LIA at TREC 2012 Web Track: Unsupervised Search Concepts Identification from General Sources of Information

Romain Deveaud[*] – Eric SanJuan[*] – Patrice Bellot[**]

[*] LIA – University of Avignon

[**] LSIS – Aix-Marseille University

UNIVERSITÉ D'AVIGNON

Introduction

- what makes a diversified result list ?
  - relevant documents
  - cover all (or at least many) sub-topics
  - prevent redundancy

- modeling query concepts (or aspects, or intents)
  - promoting documents that match one or several concepts

Introduction

- topic modeling approach for identifying concepts
  - but we need it to be query-oriented

- put it together with pseudo-relevance feedback

- rank documents (mostly) based on their likelihood of generating the concepts

Introduction

- relies on several query-dependent parameters
  - number of concepts
  - number of feedback documents
  - -> adaptive approach

- exploration of the influence of several heterogeneous sources of feedback documents

Outline

- Introduction
- Query-oriented topic modeling
- General sources of information
- Results
- Conclusions and future work

Modeling query concepts

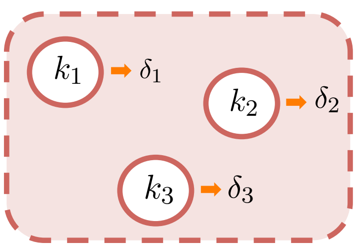
Source of information R

query Q

pseudo-relevant set of documents R_Q

topic modeling (Latent Dirichlet Allocation [Blei *et al.*, 2003])

query-related concept model



- several parameters to estimate
  - number of concepts
  - number of feedback documents
- adaptive parameters

Modeling query concepts (2)

```xml
<topic number="174" type="ambiguous">
  <query> rock art </query>
  <description> Where can I learn about rock painting or buy a rock-painting kit? </description>
  <subtopic number="1" type="inf"> Where can I learn about rock painting or buy a rock-painting kit? </subtopic>
  <subtopic number="2" type="nav"> Where can I buy tools for stone carving or engraving? </subtopic>
  <subtopic number="3" type="inf"> Find information on cave paintings in France. </subtopic>
  <subtopic number="4" type="nav"> Where can I buy rock and roll posters? </subtopic>
  <subtopic number="5" type="inf"> Find information on the artwork used on rock music album covers. </subtopic>
</topic>
```

$P(w\|k_1)$	w
0.35767049	rock
0.24763384	art
0.07159416	paintings
0.05064394	site
0.03579499	world
0.03541925	petroglyphs
0.03131438	mexico

δ_1 = 0.38223408

$P(w\|k_2)$	w
0.34878048	art
0.32767137	rock
0.04390243	cultural
0.04146341	human
0.03902439	cognitive
0.03658536	markings
0.03414634	cave

δ_2 = 0.10889479

$P(w\|k_3)$	w
0.39118065	rock
0.17496443	art
0.15647226	music
0.03982930	experimental
0.03982930	progressive
0.03982930	bands
0.02844950	pop

δ_3 = 0.20064946

$P(w\|k_4)$	w
0.36349674	art
0.30735686	rock
0.05117953	prehistoric
0.04740632	petroglyphs
0.04602233	research
0.03516577	archaeology
0.02930480	pottery

δ_4 = 0.30822165

– 4 concepts modeled from 8 feedback documents

Modeling concepts and choosing feedback documents

- the number of latent concepts depends on feedback documents
  - not the same concepts are expressed through 3 or through 10 documents
  - increasing the number of feedback documents increases diversity...
  - ... but also increases the likelihood of picking non-relevant documents

Modeling concepts and choosing feedback documents (2)

- how to accurately chose feedback documents when doing PRF?
  - dependent on the query and on the collection
  - machine learning approaches [Lv;He, CIKM'09]

- moving from a set of feedback documents to a mixture of topics
  - a set of N feedback documents can be represented by a mixture of K topics

Modeling concepts and choosing feedback documents (3)

- the « best » set of feedback documents is the one that has the higher topical coverage with all other feedback sets
  - all feedback documents discuss closely related topics
  - a marginal topic appearing in a feedback set is likely to be spam or non-relevant

Modeling concepts and choosing feedback documents (4)

