

RELEVANCE FEEDBACK IN WEB SEARCH

SERGEI VASSILVITSKII (STANFORD UNIVERSITY)
ERIC BRILL (MICROSOFT RESEARCH)

INTRODUCTION

- Web search is a non-interactive system.
 - Exceptions are spell checking and query suggestions
 - By design search engines are stateless
- But many searches become interactive:
 - query, get results back, reformulate query...
 - Can use interaction to retrieve user intent

RELEVANCE FEEDBACK

Windows Live Beta

sigir

Web News Images Local Feeds Academic Products

sigir 1-7 (54,926) +add to live.com




Special Inspector General for Iraq Reconstruction : SIGIR Homepage
Welcome to the Office of the Special Inspector General for Iraq Reconstruction (**SIGIR**), a temporary federal agency serving the American public as a watchdog for fraud ...
www.sigir.mil/Default.aspx

ACM SIGIR Special Interest Group on Information Retrieval Home Page
Welcome to the ACM **SIGIR** Web site. ACM **SIGIR** addresses issues ranging from theory to user demands in the application of computers to the acquisition, organization, storage, retrieval, and ...
www.sigir.org

SIGIR 2004
The major international forum for the presentation of new research results and demonstration of new ... **SIGIR** is the major international forum for the presentation of new research results and the ...
www.sigir.org/sigir2004

+ Show more results from this site

SIGIR 2006—Seattle
29th Annual International ACM **SIGIR** Conference on Research & Development on Information Retrieval, Seattle 2006 ...
www.sigir2006.org



USING THIS INFORMATION

- Classical methods: e.g. Rocchio's term reweighing (TFiDF) + cosine similarity scores.
- There is more information here: what can the structure of the web tell us?

HYPOTHESIS

- For a given query:
 - Relevant pages tend to point to other relevant pages.
 - ➔ Similar to Pagerank.

HYPOTHESIS

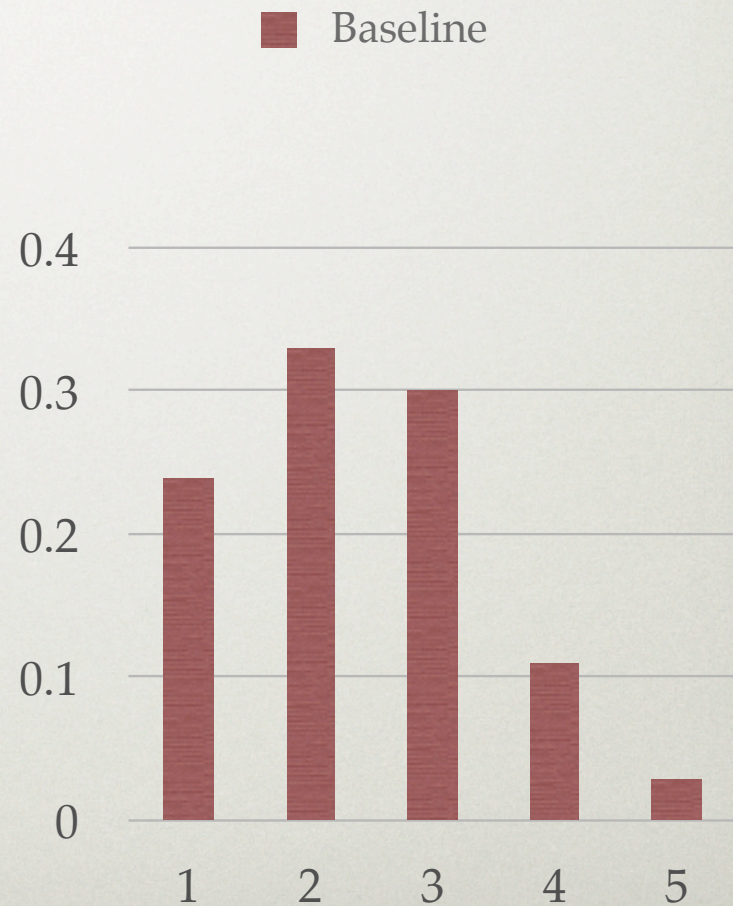
- For a given query:
 - Relevant pages tend to point to other relevant pages.
 - ➔ Similar to Pagerank.
 - Irrelevant pages tend to be pointed to by other irrelevant pages.
 - ➔ “Reverse Pagerank”
 - ➔ Those who point to web spam are likely to be spammers.

