ANR AGENCE NATIONALE DE LA RECHERCHE CAAS

Vers une représentation du contexte thématique en **Recherche d'Information**

Romain Deveaud

29 novembre 2013, Avignon

Thèse réalisée sous la direction de Patrice Bellot et Eric SanJuan

Recherche d'Information

société moderne et ultra-connectée

- production continue d'informations
- 30% du trafic internet réalisé à partir de mobiles
- ~ 8 milliards de requêtes chaque jour (Google + Yahoo! + Bing)

le monde réel s'interface avec cette « infosphère »

- moteurs de recherche comme points d'entrée
- requête ➠ information

Recherche d'Information

historiquement, RI = documents pertinents [1]

- recherche des documents contenant les mots-clés
- sortie : liste de documents ordonnée

de la RI à l'accès à l'information

- recherche de passages, de phrases ou de réponses [2,3,4]
- génération de mini-descriptions (*snippets*) [5]
- résumé multi-document orienté requête [6]

[1] D. Harman. **Information Retrieval Evaluation.** *Synthesis Lectures on Information Concepts, Retrieval, and Services.*
[2] M. Kaszkiel et J. Zobel. **Passage Retrieval Revisited.** In *Proc. of SIGIR'97*.
[3] N. Fuhr, J. Kamps, M. Lalmas, S. Malik, et A. Trotman. **Overview of the INEX 2007 Ad Hoc Track.** In *Proc. of INEX'07*.
[4] E. M. Voorhees et D. M. Tice. **The TREC-8 Question Answering Track Evaluation.** In *Proc. of TREC-8*.
[5] Y. Huang, Z. Liu, et Y. Chen. **Query Biased Snippet Generation in XML Search.** In *Proc. of SIGMOD'08*.
[6] F. Boudin et J. Torres Moreno. **Neo-cortex : A performant user-oriented multi-document summarization system.** *Computational Linguistics and Intelligent Text Processing.* (2007)

Recherche d'Information

interaction humain – machine, double défi

côté humain

- formulation précise d'un besoin d'information [7]
- réduction à un petit ensemble de mots-clés : la requête

côté machine (système)

- interprétation/compréhension du besoin d'information
- nécessité d'une certaine forme de contexte

[7] P. Ingwersen. **Polyrepresentation of Information Needs and Semantic Entities : Elements of a Cognitive Theory for Information Retrieval Interaction.** In *Proc. of SIGIR'92*.

Contexte thématique

deux types de contextes: utilisateur / thématique

- modèles des intérêts de l'utilisateur [8,9]
- modèles des thèmes liés à la requête [10,11]

contexte thématique en tant que désambiguïsation

- quels sont le ou les sens de la requête?

[8] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, et E. Ruppin. **Placing Search in Context : The Concept Revisited.** *ACM TOIS (2002).*

[9] R. W. White, P. N. Bennett, et S. T. Dumais. **Predicting Short-term Interests Using Activity-based Search Context.** In *Proc. of CIKM'10.*

[10] R. W. White, P. Bailey, et L. Chen. **Predicting User Interests from Contextual Information.** In *Proc. of SIGIR'09.*

[11] R. Kaptein et J. Kamps. **Explicit extraction of topical context.** *JASIST (2011).*

Représentation du contexte thématique

estimation fine du contexte thématique

couverture complète des informations liées à la requête (et par extension au besoin d'information)

nécessité d'avoir des modèles robustes et précis

extensions des modèles de pertinence [12,13]

[12] V. Lavrenko et W. B. Croft. **Relevance Based Language Models.** In *Proc. of SIGIR'01.*
[13] C. Zhai et J. Lafferty. **Model-based Feedback in the Language Modeling Approach to Information Retrieval.** In *Proc. of CIKM'01.*

Contributions

utilisation de différentes sources d'informations?

- combinaison de ressources [14]

modélisation des concepts implicites de la requête

- affranchissement des modèles supervisés et des ressources structurées [15]

quels sont les effets de tels concepts sur la RI? [16]

[14] R. Deveaud, E. SanJuan, et P. Bellot. **Estimating Topical Context by Diverging from External Resources.** In *Proc. of SIGIR'13.*

[15] R. Deveaud, E. SanJuan, et P. Bellot. **Unsupervised Latent Concept Modeling to Identify Query Facets.** In *Proc. of OAIR'13.*

[16] R. Deveaud, E. SanJuan, et P. Bellot. **Are Semantically Coherent Topic Models Useful for Information Retrieval ?** In *Proc. of ACL'13.*

UNIVERSITÉ D'AVIGNON

Méthodologie expérimentale

pertinence & évaluation

collections de documents

sources d'information

Pertinence

qualité de ce qui est pertinent, logique, parfaitement approprié

— Dictionnaire Larousse (2012)

Pertinence

notion centrale dans l'évaluation des systèmes de RI

« ce document répond-il à ma requête? » [17,18]

similaire à la notion d'information [19]

« quelle quantité d'information contient ce document? »

conditionnée elle aussi par la requête

notions très liées qui se confondent parfois

[17] C. J. V. Rijsbergen. **Information Retrieval** (2nd ed.). (1979)
[18] L. Schamber, M. Eisenberg, et M. S. Nilan. **A Re-examination of Relevance : Toward a Dynamic, Situational Definition.** *Information Processing & Management.* (1990)
[19] M. J. Bates. **Fundamental Forms of Information.** *Journal of the American Society for Information Science and Technology. (2006)*

Évaluation

campagnes d'évaluation au centre de la culture RI

- TREC, INEX, CLEF, NTCIR, FIRE...
- ensembles de données complets permettant évaluation et comparaison entre systèmes

collection de test

- ensemble de documents, requêtes, jugements de pertinence
- documents jugés manuellement pour leur pertinence par rapport à la requête [20]

[20] K. Jones, C. Van Rijsbergen, B. L. Research, et D. Department. **Report on the Need for and Provision of an Ideal Information Retrieval Test Collection.** (1975)

Collections de documents

collections TREC

articles journalistiques

web « général »

web « gouvernemental »

Nom	requêtes utilisées	# docs. pertinents	Par requête moyen	min.	max.
WT10g	451-550	5 980	59,8	1	519
Robust04	301-450, 601-700	17 412	69,65	3	448
GOV2	701-850	26 917	179,45	4	617
ClueWeb09-B	50-150	14 842	98,95	1	314

Nom	# documents	taille de l'index	# mots uniques	# total de mots	μ
WT10g	1 692 096	9,2 Go	5 437 563	1 043 993 839	617
Robust04	528 155	2 Go	675 713	253 367 449	480
GOV2	25 205 179	202 Go	39 286 722	23 623 611 729	937
ClueWeb09-B	50 220 423	583 Go	87 330 765	40 416 831 010	805

Sources d'information

couverture étendue du contexte thématique

« collection externe », « ressource »

sources de grande taille et de natures différentes

New York Times LDC (1987-2007)

Ressource	# documents	taille de l'index	# mots uniques	# total de mots	μ
NYT	1 855 658	11 Go	1 086 233	1 378 897 246	743
Wiki	3 406 520	12 Go	7 419 901	1 143 840 781	336
GW	4 111 240	12 Go	1 288 389	1 397 727 483	340
Web	29 038 220	336 Go	33 314 740	22 814 465 842	786

Wikipédia, version 01/12

Gigaword English LDC, dépêches journalistiques

ClueWeb09-B, sans *spam*

UNIVERSITÉ D'AVIGNON

Estimation du contexte thématique par de multiples sources d'informations

modèles de langue pour la RI

divergence à partir de sources d'information

expérimentation, résultats, discussion

Modèles de langue pour la RI

modèle probabiliste

$$P(D|Q) = \frac{P(Q|D)P(D)}{P(Q)}$$

$$P(D|Q) \propto P(Q|D)P(D)$$

modèle de la requête

$$P(D|Q) \propto P(D)P(Q|\theta_D) \propto P(D) \prod_{w \in Q} P(w|\theta_D)^{tf(w,Q)}$$

$$\log P(D|Q) \propto \log P(D) + \sum_{w \in Q} P(w|\theta_Q) \log P(w|\theta_D)$$

Motivation

retour de pertinence dit « *simulé* »

- hypothèse : les N premiers documents renvoyés par le système de RI sont considérés comme pertinents
- représentation concrète de la notion abstraite de contexte thématique

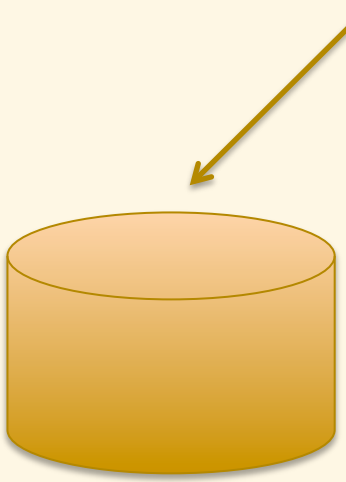
extraction de mots ou multi-mots

- synonymes ou concepts liés
- complément ajouté à la requête [12,21]

[12] V. Lavrenko et W. B. Croft. **Relevance Based Language Models.** In *Proc. of SIGIR'01.*
[21] D. Metzler et W. B. Croft. **Latent Concept Expansion Using Markov Random Fields.** In *Proc. of SIGIR'07*.

Motivation

hubble telescope achievements



collection de documents indexée

Hubble's Top Achievements - NASA
www.nasa.gov/externalflash/**hubble**_ga
Hubble's Top **Achievements** ... The **Hubb**
background of Earth after a week of repairs

Most Amazing Hubble Space Teles
www.space.com/17-amazing-**hubble**-dis
21 Apr 2011 - **The Hubble Space Telescop**
time it ... Here is a short rundown of **Hubble**

.
.
.

A Brief History of the Hubble Spa
content.time.com/time/photogallery/0,
... TIME Covers · Brief history of **the hubble space telescope** NASA ... at the cosmos since 1990. A look back at the **telescope's** remarkable life and **achievements**.

documents pseudo-pertinents

w	$P(w\|\hat{\theta}_Q)$
telescope	0.0567695577
space	0.0419802250
hubble	0.0380013632
shuttle	0.0217168275
light	0.0217168275
universe	0.0208438799
NASA	0.0195159071
mirror	0.0190680336
earth	0.0172484003
ultraviolet	0.0158515806

