Ph.D. Thesis
Defense

Web Mining Techniques for Query Log Analysis and Expertise Retrieval

Hongbo Deng

Department of Computer Science and Engineering
The Chinese University of Hong Kong

Date: Sep 2, 2009

Rich Web Data

Web pages

One trillion

unique URLs

Question answer

Yahoo! answer

Web mining

techniques

Scientific litera

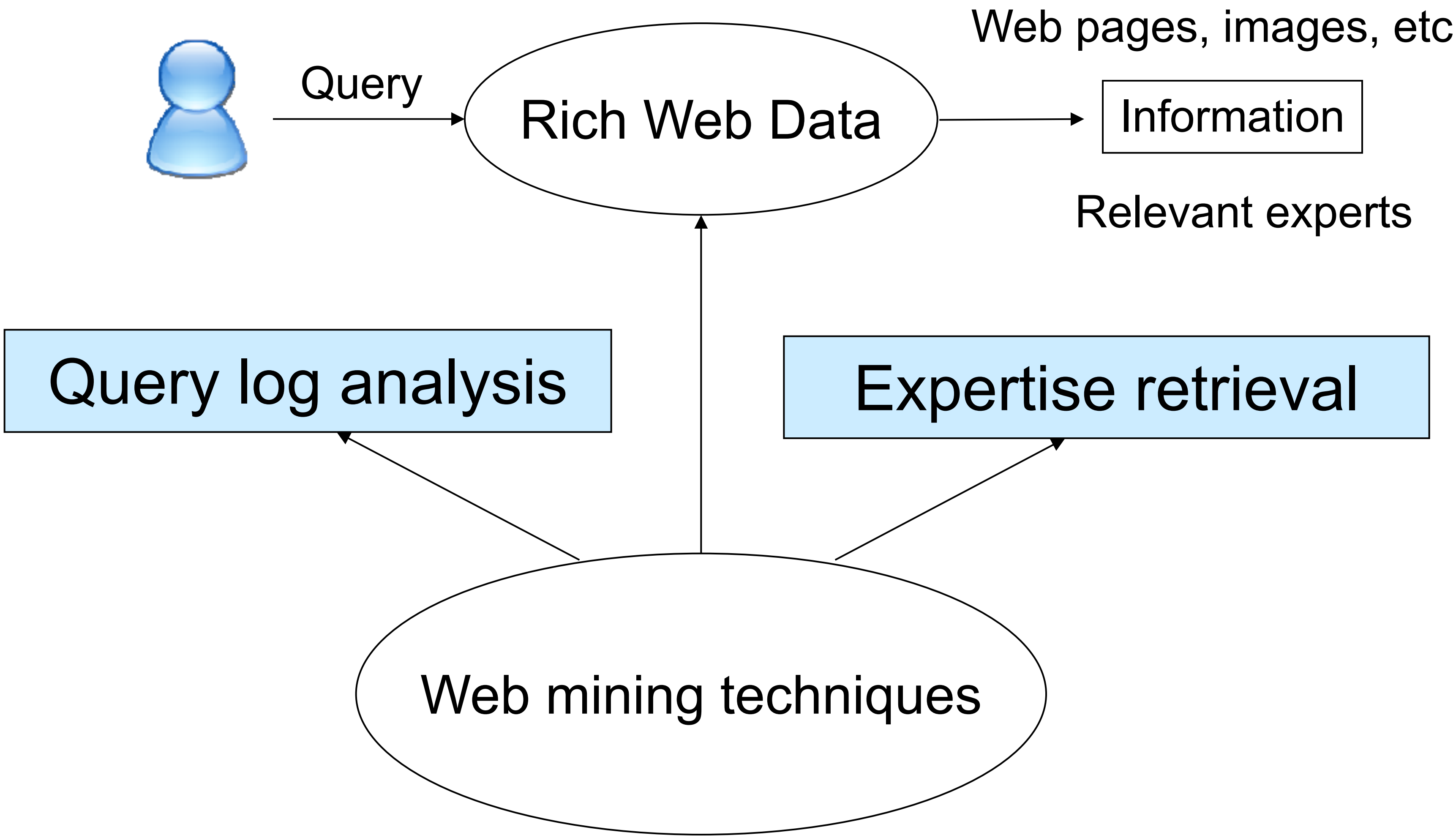
DBLP dat

Query logs

L query log

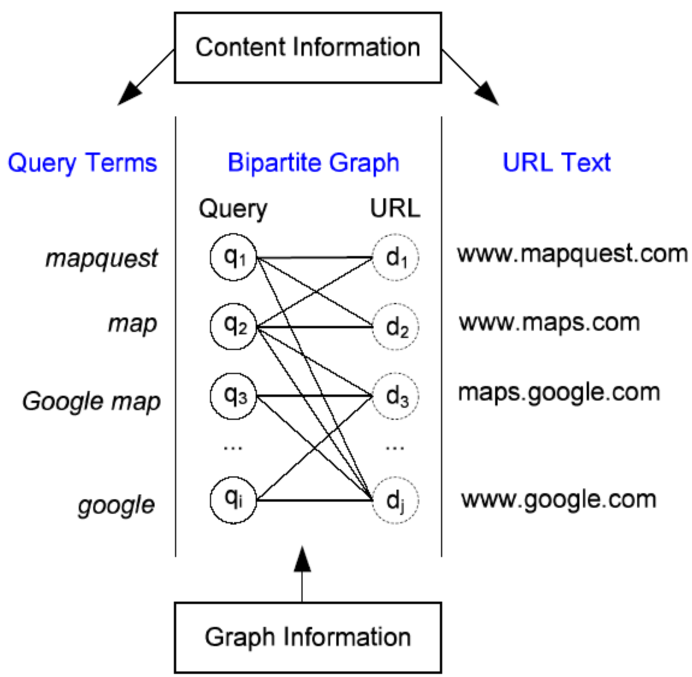
Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Overview

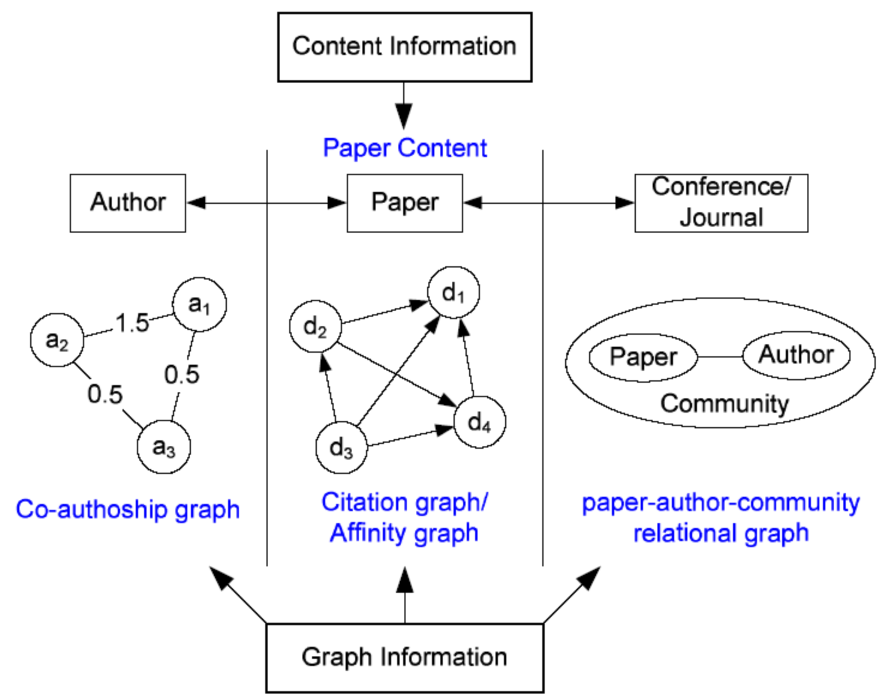

Web pages, images, etc.
Query
Rich Web Data
Information
Relevant experts
Query log analysis
Expertise retrieval
Web mining techniques

Objective

- Combine the content and the graph information
- Leverage IR, link analysis, ML in a unified way



The query log data



The expertise retrieval data

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Outline

- Background: Web Mining Techniques
  - Information retrieval, link analysis, machine learning
- Modeling Bipartite Graph for Query Log Analysis
  - Entropy-biased Models [w/ King-Lyu, SIGIR'09]
  - Co-HITS Algorithm [w/ Lyu-King, KDD'09]
- Modeling Expertise Retrieval
  - Baseline and Weighted Model [w/ King-Lyu, ICDM'08]
  - Graph-based Re-ranking Model [w/ Lyu-King, WSDM'09]
  - Enhanced Models with Community [w/ King-Lyu, CIKM'09]
- Conclusion and Future Work

Background

- Information retrieval models
  - Vector space model
  - Probabilistic model
  - Language model
- Web link analysis
  - PageRank: a link represents a vote
  - HITS: good hubs points to good authorities
  - Other variations
- Machine learning
  - Semi-supervised learning
  - Graph-based regularization framework

Modeling Bipartite Graph for Query Log Analysis

- Many Web data can be modeled as bipartite graphs

Query log data:
Query-URL
bipartite graph

Netflix data:
User-Movie
bipartite graph

DBLP data:
Author-Paper
bipartite graph

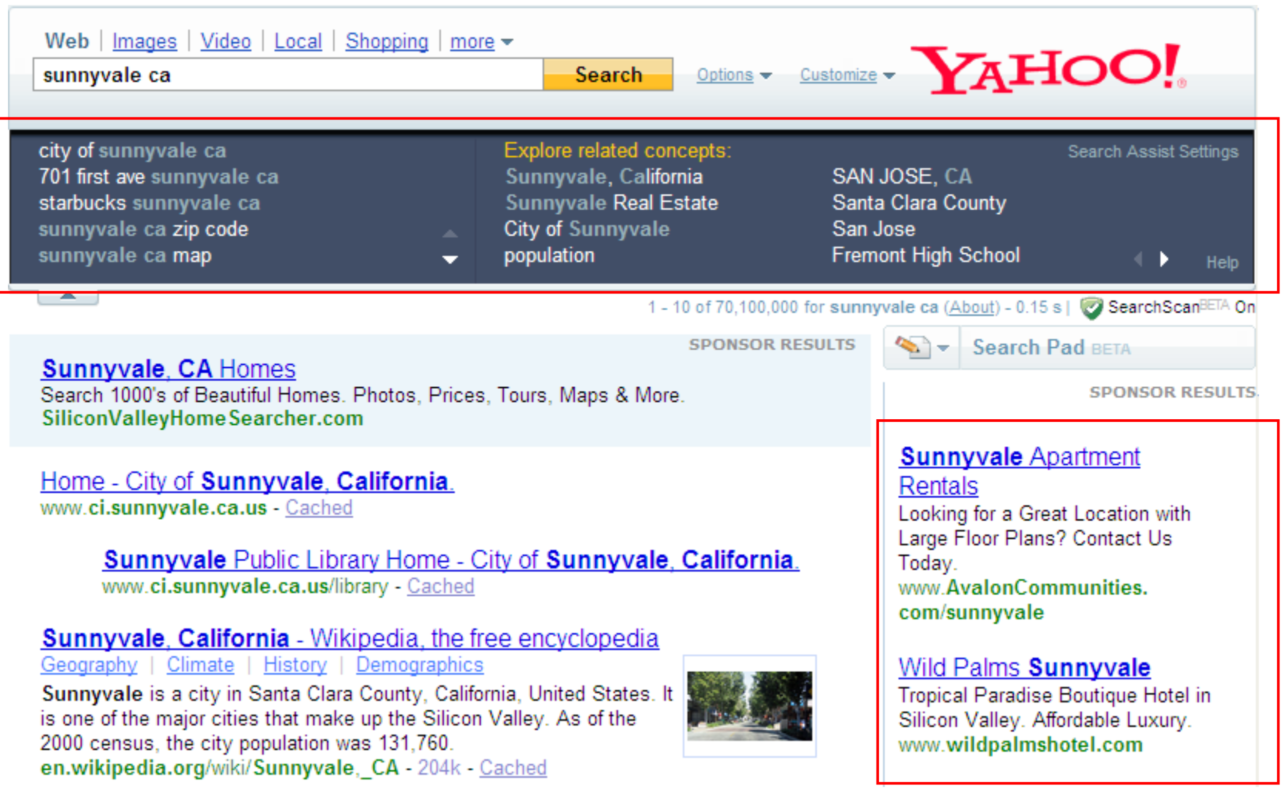
How to weigh the edges of the graph?

How to combine the graph with other info.?

It is very essential to model bipartite graphs for mining these data types.

