Query-biased Partitioning for Selective Search

Zhuyun Dai, Chenyan Xiong, Jamie Callan
Carnegie Mellon University
Language Technologies Institute

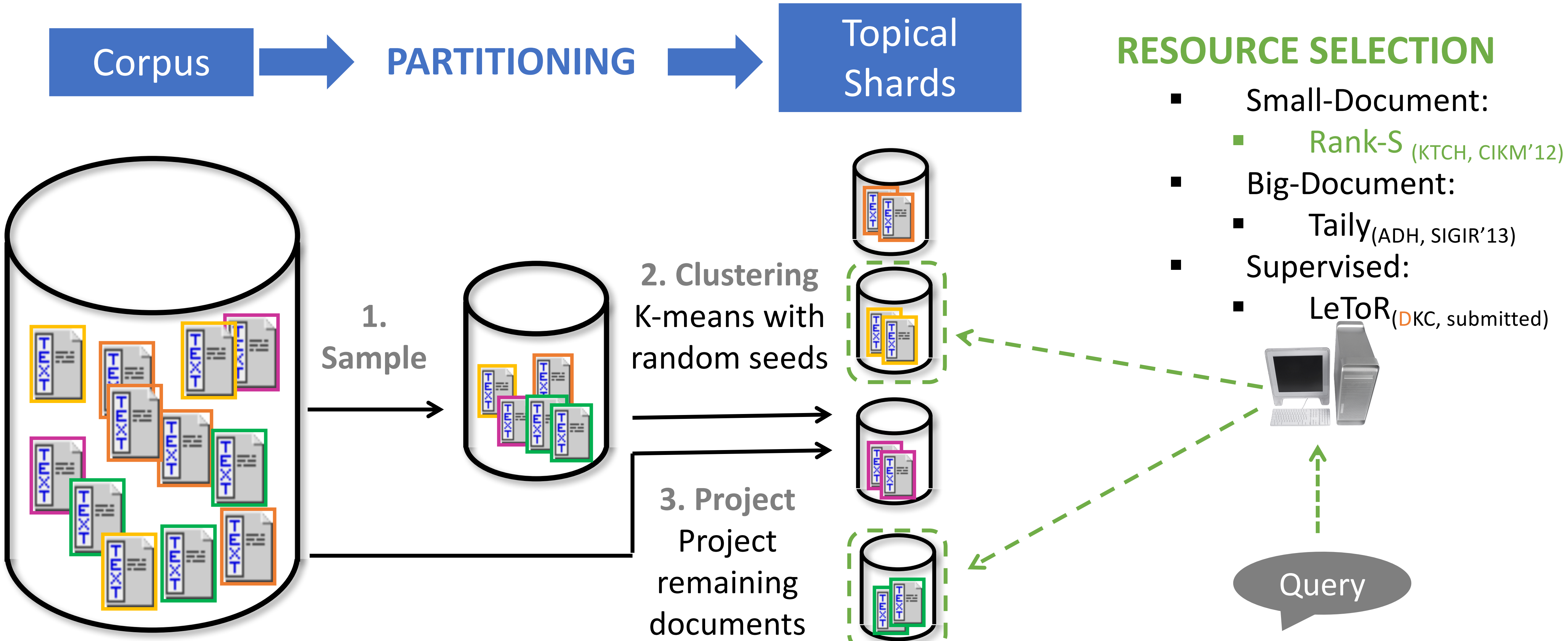
Outline

- **Background – Selective Search**
- Proposed Methods
  - Query-driven clustering initialization
  - Query-biased similarity metric
- Experiments & Analysis
- Conclusions

