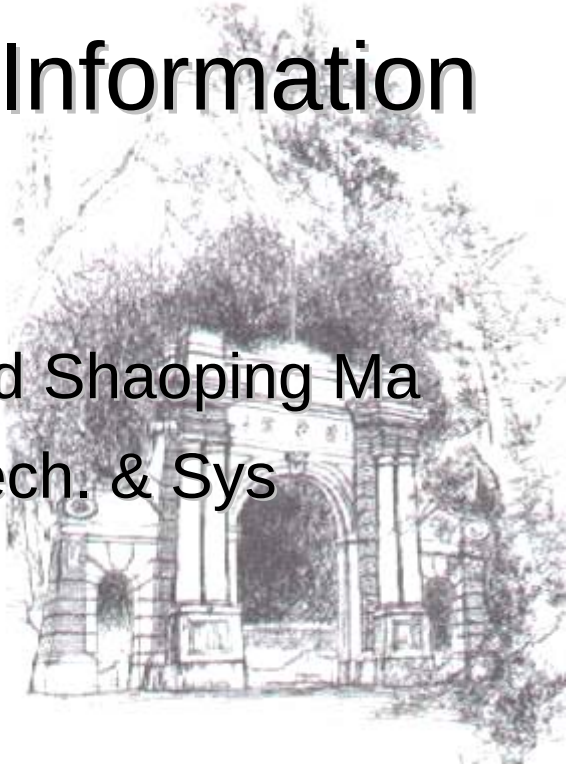
Automatic Query Type Identification Based on Click Through Information

Yiqun Liu, Min Zhang, Liyun Ru and Shaoping Ma

State Key Lab of Intelligent Tech. & Sys

Tsinghua University



Automatic Query Type Identification

- Research Background
- User analysis for query type identification
- A Query Type Identification Algorithm
- Experiments Results and Discussions

Automatic Query Type Identification

- Research Background
- User analysis for query type identification
- A Query Type Identification Algorithm
- Experiments Results and Discussions

Research Background

清华大学
Tsinghua University

- Observer user from Search Engine's prospect
  - Query stream & click through information
  - Query stream
    - Made up of queries which contain 3-4 words in English or less than 2 words in Chinese
    - Always confusing
    - Same query, different user request
    - Click through information helps us to identify users' information needs

Research Background

清华大学
Tsinghua University

- Example: 魔獸爭霸（War Craft）
  - User type 1: Users want to visit a particular web site related to the game
  - User type 2: Users want to download the corresponding computer game
  - User type 3: Users want to get a overview of the corresponding computer game
  - We cannot identify the users' information needs without the help of click through information

Research Background

清华大学
Tsinghua University

- Categories of Users' information needs
  - Proposed by Broder(IBM, 2002) & Rose(Yahoo! 2004) respectively with search engine user behavior analysis
  - Navigational
    - A specific search target page
    - Users want to know a certain web page's URL
    - "Yahoo HK", "SIGIR 04 home"
  - Informational / Transactional
    - No specific search target page
    - Users want to know something about a certain topic
    - "bird flu", "American civil war"

Research Background

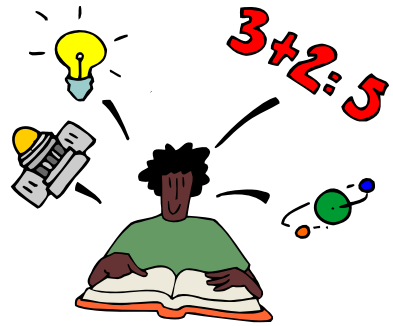
清华大学
Tsinghua University

- Why should we identify users' query types?
  - Different ranking models
    - Navigational type search: anchor text, URL information…
    - Informational type search: hyper link analysis, traditional IR models

  - Different performance
    - Navigational type search: MRR > 80%, systems can return the correct answer at 1st ranking for most queries
    - Informational type search: P@10 < 30%, systems can only return less than 3 correct answers in the top 10 results.

