Center for Intelligent Information Retrieval,
University of Massachusetts, Amherst

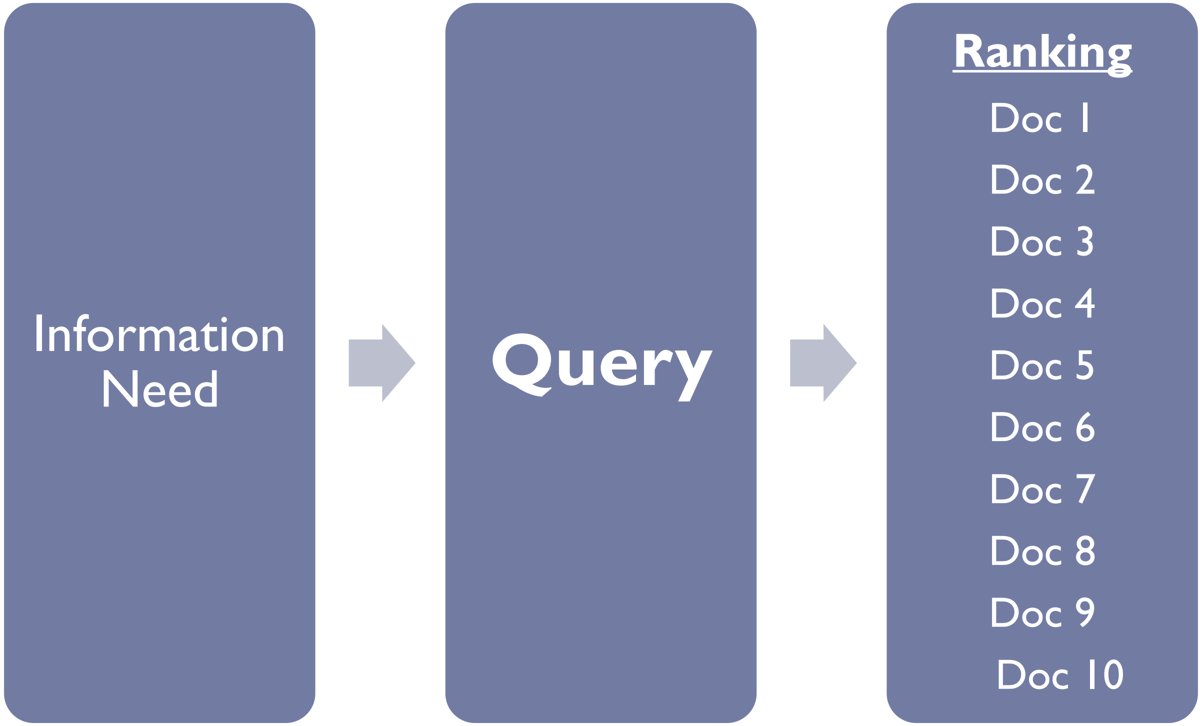
Discovering Key Concepts in Verbose Queries

Michael Bendersky, W. Bruce Croft

Outline

- Why long queries?
- Key concepts in long queries
- Key concepts identification
- Putting it all together: Retrieval with weighted structured queries

Ad hoc Information Retrieval



- Performance is measured based on
  - Explicit Relevance Judgments
    - e.g., ***mean average precision***
  - Implicit User Feedback
    - e.g., ***click patterns***

Introducing the problem

- Most research in ad hoc IR is focused on keyword queries
  - Sufficient for expressing simple information needs
  - Common in many domains, including web search
- In some domains long queries are more natural, as they can express more complex information needs
  - Q&A
  - Text reuse
  - Academic and enterprise search
  - Search-in-Context

What is a long query?

- Natural Language Queries
  - *ways in which the Federal Reserve conducts monetary policy*
  - *picture of Zephyr mythical figure that depicts wind blowing*
- Questions in Q&A archives
  - *What should I bring when traveling to Bolivia?*
- Queries with multiple keywords/noun phrases
  - *Jefferson Medical Center, Philadelphia, PA*
- "Copy-Paste" Queries
  - *required installation file could not be found SKU112.CAB*