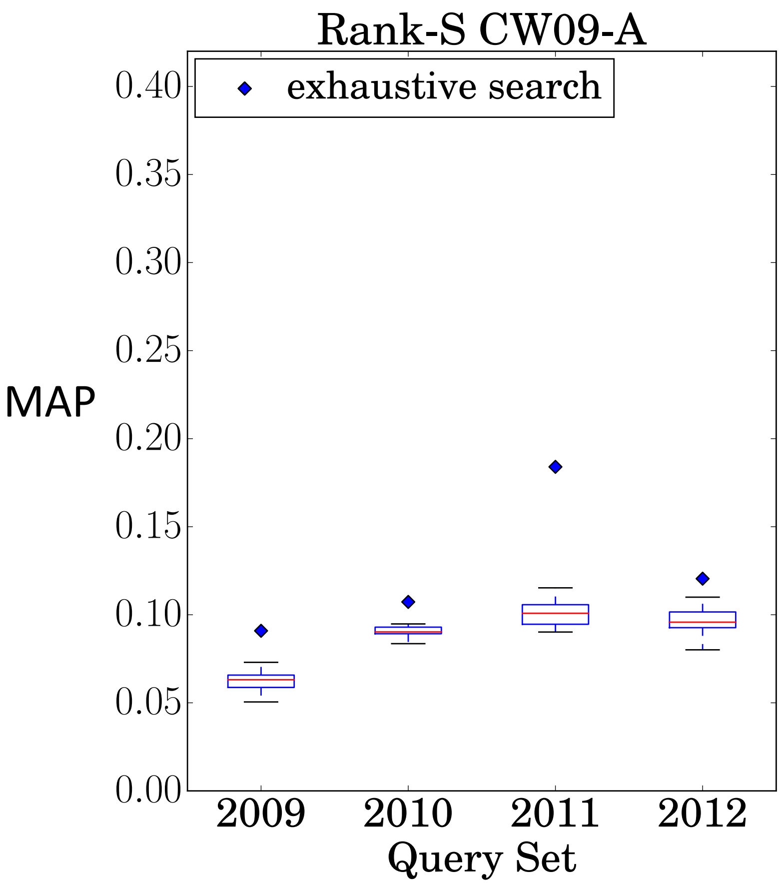
Selective Search

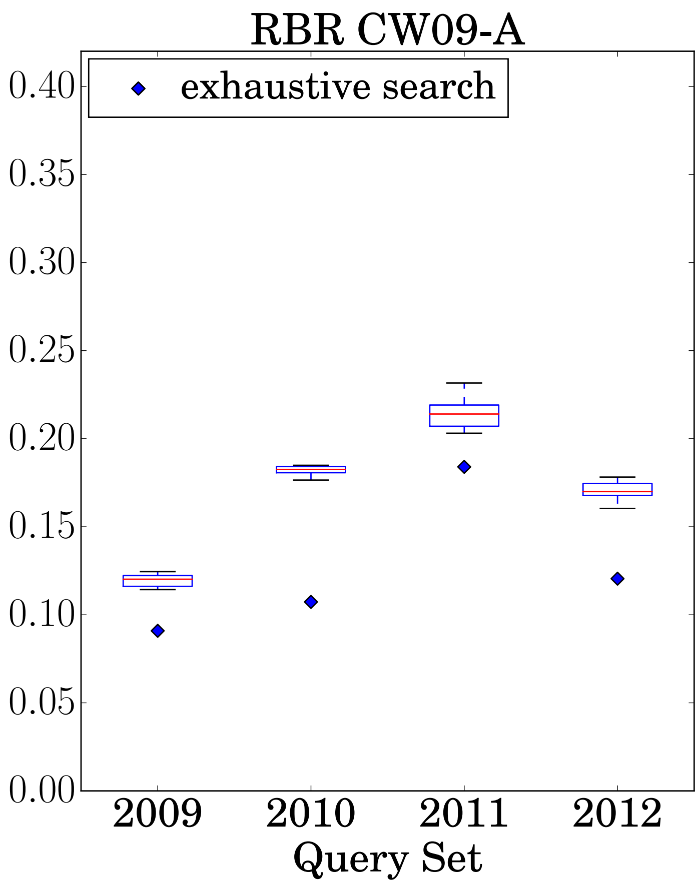
- Traditional Distributed Search
  - A document corpus => small random shards
  - Searched all shards in parallel
  - Merge results
  - Exhaustive Search
- Selective Search
  - A document corpus => topical shards.
  - The query is run against only a few shards.
  - Goal: same search accuracy as exhaustive search, but much faster

