# Axiomatic Analysis and Optimization of Information Retrieval Models

**SIGIR 2014 Tutorial**

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

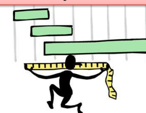
ChengXiang Zhai
Dept. of Computer Science
University of Illinois at Urbana-Champaign
USA
http://www.cs.illinois.edu/homes/czhai

TIMAN InfoLab — Gold Coast SIGIR 2014

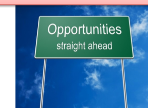
## Goal of Tutorial

Axiomatic Approaches to IR

Review major research progress



Discuss promising research directions




You can expect to learn
- Basic methodology of axiomatic analysis and optimization of retrieval models
- Novel retrieval models developed using axiomatic analysis

## Organization of Tutorial

- Motivation
- Axiomatic Analysis and Optimization: Early Work
- Axiomatic Analysis and Optimization: Recent Work
- Summary

TIMAN InfoLab 3

## Organization of Tutorial

- Motivation
- Axiomatic Analysis and Optimization: Early Work
- Axiomatic Analysis and Optimization: Recent Work
- Summary

TIMAN InfoLab 4

## Search is everywhere, and part of everyone's life

Web Search — Desk Search — Enterprise Search — Social Media Search — Site Search

TIMAN InfoLab 5

## 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 |

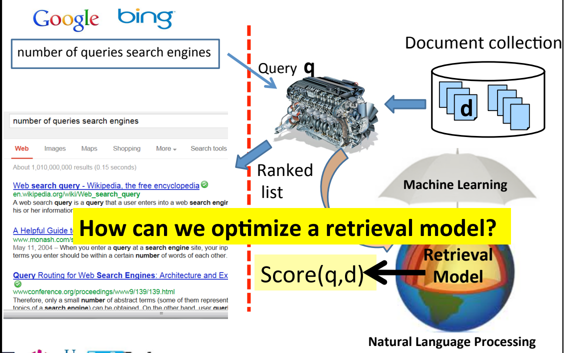
… …

**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

TIMAN InfoLab 6

## Behind all the search boxes…






## Retrieval model = computational definition of "relevance"

S("world cup schedule", d)

s("world", d) s("cup", d) s("schedule", d)

How many times does "schedule" occur in d?
**Term Frequency** (TF): c("schedule", d)

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

How often do we see "schedule" in the entire collection C?
**Document Frequency**: df("schedule")
P("schedule"|C)



## Scoring based on bag of words in general

q w d

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. 1996]
- BM25 [Robertson & Walker 1994]
- PL2 [Amati & van Rijsbergen 2002]
- Query likelihood with Dirichlet prior (DIR) [Ponte & Croft 1998], [Zhai & Lafferty 2001]

**but many others failed to work well…**



## Some 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 (divergence from randomness)**

$$\sum_{w \in q \cap d} c(w,q) \cdot \frac{tfn_w^d \cdot \log_2(tfn_w^d \cdot \lambda_w) + \log_2 e \cdot (\frac{1}{\lambda_w} - tfn_w^d) + 0.5 \cdot \log_2(2\pi \cdot tfn_w^d)}{tfn_w^d + 1}$$

$$tfn_w^d = c(w,d) \cdot \log_2(1+c \cdot \frac{avdl}{|d|}), \lambda_w = \frac{N}{c(w,C)}$$

## PIV, DIR, BM25 and PL2 tend to perform similarly.

**Performance Comparison (MAP)**

| | 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 |

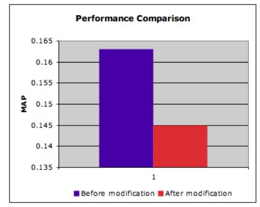
**Why do they tend to perform similarly even though they were derived in very different ways?**

TIMAN InfoLab 13

## Performance sensitive to small variations in a formula

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

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




**Why is a state of the art retrieval function better than many other variants?**

TIMAN InfoLab 14

## Additional Observations

- PIV (vector space model) **1996**
- DIR (language modeling approach) **2001**
- BM25 (classic probabilistic model) **1994**
- PL2 (divergence from randomness) **2002**

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

- "Ad Hoc IR – Not Much Room for Improvement" [Trotman & Keeler 2011]
- "Has Adhoc Retrieval Improved Since 1994?" [Armstrong et al. 2009]

**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?**

TIMAN InfoLab 15

## Suggested Answers: Axiomatic Analysis

- Why do these methods 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")**

TIMAN InfoLab 16

## Organization of Tutorial

- Motivation
- Axiomatic Analysis and Optimization: Early work
- Axiomatic Analysis and Optimization: Recent Work
- Summary

