Axiomatic Analysis and Optimization of Information Retrieval Models

ChengXiang Zhai
Dept. of Computer Science
University of Illinois at Urbana-Champaign
USA
http://www.cs.illinois.edu/homes/czhai

Hui Fang
Dept. of Electrical and Computer Engineering
University of Delaware
USA
http://www.eecis.udel.edu/~hfang

Goal of Tutorial

- Introduce the "axiomatic approach" to development of information retrieval models
- Review the major research progress in this area
- Discuss promising future research directions
- You can expect to learn
  - Basic methodology of axiomatic analysis and optimization of retrieval models
  - Novel retrieval models developed using axiomatic analysis
- Prerequisite: basic knowledge about information retrieval models is assumed

Outline

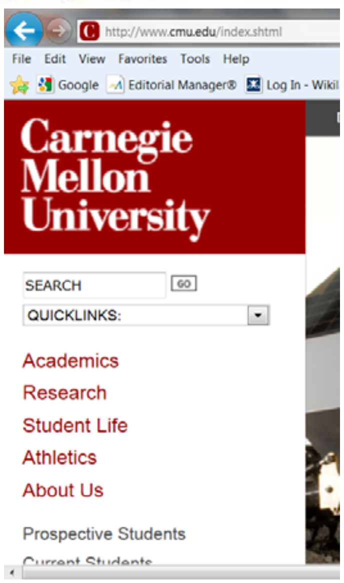
- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Search is everywhere, and part of everyone's life

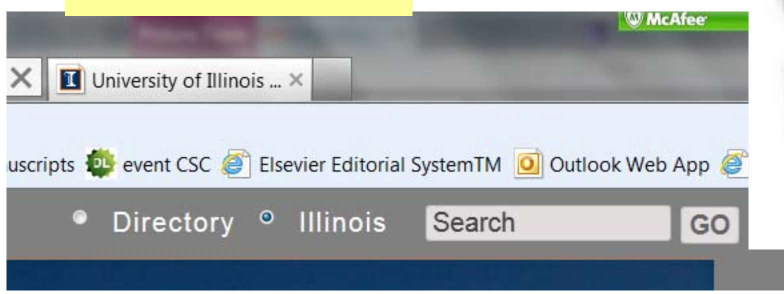
Web Search



Desk Search

Enterprise Search

Site Search



Social Media Search

... ...

Search accuracy matters!

	# Queries /Day	X 1 sec	X 10 sec
Google	4,700,000,000	~1,300,000 hrs	~13,000,000 hrs
twitter	1,600,000,000	~440,000 hrs	~4,400,000 hrs
PubMed	2,000,000	~550 hrs	~5,500 hrs

… …

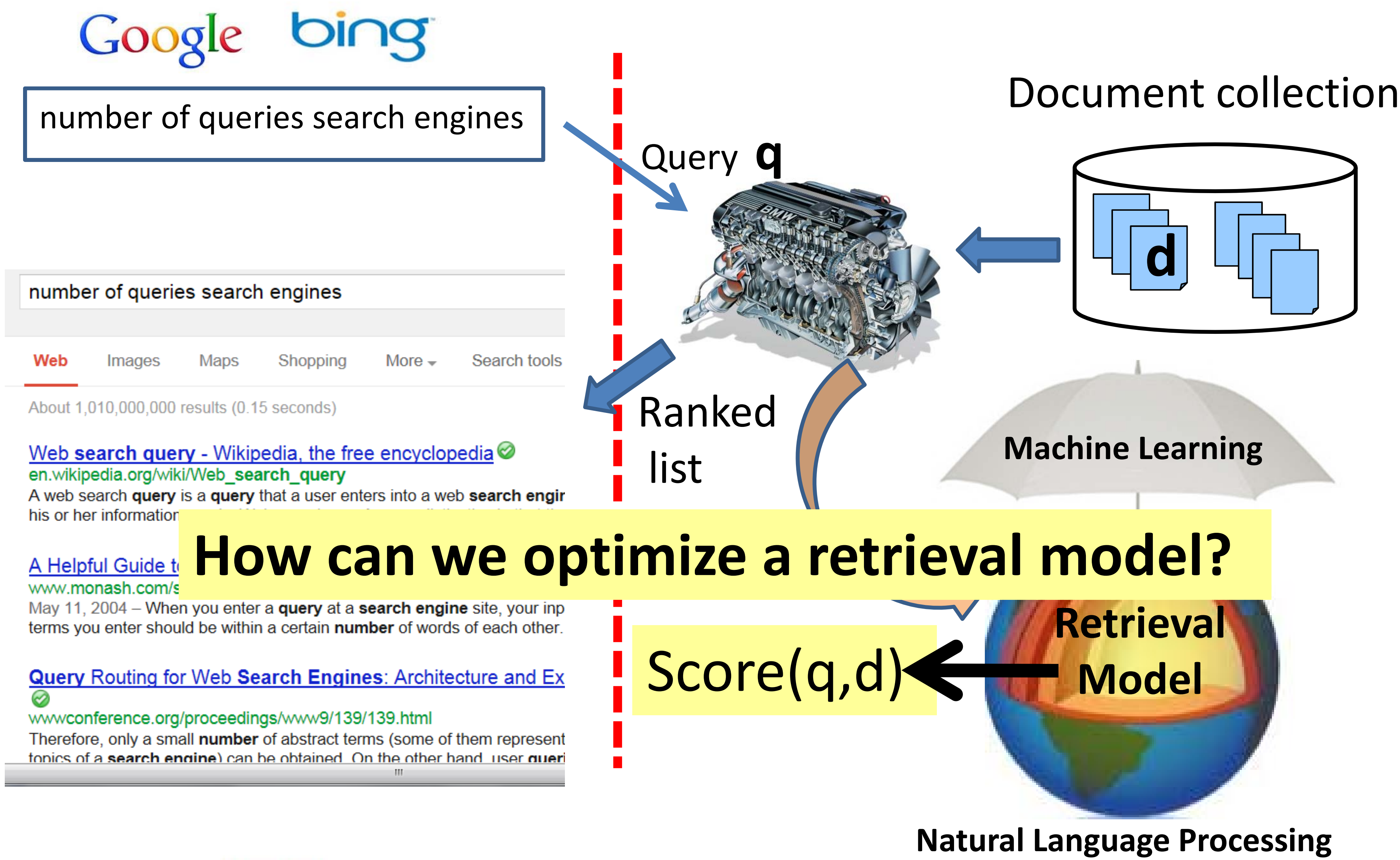
How can we improve all search engines in a general way?

Sources:

Google, Twitter: http://www.statisticbrain.com/

PubMed: http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html

Behind all the search boxes…



Retrieval model = computational definition of "relevance"

S("retrieval model tutorial", d)

s("retrieval", d) s("model", d) s("tutorial", d)

How many times does "retrieval" occur in d?

Term Frequency (TF): c("retrieval", d)

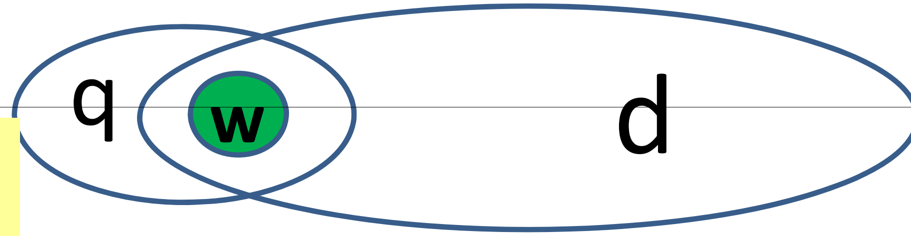
How long is d? **Document length**: |d|

How often do we see "retrieval" in the entire collection?

Document Frequency: df("retrieval")

P("retrieval"|collection)

Scoring based on bag of words in general



Sum over **matched query terms**

$$s(q,d) = f\left(\sum_{w \in q \cap d} weight(w,q,d), a(q,d) \right)$$

$$g[c(w,q), c(w,d), |d|, df(w)]$$

$$p(w|C)$$

Inverse Document Frequency (IDF)

Term Frequency (TF)

Document length

Improving retrieval models is a long-standing challenge

- Vector Space Models: [Salton et al. 1975], [Singhal et al. 1996], …
- Classic Probabilistic Models: [Maron & Kuhn 1960], [Harter 1975], [Robertson & Sparck Jones 1976], [van Rijsbergen 1977], [Robertson 1977], [Robertson et al. 1981], [Robertson & Walker 1994], …
- Language Models: [Ponte & Croft 1998], [Hiemstra & Kraaij 1998], [Zhai & Lafferty 2001], [Lavrenko & Croft 2001], [Kurland & Lee 2004], …
- Non-Classic Logic Models: [van Rijsbergen 1986], [Wong & Yao 1995], …
- Divergence from Randomness: [Amati & van Rijsbergen 2002], [He & Ounis 2005], …
- Learning to Rank: [Fuhr 1989], [Gey 1994], ...
- …

Many different models were proposed and tested

Some are working very well (equally well)

- Pivoted length normalization (PIV) [Singhal et al. 96]
- BM25 [Robertson & Walker 94]
- PL2 [Amati & van Rijsbergen 02]
- Query likelihood with Dirichlet prior (DIR) [Ponte & Croft 98], [Zhai & Lafferty]

but many others failed to work well…

State of the art retrieval models

- PIV (vector space model)

$$\sum_{w \in q \cap d} \frac{1+\ln(1+\ln(c(w,d)))}{(1-s)+s\frac{|d|}{avdl}} \cdot c(w,q) \cdot \ln\frac{N+1}{df(w)}$$

- DIR (language modeling approach)

$$\sum_{w \in q \cap d} c(w,q) \times \ln\left(1+\frac{c(w,d)}{\mu \cdot p(w|C)}\right) + |q| \cdot \ln\frac{\mu}{\mu+|d|}$$

- BM25 (classic probabilistic model)

$$\sum_{w \in q \cap d} \ln\frac{N-df(w)+0.5}{df(w)+0.5} \cdot \frac{(k_1+1)\times c(w,d)}{k_1((1-b)+b\frac{|d|}{avdl})+c(w,d)} \cdot \frac{(k_3+1)\times c(w,q)}{k_3+c(w,q)}$$

PL2 is a bit more complicated, but implements similar heuristics

Questions

- **Why do {BM25, PIV, PL2, DIR, …} tend to perform similarly even though they were derived in very different ways?**

	AP88-89	DOE	FR88-89	Wt2g	Trec7	trec8
PIV	0.23	0.18	0.19	0.29	0.18	0.24
DIR	0.22	0.18	0.21	0.30	0.19	0.26
BM25	0.23	0.19	0.23	0.31	0.19	0.25
PL2	0.22	0.19	0.22	0.31	0.18	0.26

Questions

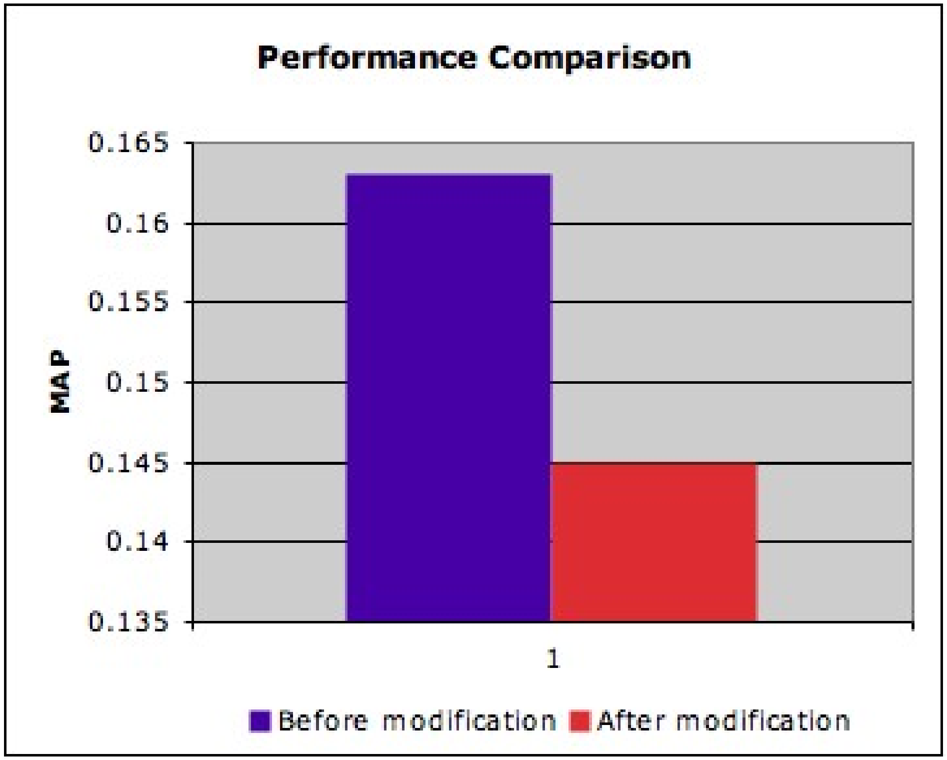
- Why do {BM25, PIV, PL, DIR, …} tend to perform similarly even though they were derived in very different ways?
- **Why are they better than many other variants?**

Is it possible to predict the performance?

