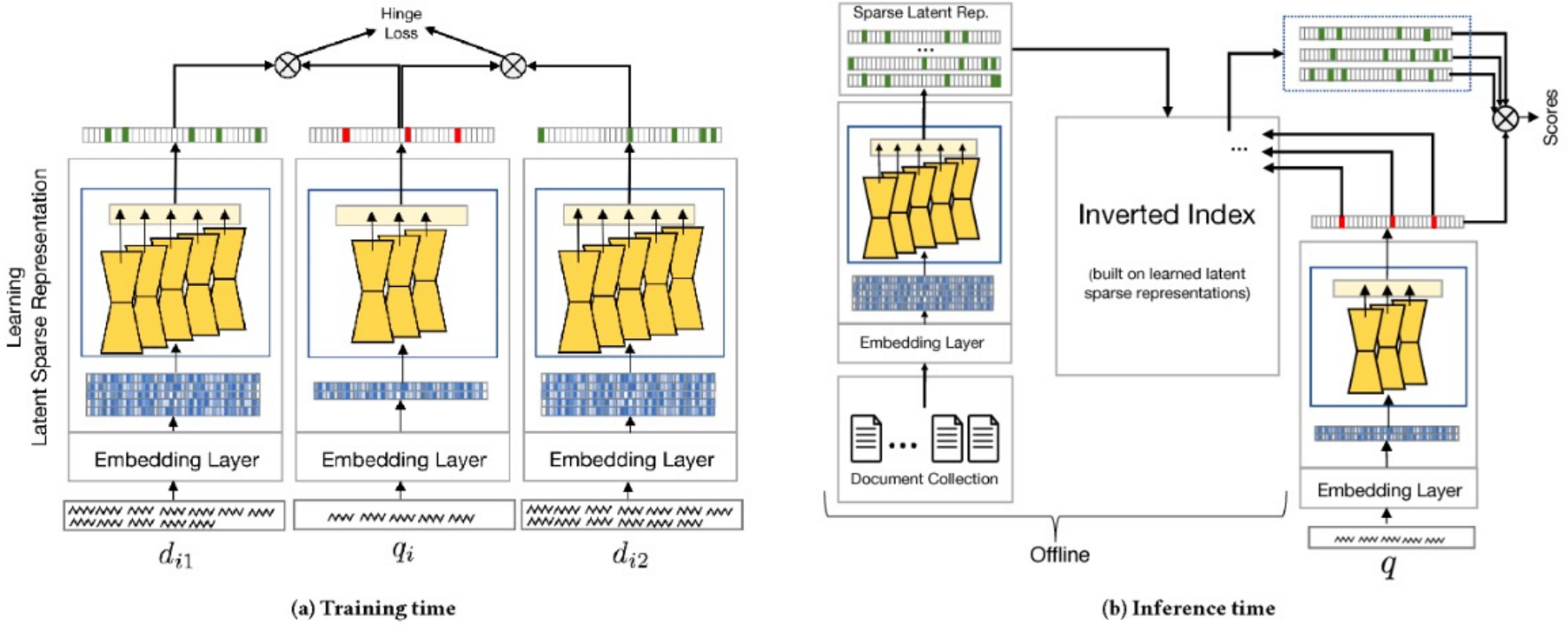


Figure 2: General schema of the proposed SNRM model.

- Offline : Apprentissage de représentation parcimonieuse pour construire l'index inverse (MLP + RELU avec une hinge loss)
- Online : Représentation parcimonieuse de la requête pour sonder l'index inverse et calculer les scores d'appariement

DeepCT

SCIENCES SORBONNE UNIVERSITÉ

→ Prédire le poids d'un terme en se basant sur sa représentation contextualisée

Table 1: Two passages that mention 'stomach' twice. Only the first passage is about the topic 'stomach'. This paper proposes a method to weight terms by their roles in a specific text context, as shown by the heatmap over terms.

In some cases, an upset **stomach** is the result of an allergic reaction to a certain type of food. It also may be caused by an irritation. Sometimes this happens from consuming too much alcohol or caffeine. Eating too many fatty foods or too much food in general may also cause an upset **stomach**.
All parts of the body (muscles , brain, heart, and liver) need energy to work. This energy comes from the food we eat. Our bodies digest the food we eat by mixing it with fluids(acids and enzymes) in the **stomach**. When the **stomach** digests food, the carbohydrate (sugars and starches) in the food breaks down into another type of sugar, called glucose.

