Effective Topic Distillation with Key Resource Pre-selection

Yiqun Liu, Min Zhang and Shaoping Ma

State Key Lab of Intelligent Tech. & Sys.
Tsinghua University, Beijing, 100084

liuyiqun03@mails.tsinghua.edu.cn

(2004/10/19)

Outline

- Why Key Resource Pre-selection?
- Possibilities of selecting key resources
- How to select key resources?
- Experiments
- Conclusion

Why Key resource selection? (1)

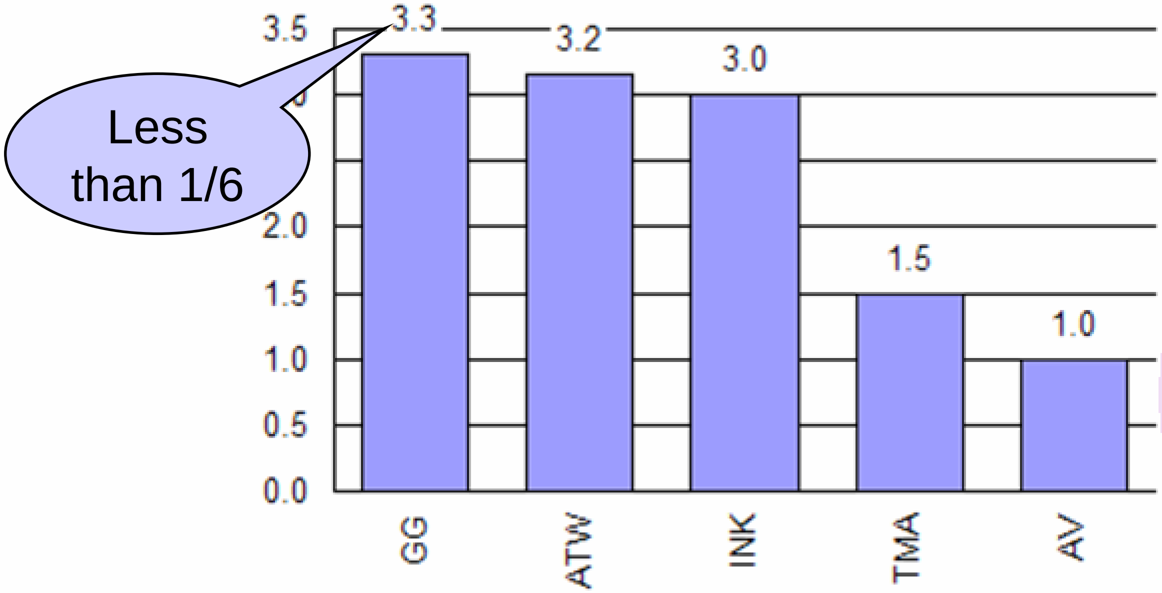
- **The amount of web pages**

Medium	2002 Internet
Surface Web	167 TB
Deep Web	91,850 TB
#Surface web pages	20 billion
#Deep web pages	130 billion

According to "How Much Information", 2003.
http://www.sims.berkeley.edu/how-much-info-2003.

Why Key resource selection? (2)

