Quantification et identification des concepts implicites d'une requête

Romain Deveaud[1] – Ludovic Bonnefoy[1] – Patrice Bellot[2]

[1] LIA – Université d'Avignon

[2] LSIS – Aix-Marseille Université

Introduction

améliorer la représentation du contexte thématique de la recherche [White *et al.*,SIGIR'09]

retour de pertinence simulé (Pseudo Relevance Feedback)

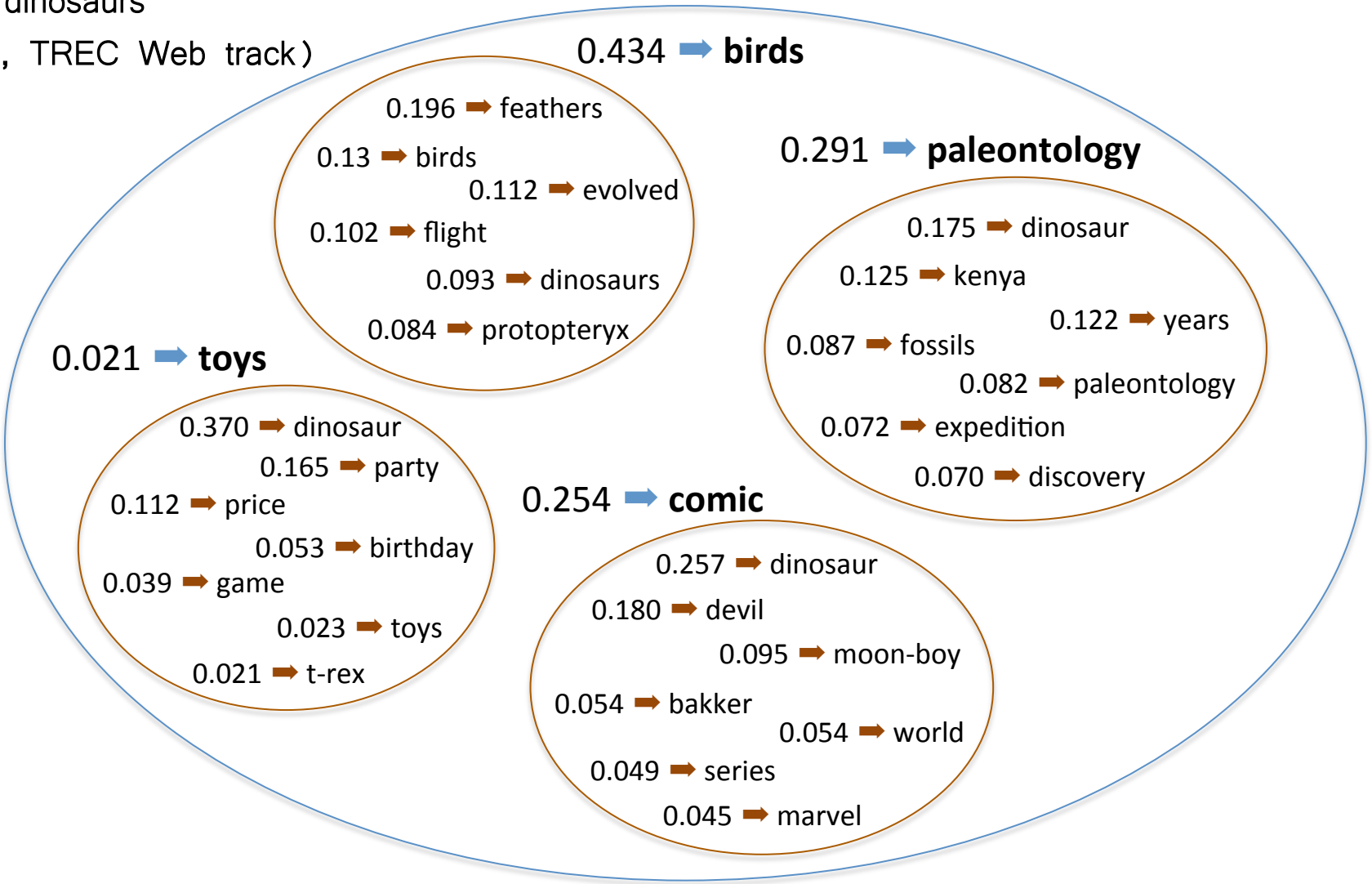
"RI conceptuelle" : les mots informatifs sont considérés comme des concepts

[Metzler & Croft,SIGIR'07;Egozi *et al.*,ACM TOIS'11]

selon [Stock,JASIST'10] un concept est défini comme une classe contenant des objets possédant certaines propriétés et attributs

Introduction

requête : dinosaurs
(topic 14, TREC Web track)



Introduction

utilisation du texte des documents de *feedback*

modélisation thématique (topic modeling)

identification des groupements de mots formant des concepts

allocation latente de Dirichlet (LDA) [Blei,JMLR'03]

quel nombre de concepts? combien de documents de feedback?

pas de supervision, peu de paramètres

quantification et identification de concepts implicites

modélisation thématique

estimer le nombre de concepts

combien de documents de *feedback*?

pondération des concepts

intégration des concepts pour la recherche documentaire

Modélisation thématique

aller au-delà du paradigme classique du sac de mots

sac de thèmes, qui sont des sacs de mots

LDA pour Latent Dirichlet Allocation, ou allocation latente de Dirichlet

LDA apprend donc à partir d'un ensemble de documents

la probabilité $\theta_{d,k}$ que le concept *k* apparaisse dans le document *d* (i.e. $P(k|d)$)

la probabilité $\phi_{k,w}$ que le mot *w* appartienne au concept *k* (i.e $P(w|k)$).

