Unsupervised Latent Concept Modeling to Identify Query Facets

Romain Deveaud[α] – Eric SanJuan[α] – Patrice Bellot[β]

[α] University of Avignon

[β] Aix-Marseille University

Introduction

« *the user's own request formulation is a representation of [her] current cognitive state concerned with an information need* »

[Ingwersen, SIGIR'94]

expressing an information need with 2-3 keywords is not a trivial task

complex information need, lack of vocabulary, lack of background knowledge

Introduction

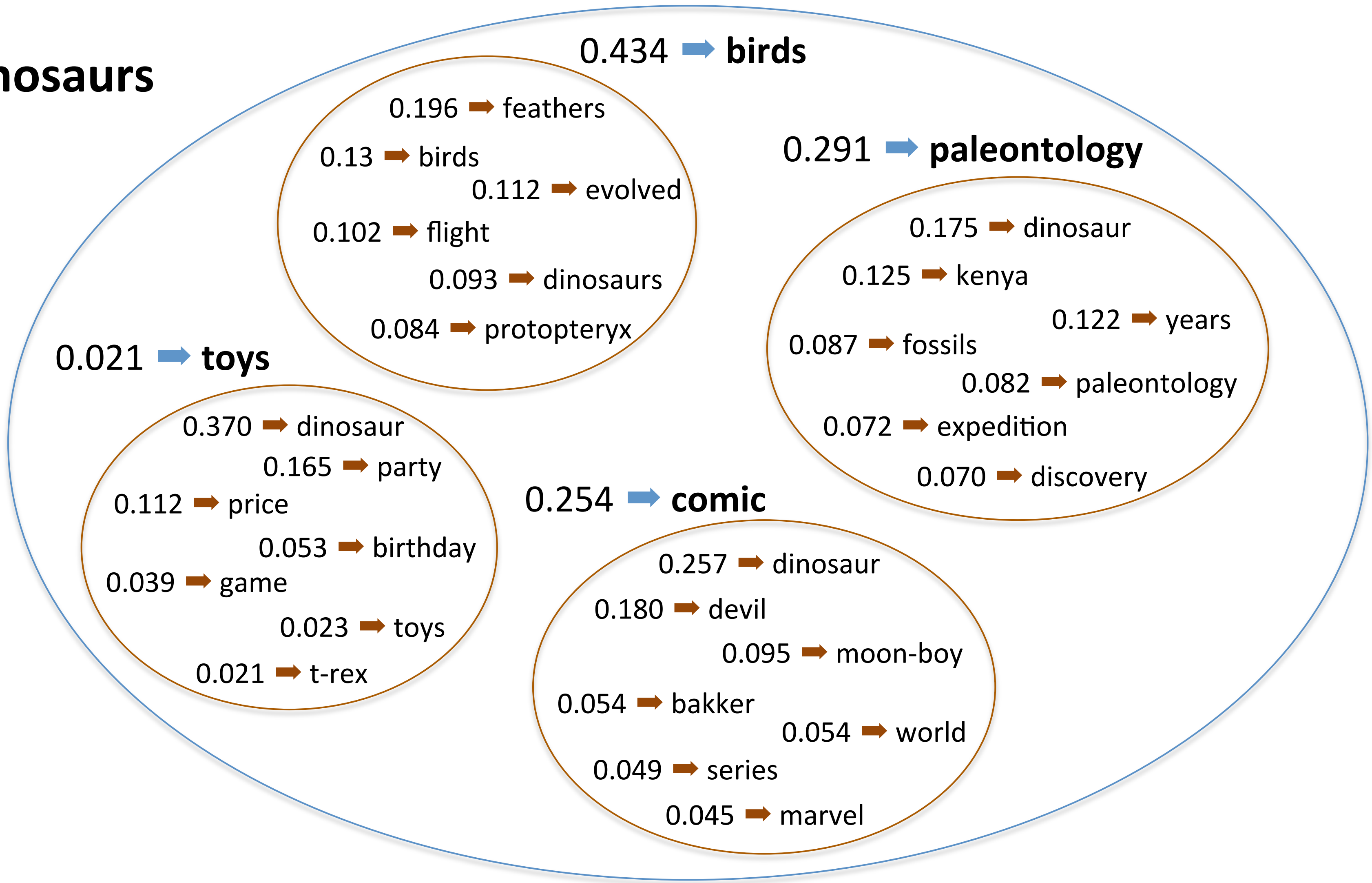
dynamically infer the concepts of the query (unlike faceted search)