- **Index amount of web search engine**



Billions Of Textual Documents Indexed

GG=Google,

ATW=AllTheWeb,

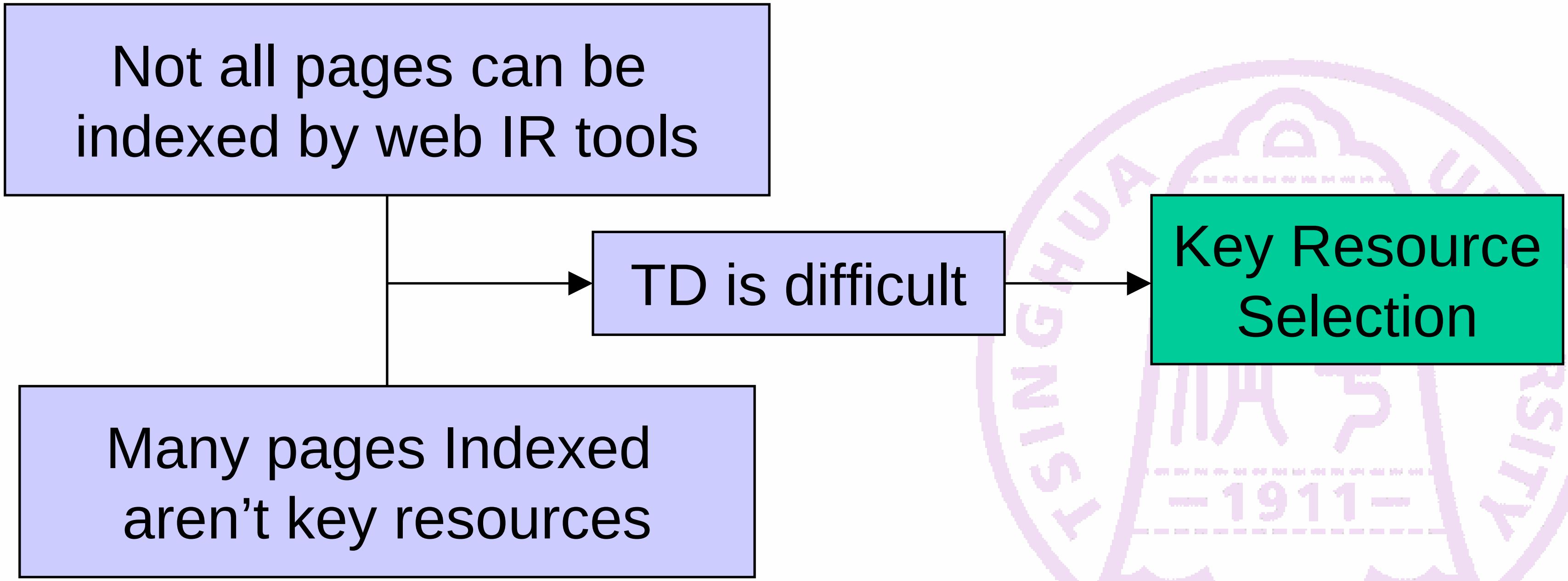
INK=Inktomi,

TMA=Teoma,

AV=AltaVista

According to a report by search engine watch website; September 2, 2003

Why Key resource selection? (3)


Not all pages can be indexed by web IR tools
Many pages Indexed aren't key resources
TD is difficult
Key Resource Selection

Definitions of TD and key resource

- **Key Resource (Key Resource Page)**
  - High-quality web pages for a particular topic
    - Offering credible information/service for this topic
    - Introducing other useful web pages for this topic
  - Key resources are only a small part of relevant pages
- **Topic Distillation (TD)**
  - To find key resources for certain topics
  - A major task for web search (it covers over 70% web search queries)

