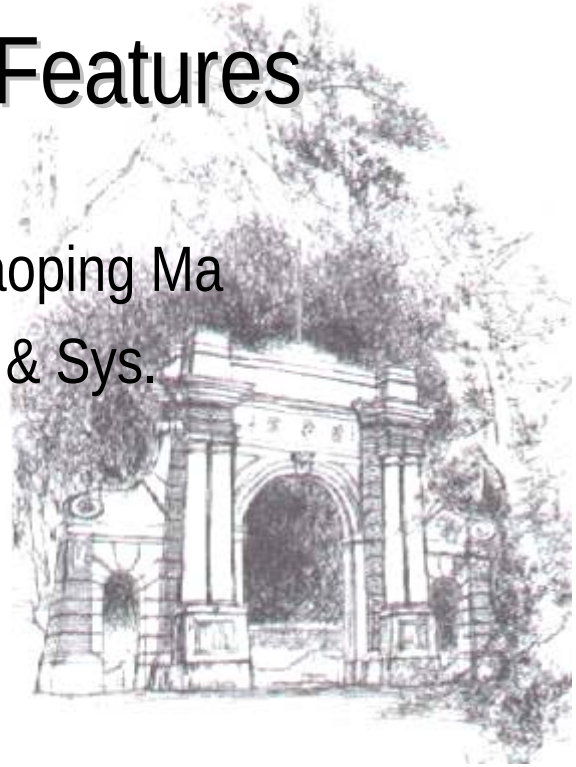
Data Cleansing for Web Information Retrieval using Query Independent Features

Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma
State Key Lab of Intelligent Tech. & Sys.
Tsinghua University

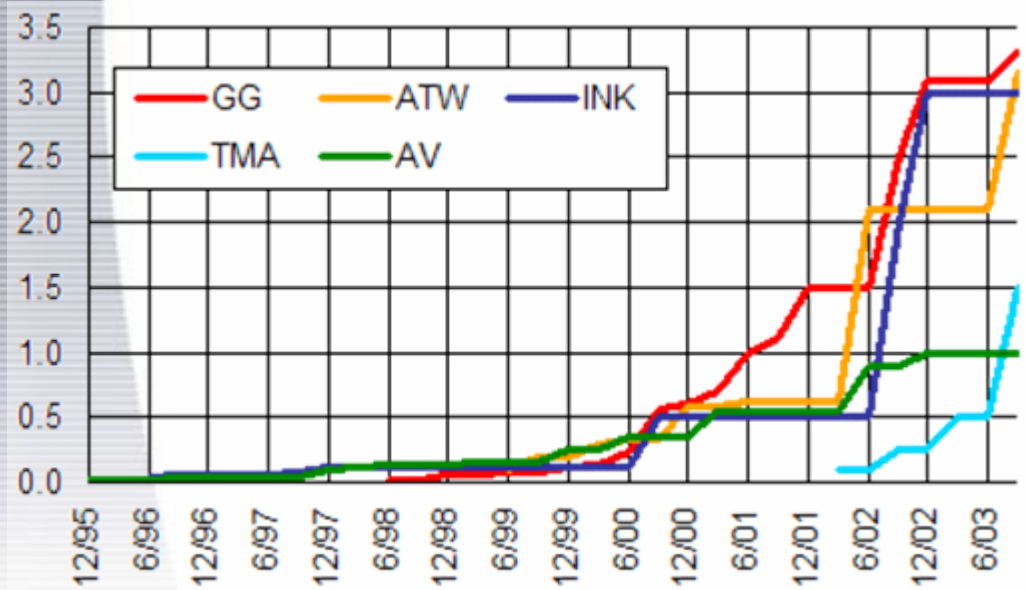


Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work

Data cleansing and its applications in Web IR

- Index Size War between Search Engines
  - Billions Of Textual Documents Indexed December 1995-September 2003



From Danny Sullivan, SearchEngineWatch web site

Data cleansing and its applications in Web IR

- Index Size War between Search Engines (cont.)

Search Engine	Reported Size	Page Depth
Google	**8.1 billion (Dec. 2004)**	101K
MSN	5.0 billion	150K
Yahoo	19.2 bilion (Aug. 2005)	500K
Ask Jeeves	2.5 billion	101K+
All the Web	**152 billion**	605K
All the Surface Web	**10 billion**	8K

From Danny Sullivan, SearchEngineWatch web site

Data cleansing and its applications in Web IR

- An end to the index size war?
  - No search engine can cover all resources on the Web

	Google	Yahoo!	MSN	Teoma
Round 1	76.30%	69.28%	62.03%	57.58%
Round 2	76.09%	69.29%	61.90%	57.69%
Round 3	76.27%	69.37%	61.87%	57.70%
Round 4	76.05%	69.30%	61.73%	57.57%
Round 5	76.11%	69.26%	61.96%	57.56%
Average	76.16%	69.32%	61.90%	57.62%