PIV:

$$1+\log(c(t,D))$$

$$S(Q,D)=\sum_{t\in D\cap Q} c(t,Q)\times\log\frac{N+1}{df(t)}\times\frac{\cancel{1+\log(1+\log(c(t,D)))}}{(1-s)+s\times\frac{|D|}{avdl}}$$



Questions

- Why do {BM25, PIV, PL, DIR, …} tend to perform similarly even though they were derived in very different ways?
- Why are they better than many other variants?
- **Why does it seem to be hard to beat these strong baseline methods?**

Questions

- Why do {BM25, PIV, PL, DIR, …} tend to perform similarly even though they were derived in very different ways?
- Why are they better than many other variants?
- Why does it seem to be hard to beat these strong baseline methods?
- **Are they hitting the ceiling of bag-of-words assumption?**
  - **If yes, how can we prove it?**
  - **If not, how can we find a more effective one?**

Suggested Answers: Axiomatic Analysis

- Why do {BM25, PIV, PL, DIR, …} tend to perform similarly even though they were derived in very different ways?

  They share some nice common properties

  These properties are more important than how each is derived

- Why are they better than many other variants?

  Other variants don't have all the "nice properties"

- Why does it seem to be hard to beat these strong baseline methods?

  We don't have a good knowledge about their deficiencies

- Are they hitting the ceiling of bag-of-words assumption?
  - If yes, how can we prove it?
  - If not, how can we find a more effective one?

Need to formally define "the ceiling" (= complete set of "nice properties")

Axiomatic Relevance Hypothesis (ARH)

- Relevance can be modeled by a set of formally defined constraints on a retrieval function
  - If a function satisfies all the constraints, it will perform well empirically
  - If function F_a satisfies more constraints than function F_b, F_a would perform better than F_b empirically
- Analytical evaluation of retrieval functions
  - Given a set of relevance constraints $C = \{c_1, \ldots, c_k\}$
  - Function F_a is analytically more effective than function F_b iff the set of constraints satisfied by F_b is a proper subset of those satisfied by F_a.
  - A function F is optimal iff it satisfies all the constraints in C.

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Different models, but similar heuristics

- PIV

$$\sum_{w \in q \cap d} \frac{1+\ln(1+\ln(c(w,d)))}{(1-s)+s\frac{|d|}{avdl}} \cdot c(w,q) \cdot \ln\frac{N+1}{df(w)}$$

Document Length Normalization
Inverse Document Frequency
Term Frequency
Parameter sensitivity

- DIR

$$\sum_{w \in q \cap d} c(w,q) \times \ln\left(1+\frac{c(w,d)}{\mu \cdot p(w|C)}\right) + |q| \cdot \ln\frac{\mu}{\mu+|d|}$$

- BM25

$$\sum_{w \in q \cap d} \ln\frac{N-df(w)+0.5}{df(w)+0.5} \cdot \frac{(k_1+1)\times c(w,d)}{k_1((1-b)+b\frac{|d|}{avdl})+c(w,d)} \cdot \frac{(k_3+1)\times c(w,q)}{k_3+c(w,q)}$$

PL2 is a bit more complicated, but implements similar heuristics

Are they performing well because they implement similar retrieval heuristics?

Can we formally capture these necessary retrieval heuristics?

[Fang et. al 2004, Fang et al 2011]

Term Frequency Constraints (TFC1)

Give a higher score to a document with more occurrences of a query term.

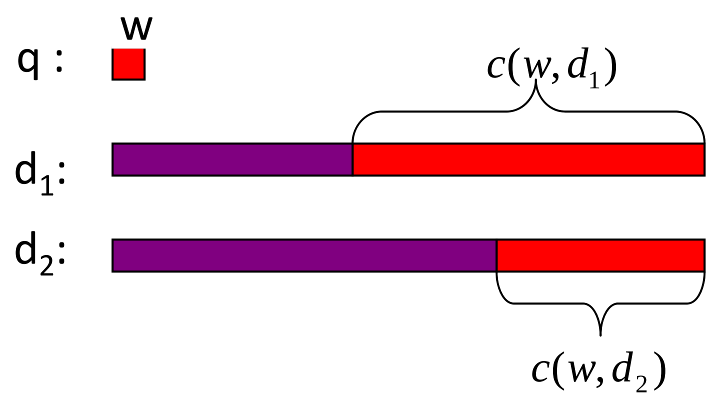
- *TFC1*

Let *q* be a query with only one term *w*.

If $|d_1| = |d_2|$

and $c(w,d_1) > c(w,d_2)$

then $f(d_1,q) > f(d_2,q)$.



$$f(d_1,q) > f(d_2,q)$$

Term Frequency Constraints (TFC3)

Favor a document with more distinct query terms.

- *TFC3*

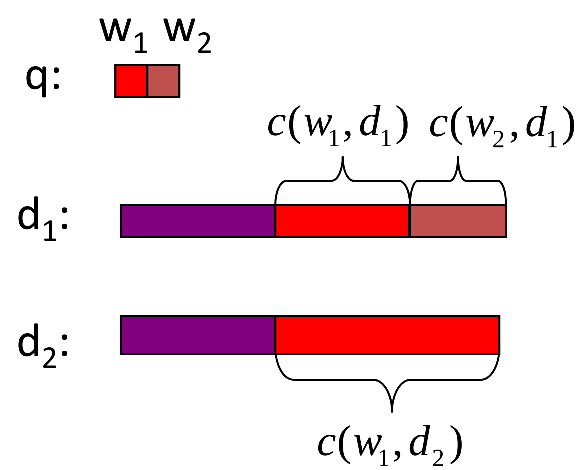
Let q be a query and w_1, w_2 be two query terms.

Assume $idf(w_1) = idf(w_2)$ and $|d_1| = |d_2|$

If $c(w_1, d_2) = c(w_1, d_1) + c(w_2, d_1)$

and $c(w_2, d_2) = 0, c(w_1, d_1) \neq 0, c(w_2, d_1) \neq 0$

then $f(d_1, q) > f(d_2, q)$.



$$f(d_1, q) > f(d_2, q)$$

Length Normalization Constraints(LNCs)

Penalize long documents(LNC1);

Avoid over-penalizing long documents (LNC2) .

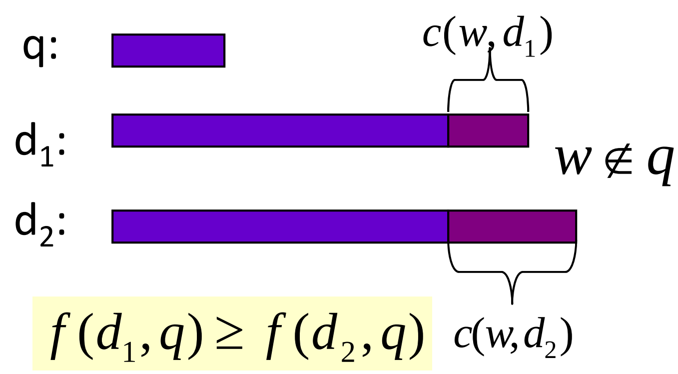
- ***LNC1***

  Let q be a query.

  If for some word $w \notin q, c(w,d_2) = c(w,d_1) + 1$

  but for other words $w, c(w,d_2) = c(w,d_1)$

  then $f(d_1,q) \geq f(d_2,q)$

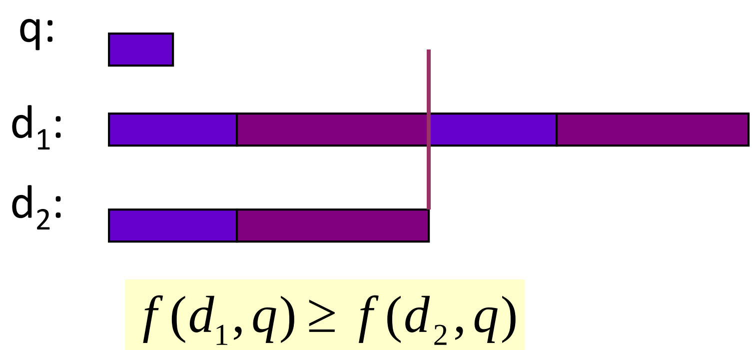


- ***LNC2***

  Let *q* be a query.

  If $\forall k > 1, |d_1| = k \cdot |d_2|$ and $c(w,d_1) = k \cdot c(w,d_2)$

  then $f(d_1,q) \geq f(d_2,q)$



TF-LENGTH Constraint (TF-LNC)

Regularize the interaction of TF and document length.

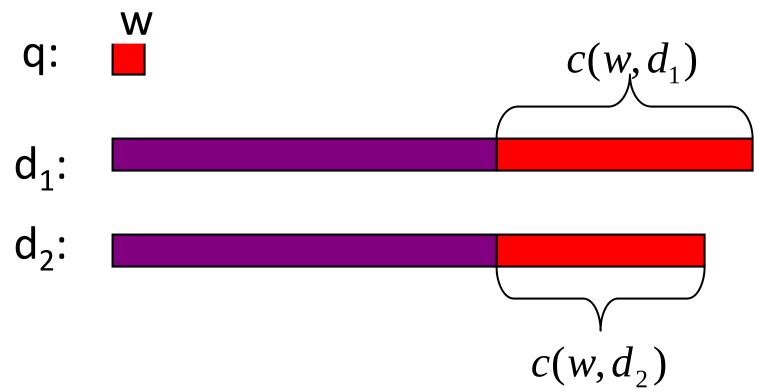
- *TF-LNC*

Let *q* be a query with only one term *w*.

If $|d_1| = |d_2| + c(w,d_1) - c(w,d_2)$

and $c(w,d_1) > c(w,d_2)$

then $f(d_1,q) > f(d_2,q)$.



$$f(d_1,q) > f(d_2,q)$$

Seven Basic Relevance Constraints

[Fang et al. 2011]

Constraints	Intuitions
TFC1	To favor a document with more occurrences of a query term
TFC2	To ensure that the amount of increase in score due to adding a query term repeatedly must decrease as more terms are added
TFC3	To favor a document matching more distinct query terms
TDC	To penalize the words popular in the collection and assign higher weights to discriminative terms
LNC1	To penalize a long document (assuming equal TF)
LNC2, TF-LNC	To avoid over-penalizing a long document
TF-LNC	To regulate the interaction of TF and document length

Hui Fang, Tao Tao, ChengXiang Zhai: Diagnostic Evaluation of Information Retrieval Models. ACM Trans. Inf. Syst. 29(2): 7 (2011)

Discussion 1: Weak or Strong Constraints?

TDC:

To penalize the words popular in the collection and assign higher weights to discriminative terms

- Our first attempt:
  - Let $Q = \{q_1, q_2\}$. Assume $|D_1| = |D_2|$ and $c(q_1, D_1) + c(q_2, D_1) = c(q_1, D_2) + c(q_2, D_2)$. If $td(q_1) \geq td(q_2)$ and $c(q_1, D_1) \geq c(q_1, D_2)$, we have $S(Q, D_1) \geq S(Q, D_2)$.

- Our second attempt (a relaxed version)
  - Let $Q = \{q_1, q_2\}$. Assume $|D_1| = |D_2|$ and D1 contains only q1 and D2 contains only q2. If $td(q_1) \geq td(q_2)$, $S(Q, D_1 \cup D) \geq S(Q, D_2 \cup D)$.

Discussion 2:
Avoid including redundant constraints

LNC1

Let q be a query.

If for some word $w \notin q, c(w,d_2) = c(w,d_1) + 1$

but for other words $w, c(w,d_2) = c(w,d_1)$

then $f(d_1,q) \geq f(d_2,q)$

TF-LNC

Let q be a query with only one term w.

If $|d_1| = |d_2| + c(w,d_1) - c(w,d_2)$

and $c(w,d_1) > c(w,d_2)$

then $f(d_1,q) > f(d_2,q)$.

Derived constraints

Let q be a query with only one term w.

If $|d_3| < |d_2| + c(w,d_3) - c(w,d_2)$

and $c(w,d_3) > c(w,d_2)$

then $f(d_3,q) > f(d_2,q)$.

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Axiomatic Relevance Hypothesis (ARH)

- Relevance can be modeled by a set of formally defined constraints on a retrieval function
  - If a function satisfies all the constraints, it will perform well empirically
  - If function F_a satisfies more constraints than function F_b, F_a would perform better than F_b empirically
- Analytical evaluation of retrieval functions
  - Given a set of relevance constraints $C = \{c_1, \ldots, c_k\}$
  - Function F_a is analytically more effective than function F_b iff the set of constraints satisfied by F_b is a proper subset of those satisfied by F_a.
  - A function F is optimal iff it satisfies all the constraints in C.

Testing the Axiomatic Relevance Hypothesis

- Is the satisfaction of these constraints correlated with good empirical performance of a retrieval function?
- Can we use these constraints to analytically compare retrieval functions without experimentation?
- "Yes!" to both questions
  - Constraint analysis reveals optimal ranges of parameter values
  - When a formula does not satisfy the constraint, it often indicates non-optimality of the formula.
  - Violation of constraints may pinpoint where a formula needs to be improved.

An Example of Constraint Analysis

PIV:

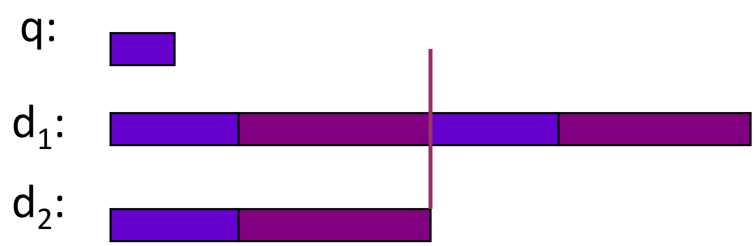
$$f(d,q)=\sum_{w\in q\cap d}\frac{1+ln(1+ln(c(w,d)))}{1-s+s\frac{|d|}{avdl}}\cdot c(w,q)\cdot ln\frac{N+1}{df(w)}$$

LNC2:

Let *q* be a query.

