Unsolved Problems in Search
(and how we approach them)

W. Bruce Croft
University of Massachusetts Amherst

Overview

- Search is everywhere
- Search works well…sometimes
- What doesn't work or doesn't exist
- Applications and testbeds
- An example research topic

The Ubiquitous Search Engine

- Most applications involve search, and many involve *search engines*
- Not just web search
  - desktop search
  - enterprise search
  - vertical search
  - social search
  - forum search
  - QA, FAQ and CQA
  - product search
  - entity and expert search
  - literature search
  - media search
  - database/XML search

The Ubiquitous Search Engine

- What do these search applications have in common?
  - goal of effective, efficient search
  - have to deal with text and other media with inexact semantics
    - noisy, inaccurate representations
    - noisy, inaccurate queries
    - "vocabulary mismatch"
  - ranking, statistics, and probabilities

Accomplishments of Search

- Major industry
  - search services, search companies, search and application startups
- Research activity very high
  - e.g. SIGIR, CIKM, WSDM, ECIR…
  - generally accepted evaluation methodology and measures
- Significant part of peoples' daily lives

Does Search Work?

- Depends on the tasks, goals, and measures of success
- Clearly current technology provides a useful level of performance
- If goal is to find web pages with relevant information in response to popular keyword queries, then effectiveness is high
  - according to, e.g., NDCG

Does Search Work?

- Experiments for more than 15 years in TREC have demonstrated significant improvements in many tasks, e.g.,
  - Ad-hoc search, routing and adaptive filtering, QA, home page and named page search, expert search
- Limited by availability of test collections, "realism" of task and evaluation

On the Other Hand..

- Current search technology is nothing like the "vision of the future" systems
  - e.g., Gray's "librarian" or HAL 9000
- On a more mundane level,
  - web search can be difficult
  - desktop search is mediocre
  - forum search is terrible
  - QA is very limited
  - proliferation of search services
  - and so on

Status

- There are many unsolved problems in search, both from a "science" and engineering perspective
  - i.e. many applications don't work all that well, and we don't understand the processes involved
  - but we are clearly making progress

The Big Picture

- No generally accepted "theory" of search
  - Some doubt about what would constitute such a theory
    - e.g., typically model documents, queries, topical relevance
    - what about users, information needs, interaction, tasks, context, structure, authority, other factors influencing relevance?
  - some consensus on probabilistic models, and feature-based models, but these "theories" are very limited

Evolutionary, not Revolutionary

- Every search application provides an environment for studying some aspect of the unsolved problems of search
- Progress is made by defining the problem being studied, building a testbed, and evaluating different approaches in the context of that application
  - e.g. TREC
- Hopefully integrate into the “big picture”

Typical Web Search Issues

- Scale
- Spam (or Adversarial IR)
- Advertising
- Coverage and freshness
- Evaluation
- Query processing
- Ranking algorithm

Studying Web Search

- Major engineering improvements
- But also significant contributions to the big picture
  - feature-based ranking
  - document structure
  - social aspect of relevance
  - understanding the user's intent
- More user data than any other application

Other Issues and Applications

- Long queries
  - web search, QA, CQA, enterprise search, vertical search, literature search
- Structure and heterogeneity
  - desktop search, enterprise search, database search
- Task and context
  - local search, exploratory search, social search, enterprise search

Testbeds

- Can't study a problem without a testbed with associated tasks and evaluation measures
- Can't wait for TREC to provide one
- Don't expect companies to provide one
  - but you may get lucky
- Some testbeds are difficult to produce (e.g. web search), others require some ingenuity
  - and some slack from reviewers!

Example: Desktop Search

- Goal is to have a testbed that can be distributed without privacy concerns
  - needs to have multiple types of “documents” with related content, appropriate queries
- Options:
  - Real desktop environments
  - Existing collections, e.g. TREC Enterprise, INEX Heterogeneous, Wikipedia
  - Simulated desktops

Example: Long Queries

- What is a long query?
- Why are they interesting?
- Looking at parts of the problem
  - Finding key concepts
  - Finding similar questions
  - Text reuse
  - Prior art search

What is a Long Query?

- TREC description query
  - e.g. *"Provide information on all kinds of material international support provided to either side in the Spanish Civil War."*
- Questions from users in Q&A services
  - e.g. *"Where can I complain about my wedding photographer?"*

What is a Long Query?

- Queries with more than one keyword or phrase from Web logs
  - e.g. *"lessons about kids in the bible", "best time of the year to visit bolivia"*
- Whole sentences or passages from documents
  - e.g. *"Process for the preparation of a zeolitic catalyst which comprises treating a zeolite of the Y-type having an alkali metal oxide/aluminium oxide molar ratio of at most 0.13 with a solution of a multi-valent metal salt having a cationic radius between 0.6 and 1.0 angstrom and combining the ion-exchanged zeolite without a calcination step with a hydrogenation component of a Group 8 and/or Group 6b metal."*