Outline

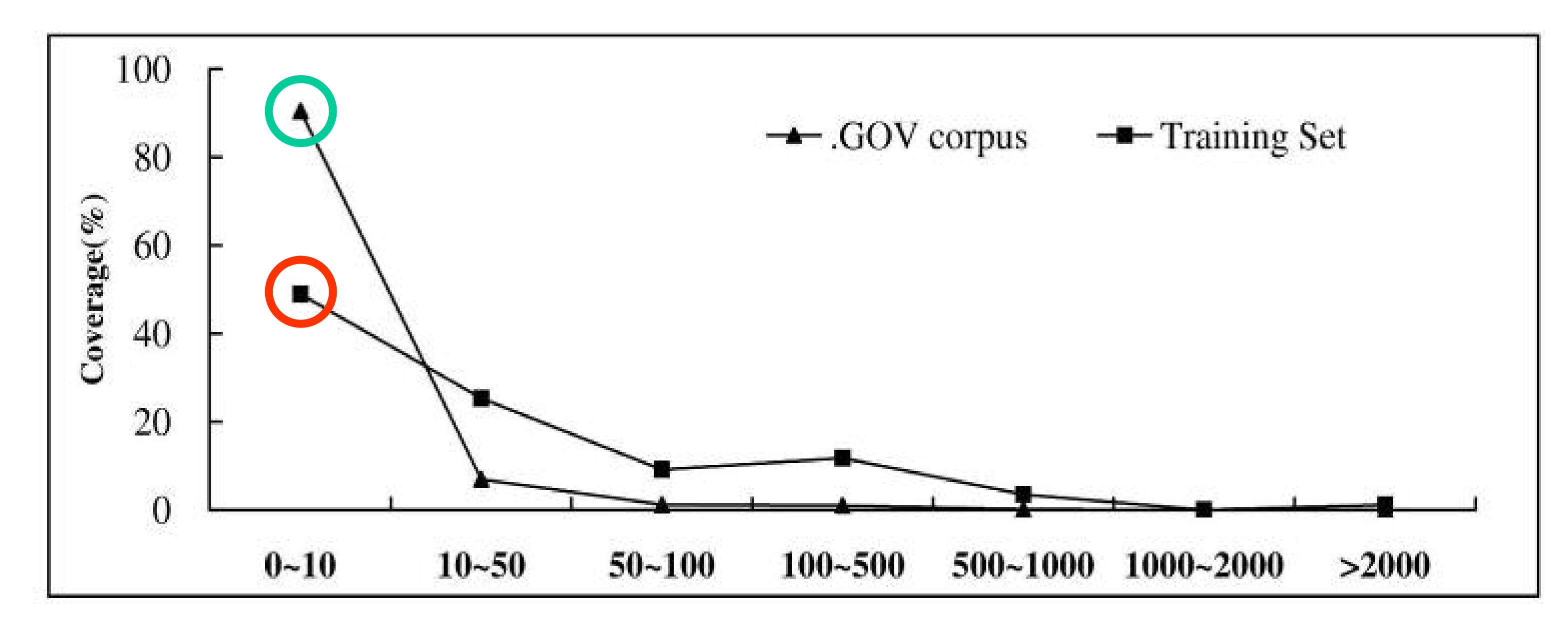
- **Selecting key resources is useful for TD**
- **Possibilities of selecting key resources**
  - Is there any difference between ordinary pages and key r esource pages?
- **How to select key resources?**
- **Experiments**
- **Conclusion**

Non-content features of key resources

- **Key resources v.s. ordinary pages (non-content features)**
  - Common-used features
    - In-degree, URL-type, Doc-length
  - Features involving site's self-link analysis
    - In-site out-link number, anchor text rate
- **Two Data sets to compare the differences**
  - Key resource page training set
    - Built with TREC 11 TD task relevant qrels
  - Ordinary page set: .GOV (over 1.2M web pages from .GOV domain)