Selective Search Pipeline



Error Analysis (DKC, SIGIR'15)





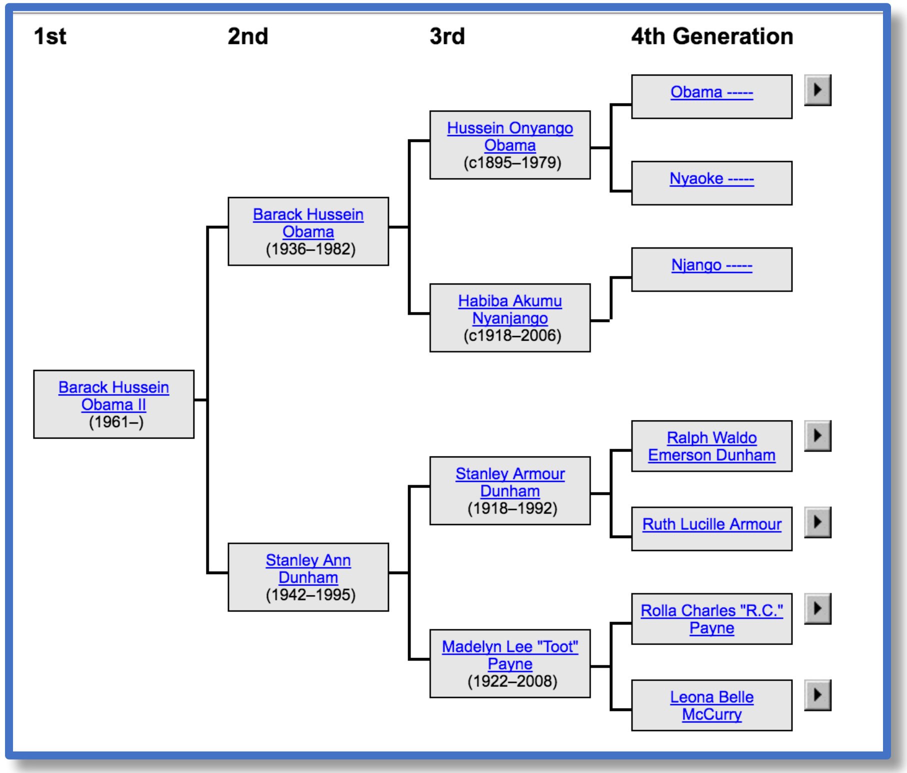
1. System Variance: from the clustering process.
2. Lower search effectiveness (MAP) than exhaustive if using a real resource selection algorithm.