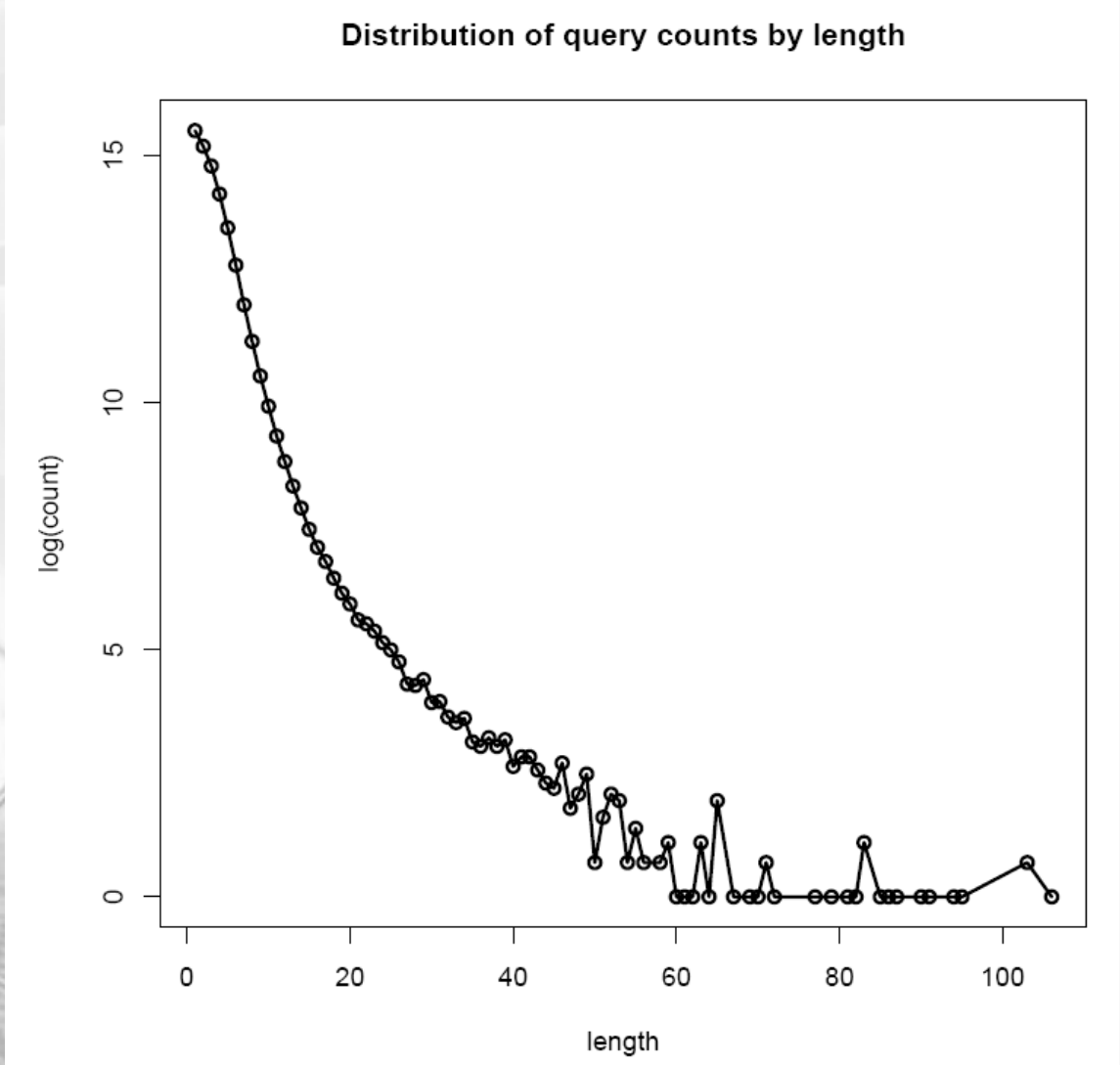
Characteristics of a Long Query

- Length (duh!)
  - Average length of Q&A questions more than 20 words and about 9 words for FAQs from Web
  - TREC descriptions are 14-20 words average vs. 2.5-5 words for title
- Grammar
  - Long queries tend to be more grammatical, sometimes full sentences
  - But, from a Q&A log:
    - "which airplne of the world has been fly the longest"
    - "who the first one fly to the spase"

Characteristics of a Long Query

- Frequency
  - Duplicates of long queries are generally rare
  - So, long queries are part of the "long tail"
  - Near-duplicates or semantically similar queries somewhat more common
- Information need
  - More complex information needs?
    - or maybe a better expression of real information needs than keywords
  - Not homepage/navigational searches

MSN Query Log



- Queries of length 4 or less account for 90.3%
- Average query length is 2.4

MSN Query Log

- Long query types
  - *Questions* (e.g., wh-)
  - *Operators* (contains query language operators)
  - *Composite* (made up of short queries)
  - *Non-Composite* (noun phrases and sentences)
  - *Exact quotes*

MSN Query Log

Total Queries: $14,921,286$		
Long Queries $(5 \leq l(q) \leq 12)$ **:** $1,423,664$		
Type	Count	% of Long
Questions	106,587	7.49
Operators	78,331	5.50
Composite	918,482	64.52
Non-Composite	320,263	22.50
Noun-Phrases	*204,823*	*14.39*
Pseudo-Sentences	*115,440*	*8.11*
Very Long Queries $(l(q) > 12)$ **:** $13,835$		
Type	Count	% of Very Long
Questions	2,941	21.26
Operators	1,964	14.20
Quotes	1,469	10.62
Pseudo-Sentences	5,945	42.97
Others	1,516	10.96

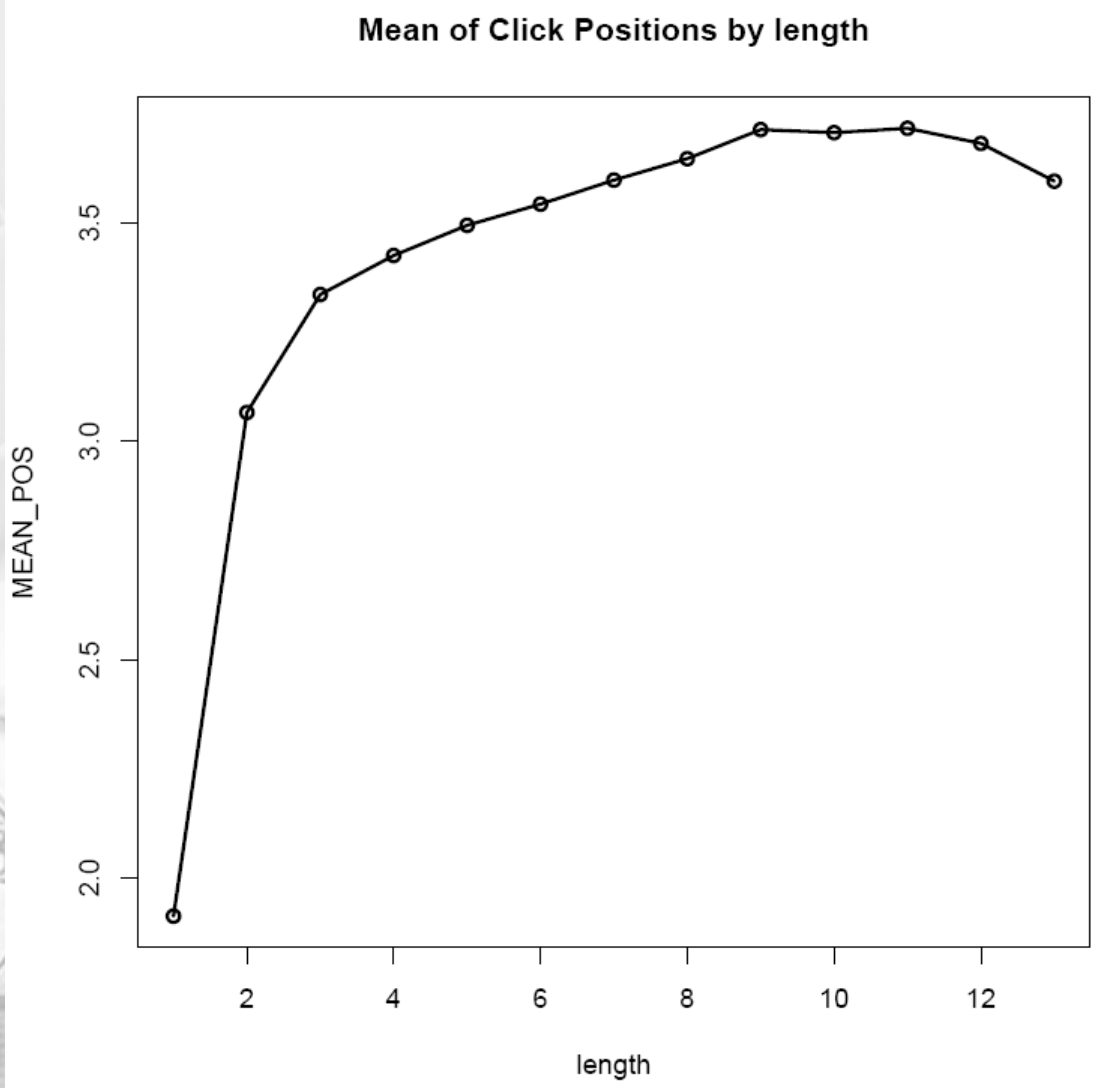
Why Long Queries?