(ultimately) full description the information need [Metzler & Croft, SIGIR'07; Egozi *et al.*, ACM TOIS'11]

human concepts are too complex to be expressed by single words [Stock, JASIST'10]

Introduction

Q = **dinosaurs**



Introduction

pseudo-relevance feedback

topic modeling (LDA [Blei, JMLR'03]) *on* feedback documents

two problems: which number of concepts? which pseudo-relevant feedback documents?

Estimating the number of concepts

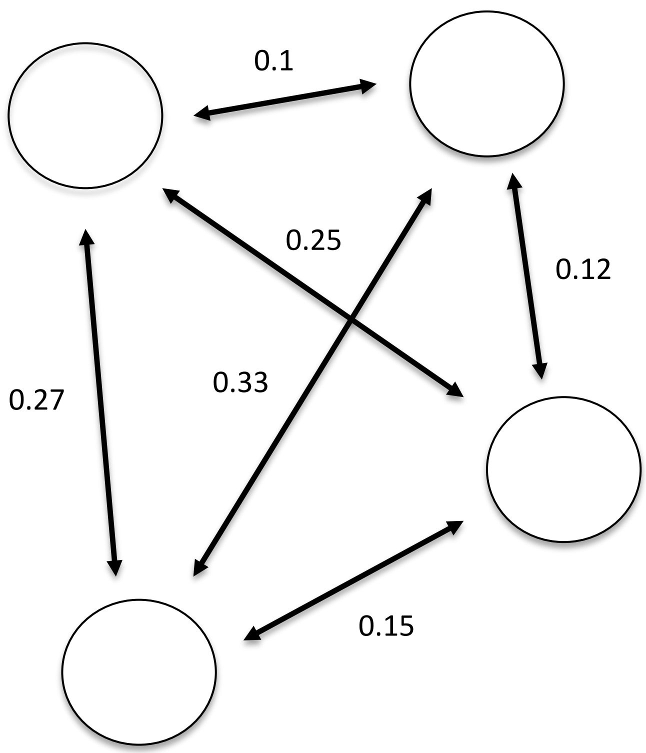
given a query Q, $\mathcal{R}_Q$ is a set of pseudo-relevant feedback documents retrieved by a state-of-the-art IR system

probabilistic topic models need a predefined number of topics

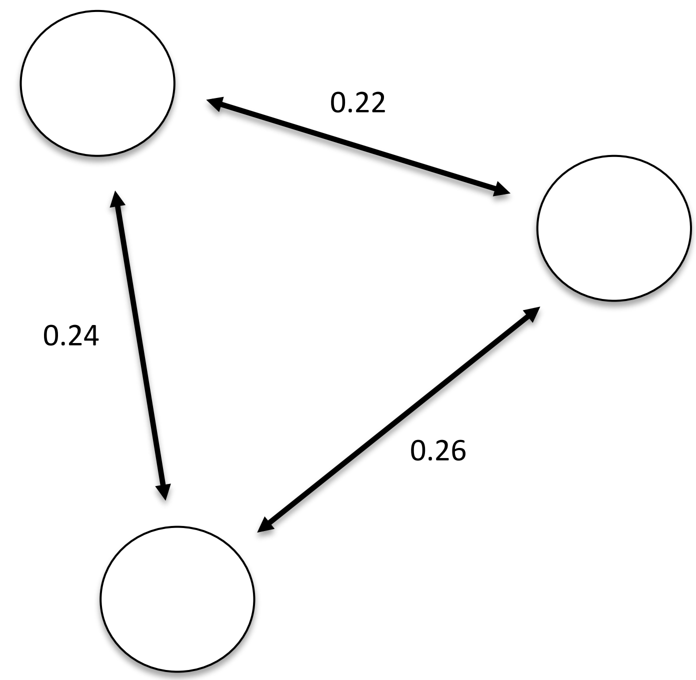
how much topics in $\mathcal{R}_Q$?