If $\forall k>1, |d_1|=k\cdot|d_2|$ and $c(w,d_1)=k\cdot c(w,d_2)$

then $f(d_1,q)\geq f(d_2,q)$



$f(d_1,q)\geq f(d_2,q)$

Does PIV satisfy LNC2?

An Example of Constraint Analysis

LNC2:

Let q be a query.

If $\forall k>1, |d_1| = k \cdot |d_2|$ and $c(w,d_1) = k \cdot c(w,d_2)$

then $f(d_1,q) \geq f(d_2,q)$

$$\frac{1+ln(1+ln(c(w,d_1)))}{1-s+s\frac{|d_1|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)} \geq \frac{1+ln(1+ln(c(w,d_2)))}{1-s+s\frac{|d_2|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)}$$

$$\frac{1+ln(1+ln(k\cdot c(w,d_2)))}{1-s+s\frac{k\cdot|d_2|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)} \geq \frac{1+ln(1+ln(c(w,d_2)))}{1-s+s\frac{|d_2|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)}$$

$$\frac{1+ln(1+ln(k\cdot c(w,d_2)))}{1-s+s\frac{k\cdot|d_2|}{avdl}} \geq \frac{1+ln(1+ln(c(w,d_2)))}{1-s+s\frac{|d_2|}{avdl}}$$

An Example of Constraint Analysis

$$\frac{1+ln(1+ln(k\cdot c(w,d_2)))}{1-s+s\frac{k\cdot|d_2|}{avdl}} \geq \frac{1+ln(1+ln(c(w,d_2)))}{1-s+s\frac{|d_2|}{avdl}}$$

$$s \leq \frac{tf_1 - tf_2}{(k\frac{|d_2|}{avdl}-1)tf_2 - (\frac{|d_2|}{avdl}-1)tf_1}$$

$$tf_1 = 1+ln(1+ln(k\cdot c(w,d_2)))$$

$$tf_2 = 1+ln(1+ln(c(w,d_2)))$$

Assuming $|d_2| = avdl,$

$$s \leq \frac{1}{k-1} \times (\frac{tf_1}{tf_2} - 1)$$

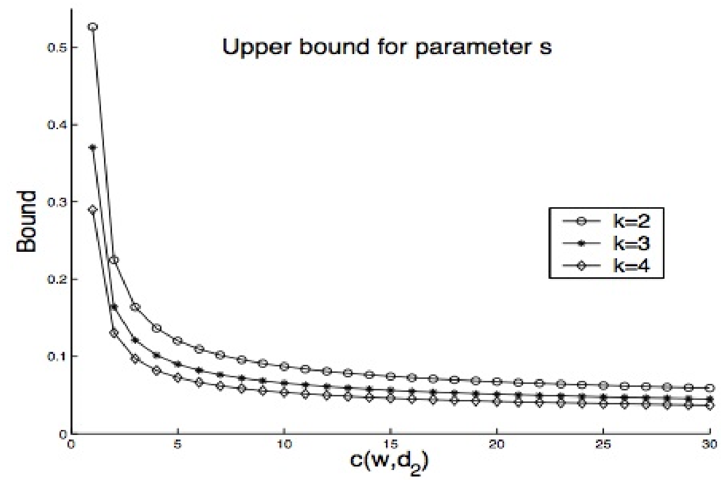


Figure 1: Upper bound of parameter s.

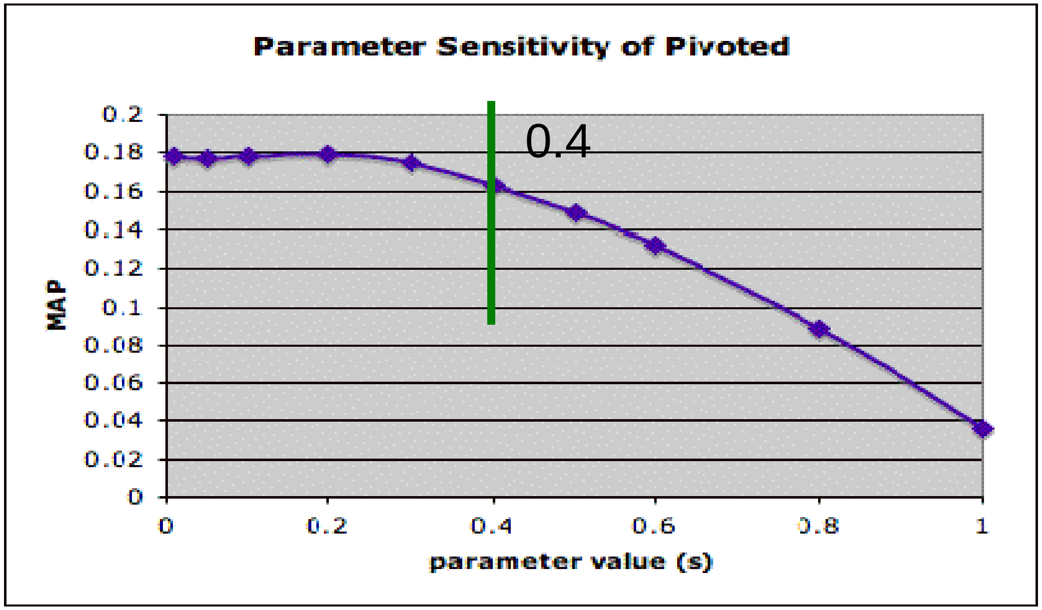
Bounding Parameters

- PIV

Optimal *s* (for average precision)

LNC2 ➔ s<0.4

	AP	DOE	FR	ADF	Web	Trec 7	Trec 8
LK	0.2	0.2	0.05	0.2	---	---	---
SK	0.01	0.2	0.01	0.05	0.01	0.05	0.05
LV	0.3	0.3	0.1	0.2	0.2	0.2	0.2
SV	0.2	0.3	0.1	0.2	0.1	0.1	0.2

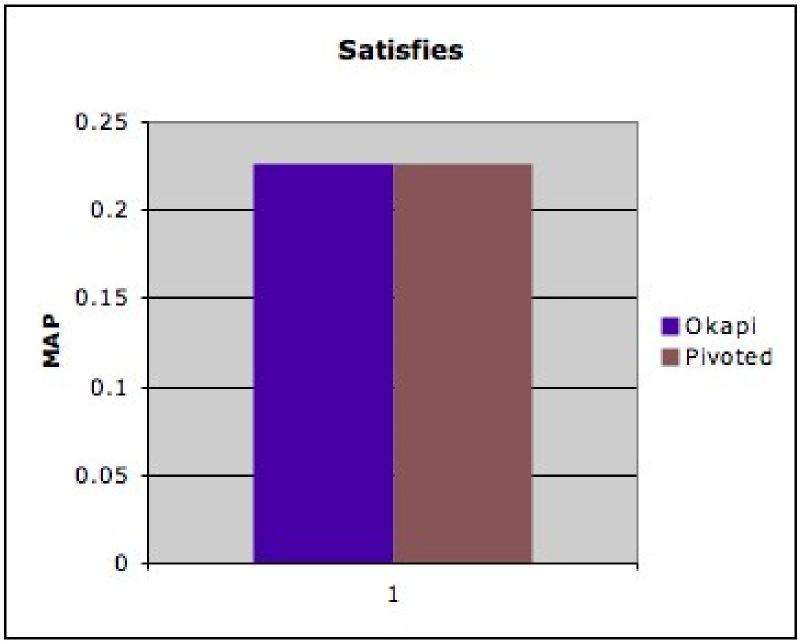


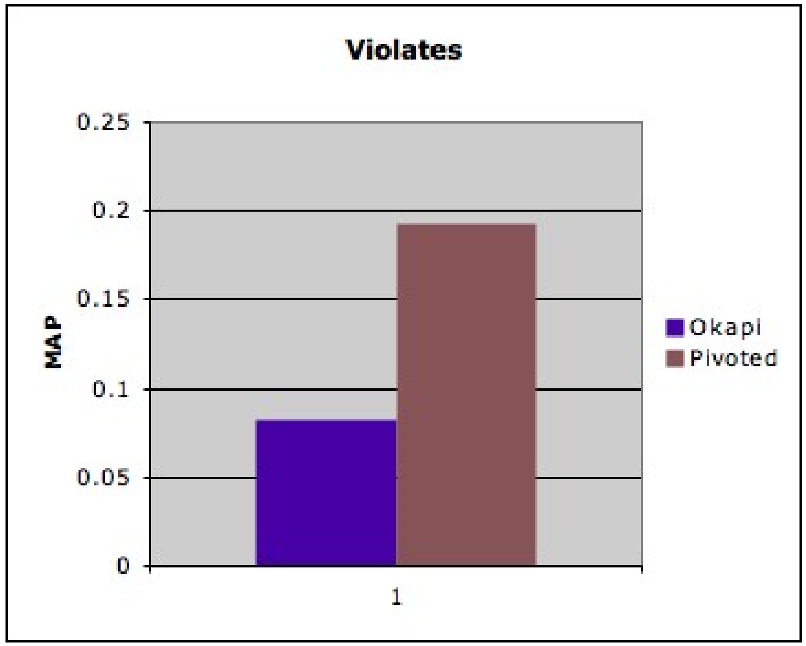
Analytical Comparison

- Okapi BM25

$$\sum_{t \in Q \cap D} \log \frac{N - df(t) + 0.5}{df(t)} \cdot \frac{(k_1 + 1) \cdot c(t,D)}{c(t,D) + k_1((1-b) + b \cdot \frac{|D|}{avdl})} \cdot \frac{(k_3 + 1) \cdot c(t,Q)}{k_3 + c(t,Q)}$$

Negative → Violates the constraints





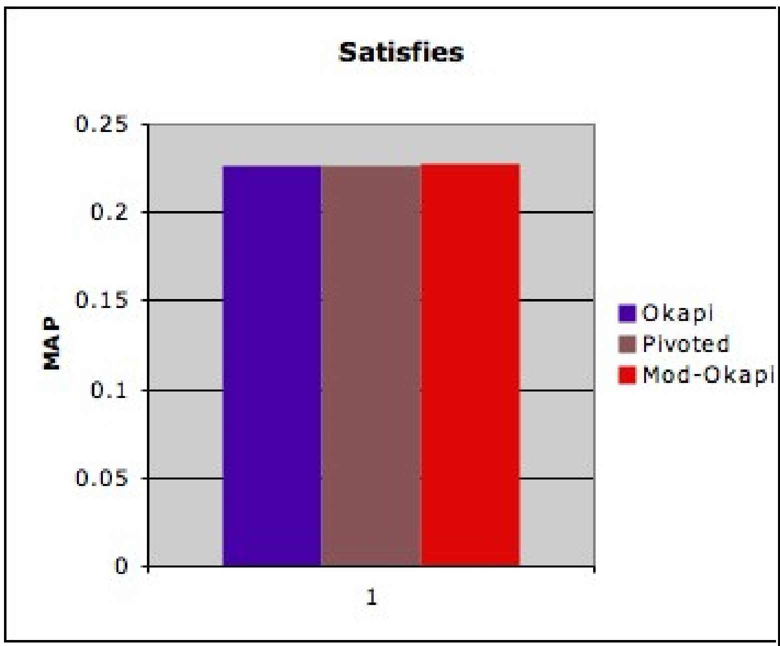
Fixing a deficiency in BM25 improves the effectiveness

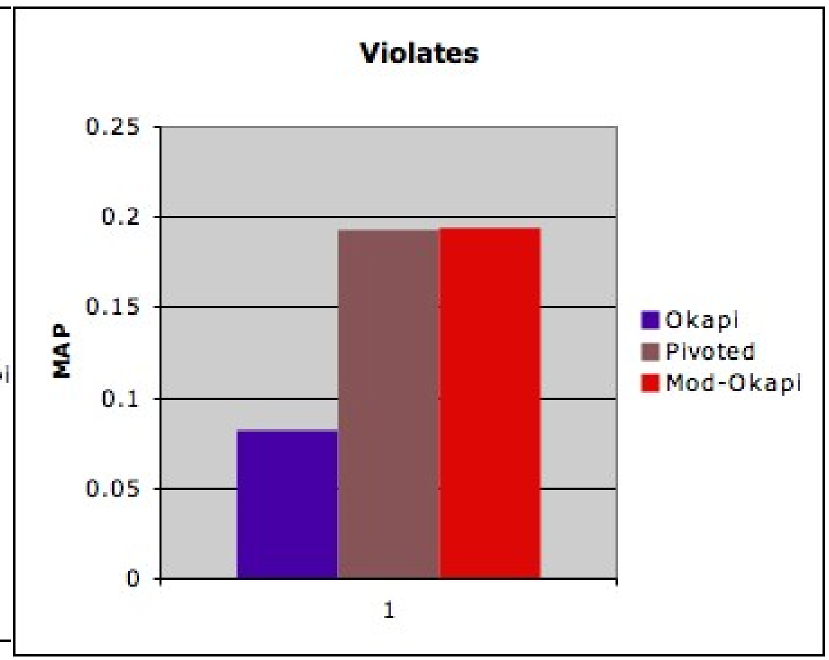
- Modified Okapi BM25

$$\log \frac{N+1}{df(t)}$$

$$\sum_{t \in Q \cap D} \log \frac{N - df(t) + 0.5}{df(t)} \cdot \frac{(k_1+1)\cdot c(t,D)}{c(t,D) + k_1((1-b) + b \cdot \frac{|D|}{avdl})} \cdot \frac{(k_3+1)\cdot c(t,Q)}{k_3 + c(t,Q)}$$