- Natural for some applications
  - e.g., Q&A, text reuse, professional/scholar
- May be the best way of expressing most information needs
  - i.e., perhaps selecting keywords is what is difficult for people
- Next step towards the goal of the "vision of the future" search engine

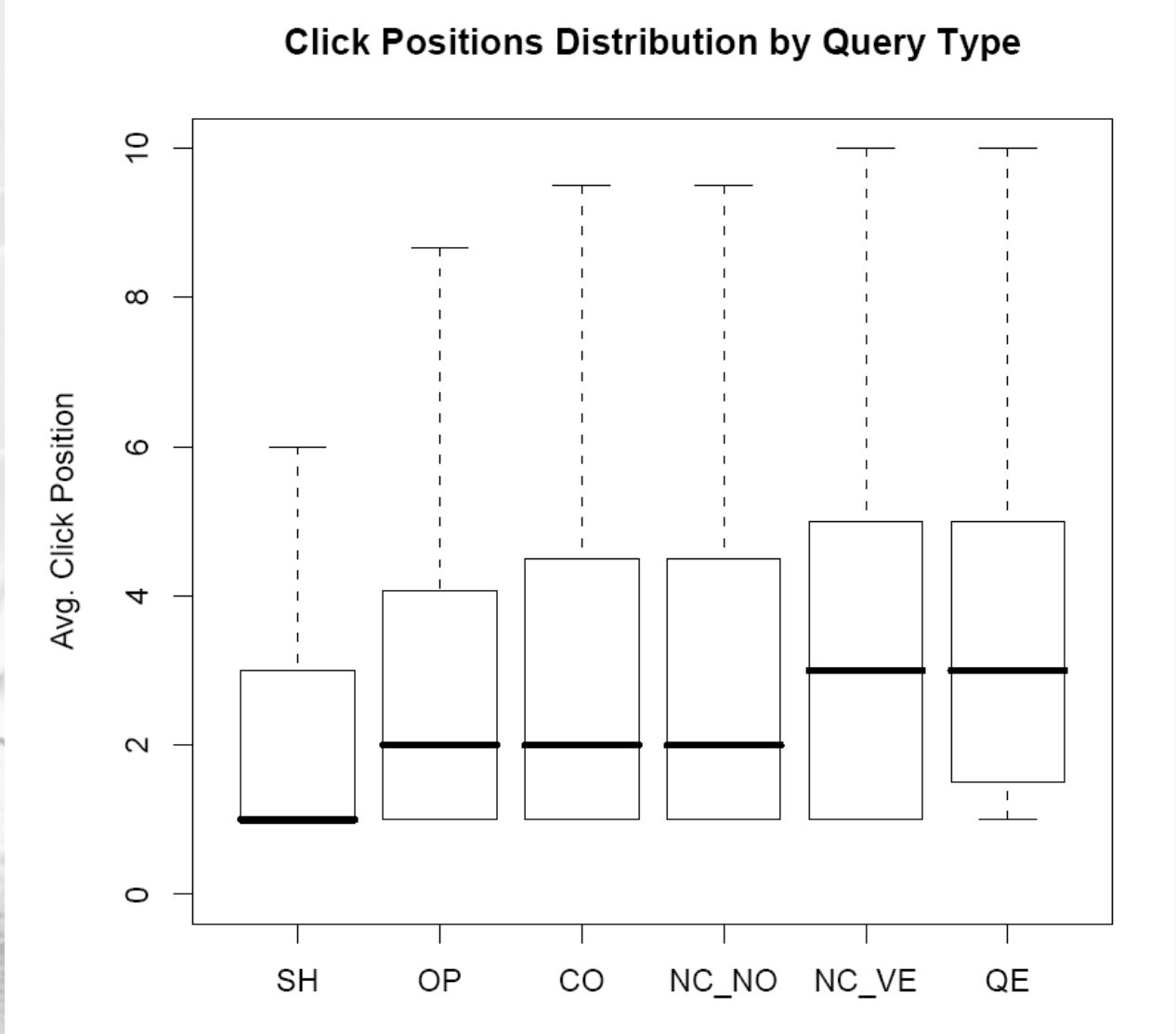
Do Long Queries Work?