- retrieving top m feedback documents (m ∈ [1,20])
  - but the number of concepts vary from one set to another

- estimating the optimal number of concepts for each feedback set of m documents

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

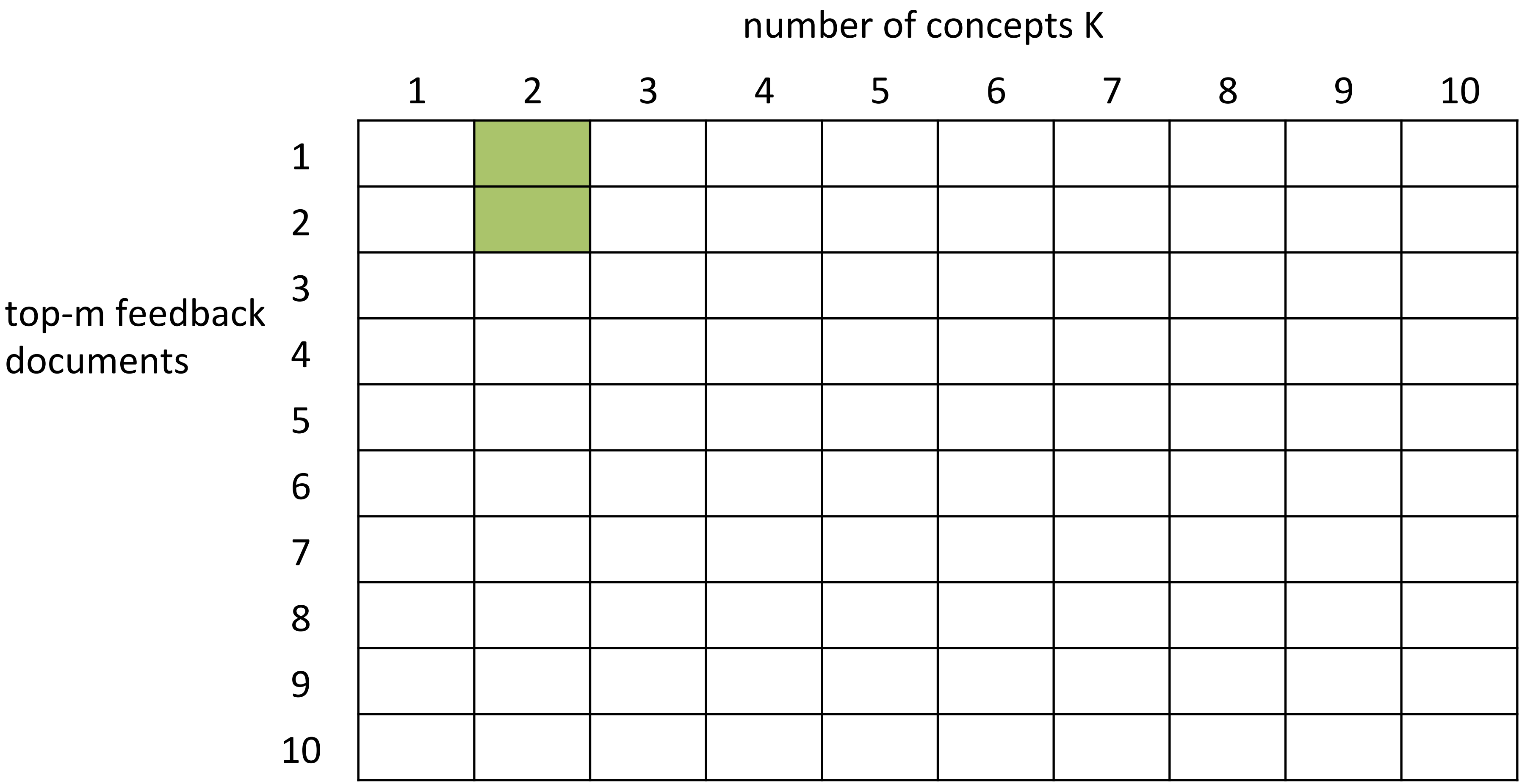
Modeling concepts and choosing feedback documents (5)