Query Log Analysis

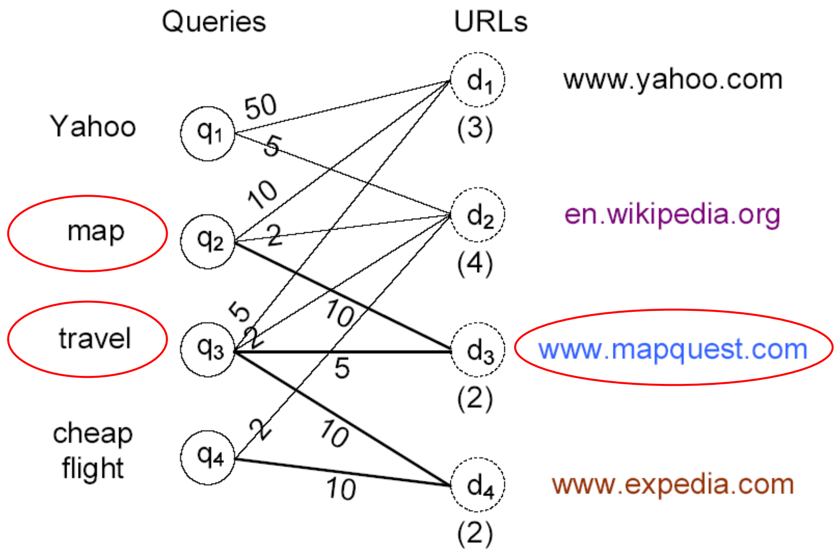
- Query log analysis – improve search engine's capabilities
  - Query suggestion
  - Query classification
  - Targeted advertising
  - Ranking



Click Graph

- Click graph – an important technique
  - A bipartite graph between queries and URLs
  - Edges connect a query with the URLs
  - Capture some semantic relations, e.g., “map” and “travel”

How to utilize and model the click graph to represent queries?



Traditional model based on the *raw click frequency (CF)*

Propose two kinds of models

- Entropy-biased framework
- Co-HITS algorithm

Outline

- Part I: Entropy-biased Framework for Modeling Click Graphs
  - Motivation and Preliminaries
  - Traditional Click Frequency Model
  - Entropy-biased Model
  - Experimental Results
  - Summary

Motivation

Is a single click on different URLs equally important?

- Basic idea
  - Various query-URL pairs should be treated differently
- Intuition
  - Common clicks on less frequent but more specific URLs are of greater value than common clicks on frequent and general URLs

Preliminaries

Query instance: $\langle q, d, u \rangle$

Query: $Q = \{q_1, q_2, ..., q_M\}$

URL: $D = \{d_1, d_2, ..., d_N\}$

User: $U = \{u_1, u_2, ..., u_K\}$

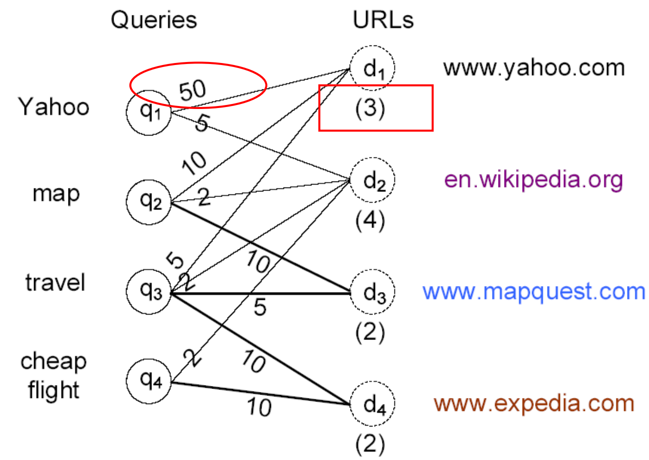


Table 1: Table of Notation.

Symbol	Meaning
C	$M \times N$ query-URL matrix
c_{ij}	Click frequency between query q_i and URL d_j, with the entry (i, j) of the matrix C
uf_{ij}	User frequency between q_i and d_j
$n(d_j)$	Number of queries associated with URL d_j
$idf(d_j)$	Importance of a certain URL d_j
$p(d_j \vert q_i)$	Transition probability from q_i to d_j
$p(q_i \vert d_j)$	Transition probability from d_j to q_i
P_{q2d}	An $M \times N$ query-URL probability matrix
P_{d2q}	An $N \times M$ URL-query probability matrix

Click frequency matrix

C	d_1	d_2	d_3	d_4
q_1	50	5	0	0
q_2	10	2	10	0
q_3	5	2	5	10
q_4	0	2	0	10

Traditional Click Frequency Model

- **Transition probability:** Normalize the *click frequency* (CF)

From query to URL:

$$p(d_j|q_i) = \frac{c_{ij}}{cf(q_i)}$$

$$P_{q2d}:\ \overrightarrow{q_i} = \langle P_{q2d}(i,1), ..., P_{q2d}(i,N)\rangle$$

From URL to query:

$$p(q_i|d_j) = \frac{c_{ij}}{cf(d_j)}$$

$$P_{d2q}:\ \overrightarrow{d_j} = \langle P_{d2q}(j,1), ..., P_{d2q}(j,M)\rangle$$

Click frequency matrix

C	d_1	d_2	d_3	d_4
q_1	50	5	0	0
q_2	$\boxed{10}$	2	$\boxed{10}$	0
q_3	5	2	5	10
q_4	0	2	0	10

$\longrightarrow$

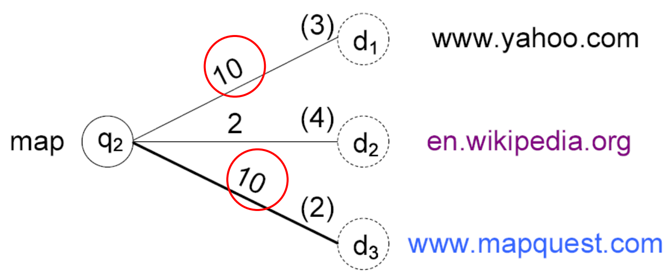
CF transition probabilities

P_{q2d}	d_1	d_2	d_3	d_4
q_1	0.909	0.091	0	0
q_2	$\boxed{0.455}$	0.091	$\boxed{0.455}$	0
q_3	0.227	0.091	0.227	0.455
q_4	0	0.167	0	0.833

- Measure the similarity between queries
  - The most similar query: q2 (“map”) → q1 (“Yahoo”)
  - More reasonable: q2 (“map”) → q3 (“travel”)

Entropy-biased model

Entropy-biased Model



It would be more reasonable to **weigh these two edges differently**

- The more general and highly ranked URL
  - Connect with more queries
  - Increase the ambiguity and uncertainty

Transition probability from a URL to a query

- The entropy of a URL: $$E(d_j) = -\sum_{i \in Q} p(q_i|d_j) \log p(q_i|d_j)$$
  - Suppose $$p(q_i|d_j) = \frac{1}{n(d_j)}$$ → The number of queries that connected with d_j

    Query frequency
  - Tend to be proportional to the *$n(d_j)$*

$$E(d_j) = \log n(d_j)$$

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Entropy → Discriminative Ability

- Entropy increase, discriminative ability decrease
  - Be inversely proportional to each other
  - A URL with a high query frequency is less discriminative overall
- Inverse query frequency
  - Measure the discriminative ability of the URL

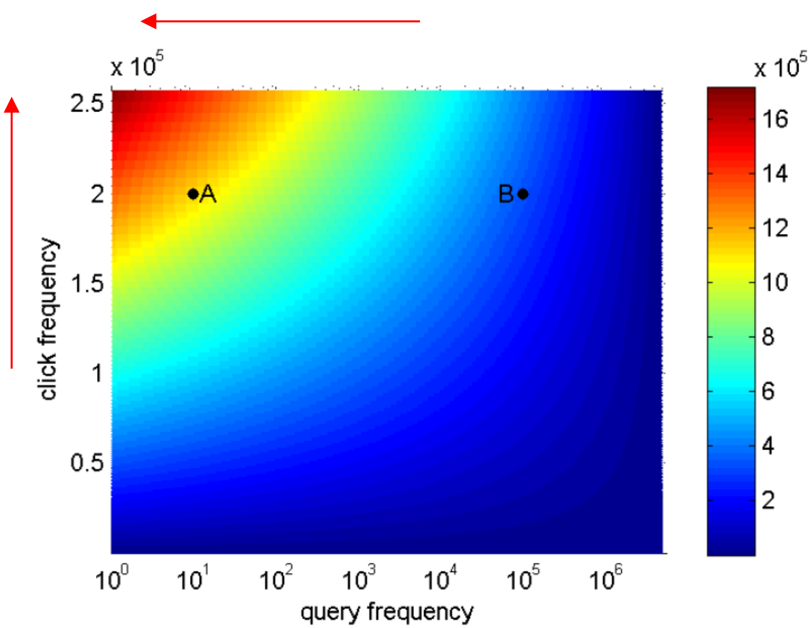
$$iqf(d_j) = \underbrace{\log |Q|}_{\text{Constant}} - \underbrace{\log n(d_j)}_{\text{Entropy}} = \log \frac{|Q|}{n(d_j)}$$

  - Benefits
    - Reduce the influence of some heavily-clicked URLs
    - Balance the bias of clicks for those highly ranked URLs
    - Incorporate with other factors to tune the model

CF-IQF Model

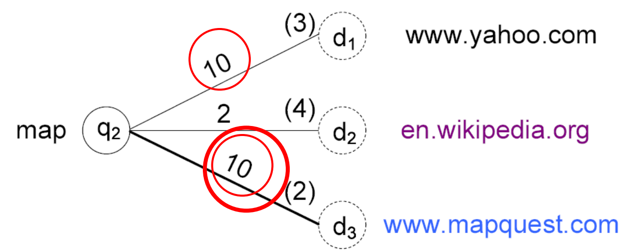
- Incorporate the IQF with the click frequency

$$cfiqf(q_i, d_j) = c_{ij} \cdot iqf(d_j)$$



The surface specified by the click frequency, query frequency and cfiqf, with color specified by the cfiqf value. The color is proportional to the surface height.

- A high click frequency
- A low query frequency
- "A" is weighted higher than "B"



Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

CF-IQF Model

- Transition probability

$$p'_c(d_j|q_i) = \frac{cfiqf(q_i, d_j)}{cfiqf(q_i)}$$

Click frequency matrix

C	d_1	d_2	d_3	d_4
q_1	50	5	0	0
q_2	10	2	10	0
q_3	5	2	5	10
q_4	0	2	0	10

→

CF-IQF transition probabilities

P'_{q2d}	d_1	d_2	d_3	d_4
q_1	1	0	0	0
q_2	0.293	0	0.707	0
q_3	0.122	0	0.293	0.586
q_4	0	0	0	1

The most similar query
q2 (“map”) → q1 (“Yahoo”)

The most similar query
q2 (“map”) → q3 (“travel”)

UF Model and UF-IQF Model

- Drawback of CF model
  - Prone to spam by some malicious clicks (if a single user clicked on a certain URL thousands of times)
- UF model
  - Utilize user frequency instead of click frequency
  - Improve the resistance against malicious clicks
- UF-IQF model