- For people, yes; for search engines, no
- Long queries give generally poor, unpredictable results with current Web search engines
- TREC description queries don't work as well as title queries
- QA techniques don't work well for more general questions

MSN Query Log



MSN Query Log



Approaches to Long Queries

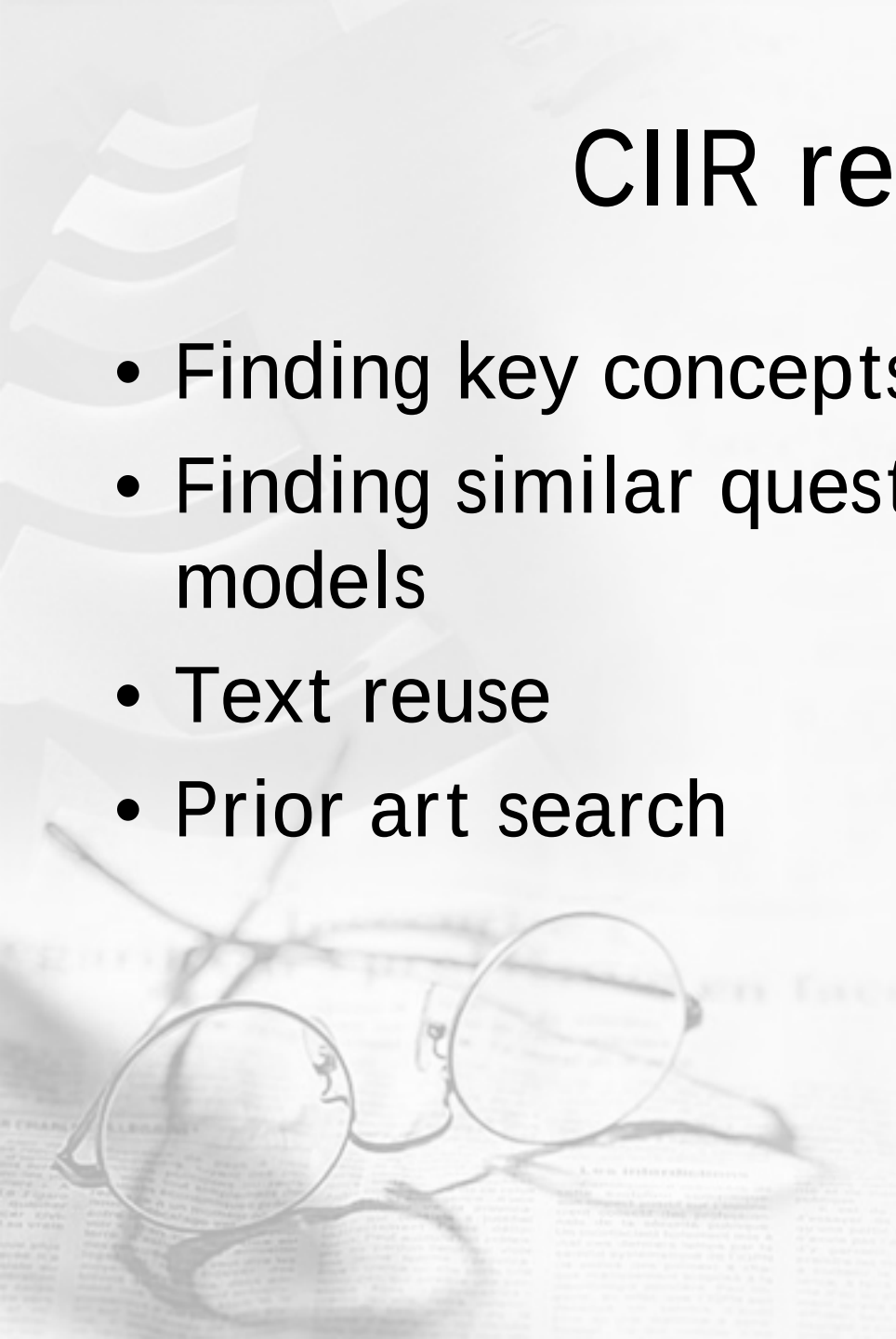
- Convert them to shorter ones
  - e.g., query segmentation, identify key concepts, ignore or reduce weight of some parts
- Find similar queries that work
  - e.g., finding answers in CQA by finding same question
  - more generally, learning how to paraphrase

Approaches to Long Queries

- Analyze query to identify additional features
  - e.g., "factoid" question answering
  - more generally, exploiting linguistic features for ranking
- New retrieval models
  - e.g. translation-based models

CIIR research

- Finding key concepts in long queries
- Finding similar questions using translation models
- Text reuse
- Prior art search



Finding key concepts

- Long or “verbose” queries mix key concepts with additional qualifications, relationships, structure
- Current search engines don’t make good use of this additional text
  - this includes web search engines and TREC search engines

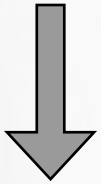
Basic Process

- Segment query to find concepts
  - NPs used for simplicity
  - cf. segmentation research for NCs
- Weight concepts using classifier
- Use a linear combination of query and all weighted concepts for ranking

$$rank(d) \propto \lambda p(q \mid d) + (1-\lambda) \sum_{c_i \in q} p(c_i \mid q)\, p(c_i \mid d)$$