try several values, and keep the topic model $\mathbb{T}_K$ which models the most scattered topics

Estimating the number of concepts



total = 0.2033



total = 0.24

Estimating the number of concepts

topics are probability distributions

measuring the average Kullback-Leibler divergence between all pairs of topics

number of latent concepts in $\mathcal{R}_Q$:

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

Maximizing conceptual coherence

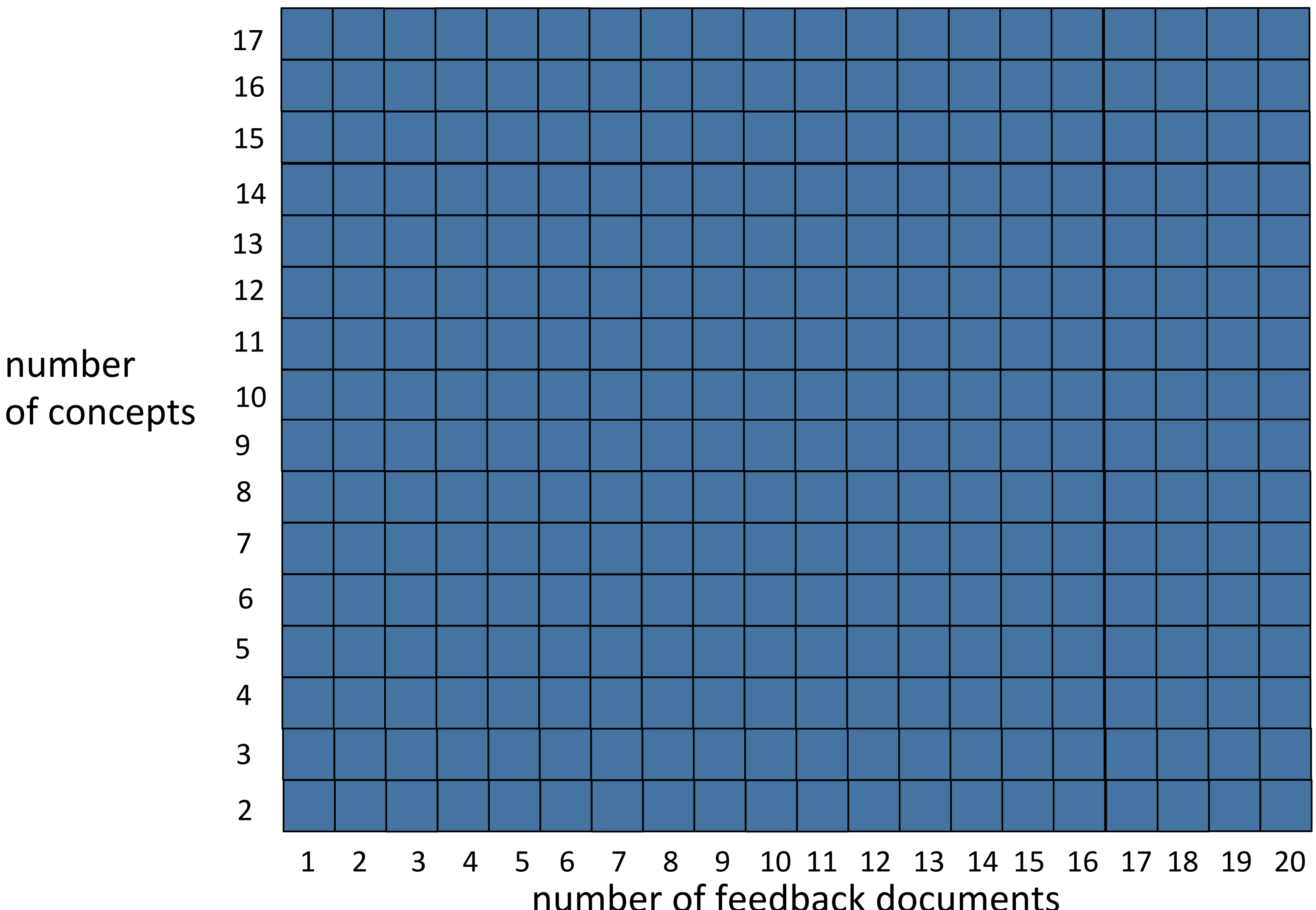
we estimate a number of concepts for a given set of documents (can be 2, 5, 10, 10,000...)