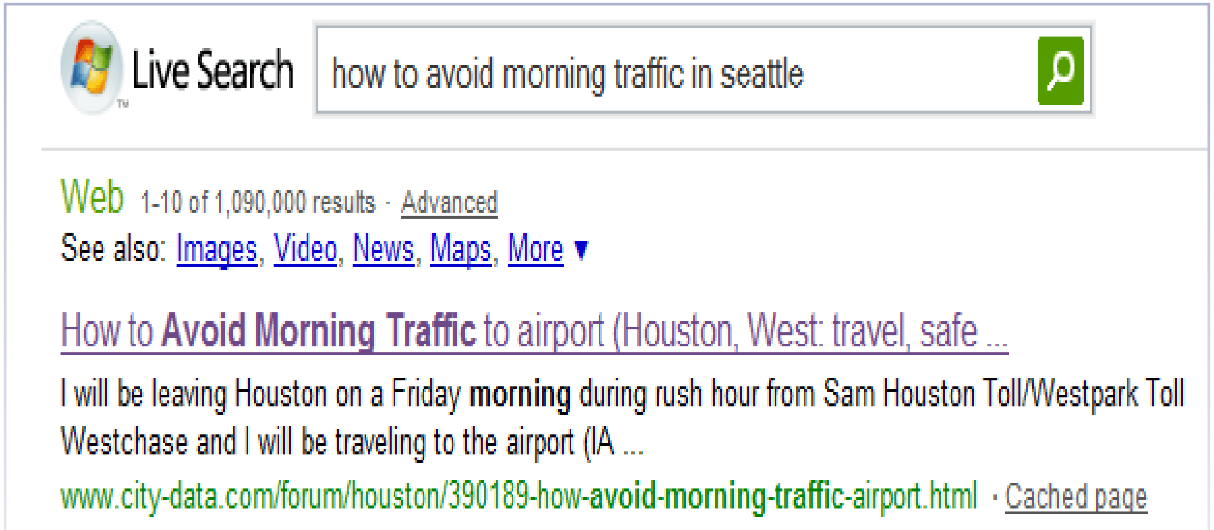
Do Long Queries Work?

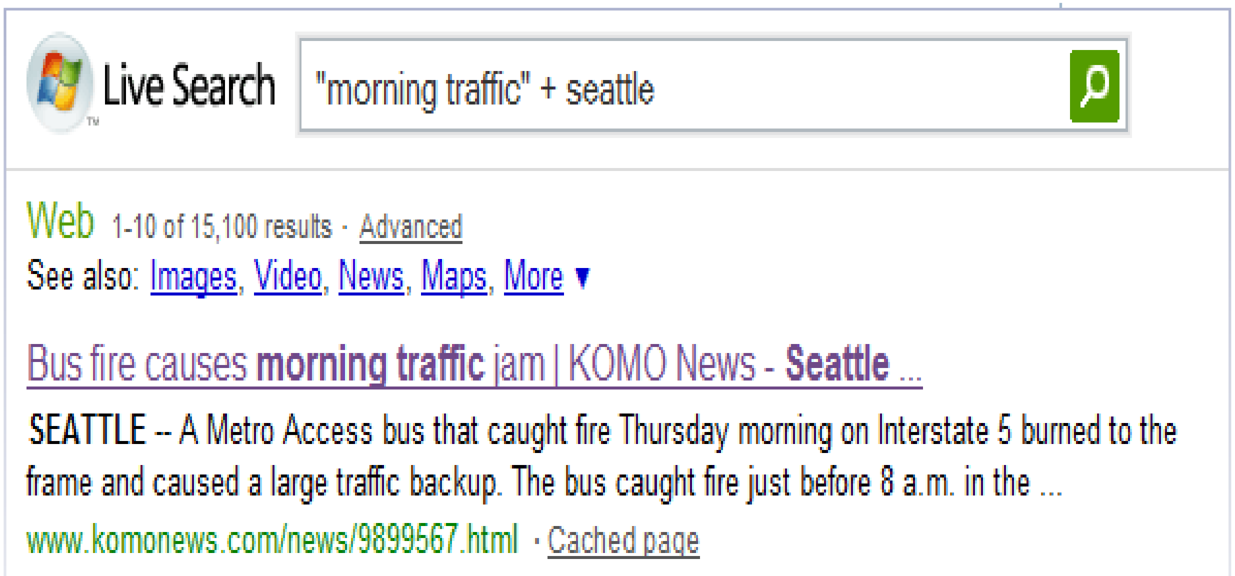
For people, yes; for search engines, no

- Unpredictable results with current web search engines
  - Sparser click-data
  - Often suffer from term mismatch
- TREC description queries don't work as well as title queries
  - More details follow
- Searching Q&A archives is not very effective