TREC example

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

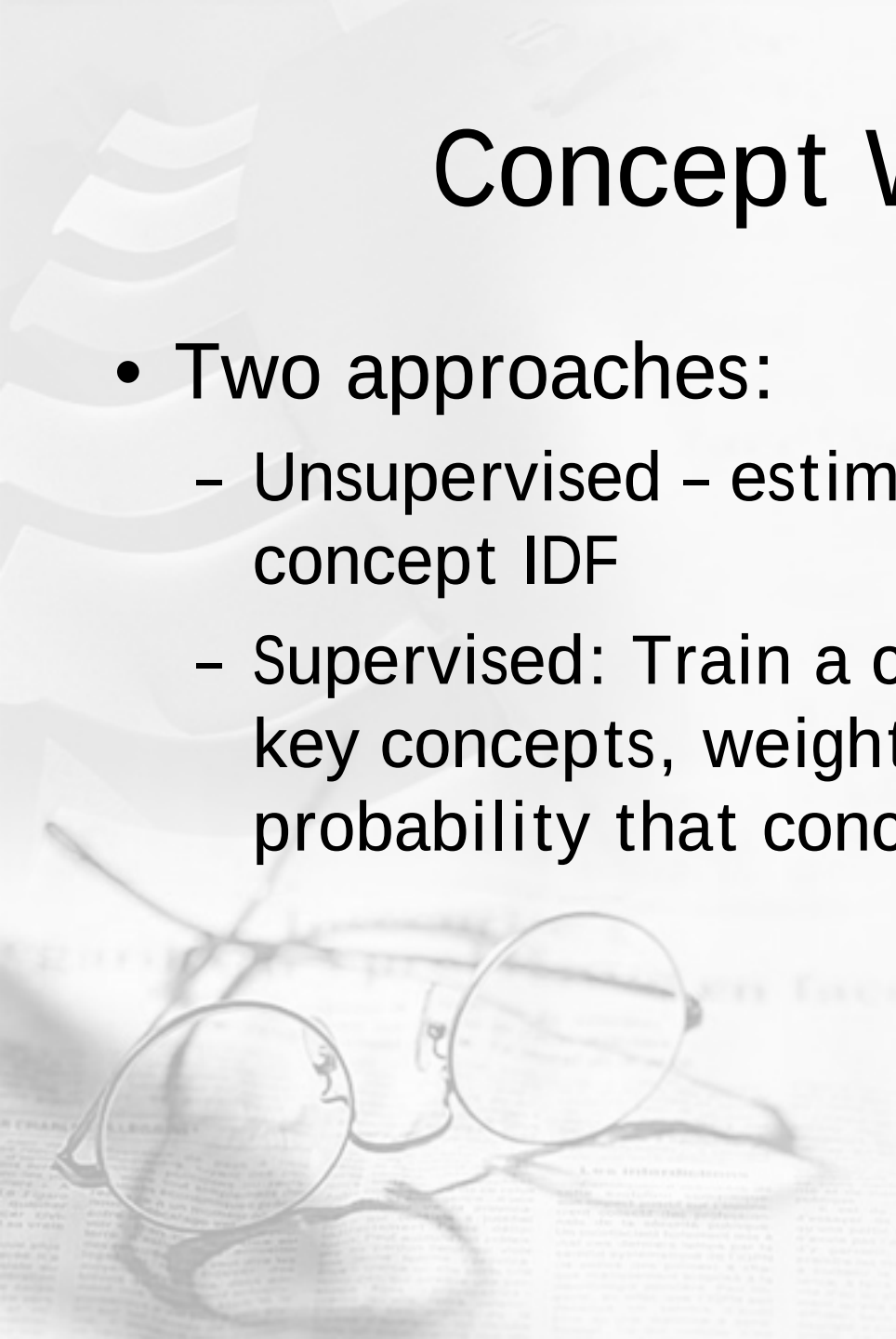


Concept extraction

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

Concept Weighting

- Two approaches:
  - Unsupervised – estimate importance using concept IDF
  - Supervised: Train a classifier to recognize key concepts, weight by estimate of probability that concept belongs to that class



Collection-based features

$is\_cap(c_i)$ - Is concept capitalized?

$tf(c_i)$ - Concept TF in the collection

$idf(c_i)$ - Concept IDF in the collection

$ridf(c_i)$ - Concept residual IDF in the collection
(Actual IDF deviation from Poisson model prediction; Church & Gale, 1995)

$wig(c_i)$ - Concept Weighted Information Gain *(Zhou & Croft, 2007)*

Collection-independent features

$g\_tf(c_i)$ - Concept frequency in *Google n-grams*. Estimates concept frequency in a large web collection

$qp(c_i)$ - Number of times a concept was used as a part of a query, extracted from *Live Search* query logs

$qe(c_i)$ - Number of times a concept was used as an exact query, extracted from *Live Search* query logs

Retrieval results

	ROBUST04		W10g		GOV2	
	prec@5	MAP	prec@5	MAP	prec@5	MAP
<*title*>	47.80	25.28	30.73_d	19.31	56.75	$\mathbf{29.67}_d$
<*desc*>	47.26	24.50	39.20^t	18.62	52.62	25.27^t
SeqDep<*desc*>	**49.11**	25.69_d	39.80^t	19.28	$\mathbf{56.88}_d$	27.53^t_d
KeyConcept[2]<*desc*>	48.54	$\mathbf{26.20}_d$	$\mathbf{40.40}^t$	$\mathbf{20.46}^t_d$	56.77_d	27.27^t_d

MAP and Precision at 5 results.

Key Concept Summary

- Identifying key concepts in queries can be done with reasonable accuracy using supervised learning with very limited training data
- Query expansion by weighted concepts improves retrieval performance for verbose queries
- More features, query processing could be used

Searching CQA archives

- A CQA service involves people answering other peoples' questions
  - e.g., Yahoo! Answers, Live QnA
- Many questions about relationships, but covers the whole range
  - including TREC QA factoid-type questions
- Much longer questions than the Web
  - because people are at the "other end"
- Latency of replies is a problem

Searching CQA archives

- Searching the archive of previously answered questions can provide good answers
- Three approaches:
  - treat answers as “mini-documents” and search
  - treat as a (factoid) question-answering problem
  - find similar questions and retrieve associated answers
- Answers found through similar questions give best performance