number of concepts K

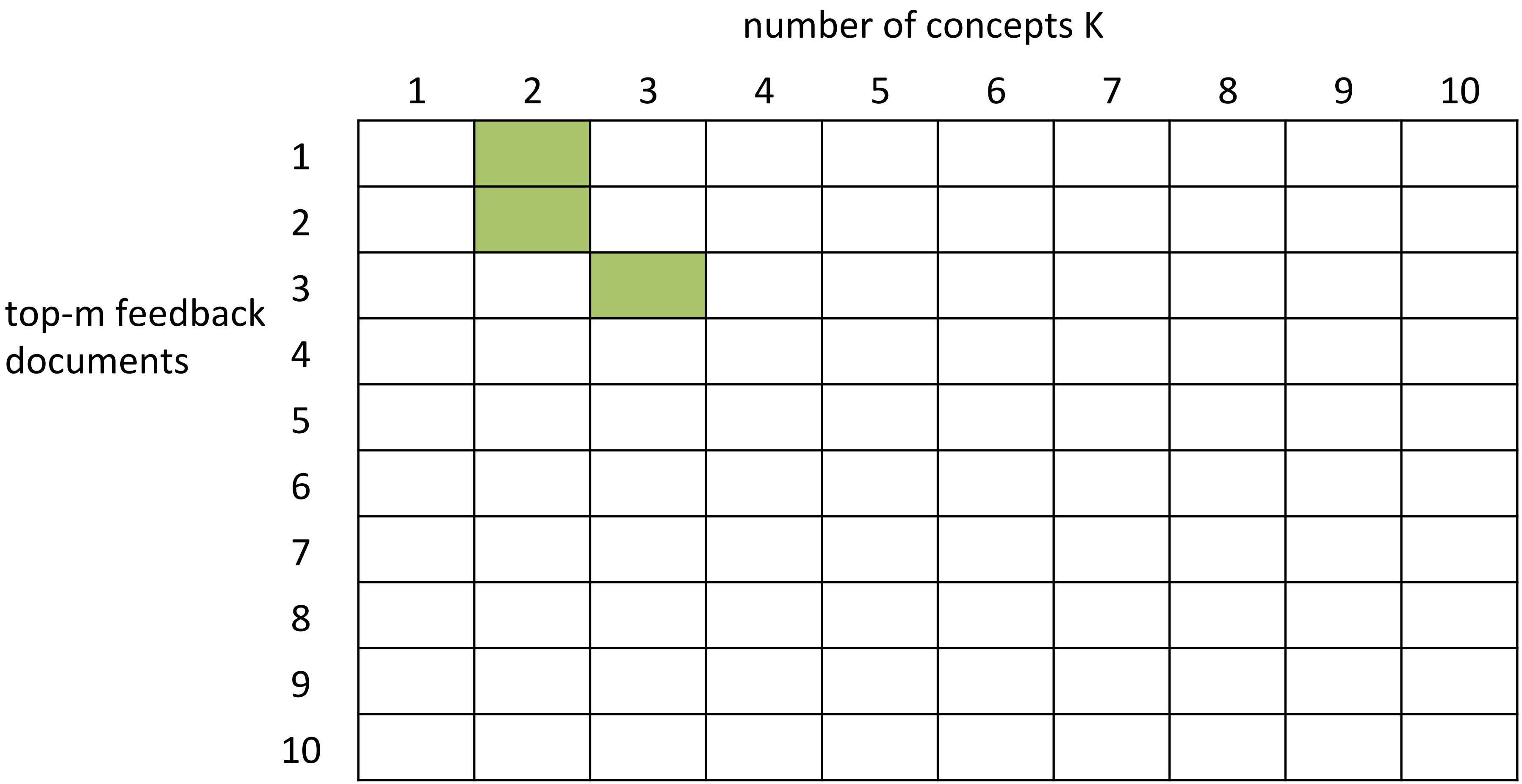
top-m feedback documents

	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										

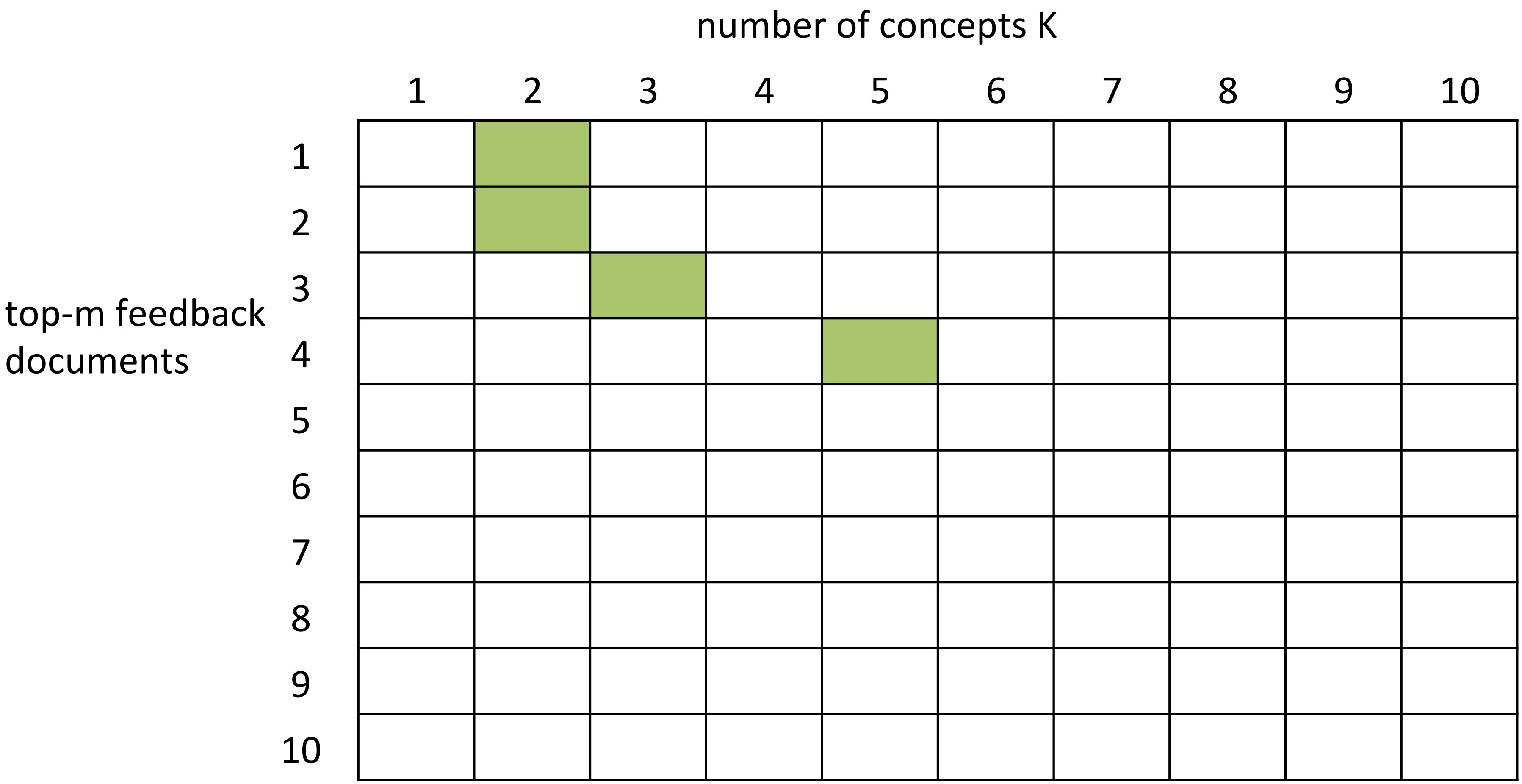
Modeling concepts and choosing feedback documents (5)



Modeling concepts and choosing feedback documents (5)



Modeling concepts and choosing feedback documents (5)



Modeling concepts and choosing feedback documents (6)

- we just computed a *concept model* for each top-m feedback documents set
  - i.e. 20 concept models (since m $\in$ [1,20])
  - and we need to chose one

- minimize marginality <=> maximize similarity

Modeling concepts and choosing feedback documents (7)

