Web Data Cleansing for Information Retrieval using Key Resource Page Selection

Yiqun Liu, Canhui Wang, Min Zhang, Shaoping Ma

State Key Lab of Intelligent Tech. & Sys.

Tsinghua University, Beijing, China

liuyiqun03@mails.tsinghua.edu.cn

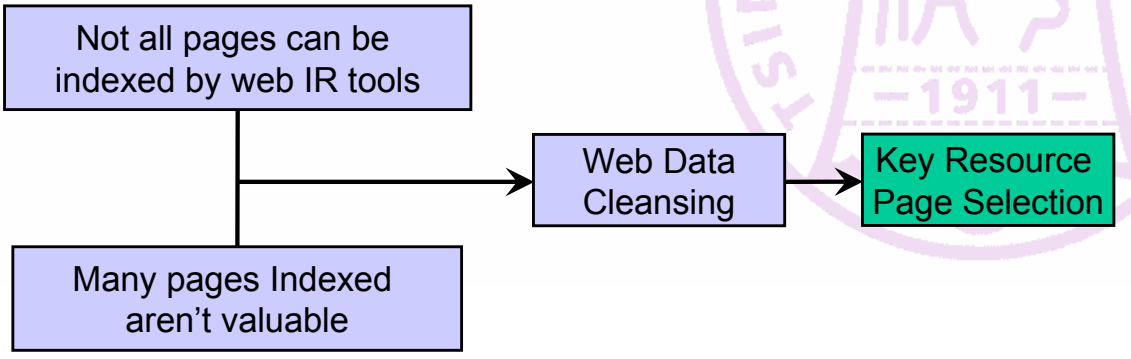
Data Cleansing is necessary for Web IR

- NO search engine can index all Web pages!

Search Engine	Reported Size	Page Depth
Google *	**8.1 billion**	101K
MSN *	5.0 billion	150K
Yahoo *	4.2 billion	500K
All the Web **	**152 billion**	605K
All the Surface Web **	**10 billion**	8K

* Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 2th, 2005.
** Danny Sullivan, Search Engine Sizes. In search engine watch website.

- Not all pages are valuable for search engines
  - **Redundancy** (30% for English Web pages, 40% for Chinese Web pages)
  - **Error** (typos, grammatical mistakes, OCR errors…)
  - **Low Quality** (Content may be old, invalid, false…)
- Data Cleansing



Key Resource Selection for Data Cleansing

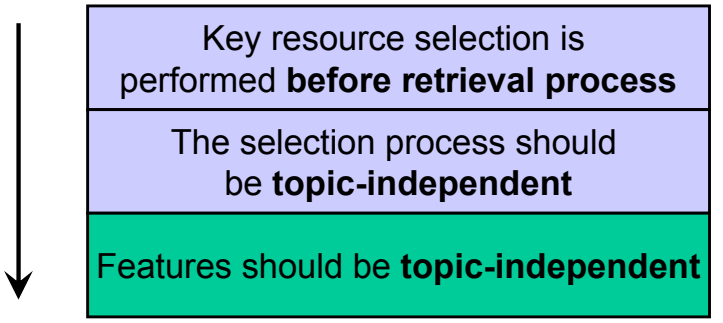
- What are key resource pages?
  - Quality Web pages that are most representative of a topic
    - Offering credible information itself
    - Providing entries to clusters of high quality pages
  - Example (topic: information retrieval)
    - Key resource:
      - Online text of the book <IR> by Dr. CJ Van Rijsbergen
      - SIGIR web site
      - Information Retrieval Links collected by a certain person
    - Not key resource
      - A conference's CFP page which is partly related to IR.
      - A product's intro page which describes IR function.
      - An ordinary (not famous) academic paper related to IR
- The differences between key resource and ordinary pages in topic-independent features

Feature	Ordinary page set	Key resource set
In-degree	9.94	153.12
URL-type	3.85	3.07
In-site out-link anchor text rate	0.06	0.12
In-site out-link number	17.58	37.70
Document Length (in words)	7037.43	9008.02

Ordinary Page Set: .GOV corpus
Key Resource Page Set: TREC 2002 web track topic distillation task answers

Difficulties in Key Resource Selection

- Difficulties in feature selection



  - Topic-independent candidates:
    - Simple features: in-degree, out-degree, URL depth, D ocument Length, Anchor text length, Page Structure…
    - Complicate features: PageRank, Hub value, Authority Value…
- Difficulties in algorithms
  - Web page classification
  - Lack of negative examples (uniform sampling is difficult and sometimes not possible)
  - Learning with unlabeled data and positive examples
  - Previous work: OSVM, PEBL: not quite suitable for learning based on topic-independent features.