in other words, we model a set of concepts for a given set of documents

using more documents provide more information...

... which could be noise

Maximizing conceptual coherence



Q = **dinosaurs**

Maximizing conceptual coherence

choosing the « best » model

maximizing similarity in order to discard marginal concepts

concept models not in the same probabilistic space (different sets of documents)

$$M = \operatorname*{argmax}_{m} \sum_{n, n \neq m} \underbrace{\sum_{k_j \in \mathbb{T}^m_{K(m)}} \sum_{k_i \in \mathbb{T}^n_{K(n)}}}_{\text{each pair of concept from different models}} \underbrace{\frac{|k_i \cap k_j|}{|k_i|} \sum_{w \in k_i \cap k_j} \log \frac{N}{df_w}}_{\text{similarity between two concepts}}$$

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

Concept weighting

reflecting the relative importance of each concept...

$$\delta_k = \sum_{D \in \mathcal{R}_Q} P(Q|D) P_{TM}(k|D)$$

... and each word

$$\hat{\phi}_{k,w} = \frac{P_{TM}(w|k)}{\sum_{w' \in \mathbb{W}_k} P_{TM}(w'|k)}$$

Document ranking

language modeling approach to IR

Dirichlet smoothing

linear interpolation of query likelihood and weighted concepts

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

balance parameter

Experiments & evaluation

4 different sources of information used for concept modeling

Resource	# documents	# unique words	# total words
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