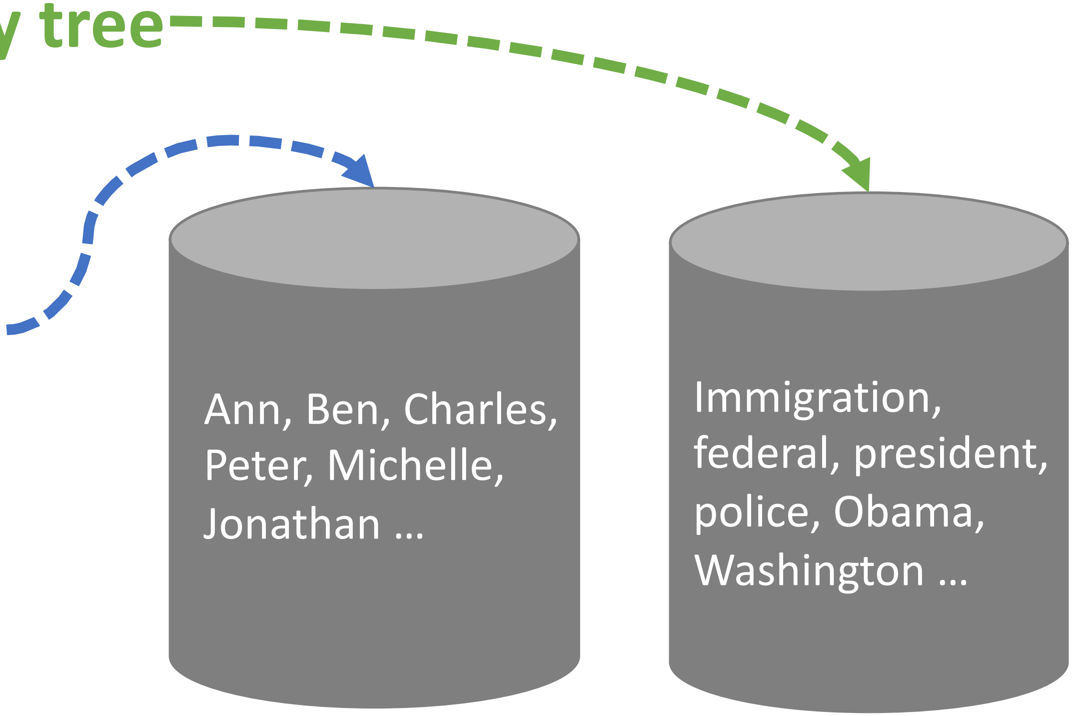
Rank-S: **Real** Resource Selection

RBR: **Oracle** Resource Selection

Why does resource selection select the wrong shards?

- Problem: The topics generated by the content-based partitioning do not match the topics searched by the users.
- Example: Query **Obama family tree**





‘People Names’ shard ‘U.S. Politics’ shard

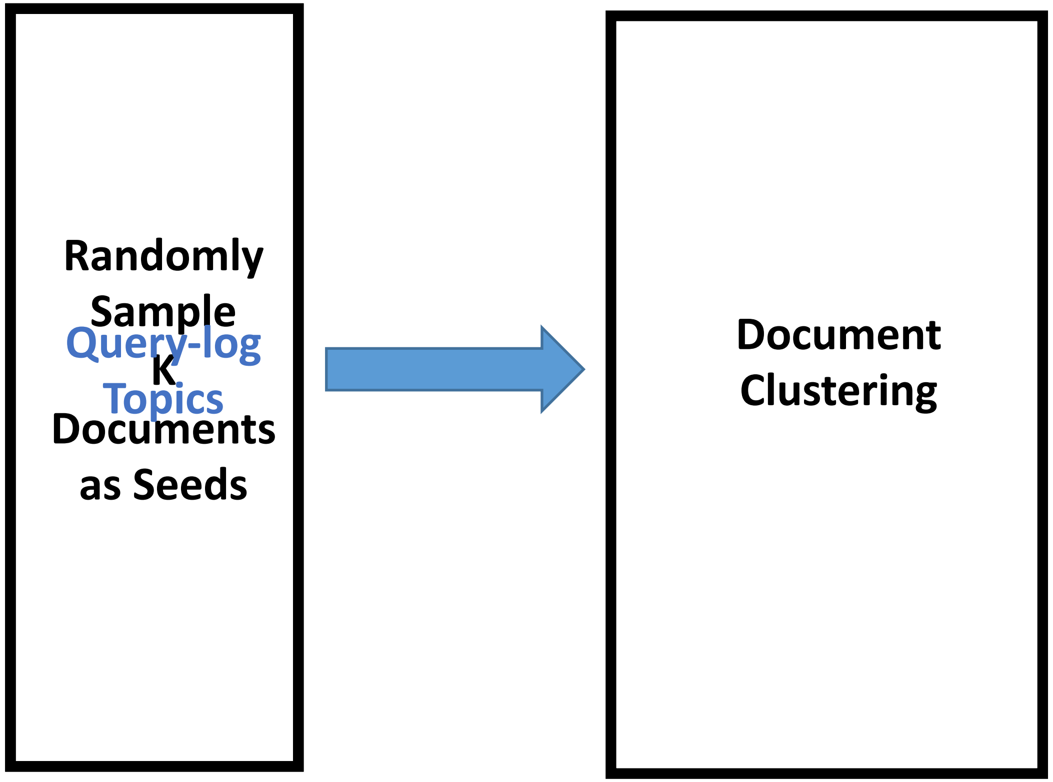
Content-based Partitioning: Topic Mismatch

- How to group together documents that satisfy the same user intent?
  - Query Logs!
- This work investigates aligning document partitioning with topics from the query logs.