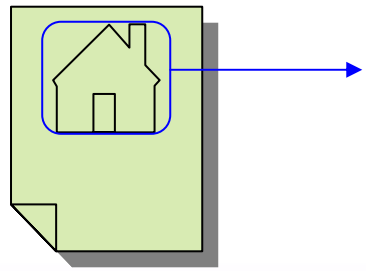
  - In Sep. 2005, Google removes the number of indexed pages because "absolute numbers are no longer useful"

Data cleansing and its applications in Web IR

- Data quality is more important than quantity for Web IR tools
  - Spams and SEOs
  - Duplicates in Web pages
  - Unreliable, out-dated data
- Current data cleansing algorithms in Web IR
  - Local scale data cleansing
  - Global scale data cleansing

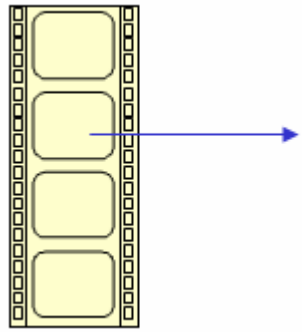
Data cleansing and its applications in Web IR

- Local scale data cleansing
  - To reduce the useless blocks / To find the important blocks inside a Web page
  - Reduce spam hyperlinks / useless hyperlinks (Kushmerick et. al.)
  - Reduce Ad. Contexts (Davison et. al.)
  - VIsion Based Page Segmentation, VIPS, MSRA
  - Site template detecting (Yossef et. al.)



Data cleansing and its applications in Web IR

- Global scale data cleansing
  - To reduce low quality pages / To locate important pages inside a given Web page corpus
  - Hyperlink structure analysis algorithms
    - PageRank, HITS
    - Hypothesis 1: Recommendation
    - Hypothesis 2: Topic locality
    - Challenged by Spam links and SEOs



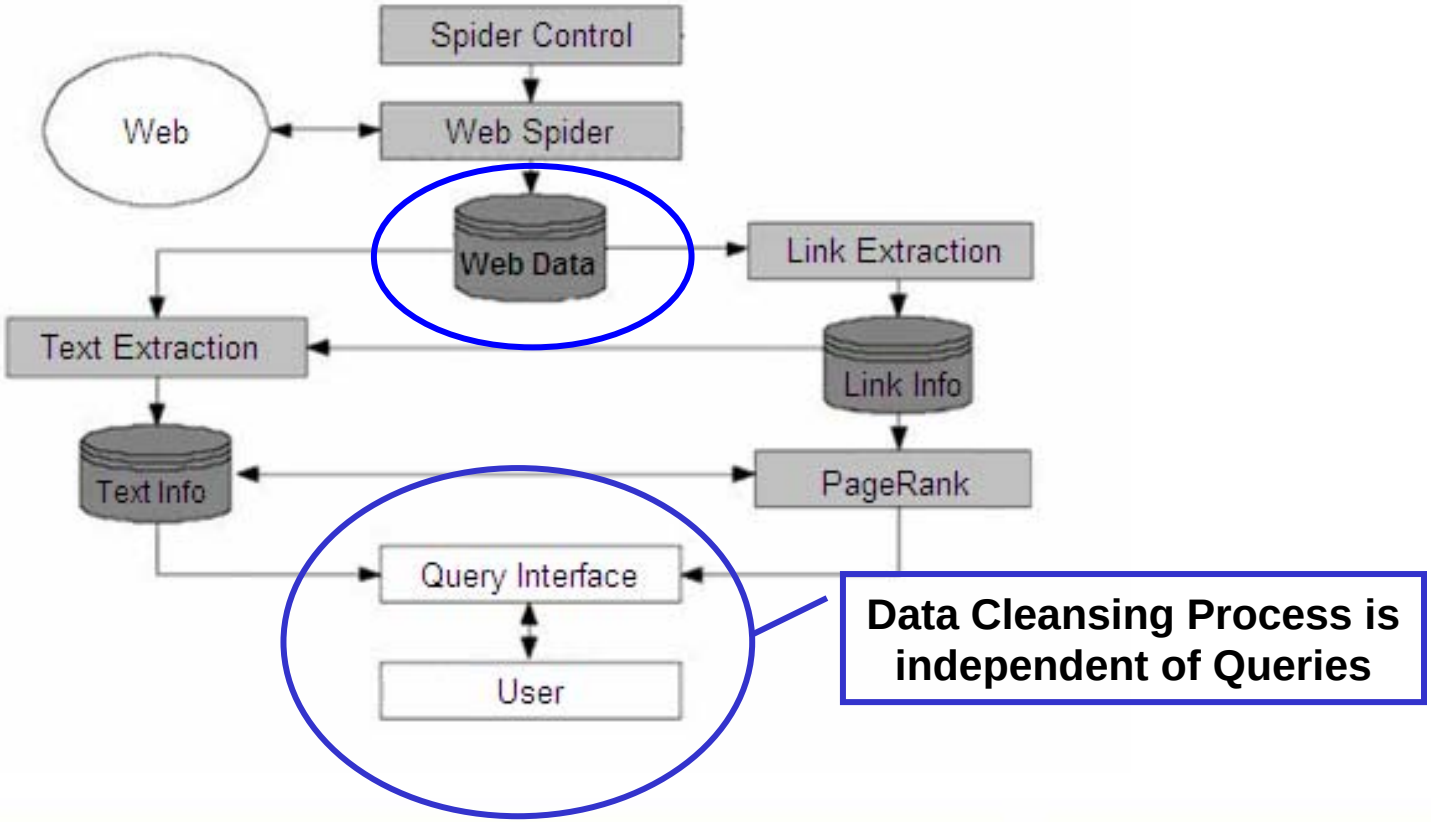
  - Monika Henzinger (Google Research Director): A better estimate of the quality of a page requires additional sources of information.