Make it satisfy constraints; expected to improve performance





TIMAN InfoLab

Systematic Analysis of 4 State of the Art Models

[Fang et al. 2011]

Function	TFCs	TDC			LNC
				C1*	C2*
				C3	
				C4	C4
(Modified)				Yes	Yes
PL2 (Original)	C5	C6*	C7	C8*	C8*
PL2 (modified)	Yes	C6*	Yes	C8*	C8*

Parameter s must be small

Problematic when a query term occurs less frequently in a doc than expected

Negative IDF

Problematic with common terms; parameter c must be large

Perturbation tests:

An empirical way of analyzing the constraints

[Fang et al 2011]

Medical Diagnosis Analogy

Non-optimal retrieval function

Better performed retrieval function

Design tests with available instruments

observe symptoms

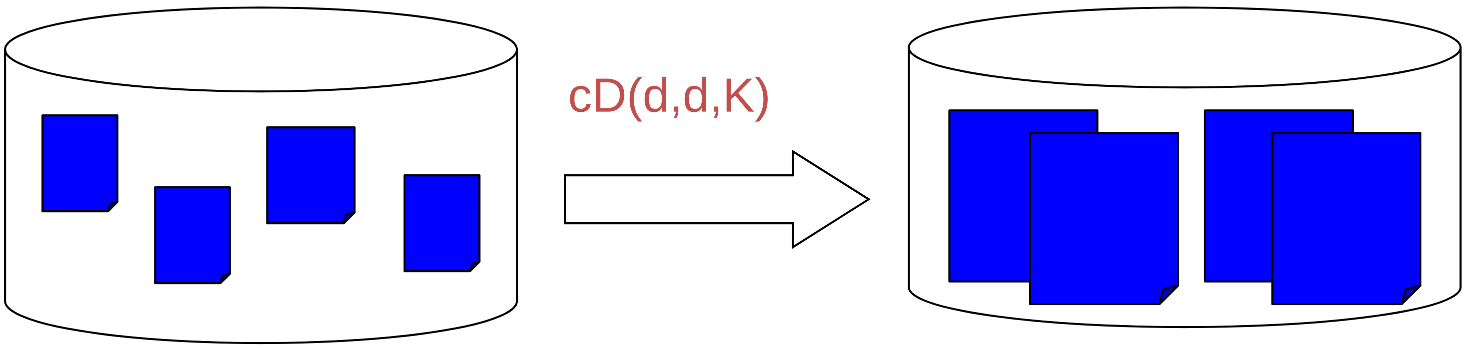
provide treatments

How to find available instruments?
How to design diagnostic tests?

Relevance-Preserving Perturbations

- Perturb term statistics
- Keep relevance status

Document scaling perturbation:



concatenate every document with itself K times

9 perturbations are designed

Relevance-Preserving Perturbations

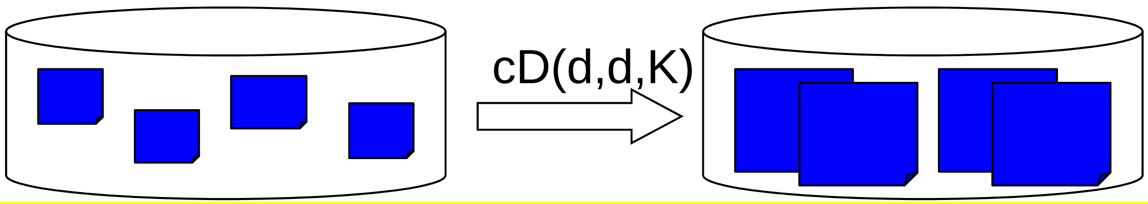
Name	Semantic
Relevance addition	Add a query term to a relevant document
Noise addition	Add a noisy term to a document
Internal term growth	Add a term to a document that original contains the term
Document scaling	Concatenate D with itself K times
Relevance document concatenation	Concatenate two relevant documents K times
Non-relevant document concatenation	Concatenate two non-relevant documents K times
Noise deletion	Delete a term from a non-relevant document
Document addition	Add a document to the collection
Document deletion	Delete a document from the collection

Length Scaling Test

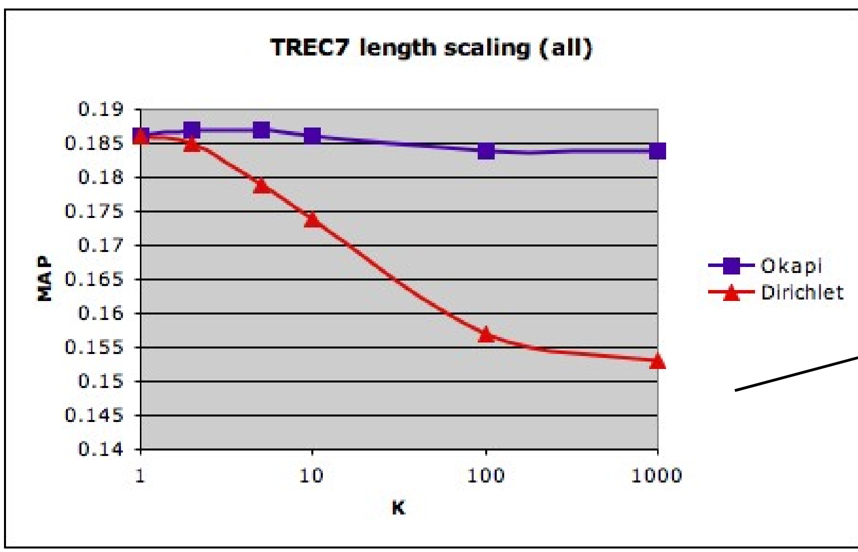
1. Identify the aspect to be diagnosed

test whether a retrieval function over-penalizes long documents

2. Choose appropriate perturbations



3. Perform the test and interpret the results



Dirichlet over-penalizes long documents!

Identifying the weaknesses makes it possible to improve the performance

	MAP			P@30		
	trec8	**wt2g**	**FR**	trec8	wt2g	FR
DIR	**0.257**	**0.302**	**0.207**	0.365	0.331	0.151
Imp.D.	**0.262**	**0.321**	**0.224**	0.373	0.345	0.166

Summary of All Tests

Tests	What to measure?
Length variance reduction	The gain on length normalization
Length variance amplification	The robustness to larger document variance
Length scaling	The ability at avoid over-penalizing long documents
Term noise addition	The ability to penalize long documents
Single query term growth	The ability to favor docs with more distinct query terms
Majority query term growth	Favor documents with more query terms
All query term growth	Balance TF and LN more appropritely

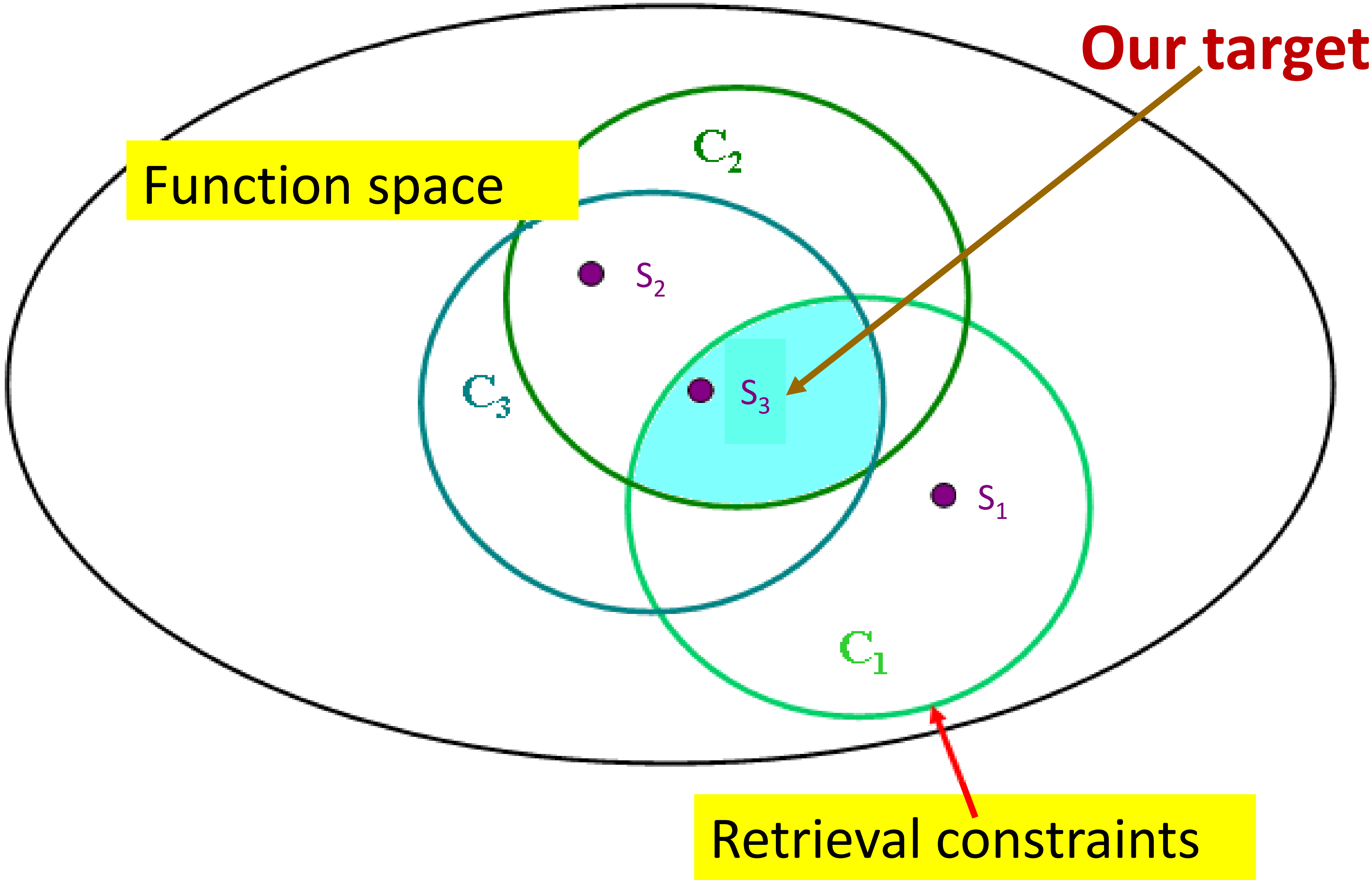
Hui Fang, Tao Tao, ChengXiang Zhai: Diagnostic Evaluation of Information Retrieval Models. ACM Trans. Inf. Syst. 29(2): 7 (2011)

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

How can we leverage constraints to find an optimal retrieval model?

Basic Idea of the Axiomatic Framework (Optimization Problem Setup)



Three Questions

- How do we define the constraints?

  We've talked about that; more later

- How do we define the function space?

  One possibility: leverage existing state of the art functions

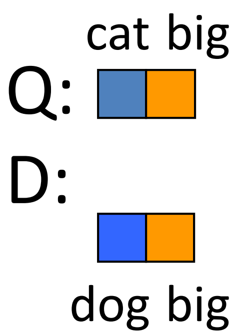
- How do we search in the function space?