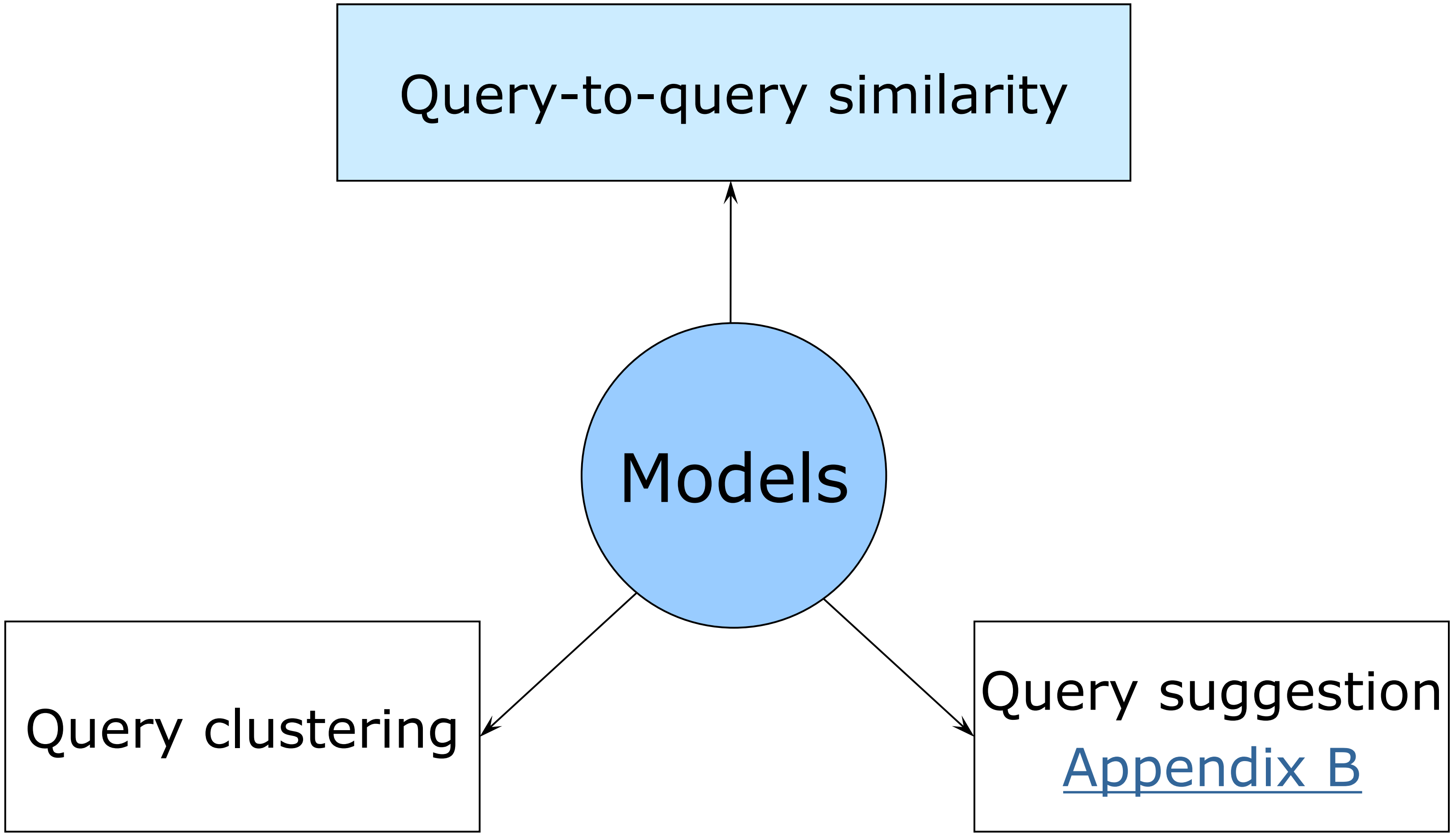
$$ufiqf(q_i, d_j) = uf_{ij} \cdot iqf(d_j)$$

$$p'_u(d_j|q_i) = \frac{ufiqf(q_i, d_j)}{ufiqf(q_i)}$$

Connection with TF-IDF

- TF-IDF
  - Successfully used in vector space model for text retrieval
  - Try to interpret IDF based on binary independence retrieval (BIR), information entropy and LM
  - TF-IDF has never been explored to bipartite graphs
- Entropy-biased framework (CF-IQF)
  - IQF is new
  - CF-IQF is a simplified version of entropy-biased model
  - Share the key point to tune the importance of an edge
  - The scheme can be applied to other bipartite graphs

Mining Query Log on Click Graph



Similarity Measurement

- Cosine similarity

$$Cos(\theta) = \frac{\overrightarrow{q_i} \cdot \overrightarrow{q_j}}{\| \overrightarrow{q_i} \| \| \overrightarrow{q_j} \|}$$

- Jaccard coefficient

$$J(\overrightarrow{q_i}, \overrightarrow{q_j}) = \frac{\sum_{n \in N} |P_{q2d}(i,n) \cap P_{q2d}(j,n)|}{\sum_{n \in N} |P_{q2d}(i,n) \cup P_{q2d}(j,n)|}$$

where $P_{q2d}(i,n)$ denotes the n-th value of $\overrightarrow{q_i}$

- The similarity results are reported and analyzed

Experimental Evaluation

- Data collection
  - AOL query log data

Table 2: Samples of the AOL query log dataset.

UserID	Query	Time	Rank	ClickURL
2722	yahoo	2006-04-25 13:03:23	1	http://www.yahoo.com
121537	map	2006-05-25 18:28:58	1	http://www.mapquest.com
123557	travel	2006-03-13 01:09:53	2	http://www.expedia.com
1903540	cheap flight	2006-05-15 00:31:43	1	http://www.cheapflights.com

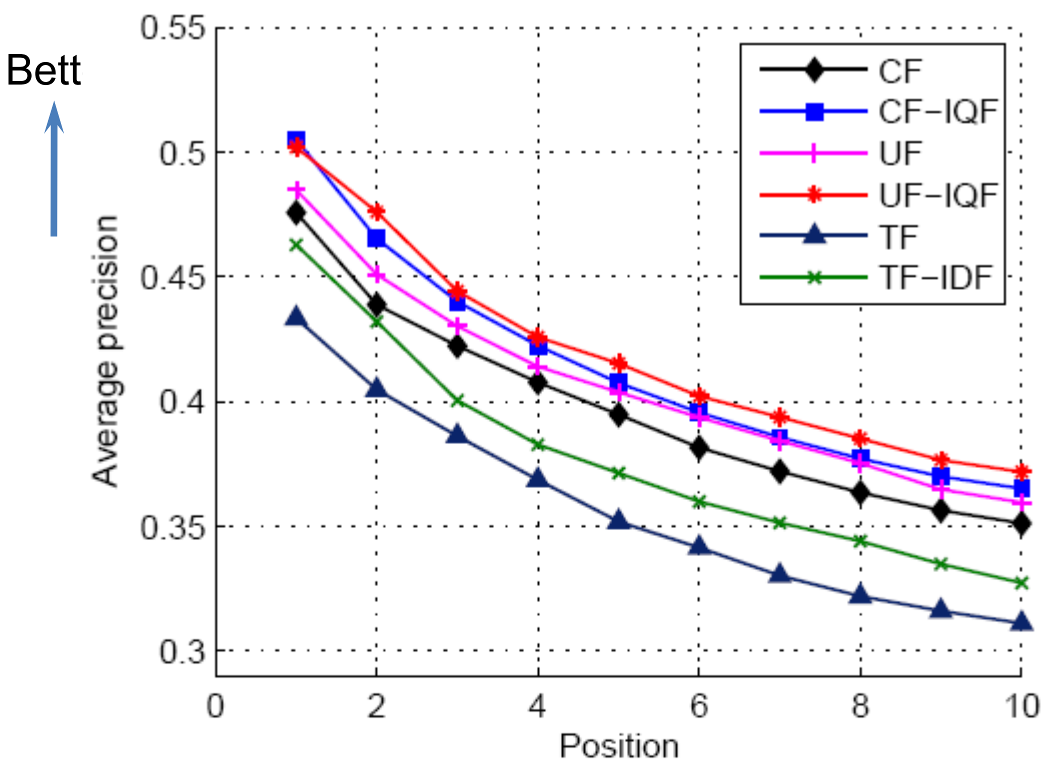
- Cleaning the data
  - Removing the queries that appear less than 2 times
  - Combining the near-duplicated queries
  - 883,913 queries and 967,174 URLs
  - 4,900,387 edges

Evaluation: ODP Similarity

- A simple measure of similarity among queries using ODP categories (query → category)
  - Definition: $Sim(Ca_i, Ca_r) = \frac{|P(Ca_i, Ca_r)|}{\max(|Ca_i|, |Ca_r|)}$
  - Example: 3/5
    - Q1: "United States" → "Regional > North America > United States"
    - Q2: "National Parks" → "Regional > North America > United States > Travel and Tourism > National Parks and Monuments"
- Precision at rank n (P@n): $P@n = \frac{\sum_{i=1}^{n} Sim(q_i, q_r)}{n}$
- 300 distinct queries

Experimental Results

- Query similarity analysis



(a) Cosine similarity

Results:

1. CF-IQF is better than CF
   UF-IQF > UF
2. UF is better than CF
   UF-IQF > CF-IQF
3. TF-IDF is better than TF

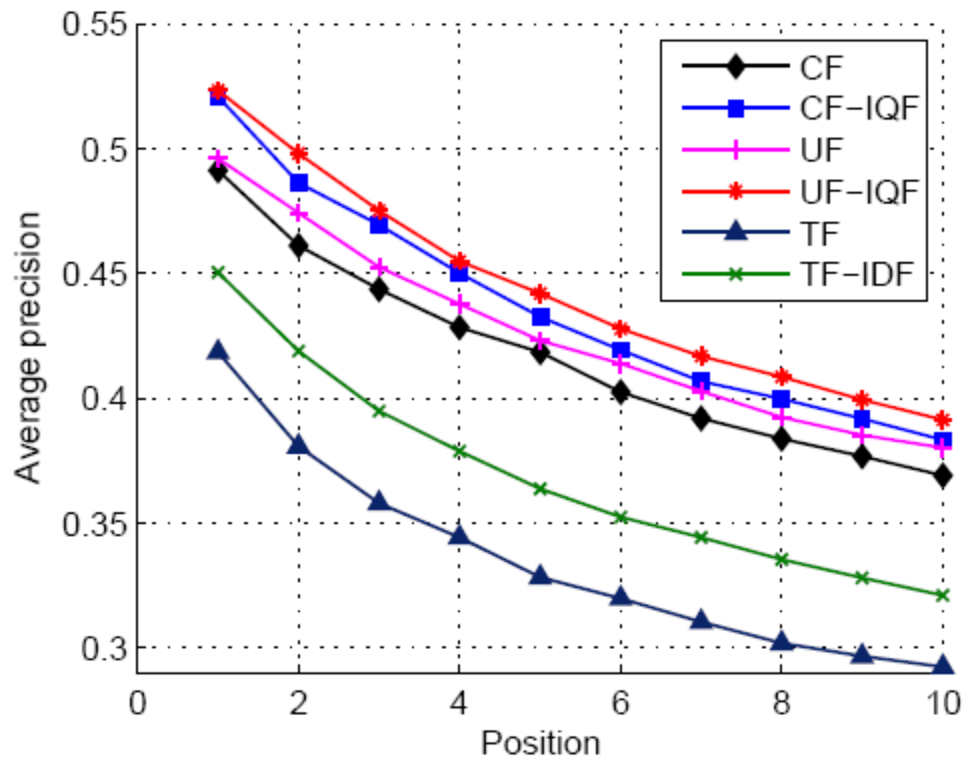
Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Experimental Results

Query similarity analysis

4. Jaccard coefficient

The improvements are consistent with the Cosine similarity



(b) Jaccard coefficient

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Experimental Results

Query similarity analysis

Table 4: Comparison of different methods by P@1 and P@10. We also show the percentage of relative improvement in the lower part.

Method	Cosine		Jaccard	
	P@1	P@10	P@1	P@10
CF	0.476	0.351	0.491	0.369
CF-IQF	**0.505**	0.365	0.521	0.383
UF	0.485	0.360	0.500	0.380
UF-IQF	0.502	**0.372**	**0.523**	**0.391**
TF	0.433	0.311	0.418	0.292
TF-IDF	0.463	0.327	0.450	0.321
CF-IQF/CF	**6.12%**	3.96%	6.01%	3.84%
UF-IQF/UF	3.52%	3.38%	**5.50%**	2.92%
UF-IQF/CF	5.49%	5.86%	6.51%	6.01%
TF-IDF/TF	6.78%	5.21%	7.63%	9.79%
CF/TF	9.76%	12.91%	17.41%	26.23%
UF/TF	11.85%	15.61%	18.53%	**30.02%**
CF-IQF/TF-IDF	9.09%	11.57%	15.65%	19.39%
UF-IQF/TF-IDF	8.44%	13.61%	16.19%	**21.89%**

5. UF-IQF achieves best performance in most cases.

6. CF and UF models > TF
CF-IQF, UF-IQF > TF-IDF

The click graph catches more semantic relations between queries than the query terms

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Summary of Part I

- Introduce the inverse query frequency (IQF) to measure the discriminative ability of a URL
- Propose the framework of the entropy-biased model for the click graph
  - IQF + CF, IQF + UF
  - Formal model to distinguish the variation on different query-URL pairs in the click graph.
- Experimental results show the improvements of the proposed models are consistent and promising

Outline

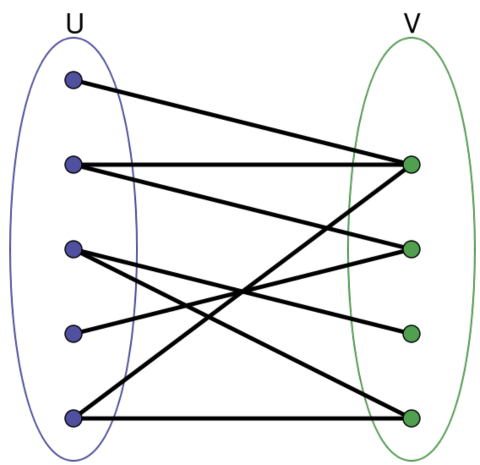
- Part II: Co-HITS Algorithm
  - Motivation
  - Co-HITS Algorithm
    - Iterative Framework
    - Regularization Framework
  - Experimental Results
  - Summary

Motivation

Two kinds of information

IR Models
for Content
- VSM
- Language Model
- etc.