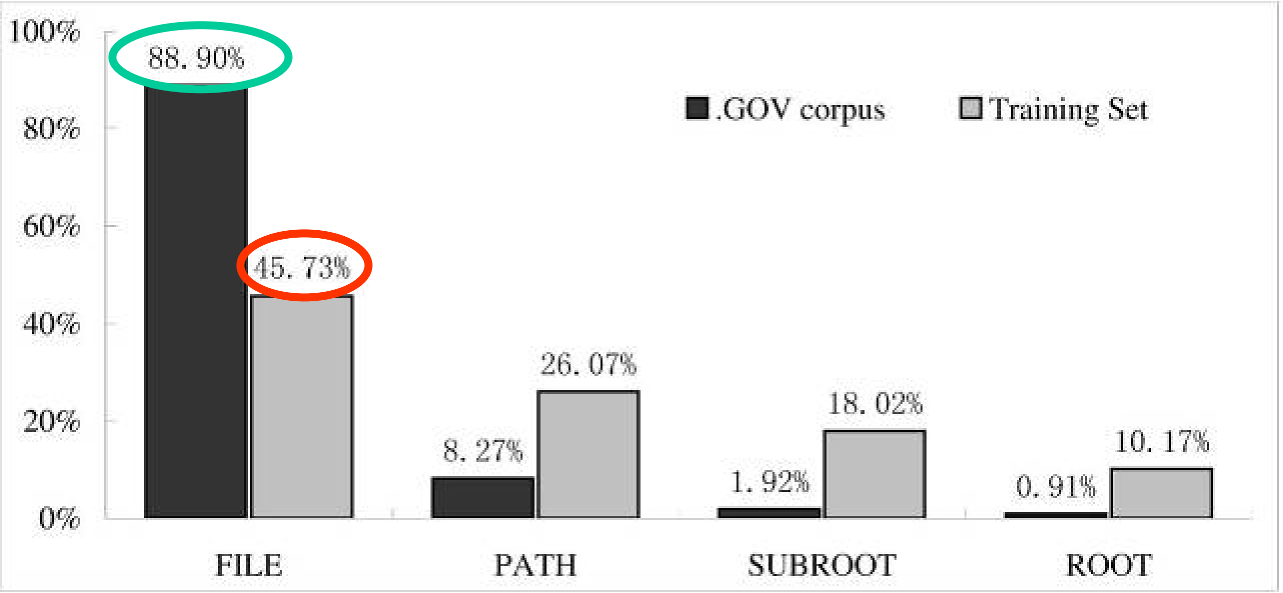
In-degree

- **Key resource pages have more in-links**



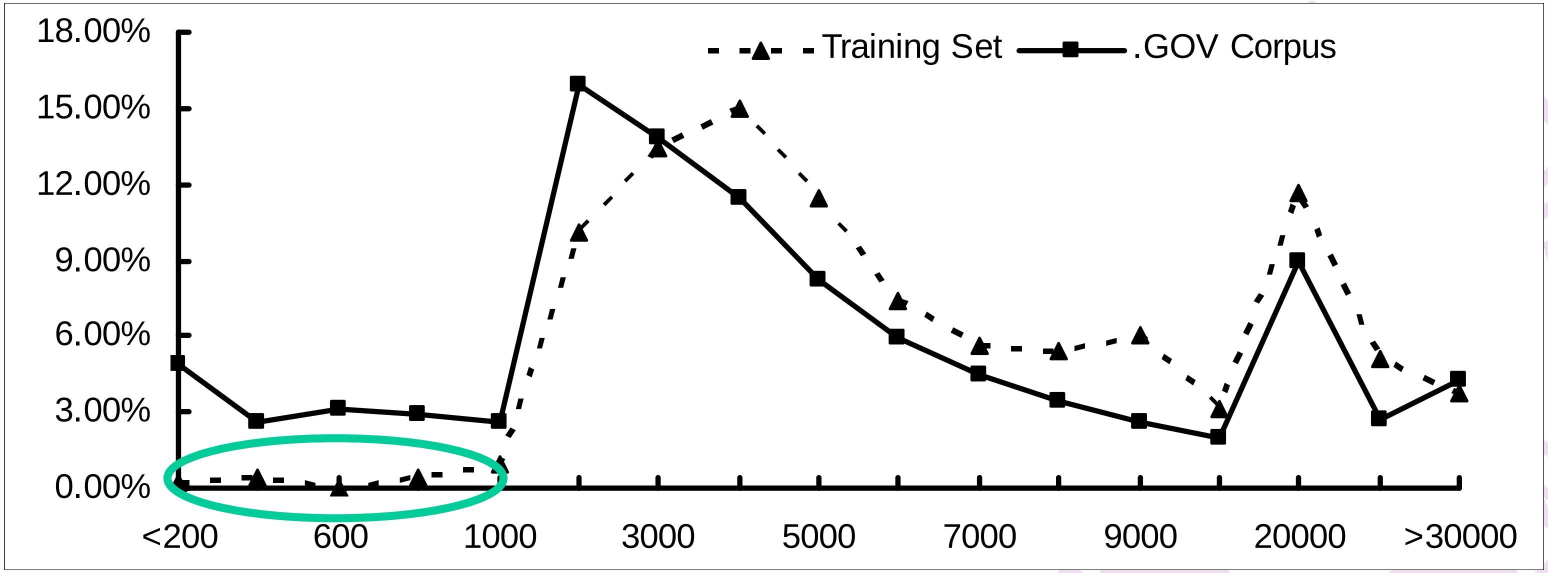
URL-type

- **Key resource pages tend to be non-FILE type**



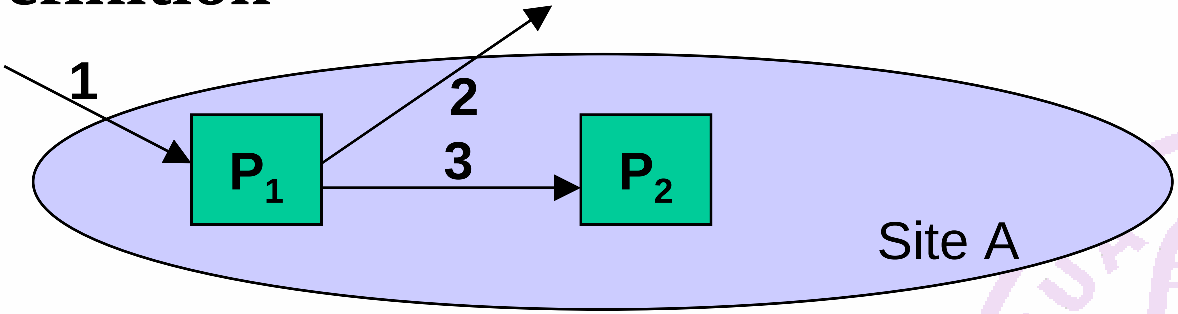
Doc-length

- **Key resources don’t have too few words**



In-site Out-link analysis

- **Definition**