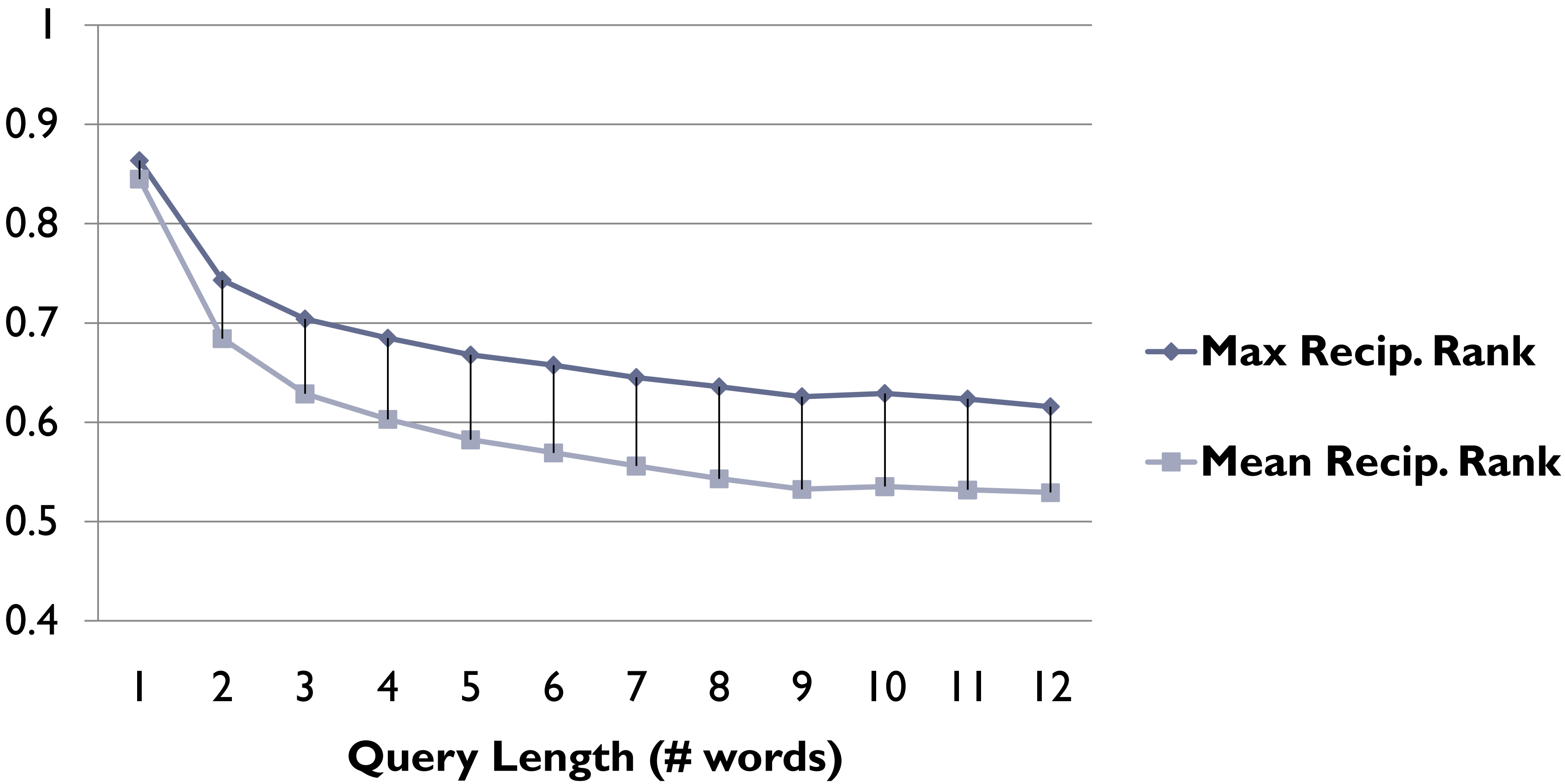
(Xue and Croft, 2008)

Long Queries on the Web





Long Queries on the Web – Click Patterns



Based on 15M query sample from a Live Search log

Long Queries on TREC

<title> *Spanish Civil War Support*

<desc> *Provide information on all kinds of material international support provided to either side in the Spanish Civil War*