Relevance



Link Analysis
for Graph
- HITS
- PageRank
- etc.

Semantic relations

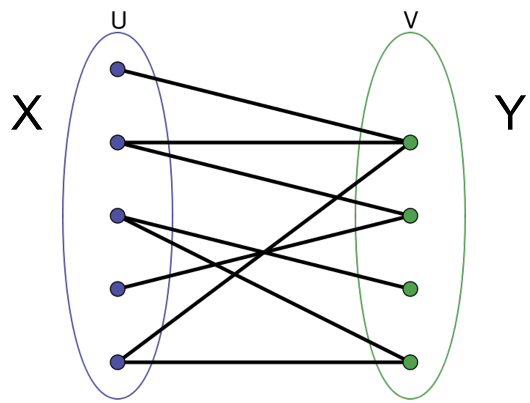
Incorporate Content with Graph
- Personalized PageRank (PPR)
- Linear Combination
- etc.

Preliminaries

Content

$$x_i^0 = f(q, u_i)$$

$$y_j^0 = f(q, v_j)$$



Graph

Explicit links:

$$W^{uv}$$

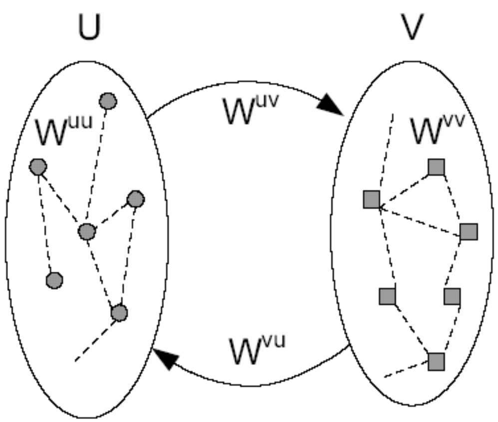
$$W^{vu}$$

Hidden links:

$$W^{uu}$$

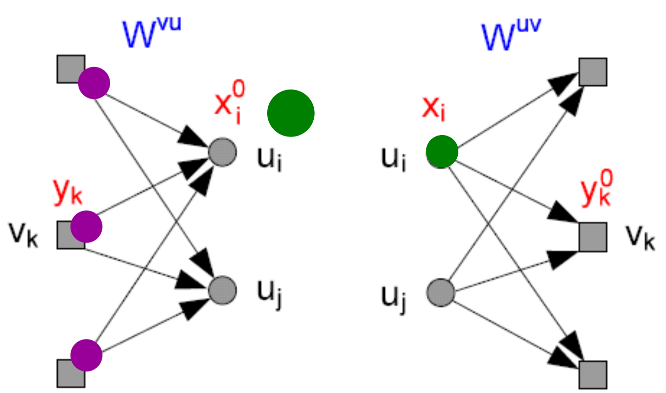
$$W^{vv}$$

$$w_{ij}^{uu} = \sum_{k \in V} w_{ik}^{uv} w_{kj}^{vu}$$



Generalized Co-HITS

- Basic idea
  - Incorporate the bipartite graph with the content information from both sides
  - Initialize the vertices with the relevance scores x^0, y^0
  - Propagate the scores (mutual reinforcement)



(a) (b)

$$
\begin{aligned}
x_i &= (1-\lambda_u)x_i^0 + \lambda_u \sum_{k \in V} w_{ki}^{vu} y_k, \\
y_k &= (1-\lambda_v)y_k^0 + \lambda_v \sum_{j \in U} w_{jk}^{uv} x_j,
\end{aligned}
$$

Initial scores

Score propagation

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Generalized Co-HITS

Iterative framework

$$\begin{aligned} x_i &= (1-\lambda_u)x_i^0 + \lambda_u(1-\lambda_v)\sum_{k\in V} w_{ki}^{vu} y_k^0 + \lambda_u\lambda_v \sum_{j\in U}\left(\sum_{k\in V} w_{jk}^{uv} w_{ki}^{vu}\right) x_j, \\ &= (1-\lambda_u)x_i^0 + \lambda_u(1-\lambda_v)\sum_{k\in V} w_{ki}^{vu} y_k^0 + \lambda_u\lambda_v \sum_{j\in U} w_{ji}^{uu} x_j. \end{aligned}$$

Iterative Framework		
λ_u	λ_v	Description
$=0$	$\in[0,1]$	Initial scores $x_i = x_i^0$
$=1$	$=1$	Original HITS
$\in(0,1)$	$=1$	Personalized PageRank
$\in(0,1)$	$=0$	One-step propagation
$\in(0,1)$	$\in(0,1)$	General Co-HITS

$$x_i = \sum_{j\in U} w_{ji}^{uu} x_j$$

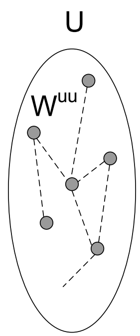
$$x_i = (1-\lambda_u)\cdot x_i^0 + \lambda_u \sum_{j\in U} w_{ji}^{uu}\cdot x_j$$

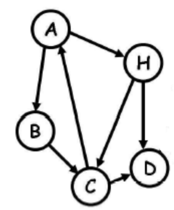
$$x_i = (1-\lambda_u)\cdot x_i^0 + \lambda_u \sum_{k\in V} w_{ki}^{vu}\cdot y_k^0$$

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Iterative → Regularization Framework

- Consider the vertices on one side





$$x_i = \boxed{(1-\lambda_u)\cdot x_i^0} + \lambda_u \boxed{\sum_{j\in U} w_{ji}^{uu}\cdot x_j}$$

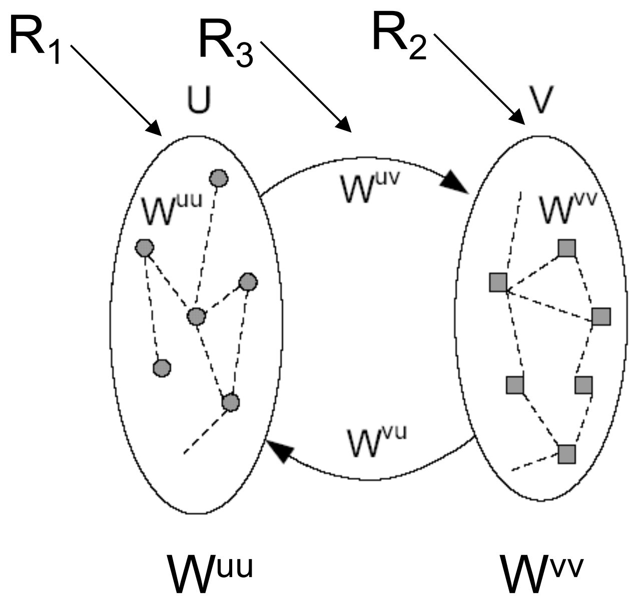
$$R_1 = \underbrace{\frac{1}{2}\sum_{i,j\in U} w_{ij}^{uu}\left\|\frac{x_i}{\sqrt{d_{ii}}}-\frac{x_j}{\sqrt{d_{jj}}}\right\|^2}_{\text{Smoothness}} + \underbrace{\mu\sum_{i\in U}\left\|x_i - x_i^0\right\|^2}_{\text{Fit initial scores}}$$

Personalized PR		Regularization Model	
$\lambda_u = 0$	Initial scores $x_i = x_i^0$	$\mu \to +\infty$	Initial scores $x_i = x_i^0$
$\lambda_u = 1$	Original HITS	$\mu = 0$	Only consider graph
$\lambda_u \in (0,1)$	Personalized PageRank	$\mu \in (0,+\infty)$	Combine content and graph

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Generalized Co-HITS

Regularization framework



$$R_1 = \frac{1}{2}\sum_{i,j\in U} w_{ij}^{uu} \left\| \frac{x_i}{\sqrt{d_{ii}}} - \frac{x_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i\in U} \left\| x_i - x_i^0 \right\|^2$$

$$R_2 = \frac{1}{2}\sum_{i,j\in V} w_{ij}^{vv} \left\| \frac{y_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i\in V} \left\| y_i - y_i^0 \right\|^2$$

$$\begin{aligned} R_3 &= \frac{1}{2}\sum_{i\in U, j\in V} w_{ij}^{uv} \left\| \frac{x_i}{\sqrt{d_{ii}}} - \frac{y_j}{\sqrt{d_{jj}}} \right\|^2 \\ &+ \frac{1}{2}\sum_{j\in V, i\in U} w_{ji}^{vu} \left\| \frac{y_j}{\sqrt{d_{jj}}} - \frac{x_i}{\sqrt{d_{ii}}} \right\|^2 \end{aligned}$$

$$R = \lambda_r(R_1 + \alpha R_2) + (1-\lambda_r)R_3$$

$$\alpha = 1$$

Intuition: the highly connected vertices are most likely to have similar relevance scores.

Generalized Co-HITS

Regularization framework

The cost function:

$$R = \lambda_r(R_1 + R_2) + (1-\lambda_r)R_3$$

Solution:

$$\boxed{\begin{aligned} F^* &= \mu_\beta (I - \mu_\alpha S)^{-1} F^0, \\ S &= D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \\ \mu_\alpha &= \frac{1}{1+\mu}, \text{ and } \mu_\beta = \frac{\mu}{1+\mu}, \end{aligned}}$$

Optimization problem:

$$\begin{aligned} \min_F \quad & \frac{1}{2}\sum_{i,j=1}^{m+n} w_{ij} \left\| \frac{f_i}{\sqrt{d_{ii}}} - \frac{f_j}{\sqrt{d_{jj}}} \right\|^2 + \mu \sum_{i=1}^{m+n} \left\| f_i - f_i^0 \right\|^2 \\ s.t. \quad & W = \begin{bmatrix} W^{uu} & \beta \cdot W^{uv} \\ \beta \cdot W^{vu} & W^{vv} \end{bmatrix} \\ & F = \begin{bmatrix} X \\ Y \end{bmatrix} \\ & \beta = (1-\lambda_r)/\lambda_r, \end{aligned}$$

Regularization Framework

μ_α, λ_r	Description
$\mu_\alpha = 0$	Initial scores $x_i = x_i^0$
$\mu_\alpha = 1$	Corresponding to HITS
$\mu_\alpha \in (0,1)$	Regularization model
$\lambda_r = 1$	Single-sided regularization
$\lambda_r \in (0,1)$	Double-sided regularization

Application to Query-URL Bipartite Graphs

- Bipartite graph construction
  - Edge weighted by the click frequency
  - Normalize to obtain the transition matrix
- Overall algorithm

Algorithm 1 Generalized Co-HITS Algorithm

Input: Given a query q and the bipartite graph
Perform:

1. Calculate the initial ranking scores based on the statistical language model and extract the top-ranked U_L and V_L as the seed sets;
2. Expand and extract the compact bipartite subgraph $\hat{G} = (\hat{U} \cup \hat{V}, \hat{E})$;
3. Get the weight matrix $\hat{W}$ or $\hat{S}$, and normalize the corresponding initial scores F^0;
4. Solve Eq. (5) or Eq. (16) and get the final scores $\hat{F}^*$.

Output: Return the ranked queries

Experimental Evaluation

- Data collection
  - AOL query log data

Table 2: Samples of the AOL query log dataset.