Modélisation thématique

rechercher les documents en fonction des thèmes qu'ils abordent (et non plus simplement les mots) est très attrayant !

- et plusieurs travaux ont utilisé LDA dans ce but [Andrzejewski & Buttler,SIGKDD'11;Park *et al.*, ECML PKDD'09;Wei & Croft,SIGIR'06]
- principe de base : apprendre la distribution thématique d'une collection de documents entière en fixant comme paramètre un grand nombre de thèmes (ou concepts)

comment modéliser des concepts fortement liés à la requête?

- modélisation thématique sur un sous-ensemble de documents liés à la requête

Estimer le nombre de concepts

soit une requête Q et $\mathcal{R}_Q$ un ensemble de documents de *feedback* récupérés par la première passe d'un système de RI état-de-l'art

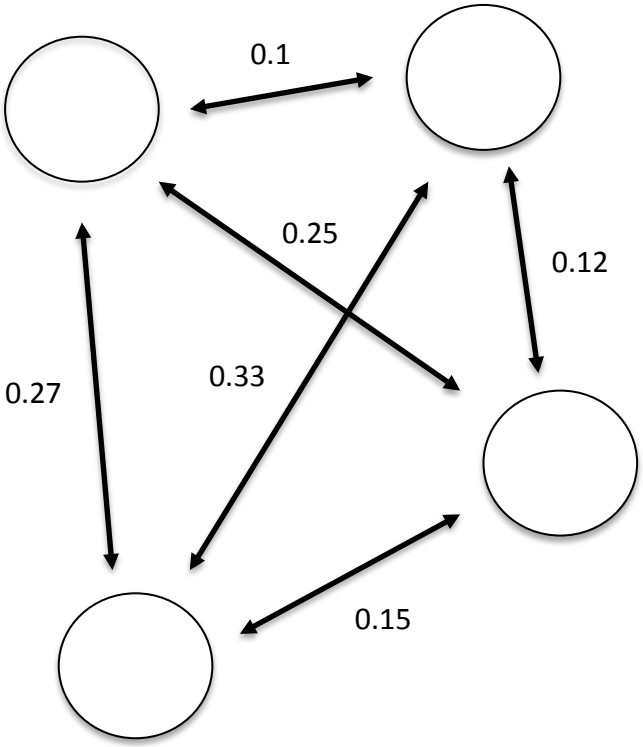
$\mathbb{T}_K$ est le modèle conceptuel appris par LDA sur l'ensemble $\mathcal{R}_Q$ avec le nombre de concepts K pour paramètre

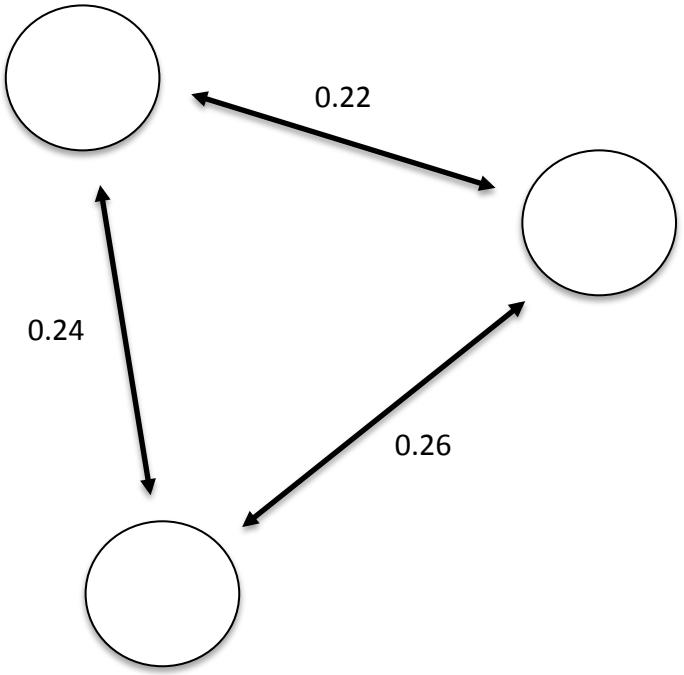
nous définissons alors le nombre de concepts implicites de Q par :

$$\hat{K} = \underset{K}{\operatorname{argmax}} \frac{1}{K(K-1)} \sum_{(k_i,k_j)\in\mathbb{T}_K} D(k_i||k_j)$$