  One possibility: search in the neighborhood of existing state of the art functions

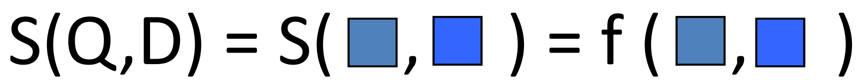
Inductive Definition of Function Space

$$S : Q \times D \rightarrow \Re \qquad Q = q_1, q_2, \ldots, q_m;\ D = d_1, d_2, \ldots, d_n$$

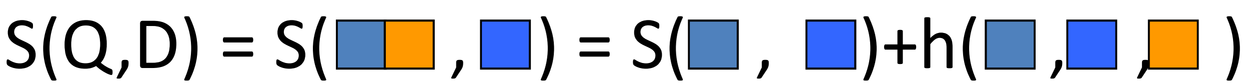
Define the function space *inductively*



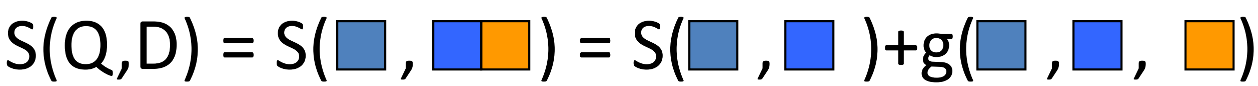
Primitive weighting function (f)



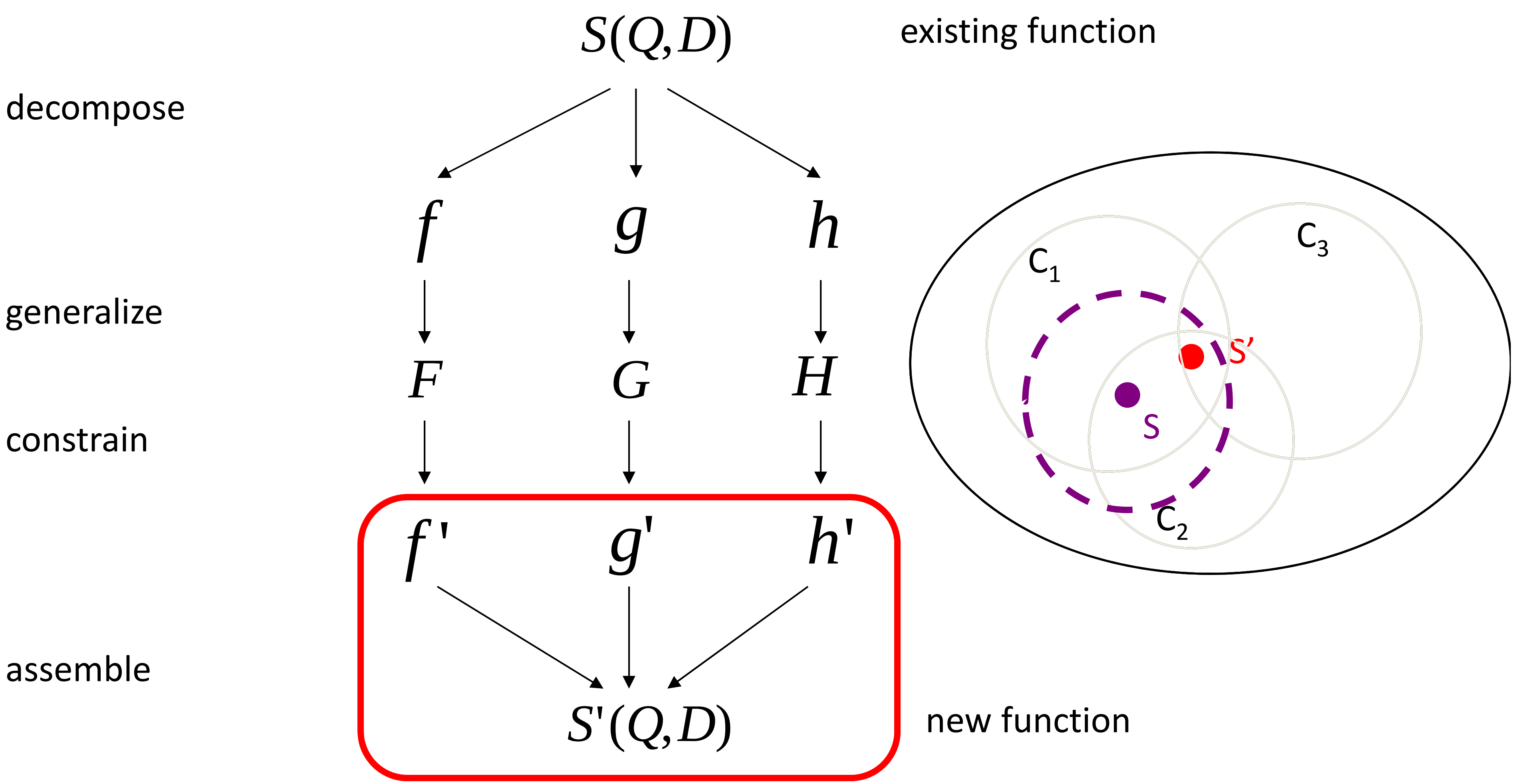
Query growth function (h)



Document growth function (g)



Derivation of New Retrieval Functions



A Sample Derived Function based on BM25

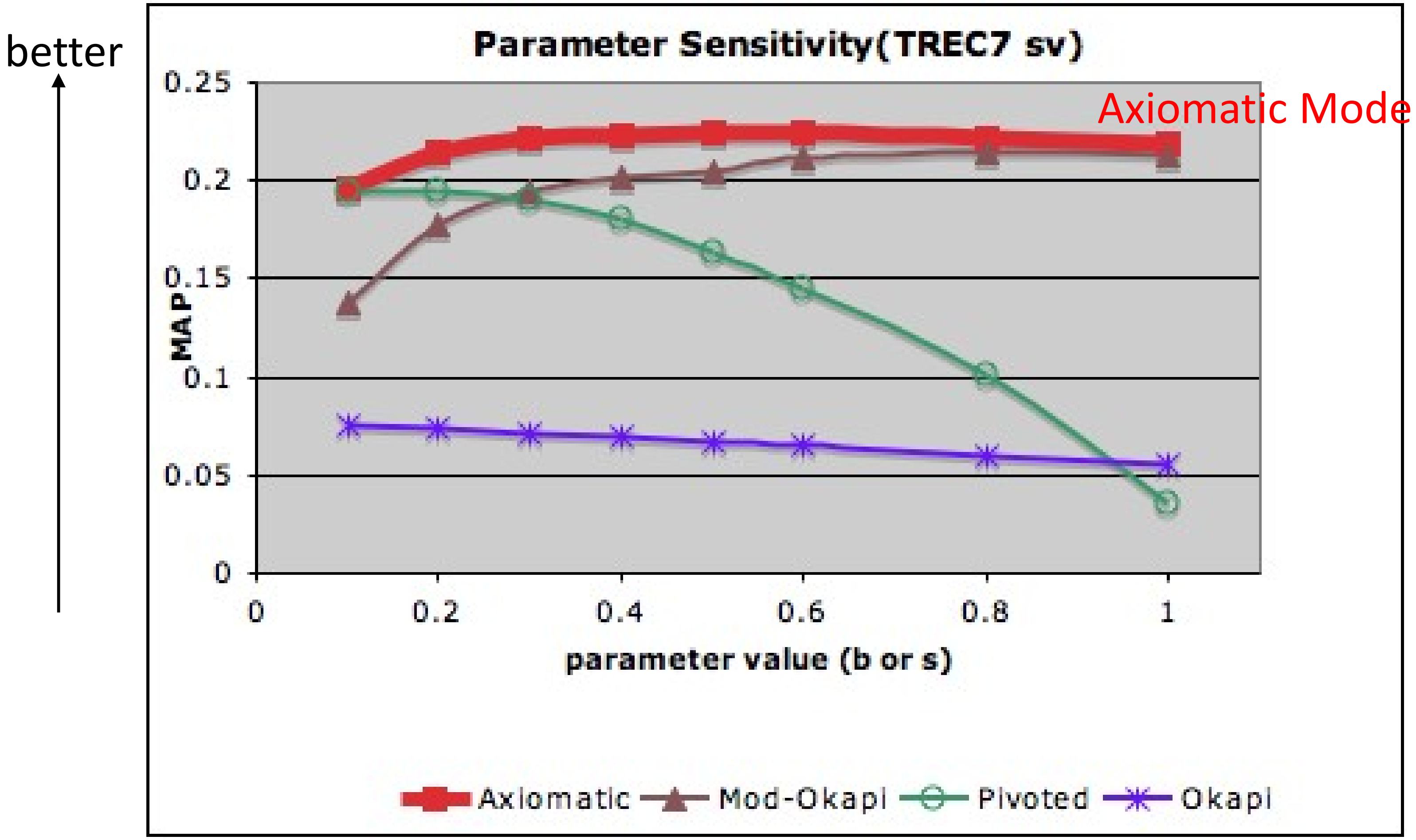
[Fang & Zhai 2005]

QTF IDF TF

$$S(Q,D) = \sum_{t \in Q \cap D} c(t,Q) \cdot \left(\frac{N}{df(t)}\right)^{0.35} \cdot \frac{c(t,D)}{c(t,D) + s + \frac{s \cdot |D|}{avdl}}$$

length normalization

The derived function is less sensitive to the parameter setting



Inevitability of heuristic thinking and necessity of axiomatic analysis

- The "theory-effectiveness gap"
  - Theoretically motivated models don't automatically perform well empirically
  - Heuristic adjustment seems always necessary
  - Cause: inaccurate modeling of relevance
- How can we bridge the gap?
  - The answer lies in axiomatic analysis
  - Use constraints to help identify the error in modeling relevance, thus obtaining insights about how to improve a model

Systematic Analysis of 4 State of the Art Models

[Fang et al. 2011]

Function	TFCs	TDC	LNC1	LNC2	TF-LNC
PIV	Yes	Yes	Yes	C1*	C2*
DIR	Yes	Yes	Yes	C3	Yes
BM25 (Original)	C4	Yes	C4	C4	C4
BM2 (Modified)	Yes	Yes	Yes	Yes	Yes

Modified BM25 satisfies all the constraints!

Without knowing its deficiency, we can't easily propose a new model working better than BM25

A Recent Success of Axiomatic Analysis: Lower Bounding TF Normalization

[Lv & Zhai 2011a]

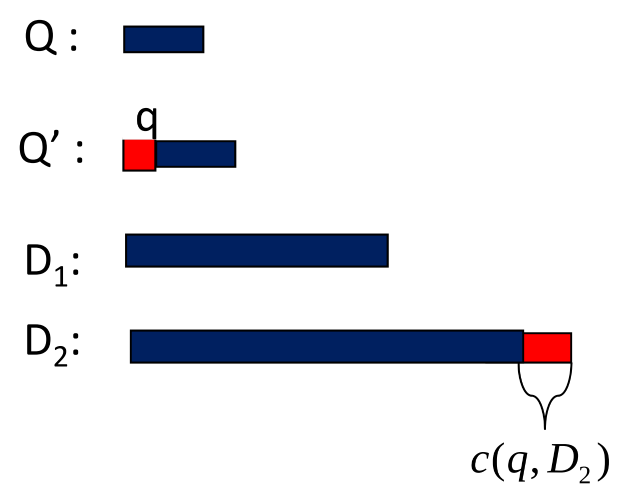
- Existing retrieval functions lack a lower bound for normalized TF with document length ➔
  - Long documents overly penalized
  - A very long document matching two query terms can have a lower score than a short document matching only one query term
- Proposed two constraints for lower bounding TF
- Proposed a general solution to fix the problem that worked for BM25, PL2, Dir, and Piv, leading to improved versions of them (BM25+, PL2+, Dir+, Piv+)

Lower Bounding TF Constraints (LB1)

The presence –absence gap (0-1 gap) shouldn't be closed due to length normalization.

- *LB1*

Let Q be a query. Assume D_1 and D_2 are two documents such that $S(Q, D_1) = S(Q, D_2)$.
If we reformulate the query by adding another term $q \notin Q$ into Q, where $c(q, D_1) = 0$ and $c(q, D_2) > 0$,
then $S(Q \cup \{q\}, D_1) < S(Q \cup \{q\}, D_2)$.



$S(Q, D_1) = S(Q, D_2)$

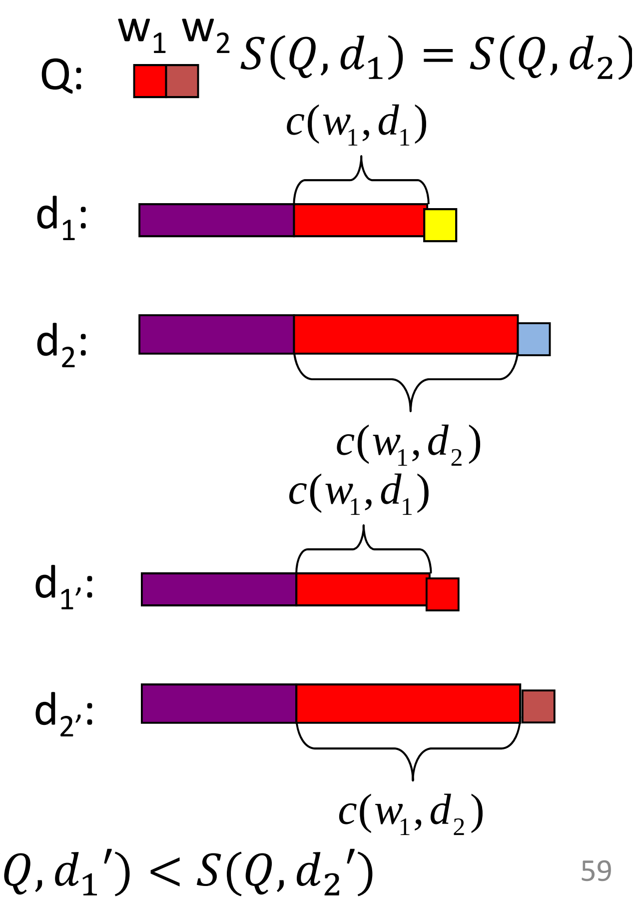
$S(Q \cup \{q\}, D_1) < S(Q \cup \{q\}, D_2)$

Lower Bounding TF Constraints (LB2)

Repeated occurrence of an already matched query term isn't as important as the first occurrence of an otherwise absent query term.

- *LB2*

Let $Q = \{w_1, w_2\}$ be a query with two terms w_1 and w_2 . Assume $td(w_1) = td(w_2)$.
If d_1 and d_2 are two documents such that $c(w_2, d_1) = c(w_2, d_2) = 0$, $c(w_1, d_1) > 0$, $c(w_1, d_2) > 0$, and $S(Q, d_1) = S(Q, d_2)$, then $S(Q, d_1 \cup \{w_1\} - \{t_1\}) < S(Q, d_2 \cup \{w_2\} - \{t_2\})$, for all t_1 and t_2 such that $t_1 \in d_1, t_2 \in d_2, t_1 \notin Q \; and \; t_2 \notin Q$.