UserID	Query	Time	Rank	ClickURL
2722	yahoo	2006-04-25 13:03:23	1	http://www.yahoo.com
121537	map	2006-05-25 18:28:58	1	http://www.mapquest.com
123557	travel	2006-03-13 01:09:53	2	http://www.expedia.com
1903540	cheap flight	2006-05-15 00:31:43	1	http://www.cheapflights.com

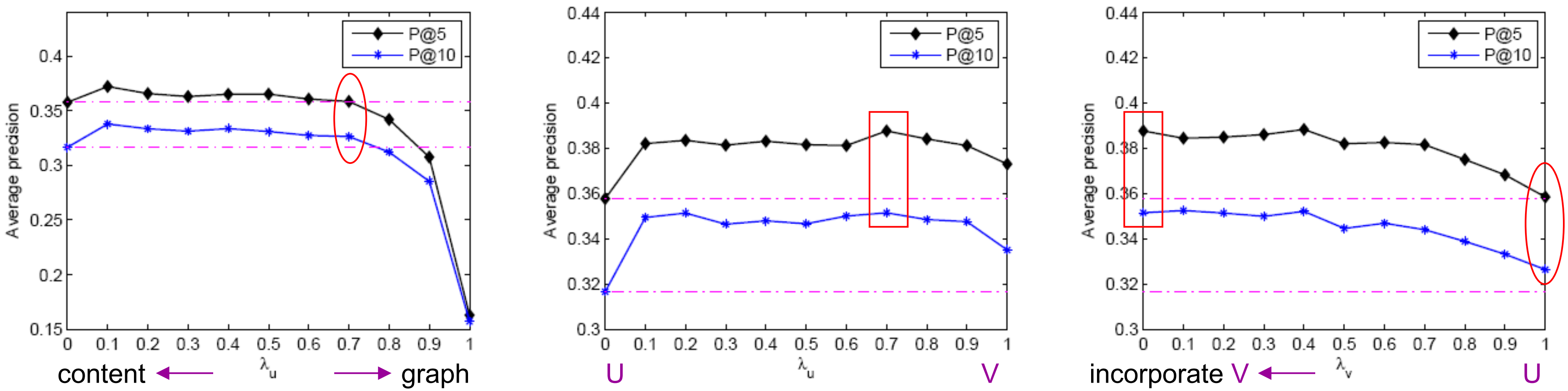
- Cleaning the data
  - Removing the queries that appear less than 2 times
  - Combining the near-duplicated queries
  - 883,913 queries and 967,174 URLs
  - 4,900,387 edges
  - 250,127 unique terms

Evaluation: ODP Similarity

- A simple measure of similarity among queries using ODP categories (query → category)
  - Definition: $Sim(Ca_i, Ca_r) = \frac{|P(Ca_i, Ca_r)|}{\max(|Ca_i|, |Ca_r|)}$
  - Example: 3/5
    - Q1: "United States" → "Regional > North America > United States"
    - Q2: "National Parks" → "Regional > North America > United States > Travel and Tourism > National Parks and Monuments"
- Precision at rank n (P@n): $P@n = \frac{\sum_{i=1}^{n} Sim(q_i, q_r)}{n}$
- 300 distinct queries

Experimental Results

Comparison of iterative framework



(a) $\lambda_v = 1$ (PPR) (b) $\lambda_v = 0$ (OSP) (c) $\lambda_u = 0.7$ (CoIter)

Iterative Framework		
λ_u	λ_v	Description
$= 0$	$\in [0,1]$	Initial scores $x_i = x_i^0$
$= 1$	$= 1$	Original HITS
$\in (0,1)$	$= 1$	Personalized PageRank
$\in (0,1)$	$= 0$	One-step propagation
$\in (0,1)$	$\in (0,1)$	General Co-HITS

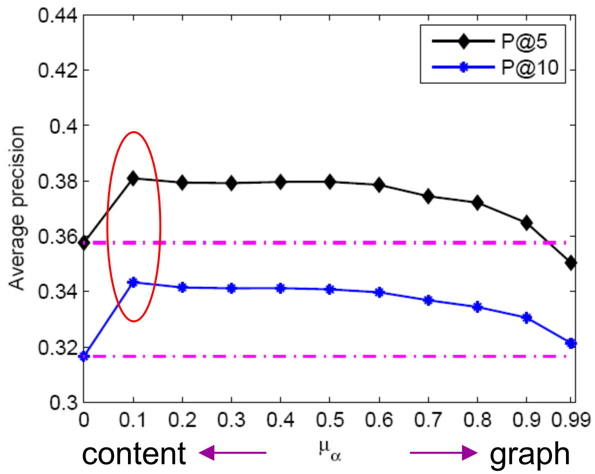
***Result 1**:*

The initial relevance scores from both sides provide valuable information. The improvements of OSP and CoIter over the baseline (the dashed line) are promising when compared to the PPR.

Experimental Results

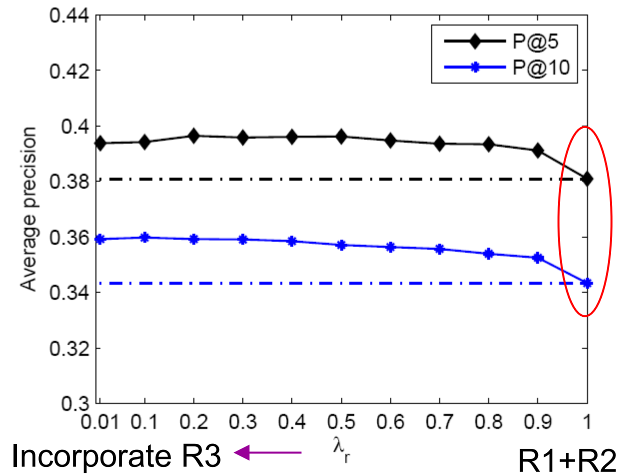
Comparison of regularization framework

single-sided regularization



(a) $\lambda_r = 1$ (SiRegu)

double-sided regularization



(b) $\mu_\alpha = 0.1$ (CoRegu)

Regularization Framework	
μ_α, λ_r	Description
$\mu_\alpha = 0$	Initial scores $x_i = x_i^0$
$\mu_\alpha = 1$	Corresponding to HITS
$\mu_\alpha \in (0,1)$	Regularization model
$\lambda_r = 1$	Single-sided regularization
$\lambda_r \in (0,1)$	Double-sided regularization

Result 2:

SiRegu can improve the performance over the baseline. CoRegu performs better than SiRegu, which owes to the newly developed cost function R_3. Moreover, CoRegu is relatively robust.

Experimental Results

Detailed results

Table 3: Comparison of different methods by P@5 and P@10. The mean precisions and the percentages of relative improvements are shown in the table.

Method	Para		Evaluation metrics	
Iter	λ_u	λ_v	P@5	P@10
Baseline	0	$\times$	0.358 (0%)	0.317 (0%)
PPR-0.1	0.1	1	0.372 (4.0%)	0.338 (6.7%)
OSP-0.7	0.7	0	0.388 (8.4%)	0.351 (11.0%)
CoIter-0.4	0.7	0.4	0.388 (8.6%)	0.352 (11.2%)
Regu	λ_r	μ_α	P@5	P@10
SiRegu-0.1	1	0.1	0.381 (6.5%)	0.343 (8.5%)
CoRegu-0.5	0.5	0.1	0.396 (10.8%)	0.357 (12.8%)

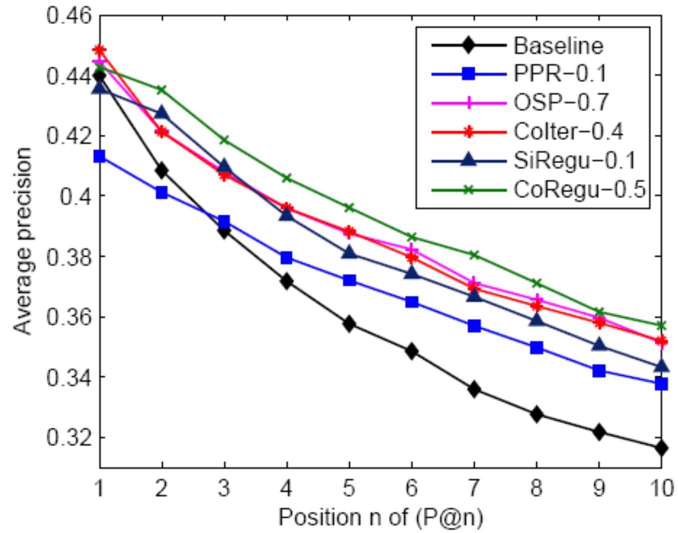


Figure 5: Comparison of six models.

***Result 3**:*

The CoRegu-0.5 achieves the best performance. It is very essential and promising to consider the double-sided regularization framework for the bipartite graph.

Summary of Part II

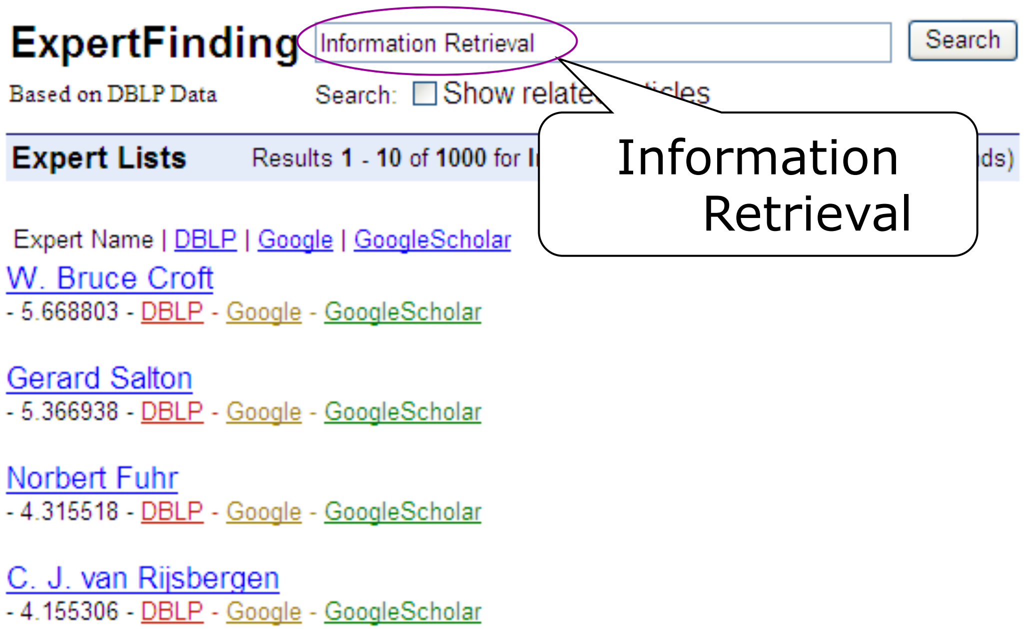
- Propose the generalized Co-HITS algorithm
  - Incorporate the bipartite graph with the content information from both sides
- Investigate two different frameworks
  - Iterative: include HITS and personalized PageRank as special cases
  - Regularization: build the connection with HITS, develop new cost functions
- Experimental results
  - CoRegu is more robust, achieves the best performance

Outline

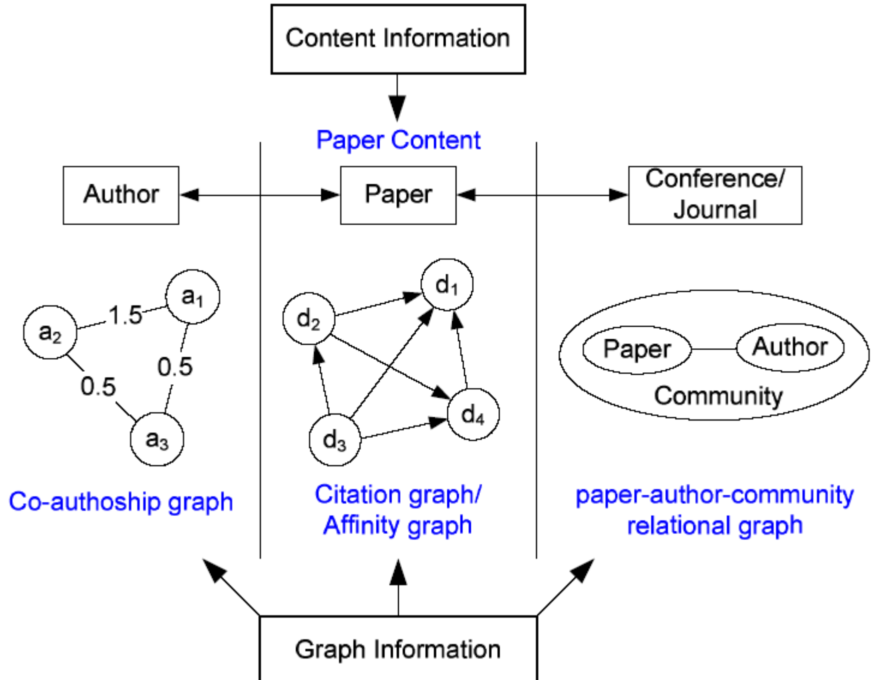
- Background: Web Mining Techniques
  - Information retrieval, link analysis, machine learning
- Modeling Bipartite Graph for Query Log Analysis
  - Entropy-biased Models [w/ King-Lyu, SIGIR'09]
  - Co-HITS Algorithm [w/ Lyu-King, KDD'09]
- Modeling Expertise Retrieval
  - Baseline and Weighted Model [w/ King-Lyu, ICDM'08]
  - Graph-based Re-ranking Model [w/ Lyu-King, WSDM'09]
  - Enhanced Models with Community [w/ King-Lyu, CIKM'09]
- Conclusion and Future Work

Modeling Expertise Retrieval

- Expertise retrieval (Expert finding) task:
  - Identify people with relevant expertise for a given query
  - A high-level information retrieval
  - DBLP bibliography and its supplemental data



Overview of Expertise Retrieval



Part III:

- Baseline model
- Weighted language model
- Graph-based re-ranking

Part IV:

- Enhanced models with community-aware strategies

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Expertise Modeling

- Expert finding
  - *p(ca|q)*: what is the probability of a candidate *ca* being an expert given the query topic *q*?
  - Rank candidates *ca* according to this probability.
- Approach:
  - Using Bayes' theorem,

$$p(ca|q) = \frac{p(ca,q)}{p(q)}$$

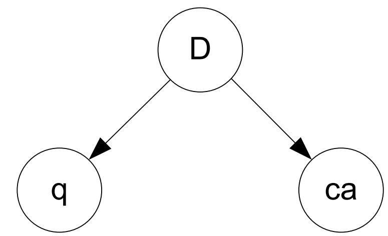
where *p(ca, q)* is joint probability of a candidate and a query, *p(q)* is the probability of a query.

$$p(ca|q) \propto p(ca,q),$$

Baseline Model (Document-based Model)

- The probability *p(ca,q)*:

$$
\begin{aligned}
p(ca,q) &= \sum_{d\in D} p(d)p(ca,q|d) \\
&= \sum_{d\in D} p(d)\underbrace{p(q|d)}_{\text{Language Model}}\underbrace{p(ca|d,q)}_{\text{Conditionally independent}}
\end{aligned}
$$



Baseline model

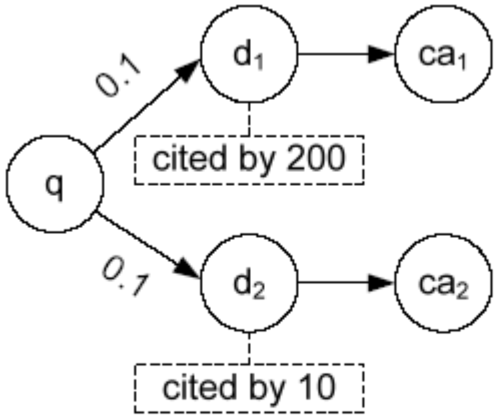
$$
\boxed{
\begin{aligned}
p(q|\theta_d) &= \prod_{t\in q} p(t|\theta_d)^{c(t,q)} \\
p(t|\theta_d) &= (1-\lambda)p(t|d) + \lambda p(t).
\end{aligned}
}
$$

$$p(ca|d,q) = p(ca|d_q)$$

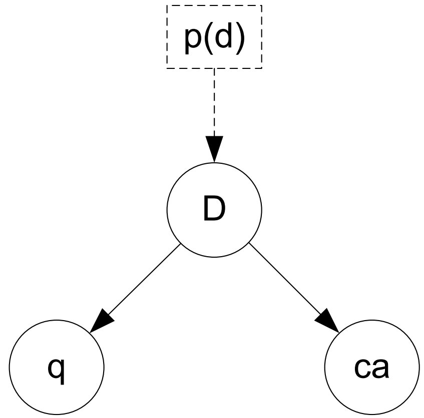
$$
p(ca|d) = \begin{cases} \frac{1}{n_d}, & (ca \text{ is the author of } d) \\ 0, & (\text{otherwise}) \end{cases}
$$

- Find out documents relevant to the query
- Aggregate the expertise of an expert candidate from the associated documents

Weighted Model



A query example



Weighted model

$$
\begin{aligned}
p(ca, q) &= \sum_{d \in D} p(d) p(ca, q|d) \\
&= \sum_{d \in D} p(d) p(q|d) p(ca|d, q)
\end{aligned}
$$

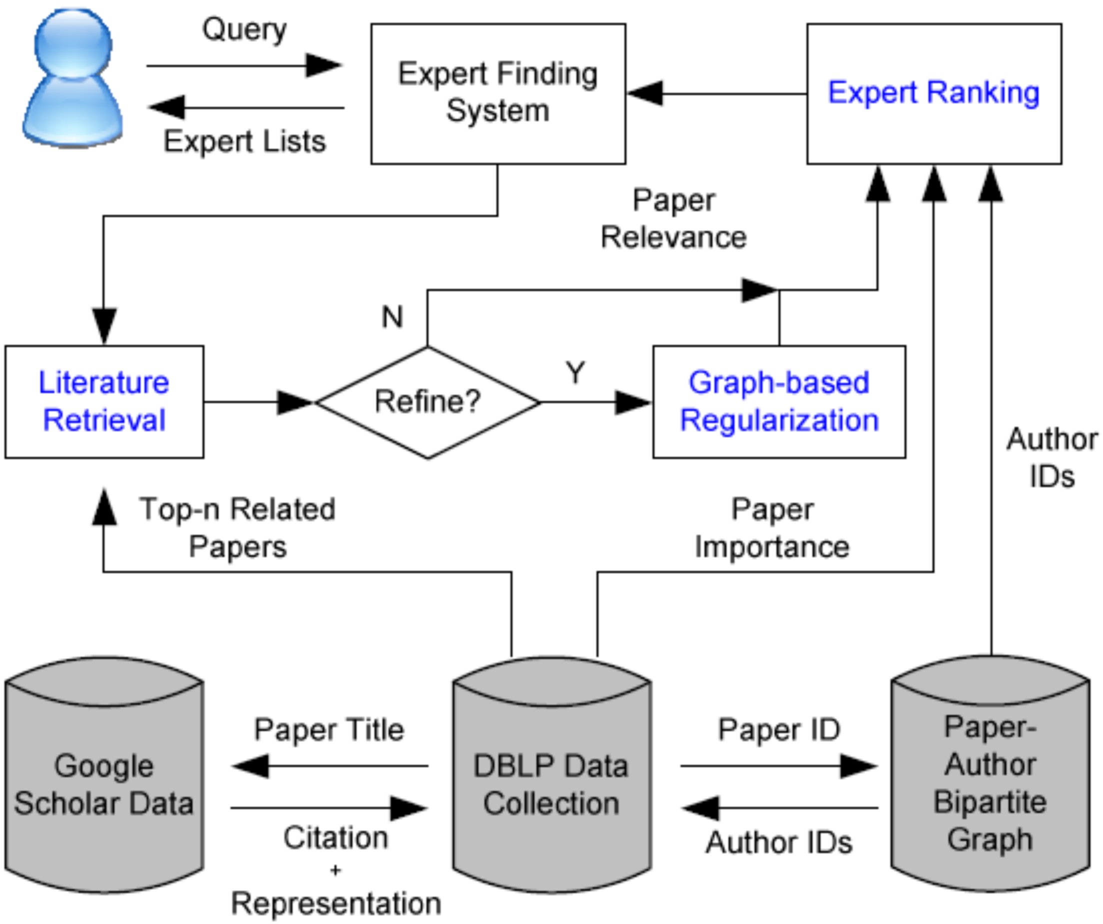
$$p(d) = \frac{w_d}{C} \propto w_d,$$

$$w_d = \begin{cases} 1, & (B1) \\ \log(10 + c_d), & (B2) \end{cases}$$

The final estimation of the weighted language model is

$$p_l(q, ca) \stackrel{rank}{=} \sum_{d \in D} w_d \left\{ \prod_{t \in q} (p(t|\theta_d))^{n(t,q)} \right\} p(ca|d).$$

General Expert Finding System



The schematic of general expert finding systems.

Graph-based Re-ranking

- Key issue for expert finding:
  - To retrieve the most relevant documents along with the relevance scores
- Intuition
  - Global consistency: Similar documents are most likely to have similar ranking scores with respect to a query
- Regularization framework

$$R(F, q, G) = \frac{1}{2}\underbrace{\sum_{i,j=1}^{n} w_{ij}\left\|\frac{f(d_i, q)}{\sqrt{D_{ii}}} - \frac{f(d_j, q)}{\sqrt{D_{jj}}}\right\|^2}_{\text{Global consistency}} + \overset{\text{Parameter}}{\mu}\underbrace{\sum_{i=1}^{n}\left\|f(d_i, q) - f^0(d_i, q)\right\|^2}_{\text{Fit initial scores}}$$

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Graph-based Re-ranking

- Optimization problem

$$F^* = arg \min_{F \in \mathbb{R}^{+n}} R(F, q, G)$$

- A closed-form solution

$$F^* = \mu_\beta (I - \mu_\alpha S)^{-1} F^0,$$
$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}},$$
$$\mu_\alpha = \frac{1}{1+\mu}, \text{ and } \mu_\beta = \frac{\mu}{1+\mu},$$

- Connection with other methods
  - $\mu_\alpha \rightarrow 0$, return the initial scores
  - $\mu_\alpha \rightarrow 1$, a variation of PageRank-based model
  - $\mu_\alpha \in (0, 1)$, combine both information simultaneously

Combination of Different Methods

Model	w_d	Refine	Meaning
LM(bas)	B1[a]	N[c]	baseline model
LM(w)	B2[b]	N	weighted language model
LM(r)	B1	Y[d]	LM(bas) with graph-based regularization
LM(w+r)	B2	Y	LM(w) with graph-based regularization

[a] uniform weight ($w_d = 1$)

[b] common logarithm weight ($w_d = \log(10 + c_d)$)

[c] without graph-based regularization

[d] with graph-based regularization

Experimental Setup

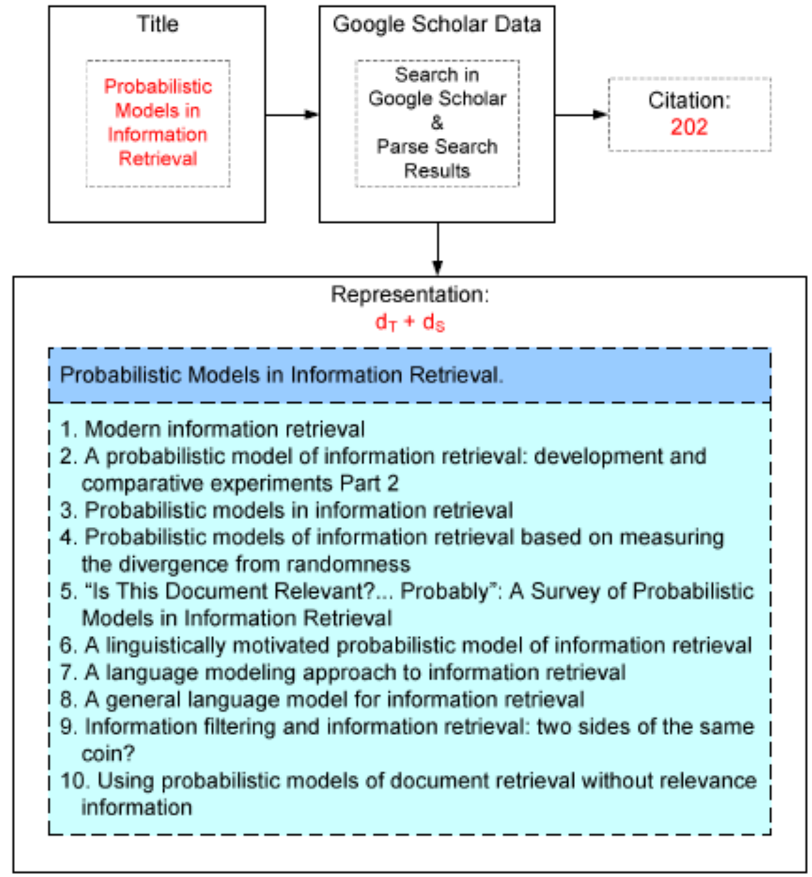
DBLP collection and representation

A sample of the DBLP XML records

```xml
<article mdate="2003-11-24" key="journals/cj/Fuhr92">
    <author>Norbert Fuhr</author>
    <title>Probabilistic Models in Information Retrieval.</title>
    <pages>243-255</pages>
    <year>1992</year>
    <volume>35</volume>
    <journal>Comput. J.</journal>
    <number>3</number>
    <url>db/journals/cj/cj35.html#Fuhr92</url>
</article>
```

Statistics of the DBLP collection

Property	#of entities
Number of papers	925,293
Number of authors	574,369
Number of terms	308,651



Experimental Setup

- Assessments
  - Manually created the ground truth through the method of pooled relevance judgments
  - 17 query topics and 17 expert lists
- Evaluation metrics
  - Precision at rank n (P@n)
  - MAP
  - Bpref (Appendix D)

Topic	#Expert
Boosting	56
Information Extraction	20
Intelligent Agents	29
Machine Learning	42
Natural Language Processing	43
Planning	34
Semantic Web	45
Support Vector Machine	31
Ontology Alignment	55
Probabilistic Relevance Model	13
Information Retrieval	23
Language Model For Information Retrieval	12
Face Recognition	21
Semi Supervised Learning	21
Reinforcement Learning	17
Privacy Preservation	17
Kernel Methods	22

Experimental Results

"Title"	P@5	P@10	P@20	R-prec	MAP	bpref
LM(bas)	61.18	51.18	44.71	40.30	27.27	33.20
LM(w)	72.94	60.59	48.53	43.22	31.91	36.79
	+19.2	+18.4	+8.55	+7.25	+17.0	+10.8
"Title+GS"	P@5	P@10	P@20	R-prec	MAP	bpref
LM(bas)	72.94	64.12	47.94	43.98	33.06	38.16
LM(w)	**81.18**	**65.29**	**53.24**	**47.93**	**37.10**	**41.60**
	+11.3	+1.84	+11.0	+8.98	+12.2	+9.01