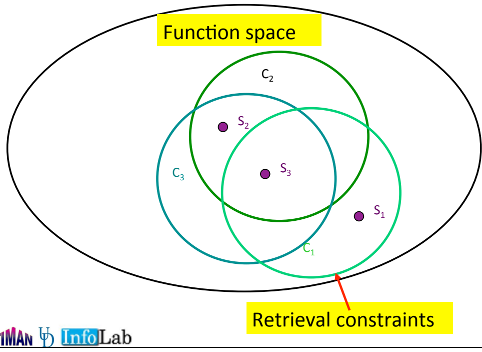
TIMAN InfoLab 17

## 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 $C$

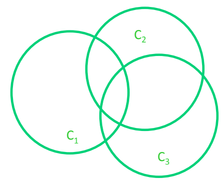
TIMAN InfoLab 18

## Axiomatic Analysis and Optimization




## Axiomatic Analysis and Optimization: Early Work – Outline

- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions






## Different functions, but similar heuristics

- **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 (divergence from randomness)**

$$\sum_{w\in q\cap d} c(w,q)\cdot\frac{tfn_w^d\cdot\log_2(tfn_w^d\cdot\lambda_w)+\log_2 e\cdot(\frac{1}{\lambda_w}-tfn_w^d)+0.5\cdot\log_2(2\pi\cdot tfn_w^d)}{tfn_w^d+1}$$

$$tfn_w^d = c(w,d)\cdot\log_2(1+c\cdot\frac{avdl}{|d|}),\ \lambda_w=\frac{N}{c(w,C)}$$

**TF weighting**

**IDF weighting**

**Length Norm.**



Are they performing well because they implement similar retrieval heuristics?

**Can we formally capture these necessary retrieval heuristics?**

For details, see
- Hui Fang, Tao Tao and ChengXiang Zhai: **A Formal Study of Information Retrieval Heuristics**, SIGIR'04.
- Hui Fang, Tao Tao and ChengXiang Zhai: **Diagnostic Evaluation of Information Retrieval Models**. ACM Transaction of Information Systems, 29(2), 2011.



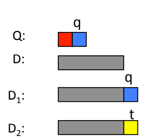
## Term Frequency Constraints (TFC1)

**TF weighting heuristic I:**
Give a higher score to a document with more occurrences of a query term.

- ***TFC1***

Let $Q$ be a query and $D$ be a document.

If $q \in Q$ and $t \notin Q$,
then $S(Q, D\cup\{q\}) > S(Q, D\cup\{t\})$




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



## Term Frequency Constraints (TFC2)

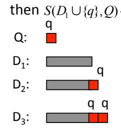
**TF weighting heuristic II:**
Require that the amount of increase in the score due to adding a query term must decrease as we add more terms.

- ***TFC2***

Let $Q$ be a query with only one query term $q$.

Let $D_1$ be a document.

then $S(D_1\cup\{q\},Q)-S(D_1,Q) > S(D_1\cup\{q\}\cup\{q\},Q)-S(D_1\cup\{q\},Q)$




$S(D_2,Q)-S(D_1,Q) > S(D_3,Q)-S(D_2,Q)$

## Term Frequency Constraints (TFC3)

**TF weighting heuristic III:**

Favor a document with more distinct query terms.

- ***TFC3***

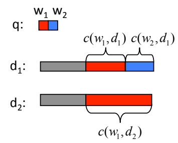
  Let $q$ be a query and $w_1$, $w_2$ be two query terms.

  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 $S(d_1, q) > S(d_2, q)$.




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



## Length Normalization Constraints (LNCs)

**Document length normalization heuristic:**

Penalize long documents(LNC1);

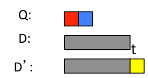
Avoid over-penalizing long documents (LNC2) .

- ***LNC1***

  Let $Q$ be a query and $D$ be a document.

  If $t$ is a non-query term,

  then $S(D \cup \{t\}, Q) < S(D, Q)$



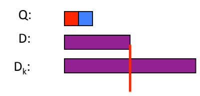

$S(Q, D') < S(Q, D)$

- ***LNC2***

  Let $Q$ be a query and $D$ be a document.

  If $D \cap Q \neq \phi$, and $D_k$ is constructed by concatenating $D$ with itself $k$ times,

  then $S(D_k, Q) \geq S(D, Q)$




$S(Q, D_k) \geq S(Q, D)$



## TF & Length normalization Constraint (TF-LNC)

**TF-LN heuristic:**

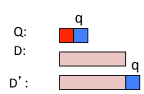
Regularize the interaction of TF and document length.

- ***TF-LNC***

  Let $Q$ be a query and $D$ be a document.

  If $q$ is a query term,

  then $S(D \cup \{q\}, Q) > S(D, Q)$.




$S(Q, D') > S(Q, D)$



## 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 |



## Disclaimers

- Given a retrieval heuristic, there could be multiple ways of formalizing it as constraints.
- When formalizing a retrieval constraint, it is necessary to check its dependency on other constraints.



## Weak or Strong Constraints?

**The Heuristic captured by 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 $td(q_1)>=td(q_2)$ and $c(q_1,D_1)>=c(q_1,D_2)$, 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 $D_1$ contains only $q_1$ and $D_2$ contains only $q_2$.
    If $td(q_1)>=td(q_1)$, we have $S(Q,D_1 \cup D) \geq S(Q,D_2 \cup D)$.