Data cleansing and its applications in Web IR

- Our data cleansing method
  - Global scale data cleansing
  - Learn from “what users need”
  - Users’ information requirement is reflected in their search target pages (pages that they want to find)
  - A better data cleansing method should judge the quality of a Web page by whether it can be a search target for a certain user query.
  - Both hyperlink structure features and other kinds of features should be considered in data cleansing

Data cleansing and its applications in Web IR

- Query-independent Data Cleansing



Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work

Query-independent features used in data cleansing

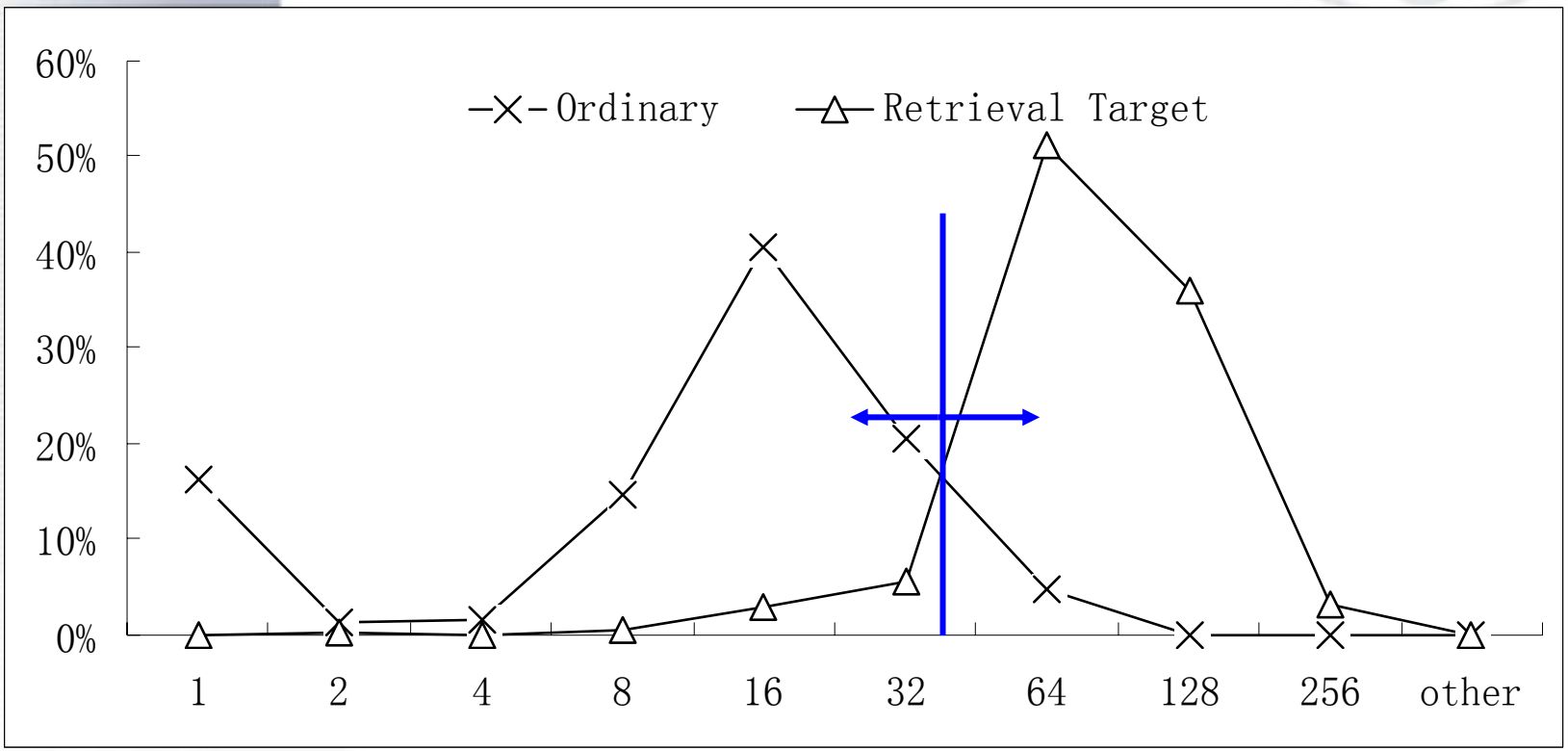
- Query-independent feature analysis of High Quality Pages
  - Corpus
    - 37M Chinese web pages collected in Nov. 2005
    - Over 0.5 Terabyte.
    - Obtained from Sogou.com
  - High Quality Page (Search Target Page)
    - Training set: 1600 pages
    - Test set: 17000 pages
    - Evaluated manually by Sogou engineers