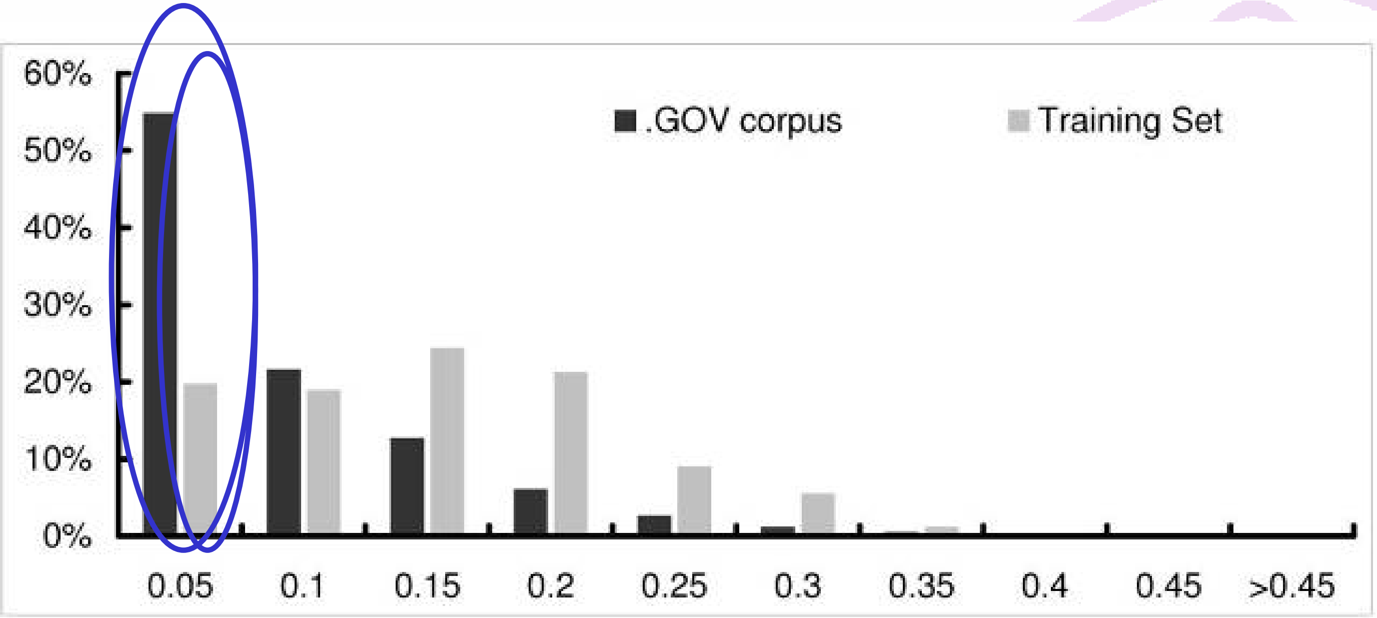


- **Feature**
  - In-site out-link number
  - In-site out-link anchor text rate

$$rate = \frac{WordCount\ (in-site\ out-link\ anchor\ )}{WordCount\ (web\ page\ full\ text\ )}$$

In-site Out-link analysis

- Key resource pages have more in-site out-links and longer in-site out-link anchor texts