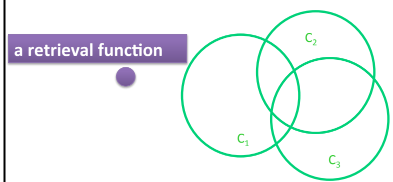
## Key Steps of Constraint Formalization

- Identify desirable retrieval heuristics
- Formalize a retrieval heuristic as reasonable retrieval constraints.
- After formalizing a retrieval constraint, check how it is related to other retrieval constraints.
  - Properties of a constraint set
    - Completeness
    - Redundancy
    - Conflict



## Axiomatic Analysis and Optimization: Early Work – Outline

- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions






## An Example of Constraint Analysis

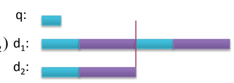
**PIV:**
$$S(d,q) = \sum_{w \in q \cap d} \frac{1 + ln(1 + ln(c(w,d)))}{1 - s + s\frac{|d|}{avdl}} \cdot c(w,q) \cdot ln\frac{N+1}{df(w)}$$

***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 $S(d_1,q) \geq S(d_2,q)$




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

**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 $S(d_1,q) \geq S(d_2,q)$

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

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

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



## 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 $|d_2| = avdl$,

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

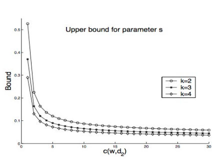
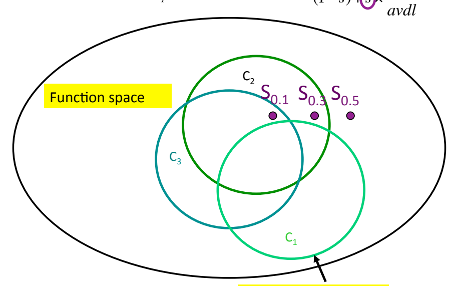



Figure 1: Upper bound of parameter s.



## An Example of Constraint Analysis

**PIV:**
$$S(q,d) = \sum_{t \in d \cap q} c(t,q) \times \log\frac{N+1}{df(t)} \times \frac{1+\log(1+\log(c(t,d)))}{(1-s)+s\times\frac{|d|}{avdl}}$$



## Review: Axiomatic Relevance Hypothesis

- 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 $C = \{c_1, ..., c_k\}$
  - 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 $C$

TIMAN InfoLab 37

## 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
  - 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.
  - Constraint analysis reveals optimal ranges of parameter values

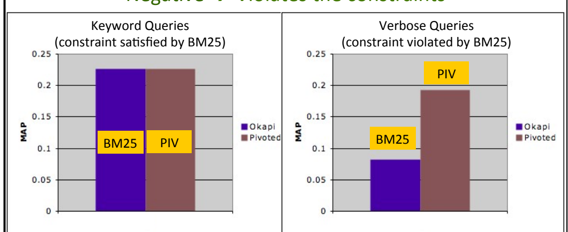
TIMAN InfoLab 38

## Violation of Constraints → Poor Performance

- **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




TIMAN InfoLab

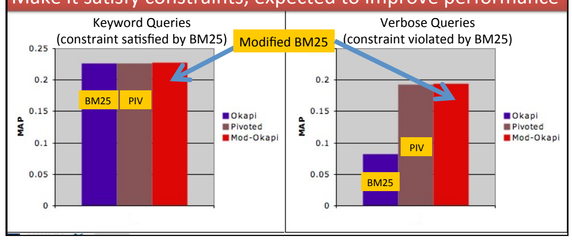
## Constraints Analysis → Guidance for Improving an Existing Retrieval Function

- **Modified Okapi BM25**

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

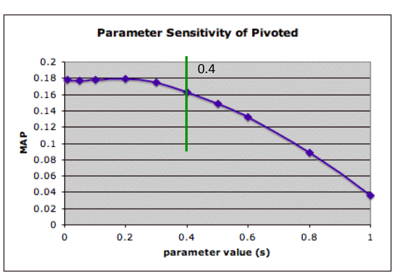
$$\boxed{\log \frac{N+1}{df(t)}}$$

Make it satisfy constraints; expected to improve performance




## Conditional Satisfaction of Constraints → Parameter Bounds

- **PIV** LNC2 → s<0.4




41

TIMAN InfoLab

## Systematic Analysis of 4 State of the Art Models

[Fang et al. 2011]

| Function | TFCs | TDC | [illegible] | [illegible] | [illegible] |
| --- | --- | --- | --- | --- | --- |
| [illegible] | [illegible] | [illegible] | [illegible] | C1* | C2* |
| [illegible] | [illegible] | [illegible] | [illegible] | C3 | [illegible] |
| [illegible] | [illegible] | [illegible] | [illegible] | C4 | C4 |
| [illegible] (Modified) | [illegible] | [illegible] | [illegible] | 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

TIMAN InfoLab 42

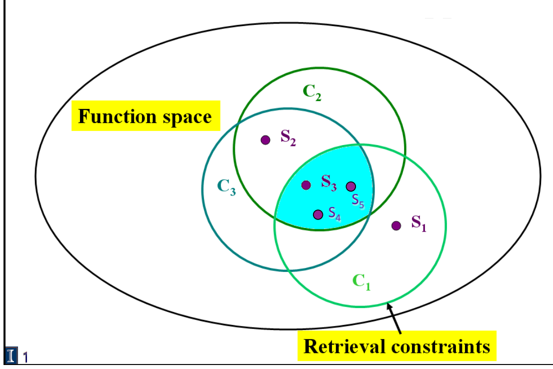
## Perturbation tests:

## An empirical way of analyzing the constraints

For details, see
- Hui Fang, Tao Tao and ChengXiang Zhai: **Diagnostic Evaluation of Information Retrieval Models**. ACM Transaction of Information Systems, 29(2), 2011.