Query-independent features used in data cleansing

- Hyperlink structure related features
  - PageRank
  - In-link number
  - In-link anchor text length
- Other features
  - Document length
  - Number of duplicates
  - URL length
  - Encode

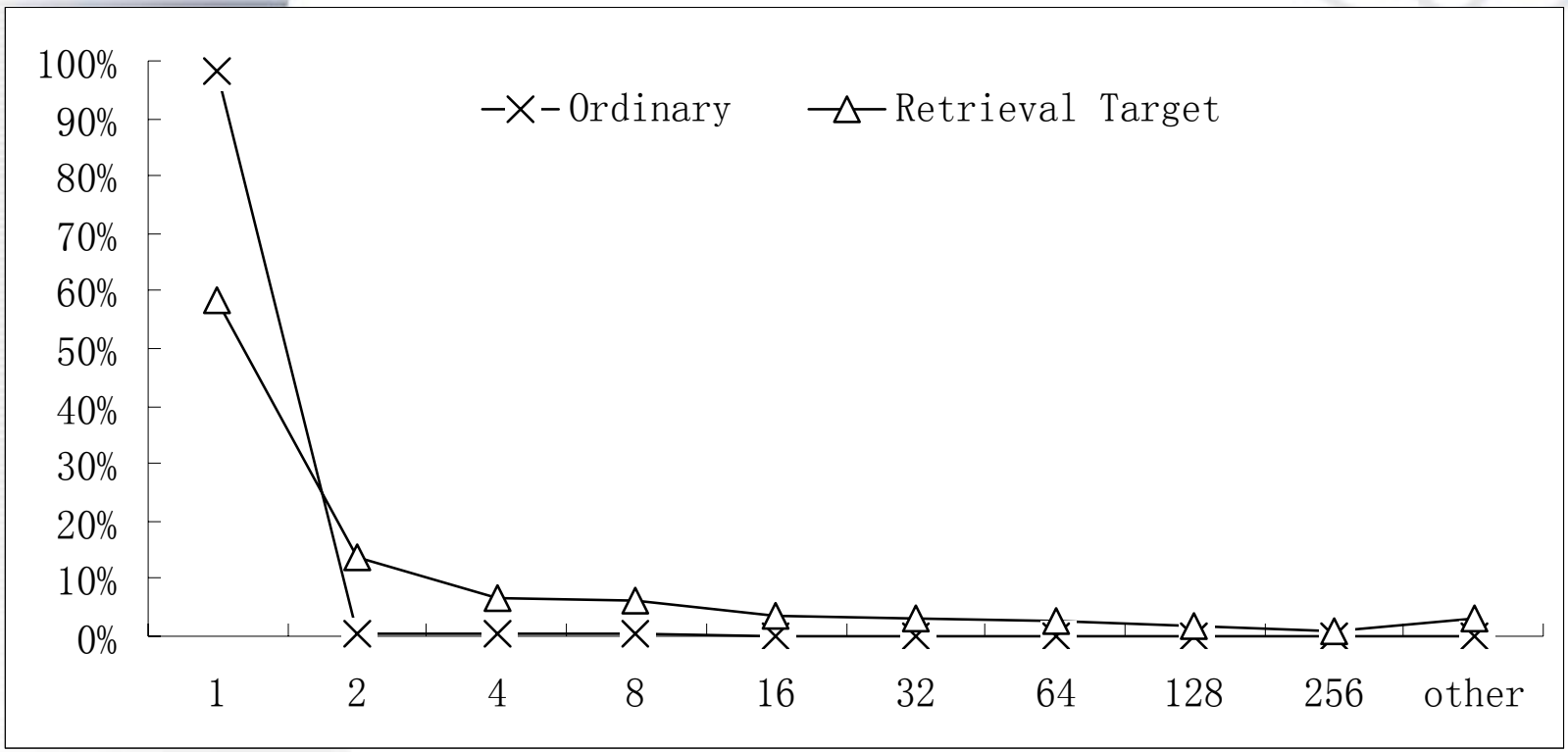
Query-independent features used in data cleansing