DATASET

- Dataset
 - 9500 queries
 - For each query 5 - 30 result URLs
 - each URL rated on a scale of 1 (poor) to 5 (perfect)
 - Total 150,000 (query, url, rating) triples
- Will use this data to simulate relevance feedback
 - Only reveal the ratings for some URLs

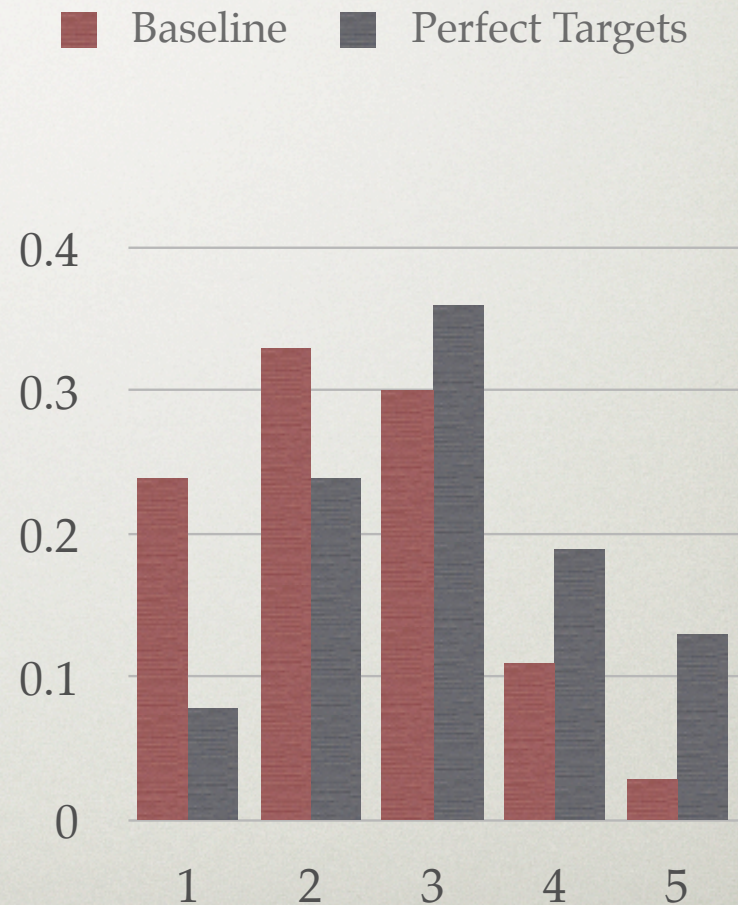
HYPOTHESIS VALIDATION

- Relevance distribution of all URLs in the dataset



HYPOTHESIS VALIDATION

- Relevance distribution of all URLs in the dataset
- Compared to the URLs that are targets of perfect results