Modèles de pertinence

enrichissement du modèle de la requête

$$P(w|\theta_Q) = \lambda P(w|\tilde{\theta}_Q) + (1-\lambda) P(w|\hat{\theta}_Q)$$

modèle de la requête originale

modèle estimé

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) \prod_{t \in Q} P(t|\theta_D)$$

documents pseudo-pertinents

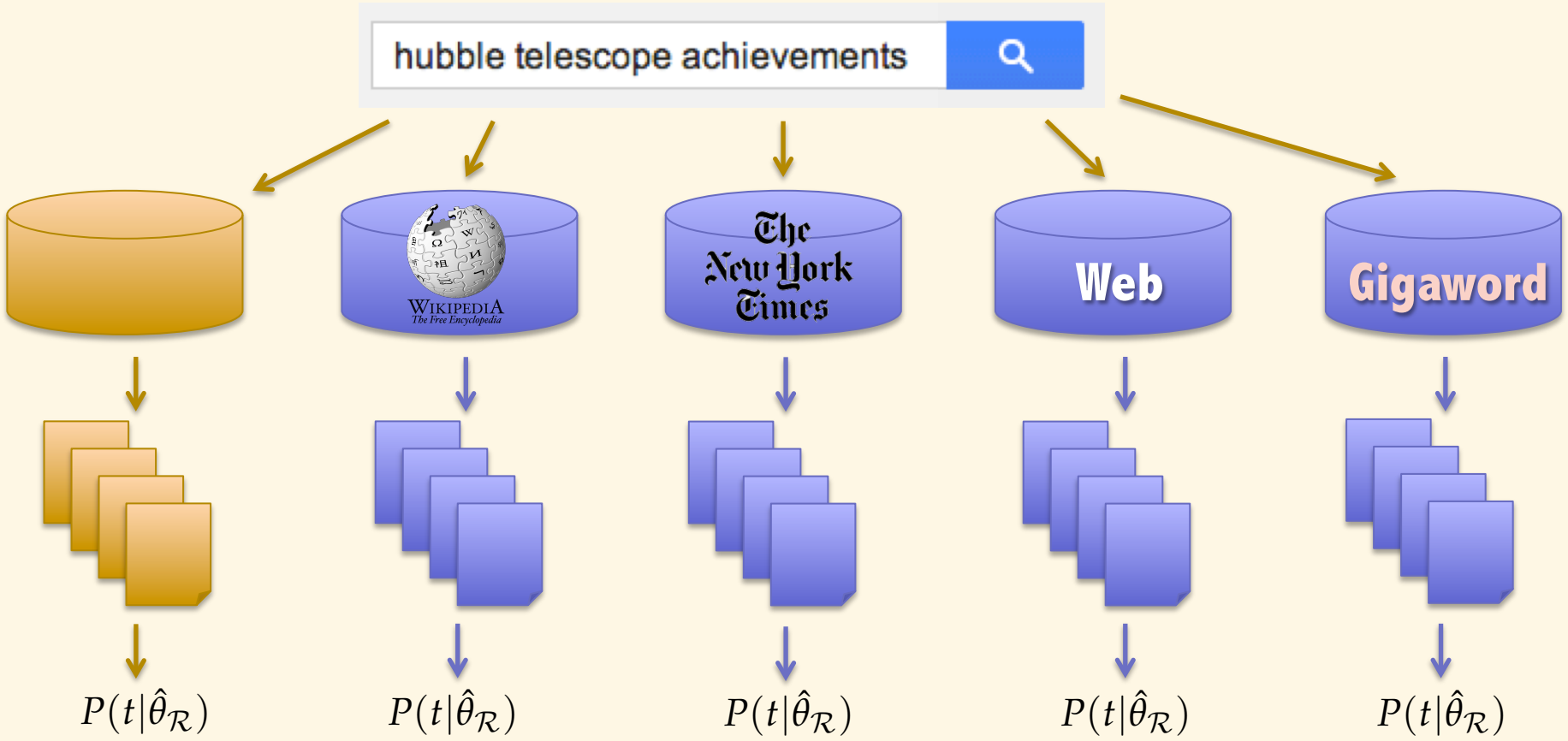
modèle du document

vraisemblance de la requête

connu dans la littérature sous le nom de RM3

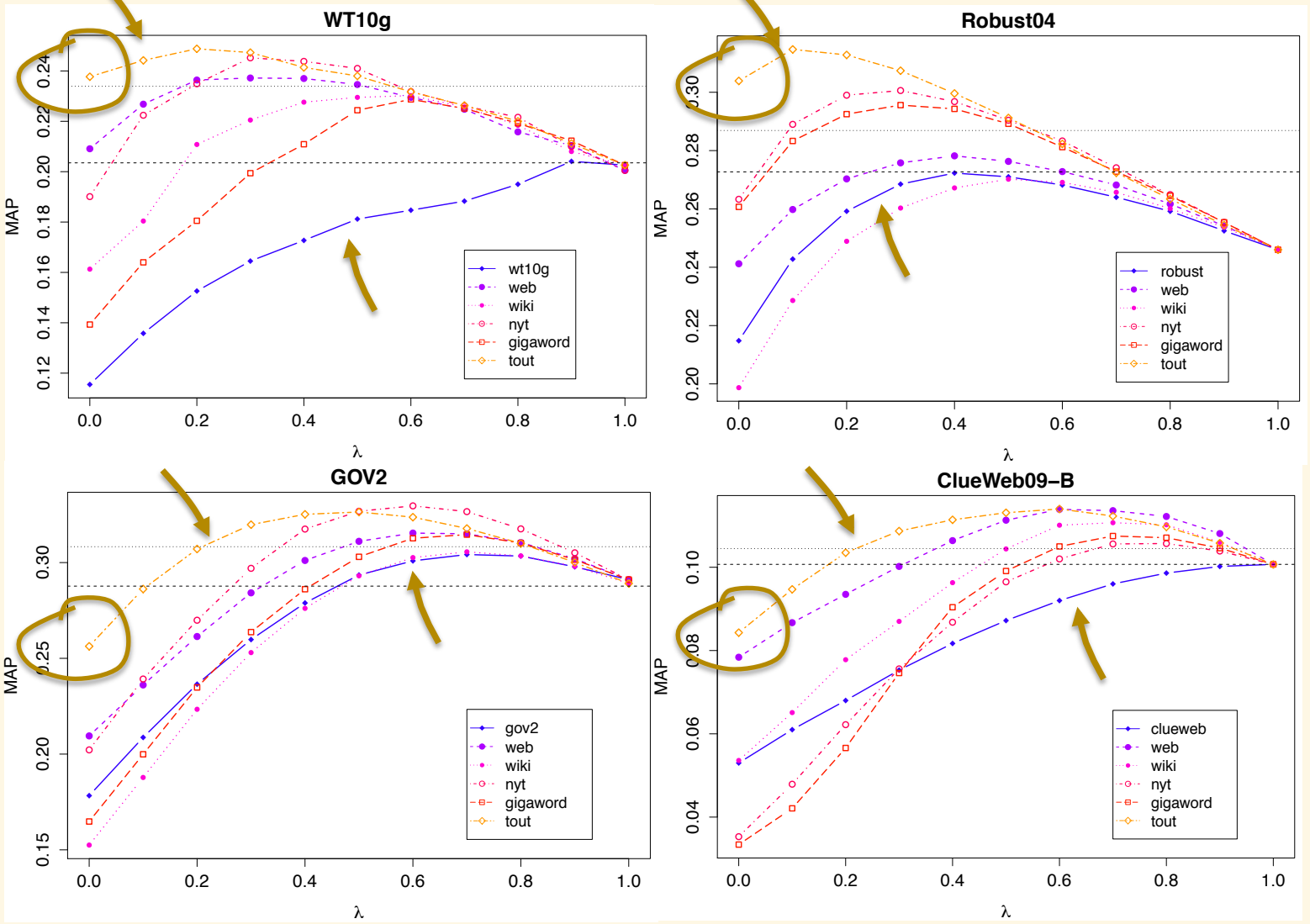
connaissances *a priori* sur le document (constantes dans notre modèle)

Divergence à partir de sources d'information



$$s(Q,D) = \lambda \log P(Q|\theta_D) - (1-\lambda) \sum_{\mathcal{R}\in\mathcal{S}} \varphi_{\mathcal{R}} \cdot KL(\hat{\theta}_{\mathcal{R}}||\theta_D)$$

Expérimentations



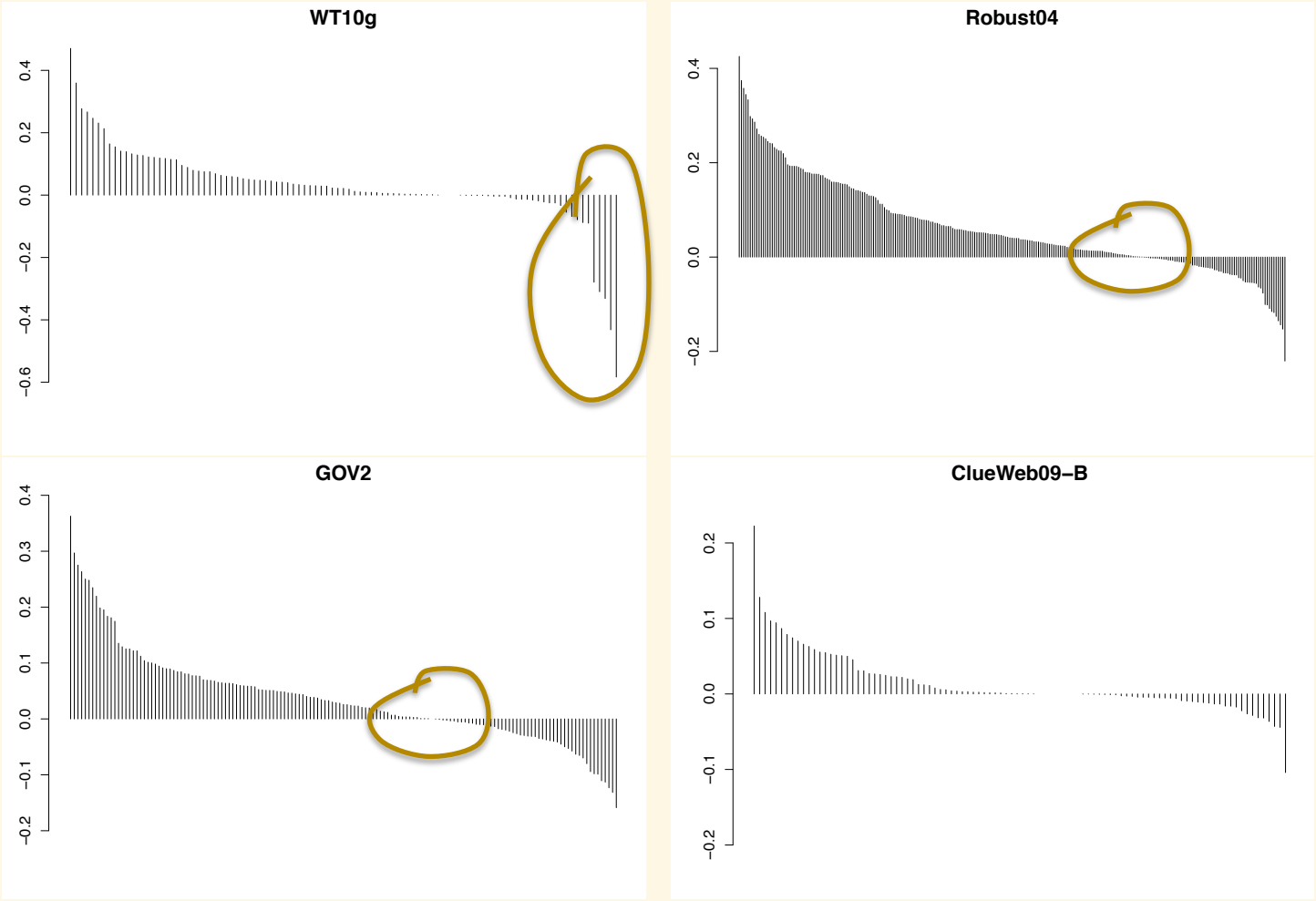
Expérimentations

enrichir la requête est peu robuste

- amélioration de certaines requêtes et dégradation d'autres
- glissement thématique (*topic drift*)
- évaluation de la robustesse de l'enrichissement par rapport à une approche sans enrichissement : QL

$$\Delta_{AP}(Q) = AP_{DfRes}(Q) - AP_{QL}(Q)$$

Expérimentations



Conclusions

estimation efficace du contexte thématique

- plusieurs sources d'information
- diversité thématique, couverture accrue

pistes d'améliorations de la robustesse

- apprentissage du paramètre λ
- approches sensibles au risque (*risk-sensitive*)

traitement des mots et multi-mots redondants [22]

[22] G. Cao, J.-Y. Nie, J. Gao, et S. Robertson. **Selecting Good Expansion Terms for Pseudo-relevance Feedback.** In *Proc. of SIGIR'08*.