- PageRank



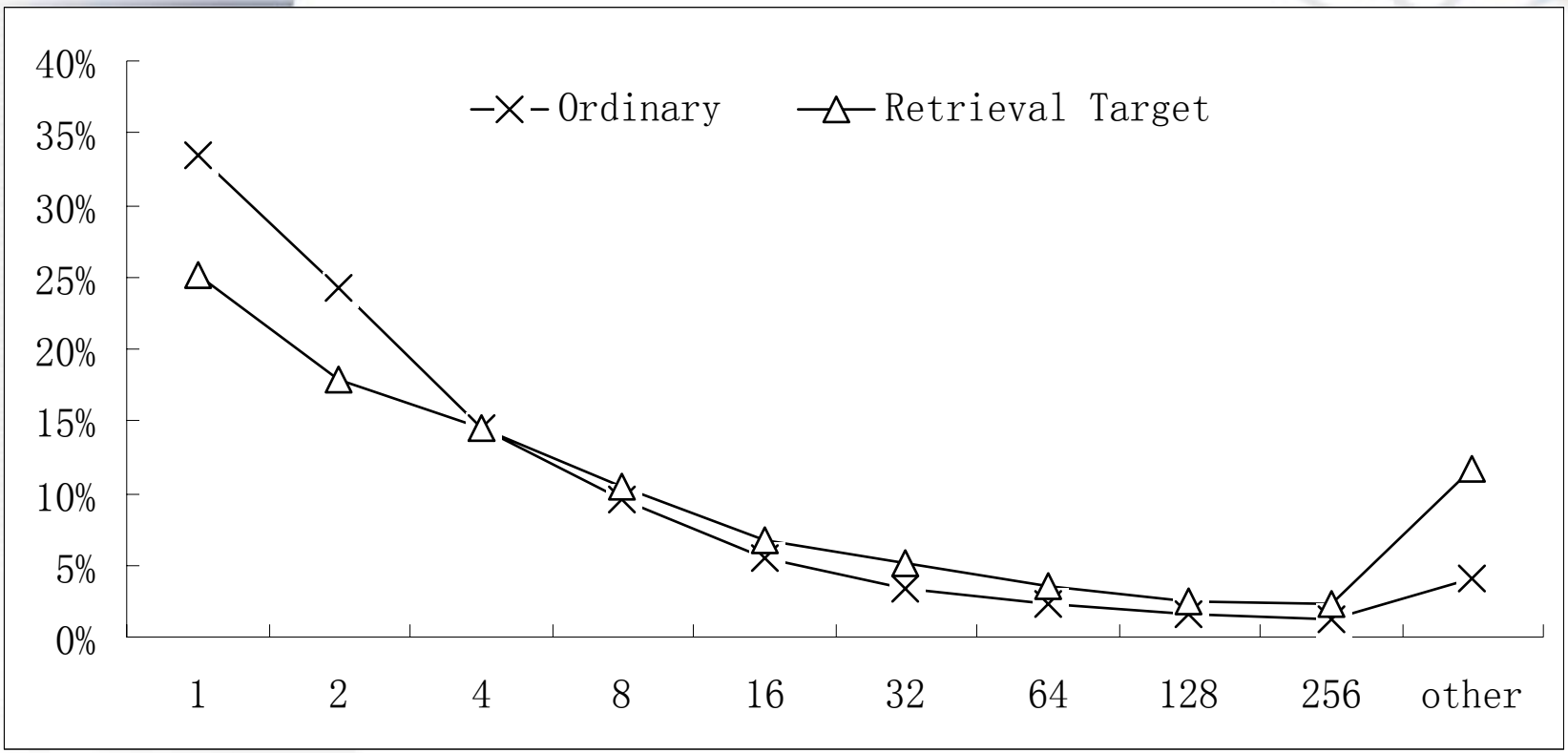
Query-independent features used in data cleansing