Outline

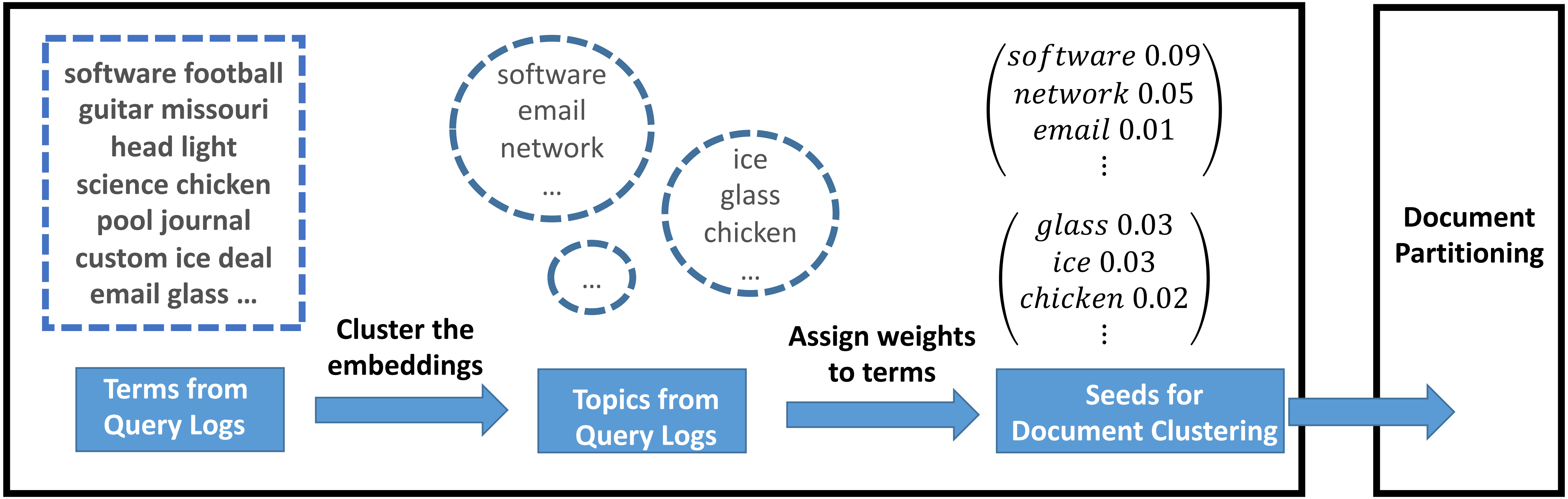
- Background
- Proposed Methods
  - **Query-driven clustering initialization**
  - Query-biased similarity metric
- Experiments & Analysis
- Conclusions

QInit: Query-driven Clustering Initialization



QInit: Query-driven Clustering Initialization

- Start the document clustering process with query-log topics.



Term Weighting in QInit

QTF * IDF: $w_q(t) = \log(tf_{t,Q} + 1) \times \log\left(\frac{|D|}{df_{t,D}} + 1\right)$

- **Query-log TF**
- $tf_{t,Q}$: Term frequency in the query log
- Promotes the importance of terms frequently used by users in search
- Log function: term distribution in the query log is very skewed

- **Collection IDF**
- $df_{t,D}$: # of documents that contain the term in the collection.
- demotes terms that are too common in the corpus

Examples of shard generated with QInit

Dataset	Top weighted Terms in Initial Seed	Relevant Queries
CW09-B	wine, tea, coffee, smoking, alcohol, drink	Starbucks, quit smoking
	animal, cock, bird, wild, egg, cat	dinosaurs, Arizona game and fish, moths,...
Gov2	tax, revenue, loans, business, bank, taxation	reverse mortgages, timeshare resales, ...
	diabetes, autism, obesity, arthritis, hypertension, celiac	aspirin cancer prevention, embryonic stem cells, ...

Outline

- Background
- Proposed Methods
  - Query-driven clustering initialization
  - **Query-biased similarity metric**
- Experiments & Analysis
- Conclusions

QKLD: Query-biased Similarity Metric

- Bias the clustering (partitioning) towards important query log terms.