- similarity between two *concept models*

$$sim(\mathbb{T}_{\hat{K}(m)}, \mathbb{T}_{\hat{K}(n)}) = \sum_{k \in \mathbb{T}_{\hat{K}(m)}} \sum_{k' \in \mathbb{T}_{\hat{K}(n)}} \underbrace{\frac{|k_i \cap k'_j|}{|k_i|}}_{\text{topics words overlap}} \sum_{w \in W} \underbrace{\boxed{p(w|k)p(w|k')}}_{\text{word probabilities for both topics}} \underbrace{\log \frac{N}{df_w}}_{\text{word IDF}}$$

- the best *concept model* is the one which is the most similar to the others

$$M = \underset{m}{\operatorname{argmax}} \sum_{n} sim(\mathbb{T}_{\hat{K}(m)}, \mathbb{T}_{\hat{K}(n)})$$

Concept representation

```xml
<topic number="174" type="ambiguous">
  <query> rock art </query>
  <description> Where can I learn about rock painting or buy a rock-painting kit? </description>
  <subtopic number="1" type="inf"> Where can I learn about rock painting or buy a rock-painting kit? </subtopic>
  <subtopic number="2" type="nav"> Where can I buy tools for stone carving or engraving? </subtopic>
  <subtopic number="3" type="inf"> Find information on cave paintings in France. </subtopic>
  <subtopic number="4" type="nav"> Where can I buy rock and roll posters? </subtopic>
  <subtopic number="5" type="inf"> Find information on the artwork used on rock music album covers. </subtopic>
</topic>
```

$P(w\|k_1)$	w
0.35767049	rock
0.24763384	art
0.07159416	paintings
0.05064394	site
0.03579499	world
0.03541925	petroglyphs
0.03131438	mexico

δ_1 = 0.38223408

$P(w\|k_2)$	w
0.34878048	art
0.32767137	rock
0.04390243	cultural
0.04146341	human
0.03902439	cognitive
0.03658536	markings
0.03414634	cave

δ_2 = 0.10889479

$P(w\|k_3)$	w
0.39118065	rock
0.17496443	art
0.15647226	music
0.03982930	experimental
0.03982930	progressive
0.03982930	bands
0.02844950	pop

δ_3 = 0.20064946

$P(w\|k_4)$	w
0.36349674	art
0.30735686	rock
0.05117953	prehistoric
0.04740632	petroglyphs
0.04602233	research
0.03516577	archaeology
0.02930480	pottery

δ_4 = 0.30822165

– 4 concepts modeled from 8 feedback documents

Concept representation (2)

- word distributions over topics are learned by LDA...
  - $p(w|k) = \phi_{k,w}$
- as well as topic distributions over documents
  - $p(k|d) = \theta_{d,k}$
- concept weighting

$$\delta_k = \sum_{d \in \mathcal{R}_Q} p(d|Q)p(k|d)$$

Outline

- Introduction
- Query-oriented topic modeling
- General sources of information and Ranking
- Results
- Conclusions and future work

General sources of information

- improve diversity by varying the sources of feedback documents

- sensitive to vocabulary mismatch
  - but we assume that the ClueWeb09 is big enough to contain words that occur in all sources

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

Setup

- Language modeling approach to IR
  - Dirichlet smoothing ($\mu = 1500$)

- all runs on ClueWeb09 cat. A
  - indexed with Indri, Krovetz stemmer, INQUERY stoplist
  - removed spammed documents (percentile < 70) using University of Waterloo's spam list