No retrieval model satisfies both constraints

Model	LB1	LB2	Parameter and/or query restrictions
BM25	Yes	**No**	*b* and k_1 should not be too large
PIV	Yes	**No**	*s* should not be too large
PL2	**No**	**No**	c should not be too small
DIR	**No**	Yes	μ should not be too large; query terms should be discriminative

Can we "fix" this problem for all the models in a general way?

Solution:
a general approach to lower-bounding TF normalization

- Current retrieval model:

Term frequency → $c(t,D)$; Document length → $|D|$

$$F\left(c(t,D), |D|, \ldots\right)$$

- Lower-bounded retrieval model:

$$\begin{cases} F\left(c(t,D), |D|, \ldots\right) + F\left(0, l, \ldots\right) & \text{If } c(t,D)=0 \\ F\left(c(t,D), |D|, \ldots\right) + F\left(\delta, l, \ldots\right) & \text{Otherwise} \end{cases}$$

Appropriate Lower Bound (the second term, $F(0, l, \ldots)$ / $F(\delta, l, \ldots)$)

Example: Dir+, a lower-bounded version of the query likelihood function

Dir:

$$\sum_{q \in Q \cap D} c(q,Q) \cdot \log\left(1 + \frac{c(q,D)}{\mu \cdot p(w|C)}\right) + |Q| \cdot \log \frac{\mu}{\mu + |D|}$$

Dir+:

$$\sum_{q \in Q \cap D} c(q,Q) \cdot \left[\log\left(1 + \frac{c(q,D)}{\mu \cdot p(w|C)}\right) + \log\left(1 + \frac{\delta}{\mu \cdot p(w|C)}\right)\right] + |Q| \cdot \log \frac{\mu}{\mu + |D|}$$

Dir+ incurs almost no additional computational cost

Example: BM25+, a lower-bounded version of BM25

BM25: $$\sum_{t \in Q \cap D} \frac{(k_3+1)\cdot c(t,Q)}{k_3+c(t,Q)} \cdot \frac{(k_1+1)\cdot c(t,D)}{k_1\left(1-b+b\frac{|D|}{avdl}\right)+c(t,D)} \cdot \log\frac{N+1}{df(t)}$$

BM25+: $$\sum_{t \in Q \cap D} \frac{(k_3+1)\cdot c(t,Q)}{k_3+c(t,Q)} \cdot \left[\frac{(k_1+1)\cdot c(t,D)}{k_1\left(1-b+b\frac{|D|}{avdl}\right)+c(t,D)} + \delta\right] \cdot \log\frac{N+1}{df(t)}$$

BM25+ incurs almost no additional computational cost

The proposed approach can fix or alleviate the problem of all these retrieval models

		LB1	LB2
Traditional retrieval models	BM25	Yes	No
	PIV	Yes	No
	PL2	No	No
	DIR	No	Yes
Lower-bounded retrieval models	BM25+	Yes	Yes
	PIV+	Yes	Yes
	PL2+	Yes	Yes
	DIR+	Cond.	Yes

BM25+ Improves over BM25

Query	Method	WT10G	WT2G	Terabyte	Robust04
Short	BM25	0.1879	0.3104	0.2931	0.2544
	BM25+	**0.1962**	0.3172	**0.3004**	**0.2553**
	BM25+ $(\delta = 1.0)$	0.1927	**0.3178**	0.2997	0.2548
Verbose	BM25	0.1745	0.2484	0.2234	0.2260
	BM25+	**0.1850**	**0.2624**	0.2336	0.2274
	BM25+ $(\delta = 1.0)$	0.1841	0.2565	**0.2339**	**0.2275**

For details, see

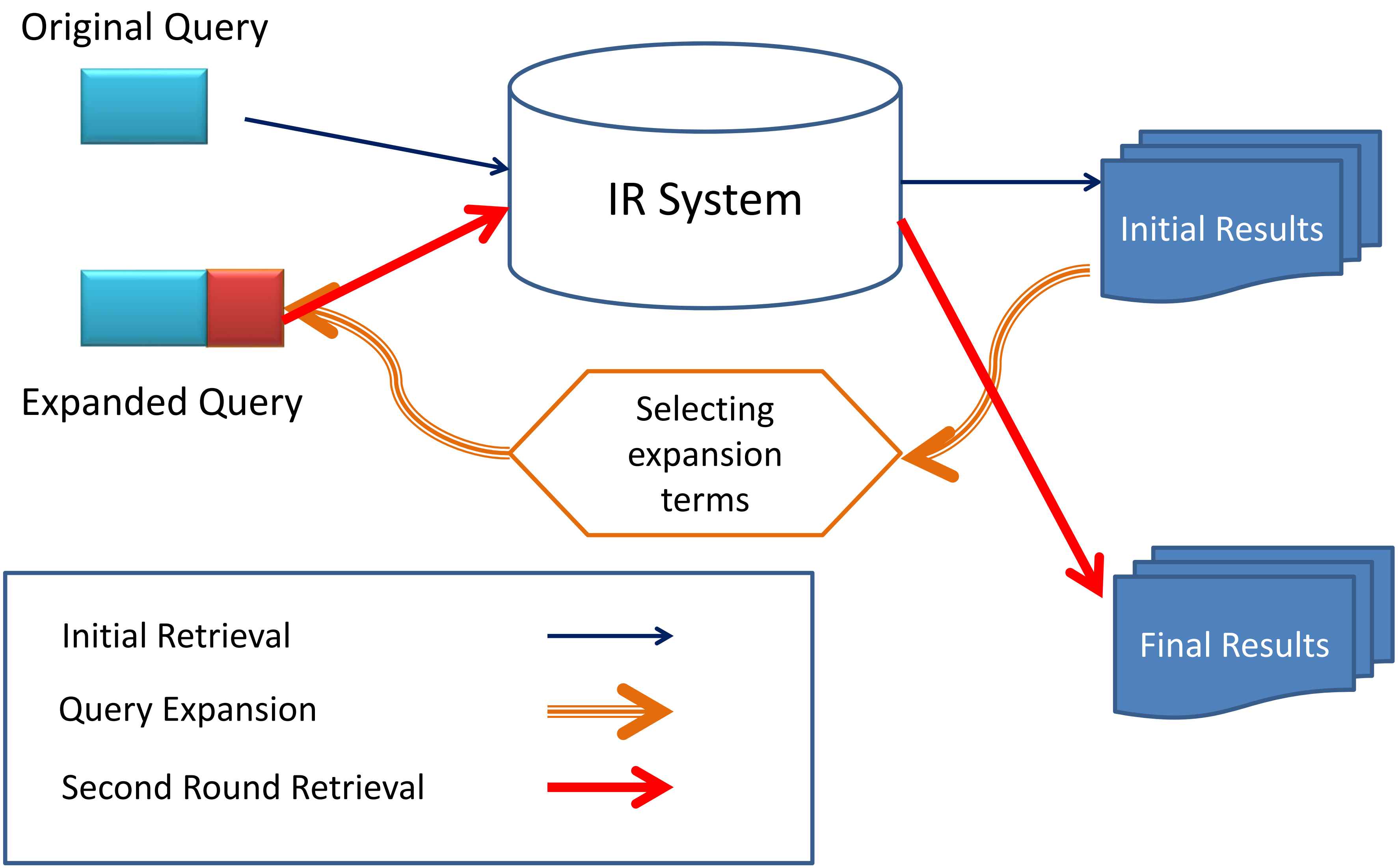
Yuanhua Lv, ChengXiang Zhai, **Lower Bounding Term Frequency Normalization**, *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (**CIKM'11**)*, page 7-16, 2011.

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Axiomatic Analysis of Pseudo-Relevance Feedback Models

Pseudo-Relevance Feedback



Existing PRF Methods

- **Mixture model [Zhai&Lafferty 2001b]**
- **Divergence minimization [Zhai&Lafferty 2001b]**
- Regularized mixture model [Tao et. al. 2006]
- Relevance model [Lavrenko et al. 2001]
- EDCM (extended dirichlet compound multinomial) [Xu&Akella 2008]
- DRF Bo2 [Amati et al. 2003]
- **Log-logistic model [Cinchant et al. 2010]**
- …

Motivation for the DF Constraint

Performance Comparison

Settings	Mixture Model	Log-logistic model	Divergence minimization
Robust-A	0.280	**0.292**	0.263

Log-logistic model is more effective because of
- **It select better feedback terms**
- **It assigns more appropriate weight for expansion terms.**

$\boldsymbol{Expansion\ terms}: \boldsymbol{intersect}$

Settings	MIX	LL	DIV
Robust-A	0.246	**0.257**	0.24
Trec-1&2-A	0.242	**0.245**	0.234
Robust-B	0.253	**0.262**	0.226
Trec-1&2-B	0.261	**0.265**	0.247

Expansion terms: diff

Settings	MIX	LL	DIV
Robust-A	0.03	**0.11**	0.009
Trec-1&2-A	0.03	**0.09**	0.009
Robust-B	0.03	**0.10**	0.015
Trec-1&2-B	0.021	**0.112**	0.005

Motivation for the DF Constraint

Performance Comparison

Settings	Mixture Model	Log-logistic model	Divergence minimization
Robust-A	0.280	**0.292**	0.263
Trec-1&2-A	0.263	**0.284**	0.254
Robust-B	0.282	**0.285**	0.259
Trec-1&2-B	0.273	**0.294**	0.257

μ(FDF)

Settings	MIX	LL	DIV
Robust-A	6.4	7.21	8.41
Trec-1&2-A	7.1	7.8	8.49
Robust-B	9.9	11.9	14.4
Trec-1&2-B	12.0	13.43	14.33

Mean IDF

Settings	MIX	LL	DIV
Robust-A	4.33	5.095	2.36
Trec-1&2-A	3.84	4.82	2.5
Robust-B	4.36	4.37	1.7
Trec-1&2-B	3.82	4.29	2.0

PRF Heuristic Constraints

[Clinchant and Gaussier, 2011a] [Clinchant and Gaussier, 2011b]

- **Document frequency constraint**
  - Feedback terms should receive higher weights when they occur more in the feedback set.

Let ϵ>0 and w_1 and w_2 two words such that

(1) IDF(w_1)= IDF(w_2)

(2) The distribution of the frequencies of w_1 and w_2 in the feedback set are given by:

T(w_1)=(x_1, x_2,…, x_j, 0,…,0)

T(w_2)=(x_1, x_2,…, x_j-ϵ, ϵ,…,0)

with $\forall x_i > 0$, and $x_j -$ ϵ>0

(hence $FTF(w_1) = FTF(w_2)$ and $FDF(w_2) = FDF(w_1) + 1$).

Then: $FW(w_1) < FW(w_2)$

Understanding the DF constraint

Performance Comparison

Settings	Mixture Model	Log-logistic model	Divergence minimization
Robust-A	0.280	**0.292**	0.263
Trec-1&2-A	0.263	**0.284**	254
Robust	0.282	**0.285**	259
Trec	0.273	**0.294**	257

Violate DF constraint

Satisfy DF constraint

Satisfy DF constraint, but IDF effect is not sufficiently enforced

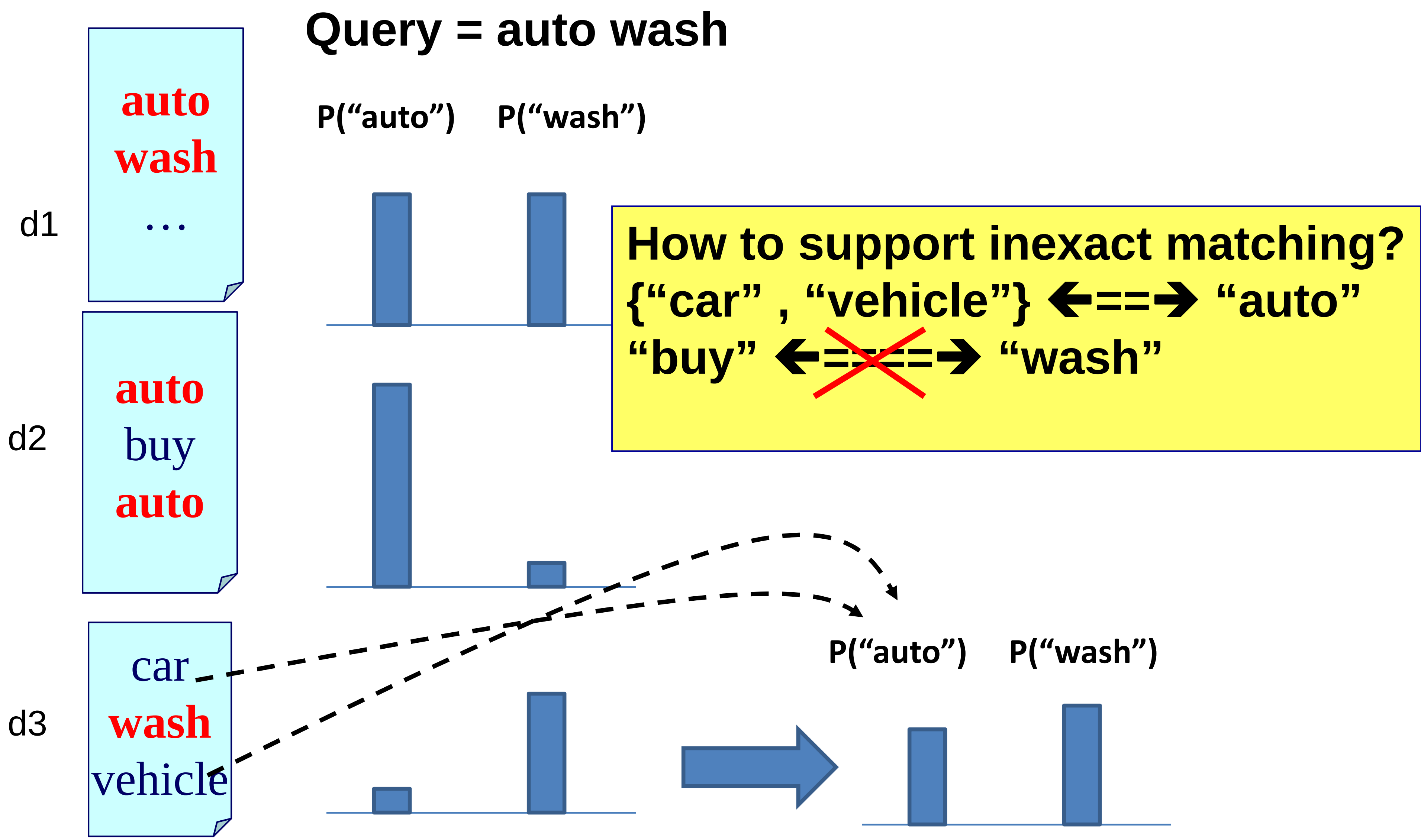
PRF Heuristic Constraints

[Clinchant and Gaussier, 2011a] [Clinchant and Gaussier, 2011b]

- **Document frequency constraint**
  - Feedback terms should receive higher weights when they occur more in the feedback set.
- **Document score constraint**
  - Document with higher score should be given more weight in the feedback weight function.
- **Proximity constraint**
  - Feedback terms should be close to query terms in documents.