- In-link anchor text length



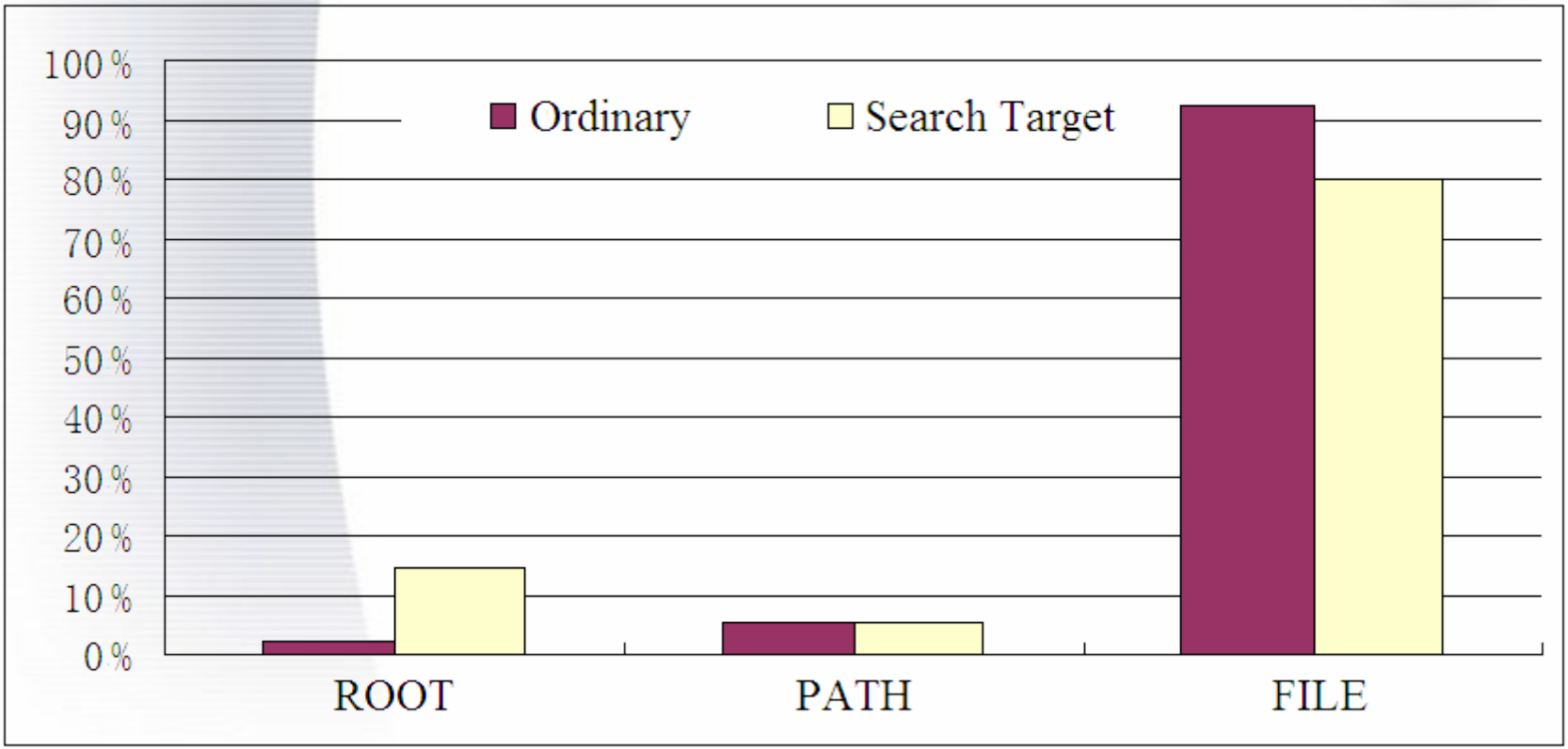
Query-independent features used in data cleansing

- Document length



Query-independent features used in data cleansing

- URL Length



Query-independent features used in data cleansing

- Other features

	Ordinary	High Quality
URL contains “?”	13.06%	1.87%
Encode is not GBK	14.04%	1.39%
Hub type page	3.78%	24.77%

- The query-independent features can separate high quality pages from ordinary pages

Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work

Algorithm and evaluation

- A learning based data cleansing algorithm
  - The possibility of one web page being a search target page is:

$$P(p \in Target\ page \mid p\ has\ feature\ A)$$

$$P(p \in Target\ page \mid p\ has\ feature\ A)$$
$$= \frac{P(p\ has\ feature\ A \mid p \in Target\ page)}{P(p\ has\ feature\ A)} \times P(p \in Target\ page)$$

$$\frac{P(p\ has\ feature\ A \mid p \in Target\ page)}{P(p\ has\ feature\ A)}$$
$$= \frac{\#(p\ has\ feature\ A \cap p \in Target\ page)}{\#(Target\ page)} \Big/ \frac{\#(p\ has\ feature\ A)}{\#(Ordinary\ page)}$$