SNIPPET 1: Family of Obama
Barack Obama was raised by his **mother**, Stanley Ann **Dunham**, called Ann, and grandparents Madelyn and Stanley **Dunham**.

SNIPPET 2: Lena Dunham
Dunham was born in New York City. Her father, Carroll **Dunham**, is a painter, and her **mother**, Laurie Simmons, is an artist and photographer.

SNIPPET 3: Obama's Education Law
President **Barack Obama** signed into law legislation that replaces the landmark No Child Left Behind education law of 2002.

QKLD: Query-biased Similarity Metric

- Bias the clustering (partitioning) towards important query log terms.

SNIPPET 1: Family of Obama
Barack Obama was raised by his **mother**, Stanley Ann **Dunham**, called Ann, and grandparents Madelyn and Stanley **Dunham**.

SNIPPET 2: Lena Dunham
Dunham was born in New York City. Her father, Carroll **Dunham**, is a painter, and her **mother**, Laurie Simmons, is an artist and photographer.

SNIPPET 3: Obama's Education Law
President **Barack Obama** signed into law legislation that replaces the landmark No Child Left Behind education law of 2002.

QKLD: Query-biased Similarity Metric

- Previous state-of-art similarity metric in selective search

$$sim_{KLD}(d,c) = \sum_{t \in d \cap c} \boxed{s_{KLD}\left(\vec{d}_t, \vec{c}_t\right)} \qquad s_{KLD}(\vec{d}_t,\vec{c}_t) = p_c(t)\log\frac{p_d(t)}{\lambda p_B(t)} + p_d(t)\log\frac{p_c(t)}{\lambda p_B(t)}$$

- Re-weight each term's similarity contribution by their importance in the query log

$$sim_{QKLD}(\vec{d},\vec{c}) = \sum_{t \in d \cap c} \boxed{(w_q(t)+b)} \times \boxed{s_{KLD}\left(\vec{d}_t, \vec{c}_t\right)}$$

- Previous state-of-art similarity metric in selective search
- Re-weight each term's similarity contribution by their importance in the query log

QTF * IDF:

- b: >0, smoothing parameter. Balance query log and document content

- b: >0, smoothing parameter. Balance query log and document content

Outline

- Background
- Proposed Methods
  - Query-driven clustering initialization
  - Query-biased similarity metric
- **Experiments & Analysis**
- Conclusions

Datasets

Document Collection	ClueWeb09-B	Gov2
Documents	50,220K	25,205K
Test Query Set (with manual relevance judgements)	200 queries (TREC)	150 queries (TREC)

Query Logs	AOL-ALL	AOL-Gov2
Queries	24,189,556	540,285
Queries after filtering	13,950,463	403,610
%Terms (w/o numbers)	978,714	69,482
%Terms after filtering	80,963	14,018

- Word Embeddings: 300-d Google word2vec trained on the corpora.
- Gov2 results are not shown in this presentation. Similar to ClueWeb09-B.

Baseline & Proposed Methods

- Partitioning methods:
  - **KLD-Rand (baseline)**
  - QKLD-Rand
  - KLD-QInit
  - QKLD-Qinit
- K (Number of clusters):
  - CW09-B: 100, Gov2: 150
  - Split big shards with another level of clustering
- Partition 10 times -> 10 different system instances
  - rule out random effects
  - evaluate system variance

Experiments

- Effectiveness: How does our method affect the clustering? Can it improve search effectiveness?
- Robustness: Is the method robust to query logs?
- Efficiency: Does it change the efficiency of the system?

Experiment 1: Clustering Analysis

- Are the query's relevant documents concentrated in a few shards?
  - Easier for resource selection algorithm to find the right shard
  - Higher recall with fewer shards searched
- Metric: coverage
- Coverage of query q:
  - sorting shards by the number of relevance documents they contain
  - The percentage of relevance documents covered by the first t% shards

$$coverage_t(q) = \frac{\sum_{i=1}^{floor(N*t\%)} R^q_{s^q_i}}{R^q}$$

- Coverage of the query set:

$$coverage_t = \frac{\sum_{q=1}^{|Q|} coverage_t(q)}{|Q|}$$

Experiment 1: Clustering Analysis (Cont.)