Figure 2: Dai and Callan, 2019. Context-Aware Sentence/Passage Term Importance Estimation For First Stage Retrieval

Doc2Query [Nogueira et al, 2019]

- ➜ Augmenter l'indexation avec des questions associées au document
- ➜ Une forme détournée pour identifier des mots importants (pas forcément dans le document)

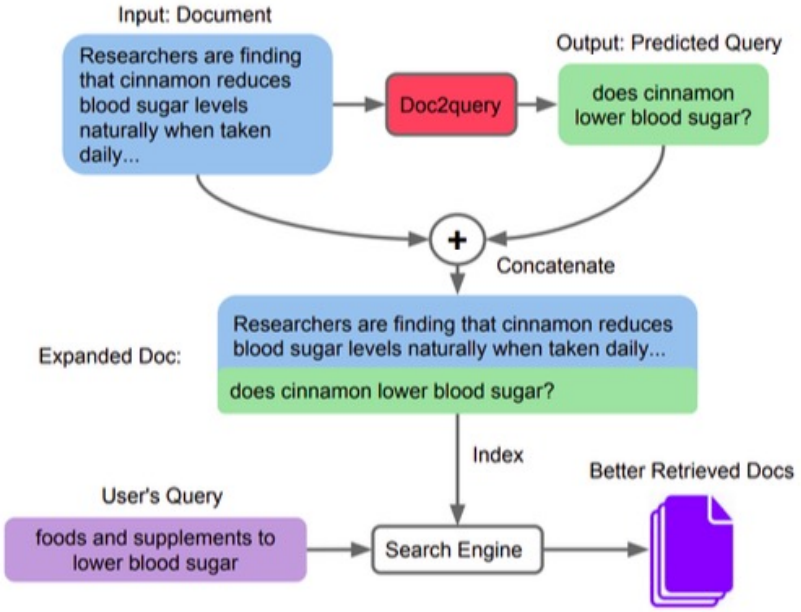


Figure 1: Given a document, our Doc2query model predicts a query, which is appended to the document. Expansion is applied to all documents in the corpus, which are then indexed and searched as before.

Stratégies d’apprentissage

Aparté sur la supervision

SCIENCES SORBONNE UNIVERSITÉ

- → Peu de données supervisées mais des modèles : Utilisons les « vieux » modèles
  Dehghani, M. et al. 2017. Neural Ranking Models with Weak Supervision.
- → Beaucoup de données mais bruitées : reformulation, clics, …
- → Et des faux négatifs : documents pertinents mais jugés non pertinents
  (peu de labels, clics, jugement humains peu nombreux)
- → Cross-encoder BERT sont assez faciles à entraîner… mais très gourmands en ressources

👉 On peut traiter le problème de de deux façons :

1. Essayer d'estimer si c'est un faux négatif avec un autre modèle
   RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering.
2. Utiliser des techniques de distillation
   👉 On minimise la distance entre les prédictions des modèles
   Gao, L. et al. 2020. Understanding BERT Rankers Under Distillation.

Distillation

- Mise en place d'architecture Teacher-Student
  - Teacher : modèle appris pour prédire des labels sur des documents (cas où on a des scores sur beaucoup de documents – processus lent)
  - Student : modèle qui s'appuie sur les labels pour apprendre l'ordonnancement