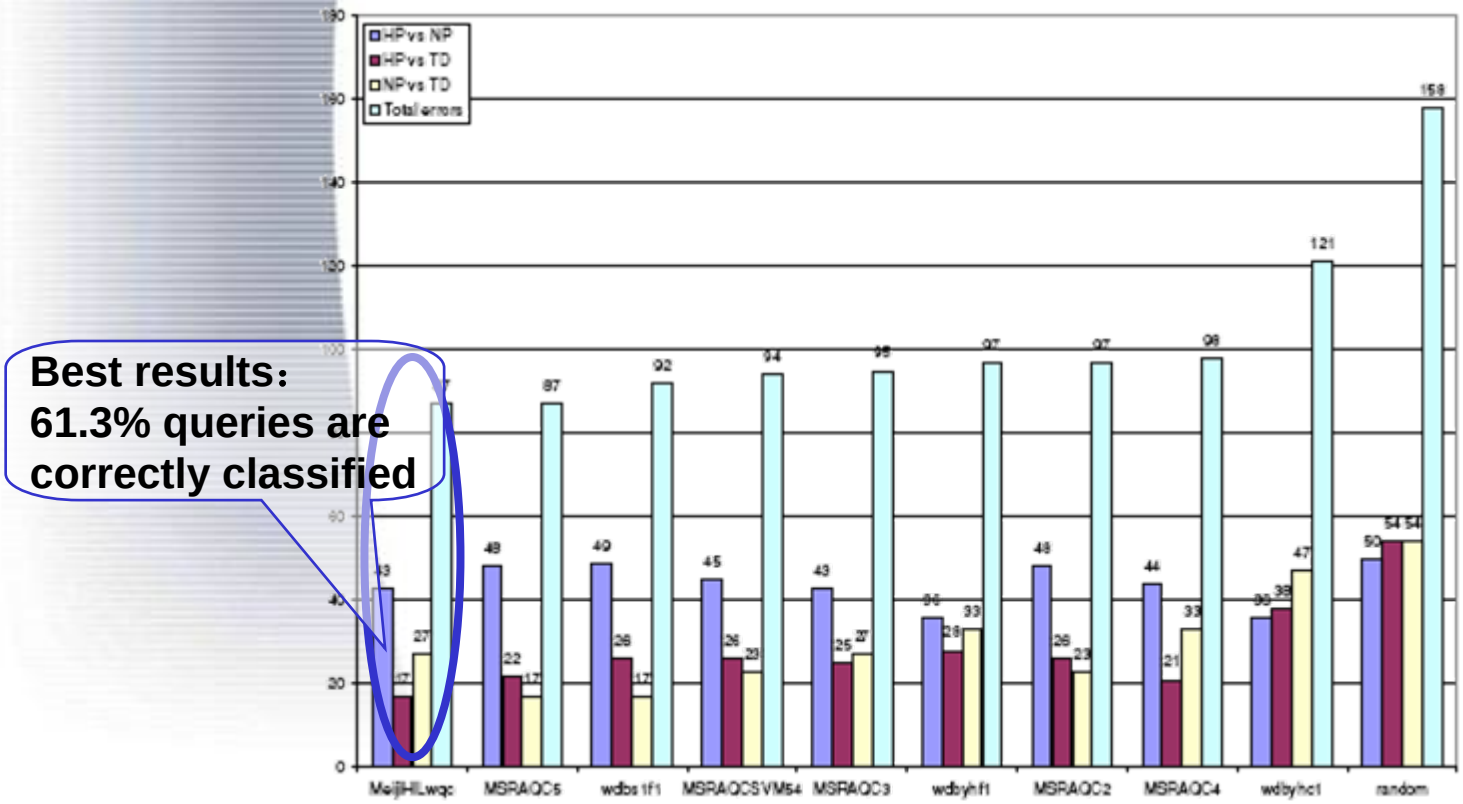
Research Background

- Features used in query type identification
  - Query content feature
    - Length, POS information, existence of Abbreviation, etc.
    - Whether the query is the anchor text for a particular page
  - Result feedback of IR system
    - The similarity between query and top-ranked documents
  - Past click-through information
    - Past click behavior



Research Background

- Related works
  - TREC2004: Query content and result feedback



Research Background

清华大学
Tsinghua University

- Related works
  - Kang et al
    - Mutual Information, POS and anchor text evidence
    - TREC data
    - Got better retrieval performance with his classification algorithm
  - Lee et al
    - Anchor text and click through information
    - UCLA campus search service data
    - 90% queries are correctly classified

Research Background

清华大学
Tsinghua University

- Major problems
  - Lack of practical search engine user analysis
    - TREC or small scale campus users' behavior are significantly different from ordinary web users
  - Lack of examination of reliability
    - Small number of special designed queries
    - How many percentages of practical queries can be classified?