Document ranking & Runs

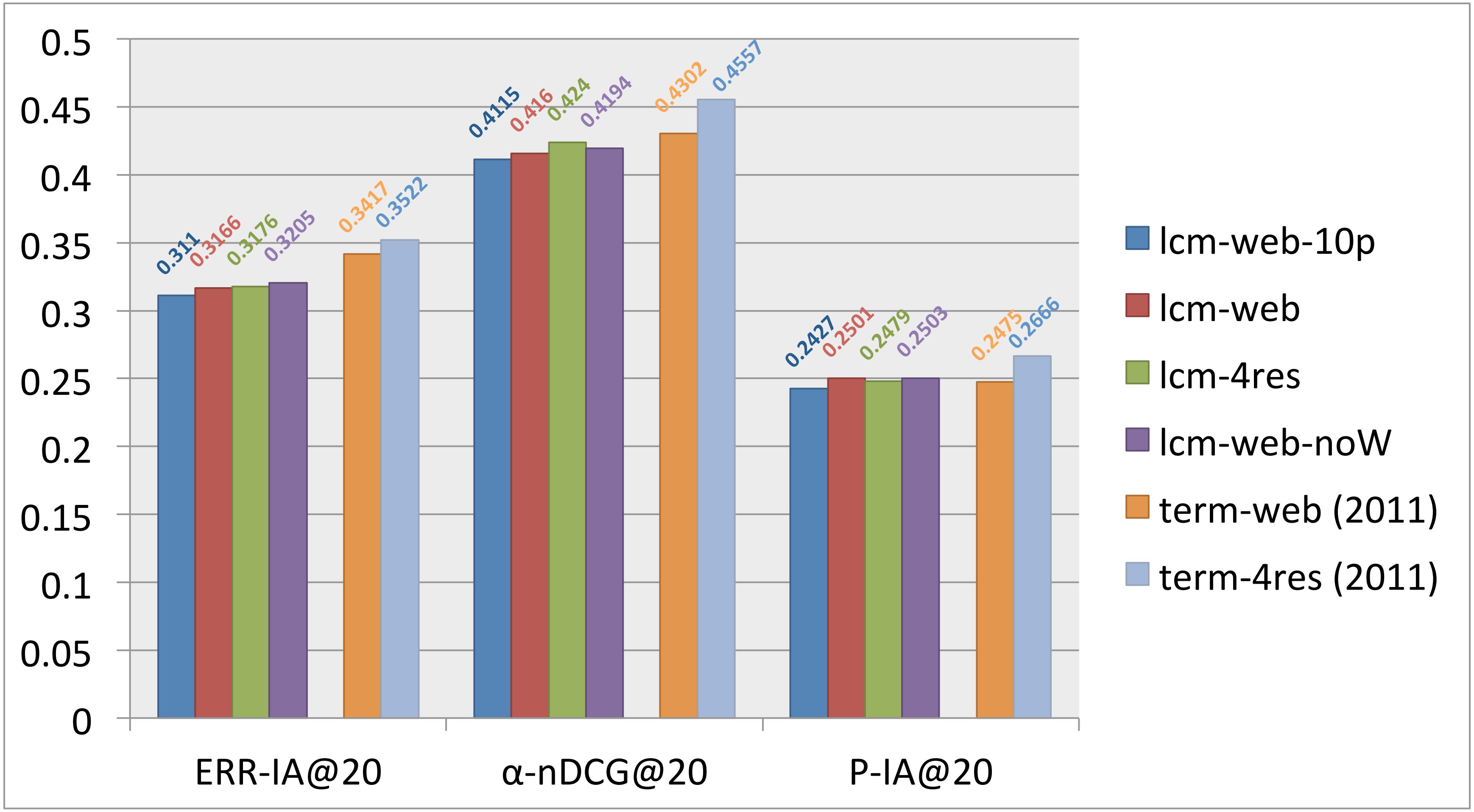
$$s(Q,d) = P(d|Q) + \frac{1}{|\mathcal{S}|} \sum_{\sigma \in \mathcal{S}} \sum_{k \in \mathbb{T}_{\hat{K},M}(\sigma)} \hat{\delta}_k \sum_{w \in \mathbb{W}_k} \hat{\phi}_{k,w} \cdot P(w|d)$$

- 4 runs
  - **lcm-web**: basic concept modeling run, uses only the web source
  - **lcm-web-10p**: same but M is fixed to 10
  - **lcm-web-noW**: same than first, without concept weighting
  - **lcm-4res**: basic concept modeling run, uses concepts from all 4 sources

Outline

- Introduction
- Query-oriented topic modeling
- General sources of information
- Results
- Conclusions and future work

Diversity



0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0.311
0.3166
0.3176
0.3205
0.3417
0.3522
0.4115
0.416
0.424
0.4194
0.4302
0.4557
0.2427
0.2501
0.2479
0.2503
0.2475
0.2666
ERR-IA@20
α-nDCG@20
P-IA@20
lcm-web-10p
lcm-web
lcm-4res
lcm-web-noW
term-web (2011)
term-4res (2011)