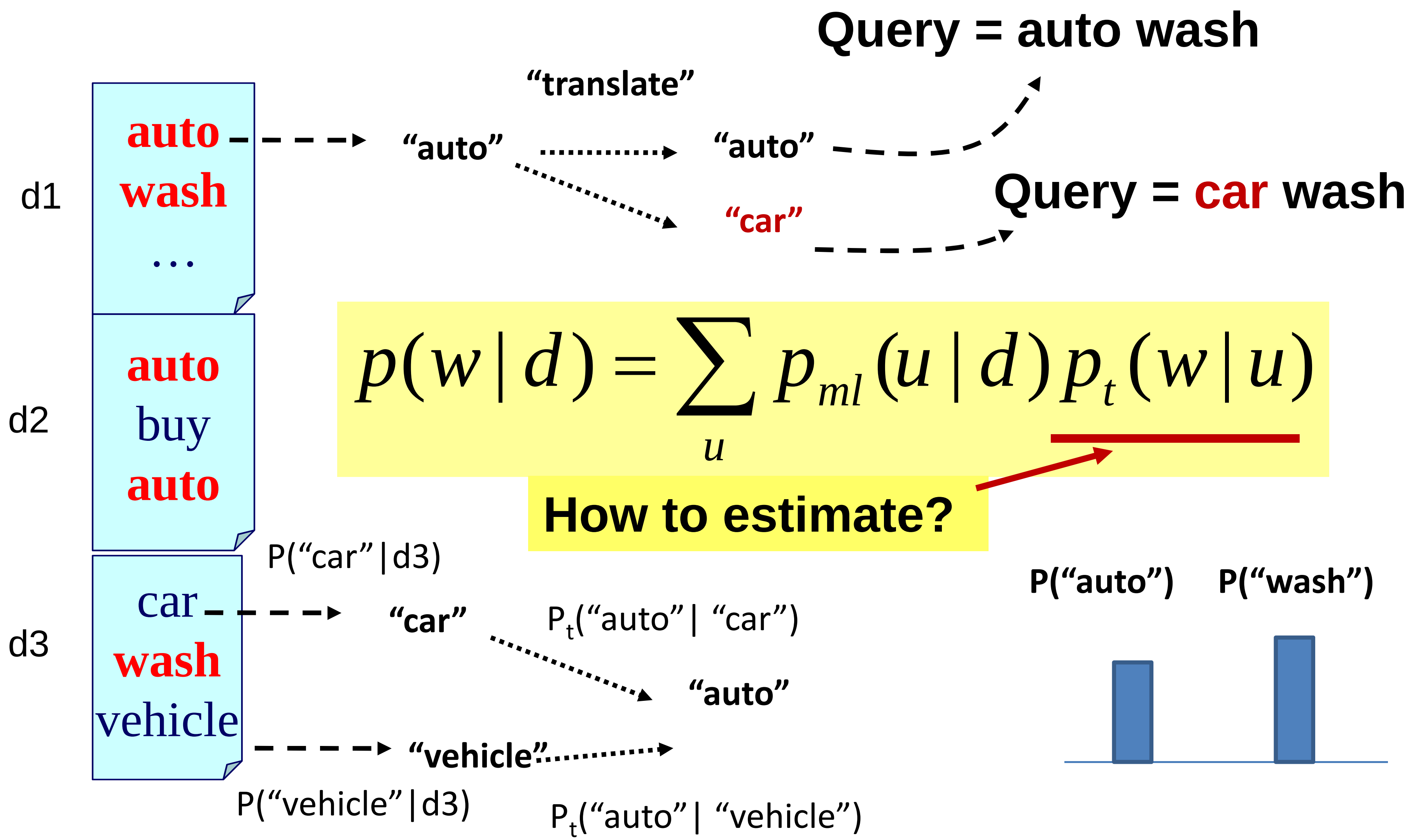
Axiomatic Analysis of Translational Model

The Problem of Vocabulary Gap



Translation Language Models for IR

[Berger & Lafferty 1999]



$$p(w|d) = \sum_{u} p_{ml}(u|d)\, p_t(w|u)$$

Estimation of Translation Model: $p_t(w|u)$

$p_t(w|u)$ = Pr(d mentions u → d is about w)

Supervised learning on (document, query) pairs:
- Synthetic queries [Berger & Lafferty 99]
- Take document title as a query [Jin et al. 02]

Limitations:
1. Can't translate into words not seen in the training queries
2. Computational complexity

Heuristic estimation based on Mutual Information: more efficient, coverage, & effective [Karimzadehgan and Zhai, SIGIR 2010].

Axiomatic Analysis of Translational Model

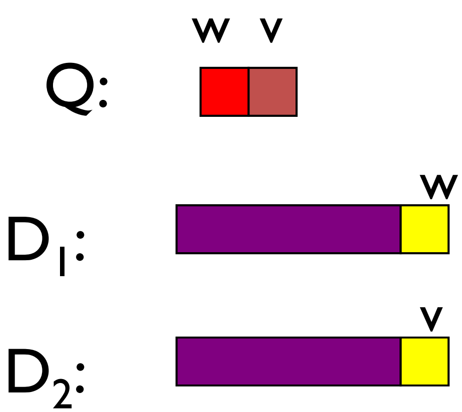
[Karimzadehgan & Zhai 2012]

- Is there a better method than Mutual Information?
- How do we know whether one estimation method is better than another one?
- Is there any better way than pure empirical evaluation?
- Can we *analytically* prove the optimality of a translation language model?

General Constraint 1: Constant Self-Trans. Prob.

***C1:* In order to have a reasonable retrieval behavior, for all translation language models, the self-translation probability should be the same (constant).**

$$\forall v \text{ and } w, p(w|w) = p(v|v)$$



$$p(w|D_1) = p(v|D_2)$$
$$\mathrm{p(v|C) = p(w|C)}$$

$$p(w,v|D_1) = [\sum_u p(u|D_1)p(w|u)] * p_{smooth}(v|C)$$
$$= p(w|D_1) * p(w|w) * p_{smooth}(v|C)$$

$$p(w,v|D_2) = p(v|D_2) * p(v|v) * p_{smooth}(w|C)$$

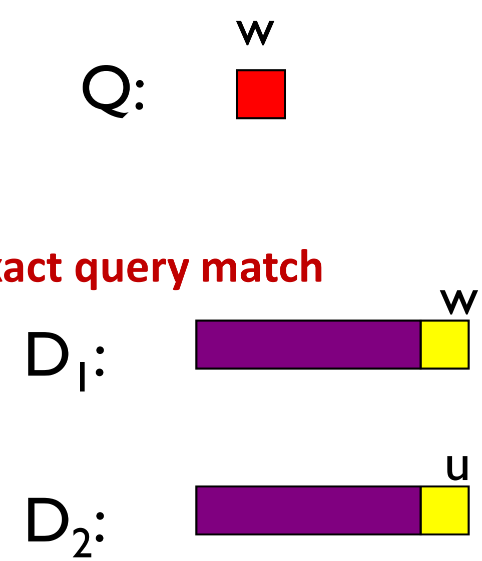
If p(w|w)>p(v|v), D1 would be (unfairly) favored

General Constraint 2

***C2:* Self-translation probability should be larger than translating any other words to this word.**

$$\forall u \text{ and } w, p(w|w) > p(w|u)$$

Exact query match



$$p(w|D_1) = p(w|D_1) * p(w|w)$$

$$p(w|D_2) = p(u|D_2) * p(w|u)$$

Since $p(w|D_1) = p(u|D_2)$ ⇒

The constraint must be satisfied to ensure a document with exact matching gets higher score.

General Constraint 3

***C3:* A word is more likely to be translated to itself than translating into any other words.**

$$\forall u \text{ and } w, p(w|w) > p(u|w)$$

Again to avoid over-rewarding inexact matches

Constraint 4 – Co-occurrence

C4: ***if word u occurs more times than word v in the context of word w and both words u and v co-occur with all other words similarly, the probability of translating word u to word w should be higher.***

$$if\ c(w,u) > c(w,v) and \sum_{w'} c(w',u) = \sum_{w'} c(w',v)$$

$$p(w|u) > p(w|v)$$

Q: "Europe"

D: … "Copenhagen …"

D': … "Chicago …"

"Europe" co-occurs more with "Copenhagen" than with "Chicago"

p(Europe | Copenhagen) > p(Europe | Chicago)

Constraint 5 – Co-occurrence

***C5:** if both u and v equally co-occur with word w but v co-occurs with many other words than word u, the probability of translating word u to word w is higher.*

$$if\ c(w,u) = c(w,v) and \sum_{w'} c(w',u) < \sum_{w'} c(w',v)$$

⬇

$$p(w|u) > p(w|v)$$

Q: "Copenhagen"

D: … "Denmark" …

D': … "Europe" …

p(Copenhagen | Denmark) > p(Copenhagen | Europe)

Analysis of Mutual Information-based Translation Language Model

$$I(w;u) = \sum_{X_w=0,1} \sum_{X_u=0,1} p(X_w, X_u) log \frac{p(X_w, X_u)}{p(X_w)p(X_u)}$$

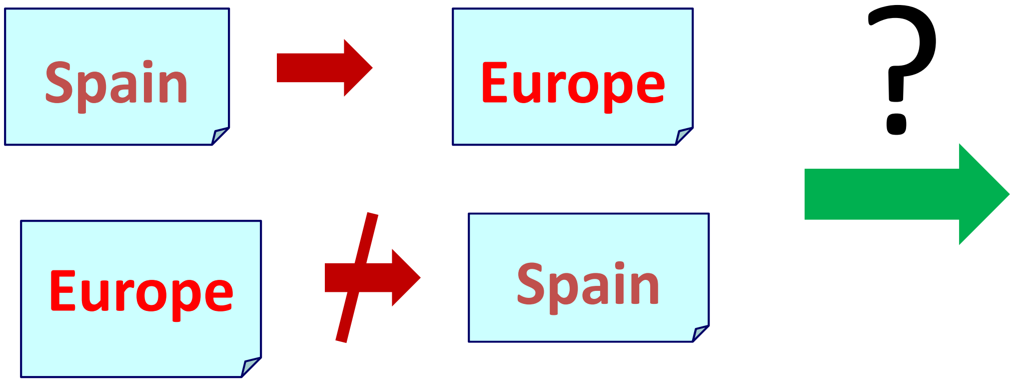
$$p_{mi}(w|u) = \frac{I(w;u)}{\sum_{w'} I(w';u)}$$

It only satisfies C3:

$$\forall u \text{ and } w, p(w|w) > p(u|w)$$

Can we design a method to better satisfy the constraints?

New Method: Conditional Context Analysis



$p(Europe|Spain)$ **high**

$p(Spain|Europe)$ **low**

Main Idea:

… … **Europe** … …. **Spain** … ….

… … **Europe** … …. **Spain** … ….

… … **Europe** … …. **Spain** … ….

… … **Europe** … …. **France** … ….

… … **Europe** … …. **France** … ….

… … … …. … ….