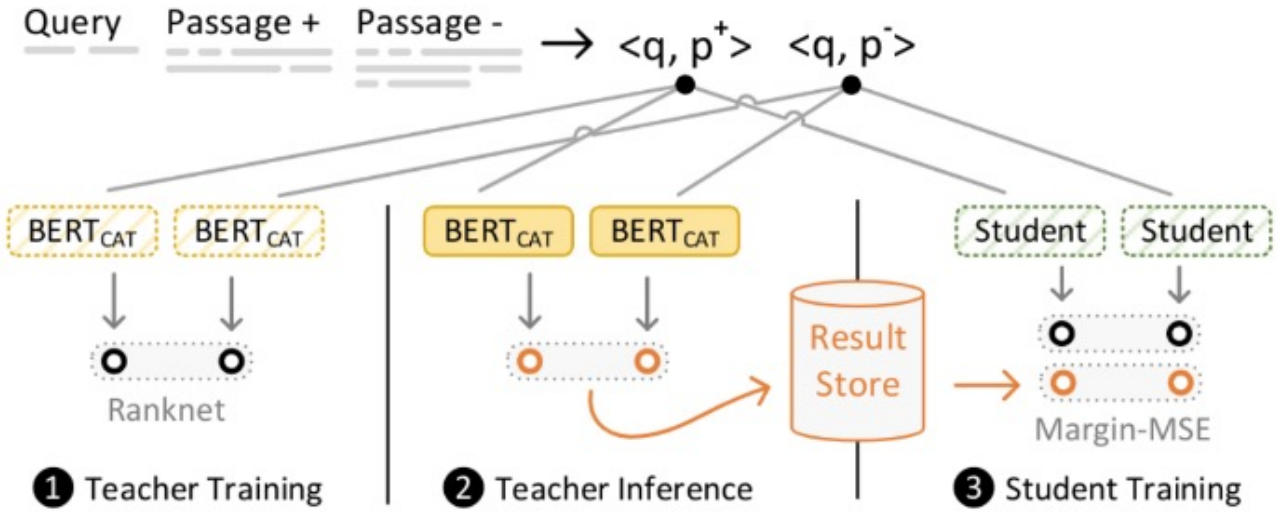


Figure 7: Hofstätter et al., 2020. Improving Efficient Neural Ranking Models with Cross-Architecture Knowledge Distillation

Pré-entraînement

Un meilleur pré-entraînement permet d'améliorer les performances

Gao, L. and Callan, J. 2021. Condenser: a Pre-training Architecture for Dense Retrieval.

- On force le CLS a encoder de l'information sémantique

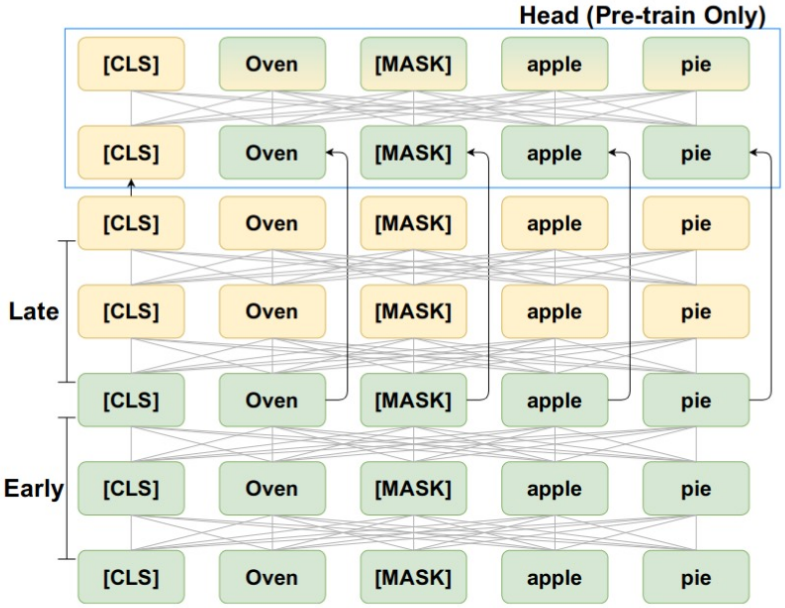


Figure 1: Condenser: Shown are 2 early and 2 late backbone layers. Our experiments each have 6 layers. Condenser Head is dropped during fine-tuning.

- On utilise des bouts de documents comme des questions

then form a training batch of coCondenser. Write a span s_{ij}'s corresponding *late* CLS representation h_{ij}, its corpus-aware contrastive loss is defined over the batch,

$$\mathcal{L}_{ij}^{co} = -\log \frac{\exp(\langle h_{i1}, h_{i2} \rangle)}{\sum_{k=1}^{n} \sum_{l=1}^{2} \mathbb{I}_{ij \neq kl} \exp(\langle h_{ij}, h_{kl} \rangle)} \quad (6)$$

Familiar readers may recognize this as the contrastive loss from SimCLR (Chen et al., 2020), for which we use random span sampling as augmentation. Others may see a connection to noise contrastive estimation (NCE). Here we provide an NCE narrative. Following the spirit of the distribu-

Jeux de données potentiels pour le pré-entraînement

SCIENCES SORBONNE UNIVERSITÉ

Dataset	Source	#Docs	Language	Latest crawl date
Books[1]	Book	74M	ENG	2015
C4[2]	web extracted text	0.3B	ENG	2019
Wikipedia[3]	Wiki text	10M	Multi-lang	monthly update
RealNews[4]	News	120GB	ENG	2019
Amazon[5]	reviews	11GB	ENG	2003
WT10G[6]	web pages	1.7M	ENG	1997
GOV2[7]	pages in .GOV	25M	ENG	2004
CWP200T	Chinese web pages	7B	CHN	2015
SogouT[8]	Sogou web pages	1.17B	CHN	2016
ClueWeb09[9]	web pages	1.04B	Multi-lang	2009
ClueWeb12[10]	web pages	0.73B	ENG	2012
MS MARCO[11]	Bing web pages	3.2M	ENG	2018

Table 7.1: Public available datasets which are potential for pre-training tasks.