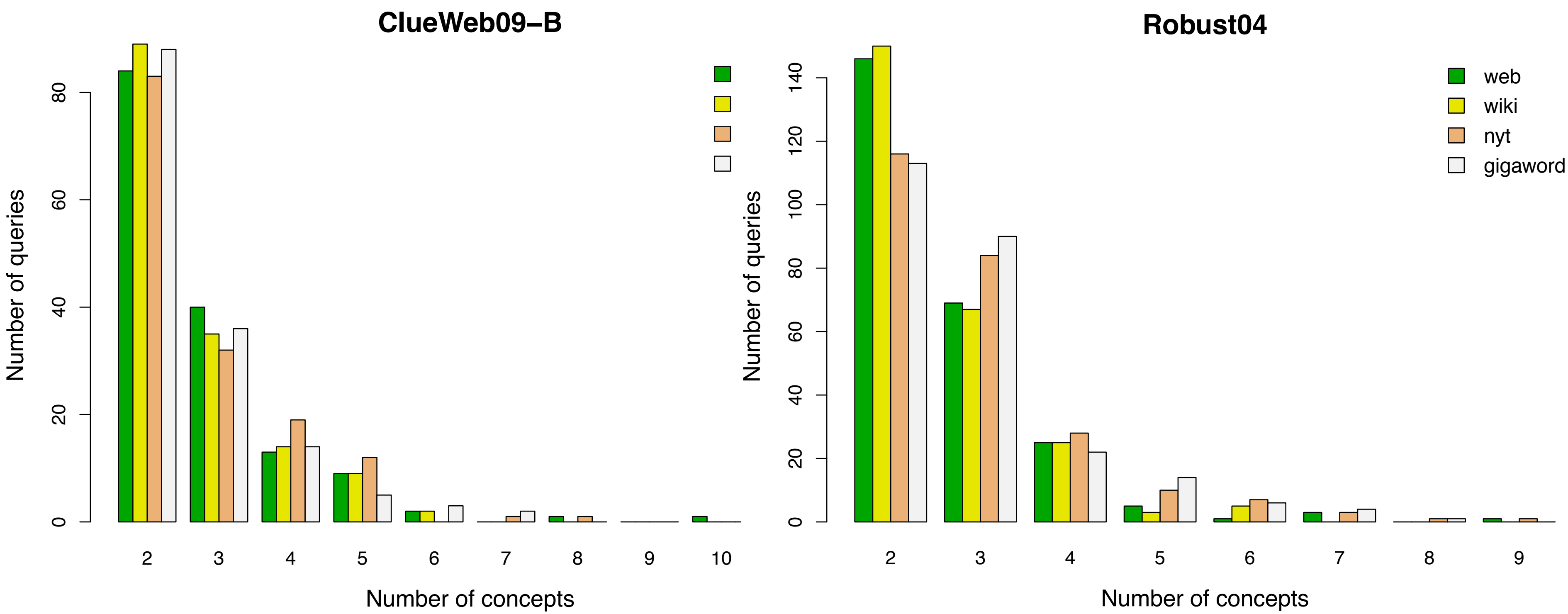
Table 2: Information about the four general sources of information used in this work.

2 test collections

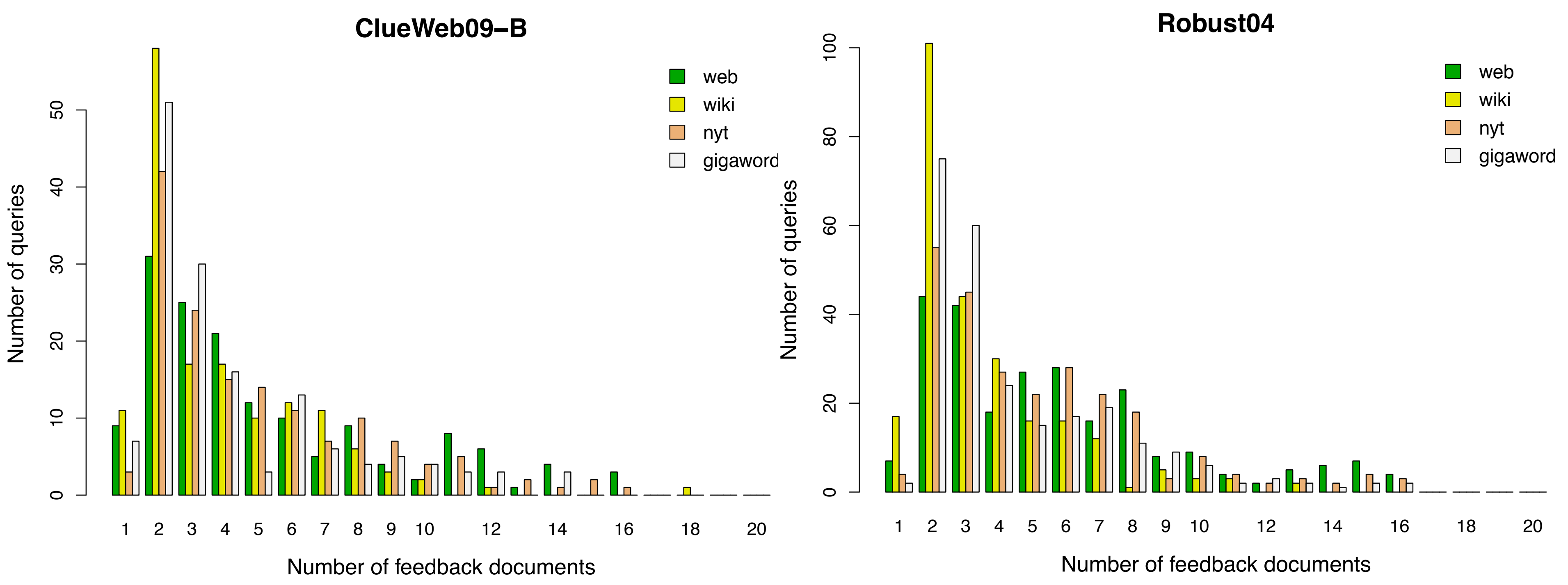
Name	# documents	Topics used
Robust04	528,155	301-450, 601-700
ClueWeb09-B	50,220,423	1-150