TOWARDS AN ALGORITHM

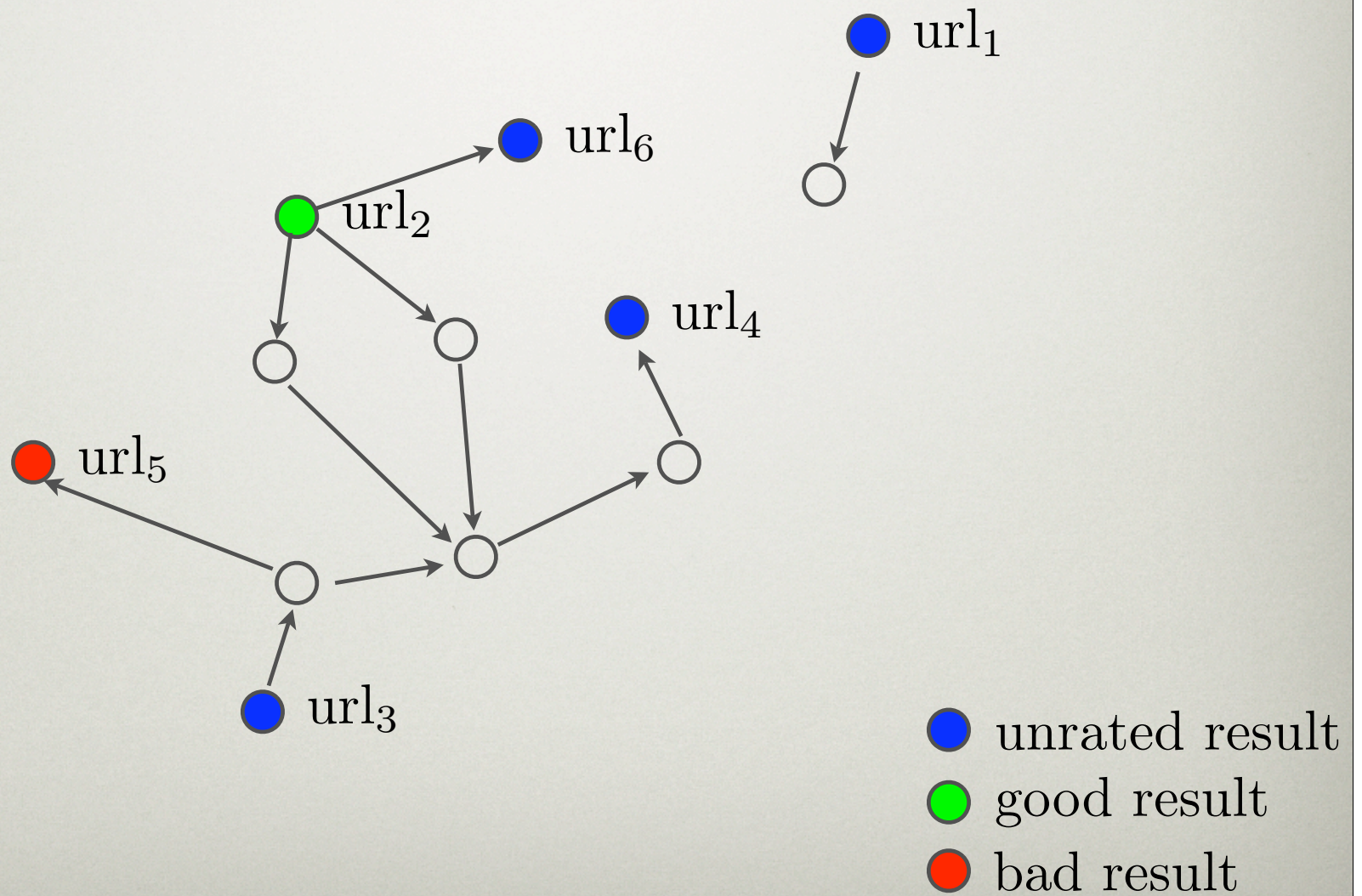
- url₁
- url₂
- url₃
- url₄
- url₅
- url₆