Fan et al, 2022. Pre-training Methods in Information retrieval

Jeux de données pour l'évaluation en RI

SCIENCES SORBONNE UNIVERSITÉ

Dataset	Subdata	Size	Source	Potential Tasks
Robust	Robust04 Robust05	0.5M docs, 250 queries 1M docs, 50 queries	TREC Robust Track	FSR, AR, QR
TREC MQ	MQ2007 MQ2008	6.5K docs, 1.7K queries 1.4K docs, 784 queries	TREC Million Query track	FSR, AR, QR
Clueweb	09-CatB 12-CatB	50M docs, 150 queries 50M docs	Web pages	FSR, AR, QR, KE
TREC web track	99-2014	See [13]	TREC web track	FSR, AR, QR
TREC DL track	2019-2021	See [14]	TREC Deep Learning track	FSR, AR
AOL	\	6M queries	AOL Query logs	AR,SS,PS,QR,QS
Sogou-QCL	\	9M docs, 0.5M queries	Sogou Query logs	AR, QR
Sogou-SRR	\	63K results, 6K queries	Sogou Query logs	AR, MMR, QR
Tiangong-ST	\	0.3M docs, 40K queries	Sogou Query logs	AR, SS, QR, QS
Qulac	\	10K question-answer pairs	TREC Web Track	AR, QR, QC
BEIR	7 IR tasks	Vary from tasks	Wiki, Quaro, Twittter, News and etc.	FSR,AR, etc.
MS MARCO	2019-20	1M queries, 8.8M passages, 3.2M docs	TREC Deep Learning Track	FSR, AR, QR
TREC CAR	\	30M paras, 2M queries	TREC Complex answer retrieval	AR, QR, KE
CNN / Daily Mail	\	0.3M docs	Human generated abstracts	DS
New York Times (NYT)	\	1.8M docs	News articles	DS
Debatepedia	\	1,303 debates	Debate key points	SG, DS
DUC	2001-07	300 clusters, See [15]	Doc understanding conference	SG, DS
WIKIREF	\	0.3M samples	QFS benchmark	SG, DS

Table 7.2: Datasets for different downstream tasks in IR. Abbreviations in potential tasks are detailed in Section 7.2.

Fan et al, 2022. Pre-training Methods in Information retrieval

RI générative vs. Classification/prédiction

- → « All NLP tasks are generation tasks: a general pretraining framework » Du et al. 2021
- → Et pourquoi pas en RI ?