Table 4: Summary of the TREC test collections used for evaluation.

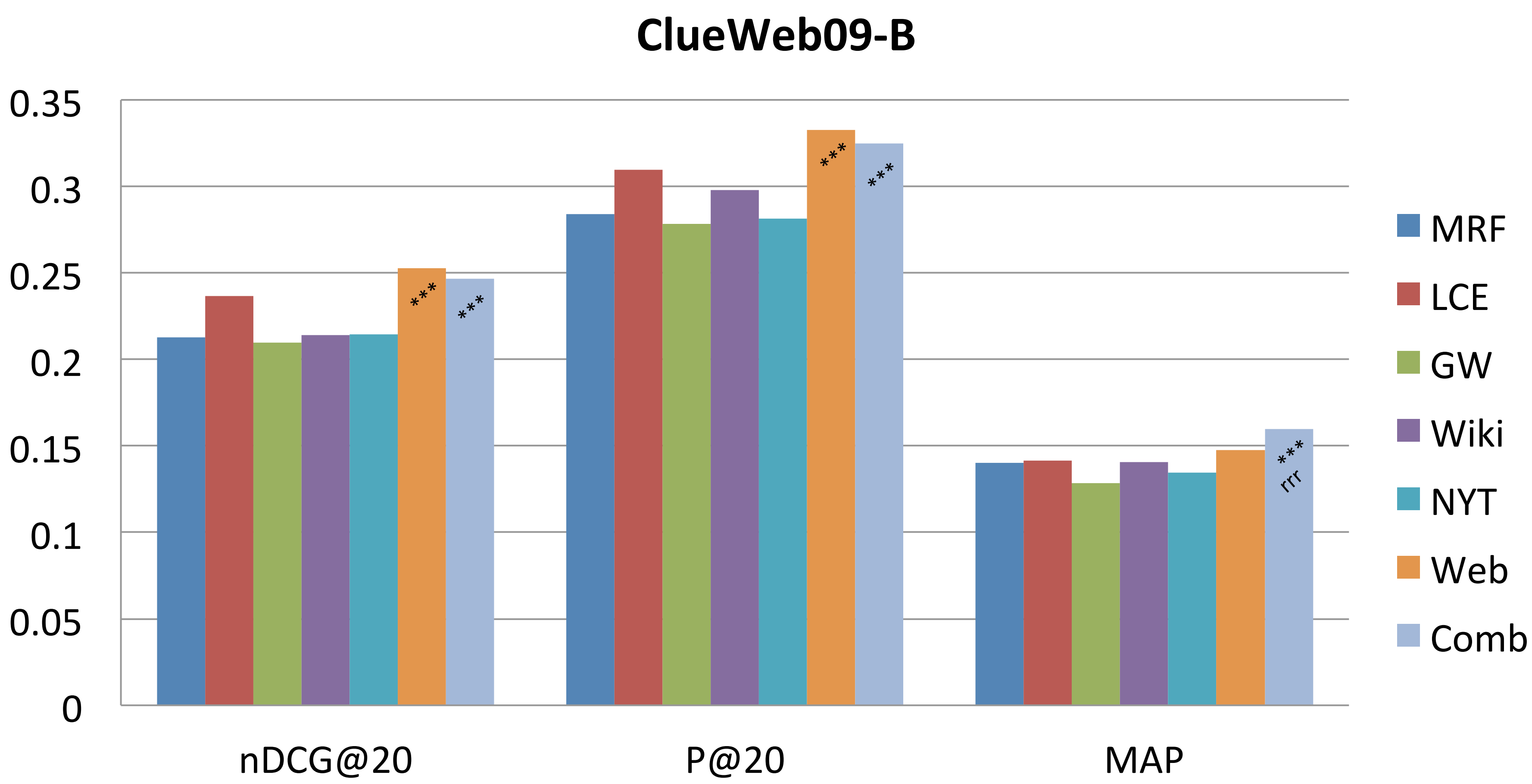
Experiments & evaluation



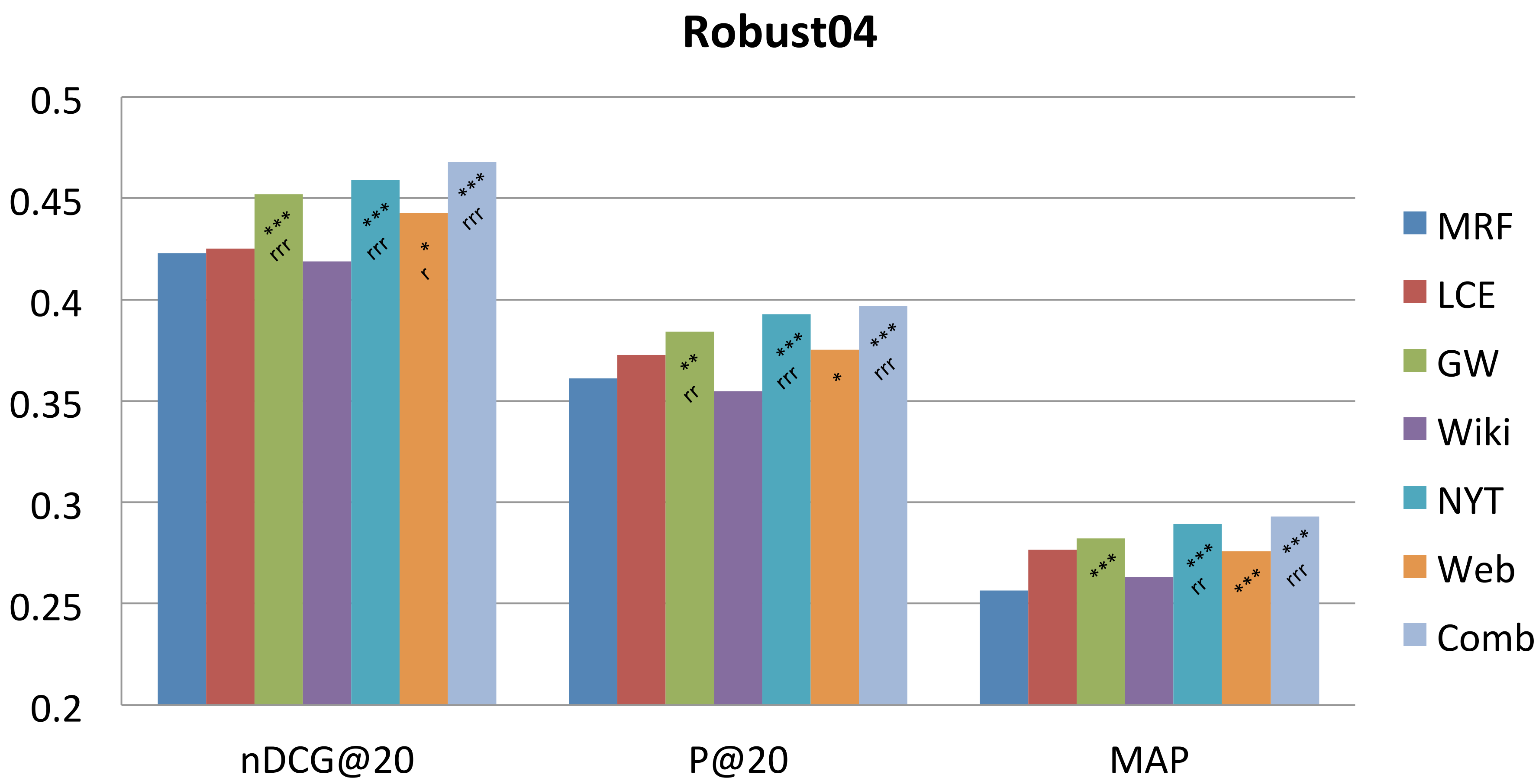
Experiments & evaluation