Automatic Query Type Identification

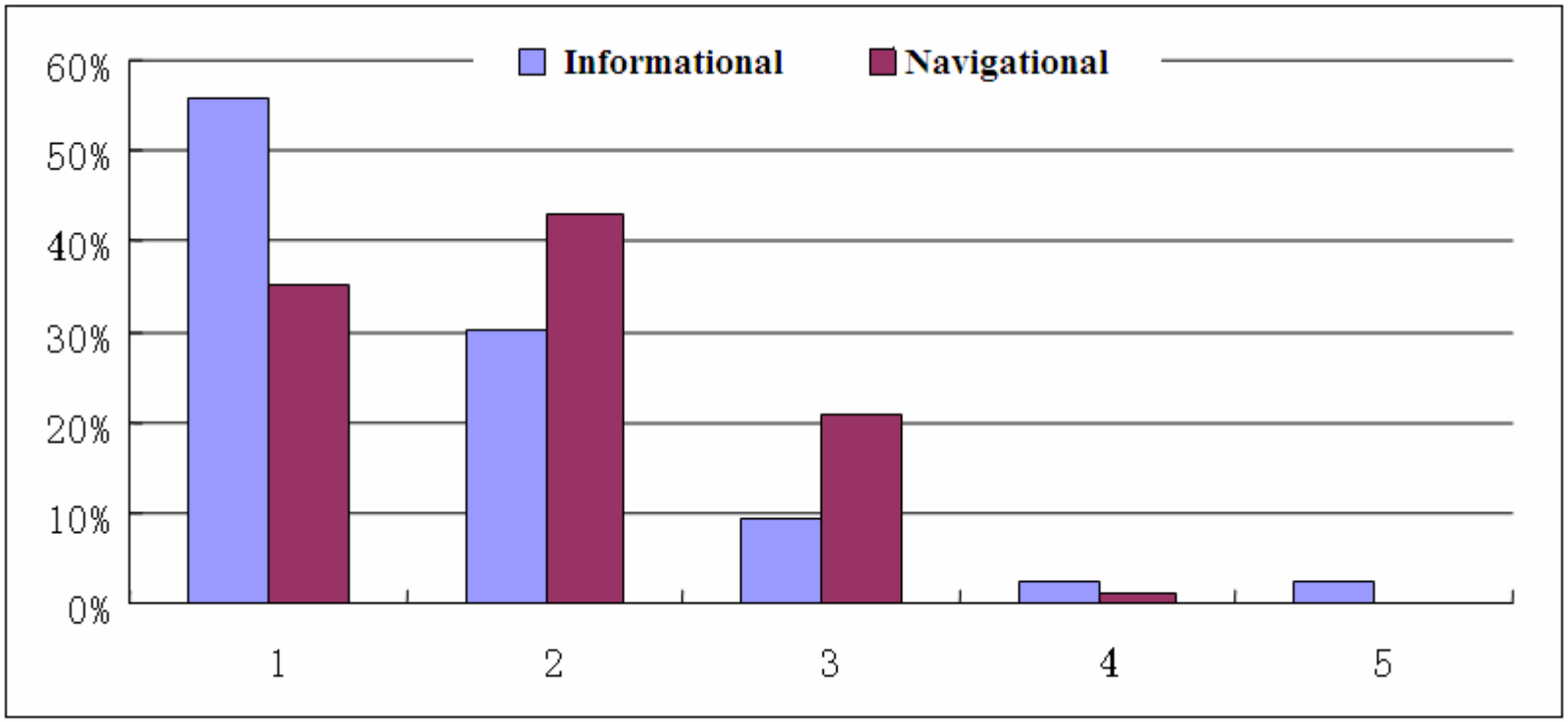
- Research Background
- User analysis for query type identification
- Query Type Identification Algorithm
- Experiments Results and Discussions

User analysis for query type identification

- Review of proposed features in query type identification
  - Practical query logs obtained from Sogou.com
    - All user queries and corresponding click through data in February 2006
    - 86538613 clicks
    - 26255952 user sessions
    - 4345557 unique user queries
  - About 200 queries are annotated by 3 assessors using voting method for training