où $D(k_i||k_j)$ est la divergence de Kullback-Leibler entre deux concepts

Estimer le nombre de concepts

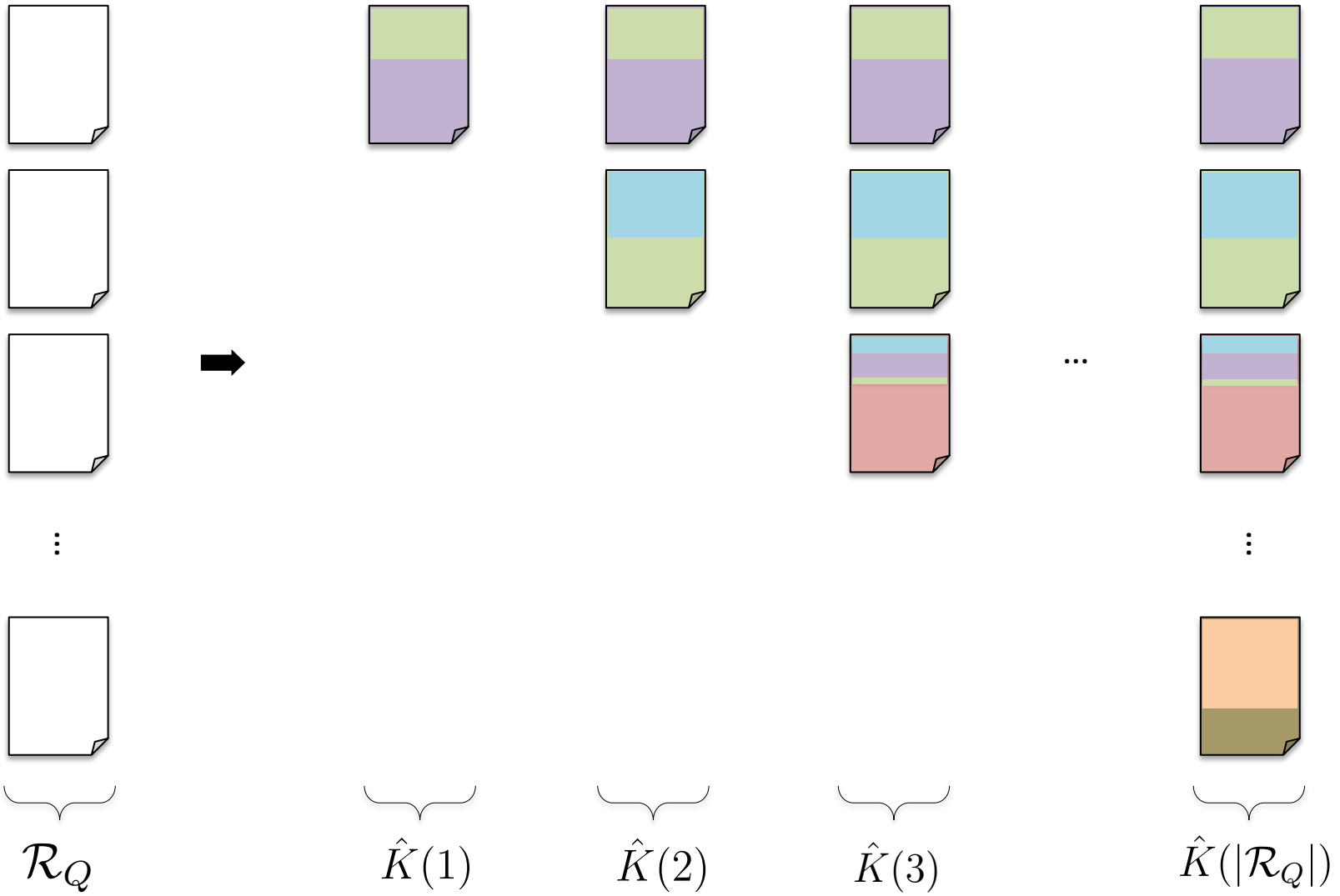




$\sum_{(k_i,k_j)\in\mathbb{T}_K} D(k_i||k_j)$ = 0.2033

$\sum_{(k_i,k_j)\in\mathbb{T}_K} D(k_i||k_j)$ = 0.24

Combien de documents de *feedback*?



Combien de documents de *feedback*?

un modèle conceptuel pour chaque sous-ensemble de documents

- éviter les modèles contenant des concepts bruités
- utilisation de la forte concentration (supposée) de documents pertinents dans les premiers rangs [He & Ounis,CIKM'09]
- les concepts apparaissant dans peu de modèles ne sont vraisemblablement pas liés à la requête

calculer la similarité entre tous les modèles conceptuels

- le modèle le plus similaire aux autres contient le moins de concepts "marginaux"
- espaces probabilistes différents => pas de divergence probabiliste

Combien de documents de *feedback*?

en pratique, on considère les 20 premiers sous-ensembles de documents de *feedback*

ensemble de 20 modèles conceptuels

il ne peut en rester qu'un...

