Quality-Biased Ranking of Web Documents

Michael Bendersky

W. Bruce Croft

Yanlei Diao

University of Massachusetts Amherst

Talk Outline

I. **Document Quality in Web Search**

II. **Ranking with Quality Bias**

III. **Evaluation**

IV. **Discussion & Conclusions**

Document Quality in Web Search

Document Quality in Web Search (I)

- Web is decentralized and heterogeneous
- Web pages vary by
  - *Authority*
  - *Goals*
  - *Credibility*
  - *Publishing Standards*
- Document quality in search engines
  - *Promote high-quality content*
  - *Demote low-quality content*

Document Quality in Web Search (II)

- Actively pursued by the commercial search engines

> *"As pure web spam has decreased over time, attention has shifted instead to sites with shallow or low-quality content. In 2010, we launched two major algorithmic changes focused on low-quality sites."*
>
> ***"Google search and search engine spam"***
> ***Matt Cutts (Official Google Blog, Jan. 2011)***

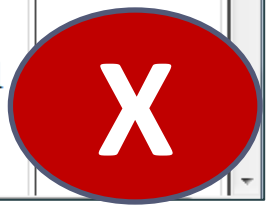
- Surprisingly little publicly available research
  - *Most research focuses on spam detection and not evaluation in the retrieval context* ***(Castillo & Davison, 2011)***

"LOW WHITE BLOOD CELL COUNT" (1ST RANK)