P(Spain|Europe)=3/5
P(Europe|Spain) =3/3

Conditional Context Analysis: Detail

- **Use the frequency of seeing word *w* in the context of word *u* to estimate *p(w|u)*.**
- **See *w* often in the context of *u* → high *p(w|u)***

$$p(w|u) = \frac{c(w,u)+1}{\sum_{w'} c(w',u) + |V|}$$

Satisfies more constraints than MI

However, C1 is not satisfied by either method

$$\forall v \text{ and } w, p(w|w) = p(v|v)$$

Heuristic Adjustment of Self-Translation Probability

Old way (non-constant self translation)

$$p_t(w \mid u) = \begin{cases} \alpha + (1-\alpha)\, p(u \mid u) & \text{w = u} \\ (1-\alpha)\, p(w \mid u) & w \neq u \end{cases}$$

New way (constant self translation)

$$p'(u|u) = s (s \geq 0.5)$$

$$p'(w|u) = \frac{(1-s)p(w|u)}{\sum_{v \neq u} p(v|u)}$$

Conditional-based Approach Works better than Mutual Information-based

Cross validation results

Data	MAP				Precision @10			
	BL	MI	Cond		BL	MI	Cond	
TREC7	0.1852	0.1854	0.1864*+		0.4180	0.42	0.418	
WSJ	0.2600	0.2658	0.275*+		0.424	0.44	0.448	
DOE	0.1740	0.1750	0.1758*		0.1913	0.1956	0.2043	

Upper bound results

Data	MAP				Precision @10			
	BL	MI	Cond		BL	MI	Cond	
TREC7	0.1852	0.1885	0.1887*		0.4180	0.42	0.446	
WSJ	0.2600	0.2708	0.2778*+		0.424	0.44	0.448	
DOE	0.1740	0.1813	0.1868*+		0.1913	0.1956	0.2086	

Constant Self-Translation Probability Improves Performance

Cross validation results

Data	MAP				Precision @10			
	MI	CMI	Cond	CCond	MI	CMI	Cond	CCond
TREC7	0.1854	0.1872+	0.1864	0.1920*^	0.42	0.408	0.418	0.418
WSJ	0.2658	0.267+	0.275	0.278*^	0.44	0.442	0.448	0.448
DOE	0.1750	0.1774+	0.1758	0.1844*^	0.1956	0.2	0.2043	0.2

Upper bound results

Data	MAP				Precision @10			
	MI	CMI	Cond	CCond	MI	CMI	Cond	CCond
TREC7	0.1885	0.1905+	0.1887	0.1965*^	0.42	0.41	0.418	0.418
WSJ	0.2708	0.2717+	0.2778	0.2800*^	0.44	0.448	0.448	0.45
DOE	0.1813	0.1841+	0.1868	0.1953*^	0.1956	0.2043	0.2086	0.2086

Outline

- Motivation
- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions
- Beyond Basic Retrieval Models
- Summary

Updated Answers

- Why do {BM25, PIV, PL, DIR, …} tend to perform similarly even though they were derived in very different ways?

  They shar

  Relevance more accurately modeled with constraints

  These properties are more important than how each is derived

- Why are they better than many other variants?

  Other variants don't have all the "nice properties"

- Why does it seem to be hard to beat these strong baseline methods?

  We don't h

  We didn't find a constraint that they fail to satisfy

- Are they hitting the ceiling of bag-of-words assumption?
  - If yes, how can we prove it?
  -

  No, they have NOT hit the ceiling yet!

Need to formally define "the ceiling" (= complete set of "nice properties")

Summary: Axiomatic Relevance Hypothesis

- Formal retrieval function constraints for modeling relevance
- Axiomatic analysis as a way to assess optimality of retrieval models
- Inevitability of heuristic thinking in developing retrieval models for bridging the theory-effectiveness gap
- Possibility of leveraging axiomatic analysis to improve the state of the art models
- Axiomatic Framework = constraints + constructive function space based on existing or new models and theories

What we've achieved so far

- A large set of formal constraints on retrieval functions
- A number of new functions that are more effective than previous ones
- Some specific questions about existing models that may potentially be addressed via axiomatic analysis
- A general axiomatic framework for developing new models
  - Definition of formal constraints
  - Analysis of constraints (analytical or empirical)
  - Improve a function to better satisfy constraints

For a comprehensive list of the constraints propose so far, check out:

http://www.eecis.udel.edu/~hfang/AX.html

You are invited to join the mailing list of axiomatic analysis for IR!!!

groups.google.com/forum/#!forum/ax4ir

Mailing list: AX4IR@googlegroup.com

Two unanswered "why questions" that may benefit from axiomatic analysis

- The derivation of the query likelihood retrieval function relies on 3 assumptions: (1) query likelihood scoring; (2) independency of query terms; (3) collection LM for smoothing; however, it can't explain why some apparently reasonable smoothing methods perform poorly
- No explanation why other divergence-based similarity function doesn't work well as the asymmetric KL-divergence function D(Q||D)

Open Challenges

- Does there exist a complete set of constraints?
  - If yes, how can we define them?
  - If no, how can we prove it?
- How do we evaluate the constraints?
  - How do we evaluate a constraint? (e.g., should the score contribution of a term be bounded? In BM25, it is.)
  - How do we evaluate a set of constraints?
- How do we define the function space?
  - Search in the neighborhood of an existing function?
  - Search in a new function space?

Open Challenges

- How do we check a function w.r.t. a constraint?
  - How can we quantify the degree of satisfaction?
  - How can we put constraints in a machine learning framework? Something like maximum entropy?
- How can we go beyond bag of words? Model pseudo feedback? Cross-lingual IR?
- Conditional constraints on specific type of queries? Specific type of documents?

Possible Future Scenario 1: Impossibility Theorems for IR

- We will find inconsistency among constraints
- Will be able to prove impossibility theorems for IR
  - Similar to Kleinberg's impossibility theorem for clustering

J. Kleinberg. An Impossibility Theorem for Clustering. Advances in Neural Information Processing Systems (NIPS) 15, 2002

Future Scenario 2: Sufficiently Restrictive Constraints

- We will be able to propose a comprehensive set of constraints that are sufficient for deriving a unique (optimal) retrieval function
  - Similar to the derivation of the entropy function

C. E. Shannon**,** A mathematical theory of communication**,** *Bell system technical journal*, Vol. 27 (1948) Key: citeulike:1584479

Future Scenario 3 (most likely): Open Set of Insufficient Constraints

- We will have a large set of constraints without conflict, but insufficient for ensuring good retrieval performance
- Room for new constraints, but we'll never be sure what they are
- We need to combine axiomatic analysis with a constructive retrieval functional space and supervised machine learning

References

Axiomatic Approaches (1)

- [Bruza&Huibers, 1994] Investigating aboutness axioms using information fields. P. Bruza and T. W. C. Huibers. SIGIR 1994.
- [Fang, et. al. 2004] A formal study of information retrieval heuristics. H. Fang, T. Tao and C. Zhai. SIGIR 2004.
- [Fang&Zhai, 2005] An exploration of axiomatic approaches to information retrieval. H. Fang and C. Zhai, SIGIR 2005.
- [Fang&Zhai, 2006] Semantic term matching in axiomatic approaches to information retrieval. H. Fang and C. Zhai, SIGIR 2006.
- [Tao&Zhai, 2007] An exploration of proximity measures in information retrieval. T. Tao and C. Zhai, SIGIR 2007.
- [Cummins&O'Riordan, 2007] An axiomatic comparison of learned term-weighting schemes in information retrieval: clarifications and extensions, Artificial Intelligence Review, 2007.
- [Fang, 2008] A Re-examination of query expansion using lexical resources. H. Fang. ACL 2008.
- [Na et al., 2008] Improving Term Frequency Normalization for multi-topical documents and application to language modeling approaches. S. Na, I Kang and J. Lee. ECIR 2008.
- [Gollapudi&Sharma, 2009] An axiomatic approach for result diversification. S. Gollapudi and Sharma, WWW 2009.
- [Zheng&Fang, 2010] Query aspect based term weighting regularization in information retrieval. W. Zheng and H. Fang. ECIR 2010.

Axiomatic Approaches (2)

- [Clinchant&Gaussier,2010] Information-based models for Ad Hoc IR. S. Clinchant and E. Gaussier, SIGIR 2010.
- [Clinchant&Gaussier, 2011] Retrieval constraints and word frequency distributions a log-logistic model for IR. S. Clinchant and E. Gaussier. Information Retrieval. 2011.
- [Fang et al., 2011] Diagnostic evaluation of information retrieval models. H. Fang, T. Tao and C. Zhai. TOIS, 2011.
- [Lv&Zhai, 2011a] Lower-bounding term frequency normalization. Y. Lv and C. Zhai. CIKM 2011.
- [Lv&Zhai, 2011b] Adaptive term-frequency normalization for BM25. Y. Lv and C. Zhai. CIKM 2011. [Lv&Zhai, 2011] When documents are very long, BM25 fails! Y. Lv and C. Zhai. SIGIR 2011.
- [Clinchant&Gaussier, 2011a] Is document frequency important for PRF? S. Clinchant and E. Gaussier. ICTIR 2011.
- [Clinchant&Gaussier, 2011b] A document frequency constraint for pseudo-relevance feedback models. S. Clinchant and E. Gaussier. CORIA 2011.
- [Zhang et al., 2011] How to count thumb-ups and thumb-downs: user-rating based ranking of items from an axiomatic perspective. D. Zhang, R. Mao, H. Li and J. Mao. ICTIR 2011.
- [Lv&Zhai, 2012] A log-logistic model-based interpretation of TF normalization of BM25. Y. Lv and C. Zhai. ECIR 2012.
- [Wu&Fang, 2012] Relation-based term weighting regularization. H. Wu and H. Fang. ECIR 2012.

Axiomatic Approaches (3)

- [Li&Gaussier, 2012] An information-based cross-language information retrieval model. B. Li and E. Gaussier. ECIR 2012.
- [Karimzadehgan&Zhai, 2012] Axiomatic analysis of translation language model for information retrieval. M. Karimzadehgan and C. Zhai. ECIR 2012.

Other References (1)

- [Salton et al. 1975] A theory of term importance in automatic text analysis. G. Salton, C.S. Yang and C. T. Yu. Journal of the American Society for Information Science, 1975.
- [Singhal et al. 1996] Pivoted document length normalization. A. Singhal, C. Buckley and M. Mitra. SIGIR 1996.
- [Maron&Kuhn 1960] On relevance, probabilistic indexing and information retrieval. M. E. Maron and J. L. Kuhns. Journal o fhte ACM, 1960.
- [Harter 1975] A probabilistic approach to automatic keyword indexing. S. P. Harter. Journal of the American Society for Information Science, 1975.
- [Robertson&Sparck Jones 1976] Relevance weighting of search terms. S. Robertson and K. Sparck Jones. Journal of the American Society for Information Science, 1976.
- [van Rijsbergen 1977] A theoretical basis for the use of co-occurrence data in information retrieval. C. J. van Rijbergen. Journal of Documentation, 1977.
- [Robertson 1977] The probability ranking principle in IR. S. E. Robertson. Journal of Documentation, 1977.

Other References (2)

- [Robertson 1981] Probabilistic models of indexing and searching. S. E. Robertson, C. J. van Rijsbergen and M. F. Porter. Information Retrieval Search, 1981.
- [Robertson&Walker 1994] Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. S. E. Robertson and S. Walker. SIGIR 1994.
- [Ponte&Croft 1998] A language modeling approach to information retrieval. J. Ponte and W. B. Croft. SIGIR 1998.
- [Hiemstra&Kraaij 1998] Twenty-one at TREC-7: ad-hoc and cross-language track. D. Hiemstra and W. Kraaij. TREC-7. 1998.
- [Zhai&Lafferty 2001] A study of smoothing methods for language models applied to ad hoc information retrieval. C. Zhai and J. Lafferty. SIGIR 2001.
- [Lavrenko&Croft 2001] Relevance-based language models. V. Lavrenko and B. Croft. SIGIR 2001.
- [Kurland&Lee 2004] Corpus structure, language models, and ad hoc information retrieval. O. Kurland and L. Lee. SIGIR 2004.

Other References (3)

- [van Rijsbergen 1986] A non-classical logic for information retrieval. C. J. van Rijsbergen. The Computer Journal, 1986.
- [Wong&Yao 1995] On modeling information retrieval with probabilistic inference. S. K. M. Wong and Y. Y. Yao. ACM Transactions on Information Systems. 1995.
- [Amati&van Rijsbergen 2002] Probabilistic models of information retrieval based on measuring the divergence from randomness. G. Amati and C. J. van Rijsbergen. ACM Transactions on Information Retrieval. 2002.
- [He&Ounis 2005] A study of the dirichlet priors for term frequency normalization. B. He and I. Ounis. SIGIR 2005.
- [Gey 1994] Inferring probability of relevance using the method of logistic regression. F. Gey. SIGIR 1994.
- [Zhai&Lafferty 2001] Model-based feedback in the language modeling approach to information retrieval. C. Zhai and J. Lafferty. CIKM 2001.
- [Tao et al. 2006] Regularized estimation of mixture models for robust pseudo-relevance feedback. T. Tao and C. Zhai. SIGIR 2006.

Other References (4)

- [Amati et al. 2003] Foundazione Ugo Bordoni at TREC 2003: robust and web track. G. Amati and C. Carpineto, G. Romano and F. U. Bordoni. TREC 2003.
- [Xu and Akella 2008] A new probabilistic retrieval model based on the dirichlet compound multinomial distribution. Z. xu and R. Akella. SIGIR 2008.
- [Berger&Lafferty 1999] Information retrieval as statistical translation. A. Berger and J. Lafferty. SIGIR 1999.
- [Kleinberg 2002] An Impossibility Theorem for Clustering. J. Kleinberg. Advances in Neural Information Processing Systems, 2002
- [Shannon 1948] A mathematical theory of communication. C. E Shannon. *Bell system technical journal*, 1948.