In-site out-link anchor text rate

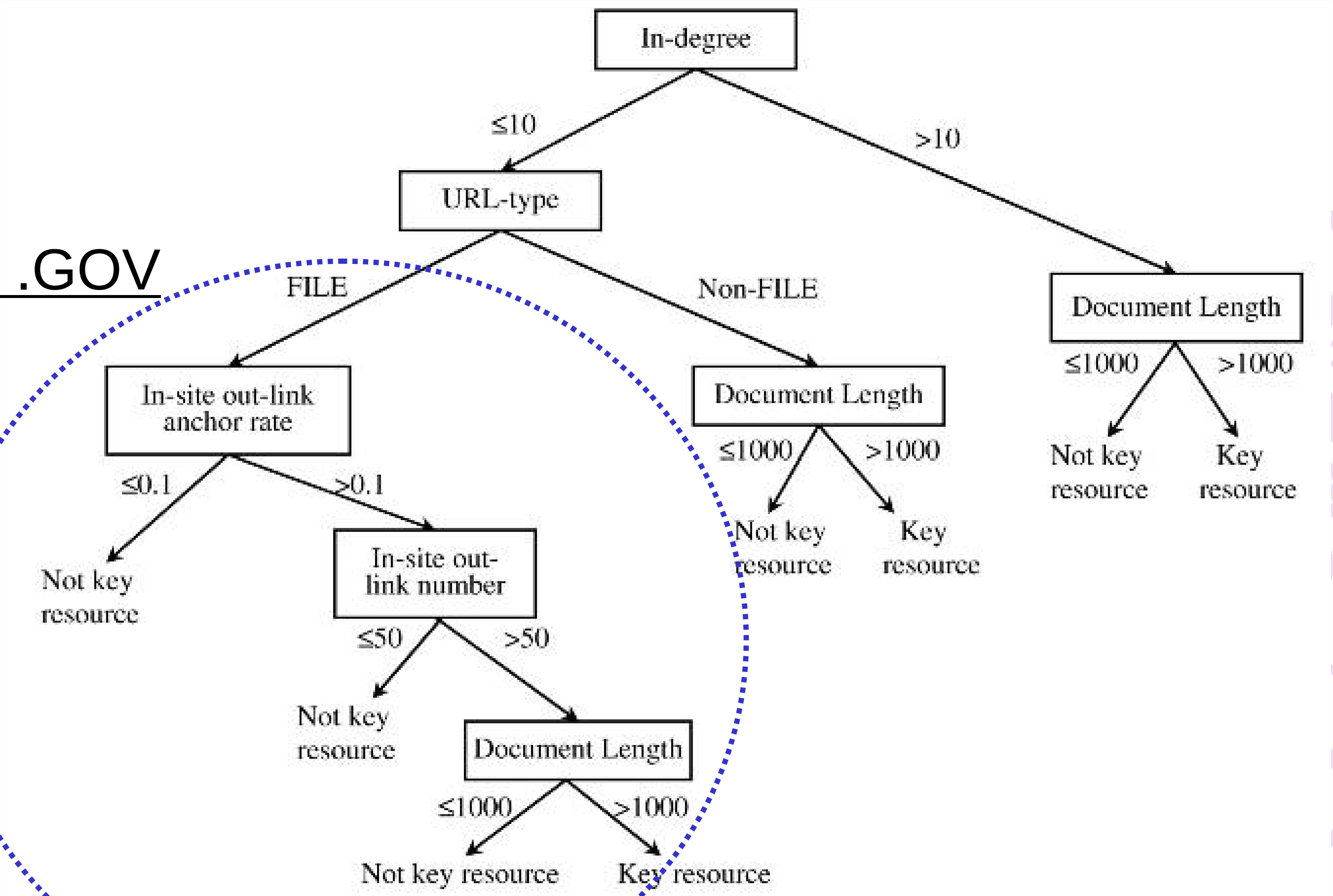
Outline

- **Selecting key resources is useful for TD**
- **Possibilities of selecting key resources**
- **How to select key resources?**
  - Construction of a key resource decision tree
- **Experiments**
- **Conclusion**

Construction of a key resource decision tree

- **Why decision tree?**
  - – The most effective and efficient classifier when there are small number of features
    - 5 non-content features
  - – Providing a metric to estimate quality of these features in the fo rm of
    - Information gain (ID3)
    - Information ratio (C4.5)