* indicates statistically significant difference with KLD-Rand

Dataset	Method	Percentage of Shards (t)			
		1%	3%	5%	10%
CW09-B	*KLD-Rand*	*0.60*	*0.86*	*0.96*	*0.99*
	KLD-QInit	0.60	0.86	0.95	0.99
	QKLD-Rand	0.65*	0.89	0.97	0.99
	QKLD-QInit	**0.67***	**0.90***	**0.97***	**1.00**

- Similarity metric: QKLD > KLD
- Initialization: QInit > Rand when combined with QKLD
- Best: QKLD-QInit

Experiment 2:Search Effectiveness

- CW09-B Results:
  - *: statistically significant difference with KLD-Rand; **: statistically significant difference with QKLD-Rand

Metric	Mean				Standard Deviation (*10^-3)		
	Exhaustive	KLD-Rand	QKLD-Rand	QKLD-QInit	KLD -Rand	QKLD -Rand	QKLD -QInit
P@10	0.253	0.275	0.284* (+3%)	**0.290** (+5%)**	7.50	9.74	**6.58**
NDCG@100	0.286	0.254	0.273* (+7%)	**0.279* (+10%)**	9.92	5.39	**5.07**
MAP@1000	0.186	0.155	0.172* (+11%)	**0.178** (+15%)**	8.77	**3.75**	5.22

Experiment 2:Search Effectiveness (Cont.)

- Effects on Recall and Precision

  Relative gains over baseline at different document rankings

Gain of NDCG@Rank	CW09-B	
	QKLD -Rand	QKLD -QInit
10	3.77%	5.70%
30	5.49%	6.69%
100	7.70%	10.03%
500	9.44%	12.37%
1000	9.87%	13.26%

- Proposed methods improved recall
- Selective search rarely hurt Precision, sometimes even better
  - Filtering out false-positives
- Recall is harder to improve
  - Searching fewer shards will miss relevant documents in other shards
  - Important in re-ranking schema

Experiment 3: Search Robustness
Query Log Influences

- Robustness: Does temporal gaps between training queries and testing queries affect the proposed methods?

Experiment 3: Search Robustness Query Log Influences (Cont.)

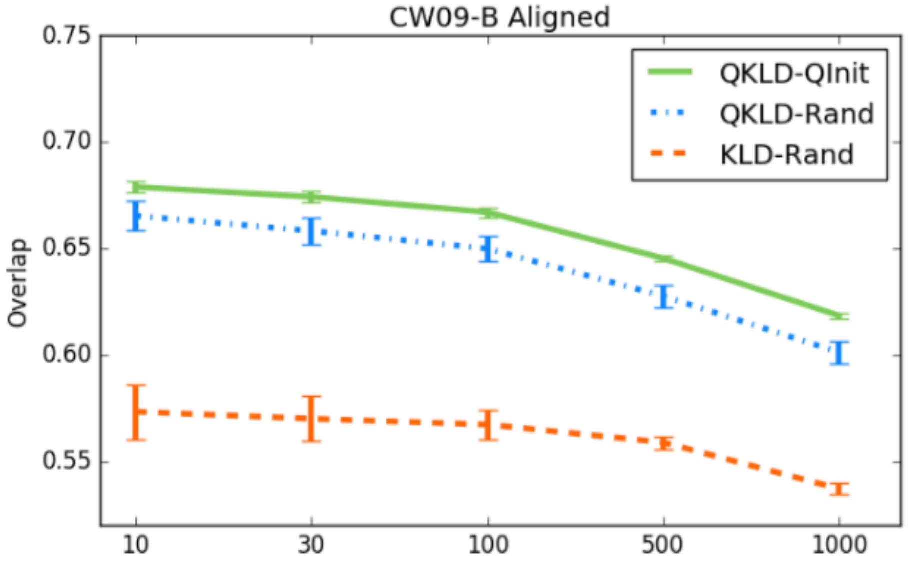
- Temporal Mismatch:
  - Training: AOL query logs (2006)
  - Testing: TREC queries (Gov2: 2004-2006; CW09-B: 2009-2012)
- Compare 2 temporal conditions