Modélisation des concepts implicites d'une requête

estimation du nombre de concepts d'une requête

estimation du nombre de documents pseudo-pertinents

expériences, analyse

Introduction

objectif : représentation des concepts sous-jacents d'une requête

- affranchissement des ontologies et taxonomies
- détection non-supervisée

comment représenter un concept?

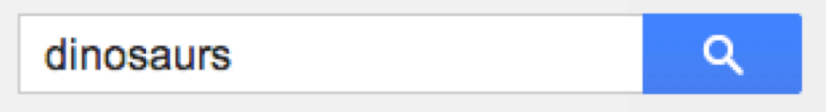
- classe contenant des objets possédants certaines propriétés et attributs [23]

utilisation des documents pseudo-pertinents

- porteurs d'informations contextuelles et conceptuelles par rapport à la requête
- aucun paramétrage
- estimation des nombres de concepts et de documents

[23] W. G. Stock. **Concepts and semantic relations in information science.** In *JASIST (2010)*.

Exemple



0.434 ➡ **birds**	0.291 ➡ **paleontology**	0.254 ➡ **comic**	0.021 ➡ **toys**
0.196 ➡ feathers	0.175 ➡ dinosaur	0.257 ➡ dinosaur	0.370 ➡ dinosaur
0.130 ➡ birds	0.125 ➡ kenya	0.180 ➡ devil	0.165 ➡ party
0.112 ➡ evolved	0.122 ➡ years	0.095 ➡ moon-boy	0.112 ➡ price
0.102 ➡ flight	0.087 ➡ fossils	0.054 ➡ bakker	0.053 ➡ birthday
0.093 ➡ dinosaurs	0.082 ➡ paleontology	0.054 ➡ world	0.039 ➡ game
0.084 ➡ protopteryx	0.070 ➡ discovery	0.045 ➡ marvel	0.023 ➡ toys

Quantification et identification de concepts implicites

modélisation thématique *sur* les documents pseudo-pertinents

- à l'opposé de la collection entière
- allocation de Dirichlet latente (LDA) [24]
- modélisation des thèmes latents : les concepts
- modèle conceptuel

deux distributions de probabilités

- thèmes sur les documents $P_{TM}(k|D, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)$
- mots sur les thèmes $P_{TM}(w|k, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)$

[24] D. M. Blei, A. Y. Ng, et M. I. Jordan. **Latent Dirichlet Allocation.** In *Journal of Machine Learning Research (2003)*.

Estimation du nombre de concepts

le nombre de concepts est un paramètre de LDA

- mais différentes requêtes font référence à des nombres de concepts différents
- requête spécifique vs. ambiguë

estimation automatique

- construction de plusieurs modèles thématiques avec différents nombre de concepts
- choix du modèle pour lequel les concepts sont les mieux délimités

Estimation du nombre de concepts

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{i=1}^{K} \sum_{j=i+1}^{K} D(k_i||k_j)$$

divergence de Kullback-Leibler

maximisation de la divergence moyenne entre toutes les paires de concepts

$$D(k_i||k_j) = \sum_{w \in \mathcal{W}_{inter}} P_{TM}(w|k_i, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K) \log \frac{P_{TM}(w|k_i, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)}{P_{TM}(w|k_j, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)} +$$

$$\sum_{w \in \mathcal{W}_{inter}} P_{TM}(w|k_j, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K) \log \frac{P_{TM}(w|k_j, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)}{P_{TM}(w|k_i, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)}$$

Combien de documents pseudo-pertinents?

problème récurrent en RI

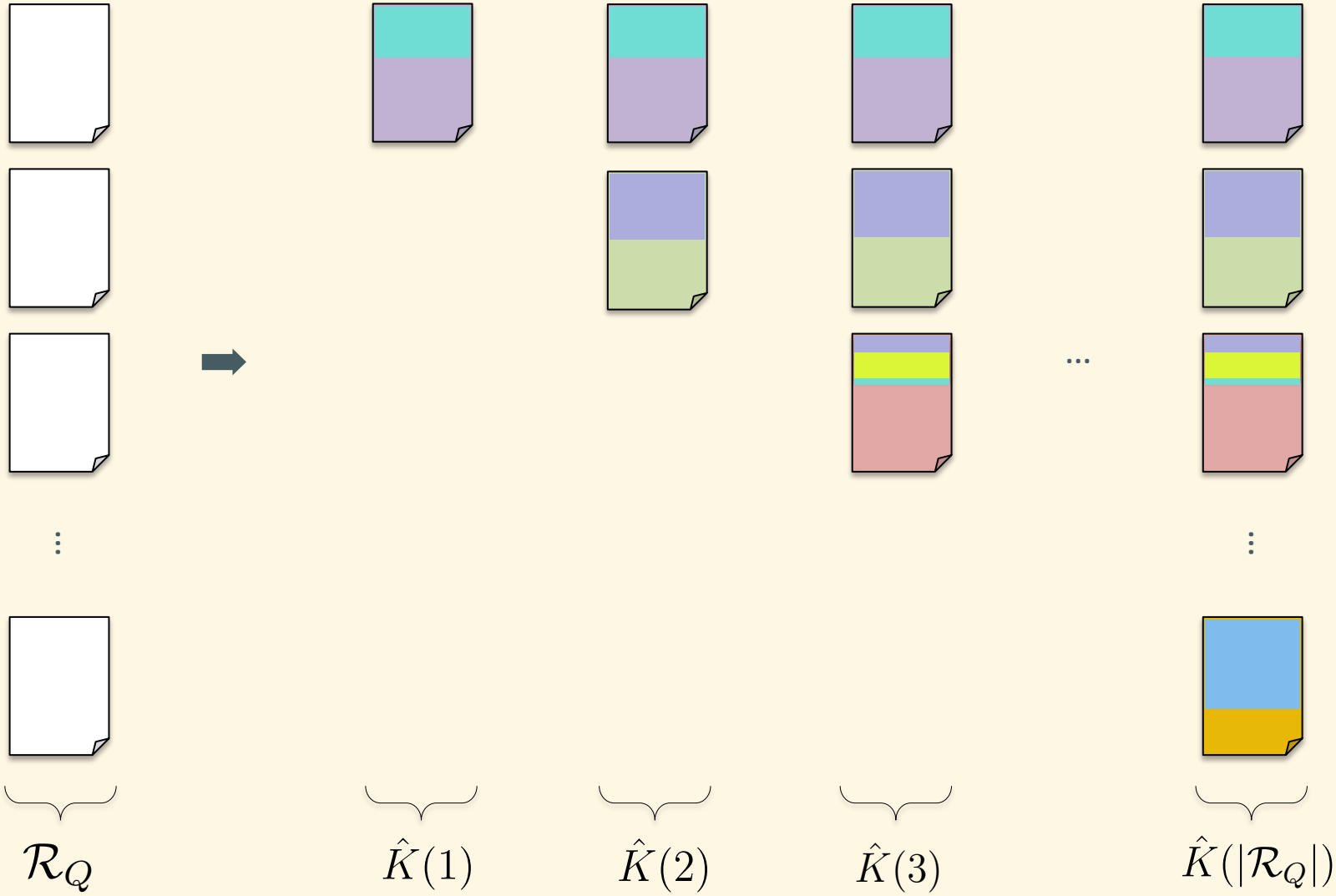
- nombres de documents pertinents varient selon les requêtes
- modèles génératifs [25] et discriminatifs [21]

approche au niveau conceptuel

- contrairement à une approche centrée document
- « quel modèle conceptuel contextualise le mieux ma requête? »

[21] B.He et I.Ounis. **Finding Good Feedback Documents.** In *Proc. of CIKM'09.*
[25] T. Tao et C. Zhai. **Regularized Estimation of Mixture Models for Robust Pseudo-relevance Feedback.** In *Proc. of SIGIR'06.*

Combien de documents pseudo-pertinents?



Combien de documents pseudo-pertinents?

« centroïde » des modèles conceptuels

- éviter les modèles bruités
- modèles appris dans des espaces différents
- pas de divergence probabiliste
- similarité textuelle sur les mots des concepts

$$sim(\mathcal{T}_{\Theta_m}^{\hat{K}(m)}, \mathcal{T}_{\Theta_n}^{\hat{K}(n)}) = \frac{1}{\eta} \sum_{k \in \mathcal{T}_{\Theta_m}^{\hat{K}(m)}} \sum_{k' \in \mathcal{T}_{\Theta_n}^{\hat{K}(n)}} \frac{|k \cap k'|}{\hat{K}(m)} \sum_{w \in k} \log \frac{N}{df_w}$$

$$\hat{M} = \operatorname{argmax}_m \sum_n sim(\mathcal{T}_{\Theta_m}, \mathcal{T}_{\Theta_n})$$

facteur de normalisation

recouvrement (en mots) entre deux concepts appartenant à deux modèles conceptuels différents

IDF dans la collection cible

Expériences

le nombre estimé de concepts est-il correct?

- pas de référence disponible
- annotation manuelle très subjective

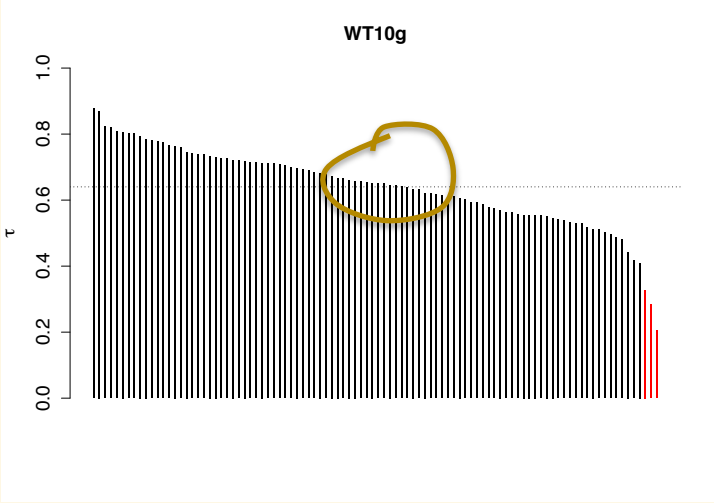
corrélation avec une modélisation thématique hiérarchique (HDP) [26]

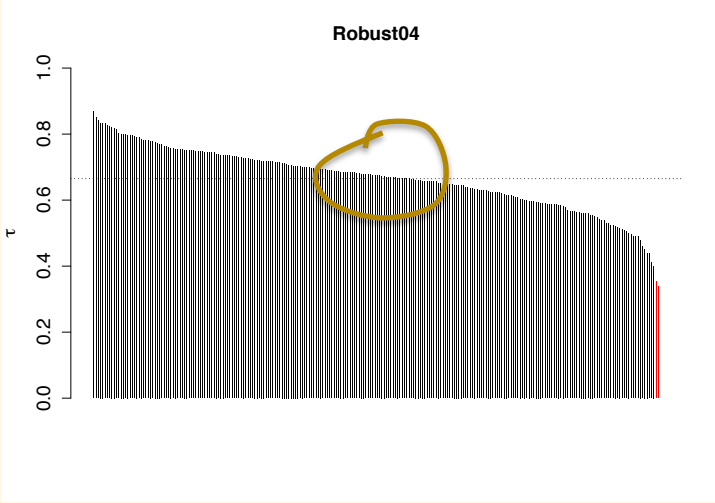
- généralisation de LDA
- non-paramétrique... ou presque