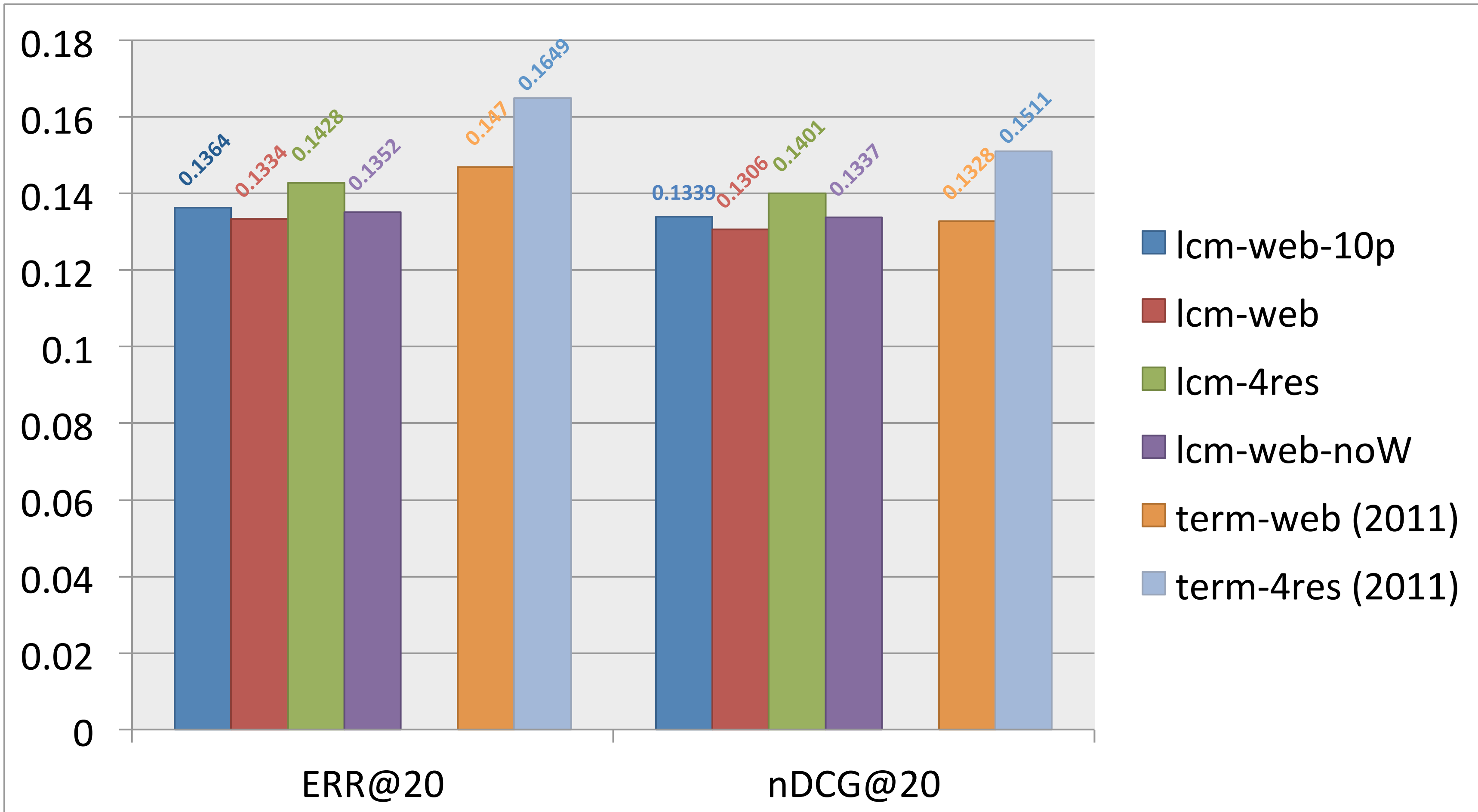
Diversity (2)

- all runs perform roughly the same on average
  - but using the 4 sources reduces the rate of failure
  - but still 1 null topic (0 on all metrics)

- all runs around median

- outperformed by our last year approach
  - single term query expansion
  - although with no statistically significant differences