- Weighted model LM(w) outperforms baseline model LM(bas)

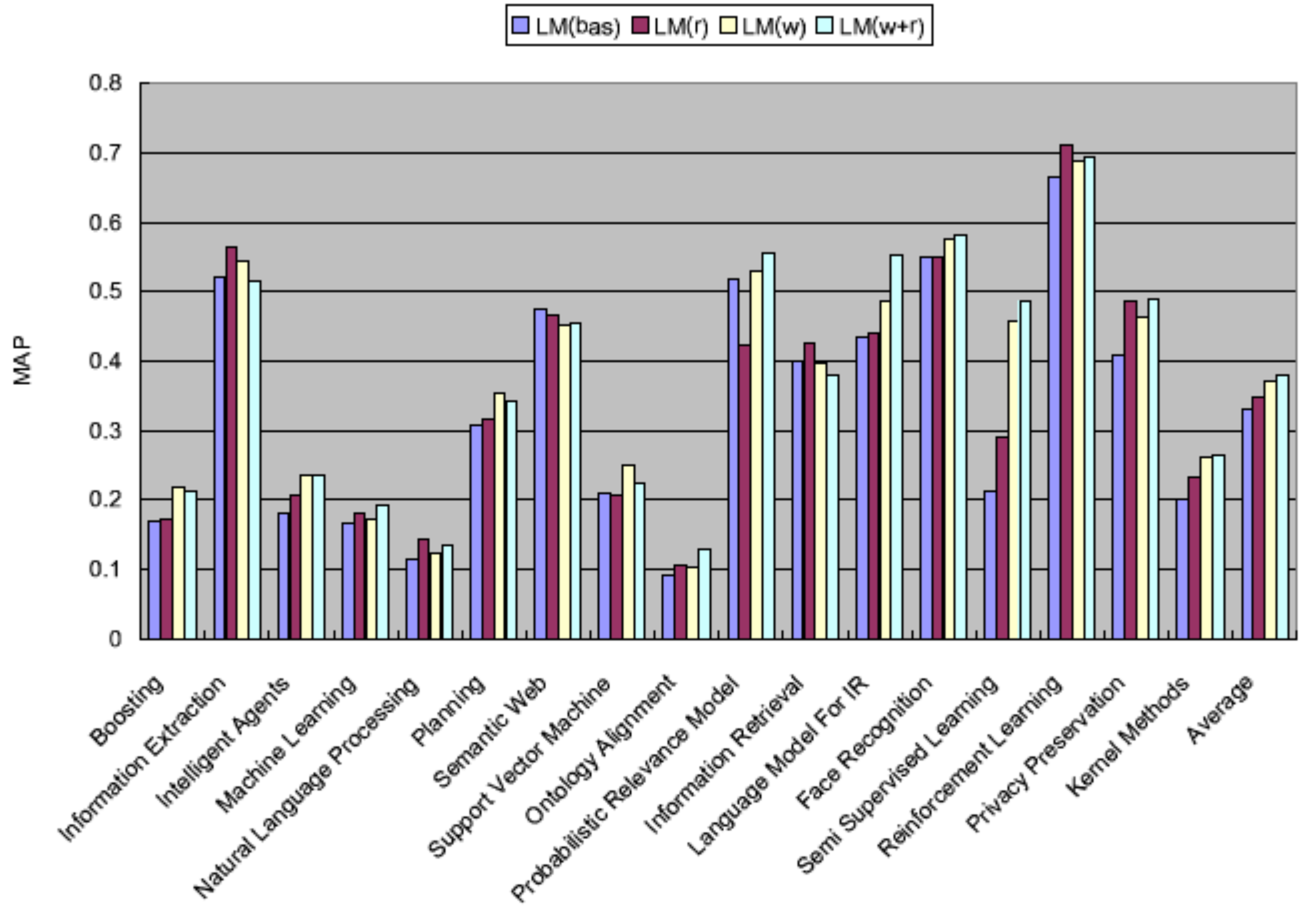
Experimental Results

Comparison of different methods (%). The percentages of relative improvements are shown in the lower part

Method	P@5	P@10	P@20	R-prec	MAP	bpref
LM(bas)	72.94	64.12	47.94	43.98	33.06	38.16
LM(r) ($\mu_\alpha = 0.5$)	77.65	65.29	51.18	46.25	34.86	39.97
LM(w)	81.18	65.29	53.24	47.93	37.10	41.60
LM(w+r) ($\mu_\alpha = 0.5$)	82.35	68.24	55.59	48.88	37.89	42.60
LM(r) / LM(bas)	+6.45%	+1.83%	+6.75%	+5.15%	+5.42%	+4.75%
LM(w+r) / LM(w)	+1.45%	+4.50%	+4.42%	+1.97%	+2.13%	+2.40%
LM(w) / LM(bas)	+11.29%	+1.84%	+11.04%	+8.98%	+12.22%	+9.01%
LM(w+r) / LM(bas)	**+12.90%**	**+6.42%**	**+15.95%**	**+11.13%**	**+14.61%**	**+11.63%**

- The performance can be boosted with Graph-based regularization

Experimental Results



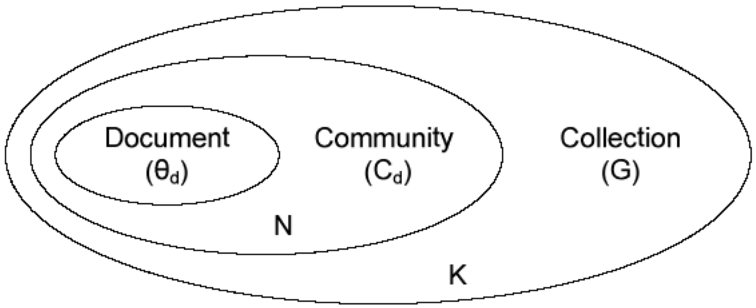
MAP

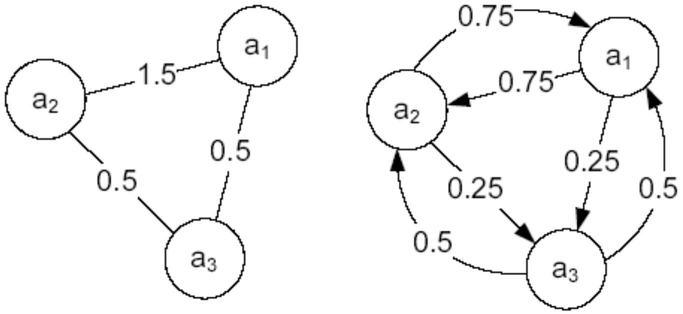
Summary of Part III

- Present the weighted model for expert finding
  - Take into account both the relevance scores and the importance of the documents
- Investigate and integrate the graph-based regularization method with the weighted model
- Experimental results are presented to show the effectiveness of proposed models
  - The performance is further boosted by refining the relevance scores of the documents

Summary of Other Contributions

- Enhancing Expertise Retrieval
  - Communities could provide valuable insight and distinctive information
  - A new smoothing method using the community context
  - Ranking refinement based on community co-authorship
  - Details in Appendix A





Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Conclusion

Model	Data	Techniques
Entropy-biased Model	Bipartite Graph	LA + IR
Co-HITS Algorithm	Bipartite Graph + Content	LA + IR + ML
Weighted Language Model	Content + Citation	IR
Graph-based Re-ranking Model	Content + Affinity Graph	LA + IR +ML
Enhanced Model with Communities	Content + Community	LA + IR +ML

Future Work

- Query log analysis
  - Personalization
    - General click graph, user's click graph, session info.
  - Incorporate with other information
    - Query-flow model, user modeling
- Expertise retrieval on the Web
  - Beyond a particular domain or intranet
  - Identify relevant experts/trusted people
  - Create a global expert and friend recommendation
- Apply to other applications
  - Entity retrieval
  - Online social media search
  - …

Selected Publications

- **Hongbo Deng**, Irwin King, Michael R. Lyu. "Entropy-biased Models for Query Representation on the Click Graph." *Proceedings of the 32nd Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009)*. Pages 339-346, Boston, MA, July 19-23, 2009. [Acceptance rate = 78/494 = 15.8%]

- **Hongbo Deng**, Michael R. Lyu and Irwin King. "A Generalized Co-HITS Algorithm and Its Application to Bipartite Graphs." *Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009)*. Pages 239-248, Paris, France, June 28th - July 1st, 2009. [Acceptance rate = 105/561 = 18.7%]

- **Hongbo Deng**, Irwin King, Michael R. Lyu. "Enhancing Expertise Retrieval Using Community-aware Strategies." *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. Hong Kong, China, Nov. 2-6, 2009. Short paper. [Acceptance rate = 294/847 = 34.7%]

- **Hongbo Deng**, Michael R. Lyu and Irwin King. "Effective Latent Space Graph-based Re-ranking Model with Global Consistency." *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM 2009)*. Pages 212-221, Barcelona, Spain, Feb. 9-12, 2009. [Acceptance rate = 29/170 = 17%]

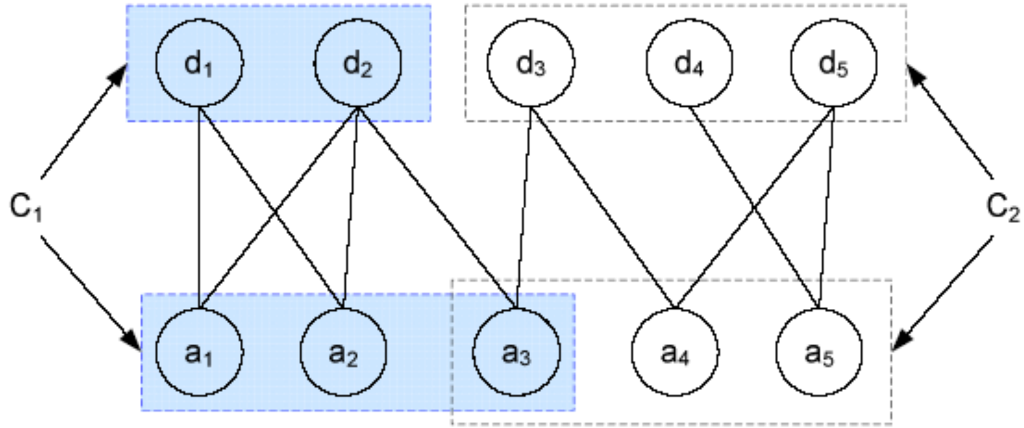
- **Hongbo Deng**, Irwin King and Michael R. Lyu. "Formal Models for Expert Finding on DBLP Bibliography Data." *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*. Pages 163-172, Pisa, Italy, Dec. 15-19, 2008. [Acceptance rate = 70/723 = 10%]

Q&A

Thanks!