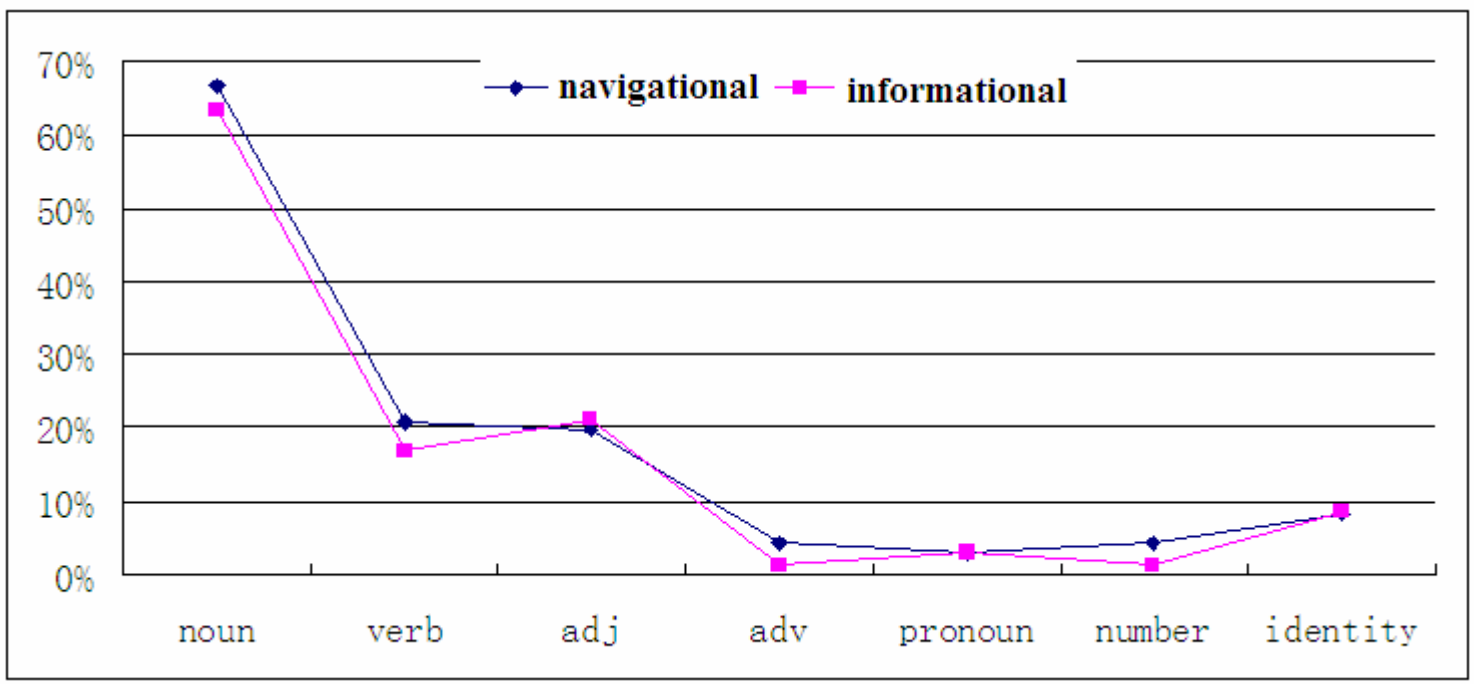
User analysis for query type identification

- Query Length
  - Distribution of query length for different query types



User analysis for query type identification

- Part of speech tagging
  - POS feature of different types of queries



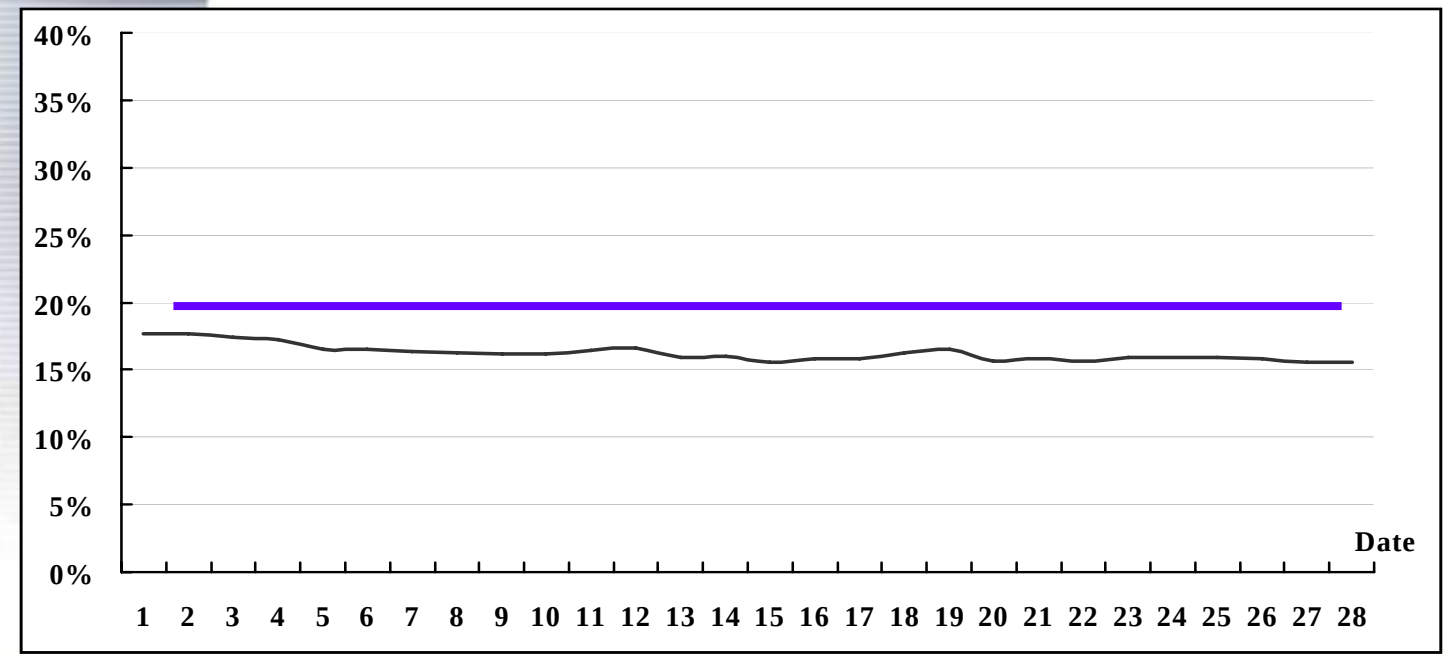
User analysis for query type identification