A Key Resource Selection Algorithm based on K-means

- Why is k-means used here?
  - Learn without negative examples
  - Independent of prior positive proportion knowledge
- Differences with traditional K-means
  - Fixed cluster number: true or not.
  - Initial positive example centroid is provided
- **Algorithm**

S_{key} : key resource training set

R : estimated proportion of the positive examples

1. Choose 2 initial cluster centroids:
   - Positive centroid: $M_1 = \frac{1}{S_{key}} \sum_{X \in S_{key}} X$
   - Negative centroid: $M_2 = \frac{M(Whole\ Collection) - R \times M_1}{1-R}$
2. In the kth iterative, instance X will be assigned to the jth cluster $S_j^{(k)}$ if:

$$||X - M_j^{(k)}|| = min(||X - M_1^{(k)}||, ||X - M_2^{(k)}||) \qquad (j = 1, 2)$$

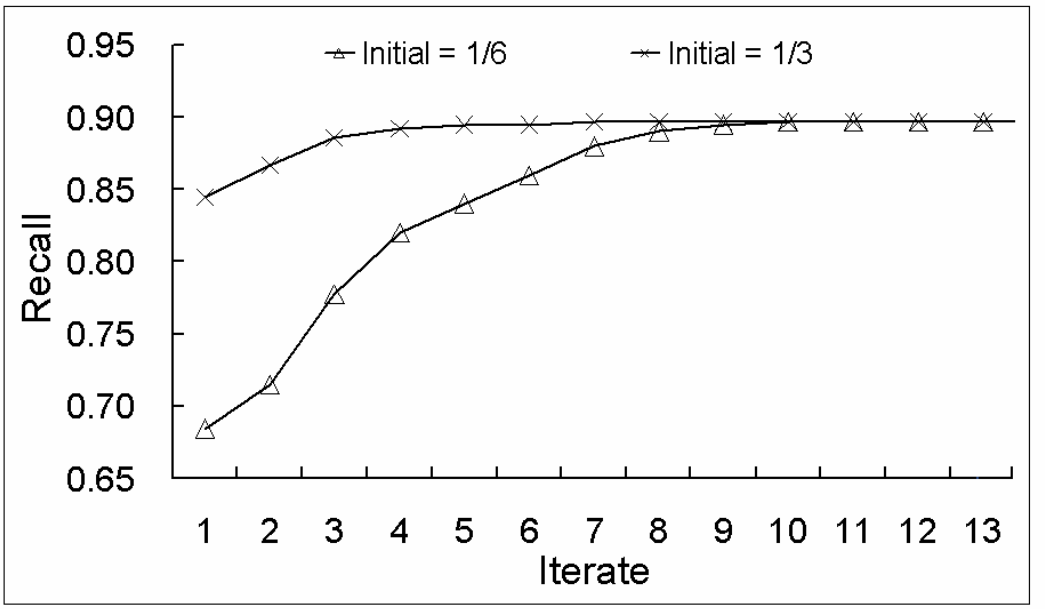
3. For $S_j^{(k)}$, caculate $M_j^{(k)}$, which is defined as:

$$M_j^{(k+1)} = \frac{1}{N_j} \sum_{X \in S_j^{(k)}} X \qquad (j = 1, 2)$$

4. If $M_1^{(k+1)} = M_1^{(k)}$, exit. Else go to 2.

A Key Resource Selection Algorithm based on K-means

- Algorithm converges with different initial *R*
  - Algorithm doesn't require prior knowledge of *R*



- Key resource result set
  - Algorithm can cover almost all key resource pages with less than half whole collection size