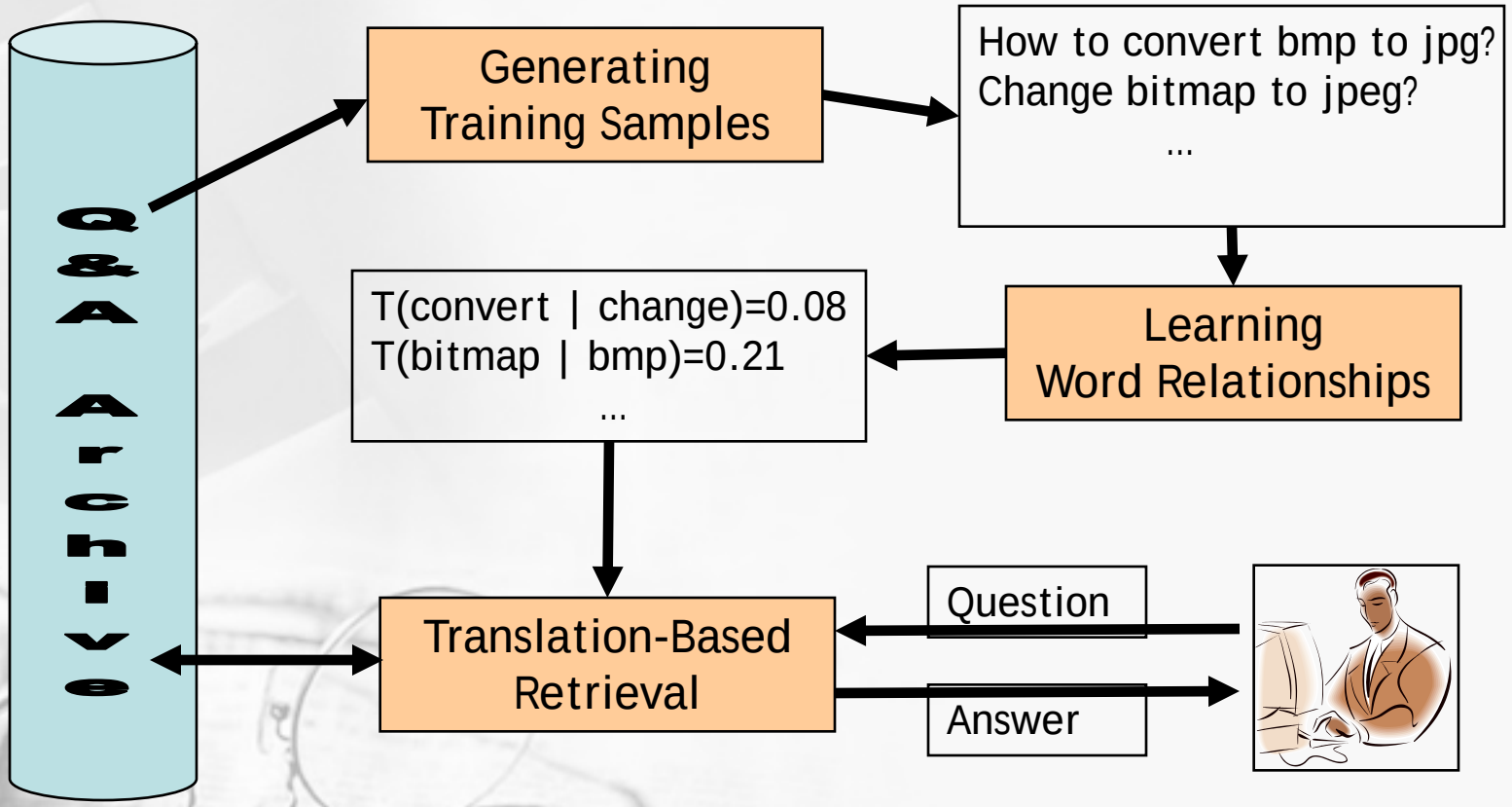
Data Sources

Collection	Naver	Wondir	WebFAQ
Provider	**naver.com**	**wondir.com**	**U of Amsterdam**
Q&A Source	**Community-based QA service**	**Community-based QA service**	**FAQs from the web, robot crawler**
Language	**Korean**	**English**	**English, Dutch**
#(Q&A Pairs)	**8 million**	**1 million**	**3 million**
#(Uniq Terms)	**9,354,612**	**176,078**	**1,978,238**
Fields (Avg Length)	**Question Title(6) Question Body(53) Answer(187)**	**Question(27) Answer(28)**	**Question(9) Answer(101)**

Question Retrieval

- Our approach uses a translation-based retrieval model
  - Statistical translation model ranks likely *reformulations*
  - Extension of query likelihood retrieval model
    - IR as Statistical Translation, Berger and Lafferty, SIGIR 99.
  - Simple model captures word relationships
  - Estimating translation probabilities is the major problem

Overview of CQA search



More detail: retrieval model

- Mixture of query likelihood and translation model
- Deals with "self-translation"

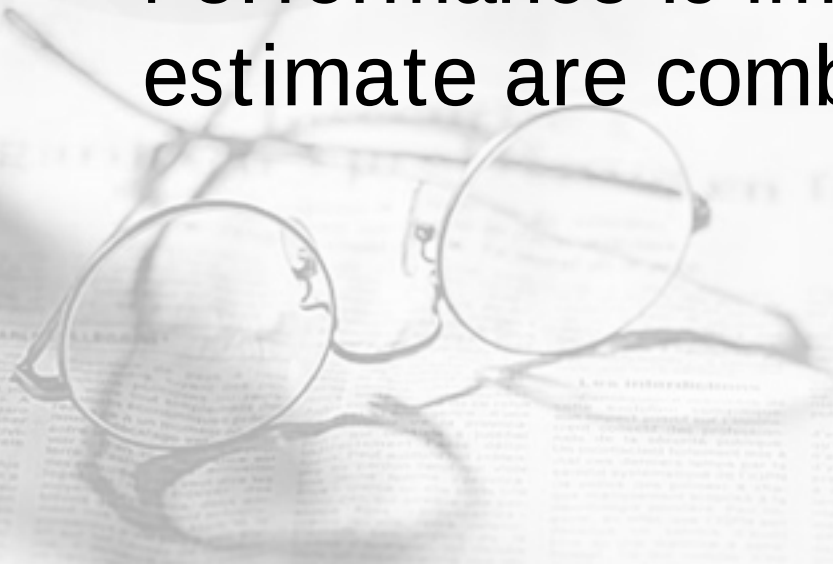
$$P(Q \mid D) = \prod_{w \in Q} P(w \mid D)$$

$$P(w \mid D) = \frac{|D|}{|D| + \lambda} P_{mx}(w \mid D) + \frac{\lambda}{|D| + \lambda} P_{ml}(w \mid C)$$

$$P_{mx}(w \mid D) = (1 - \beta) P_{ml}(w \mid D) + \beta \sum_{t \in D} P(w \mid t) P_{ml}(t \mid D)$$

Translation probabilities

- Basic approach uses EM-based algorithm from IBM model 1
- In case of Q&A pairs, either question or answer can be used as source or target
- Performance is improved if both forms of estimate are combined



Translation examples

P(A|Q)	P(Q|A)	P_{pool}
everest	mountain	everest
29,035	tallest	mountain
ft	everest	tallest
mount	highest	29,035
8,850	mt	highest
feet	discover	mt
measure	hillary	ft
expedition	edmund	measure
height	mountin	feet
nepal	biggest	mount

Top 10 translations for "everest" estimated from Wondir data

Question retrieval results

Model	Trans. Prob.	MAP	P@10
LM	-	0.322	0.221
RM	-	0.340	0.240
TransLM	P(A|Q)	0.406	0.268
TransLM	P(Q|A)	0.379	0.266
TransLM	P_{pool}	0.424	0.287