- In-link anchor information
  - Assumption:

    If one query *Q* shares the same content as a anchor text linking to a page *A*, *Q* is likely to be a navigational type query whose target page is *A.*
  - *A* has a lot of anchors whose content is *Q* -> *Q* is a navigational type query
  - Adopted by Kang (2004) and Lee (2005)

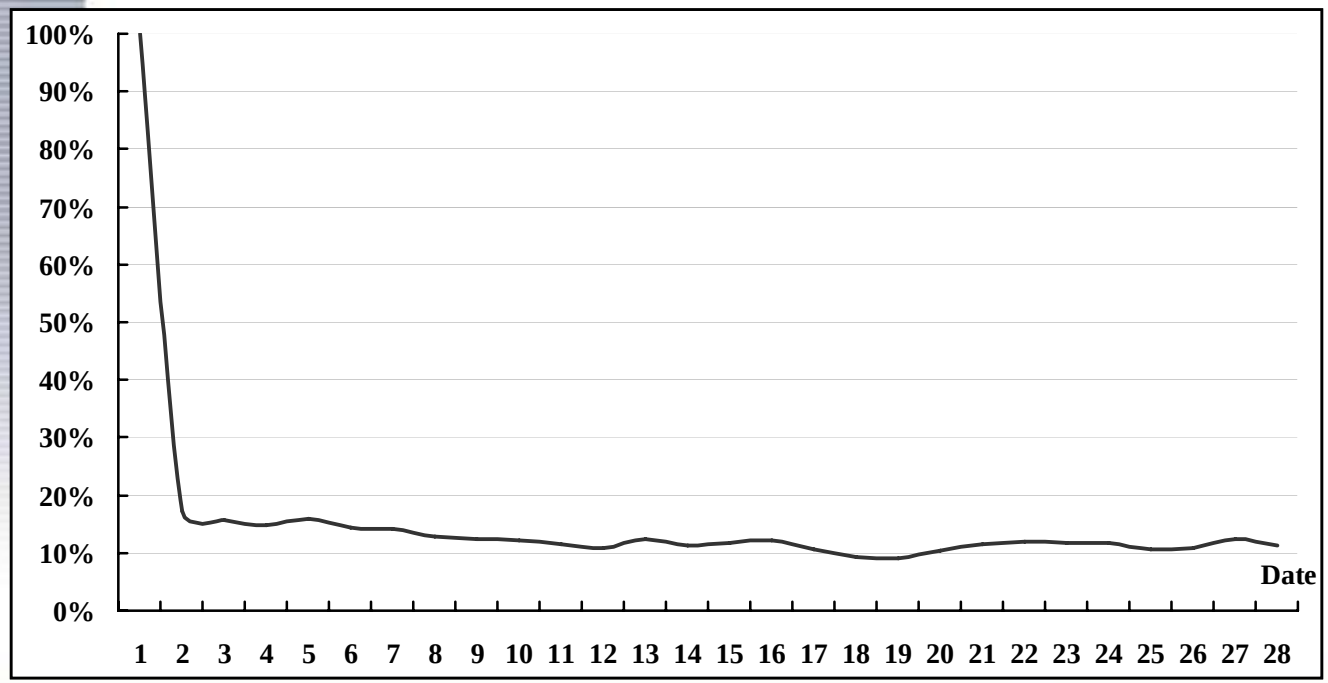
User analysis for query type identification

- How many queries can be identified using anchor text information?
  - Not all queries have a page which shares a same anchor



User analysis for query type identification

- How many queries can be identified using past click through information?
  - About 90% queries have been proposed and clicked every day.