TOWARDS AN ALGORITHM

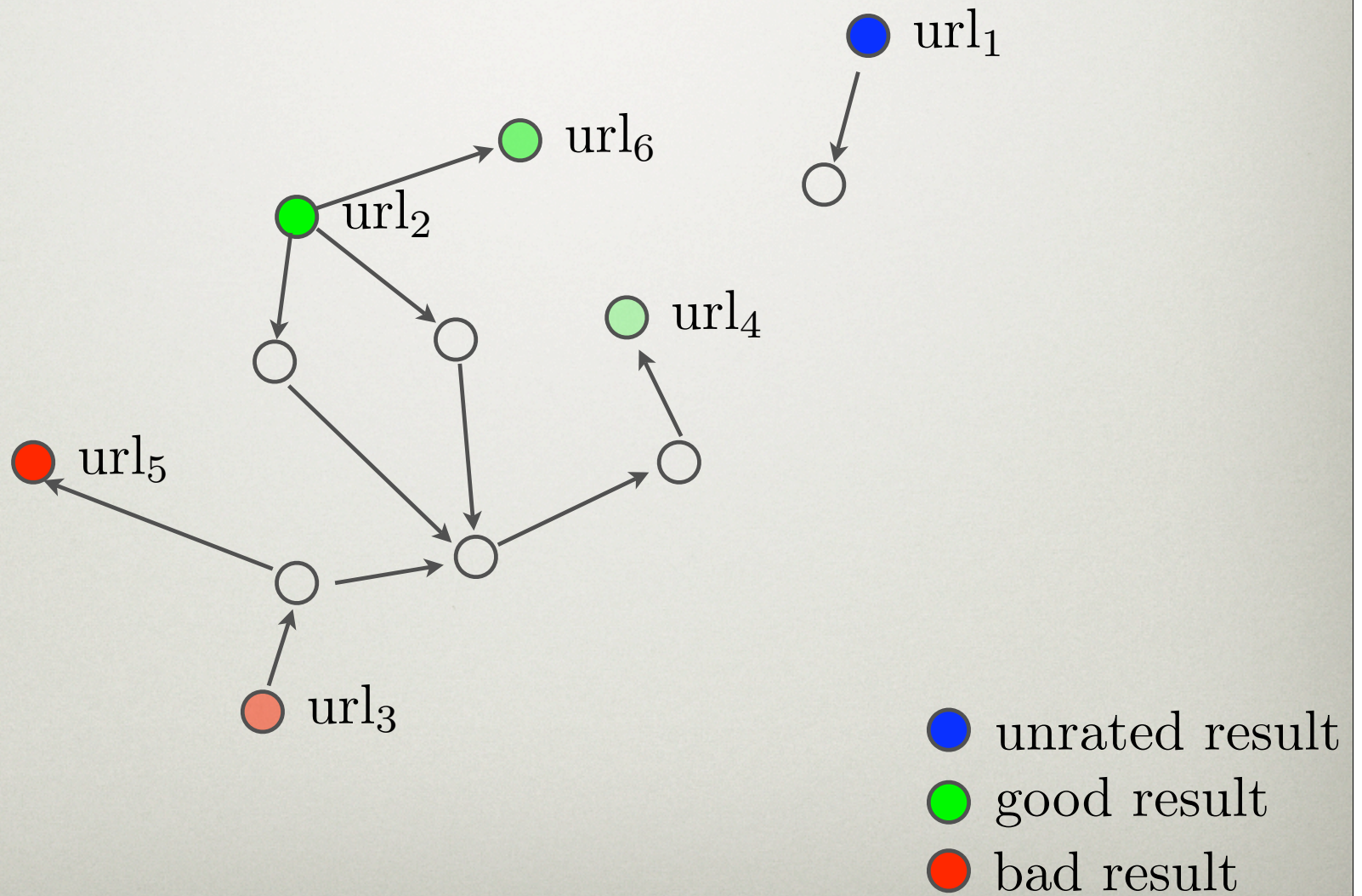
- url₁
- url₂
- url₃
- url₄
- url₅
- url₆

- unrated result
- good result
- bad result

TOWARDS AN ALGORITHM

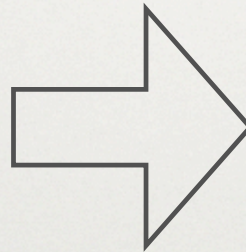


TOWARDS AN ALGORITHM



TOWARDS AN ALGORITHM

- url₁
- url₂
- url₃
- url₄
- url₅
- url₆



- url₂
- url₆
- url₁
- url₄
- url₃
- url₅

- unrated result
- good result
- bad result

PERCOLATING THE RATINGS

- Calculate the effect on u
 - Begin with a probability distribution on relevance of u (Baseline histogram)
 - For all highly rated documents v
 - If there exists a short $v \rightarrow u$ path, update u .
 - For all irrelevant documents v
 - If there exists a short $u \rightarrow v$ path, update u .
- Combine the static score together with the relevance information