Algorithm and evaluation

清华大学
Tsinghua University

- General information of the cleansed corpus

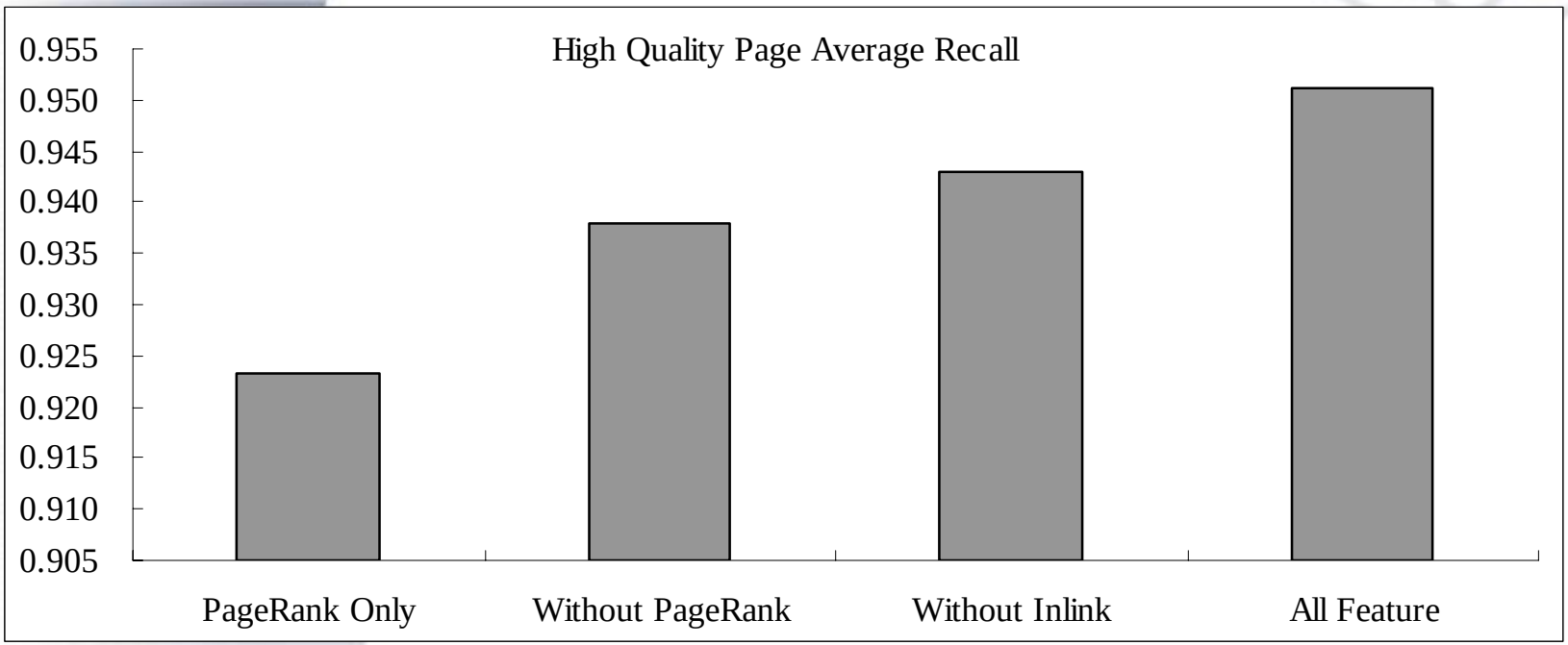
	Current Size / Original Size	High Quality Recall (Training Set)	High Quality Recall (Test Set)
Reduced Page Set	95.04%	7.27%	7.63%
Cleansed Corpus	4.96%	92.73%	92.37%

- The cleansed corpus contains about 5% pages in the original corpus, but can meet 92% user needs.