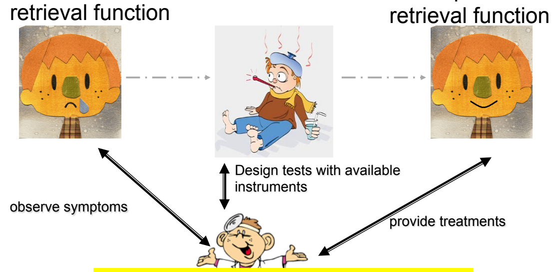


## What if constraint analysis is NOT sufficient?






## Medical Diagnosis Analogy



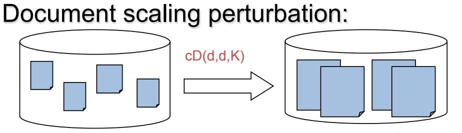

How to find available instruments?
How to design diagnostic tests?



## Relevance-Preserving Perturbations

- Perturb term statistics
- Keep relevance status

Document scaling perturbation:




concatenate every document with itself K times



## 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 |

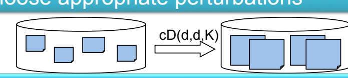


## Length Scaling Test (LV3)

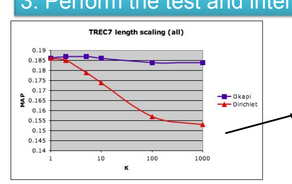
1. Identify the aspect to be diagnosed

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!

## Summary of All Tests

| Tests | What to measure? |
|---|---|
| **Length variance reduction (LV1)** | The gain on length normalization |
| **Length variance amplification (LV2)** | The robustness to larger document variance |
| **Length scaling (LV3)** | The ability at avoid over-penalizing long documents |
| **Term noise addition (TN)** | The ability to penalize long documents |
| **Single query term growth (TG1)** | The ability to favor docs with more distinct query terms |
| **Majority query term growth (TG2)** | Favor documents with more query terms |
| **All query term growth (TG3)** | Balance TF and LN more appropriately |

TIMAN InfoLab 49

## Diagnostic Results for DIR

$$S(Q,D)=\sum_{t\in Q\cap D} c(t,Q)\cdot \log\left(1+\frac{c(t,D)}{\mu\cdot p(t\mid C)}\right) - \log\left(1+\frac{|D|}{\mu}\right)\cdot |Q|$$

- Weaknesses
  - over-penalizes long documents (TN, LV3)
  - fails to implement one desirable property of TF (TG1)
- Strengths
  - performs better in a document with higher document length variance (LV2)
  - implements another desirable property of TF (TG2)

TIMAN InfoLab 50

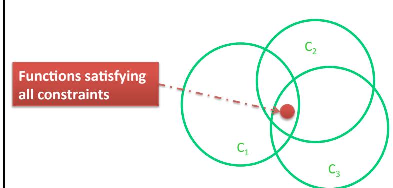
## Identifying the weaknesses makes it possible to improve the performance

| | MAP | | | P@30 | | |
|---|---|---|---|---|---|---|
| | trec8 | wt2g | FR | trec8 | wt2g | FR |
| DIR | 0.257 | 0.302 | 0.202 | 0.365 | 0.331 | 0.151 |
| Imp.D. | 0.263 | 0.323 | 0.228 | 0.373 | 0.345 | 0.166 |

TIMAN InfoLab 51

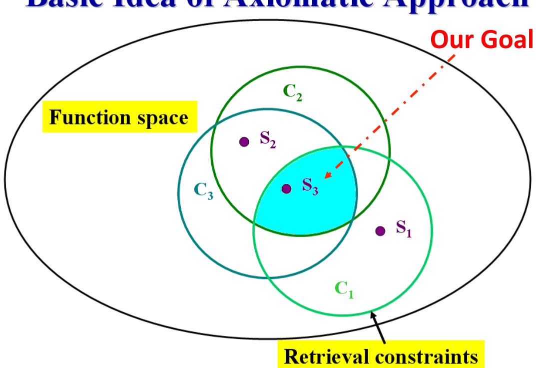
## Axiomatic Analysis and Optimization: Early Work – Outline

- Formalization of Information Retrieval Heuristics
- Analysis of Retrieval Functions with Constraints
- Development of Novel Retrieval Functions




TIMAN InfoLab 52

## Basic Idea of Axiomatic Approach




## Three Questions

- How to define the constraints?

  **We've talked about that; more later**
- How to define the function space?

  **One possibility: leverage existing state of the art functions**
- How to search in the function space?

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

For details, see
- Hui Fang and ChengXiang Zhai: **An Exploration of Axiomatic Approaches to Information Retrieval**, SIGIR'05.