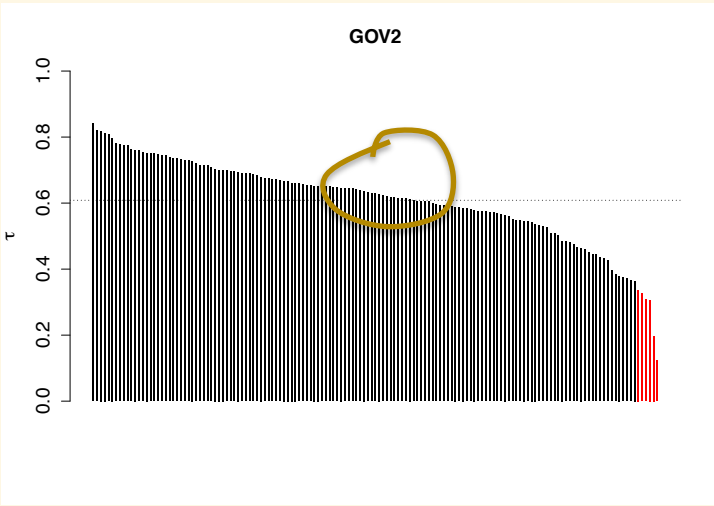
[26] Y. W. Teh, M. I. Jordan, M. J. Beal, et D. M. Blei. **Hierarchical Dirichlet Processes.** In *Journal of the American Statistical Association (2006)*.

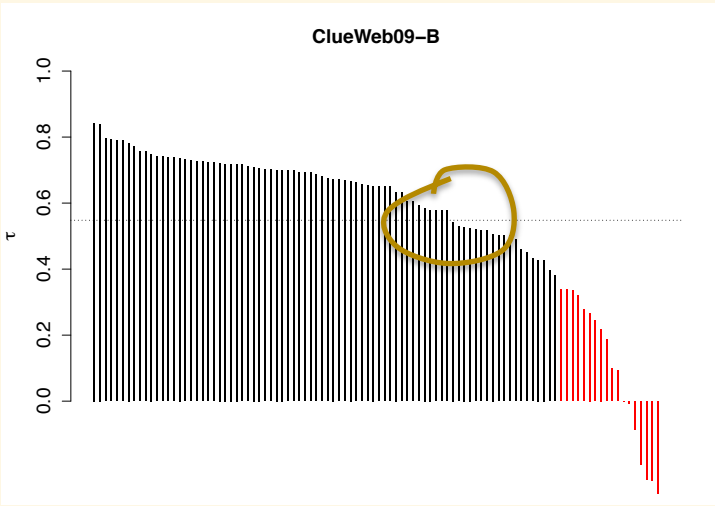
Expériences

taux de corrélation de Kendall









Expériences

mais... les concepts sont-ils significatifs?

apportent-ils vraiment de l'information?

sont-ils cohérents?

cohérence sémantique des modèles thématiques

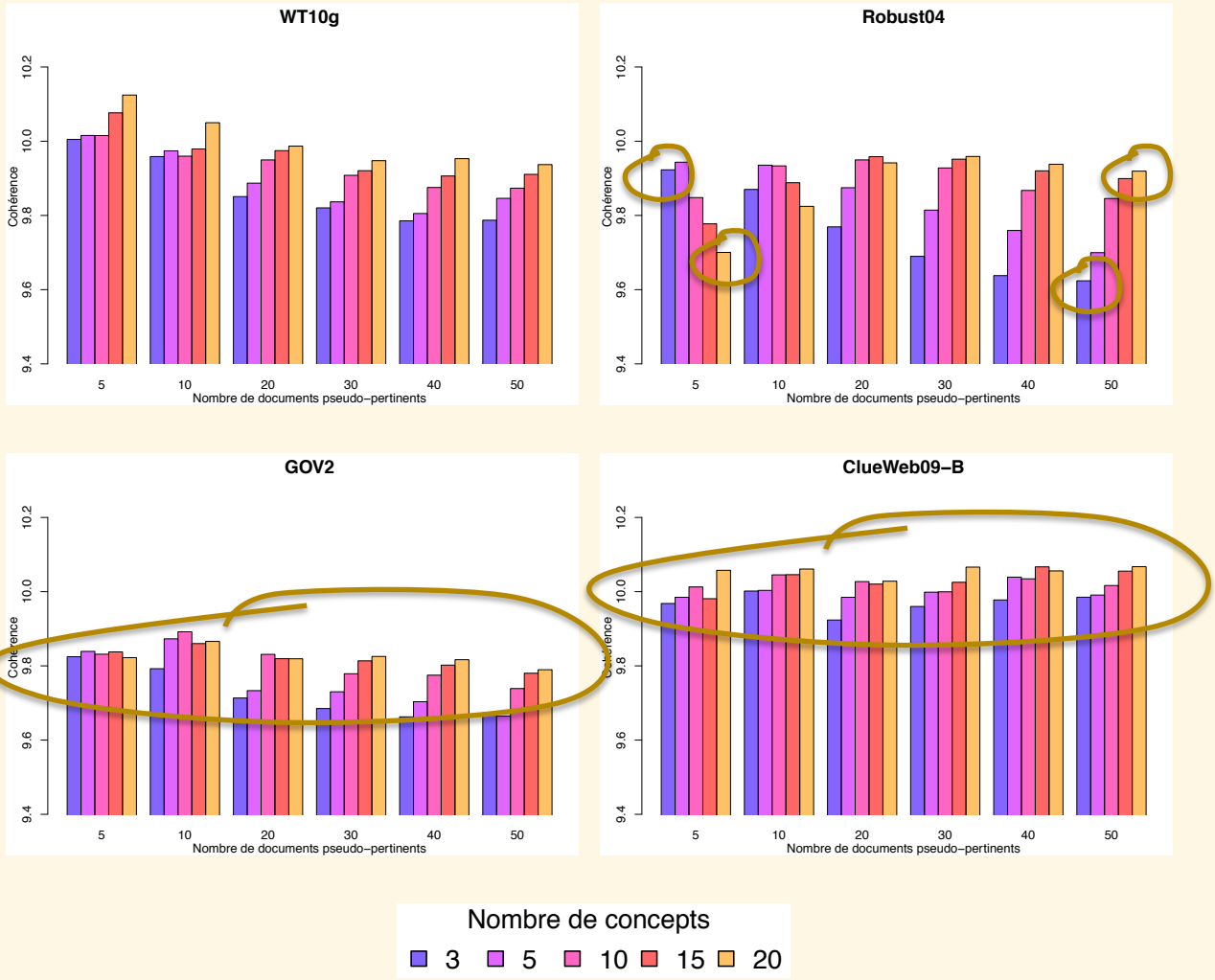
mesure les cooccurrences des mots des concepts [27]

$$c(\mathcal{T}_\Theta^K) = \frac{1}{K}\sum_{i=1}^{K}\sum_{(w,w')\in k_i}\log\frac{P(w,w')+\epsilon}{P(w)P(w')}$$

information mutuelle ponctuelle (PMI)

[27] K. Stevens, P. Kegelmeyer, D. Andrzejewski, et D. Buttler. **Exploring Topic Coherence over Many Models and Many Topics.** In *Proc. of EMNLP'12*.

Expériences



Conclusions

modélisation des concepts implicites

- automatique et non supervisée
- concepts sémantiquement cohérents

quel est leur effet sur les performances d'un système de RI?

- intégration dans des nouveaux modèles de pertinence

Modèles de pertinence conceptuels

modèle thématique de la requête

modèles de pertinence conceptuels adaptatifs

combinaison de modèles

expérimentation, résultats, discussion

Motivation

modélisation thématique et documents pseudo-pertinents pour la RI [28,29,30,31]

- efficaces pour enrichir la requête
- modélisation *a priori* de la collection entière

triple contribution

- modèles de pertinence conceptuels (TDRM)
- modèles de pertinence conceptuels adaptatifs (ATDRM)
- combinaison de ATDRMs

[28] L. A. Park et K. Ramamohanarao. **The Sensitivity of Latent Dirichlet Allocation for Information Retrieval.** In *Proc. of ECML PKDD'09*.
[29] X. Yi et J. Allan. **A Comparative Study of Utilizing Topic Models for Information Retrieval.** In *Proc. of ECIR'09*.
[30] D. Andrzejewski et D. Buttler. **Latent Topic Feedback for Information Retrieval.** In *Proc. of KDD'11.*
[31] Y. Lu, Q. Mei, et C. Zhai. **Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA.** In *Information Retrieval (2011)*.

Topic-Driven Relevance Models

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) \qquad \prod_{t \in Q} P(t|\theta_D)$$

marginalisation sur l'espace de concepts

$$P_{TM}(w|k, D, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K) = \sum_{k \in \mathcal{T}_\Theta^K} P'_{TM}(w|k, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K) \cdot P_{TM}(k|D, \boldsymbol{\theta}_M, \boldsymbol{\phi}_K)$$

probabilité que le mot w appartienne au concept k

probabilité que le concept k apparaisse dans le document D

Adaptive Topic-Driven Relevance Models

méthodes de la contribution précédente

estimation du « meilleur » modèle conceptuel

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) P_{TM}(w|k, D, \boldsymbol{\theta}_{\hat{M}}, \boldsymbol{\phi}_{\hat{K}}) \prod_{t \in Q} P(t|\theta_D)$$