$$M = \underset{m}{\operatorname{argmax}} \sum_{n,n\neq m} \underbrace{\sum_{k_j \in \mathbb{T}^m_{K(m)}} \sum_{k_i \in \mathbb{T}^n_{K(n)}}}_{\text{chaque paire de concepts appartenant à des modèles différents}} \underbrace{\frac{|k_i \cap k_j|}{|k_i|} \sum_{w \in k_i \cap k_j} \log \frac{N}{df_w}}_{\text{similarité entre deux concepts (représentés comme des sacs de mots)}}$$

[Metzler *et al.*,CIKM'05]

Pondération des concepts

les concepts n'ont pas tous la même importance par rapport à une requête

certains concepts peuvent être peu pertinents, ou être bruités

$$\delta_k = \sum_{D \in \mathcal{R}_Q} P(Q|D) \underbrace{P(k|D)}_{\text{distribution } \theta_d}$$

de la même façon, les mots n'ont pas tous la même importance au sein d'un même concept

$$\hat{\phi}_{k,w} = \frac{\overbrace{P(w|k)}^{\text{distribution } \phi_k}}{\sum_{w' \in \mathbb{W}_k} P(w'|k)}$$

Intégration des concepts pour la recherche documentaire

$$s(Q,D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \underbrace{\prod_{k \in \mathbb{T}_{\hat{K}(M)}} \hat{\delta}_k}_{\text{pondération de tous les concepts}} \underbrace{\prod_{w \in \mathbb{W}_k} \hat{\phi}_{k,w} \cdot P(w|D)}_{\text{recherche de l'ensemble des mots d'un concept}}$$

interpolation des concepts avec la requête originale

meilleures performances [Lavrenko & Croft,SIGIR'01]

$$P(Q|D) = \prod_{w \in Q} P(w|D)$$

lissage de Dirichlet (μ = 1500)

évaluation

protocole expérimental

sources d'information pour l'identification de concepts

recherche conceptuelle de documents

Evaluation

utilisations de deux collections majeures de TREC

- Robust04, reprend les requêtes jugées difficiles des “anciennes” tâches ad-hoc
- ClueWeb09-B, large collection de pages web utilisée pour la tâche Web

Nom	# documents	Topics utilisés
Robust04	528 155	301-450, 601-700
ClueWeb09-B	50 220 423	1-150

Tableau 3. *Résumé des collections de test de TREC utilisées pour notre évaluation.*

indexation et recherche documentaire avec Indri

- racinisation de Krovetz
- 10 mots par concept

Evaluation

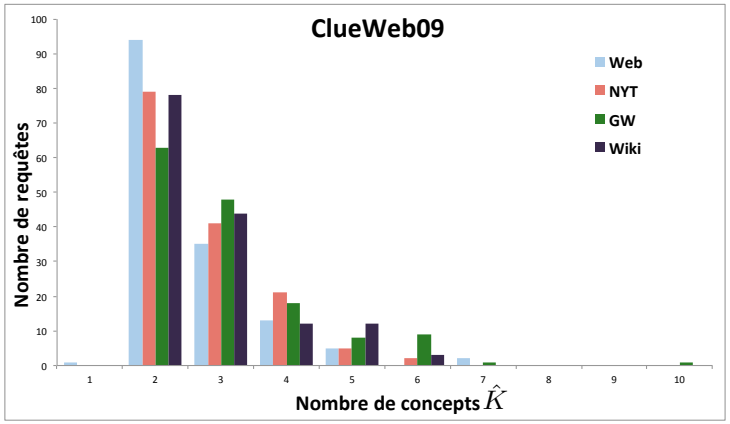
utilisation de plusieurs sources d'information pour l'extraction de concepts

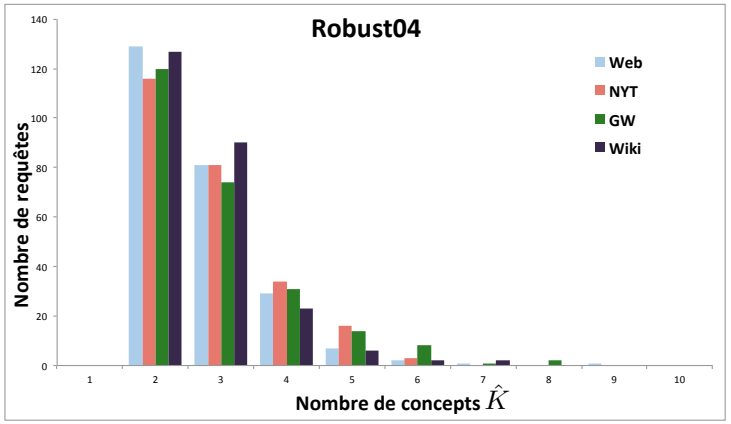
- Wikipédia
- New Yorks Times LDC
- GigaWord LDC
- Web (ClueWeb09-B nettoyé de plus de la moitié de ses documents)