Appendix A: Enhancing Expertise Retrieval

- Motivation
  - Communities could provide valuable insight and distinctive information



An example graph with two communities

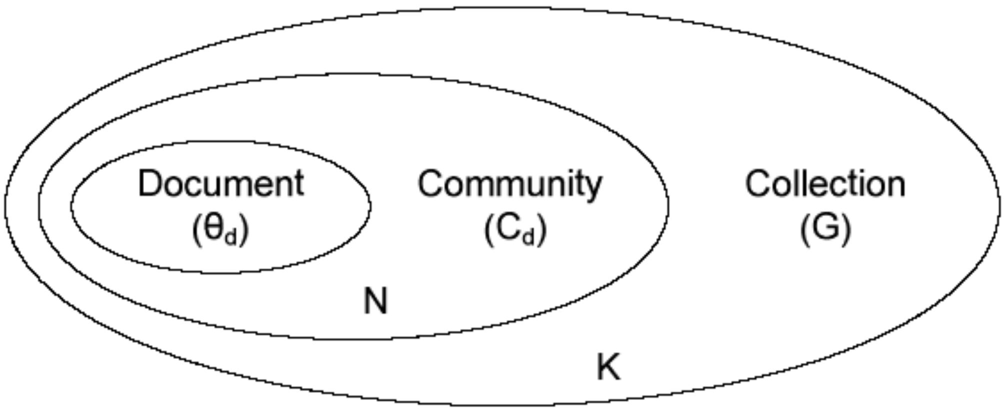
- Community-aware strategies
  - A new smoothing method using the community context
  - Ranking refinement based on community co-authorship

Document-based Model

- Key challenge
  - Compute the relevance between query and document
- Statistical language model
  - Smoothing $p(t|\theta_d)$ with the community language model $p(t|C_d)$ instead of the collection language model $p(t|G)$

$$p(t|\theta_d) = (1-\lambda)\frac{n(t,d)}{|d|} + \lambda p(t|G)$$

$$p(t|\theta_d) = (1-\lambda)\frac{n(t_i,d)}{|d|} + \lambda p(t_i|C_d)$$



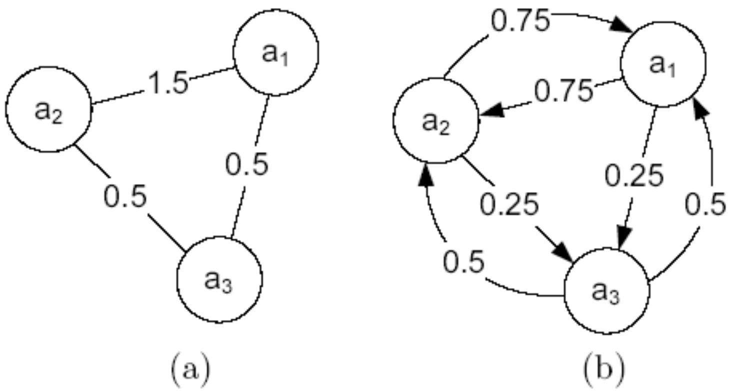
Discovering Authorities in a Community

- Co-authorship frequency

$$f_{ij} = \sum_{k=1}^{N} \frac{\delta_i^k \delta_j^k}{n_k - 1}$$

- Normalized weight

$$w_{ij} = \frac{f_{ij}}{\sum_{k=1}^{n} f_{ik}}$$



Co-authorship graph with: (a) co-authorship frequency, and (b) normalized weight

- AuthorRank
  - Measure the authority for the authors within a community

$$p(a_i|C_k) = (1-\alpha)\frac{1}{N_a(C_k)} + \alpha \sum_{j=1}^{N_a(C_k)} w_{ij} \cdot p(a_j|C_k)$$

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Community-Sensitive AuthorRank

- Relevance

$$p(C_k|q) = \frac{p(C_k) \cdot p(q|C_k)}{p(q)} \propto p(C_k) \prod_{t_i \in q} p(t_i|C_k)$$

- The quantity *p(C_k)*

$$p(C_k) \propto N_a(C_k) \cdot \log(10 + N_c(C_k))$$

- Community Sensitive AuthorRank

$$p(a_i|q) \propto \sum_k p(C_k)p(q|C_k)p(a_i|C_k)$$

  - Suppose C_k be a "virtual" document, it becomes the document-based model
  - Capture the high-level and general aspects for a query

Ranking Refinement Strategy

- Two kinds of ranking results
  - *Rd*: capture more specific and detailed aspects
  - *Rc*: reflect more general and abstract aspects
- Measure the similarity and diversity

$$J = \frac{\left|\vec{Rd} \bigcap \vec{Rc}\right|}{\left|\vec{Rd} \bigcup \vec{Rc}\right|}$$

- Ranking refinement

$$S(a_i) = \frac{1}{Rd(a_i)} + \delta(a_i) \cdot J \cdot \frac{1}{\hat{Rc}(a_i)}$$

Experimental Results

- Comparison of Document-based Models

Method	P@10	P@20	P@30	R-prec	MAP	bpref
DM(b)	0.5353	0.45	0.3726	0.4316	0.2897	0.3524
DM(bc)	0.5588	0.4647	0.3824	0.4417	0.3015	0.3621
DM(w)	0.6882	0.5029	0.4235	0.4845	0.3633	0.4159
DM(wc)	**0.6882**	**0.5265**	**0.4314**	**0.4943**	**0.3771**	**0.4279**
DM(bc)/DM(b)	+4.40%	+3.27%	+2.63%	+2.34%	+4.09%	+2.78%
DM(wc)/DM(w)	0%	+4.68%	+1.85%	+2.03%	+3.79%	+2.89%
DM(w)/DM(b)	+28.57%	+11.76%	+13.68%	+12.26%	+25.43%	+18.02%
DM(wc)/DM(b)	+28.57%	+16.99%	+15.79%	+14.53%	+30.19%	+21.44%

Experimental Results

- Comparison of Enhanced Models

Method	P@10	P@20	P@30	R-prec	MAP	bpref
EDM(b)	0.5882	0.4971	0.4196	0.4716	0.3228	0.38933
EDM(bc)	0.5941	0.5059	0.4275	0.4803	0.3342	0.39879
EDM(w)	0.7059	0.55	0.4608	0.5317	0.403	0.45839
EDM(wc)	**0.7118**	**0.5677**	**0.4628**	**0.5332**	**0.4089**	**0.46241**
EDM(b)/DM(b)	+9.89%	+10.46%	+12.63%	+9.28%	+11.44%	+10.49%
EDM(bc)/DM(bc)	+6.31%	+8.86%	+11.79%	+8.75%	+10.85%	+10.12%
EDM(w)/DM(w)	+2.56%	+9.36%	+8.79%	+9.75%	+10.92%	+10.22%
EDM(wc)/DM(wc)	+3.42%	+7.82%	+7.27%	+7.86%	+8.43%	+8.06%

Experimental Results

Discussion and Detailed Results

The detailed results of the community-sensitive AuthorRank for the query “machine learning.”

journals/ML	conf/ICML	conf/NIPS	journals/JMLR	conf/ECML
Pat Langley	Andrew W. Moore	Terrence J. Sejnowski	Michael I. Jordan	Saso Dzeroski
Robert E. Schapire	Sridhar Mahadevan	Michael I. Jordan	Yoram Singer	Johannes Frnkranz
Manfred K. Warmuth	Thomas G. Dietterich	Geoffrey E. Hinton	Tong Zhang	Gerhard Widmer
Thomas G. Dietterich	Prasad Tadepalli	Peter Dayan	Francis R. Bach	Ivan Bratko
Yoram Singer	Michael L. Littman	Christof Koch	Olivier Bousquet	Enric Plaza
Ryszard S. Michalski	Pat Langley	Klaus-Robert Mller	Klaus-Robert Mller	Pavel Brazdil
Michael J. Pazzani	Andrew McCallum	Zoubin Ghahramani	Bernhard Schlkopf	Birgit Tausend
Dana Angluin	Thorsten Joachims	Michael Mozer	Andr Elisseeff	Stephen Muggleton
Avrim Blum	Satinder P. Singh	Bernhard Schlkopf	Koby Crammer	Floriana Esposito
Leo Breiman	Michael I. Jordan	Satinder P. Singh	Ingo Steinwart	Stan Matwin

Experimental Results

DM(wc)	Authorities	EDM(wc)
Pat Langley	Pat Langley	Pat Langley
Thomas G. Dietterich	Robert E. Schapire	Thomas G. Dietterich
Sumio Watanabe	Manfred K. Warmuth	Sumio Watanabe
David E. Goldberg	Yoram Singer	David E. Goldberg
Tom M. Mitchell	Thomas G. Dietterich	Avrim Blum
Avrim Blum	Michael I. Jordan	Tom M. Mitchell
Ivan Bratko	Satinder P. Singh	Sanjay Jain
Donald Michie	Sanjay Jain	Ivan Bratko
Carl H. Smith	John Shawe-Taylor	Donald Michie
J. Ross Quinlan	Michael J. Pazzani	Michael I. Jordan

Summary

- Investigate the smoothing method using community context instead of the whole collection
- Introduce the community-sensitive AuthorRank for determining the query-sensitive authorities
- Develop an adaptive ranking refinement strategy to aggregate the ranking results
- Experimental results shows a significant improvement over the baseline method
- Return from Appendix A

Appendix B: Graph-based Random Walk

- Query-to-query graph
  - The transition probability from q_i to q_j

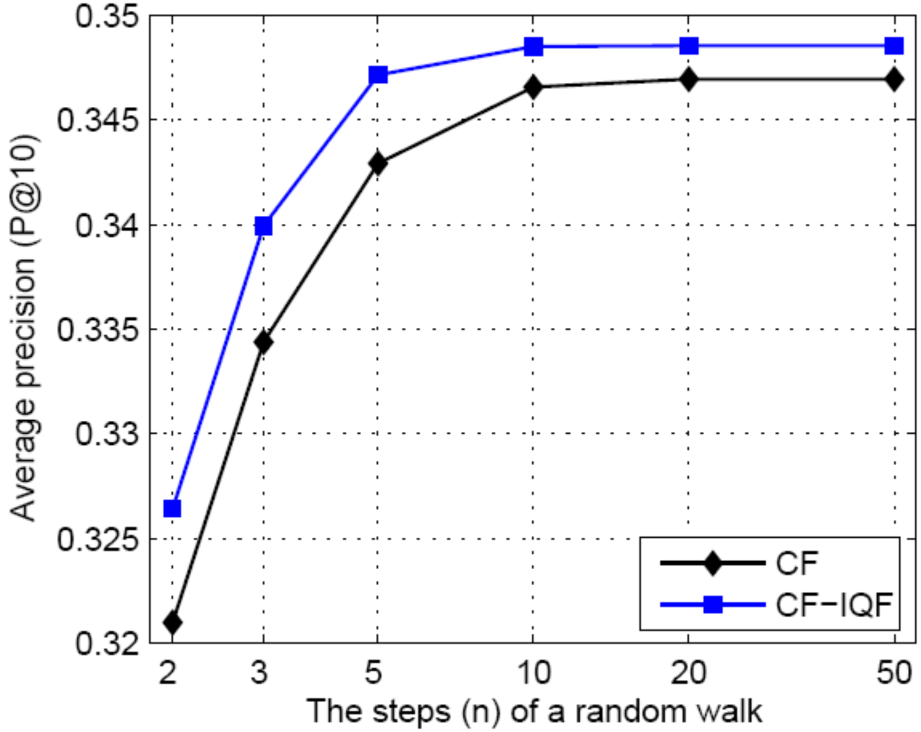
$$p(q_j|q_i) = \sum_{k \in D} p(d_k|q_i)p(q_j|d_k)$$

- The personalized PageRank

$$R_j^{n+1} = (1-\alpha)R_j^{(0)} + \alpha \cdot \sum_i p(q_j|q_i)R_i^n$$

Experimental Results

Random Walk Evaluation



Results:

1. With the increase of n, both models improve their performance.
2. CF-IQF model always performs better than the CF model.

Experimental Results

Random Walk Evaluation

CF model	CF-IQF model
Query = aa	
american airlines	american airlines
alcoholics anonymous	alcoholics anonymous
aa.com	aa.com
airlines	airlines
Query = east texas real estate	
google	east texas acreage
east texas acreage	tyler real estate
texas real estate	tyler texas realtors
tyler real estate	texas real estate
Query = home gym equipment	
home gyms	home gyms
gym equipment	gym equipment
treadmills	treadmills
buy.com	edge 329 upright exercise bike

In general, the results generated by the CF and the CF-IQF models are similar, and mostly semantically relative to the original query, such as "American airline".

CF-IQF model can boost more relevant queries as suggestion and reduce some irrelevant queries.

Appendix C: Optimization Problem

Optimization problem:

$$\min_F \quad \frac{1}{2}\sum_{i,j=1}^{m+n} w_{ij}\left\|\frac{f_i}{\sqrt{d_{ii}}}-\frac{f_j}{\sqrt{d_{jj}}}\right\|^2+\mu\sum_{i=1}^{m+n}\left\|f_i-f_i^0\right\|^2$$

$$s.t. \quad W=\begin{bmatrix} W^{uu} & \beta\cdot W^{uv} \\ \beta\cdot W^{vu} & W^{vv}\end{bmatrix}$$

$$F=\begin{bmatrix} X \\ Y \end{bmatrix}$$

$$\beta=(1-\lambda_r)/\lambda_r,$$

Differentiating and simplifying:

$$\left.\frac{dR}{dF}\right|_{F=F^*}=F^*-SF^*+\mu(F^*-F^0)=0,$$

$$F^*-\frac{1}{1+\mu}SF^*-\frac{\mu}{1+\mu}F^0=0$$

Solution:

$$\boxed{\begin{aligned} &F^*=\mu_\beta(I-\mu_\alpha S)^{-1}F^0,\\ &S=D^{-\frac{1}{2}}WD^{-\frac{1}{2}}\\ &\mu_\alpha=\frac{1}{1+\mu},\ \text{and}\ \mu_\beta=\frac{\mu}{1+\mu},\end{aligned}}$$

Hongbo Deng
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Appendix D: Evaluation Metrics

- Precision at rank n (P@n):

$$P@n = \frac{\#\text{ relevant candidates in top } n \text{ results}}{n}$$

- Mean Average Precision (MAP):

$$AP = \frac{\sum_{n=1}^{N}(P@n * \mathrm{rel}(n))}{R}$$

- Bpref: The score function of the number of non-relevant candidates

$$\mathrm{bpref} = \frac{1}{R}\sum_{r=1}^{N}(1 - \frac{\#n \text{ ranked higher than } r}{R})$$