Automatic Query Type Identification

- Research Background
- User analysis for query type identification
- Query Type Identification Algorithm
- Experiments Results and Discussions

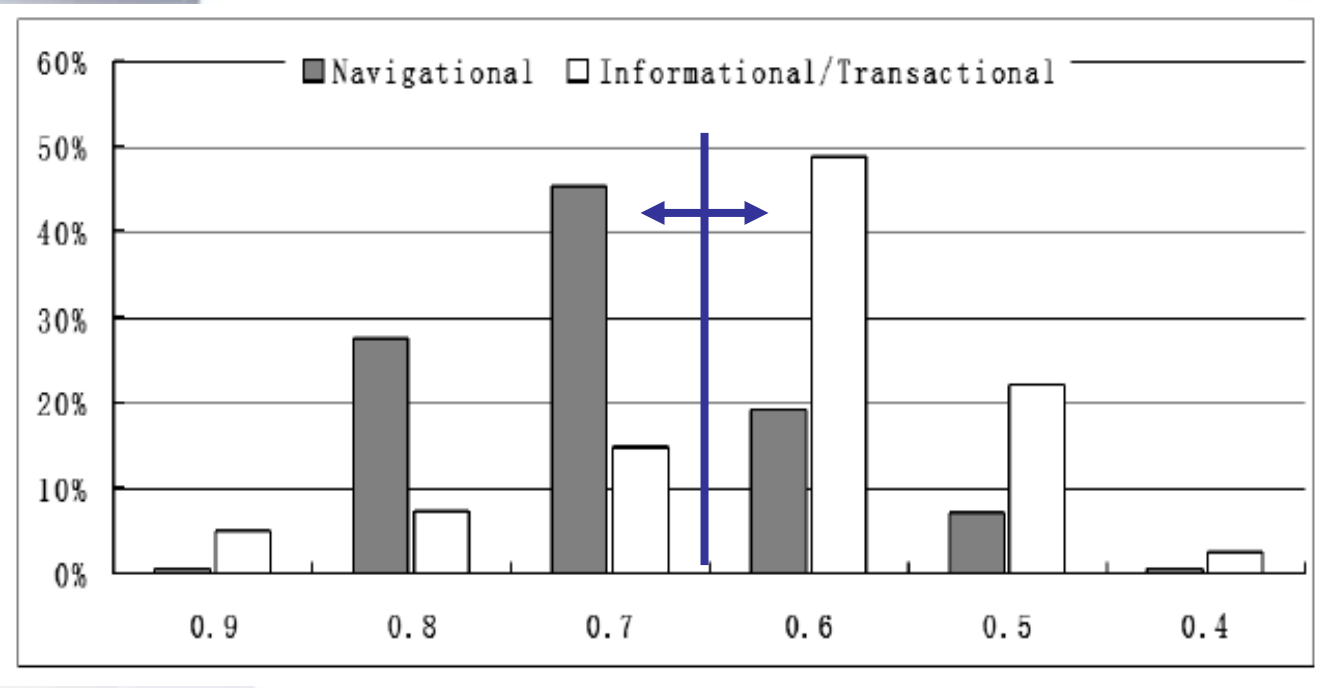
Query Type Identification Algorithm

- N-click satisfied rate
  - Assumption 1(懒鬼假設): When user submits a navigational type query, he clicks a small number of result URLs.
    - User has a specified search target in navigational searches
    - He is intended to click the highly-related results only.
  - N-click satisfied rate

$$nCS(Query\ q) = \frac{\#(Session\ of\ q\ that\ involves\ less\ than\ n\ clicks)}{\#(Session\ of\ q)}.$$