Wondir data, 50 TREC QA queries

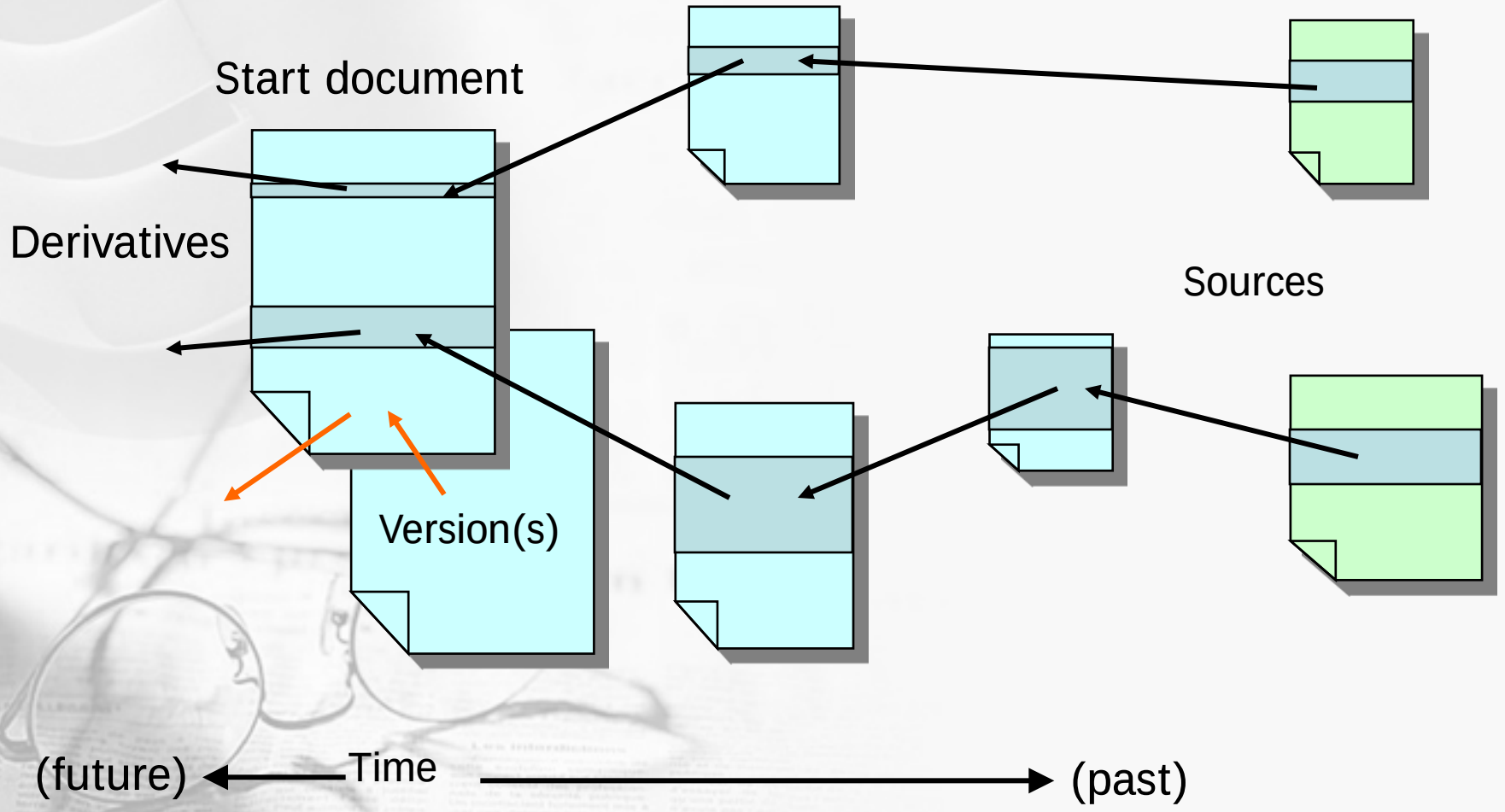
Examples of Q&A pairs retrieved

LM	TransLM+QL
Who is the leader of India?	
who is agashthy	who is the prime minister of india
who is veerappan	who is the current vice prime minister of india
how is father of india	who is the army chief of india
who is the general secretary of india	who is the first prime minister of india
Who made the first airplane that could fly?	
what is the oldest airline that still fly airplane	what is the oldest airline that still fly airplane
who is bin ladin	who was the first one who fly with plane
which airplne of the world has been fly the longest	who was the first person to fly a plane
what has 4 wheel and fly	who the first one fly to the spase
how do airplane fly	who the first one to fly to sky

CQA summary

- Q&A archives are a valuable resource for learning how to reformulate long queries
- Translation models are an effective technique for finding semantically similar queries that may be syntactically very different
- Need better training data for general retrieval
  - mine Web for "translation pairs"