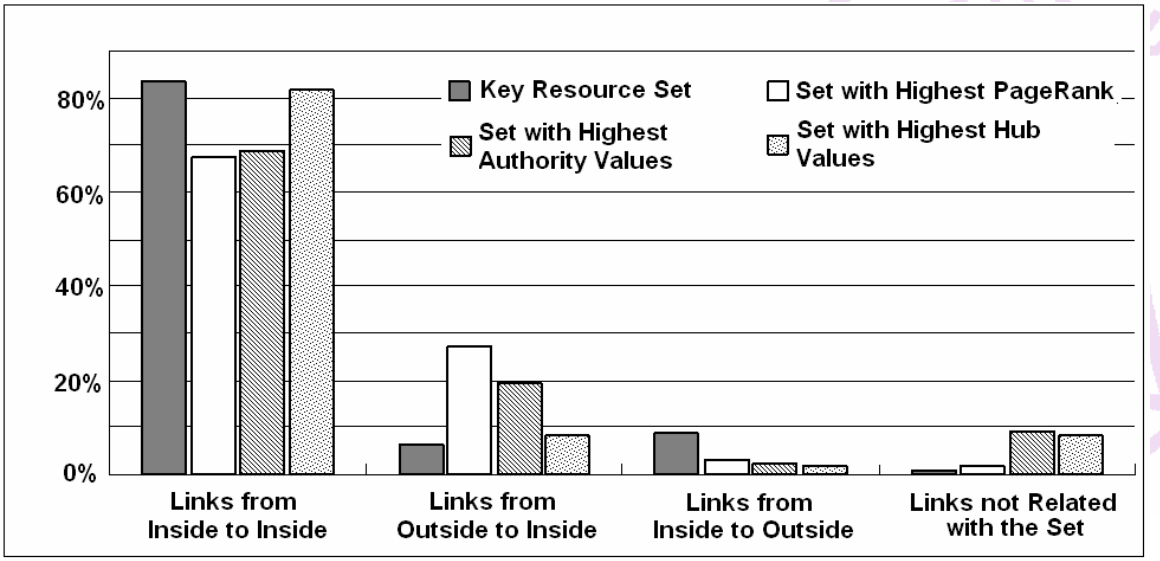
	K-means Clustering
Whole Collection (.GOV) Coverage	44.30%
Key Resource Test Set Recall	89.70%
Key Resource Test Set Precision	67.50%
F2-measure	53.89%

Experiments
(Key resource set properties)

- Key resource pages get high scores in link anal ysis algorithms
  - A large part of top-ranked pages with link analysis metrics are key resources

Link Analysis Criteria	Percentage of key resources in top 10% pages
PageRank	88.29%
HITS hub value	73.81%
HITS authority value	76.12%

- Key resource page set retains almost all hyperlink structure information



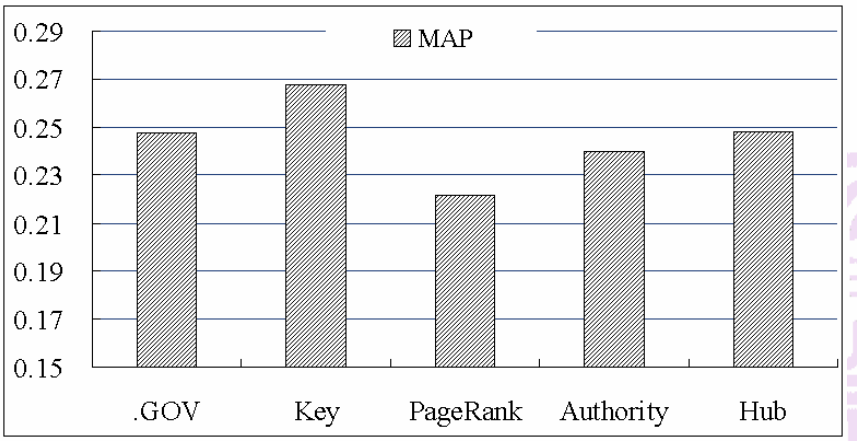
Experiments
(Retrieval performance)

- Experiment settings

Type of query	Query log analysis*	Test query set**
Navigational	20%	20%
Transactional	30%	80%
Informational	50%	

* According to Andrei Broder's AltaVista log analysis in SIGIR forum 2003
**Selected from TREC 2004 topics and qrels

- Retrieval Results



	P@10 for Topic Distillation queries	MRR for Navigational query
Whole Collection	0.1025	0.7443
K-means	0.1275	0.7278
PageRank	0.1134	0.6533
Authority	0.1100	0.6700
Hub	0.1250	0.6357

Conclusions

- Conclusions
  - Data cleansing based on key resource selection is effective in reducing unimportant pages.
  - Result set (half size of total collection) retains useful information as well as hyperlink structure of the Web collection.
  - Retrieval on result set gets better overall retrieval performance than the whole collection.
- Future work
  - Can we find better features/algorithms to improve selection precision?
  - Whether other useful pages besides key resources can be selected topic-independently?
  - How well does this method work for a page set with billions of pages? How much effort will it take to fini sh cleansing?
  - Is it possible to find a best trade-off between this relatively modest disk savings and the potential loss of effectiveness?