Query Type Identification Algorithm

- Distribution of nCS for search engine queries



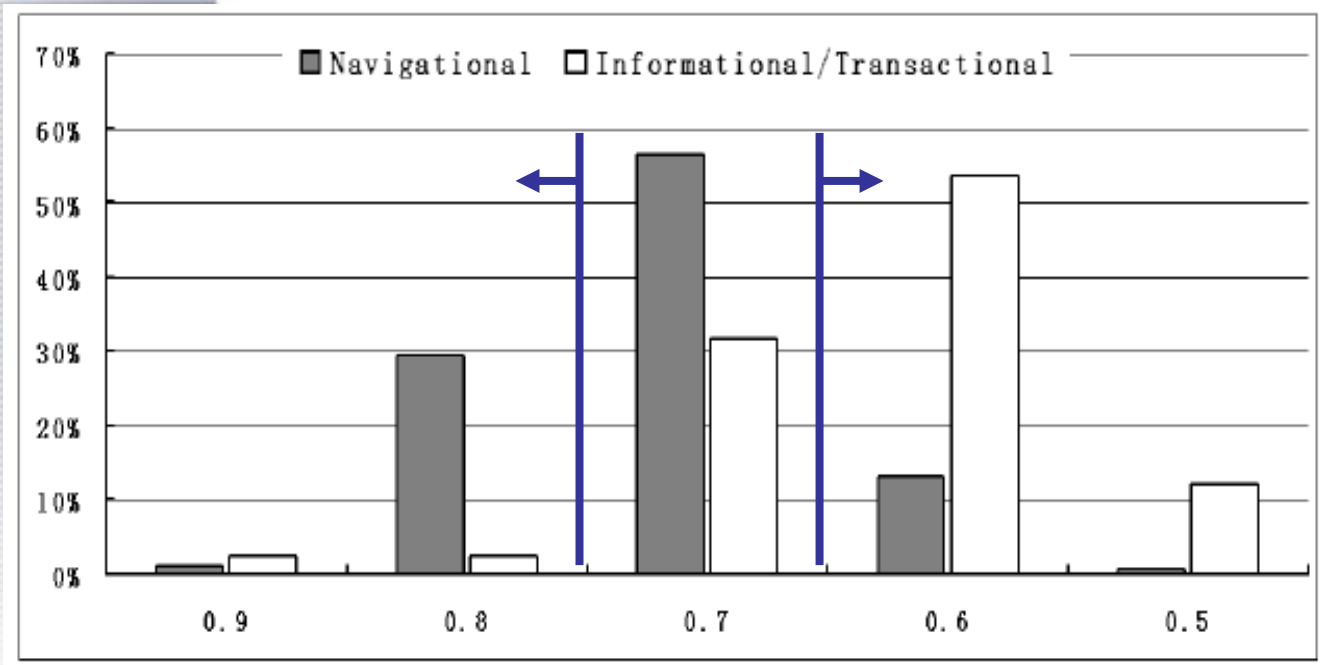
Query Type Identification Algorithm

- Top-n-result satisfied rate
  - Assumption 2(封面假設):When user submits a navigational type query, he only clicks the top-ranked result URLs.
    - Navigational type search has good performance (usually over 80% correct answers are returned at top 1 ranking result)
    - It is not necessary for him to click other results
  - Top-n-result satisfied rate

$$nRS(Query\ q) = \frac{\#(Session\ of\ \boldsymbol{q}\ that\ involves\ clicks\ only\ on\ top\ \boldsymbol{n}\ results)}{\#(Session\ of\ \boldsymbol{q})}.$$

Query Type Identification Algorithm

- Distribution of nRS for search engine queries



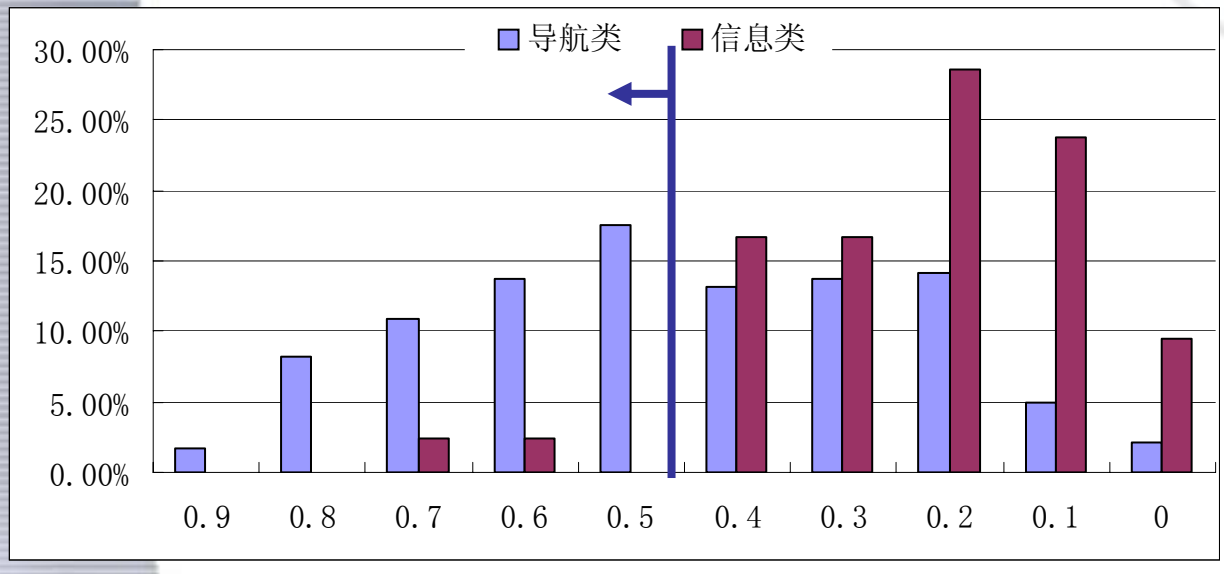
Query Type Identification Algorithm

- Click Distribution
  - Assumption 3(焦點假設):When different users submit a same navigational type queries, they intend to click the same result URL.
    - Navigational type queries have specific search targets
    - If this target appears in the result URL list, users will focus on it.
  - Click Distribution

$$CD(Query\ q) = \frac{\#(Session\ of\ q\ that\ involves\ clicks\ on\ the\ most\ frequently\ clicked\ results)}{\#(Session\ of\ q)}$$