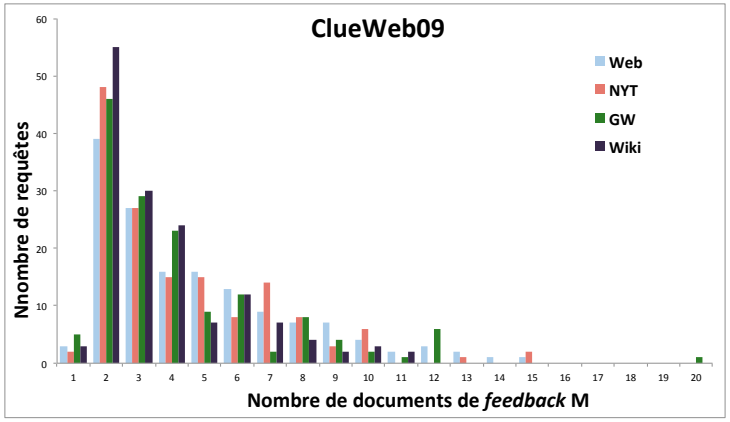
Ressource	# documents	# mots unique	# total de mots
NYT	1 855 658	1 086 233	1 378 897 246
Wiki	3 214 014	7 022 226	1 033 787 926
GW	4 111 240	1 288 389	1 397 727 483
Web	29 038 220	33 314 740	22 814 465 842

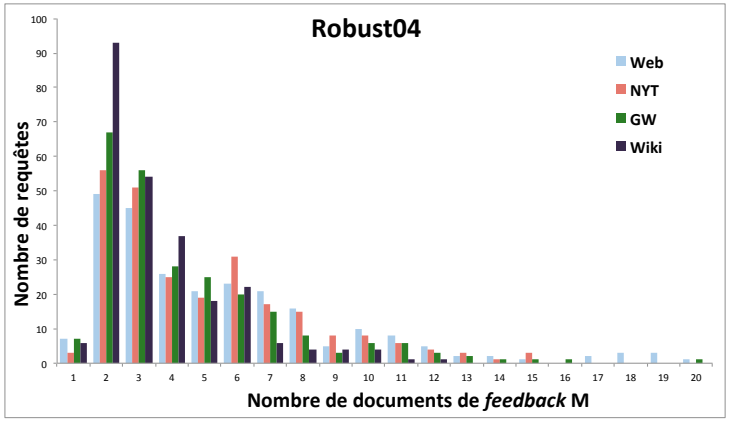
Tableau 2. *Récapitulatif des quatre sources d'information générales utilisées.*