(TREC Topic 829)

Text Retrieval Conference

"encourages research in information retrieval and related applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results"

http://trec.nist.gov/

CIIR

Long Queries on TREC – Avg. Precision

	ROBUST04		W10g		GOV2	
	MAP	w/q	MAP	w/q	MAP	w/q
<title>	25.3	2.7	19.3	4.2	29.7	3.1
<desc>	24.5	8.3	18.6	6.4	25.3	6.1

Mean Average Precision vs. Words Per Query

Past Work On Long Queries in TREC

- **(Allan et al., 1997; Callan et al., 1995)**
  - Improving performance of long TREC queries
- **(Murdock & Croft, 2005; Balasubramanian et. al. 2007)**
  - Sentence Retrieval
- **(Kumaran & Allan, 2006; Kumaran & Allan 2008)**
  - Interactive reduction/expansion of long queries

Hypothesis

Identification of the *key query concepts* will have a (significant) positive impact on the retrieval performance for verbose queries

Hypothesis Motivated

- Verbose queries tend to mix key (***Spanish Civil War***) and complementary (*material international support*) concepts
- Current retrieval techniques tend to treat these equally
- Potentially, this results in a loss of focus on the main query topic(s)

Concept Identification – *The Ideal*

- **Everything is a potential concept**

  (Bentivogli & Pianta, 2003)

  - Single words: *dog, cat*
  - Phrasal verbs: *catch up, come on*
  - Idioms: *break a leg, spend time*
  - Open compounds: *science fiction*
  - Named entities: *Spanish Civil War, Steve Jobs*
  - Free word combinations: *long queries*

CIIR

Noun Phrases as Concepts

- In this work, we approximate concept identification by noun-phrase extraction
- Reasonable approximation for the task at hand: nouns usually serve as query topics
- Works well in practice
- Used in a previous work involving key phrases extraction
  - *Allan et al. (1997) – Core concepts in TREC queries*
  - *Hulth (2003) – Keywords in scientific abstracts*
  - *Yih et. al (2006) – Keywords for web advertisement*

Back to Topic 829

Provide information on all kinds of material international support provided to either side in the Spanish Civil War

Concept Extraction

[information, kinds, material international support, side, Spanish Civil War]

Concept Weighting Principle [1]

- Not all concepts are equally important for the query
- Weigh concept c_i by $p(c_i \mid q)$
  - how well concept c_i represents query q.

Concept Weighting Principle [2]

- Either
  - Estimate $p(c_i \mid q)$ directly from the query
- Or
  - Leverage non-query specific information to estimate $p(c_i \mid q)$
- We choose the second option
  - Queries do not provide enough context
  - This is what we humans do

Concept Weighting Principle [3]

Assumption A

Each concept c_i can be assigned to one of the mutually exclusive classes

- ***KC*** *(key concepts class)*
- ***NKC*** *(non-key concepts class)*

Assumption B

A global function $h_k(c_i)$ indicates the confidence that concept c_i belongs to class ***KC***

Concept Weighting Principle [4]

- Following the assumptions, weigh each query concept

$$\hat{p}(c_i \mid q) = \frac{h_k(c_i)}{\sum_{c_i \in q} h_k(c_i)}$$

  - **That is, we *rank query concepts***
  - **Concepts which have the highest confidence in membership in class KC are regarded as the best query representatives**

Estimating $h_k(c_i)$

- As $h_k(c_i)$ is query-independent, we can
  a) Take an unsupervised approach to estimate it
    - e.g., use concept ***IDF***
  b) Try to learn it using a set of given concepts and features

- What kind of features?
  - As $h_k(c_i)$ is query-independent, we can use any concept related features

Query-Based Features

1. ***is_cap(c_i)*** Is concept capitalized in the query?
   - ***If TREC queries were not capitalized, we could resort to corpus-based capitalization***

CIIR

Collection-Based Features [1]

2. ***cf*(c_i)** Concept frequency in the collection

3. ***idf*(c_i)** Concept IDF in the collection

$$idf(c_i) = \log_2 \frac{N}{df(c_i)}$$

- ***N*** – number of documents in the collection
- ***df*(c_i)** – number of documents in the collection containing c_i

Collection-Based Features [3]