TIMAN InfoLab 54

## Inductive Definition of Function Space

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

Define the function space *inductively*

Q: cat

D: dog big

Primitive weighting function (f)

S(Q,D) = S( , ) = f ( , )

Query growth function (h)

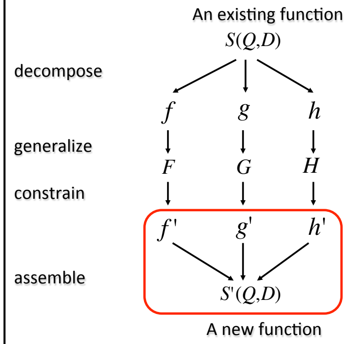
S(Q,D) = S( , ) = S( , )+h( , , )

Document growth function (g)

S(Q,D) = S( , ) = S( , )+g( , , )

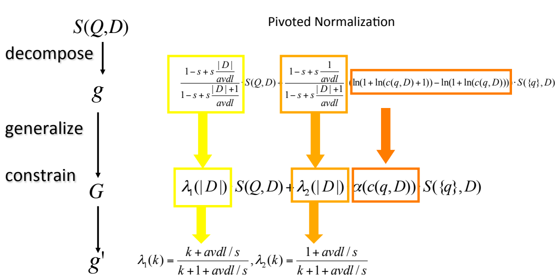


## Derivation of New Retrieval Functions






## Derivation of New Document Growth Function



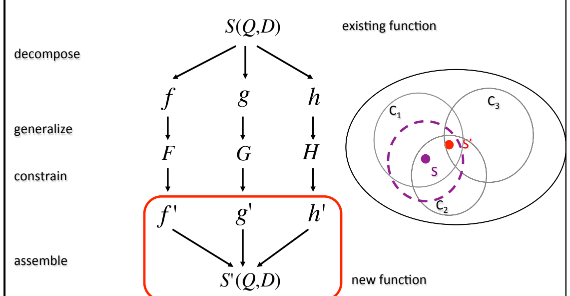

$$\frac{1-s+s\frac{|D|}{avdl}}{1-s+s\frac{|D|+1}{avdl}} \cdot S(Q,D) + \frac{1-s+s\frac{1}{avdl}}{1-s+s\frac{|D|+1}{avdl}} \cdot (\ln(1+\ln(c(q,D)+1)) - \ln(1+\ln(c(q,D)))) \cdot S(\{q\},D)$$

$$\lambda_1(|D|) \cdot S(Q,D) + \lambda_2(|D|) \cdot \alpha(c(q,D)) \cdot S(\{q\},D)$$

$$\lambda_1(k) = \frac{k + avdl/s}{k+1+avdl/s}, \lambda_2(k) = \frac{1+avdl/s}{k+1+avdl/s}$$



## Derivation of New Retrieval Functions






## A Sample Derived Function based on BM25

[Fang & Zhai 2005]

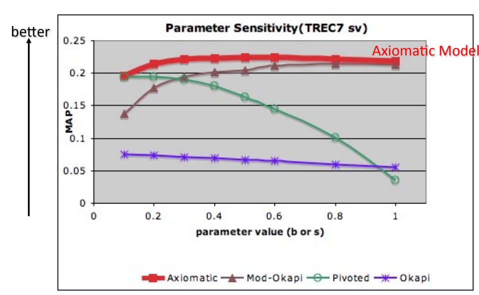
QTF IDF TF

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

length normalization



## The derived function is less sensitive to the parameter setting



## Organization of Tutorial

- Motivation
- Axiomatic Analysis and Optimization: Early Work
- Axiomatic Analysis and Optimization: Recent Work
- Summary



## Axiomatic Analysis and Optimization: Recent Work – Outline

- Lower-bounding TF Normalization
- Axiomatic Analysis of Pseudo-Relevance Feedback Models
- Axiomatic Analysis of Translational Model

For details, see
- Yuanhua Lv and ChengXiang Zhai: **Lower Bounding Term Frequency Normalization**, CIKM'11.



## Review: Constraint Analysis Results

[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**



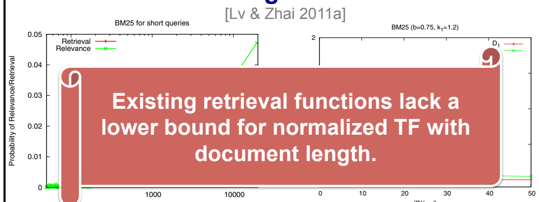
## How to identify more deficiencies?

- We need more constraints!
- But how?



## 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 are overly penalized!**

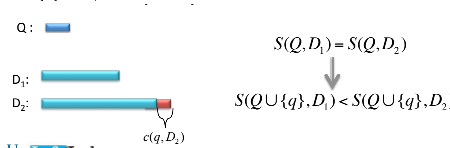
A very long document matching two query terms can have a lower score than a short document matching only one query term



## Lower Bounding TF Constraints (LB1)

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

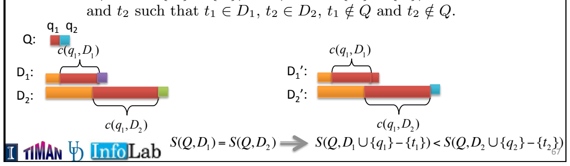
**LB1:** Let $Q$ be a query. Assume $D_1$ and $D_2$ are two documents such that $S(Q, D_1) = S(Q, D_2)$. If we reformulate the query by adding another term $q \notin Q$ into $Q$, where $c(q, D_1) = 0$ and $c(q, D_2) > 0$, then $S(Q \cup \{q\}, D_1) < S(Q \cup \{q\}, D_2)$.