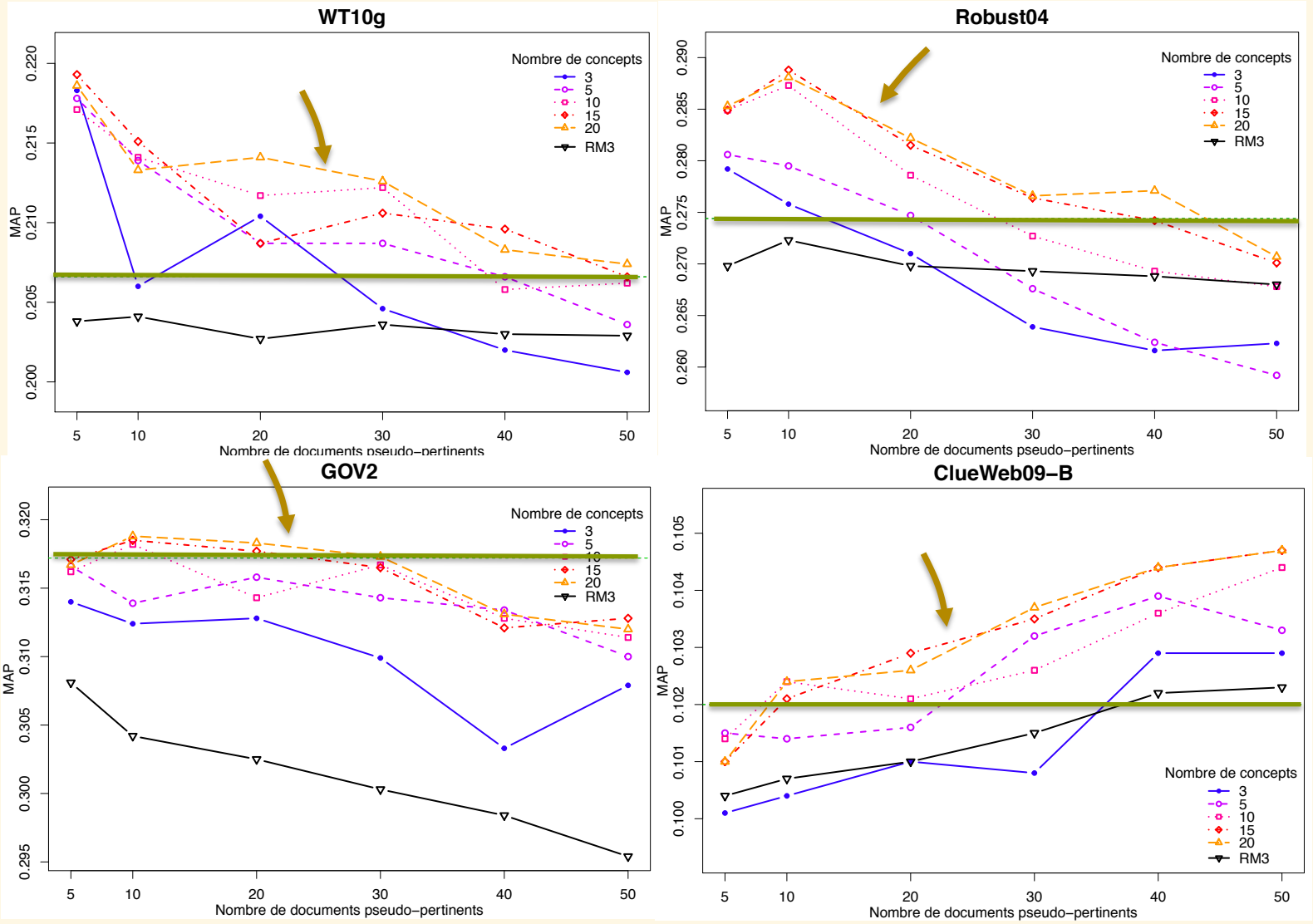
Combinaison de ATDRMs

plusieurs sources d'information

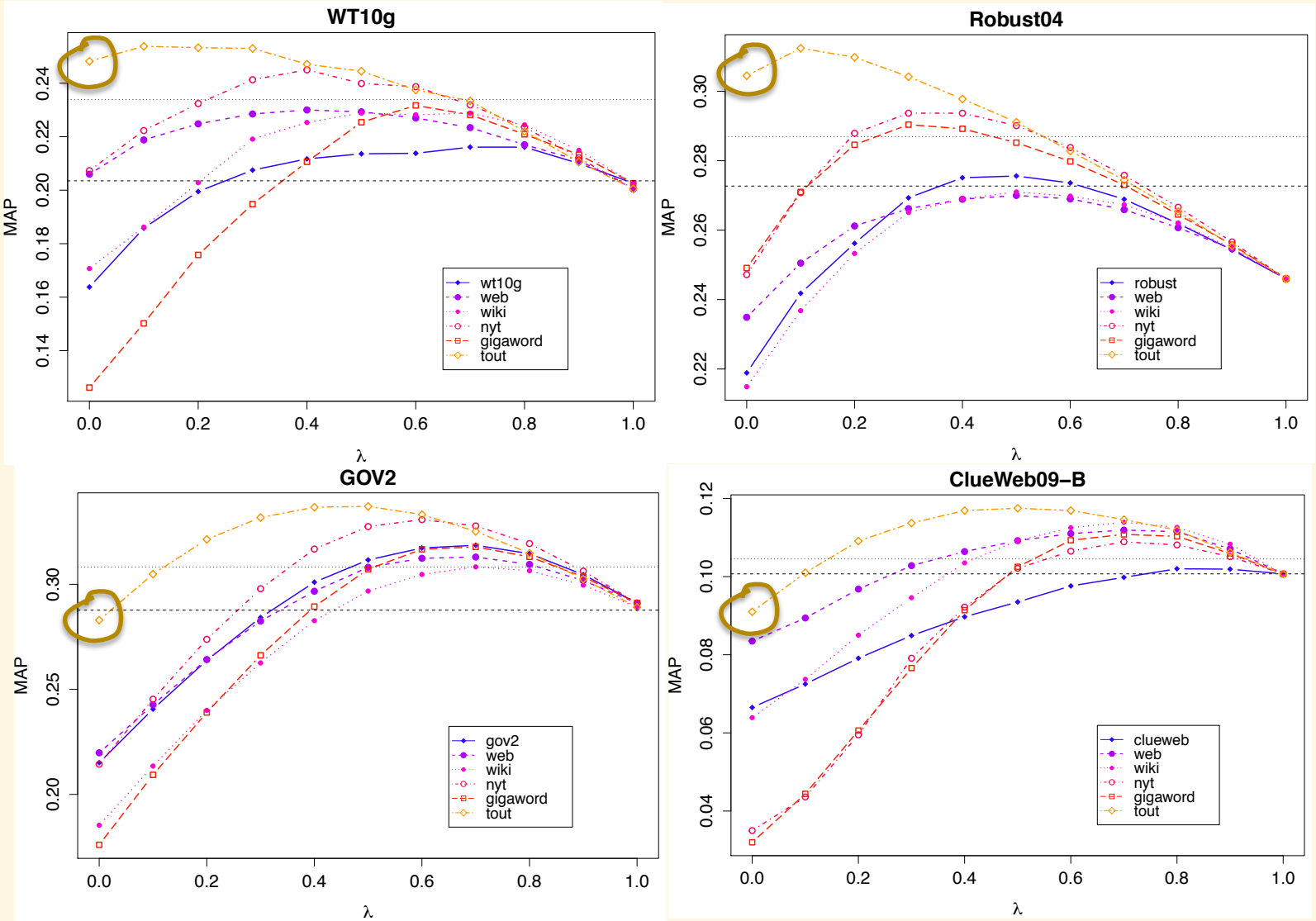
vérification de nos observations du premier chapitre

$$P(w|\hat{\theta}_Q) \propto \sum_{\theta_D \in \Theta} P(\theta_D) P(w|\theta_D) P_{TM}(w|k, D, \boldsymbol{\theta}_{\hat{M}}, \boldsymbol{\phi}_{\hat{K}}) \prod_{t \in Q} P(t|\theta_D)$$

Expérimentations (TDRM + ATDRM)



Expérimentations (MoATDRM)



Conclusions

nouveau modèle de pertinence

- basé sur une modélisation thématique
- efficace et précis

ouvert à de nouvelles extensions

- modèle intégrant directement la cohérence sémantique
- autres algorithmes de modélisation thématique probabiliste (pLSA, HDP, ...)

Conclusions

et perspectives...

Résultats

« est-il possible de représenter de façon entièrement automatique les thématiques liées au besoin d'information d'un utilisateur, exprimé uniquement par une requête ? »

- 2 variantes des modèles de pertinence
- utilisation de 4 sources d'information externes
- modélisation thématique sur des documents pseudo-pertinents

Résultats

combinaison de sources d'information

- meilleure couverture de la requête
- meilleure contextualisation, quel que soit le scénario

identification de concepts implicites

- non supervisée, non dépendante d'un type de données
- précise et produisant des concepts très cohérents

évolution des modèles de pertinence

- tirant parti des précédentes contributions
- très efficaces et robustes
- représentation précise du contexte thématique

Perspectives

estimation du risque associé à l'enrichissement

- stratégie de repli, de pondération...
- estimation jointe des différents paramètres, *deep learning*...

modèles conceptuels comme *feedback*

- guider l'utilisateur dans sa recherche
- raffinage des résultats, exploration conceptuelle...

application à la prédiction de difficulté des requêtes

- suivant la cohérence des concepts modélisés

merci de votre attention

et encore un grand merci à tous les membres du jury

Mme Josiane MOTHE
M. Jian-Yun NIE
M. Philippe MULHEM
M. Jacques SAVOY
M. Jaap KAMPS
M. Benjamin PIWOWARSKI
M. Eric SANJUAN
M. Patrice BELLOT

Contexte thématique

retour de pertinence [11]

- idée : impliquer directement l'utilisateur
- processus long, peu compatible avec les standard actuels
- estimation automatique

retour de pertinence dit « *simulé* »

- hypothèse : les N premiers documents renvoyés par le système de RI sont considérés comme pertinents
- représentation concrète de la notion abstraite de contexte thématique

[11] J. Koenemann et N. J. Belkin. **A Case for Interaction : A Study of Interactive Information Retrieval Behavior and Effectiveness.** In *Proc. of CHI'96.*