4. ***ridf(c_i)*** Concept residual IDF in the collection

- *Deviation of an actual IDF from Poisson model prediction*

  (Church & Gale, 1995)

$$ridf(c_i) = idf(c_i) - \log_2 \frac{1}{1 - e^{\theta_i}}$$

- θ_i – average number of occurrences of concept c_i per document

Collection-Based Features [4]

5. ***wig(c_i)*** Concept Weighted Information Gain

- *Information gain from a state where only average document is retrieved*

(Zhou & Croft, 2007)

$$wig(c_i) = \frac{\frac{1}{|T|}\sum_{d \in T} \log p(c_i|d) - \log p(c_i|C)}{-\log p(c_i|C)}$$

- ***T** – a set of top 50 documents retrieved from a collection in response to concept **c_i***

Collection-Independent Features

6. ***g_cf(c_i)*** Concept frequency in *Google n-grams*.
   - Estimates concept frequency in a large web collection

7. ***l_qp(c_i)*** Number of times a concept was used as a part of a web search query
   - Extracted from an excerpt of *MSN* search log

8. ***l_qe(c_i)*** Number of times a concept was used as an exact query
   - Extracted from an excerpt of *MSN* search log

Collections

Collection	# Docs	# Topics
ROBUST04	*528,155*	*250*
W10g	*1,692,096*	*100*
GOV2	*25,205,179*	*150*

Concept Classification: The Task

- **Task: identifying key concepts**
- Simplifying assumption: a single key concept per query
- Train an ***AdaBoost.M1*** classifier on a set of labelled concept instances: $\boldsymbol{x}_i \in \{\boldsymbol{KC}, \boldsymbol{NKC}\}$
- Rank concepts for each query in the test-set according to their confidence in membership in class ***KC***