## Lower Bounding TF Constraints (LB2)

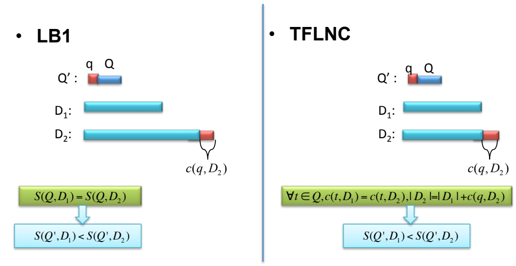
**Repeated occurrence of an already matched query term isn't as important as the first occurrence of an otherwise absent query term.**

**LB2:** Let $Q=\{q_1,q_2\}$ be a query with two terms $q_1$ and $q_2$. Assume $td(q_1)=td(q_2)$, where $td(t)$ can be any reasonable measure of term discrimination value. If $D_1$ and $D_2$ are two documents such that $c(q_2,D_1)=c(q_2,D_2)=0$, $c(q_1,D_1)>0$, $c(q_1,D_2)>0$, and $S(Q,D_1)=S(Q,D_2)$, then $S(Q,D_1\cup\{q_1\}-\{t_1\})<S(Q,D_2\cup\{q_2\}-\{t_2\})$, for all $t_1$ and $t_2$ such that $t_1\in D_1$, $t_2\in D_2$, $t_1\notin Q$ and $t_2\notin Q$.





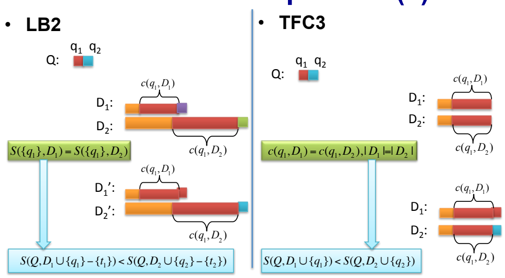

## Constraint Comparison (1)




Both constraints are designed to avoid over-penalizing long documents. However, LB1 is more general since it puts less restriction on the document length.



## Constraint Comparison (2)




Both constraints are designed to favor documents covering more distinct query terms. However, LB2 is more general since it puts less restriction on the document length.



## No retrieval model satisfies both LB 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 | $\mu$ should not be too large; query terms should be discriminative |



## Solution: a general approach to lower-bounding TF normalization

- Current retrieval model:

Term frequency — Document length

$$F\big(c(t,D),|D|,\ldots\big)$$

- Lower-bounded retrieval model:

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

Appropriate Lower Bound



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

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**



## 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** |
| Verbose | BM25 | 0.1745 | 0.2484 | 0.2234 | 0.2260 |
| | BM25+ | **0.1850** | **0.2624** | **0.2336** | **0.2274** |



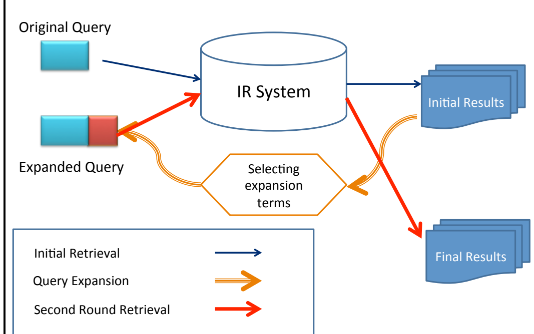
## Axiomatic Analysis and Optimization: Recent Work – Outline

- Lower-bounding TF Normalization
- Axiomatic Analysis of Pseudo-Relevance Feedback Models
- Axiomatic Analysis of Translational Model

For details, see
- Stephane Clinchant and Eric Gaussier: **A Theoretical Analysis of Pseudo-Relevance Feedback Models**, ICTIR'13.



## Pseudo-Relevance Feedback






## Existing PRF Methods

- Mixture model [Zhai&Lafferty 2001b]
- Divergence minimization [Zhai&Lafferty 2001b]
- Geometric 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 PRF Constraints

[Clinchant and Gaussier, 2011a] [ Clinchant and Gaussier, 2011b][ Clinchant and Gaussier, 2013]

**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 |

Robust-A

| Settings | MIX | LL | DIV |
|---|---|---|---|
| Avg (tf) | 62.9 | **46.7** | 53.9 |
| Avg (df) | 6.4 | **7.21** | 8.6 |
| Avg (idf) | 4.3 | **5.1** | 2.2 |

**Log-logistic model is more effective because it selects terms**
- **that are not too common (high IDF and small TF)**
- **that still occur in sufficient number o feedback documents (average DF)**

## PRF Heuristic Constraints

[Clinchant and Gaussier, 2013]

- **TF effect**
  - The feedback weight should increase with the term frequency.
- **Concavity effect**
  - The above increase should be less marked in high frequency ranges.
- **IDF effect**
  - When all other things being equal, the feedback weight of a term with higher IDF value should be larger.
- **Document length effect**
  - The number of occurrences of feedback terms should be normalized by the length of documents they appear in.
- **DF effect**
  - When all other things being equal, terms occurring in more feedback documents should receive higher feedback weights.