Mesures d'évaluation

précision et rappel

mesures classiques en classification

$P(\text{pertinent}|\text{renvoyé})$ et $P(\text{renvoyé}|\text{pertinent})$

$$AP = \frac{1}{|R|} \sum_{k=1}^{n} P@k \times rel(d_k)$$

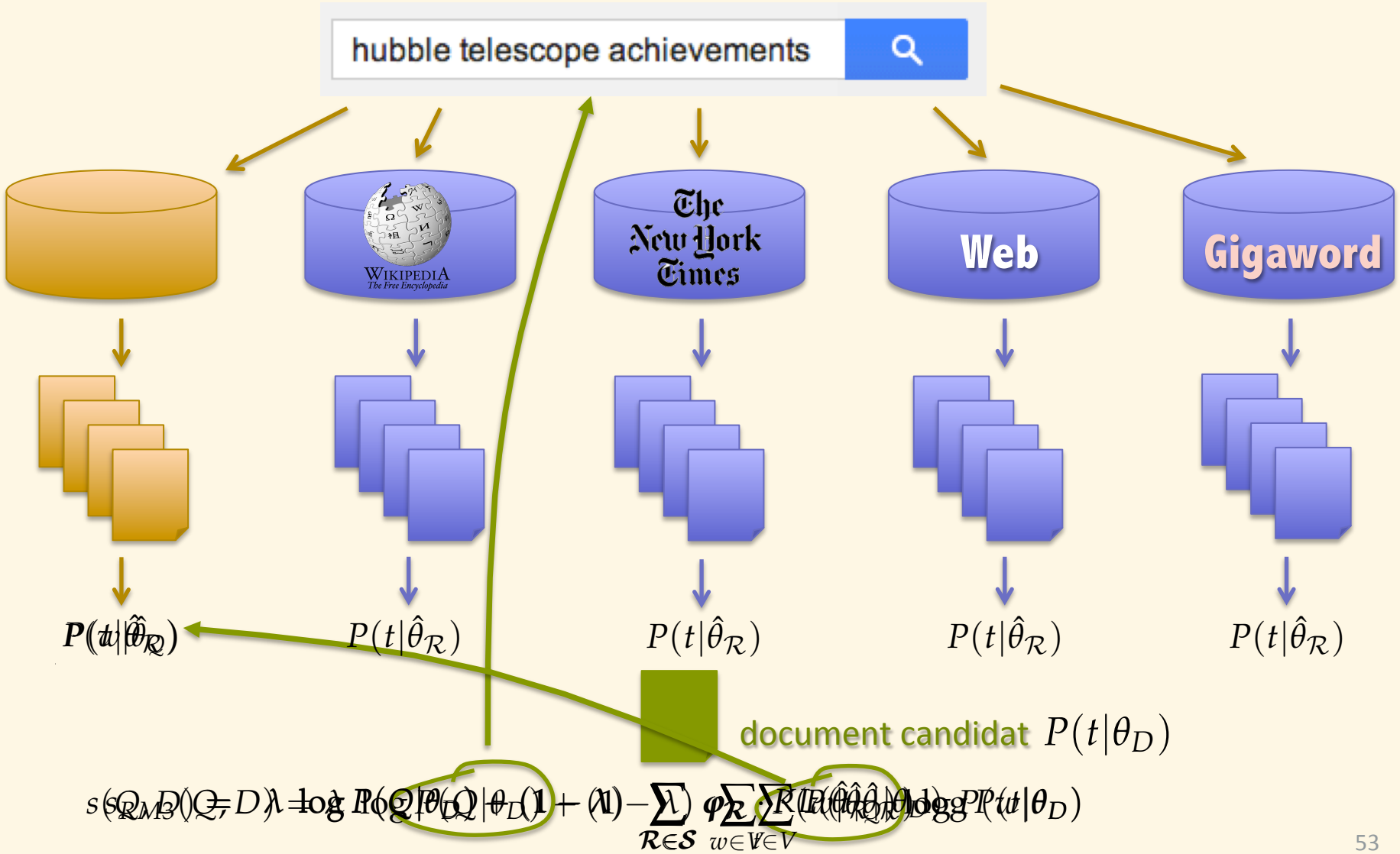
pertinence du k-ième document

nombre de documents pertinents

précision à k documents

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} AP(Q_i)$$

Divergence à partir de sources d'information



Divergence à partir de sources d'information

$$
\begin{aligned}
KL(\theta_{\mathcal{R}}||\theta_D) &= \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log \frac{P(t|\theta_{\mathcal{R}})}{P(t|\theta_D)} \\
&= \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_{\mathcal{R}}) - \sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_D) \\
&\propto -\sum_{t\in V} P(t|\theta_{\mathcal{R}}) \log P(t|\theta_D)
\end{aligned}
$$

$$
P(t|\hat{\theta}_{\mathcal{R}}) \propto \sum_{\theta_D\in\Theta} P(\theta_D) H_\Theta(t) \prod_{t\in Q} P(t|\theta_D)
$$

$$
H_\Theta(t) = -\sum_{w\in t} P(w|\Theta) \log P(w|\Theta)
$$

Divergence à partir de sources d'information

$$KL(\hat{\theta}_R||\theta_D) = -\sum_{t \in V} P(t|\hat{\theta}_{\mathcal{R}}) \log P(t|\theta_D)$$

$$\begin{aligned} s(Q,D) &= \lambda \log P(Q|\theta_D) - (1-\lambda) \sum_{\mathcal{R} \in \mathcal{S}} \varphi_{\mathcal{R}} \cdot KL(\hat{\theta}_{\mathcal{R}}||\theta_D) \\ &= \lambda \log P(Q|\theta_D) + (1-\lambda) \sum_{\mathcal{R} \in \mathcal{S}} \varphi_{\mathcal{R}} \sum_{t \in V} P(t|\hat{\theta}_{\mathcal{R}}) \log P(t|\theta_D) \end{aligned}$$

$$s_{RM3}(Q,D) = \lambda \log P(Q|\theta_D) + (1-\lambda) \sum_{w \in V} P(w|\hat{\theta}_Q) \log P(w|\theta_D)$$

Expérimentations (divergence)

recherche de documents, collections TREC

validation croisée pour λ; $N = 10$ et $k = 20$

utilisation de toutes les sources d'information

poids de chaque ressource

intuitivement: les chances que la ressource $\mathcal{R}$ (utilisée individuellement) soit meilleure que les autres ressources (utilisées individuellement)

$$\varphi_{\mathcal{R}} = \frac{\sum_{i=1}^{M} max_{MAP}(Q_i, \mathcal{R})}{M}$$

Expérimentations

	QL		RM3		MoRM		DfRes	
	MAP	P@20	MAP	P@20	MAP	P@20	MAP	P@20
WT10g	0,2026	0,2429	0,2035	0,2449	$0,2339^{\alpha,\beta}$	$0,2833^{\alpha,\beta}$	$0,2463^{\alpha,\beta}$	$0,2954^{\alpha,\beta}$
Robust04	0,2461	0,3528	$0,2727^{\alpha}$	0,3677	$0,2869^{\alpha,\beta}$	$0,3799^{\alpha,\beta}$	$0,3147^{\alpha,\beta,\gamma}$	$0,4024^{\alpha,\beta,\gamma}$
GOV2	0,2911	0,5145	0,2877	0,5074	$0,3083^{\alpha,\beta}$	$0,5409^{\alpha,\beta}$	$0,3257^{\alpha,\beta,\gamma}$	$0,5638^{\alpha,\beta,\gamma}$
ClueWeb09-B	0,1007	0,2347	0,1007	0,2260	0,1045	0,2250	$0,1140^{\alpha,\beta,\gamma}$	$0,2770^{\alpha,\beta,\gamma}$

systèmes état-de-l'art

modèle de pertinence RM3

combinaison de RM3 : MoRM [23]

combinaison très efficace

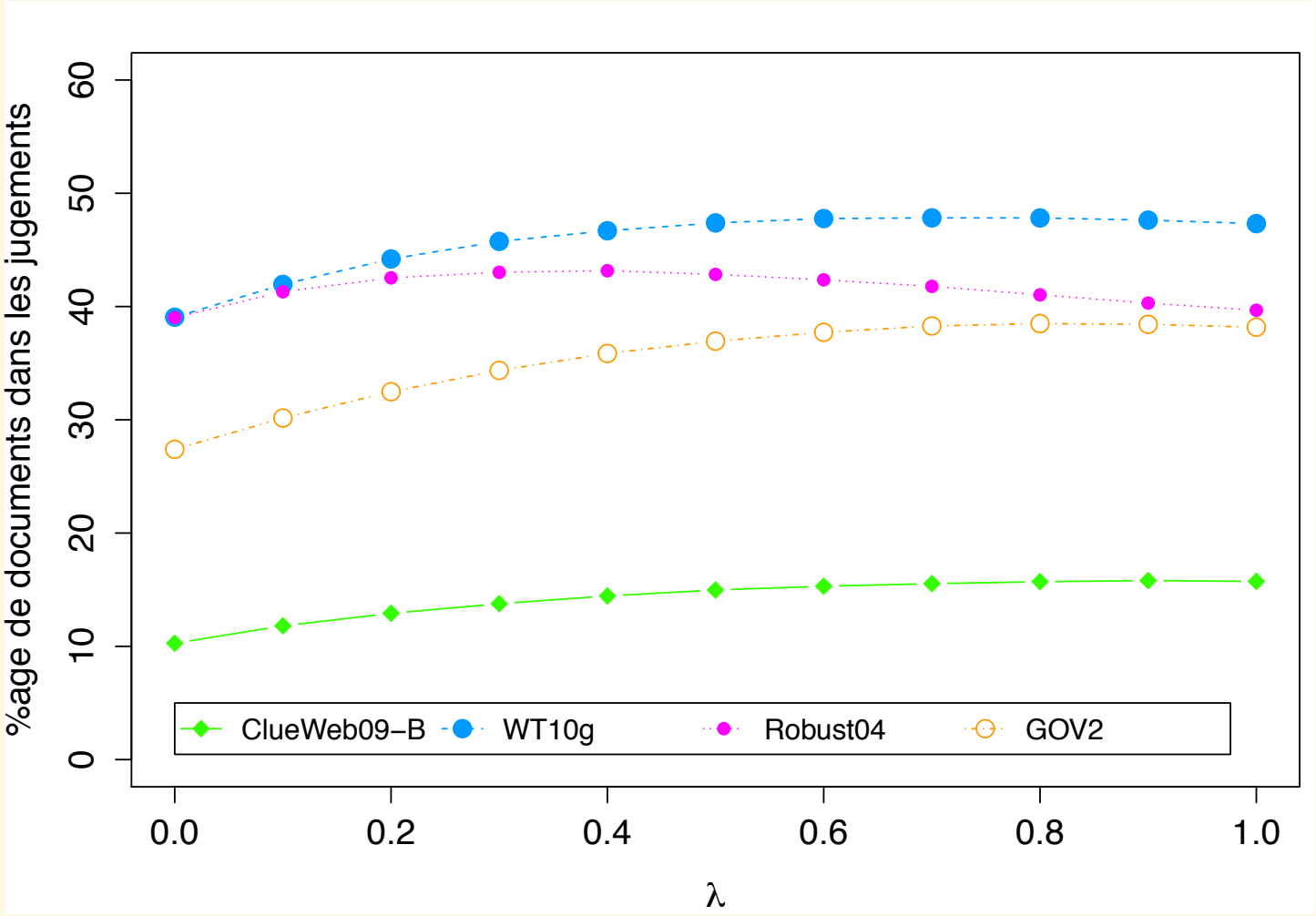
y compris là où MoRM échoue

[23] F. Diaz et D. Metzler. **Improving the Estimation of Relevance Models Using Large External Corpora.** In *Proc. of SIGIR'06.*

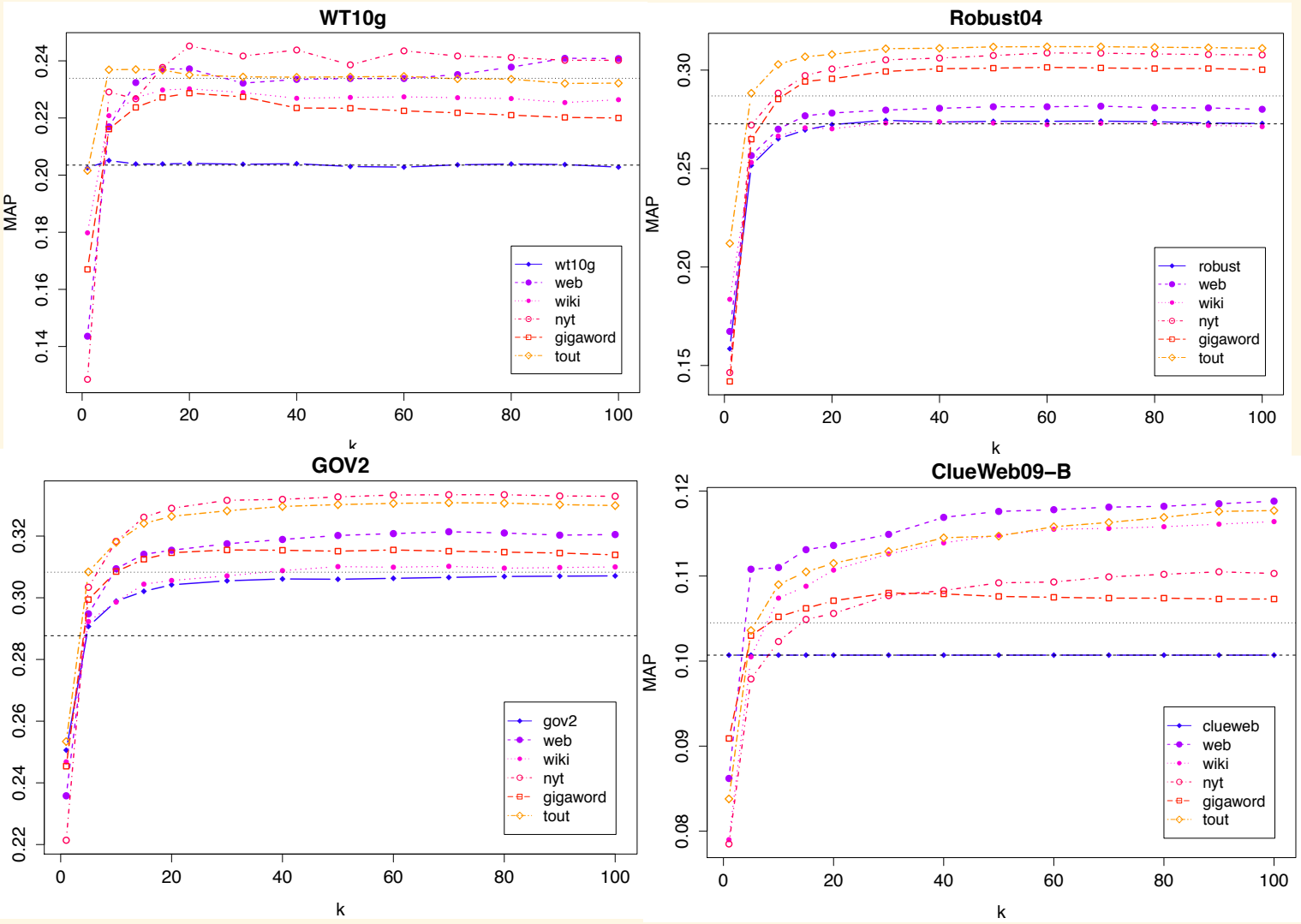
Influence des sources d'informations (DfRes)