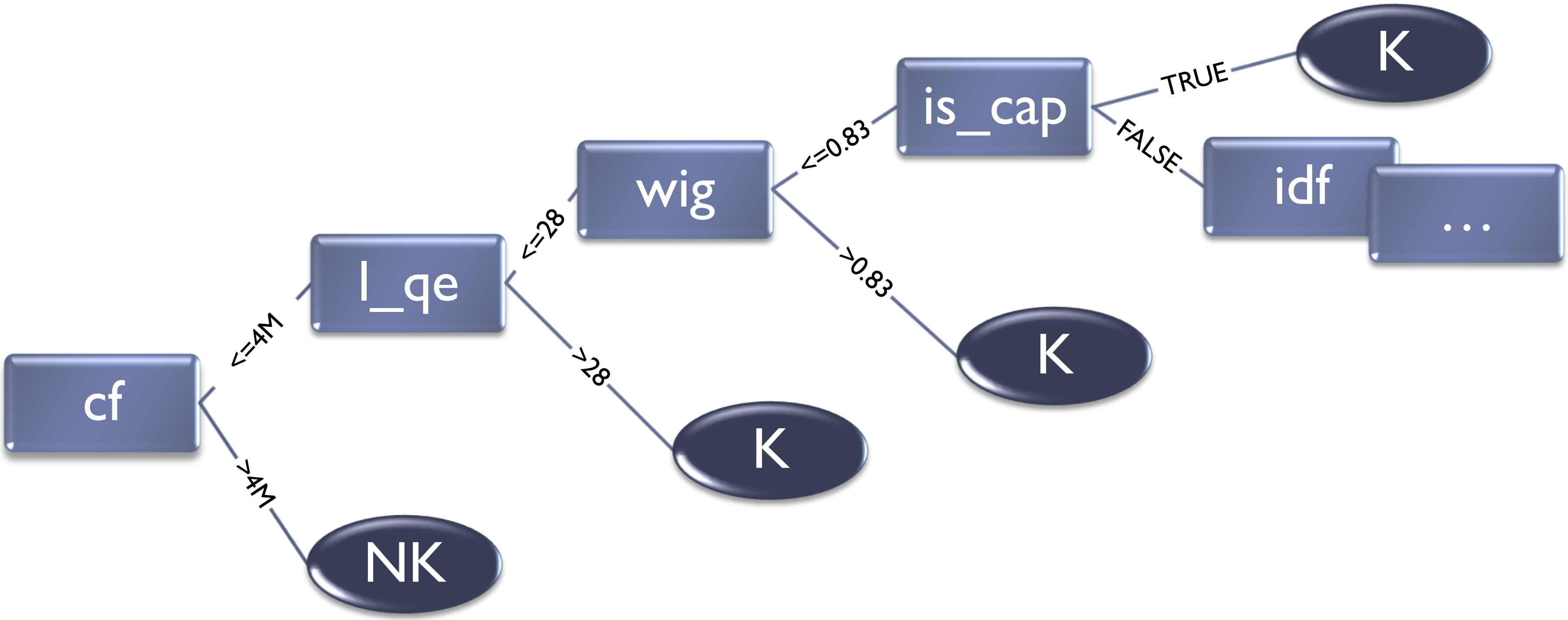
Concept Classification: Results

	AdaBoost		$idf(c_i)$	
	Accuracy	MRR	Accuracy	MRR
ROBUST04	***76.4***	***84.5***	*56.4*	*74.2*
W10g	***81.0***	***85.3***	*66.0*	*78.6*
GOV2	***84.0***	***88.9***	*74.7*	*85.7*

Accuracy and MRR results:

*3-fold cross-validation with **AdaBoost.M1** vs. **IDF***

What Makes a Key Concept?



<u>Example:</u>

A high-weight decision tree for key concept classification for GOV2

What About Retrieval?

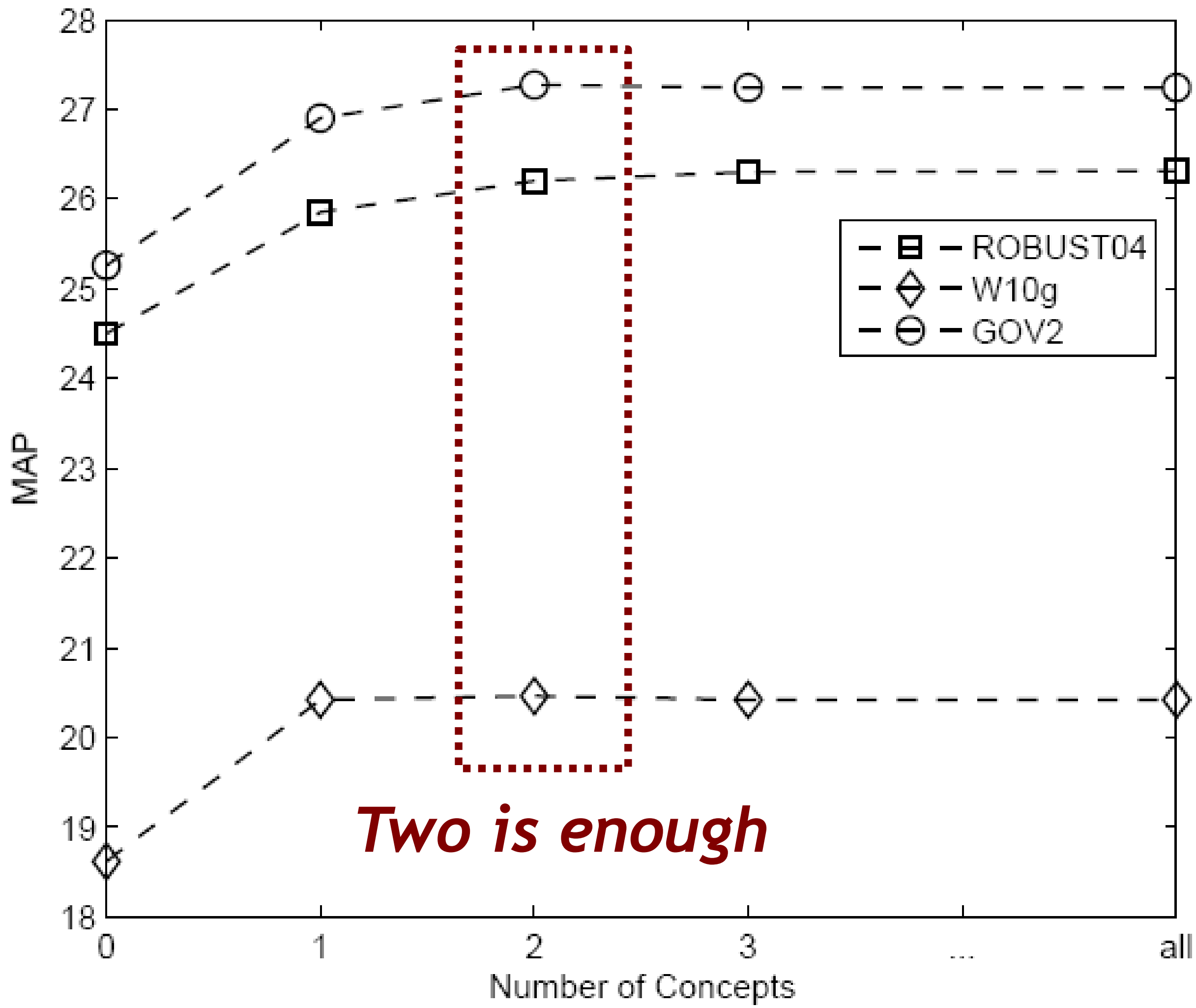
- Does identifying key concepts help at all?
- Does the concept weighting help?