Text Reuse and Information Flow

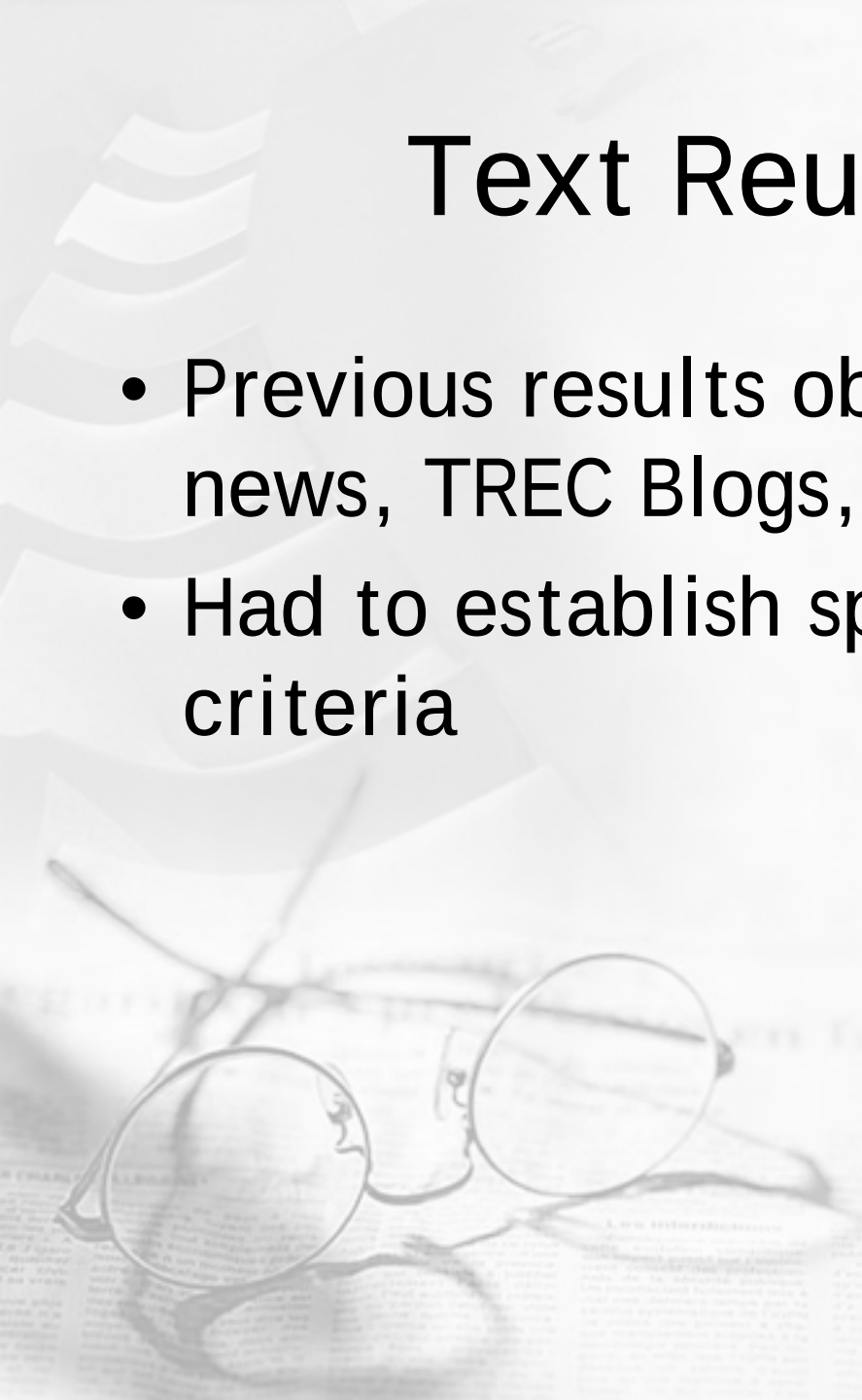


Current Research

- Develop techniques to detect local text reuse
  - Sentence similarity measures
  - Passage similarity measures
  - Variations of fingerprinting
- Develop a web-based tool for tracking information flow
  - find pool of documents using partial queries
  - apply text reuse detection at sentence level
  - identify timestamps for flow analysis

Text Reuse Testbeds

- Previous results obtained using TREC news, TREC Blogs, and now Web
- Had to establish specific relevance criteria



Prior Art Search

- Given a patent (or application), find relevant prior art
  - Query is entire patent
- Testbed for studying syntactic features and “obfuscation”
- Currently using LTR approach
  - extract features
  - train a ranking function using cited prior art

Long Query Summary

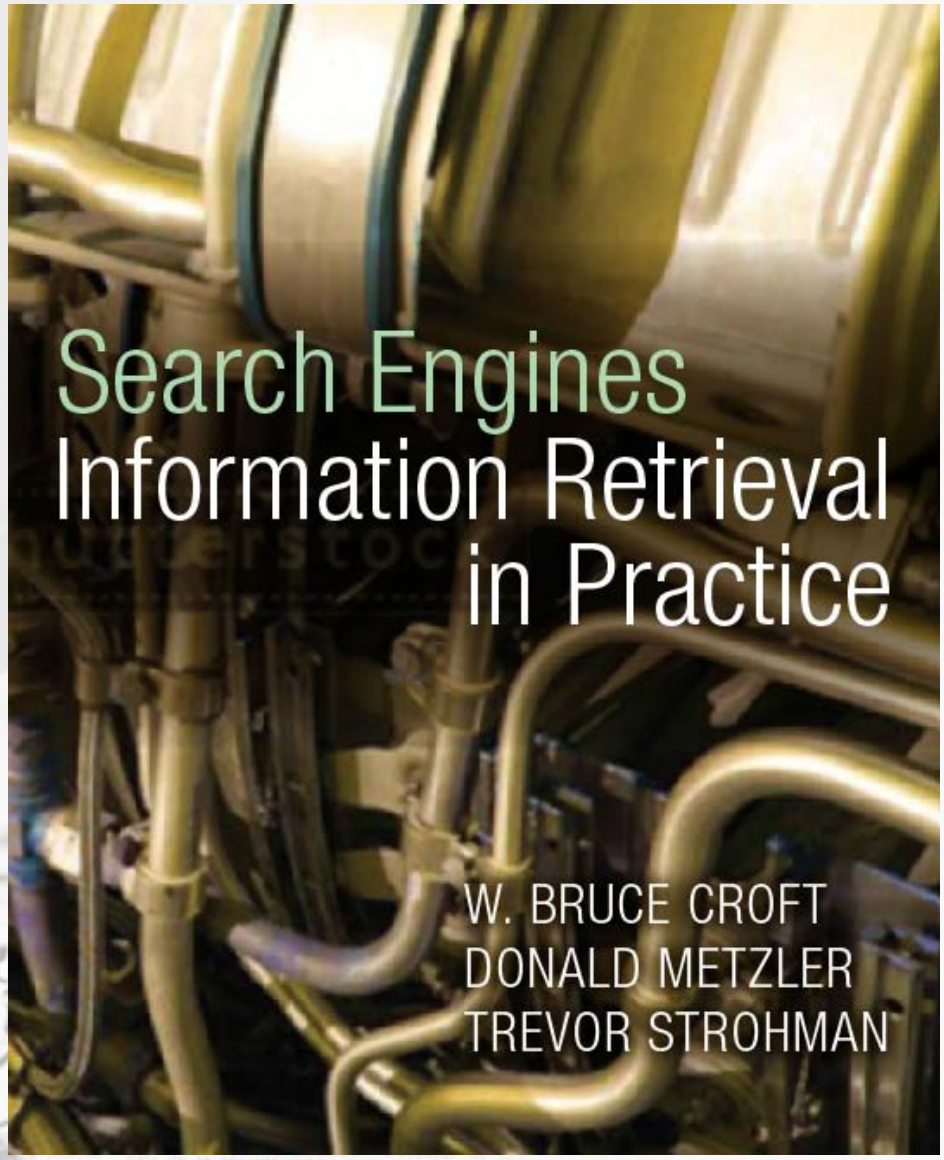
- Each application and testbed provides some insight to long queries
- Common foundations should develop out of application-based approaches
  - e.g., models of concept importance, query/sentence transformation models which may unify many aspects of the problem

Conclusions

- There are *many* unsolved problems in search
  - shown by mediocre effectiveness in a range of search applications
- Progress is made by defining problem and measuring performance in the context of an application
  - but should relate this to the big picture

Conclusions

- The problem of long queries is an example of studying an issue using multiple testbeds and perspectives
- We need more testbeds from more applications
  - and more graduate students


Search Engines
Information Retrieval
in Practice
W. BRUCE CROFT
DONALD METZLER
TREVOR STROHMAN

(Jan. 09)