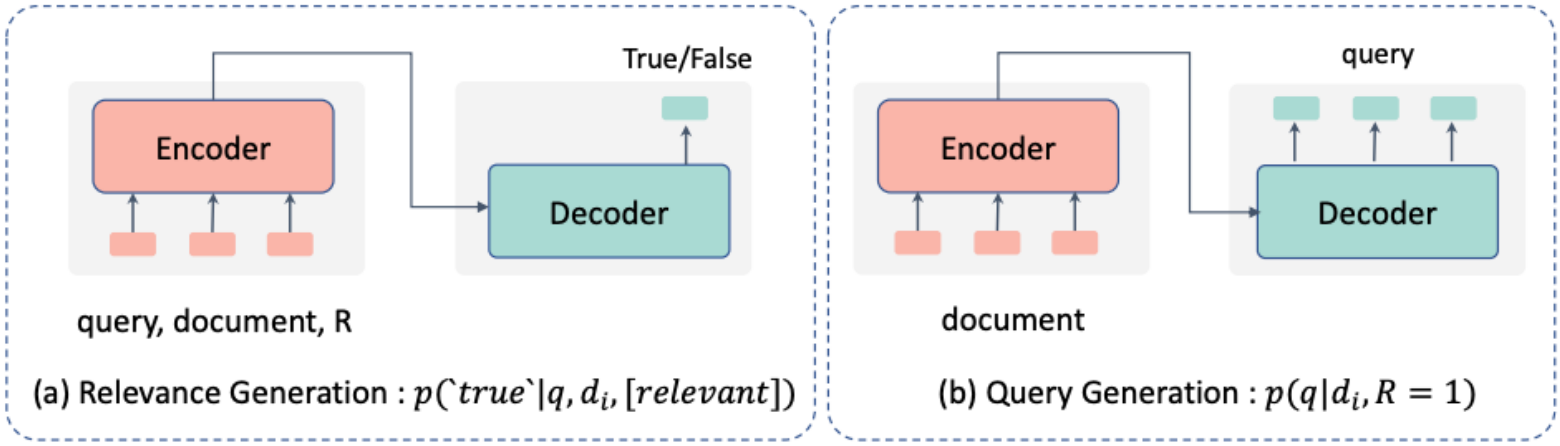


Figure 4.3: Two categories of generative ranking models.

Fan et al, 2022. Pre-training Methods in Information retrieval

Et ChatGPT dans tout cela ?

ChatGPT : qu'en disent-ils ?

SCIENCES SORBONNE UNIVERSITÉ

UN modèle ?!

- IA généraliste vs. IA étroite
- Données, architecture, entraînement ?
- Evaluation ?

Quel type d'erreurs ?

Quelles questions ?

Inapproprié ?

Voir le cours de Karën Fort

Pourquoi ChatGPT n'est pas un moteur de recherche ?

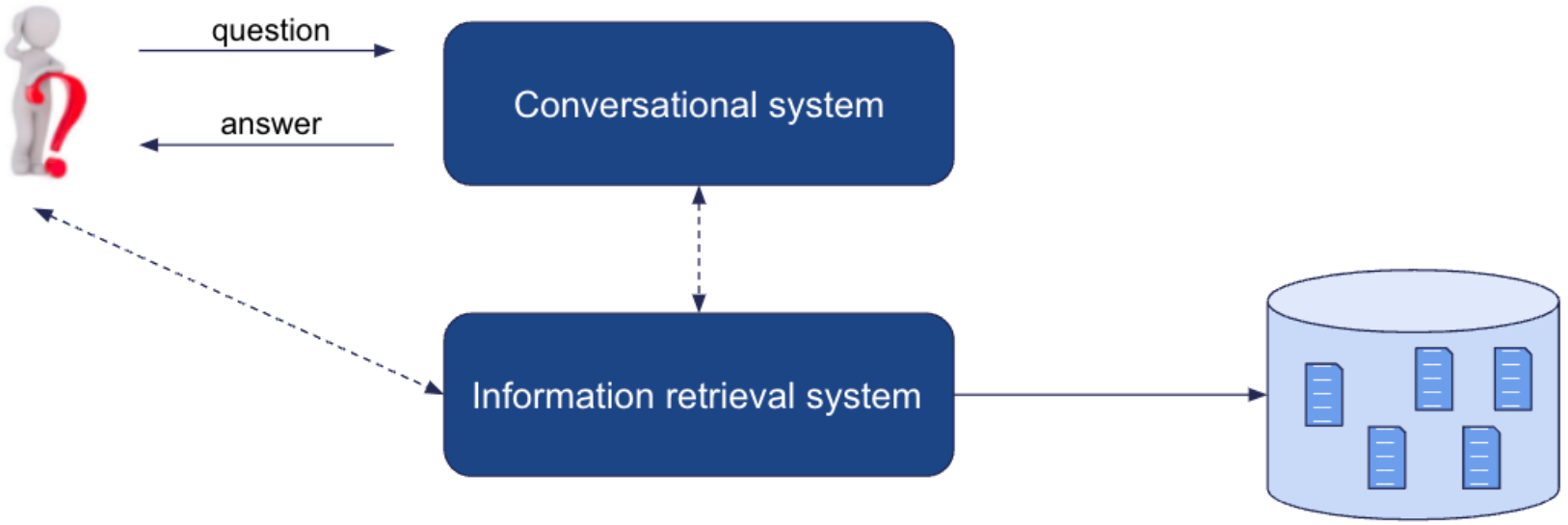
SCIENCES SORBONNE UNIVERSITÉ

→ Recherche d'information :

"Information retrieval (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)." C. Manning