ALGORITHM PARAMETERS

- If there exists a “short” path...
 - Strength of signal decreases with length
 - Recall of the system increases with length
 - Computational considerations
 - Looked at paths of 4 hops or less

ALGORITHM PARAMETERS

- If there exists a “short” path...
 - Strength of signal decreases with length
 - Recall of the system increases with length
 - Computational considerations
 - Looked at paths of 4 hops or less
- ...update u .
 - Maintain a probability distribution on the relevance of u .

EXPERIMENTAL SETUP

- For each query in the dataset split the URLs into
 - Train: the relevance is revealed to the algorithm
 - Test: Only the static score is revealed
- Compare the ranking of the test URLs by their static score vs. static + RF scores.

EVALUATION MEASURE

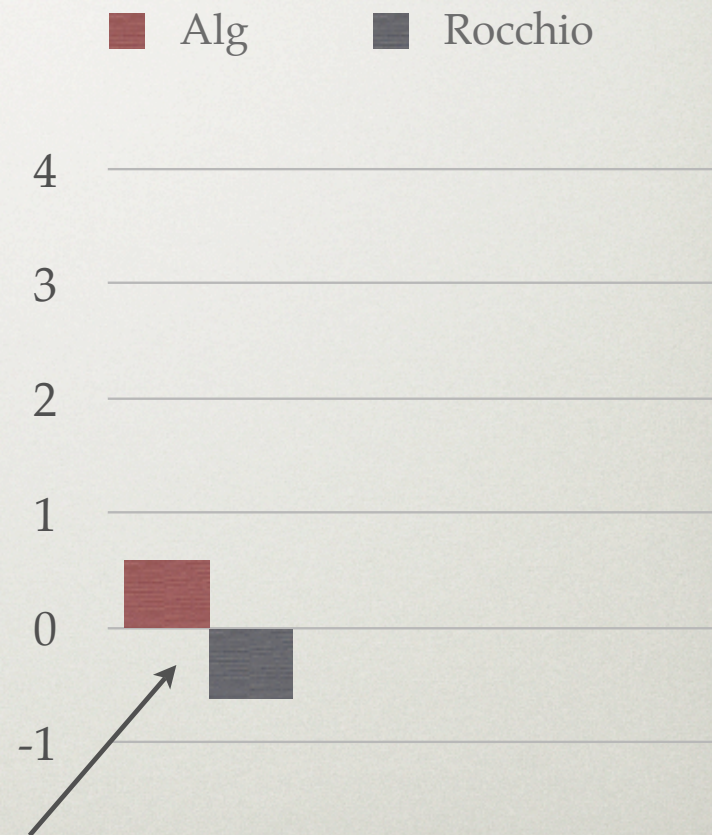
- Measure: NDCG (Normalized Discounted Cumulative Gain):

$$NDCG \propto \sum_i \frac{2^{rel(i)} - 1}{\log(1 + i)}$$

- Why NDCG?
 - sensitive to the position of highest rated page
 - Log-discounting of results
 - Normalized for different lengths lists