Evaluation

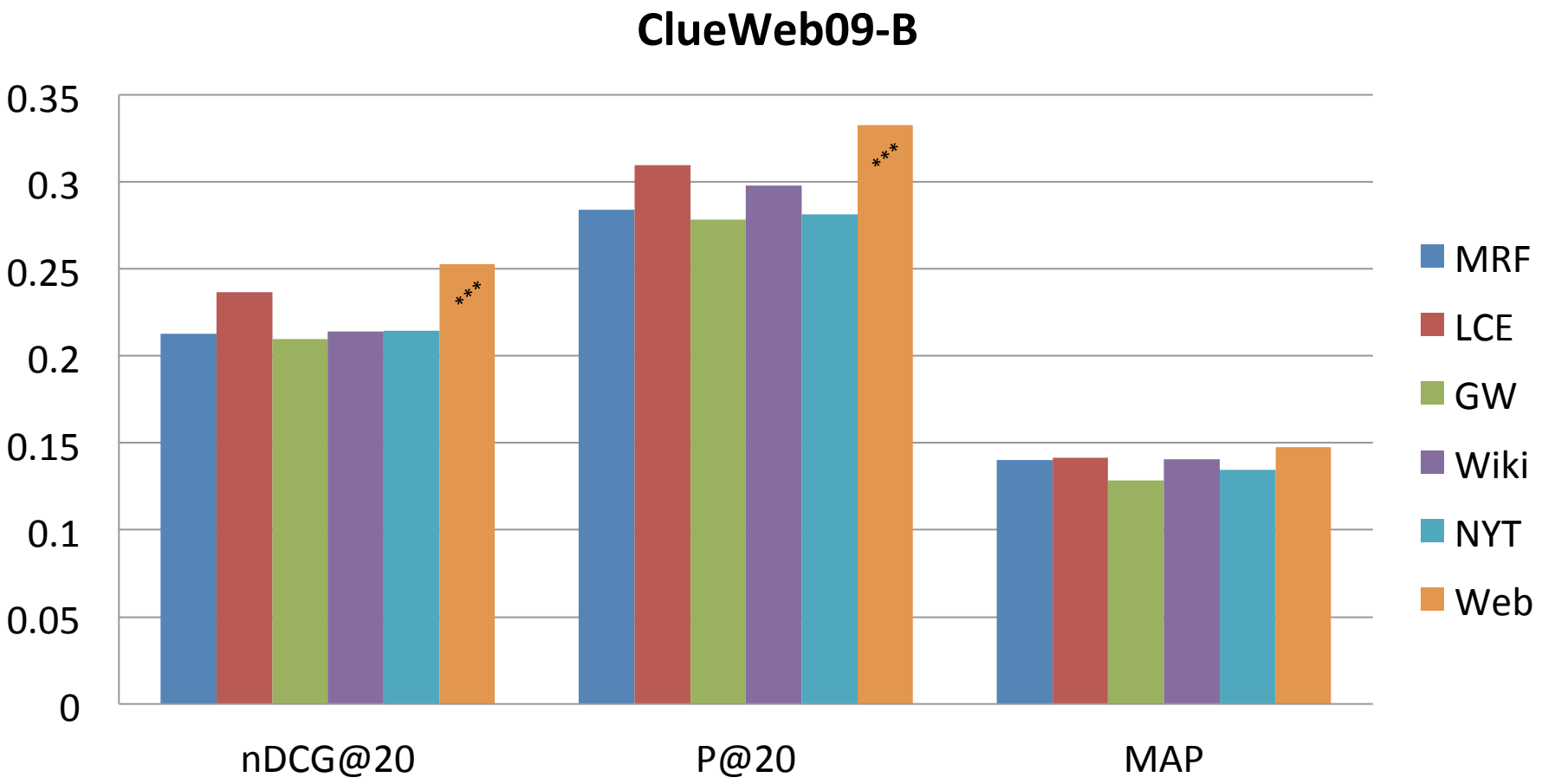

ClueWeb09
Web
NYT
GW
Wiki
Nombre de requêtes
Nombre de concepts $\hat{K}$


Robust04
Web
NYT
GW
Wiki
Nombre de requêtes
Nombre de concepts $\hat{K}$

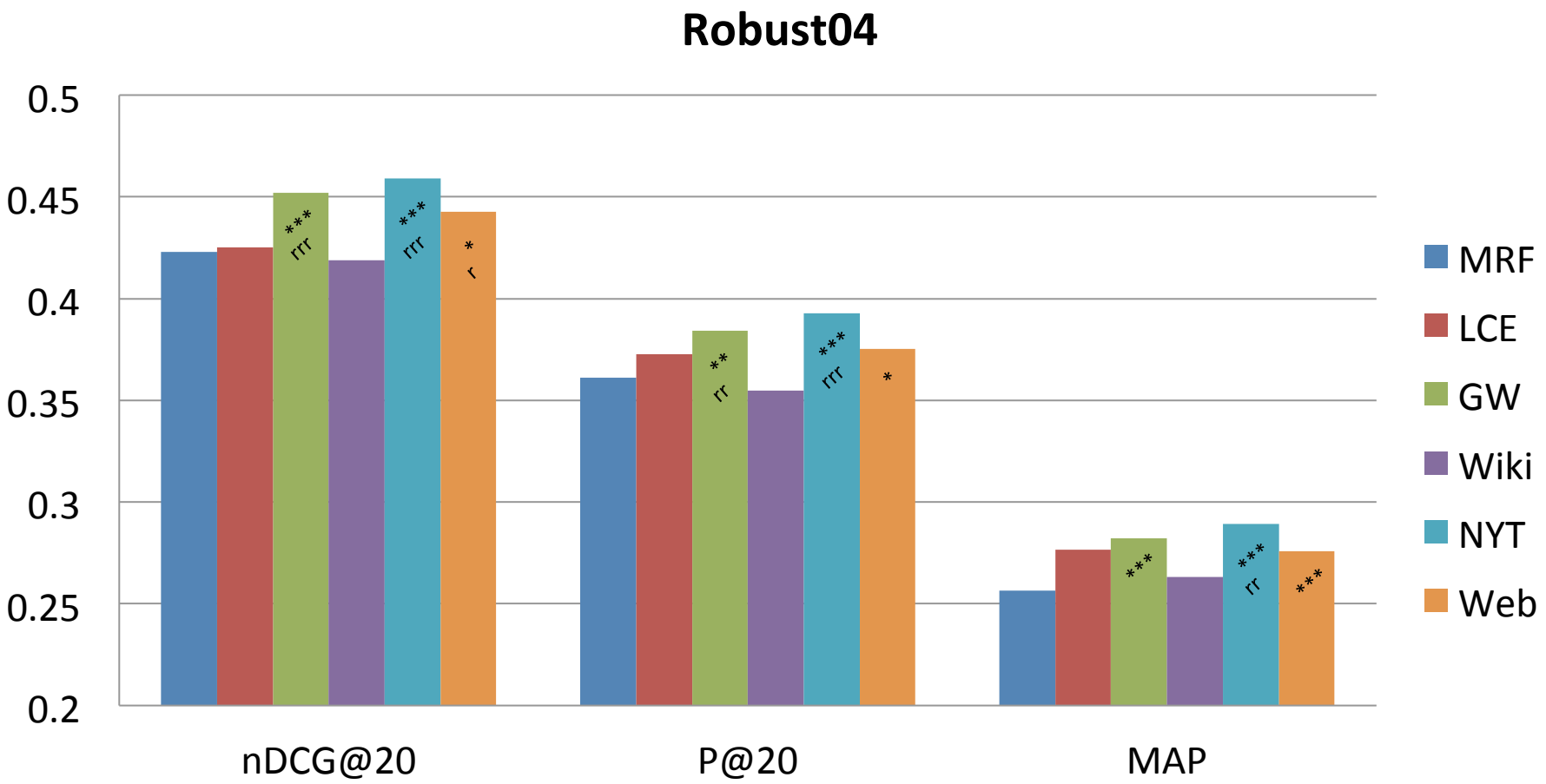

ClueWeb09
Web
NYT
GW
Wiki
Nnombre de requêtes
Nombre de documents de *feedback* M