Experiments & evaluation



Experiments & evaluation



Experiments & evaluation

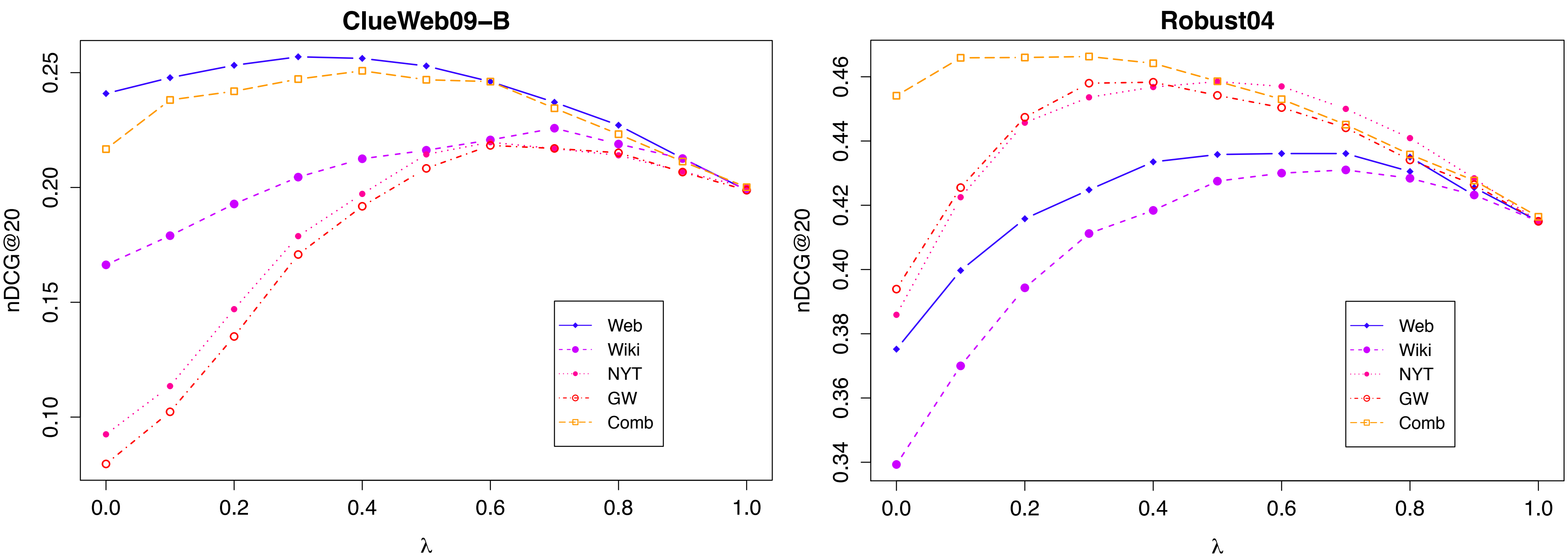


Figure 2: Retrieval performance (in nDCG@20) as a function of parameter λ.

Conclusion

unsupervised approach to identify query concepts

integration of several sources of information

may benefit from supervised training

entity linking

thank you for your attention