	nyt	wiki	gigaword	web	WT10g	Robust04	GOV2	ClueWeb09-B
WT10g	0,303	0,162	0,121	**0,313**	0,101	-	-	-
Robust04	**0,309**	0,076	0,281	0,149	-	0,185	-	-
GOV2	0,213	0,121	0,179	0,219	-	-	**0,261**	-
ClueWeb09-B	0,195	0,215	0,127	**0,351**	-	-	-	0,108

Complétude des jugements de pertinence



Expérimentations



Introduction

comment représenter un concept?

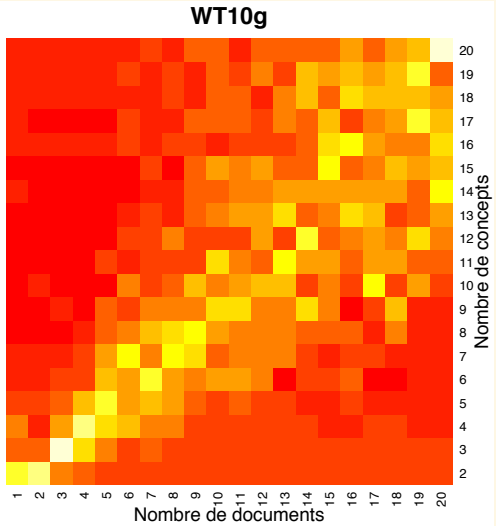
- classe contenant des objets possédants certaines propriétés et attributs [25]

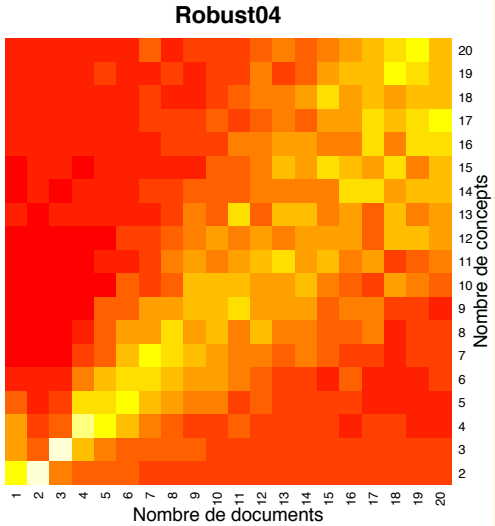
recherche par facettes (*Faceted Search*) [26]

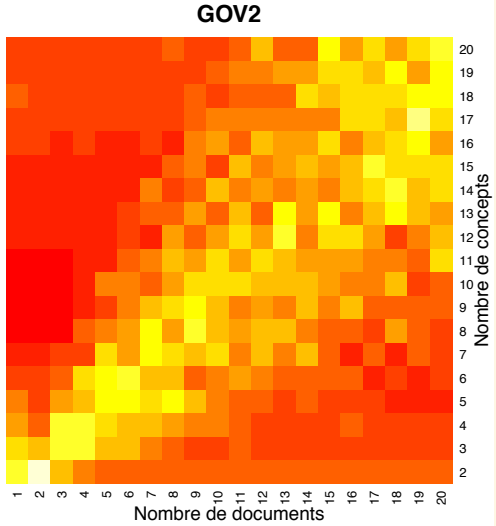
- nombre prédéfini de dimensions : taxonomie et apprentissage supervisé
- efficace en domaine fermé, le Web évolue vite !

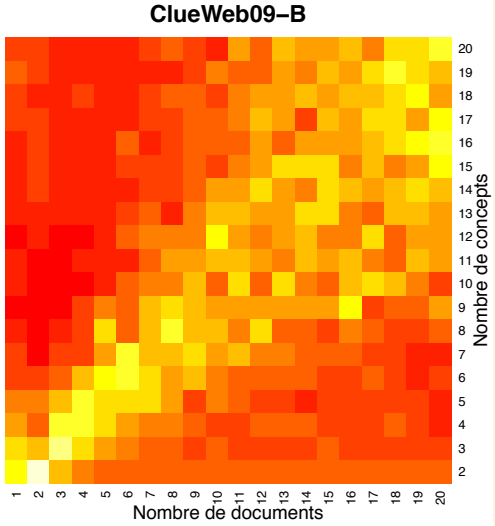
[25] W. G. Stock. **Concepts and semantic relations in information science.** In *JASIST (2010)*.
[26] D. Tunkelang. **Faceted search.** In *Synthesis Lectures on Information Concepts, Retrieval, and Services (2009)*.