Robust04
Web
NYT
GW
Wiki
Nombre de requêtes
Nombre de documents de *feedback* M

Evaluation


ClueWeb09-B
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
***
***
nDCG@20
P@20
MAP
MRF
LCE
GW
Wiki
NYT
Web

Evaluation


Robust04
0.5
0.45
0.4
0.35
0.3
0.25
0.2
nDCG@20
P@20
MAP
MRF
LCE
GW
Wiki
NYT
Web

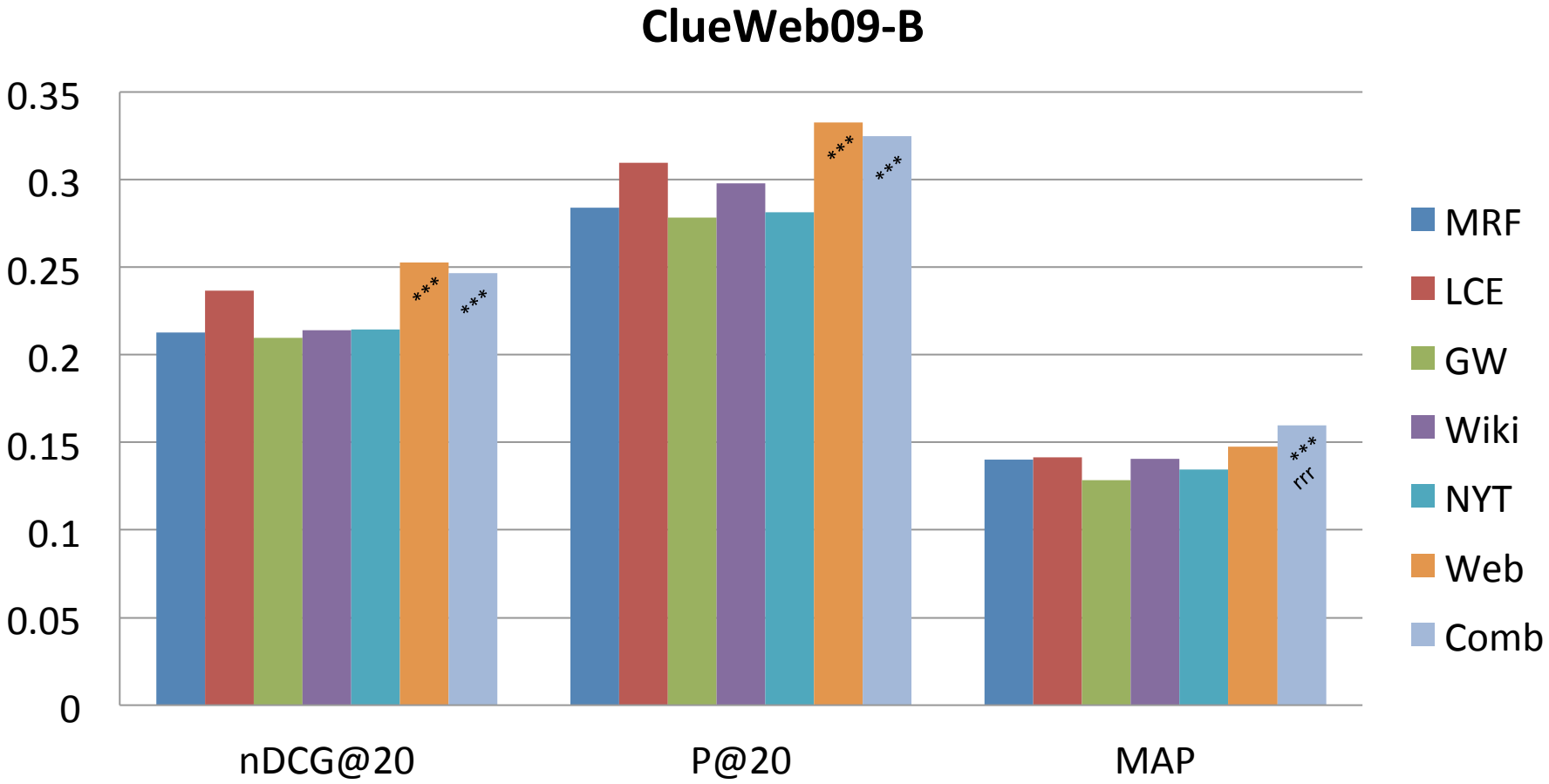
Evaluation

combinaison des concepts implicites provenant des différentes sources

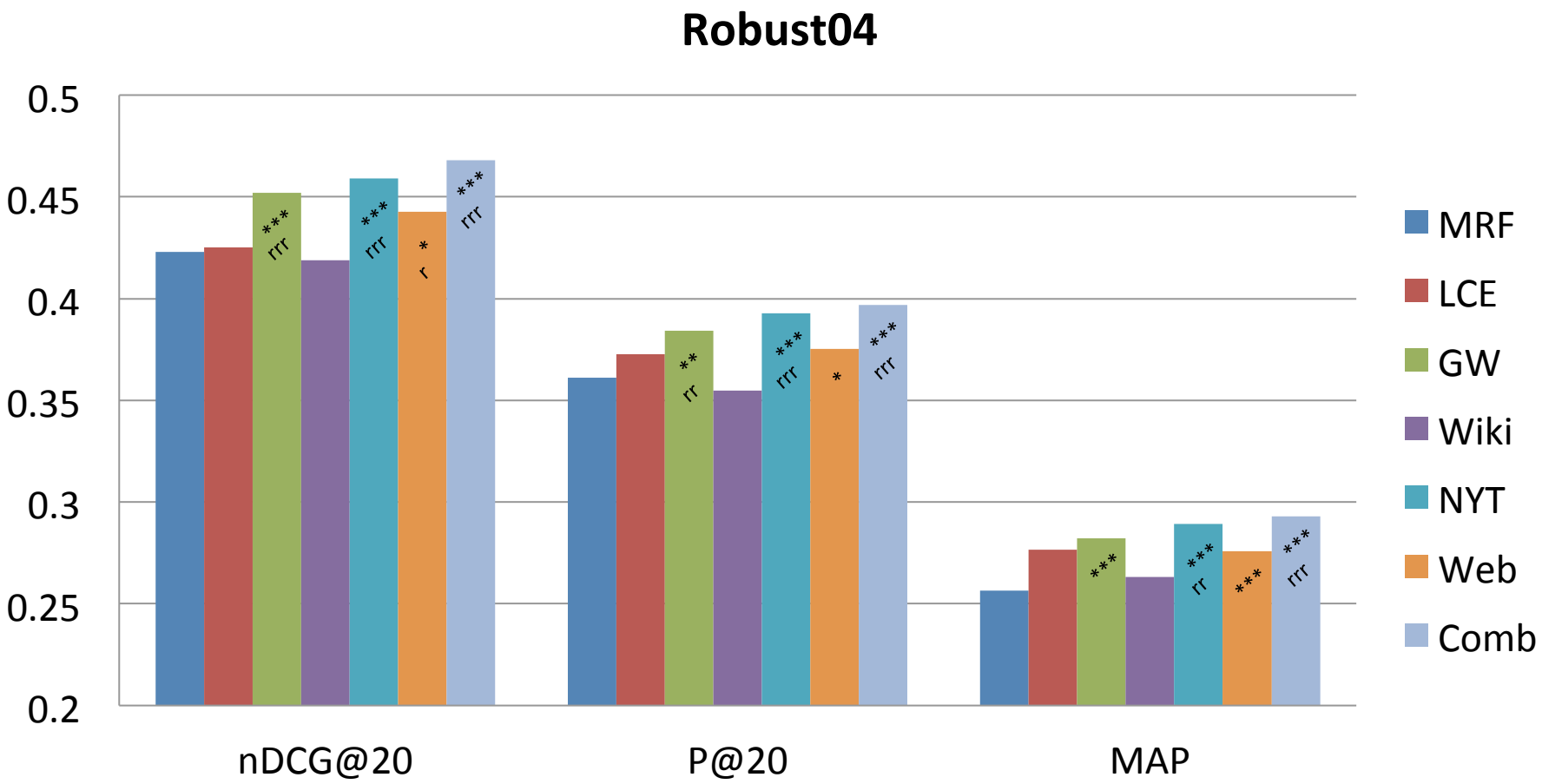
$$s(Q,D) = \lambda \cdot P(Q|D) + (1-\lambda) \cdot \frac{1}{|\mathcal{S}|} \prod_{\sigma \in \mathcal{S}} \prod_{k \in \mathbb{T}^{\sigma}_{\hat{K}(M)}} \hat{\delta}_k \prod_{w \in \mathbb{W}_k} \hat{\phi}_{k,w} \cdot P(w|D)$$

moyenne des concepts extraits de différentes sources d'information

Evaluation


ClueWeb09-B
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
nDCG@20
P@20
MAP
***
***
***
***
***
rrr
MRF
LCE
GW
Wiki
NYT
Web
Comb

Evaluation


Robust04
0.5
0.45
0.4
0.35
0.3
0.25
0.2
nDCG@20
P@20
MAP
MRF
LCE
GW
Wiki
NYT
Web
Comb

Conclusion

représenter explicitement les concepts implicites permet d'améliorer significativement la pertinence des documents renvoyés

- moins clair pour une tâche de recherche Web...
- ... mais des travaux en cours utilisant d'autres collections valident ces résultats

la nature des sources d'information utilisées influence logiquement la qualité des concepts

la combinaison (moyenne) des ressources améliore également les résultats

- ou en tout cas améliore la significativité
- ce qui suggère une plus grande robustesse (également confirmé par des travaux en cours)

merci de votre attention