## Summary of Constraint Analysis

| | TF | Concave | IDF | Doc Len | DF |
|---|---|---|---|---|---|
| **Mixture** | Y | Cond. | Y | N | N |
| **Div Min** | Y | Y | Cond. | Y | Y |
| **G. Rel.** | Y | Y | N | Y | Y |
| **Bo** | Y | N | Cond. | N | N |
| **Log-Logistic** | Y | Y | Y | Y | Y |

The authors also discussed how to revise the mixture model and geometric relevance model to improve the performance.



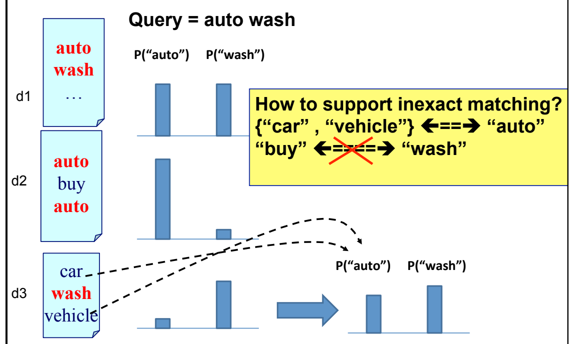
## Axiomatic Analysis and Optimization: Recent Work – Outline

- Lower-bounding TF Normalization
- Axiomatic Analysis of Pseudo-Relevance Feedback Models
- Axiomatic Analysis of Translational Model

For details, see
- Maryam Karimzadehgan and ChengXiang Zhai: **Axiomatic Analysis of Translation Language Model for Information Retrieval**, ECIR'12.



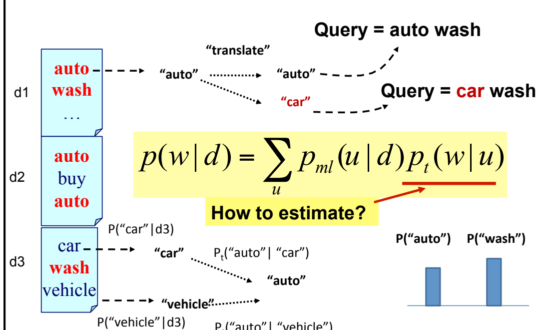
## The Problem of Vocabulary Gap






## Translation Language Models for IR

[Berger & Lafferty 1999]




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



## Axiomatic Analysis of Translational Model

[Karimzadehgan & Zhai 2012]

Estimation of translation model

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

- 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).*

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

Q: w v

$D_1$: w

$D_2$: v

$p(w|D_1) = p(v|D_2)$
$p(v|C) = p(w|C)$

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

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

**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.*

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

Q: w

Exact query match

$D_1$: w

$D_2$: u

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

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

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.*

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

⬇

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

Q: "Australia"

D: … "Brisbane …"

D': … "Chicago …"

"Australia" co-occurs more with "Brisbane" than with "Chicago" ➡

p(Australia | Brisbane) > p(Australia | Brisbane)



## 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.*

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

⬇

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

Q: "Brisbane"

D: … "Queensland"

D': … "Australia" …

p(Brisbane | Queensland) > p(Brisbane | Australia)



## 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

Spain → Europe ? $p(Europe|Spain)$ high

Europe ↛ Spain $p(Spain|Europe)$ 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)

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

New way (constant self translation)

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

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



## 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 |

- **Conditional-based Approach Works better than Mutual Information-based**
- **Constant Self-Translation Probability Improves Performance**



## Organization of Tutorial

- Motivation
- Axiomatic Analysis and Optimization: Early Work
- Axiomatic Analysis and Optimization: Recent Work
- Summary



## Updated Answers: Axiomatic Analysis

- Why do these methods tend to perform similarly even though they were derived in very different ways?
  **Relevance more accurately modeled with constraints**
- 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 didn't find a constraint that they fail to satisfy**
- Are they hitting the ceiling of bag-of-words assumption?
  –
  –
  Nee
  pro

  **No, they have NOT hit the ceiling yet!**

## 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



## 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



## 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?

TIMAN InfoLab 103

# 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

TIMAN InfoLab 104

# 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

TIMAN InfoLab 105

# 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

TIMAN InfoLab 106

# Generalization of the axiomatic analysis process (beyond IR)

1. Set an objective function, e.g.,
   - Ranking: S(Q,D)
   - Diversification: f(D, q, w(), dsim())
2. Identify variables that have impacts to the objective function
3. Formalize constraints based on the variables
   - For each variable, figure out its desirable behavior with respect to the objective function, and these desirable properties would be formalized as axioms (i.e., constraints).
     - Exploratory data analysis
   - Study the relations among multiple variables and formalize the desirable properties of these relations as additional constraints.