RESULT SUMMARY

- NDCG change for three subsets of pages.
- Complete Dataset



Rocchio: Demotes the best result

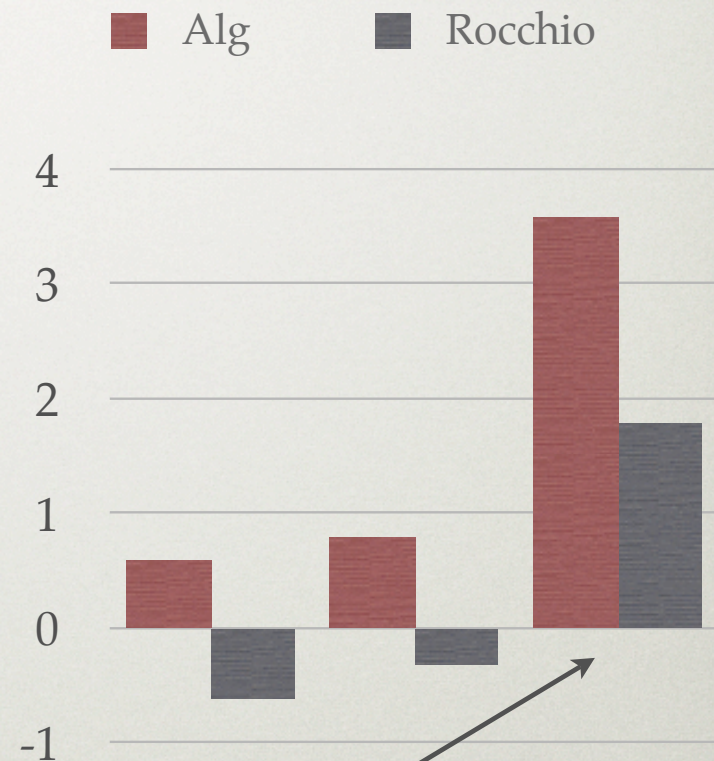
RESULT SUMMARY

- NDCG change for three subsets of pages.
- Complete Dataset
- Only queries with $NDCG < 100$



RESULT SUMMARY

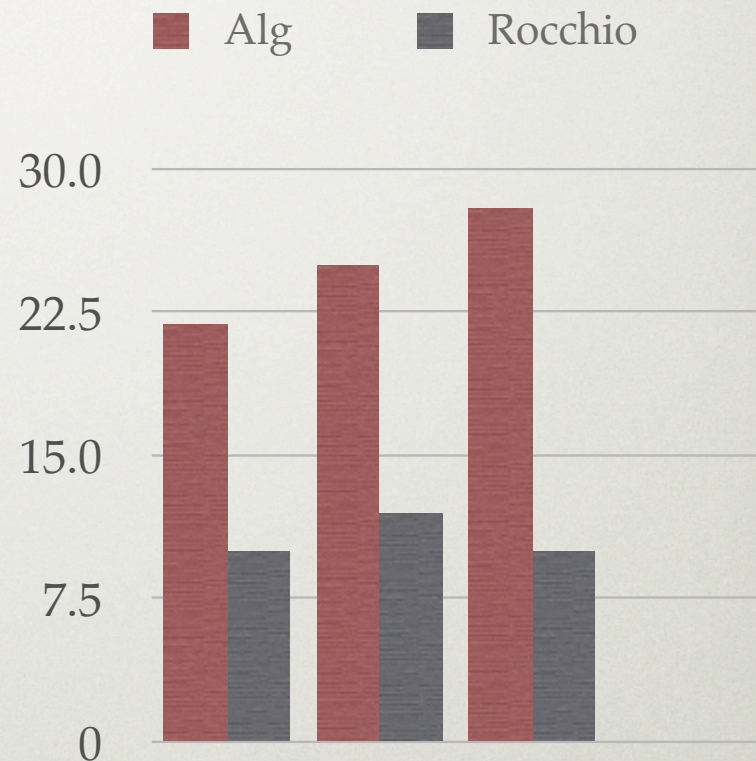
- NDCG change for three subsets of pages.
- Complete Dataset
- Only queries with $\text{NDCG} < 100$
- Only queries with $\text{NDCG} < 85$



Increased performance for harder queries

RESULT SUMMARY (2)

- Recall for the three datasets.
- Complete Dataset
- Only Queries with $NDCG < 100$
- Only Queries with $NDCG < 85$



RESULTS SUMMARY (3)

- Many more experiments:
 - How does the number of URLs rated affect the results?
 - Are some URLs better to rate than others?
 - Can we predict when recall will be low?

FUTURE WORK

- Hybrid Systems: Combining text based and link based RF approaches
- Learning feedback based on clickthrough data
- Large scale experimental evaluation of different RF approaches

THANK YOU

ANY QUESTIONS?