



Dr. Searcher and Mr. Browser: A unified hyperlink-click graph

Barbara Poblete, Carlos Castillo, Aristides Gionis

Yahoo! Research

Taller Web 2009

In this talk

- New unified web graph: **hyperlink-click** graph
- Union of the hyperlink graph and the click graph
- Random walk generates robust results that match or are better than the results obtained on either individual graph

Motivation

- Two types of web graphs: **click** and **hyperlink** graph
- Represent two most common tasks of users:
searching and **browsing**
- Correspond to two prototypical ways of looking for information: **asking** and **exploring**
- Edges capture semantic relations among nodes:
e.g., similarity and authority endorsement

Motivation...

Drawbacks:

- ☹ Hyperlink graph: adversarial increase in PageRank scores, i.e., spam or link farms
- ☹ Click graph: sparsity, inherent bias in web-search engine ranking, dependency on textual match and click spam

Our claim: Hyperlink and click graphs are complementary and can be used to alleviate the shortcomings of each other

Motivation...

Drawbacks:

- ☹ Hyperlink graph: adversarial increase in PageRank scores, i.e., spam or link farms
- ☹ Click graph: sparsity, inherent bias in web-search engine ranking, dependency on textual match and click spam

Our claim: Hyperlink and click graphs are complementary and can be used to alleviate the shortcomings of each other

Contribution

Introduce the hyperlink-click graph,
union of hyperlink and click graph

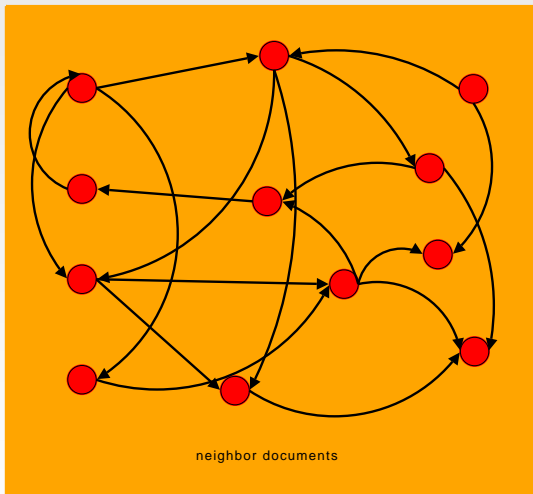
Consider random walk on the hyperlink-click graph

- 1 Model user behavior more accurately
- 2 Show that ranking with the hyperlink-click graph is similar to the best performance by either of the two graphs
- 3 The unified graph compensates where either graph performs poorly
- 4 more robust and fail-safe

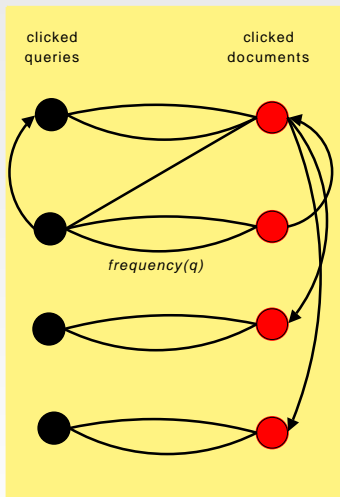
Uses of the hyperlink-click graph:

- Ranking of documents
- Query ranking and query recommendation
- Similarity search
- Spam detection

hyperlink graph

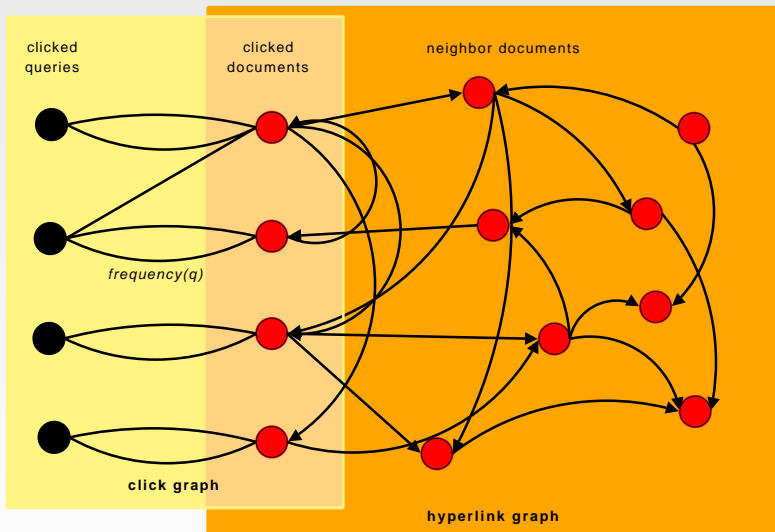


click graph



Web graphs...

hyperlink-click graph



Random walks on web graphs

Random walk on the hyperlink graph

- $\mathbf{P}_H = \alpha \mathbf{N}_H + (1 - \alpha) \mathbf{1}_H$

Random walk on the click graph

- $\mathbf{P}_C = \alpha \mathbf{N}_C + (1 - \alpha) \mathbf{1}_C$

Random walk on the hyperlink-click graph

- $\mathbf{P}_{HC} = \alpha \beta \mathbf{N}_C + \alpha(1 - \beta) \mathbf{N}_H + (1 - \alpha) \mathbf{1}$