Algorithm and evaluation

清华大学
Tsinghua University

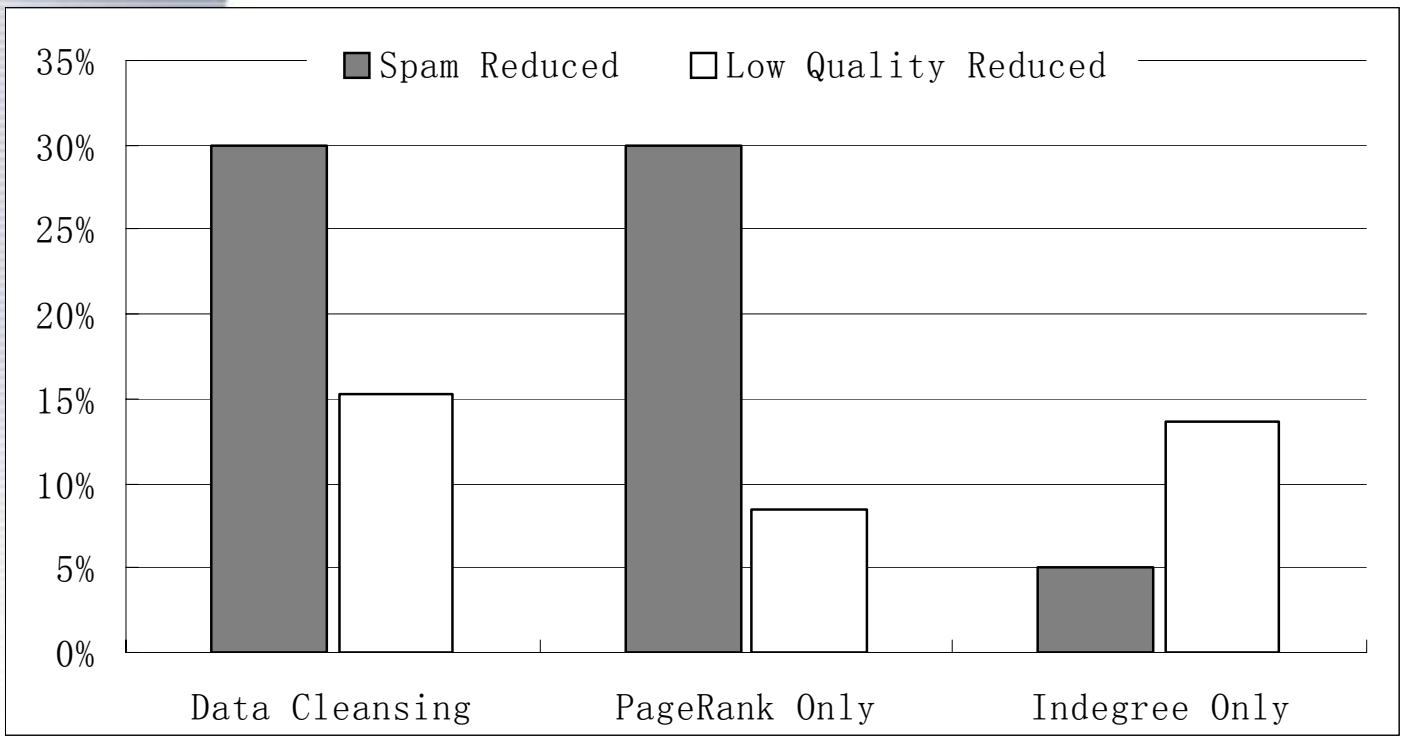
- Function of different features in our algorithm



- Although PageRank plays an important role in the algorithm, we don't rely on this single feature.

Algorithm and evaluation

- The possibility of reducing spam/low quality pages using our data cleansing algorithm



清华大学
Tsinghua University

Outlines

- Data cleansing and its applications in Web IR
- Query-independent features used in data cleansing
- Algorithm and evaluation
- Conclusions and future work

Conclusions and future work

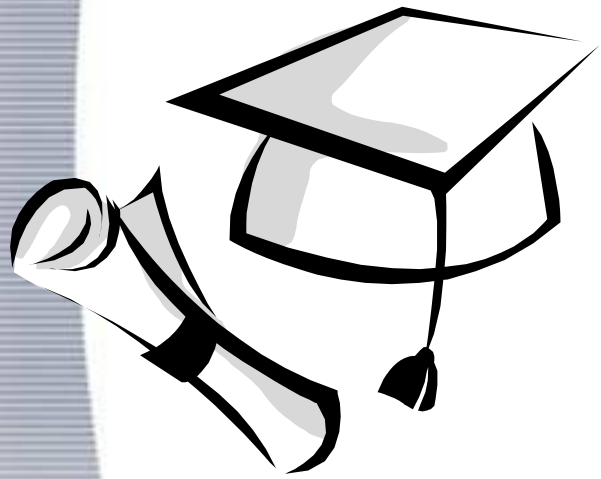
清华大学
Tsinghua University

- Conclusions:
  - Query-independent features can separate Search Target Pages from ordinary pages
  - It is possible to reduce 95% web pages with a small loss in key information
  - The data cleansing algorithm can also reduce part of spam pages / low quality pages

Conclusions and future work

清华大学
Tsinghua University

- Future work
  - Retrieval in the cleansed corpus
  - Hyper link analysis in the cleansed corpus
  - A learn-based algorithm to reduce spam pages / low quality pages
  - Personalized search



Thank you!

Questions or comments?