Texas Department of Health - Licensure and Enforcement Division

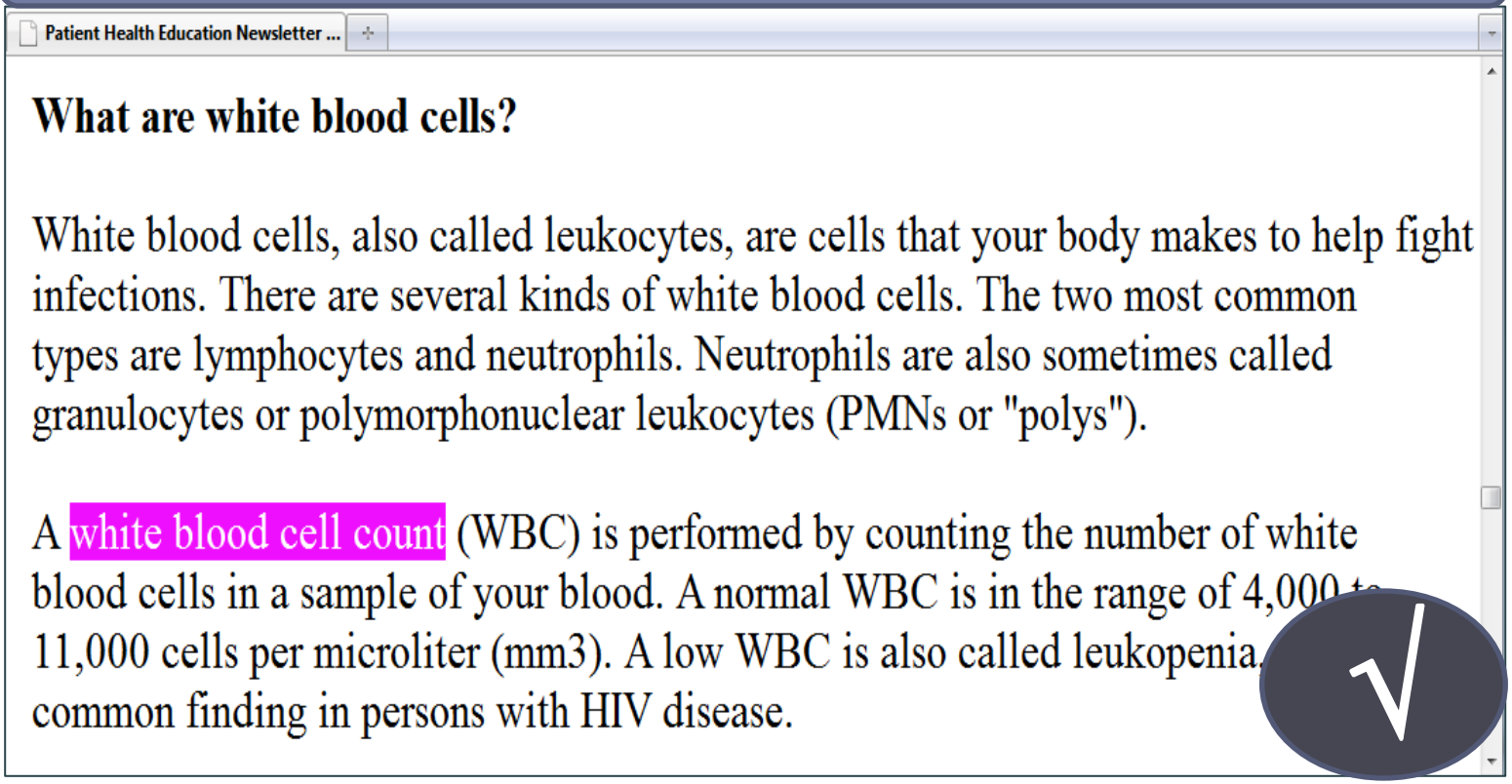
Texas Department of Health

License/Permit Number/CF#	Location of Facility	Nature of Violation	Discipline/Enforcement Action	Date Degraded	Date Regraded (if applicable)
078904	Upshur County	Allegations relating to: 1) Systemic white blood cell count.	Suspension of permit.	12/05/01	12/07/01
016390	Tom Green County	Allegations relating to: 1) Systemic white blood cell count	Suspension of permit.	12/06/01	



"LOW WHITE BLOOD CELL COUNT" (3RD RANK)

Dept. of Veterans Affairs - Patient Health Education Newsletter


Patient Health Education Newsletter ...

What are white blood cells?

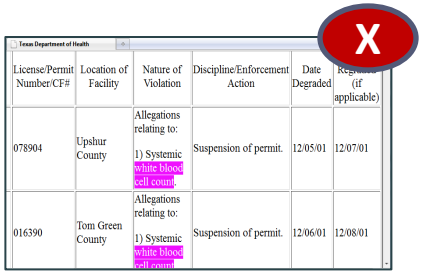
White blood cells, also called leukocytes, are cells that your body makes to help fight infections. There are several kinds of white blood cells. The two most common types are lymphocytes and neutrophils. Neutrophils are also sometimes called granulocytes or polymorphonuclear leukocytes (PMNs or "polys").

A white blood cell count (WBC) is performed by counting the number of white blood cells in a sample of your blood. A normal WBC is in the range of 4,000 t
11,000 cells per microliter (mm3). A low WBC is also called leukopenia.
common finding in persons with HIV disease.

√

"LOW WHITE BLOOD CELL COUNT" QUALITY BIAS EFFECT

#1

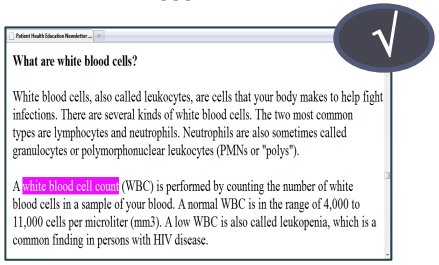


License/Permit Number/CF#	Location of Facility	Nature of Violation	Discipline/Enforcement Action	Date Degraded	Reinstated (if applicable)
078904	Upshur County	Allegations relating to: 1) Systemic white blood cell count	Suspension of permit.	12/05/01	12/07/01
016390	Tom Green County	Allegations relating to: 1) Systemic white blood cell count	Suspension of permit.	12/06/01	12/08/01

X

...

#3



What are white blood cells?