Random walks on web graphs

Random walk on the hyperlink graph

- $\mathbf{P}_H = \alpha \mathbf{N}_H + (1 - \alpha) \mathbf{1}_H$

Random walk on the click graph

- $\mathbf{P}_C = \alpha \mathbf{N}_C + (1 - \alpha) \mathbf{1}_C$

Random walk on the hyperlink-click graph

- $\mathbf{P}_{HC} = \alpha \beta \mathbf{N}_C + \alpha(1 - \beta) \mathbf{N}_H + (1 - \alpha) \mathbf{1}$

Random walks on web graphs

Random walk on the hyperlink graph

- $\mathbf{P}_H = \alpha \mathbf{N}_H + (1 - \alpha) \mathbf{1}_H$

Random walk on the click graph

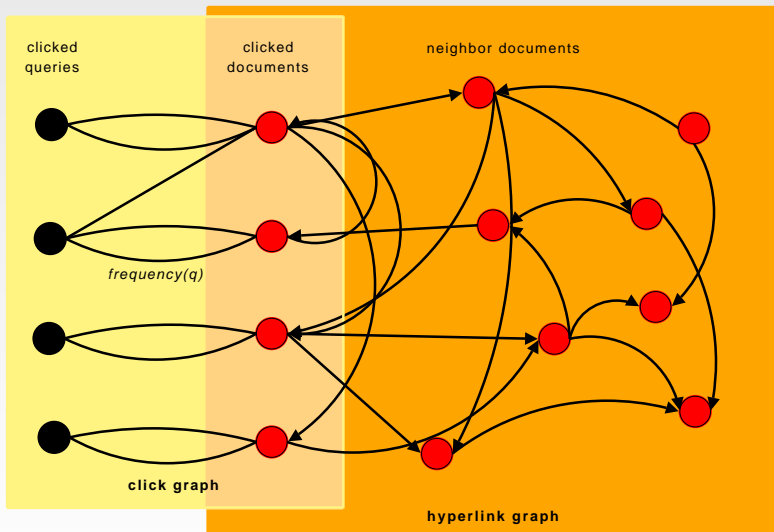
- $\mathbf{P}_C = \alpha \mathbf{N}_C + (1 - \alpha) \mathbf{1}_C$

Random walk on the hyperlink-click graph

- $\mathbf{P}_{HC} = \alpha \beta \mathbf{N}_C + \alpha(1 - \beta) \mathbf{N}_H + (1 - \alpha) \mathbf{1}$

Web graphs

hyperlink-click graph



Evaluation and results

- Validate utility of the random walk scores on the hyperlink-click graph
- Compare scores with those produced from the hyperlink and the click graph

Evaluation and results...

We focus on two **tasks** in which a good ranking method should perform well:

- ranking high-quality documents and
- ranking pairs of documents

The evaluation is centered on analyzing the dissimilarities among the different models

Evaluation and results...

Dataset:

- Yahoo! query log
- 9 K seed documents extracted from the query log
 - documents with at least 10 clicks
- 61 K queries with at least one click to the seed documents
- Using a web crawl expanded to 144 million documents
- The expansion considers all in- and out-neighbors of all documents in the seed set seed set.

Evaluation and results...

Two types of data:

- Without sponsored results: query log only with *algorithmic* results
- With sponsored results:

Random walk evaluation

- we compute scores on the complete datasets, but
- we evaluate only on documents in the intersection of the three graphs
- Two tasks:
 - ① ranking high-quality documents (DMOZ),
 - ② ranking pairs of documents (user evaluation).

Evaluation and results...

DMOZ:

Π_Z : Our first measure is the normalized sum of the π scores of D_Z documents.

$$\Pi_Z = \frac{\sum_{d \in D_Z} \pi(d)}{\sum_{d \in D_C} \pi(d)}$$

Γ_Z : *Goodman-Kruskal Gamma* Γ measure between the rankings

$$\Gamma = \frac{D - A}{D + A}$$

Evaluation and results...

User study:

- 13 users
- 1 710 assessments
- users expressed not neutral opinion in 32% of document pairs
- Use Γ to measure pairs of documents on which rankings agree with users

Evaluation and results...

DMOZ:

- **Macro-evaluation:** captures the overall scores of high-quality documents

Π_Z and Γ_Z are computed considering all the documents in the evaluation set

- **Micro-evaluation:** at query level

Π_Z and Γ_Z in the evaluation set are reduced to only those documents clicked from a particular query.

(User study: micro-evaluation)

Evaluation and results...

DMOZ:

- **Macro-evaluation:** captures the overall scores of high-quality documents

Π_Z and Γ_Z are computed considering all the documents in the evaluation set

- **Micro-evaluation:** at query level

Π_Z and Γ_Z in the evaluation set are reduced to only those documents clicked from a particular query.

(User study: micro-evaluation)

Evaluation and results...