	Moteur de recherche	ChatGPT
Tâche	Recherche d'information	? Générateur de texte ?
D'où est extraite la « connaissance »	Appariement requêtes-documents	Paramètres du modèle de langue
Evaluation	Précision / Rappel Pertinence des ordonnancements	?
Interactions	Requêtes, clics, …	Langage naturel

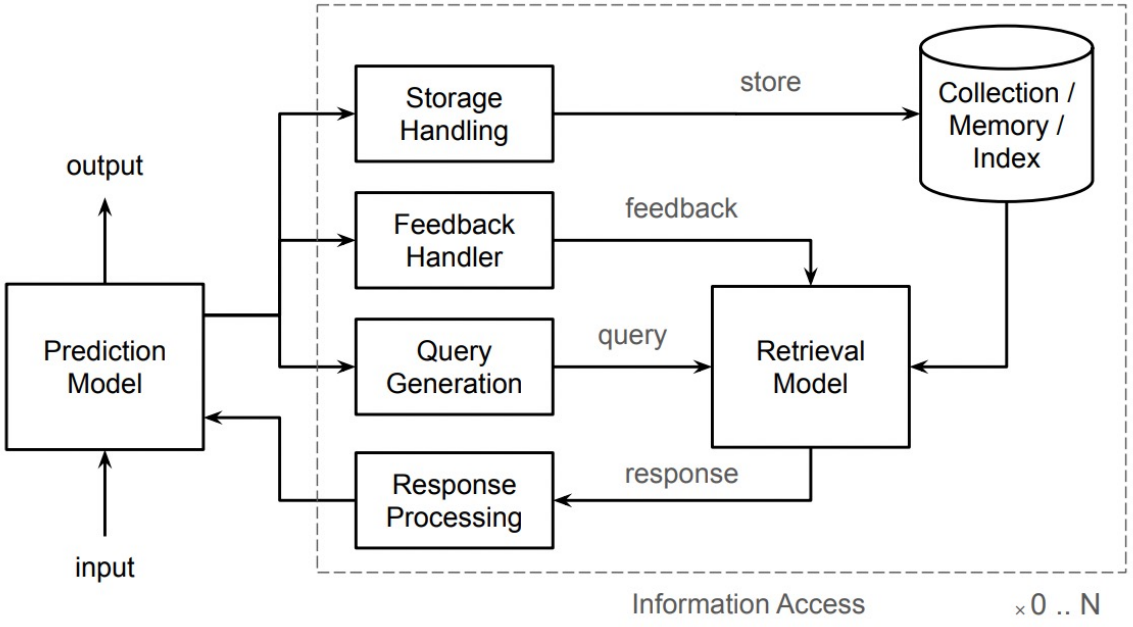
Vers la recherche conversationnelle ?



- → Importance de la source de l'information : nécessité d'identifier les documents pertinents
- → En rapport avec une conversation
- → Dialogue avec l'utilisateur de façon proactive :
  - → clarifier son besoin,
  - → suggérer des trajectoires de découverte de documents

Apprentissage augmenté par la RI

- Motivations :
  - Généralisation
  - Scalabilité
  - Mise à jour des collections et aspects temporels
  - Interprétabilité et explicabilité
  - RI sur le matériel (vie privée)



Quelques pointeurs pour vos lectures

SCIENCES SORBONNE UNIVERSITÉ

- Conférences internationales RI : SIGIR, ECIR, ICTIR, CIKM, WSDM, WWW / nationale : CORIA
- Mais aussi les conférences NLP : EMNLP, ACL, COLING, ...
- Journaux internationaux : ACM TOIS, ACM Computing Surveys, IP&M, JIR, JASIST
- Campagnes d'évaluation : TREC, NIST, CLEF, FIRE
- Livres :
  – Christopher D. Manning, Prabhakar Raghavan and Hinrich Schutze, Introduction to Information Retrieval, Cambridge University Press. 2008.
  – Massih-Reza Amini, Eric Gaussier, Recherche d'information, Applications, modeles et algorithmes, Eyrolles 2013-2018
  – W. Bruce Croft, Donald Metzler, Trevor Strohman, Search Engines Information Retrieval in Practice, Addison Wesley, 2009
- Synthèses:
  - Sur la Recherche d'Information Neuronale (en particulier les loss) : Mitra, B. and Craswell, N. 2017. An Introduction to Neural Information Retrieval.
  - Sur la RI et les transformers (très complet) : Lin, J. et al. 2020. Pretrained Transformers for Text Ranking: BERT and Beyond.

Thank you for your attention