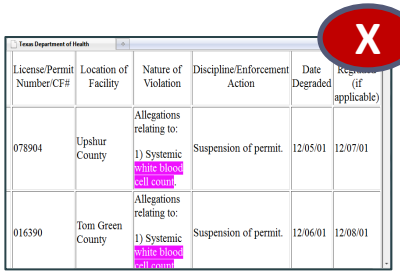
White blood cells, also called leukocytes, are cells that your body makes to help fight infections. There are several kinds of white blood cells. The two most common types are lymphocytes and neutrophils. Neutrophils are also sometimes called granulocytes or polymorphonuclear leukocytes (PMNs or "polys").

A white blood cell count (WBC) is performed by counting the number of white blood cells in a sample of your blood. A normal WBC is in the range of 4,000 to 11,000 cells per microliter (mm3). A low WBC is also called leukopenia, which is a common finding in persons with HIV disease.

√

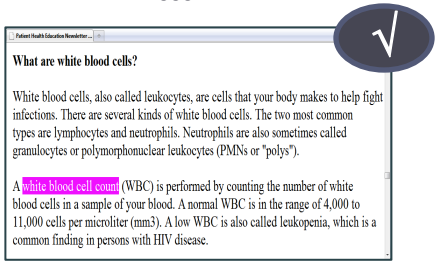
"LOW WHITE BLOOD CELL COUNT"
QUALITY BIAS EFFECT

#1

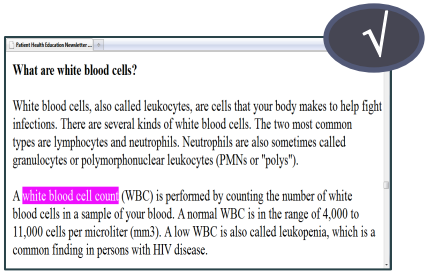


...

#3



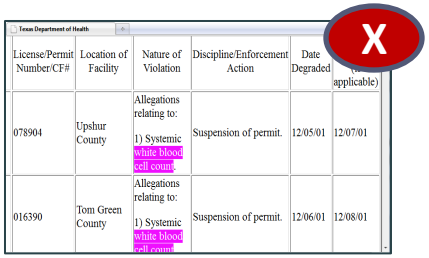
#1



...

...

#11



Document Quality and Spam

- Quality of a web page is determined by many factors
  - *Original, up-to-date content of genuine value*
  - *Links to related resources*
  - *Layout for easy reading and navigation*

- Accordingly, quality should be viewed as a ***continuous spectrum*** ranging from high-quality pages to spam

- Most web documents are somewhere between these two extremes

Document Quality and Priors

- High quality of the web document content increases the a-priori probability of the document being relevant
  - *a.k.a. document prior*
- Quality factors should be combined in a way that directly improves the retrieval effectiveness
  - *e.g., nDCG or MAP*
- If possible, should be combined with other document priors
  - *Link-Based – PageRank* ***(Brin & Page, 1998)***, *SALSA* ***(Najork, 2007)***
  - *User-Based – BrowseRank* ***(Liu et al., 2008)***, ***(Richardson et al., 2006)***
  - *Length-Based – Pivoted length normalization* ***(Singhal et al., 1996)***

Ranking with Quality Bias

II

MRF-IR MODEL *(METZLER & CROFT, 2005)*

- ***Score(Q,D)*** is determined by a Markov Random Field that describes the dependency between document ***D*** and query ***Q***

- Principled way to capture
  - *Single term matches*
  - *Exact phrase matches*
  - *Proximity matches*

- State-of-the-art retrieval effectiveness
  - *Web Search* ***(METZLER & CROFT, 2005, BENDERSKY ET AL. 2010)***
  - *Blog Search* ***(ELSAS ET AL., 2008)***
  - *Text REtrieval Conference* ***(WEB TRACK 2009-2010)***