DMOZ

metric	macro	
	without sponsored results	with sponsored results
Γ_Z	$G_C \approx G_{HC} > G_H$	$G_C > G_{HC} > G_H$
Π_Z	$G_H \approx G_{HC} > G_C$	$G_H \approx G_{HC} > G_C$

metric	micro	
	without sponsored results	with sponsored results
Γ_Z	$G_C \approx G_{HC} > G_H$	$G_{HC} \approx G_C > G_H$
Π_Z	$G_{HC} \approx G_C > G_H$	$G_{HC} \approx G_H \approx G_C$

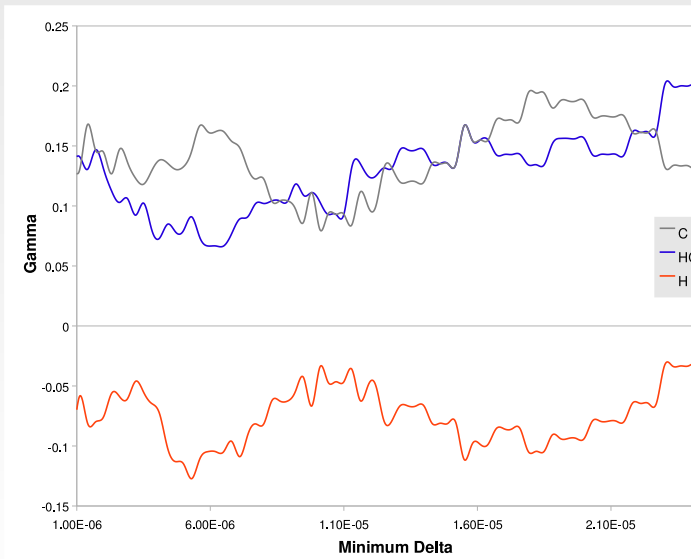
Evaluation and results...

User study

metric	without sponsored results	with sponsored results
Γ	$G_C > G_{HC} > G_H$	$G_H > G_{HC} > G_C$

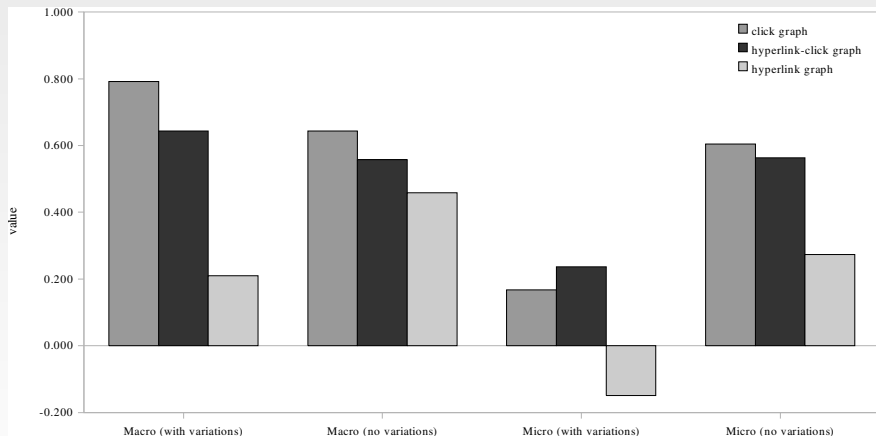
Evaluation and results...

Exclude documents with very similar scores



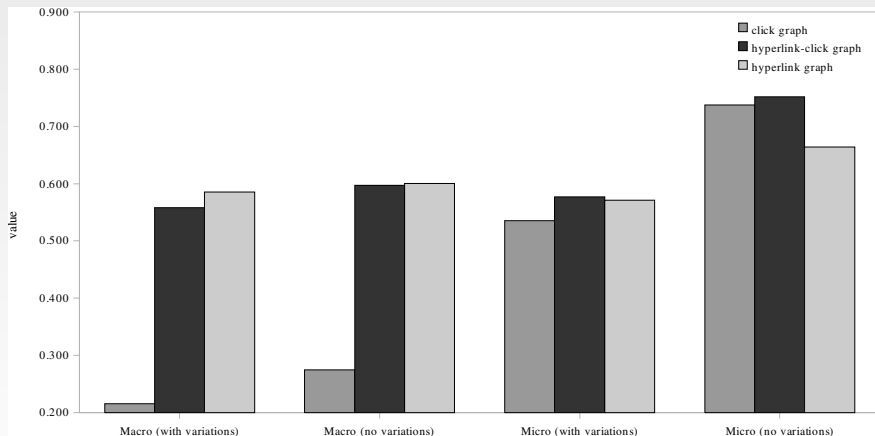
Evaluation and results...

Summary of Γ_Z values for DMOZ:



Evaluation and results...

Summary of Π_Z values for DMOZ:



Conclusions

- We study random walk on a unified web graph
- Intended to model user searching and browsing behavior
- Several combinations of metrics are used for evaluation

Experimental evaluation shows:

- The unified graph is always close to the best performance of either the click or the hyperlink graph

Future work

- Analyze how to deal with the inherent bias that exists in any ranking technique based on usage mining
- Study other web mining applications
 - 1 Link and click spam detection
 - 2 Similarity search
 - 3 Query recommendation



Questions?