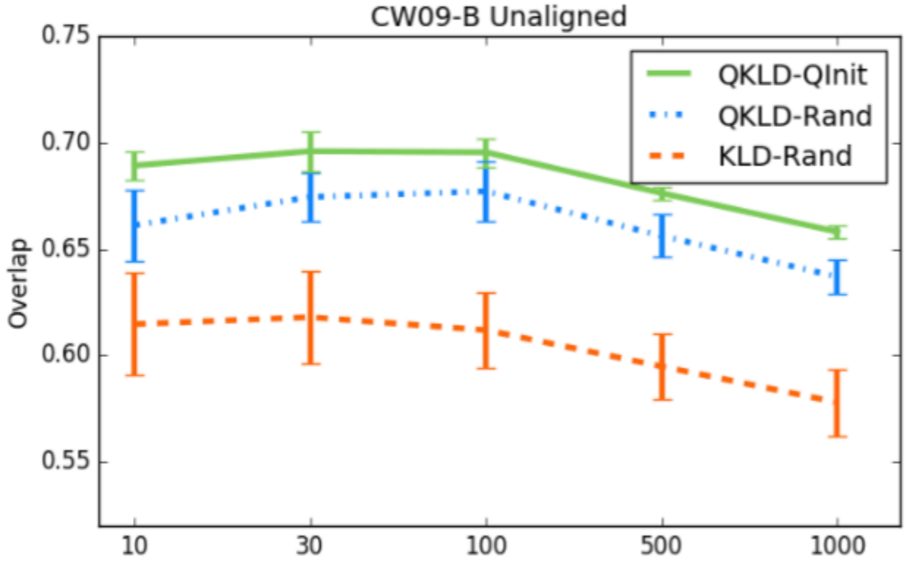
Condition	Training query log	Testing query set
Unaligned	AOL, first 2 months	TREC
Aligned	AOL, first 2 months	AOL, last 1 months

- Evaluation: overlap between exhaustive search and selective search (the high the better)

$$overlap_k = \frac{|D_k^{exh} \cap D_k^{sel}|}{k}$$

Experiment 3: Search Robustness
Query Log Influences (Cont.)





- Unaligned in general had higher overlap
  - TREC queries, less noisy
- Same trend:
  - QKLD-QInit > QKLD-Rand > KLD-Rand
- Same relative gain over baseline: difference is not statistically significant
- Not Sensitive to the temporal mismatch.

Experiment 4: Search Efficiency

- Efficiency: Whether query-biased partitioning changes selective search efficiency ?

Experiment 4: Search Efficiency

	CW09-B		Gov2	
	C_{RES}	C_{LAT}	C_{RES}	C_{LAT}
Exhaustive Search	5.24	0.33	2.89	0.18
KLD-Rand	0.53	0.24	0.29	0.11
QKLD-Rand	0.52	0.24	0.29	0.11
QKLD-QInit	0.52	0.23	0.28	0.11

- Metrics:
  - Total resource usage C_{RES}: $C_{RES}(q) = |D^q_{CSI}| + \sum^{T} |D^q_{S_i}|.$
  - Query Latency C_{LAT}: $C_{LAT}(q) = |D^q_{CSI}| + \max_{i=1}^{T_q} |D^q_{S_i}|.$
- Query-biased partitioning does NOT increase search cost.

Outline

- Background
- Proposed Methods
  - Query-driven clustering initialization
  - Query-biased similarity metric
- Experiments & Analysis
- **Conclusion**

Conclusion

- Proposed a query-biased partitioning strategy for selective search
  - Previous clustering: un-supervised.
  - Use query-logs as a weak supervision for the clustering.
- Evaluation & Analysis:
  - Improves search effectiveness and reduce variance:
    - Concentrates relevant documents together
  - Not sensitive to temporal difference between training & testing queries
    - Queries change over time, but the general topics are stable.
    - Do not need a perfect query log!

Thank You!

Q&A