MRF-IR Ranking

$$Score(Q,D) = \lambda_T f_T(q_i, D) +$$
Term Match

$$\lambda_O f_O(q_i q_{i+1}, D) +$$
Exact Match

$$\lambda_U f_U(q_i q_{i+1}, D)$$
Proximity Match

MRF-IR Ranking with Quality Bias

$$Score(Q,D) = \lambda_T f_T(q_i, D) +$$
$$\lambda_O f_O(q_i q_{i+1}, D) +$$
$$\lambda_U f_U(q_i q_{i+1}, D) +$$
$$\sum_{l \in L} \lambda_l f_l(D)$$

Term Match

Exact Match

Proximity Match

Quality Bias

Learning to Rank with Quality Bias (I)

- No explicit grading of documents by quality

$$Score(Q,D) = \lambda_T f_T(q_i, D) + \lambda_O f_O(q_i q_{i+1}, D) + \lambda_U f_U(q_i q_{i+1}, D) + \sum_{l \in L} \lambda_l f_l(D)$$

- No need in labeled data beyond relevance judgments

- Document quality bias is an integral part of ranking

Learning to Rank with Quality Bias (II)

- Learn the weights Λ to directly optimize retrieval metric
  - *MAP*
  - *nDCG*

$$Score(Q,D) = \lambda_T f_T(q_i, D) + \lambda_O f_O(q_i q_{i+1}, D) + \lambda_U f_U(q_i q_{i+1}, D) + \sum_{l \in L} \lambda_l f_l(D)$$

- We use a ***coordinate ascent algorithm*** *(Metzler & Croft, 2007)*
  - *Efficient for a relatively small number of parameters*
  - *Empirically good performance*
- However, most other LR4IR methods can be adopted
  - *SVM-Rank, Lambda-Rank, …*

Quality Factors (I)

- **Visible Text**
  - *Number of visible terms*
  - *Fraction of visible text page* ***(Zhu & Gauch, 2000)***
- **Readability**
  - *Average term length* ***(Kanungo & Orr, 2009)***
  - *Stopwords ratio* ***(Ntoulas et al., 2006)***
  - *Stopwords coverage* ***(Ntoulas et al., 2006)***

Quality Factors (II)

- **Cohesiveness**
  - *Entropy of the page content*
- **Ease of Navigation**
  - *Numbers of terms in the page title*
  - *The depth of the URL path* ***(Kraaj et al., 2002)***
  - *Fraction of table text*
- **Link Provision**
  - *Fraction of anchor text* ***(Ntoulas et al., 2006)***