Expériences

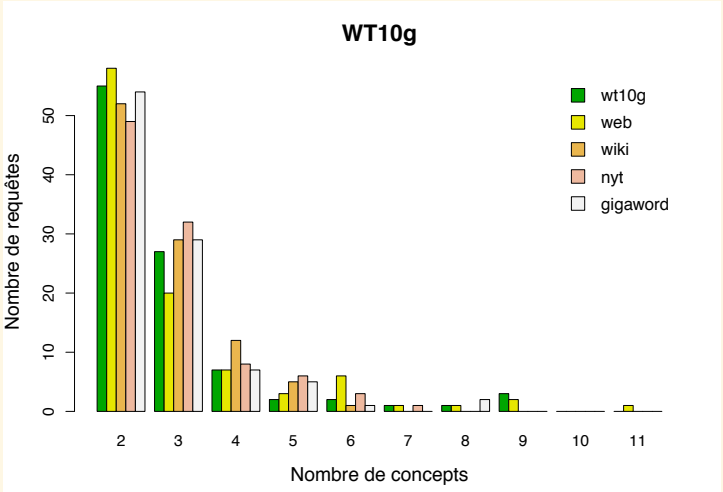


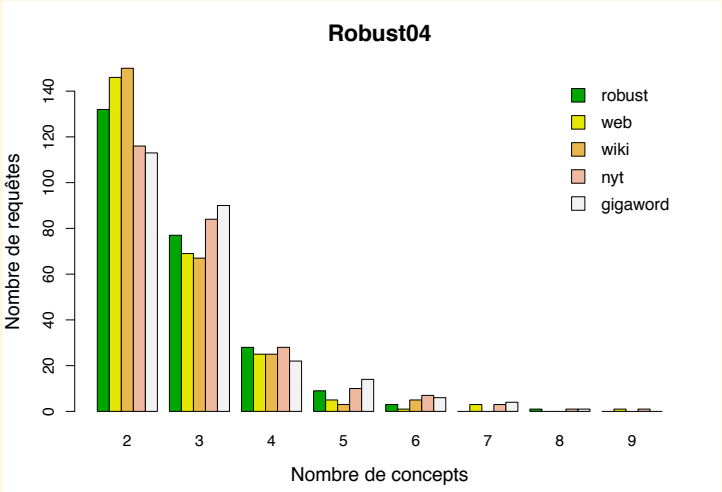


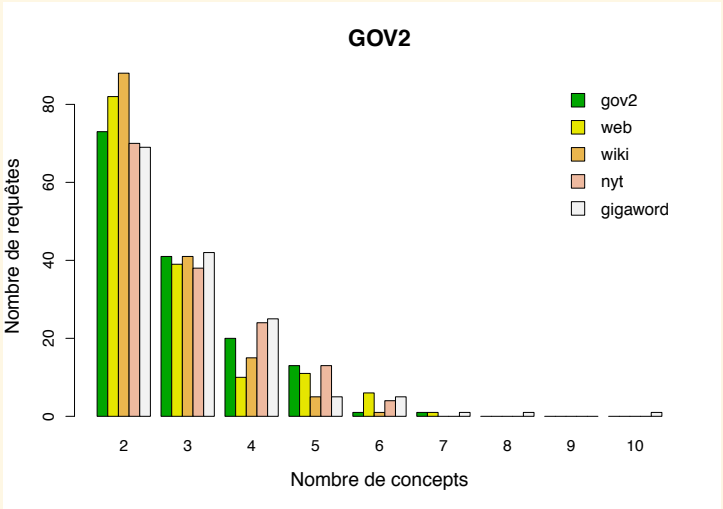


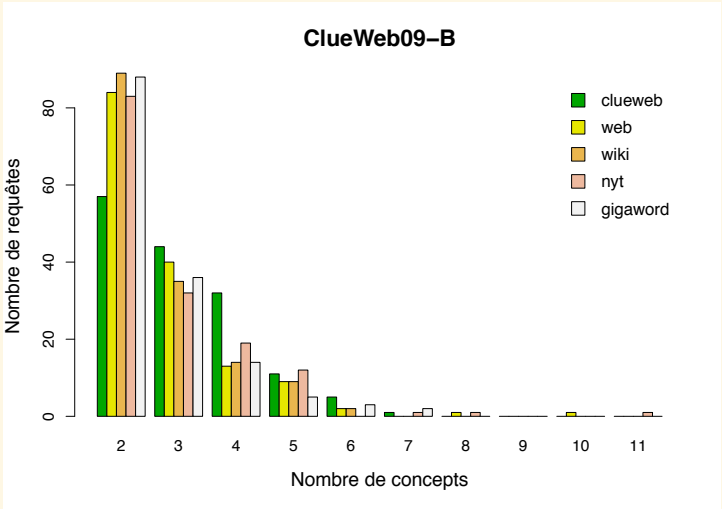


Expériences

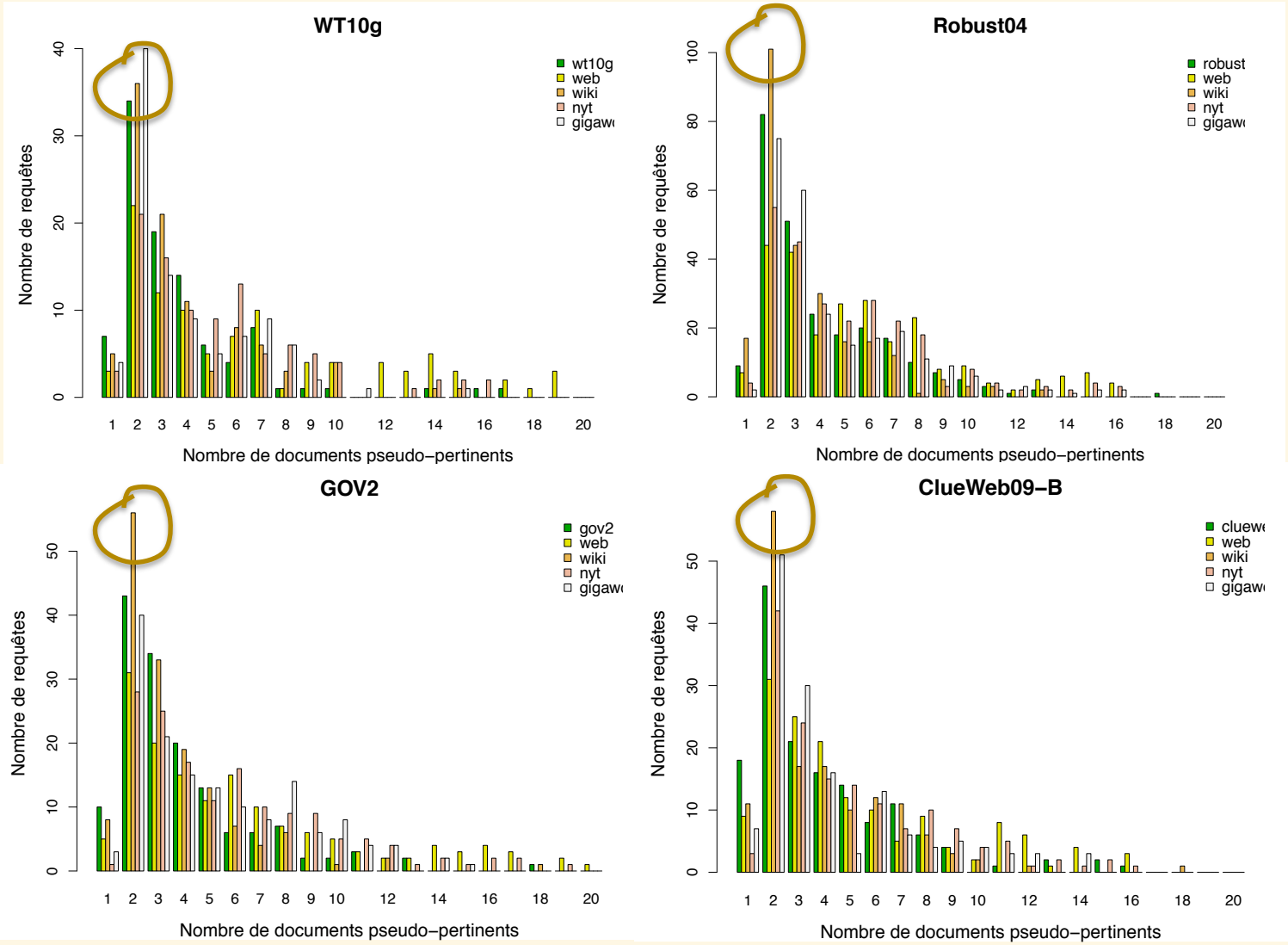




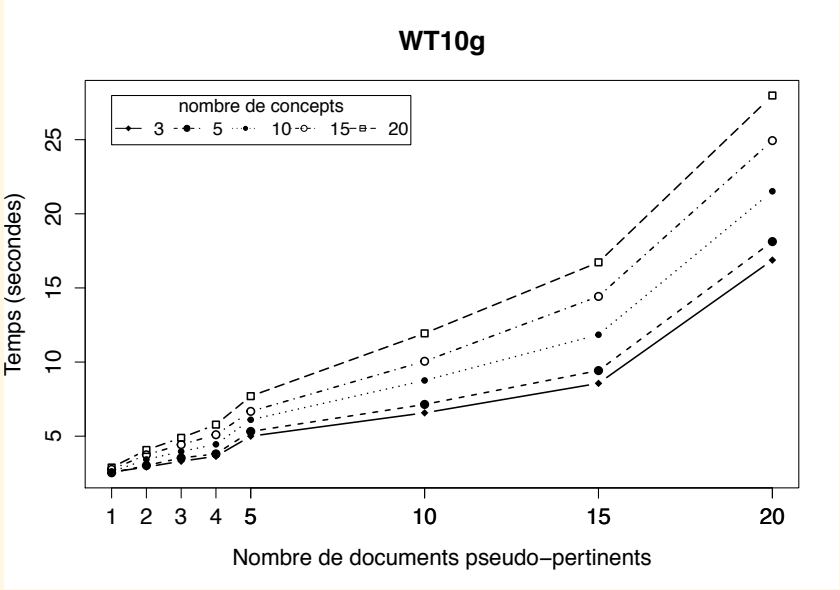


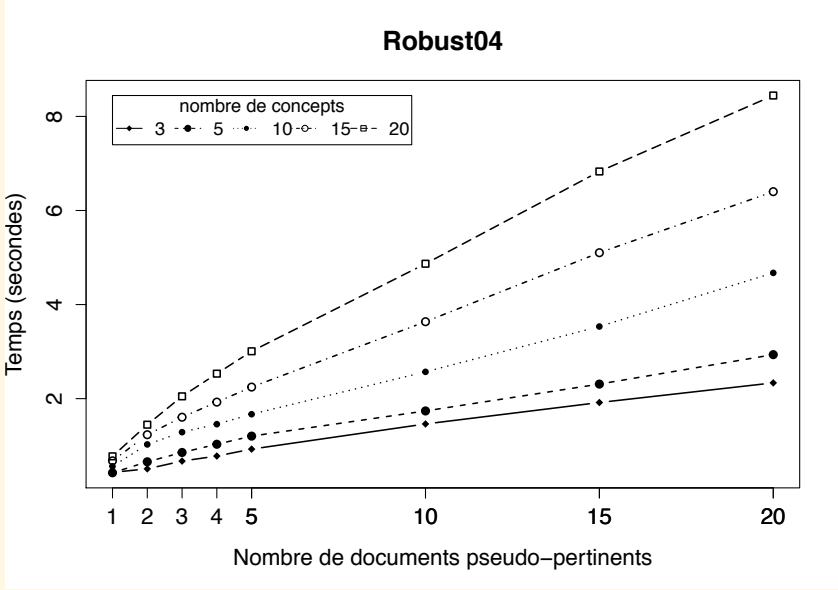


Expériences



Temps d'exécution





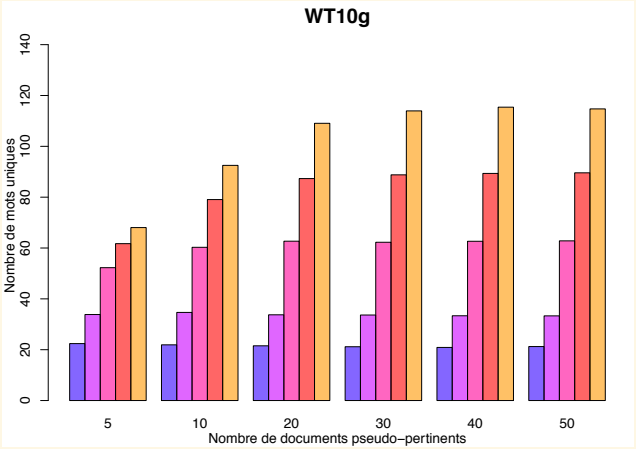
Distribution des mots sur les concepts

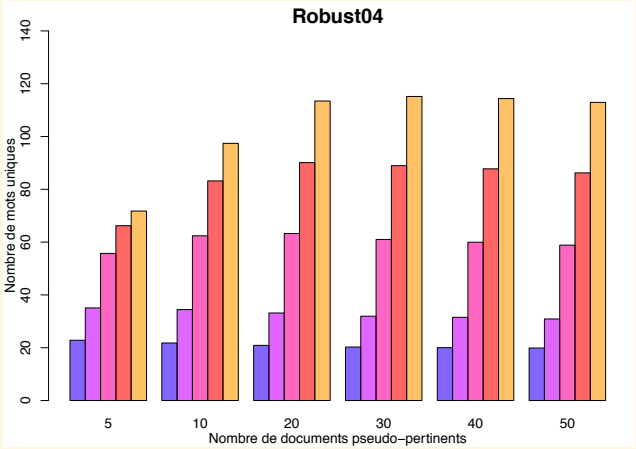
nous ne considérons que les 10 mots de plus fortes probabilités pour chaque concept

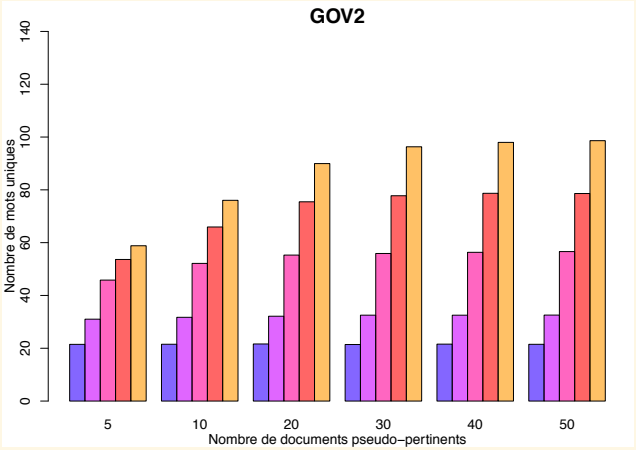
normalisation de la distribution

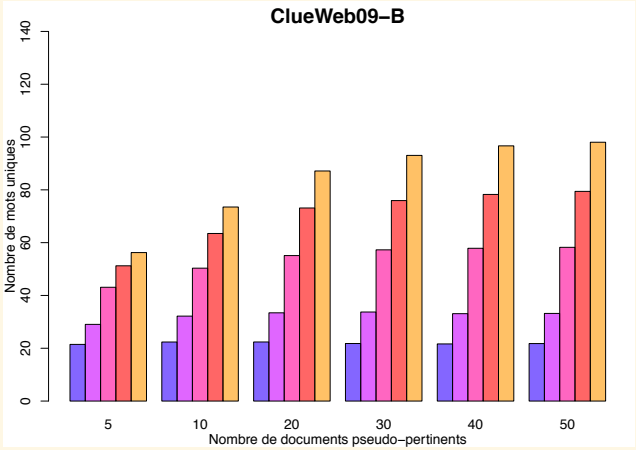
$$P'_{TM}(w|k,\boldsymbol{\theta}_{\hat{M}},\boldsymbol{\phi}_{\hat{K}}) = \frac{P_{TM}(w|k,\boldsymbol{\theta}_{\hat{M}},\boldsymbol{\phi}_{\hat{K}})}{\sum_{w'\in\mathcal{W}_k} P_{TM}(w'|k,\boldsymbol{\theta}_{\hat{M}},\boldsymbol{\phi}_{\hat{K}})}$$

Nombre de mots dans les concepts









Nombre de concepts
3 5 10 15 20

Influence des sources d'informations (MoATDRM)

	nyt	wiki	gigaword	web	WT10g	Robust04	GOV2	ClueWeb09-B
WT10g	**0,343**	0,090	0,160	0,313	0,089	-	-	-
Robust04	0,273	0,100	**0,309**	0,116	-	0,201	-	-
GOV2	0,247	0,188	0,187	0,093	-	-	**0,280**	-
ClueWeb09-B	0,142	0,173	0,202	**0,369**	-	-	-	0,113