Construction of a key resource decision tree



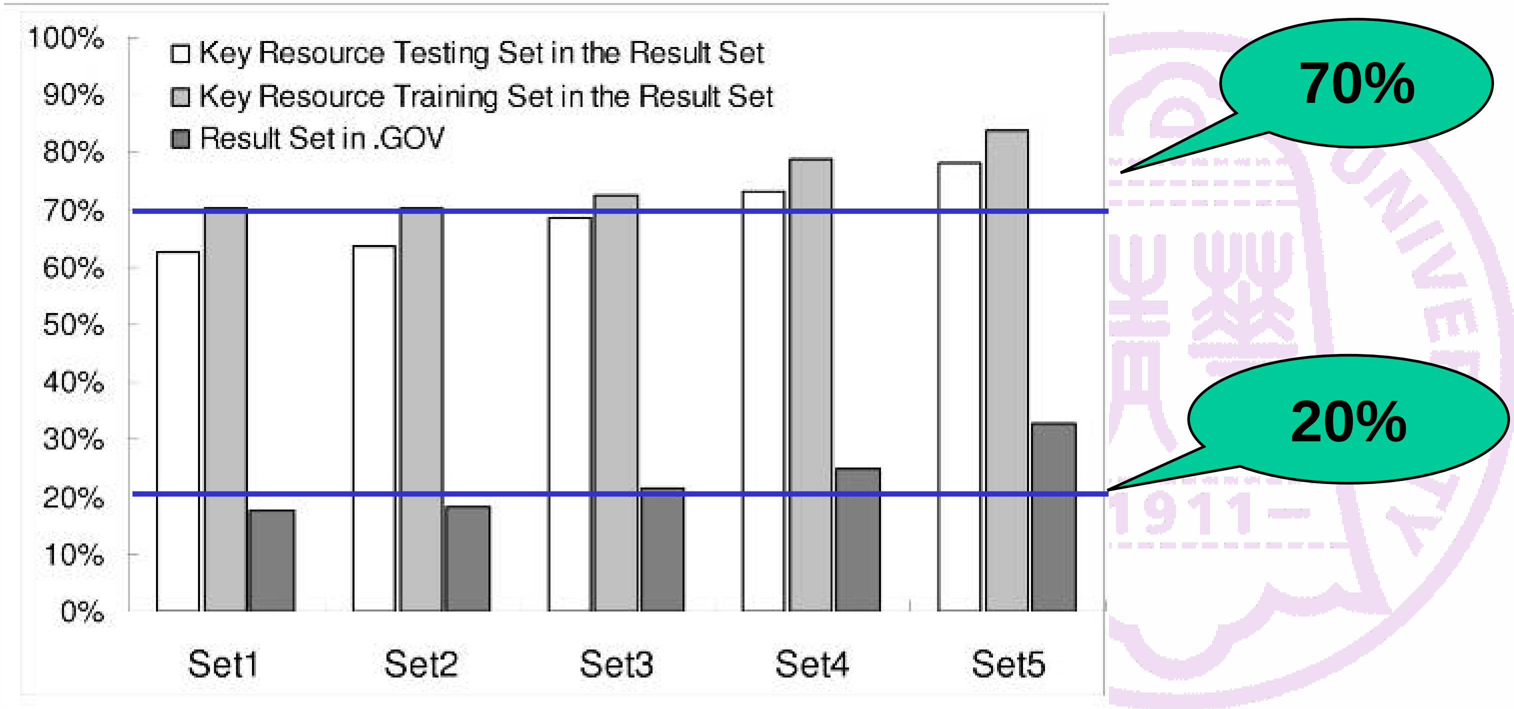
68.53% of .GOV

Outline

- **Selecting key resources is useful for TD**
- **Possibilities of selecting key resources**
- **How to select key resources?**
- **Experiments**
  - Is this key resource selection process effective?
  - Does TD perform better on the key resource result set?
- **conclusion**

Is this key resource selection process e ffective?