Evaluation

III

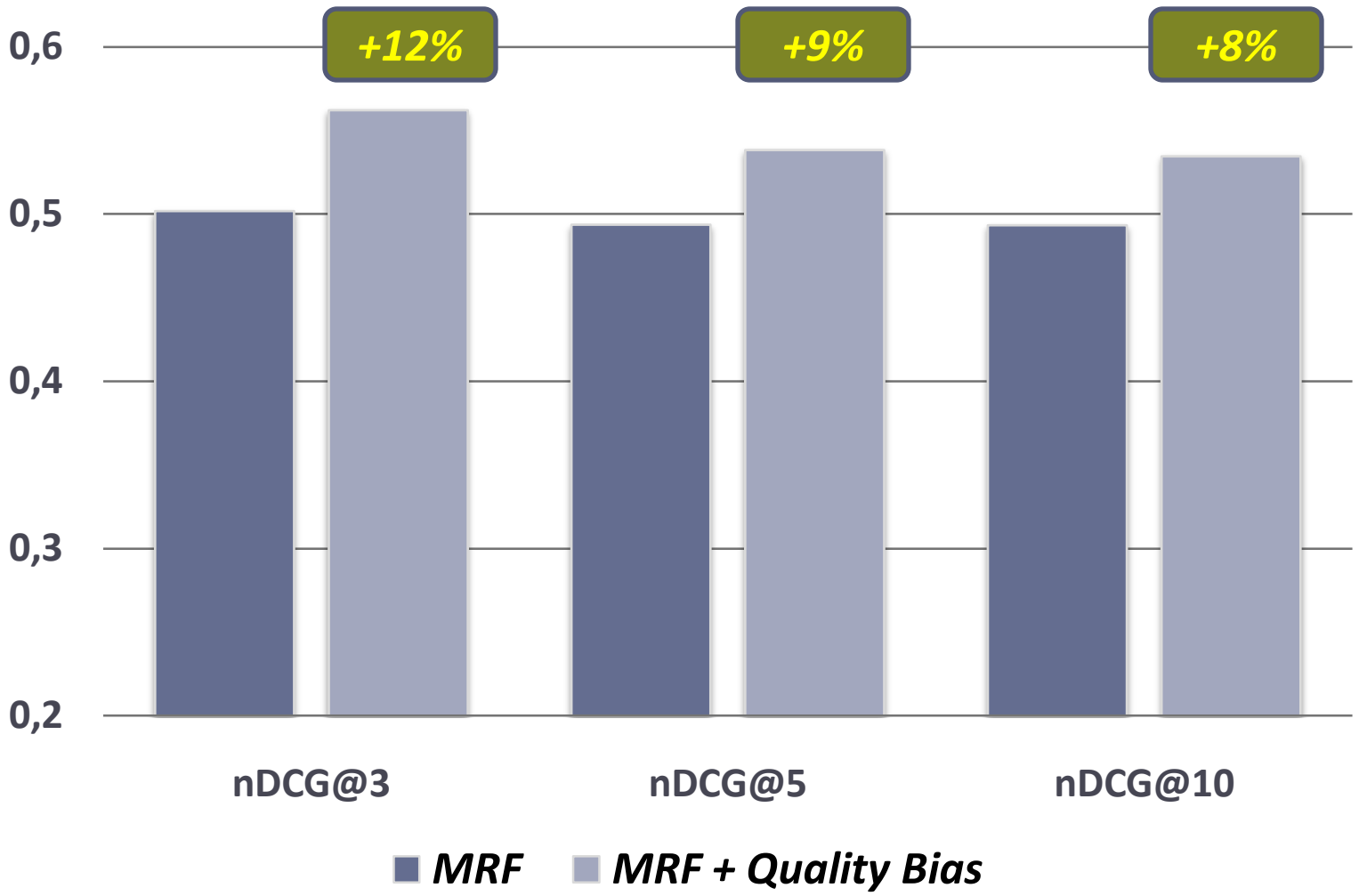
Document Collections

- Focus on collections with ***no explicit spam***

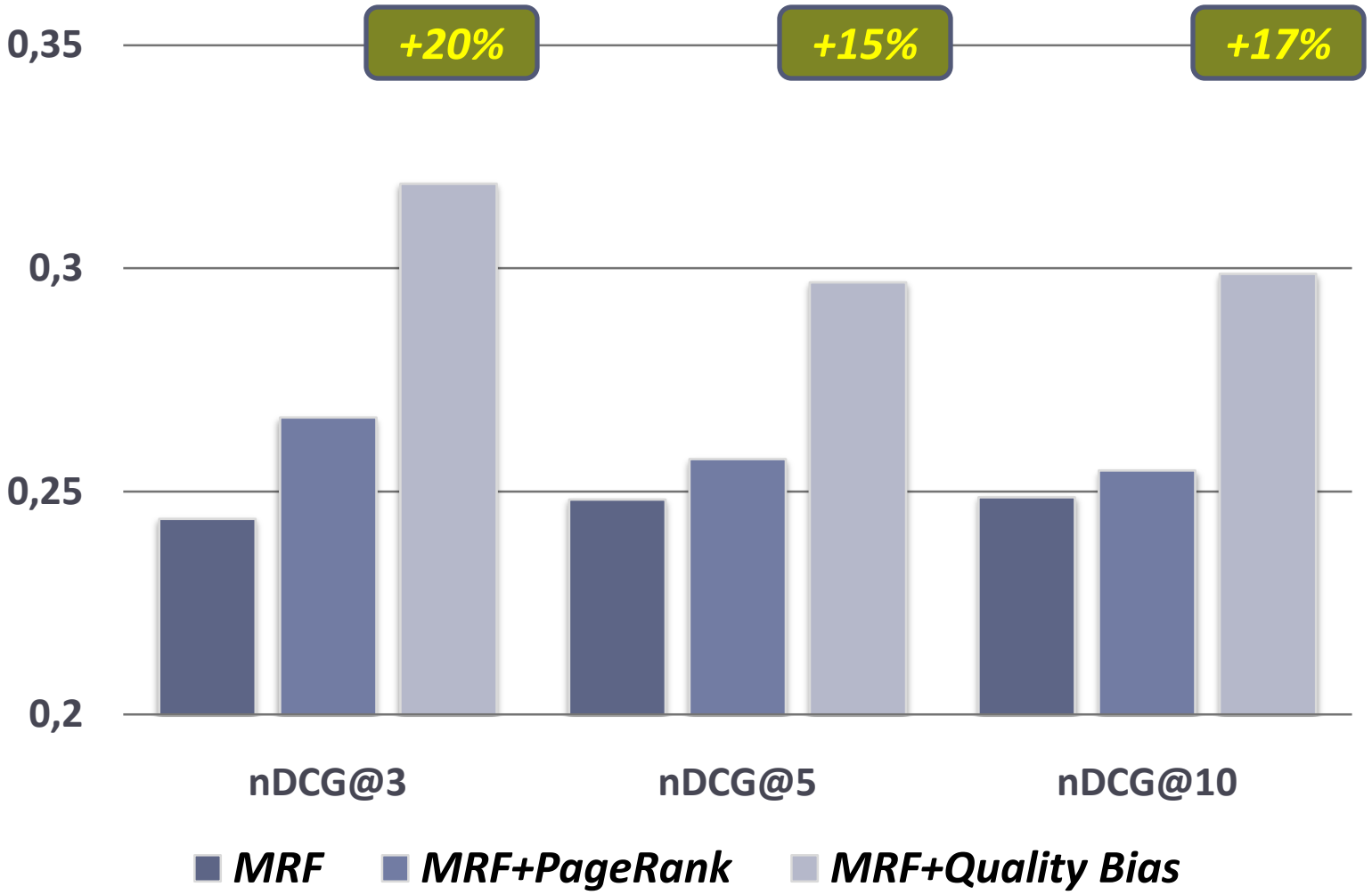
- **GOV2 collection**
  - *Crawl of .gov domain and associated topics for evaluation*
  - *Specialized web collection*
  - *No explicit spam pages*

- **ClueWeb collection**
  - *500M web pages and associated topics for evaluation*
  - *General web collection*
  - *Spam filter applied* ***(Cormack et al., 2010)***

GOV2



ClueWeb


0,35
0,3
0,25
0,2
+20%
+15%
+17%
nDCG@3
nDCG@5
nDCG@10
MRF
MRF+PageRank
MRF+Quality Bias

Discussion & Conclusions

IV

Discussion

- **Quality biased ranking in GOV2 collection**
  - *Significantly improves nDCG & MAP*
  - *Significantly demotes non-informative pages*
    - *Long tables and lists*
    - *Data dumps*
    - *Link listings*

- **Quality biased ranking in ClueWeb collection**
  - *Significantly improves nDCG & MAP even after spam filtering*
  - *Implicitly promotes Wikipedia pages*
    - *5 times more likely to retrieve a Wikipedia page in top-10 results*

Conclusions

- **A straightforward approach for introducing quality bias into state-of-the-art retrieval model**
- **Simple-to-compute content-based features**
  - *E.g., readability, page layout, navigation*
  - *Easy to extend with more sophisticated quality features*
- **Significant retrieval effectiveness improvements**
  - *Specialized web collection containing documents of differing quality*
  - *General web collection with spam filter*

Thank you!