TIMAN InfoLab 107

# Generalization of the axiomatic analysis process (beyond IR) (cont.)

4. For all the formalized constraints, study their dependencies and conflicts, and remove redundant constraints.
5. Function Derivation
   - If no conflict constraints, find instantiations of the objective function that can satisfy all constraints.
     - Derive new functions
     - Modify existing ones
   - If there are conflict constraints, study the trade-off and identify scenarios that requires a subset of non-conflict constraints, and then derive functions based on these constraints.

TIMAN InfoLab 108

## Towards General Axiomatic Thinking

- Given a task of designing a function to solve a problem: Y=f(X)
  - Identify properties function f should satisfy
  - Formalize such properties with mathematically well defined constraints
  - Use the constraints to help identify the best function
- Potentially helpful for designing any function
- Constraints can be of many different forms (inequality, equality, pointwise, listwise, etc)
  - Pointwise: For all "a" that satisfies a certain condition, f(a)=b
  - Pairwise: For all a and b that satisfy a certain condition, f(a)>f(b) (or f(a)=f(b))
  - Listwise: For all a1, a2, … and ak that satisfy a certain condition, then f(a1)>f(a2)>…. >f(ak ) (or f(a1)=….=f(ak))



## Axiomatic Thinking & Machine Learning

- Learn f using supervised learning = constrain the choice of f with an empirical objective function (minimizing errors on training data)
- However, the learned functions may violate obvious constraints due to limited training data (the data is almost always limited!)
- Axiomatic thinking can help machine learning by regularizing the function space or suggesting a certain form of the functions
- For example, f(X)=a1*x1+a2*x2+…+ak*xk
  - A simple constraint can be if x2 increases, f(X) should increase (derivative w.r.t. x2 is positive) ➔ a2>0
  - Another constraint can be: the second derivative w.r.t. x2 is negative (i.e., "diminishing return") ➔ the assumed function form is non-optimal; alternative forms should be considered



## Some Examples of Axiomatic Thinking outside IR (1)

- **ProWord: An Unsupervised Approach to Protocol Feature Word Extraction**, by Zhuo Zhang, Zhibin Zhang, Patrick P. C. Lee, Yunjie Liu and Gaogang Xie. INFOCOM, 2014.
  - "*Our idea is inspired by the heuristics in information retrieval such as TF-IDF weighting, and we adapt such heuristics into traffic analysis. ProWord uses a ranking algorithm that maps different dimensions of protocol feature heuristics into different word scoring functions and uses the aggregate score to rank the candidates.*"



## Some Examples of Axiomatic Thinking outside IR (2)

- **A Formal Study of Feature Selection in Text Categorization**, by Yan Xu, Journal of Communication and computer, 2009
  - "*In this paper, we present a formal study of Feature selection (FS) in text categorization. We first define three desirable constraints that any reasonable FS function should satisfy, then check these constraints on some popular FS methods …. Experimental results indicate that the empirical performance of a FS function is tightly related to how well it satisfies these constraints*"



## Some Examples of Axiomatic Thinking outside IR (3)

- **eTuner: Tuning Schema Matching Software Using Synthetic Scenarios**, by Yoonkyong Lee, Mayssam Sayyadian, Anhai Doan and Arnon S. Rosenthal. *VLDB Journal, 2007.*
  - Using constraints to help generate test cases for schema matching
  - Cited [Fang & Zhai 2004] as a relevant work



## The End

# 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.




# Axiomatic Approaches (2)


- [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.
- [Cummins & O'Riordan 2009] Ronan Cummins and Colm O'Riordan. Measuring constraint violations in information retrieval, SIGIR 2009.
- [Zheng&Fang, 2010] Query aspect based term weighting regularization in information retrieval. W. Zheng and H. Fang. ECIR 2010.
- [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.




# Axiomatic Approaches (3)


- [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.
- Shima Gerani, ChengXiang Zhai, Fabio Crestani: Score Transformation in Linear Combination for Multi-criteria Relevance Ranking. ECIR 2012: 256-267




# Axiomatic Approaches (4)


- [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.
- [Cummins and O'Riodan 2012] Ronan Cummins and Colm O'Riordan. A Constraint to Automatically Regulate Document-Length Normalisation, 21st ACM International Conference on Information and Knowledge Management (CIKM), Oct 29 - Nov 2,2012, Maui, Hawaii, USA
- [Clinchant&Gaussier, 2013] A Theoretical Analysis of Pseudo-Relevance Feedback Models. ICTIR 2013.




# 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.


TIMAN InfoLab 121

## 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.


TIMAN InfoLab 122

## 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.
- [Trotman & Keeler 2011] Ad Hoc IR – Not Much Room for Improvement. A. Trotman and D. Keeler, SIGIR 2011.
- [Armstrong et al 2009] Has Adhoc Retrieval Improved Since 1994? T. G. Armstrong, A. Moffat, W. Webber and J. Zobel, SIGIR 2009.


TIMAN InfoLab 123