Concept Weighting for Ranking

- Having estimated ***p(c_i|q)*** we may use a linear combination of query and all weighted concepts for ranking

$$rank(d) \propto \underbrace{\lambda \log p(q \mid d)}_{\text{Query Score}} + (1-\lambda)\sum_{c_i \in q} \underbrace{\log p(c_i \mid d)}_{\text{Concept Score}} \overbrace{p(c_i \mid q)}^{\text{Concept Weight}}$$

How Many Concepts?



Indri Query Language: Crash Course

- Simple bag-of-words query
  - **#combine** *(spanish civil war)*
- Phrase Operators
  - ***#1*** *(spanish civil war)*
  - ***#uw8*** *(spanish civil war)*
- Weights
  - ***#weight*** *(0.8* ***#combine****(spanish civil war)*
    0.1 ***#1****(spanish civil war)*
    0.1 ***#uw8****(spanish civil war))*

Indri – Open Source search engine

"The Indri query language… allows complex phrase matching, synonyms, weighted expressions, Boolean filtering, numeric fields, and the extensive use of document structure…"

http://www.lemurproject.org/

CIIR

<title> and <desc> - #COMBINE query

```
#combine( Spanish Civil War support )
```

```
#combine( information kinds material international support provided side Spanish Civil War )
```

<desc> - Sequential Dependence Model *(Metzler & Croft, 2005)*

```
#weight(
  0.85 #combine( information kinds material international support provided side Spanish Civil War )
  0.10 #combine( #1(information kinds) #1(kinds material) #1(material international)
                 #1(international support) #1(support provided) #1(provided side)
                 #1(side Spanish) #1(Spanish Civil) #1(Civil War) )
  0.05 #combine( #uw8(information kinds) #uw8(kinds material) #uw8(material international)
                 #uw8(international support) #uw8(support provided) #uw8(provided side)
                 #uw8(side Spanish) #uw8(Spanish Civil) #uw8(Civil War) ))
```

<desc> - Key Concepts Expanded

```
#weight(
  0.8 #combine( information kinds material international support
                provided side Spanish Civil War  )
  0.2 #weight( 0.99994 #combine ( Spanish Civil War )
               0.00006 #combine ( material international support )))
```

Retrieval Results [1]

	ROBUST04	W10g	GOV2
	MAP	*MAP*	*MAP*
<title>	25.28	19.31	**29.67**$_d$
<desc>	24.50	18.62	25.27^t
SeqDep<desc>	25.69$_d$	19.28	27.53^t_d
KeyConcept[2]<desc>	**26.20**$_d$	**20.46**^t_d	27.27^t_d

Comparison of methods performance (Mean Average Precision)

Retrieval Results [2]

	ROBUST04	W10g	GOV2
	MAP	*MAP*	*MAP*
<title>	25.28	19.31	**29.67**$_d$
<desc>	24.50	18.62	25.27^t
SeqDep<desc>	25.69$_d$	19.28	27.53^t_d
KeyConcept[2]<desc>	**26.20**$_d$	**20.46**^t_d	27.27^t_d

- **Query expansion by key concepts**
  a) always outperforms the original description queries
  b) comparable performance to ***SeqDep*** model
  c) more efficient than ***SeqDep*** model

Future (and Present) Work

- Finding text reuse on the web
  - Did I see this story somewhere else?

(Bendersky & Croft, To appear in WSDM 2009)

- "Learning to Rank" is an active field in IR
  - Can we reweight query terms based on available relevance judgments?

(Lease et al., To appear in ECIR 2009)

- Investigating long queries in web search logs
  - Insight into how and why people formulate long queries
  - Possibly leverage the insights into other domains

CIIR