Ad Hoc



0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0.1364
0.1334
0.1428
0.1352
0.147
0.1649
0.1339
0.1306
0.1401
0.1337
0.1328
0.1511
ERR@20
nDCG@20
lcm-web-10p
lcm-web
lcm-4res
lcm-web-noW
term-web (2011)
term-4res (2011)

Ad Hoc (2)

- all runs significantly better than (unofficial) MRF-IR [Metzler, SIGIR'05]

- estimating the number of feedback documents does not seem to important
  - w.r.t to setting M = 10
  - consistent with results reported by [He, CIKM'09]

Examples of successes

```
<topic number="194" type="faceted">
  <query> designer dog breeds </query>
  <description> What breeds of small or toy dog hybrids are there? </description>
  <subtopic number="1" type="inf"> What breeds of small or toy dog hybrids are there? </subtopic>
  <subtopic number="2" type="inf"> Find puppies of designer dog breeds for sale. </subtopic>
  <subtopic number="3" type="inf"> Find pictures of various designer dog breeds. </subtopic>
</topic>
```

- hard topic

- concepts/entities about
  - purebred dog
  - genetics (characteristics, traits)
  - canine reproduction

Examples of successes (2)

```
<topic number="174" type="ambiguous">
  <query> rock art </query>
  <description> Where can I learn about rock painting or buy a rock-painting kit? </description>
  <subtopic number="1" type="inf"> Where can I learn about rock painting or buy a rock-painting kit? </subtopic>
  <subtopic number="2" type="nav"> Where can I buy tools for stone carving or engraving? </subtopic>
  <subtopic number="3" type="inf"> Find information on cave paintings in France. </subtopic>
  <subtopic number="4" type="nav"> Where can I buy rock and roll posters? </subtopic>
  <subtopic number="5" type="inf"> Find information on the artwork used on rock music album covers. </subtopic>
</topic>
```

- medium topic

- concepts/entities about
  - prehistoric art
  - twykelfontein
  - experimental music

Examples of failures

```xml
<topic number="183" type="faceted">
  <query> kansas city mo </query>
  <description> What are some Kansas City, MO tourist attractions? </description>
  <subtopic number="1" type="inf"> What are some Kansas City, MO tourist attractions? </subtopic>
  <subtopic number="2" type="inf"> What hotels are near the Kansas City airport? </subtopic>
  <subtopic number="3" type="nav"> Find the Kansas City Chiefs homepage. </subtopic>
  <subtopic number="4" type="inf"> What casinos are in Kansas City, Missouri? </subtopic>
  <subtopic number="5" type="inf"> Find information on the Hallmark Visitors Center in Kansas City, MO. </subtopic>
</topic>
```

- medium topic (0 for all runs, 4res imroved a bit but results stay below median)

- concepts only are about Kansas City, Missouri

- only 1 feedback document was selected

Examples of failures (2)

```
<topic number="154" type="faceted">
  <query> figs </query>
  <description> Find information on nutritional or health benefits of figs. </description>
  <subtopic number="1" type="inf"> Find information on nutritional or health benefits of figs. </subtopic>
  <subtopic number="2" type="nav"> Find recipes that use figs. </subtopic>
  <subtopic number="3" type="inf"> Find information on the different varieties of figs. </subtopic>
  <subtopic number="4" type="inf"> Find information on growing figs. </subtopic>
</topic>
```

- easy topic for all participants
- concepts are all about Ficus (Figs tree)
- again only 1 feedback document was used to model the concepts

Examples of failures (3)

- feedback documents selection seems to be essential (more than the number of topics)
  - it needs more exploration though

- already encountered in the INEX Tweet Contextualization track

Outline

- Introduction
- Query-oriented topic modeling
- General sources of information
- Results
- Conclusions and future work

Conclusions and future work

- tried an unsupervised method for latent search concepts identification
  - weighted bags of weighted words
  - incorporation into the document ranking function
- lots of things to sort out
  - feedback documents selection
  - trade-off between query and concepts

Conclusions and future work (2)

- provide human-readable feedback for better query refinement/rewriting
  - displaying concepts, entities, facets...

- prediction of query intents or subtopics thank you for your attention

questions?