- **Key resource selection algorithm is effective**

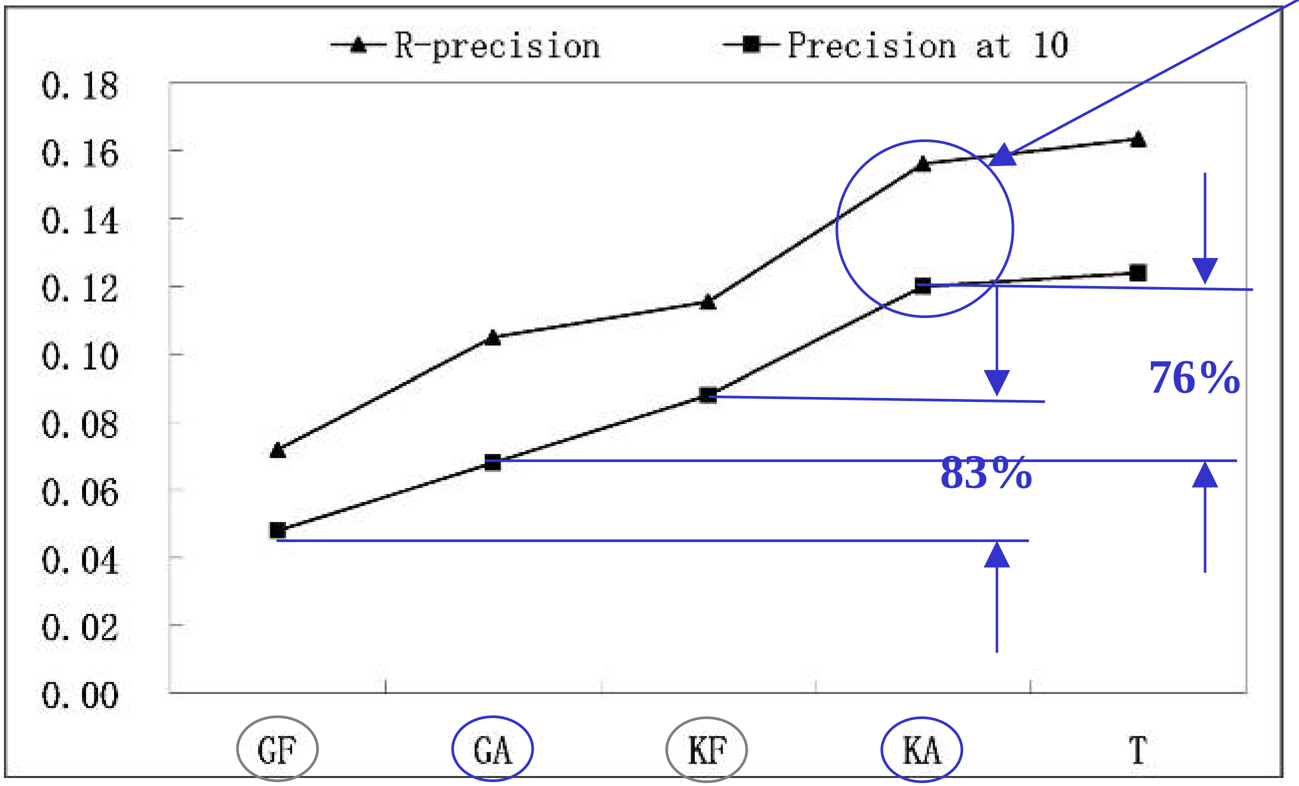


Does TD perform better on the key res ource result set?

- **Test set:**
  - From TREC 2003 TD task
  - 50 topics and corresponding relevant qrels
- **Evaluation Metrics:**
  - Precision at 10 documents
  - R-precision (precision at #relevant documents)
- **Weighting**
  - BM2500 ranking, default parameters

Does TD perform better on the key resource result set?

- Text retrieval on different data set



24.89% .GOV data

G = .GOV corpus

K = Key resource set

F = Full text

A = Anchor text

T = Trec 2003 best run

Conclusion

- **Key resource pre-selection is needed for TD**
  - Finding high quality pages independent of a given user request
- **A new type of non-content features**
  - In-site out-link analyses
- **Algorithm of using decision tree to find key resources**
- **Key resource page set:**
  - use less than 20% .GOV pages
  - cover more than 70% key resource information
  - get better performance than whole page set

    (There is 76% performance improvement in p@10)

Thank you!

Questions and comments?

Welcome to contact me:

liuyiqun03@mails.tsinghua.edu.cn