Query Type Identification Algorithm

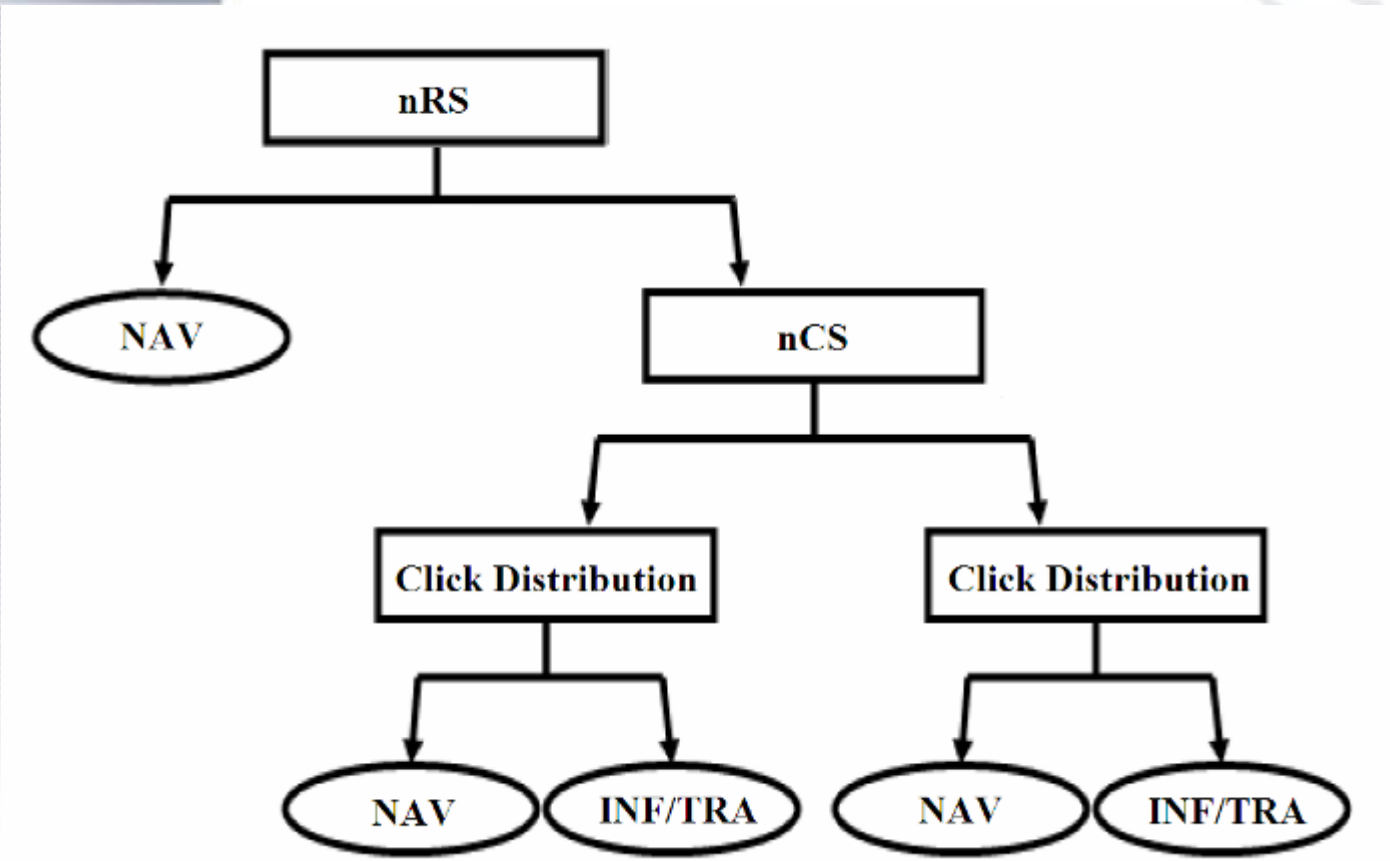
- Distribution of CD for search engine queries



Queries	Focus URL
讀寫網	www.duxie.net/
南方都市報	www.nanfangdaily.com.cn/
卓越網	www.joyo.com/

Query Type Identification Algorithm

- A query type identification decision tree



Query Type Identification for Web Search Engines

清华大学 Tsinghua University

- Research Background
- User analysis for query type identification
- Query Type Identification Algorithm
- Experiments Results and Discussions

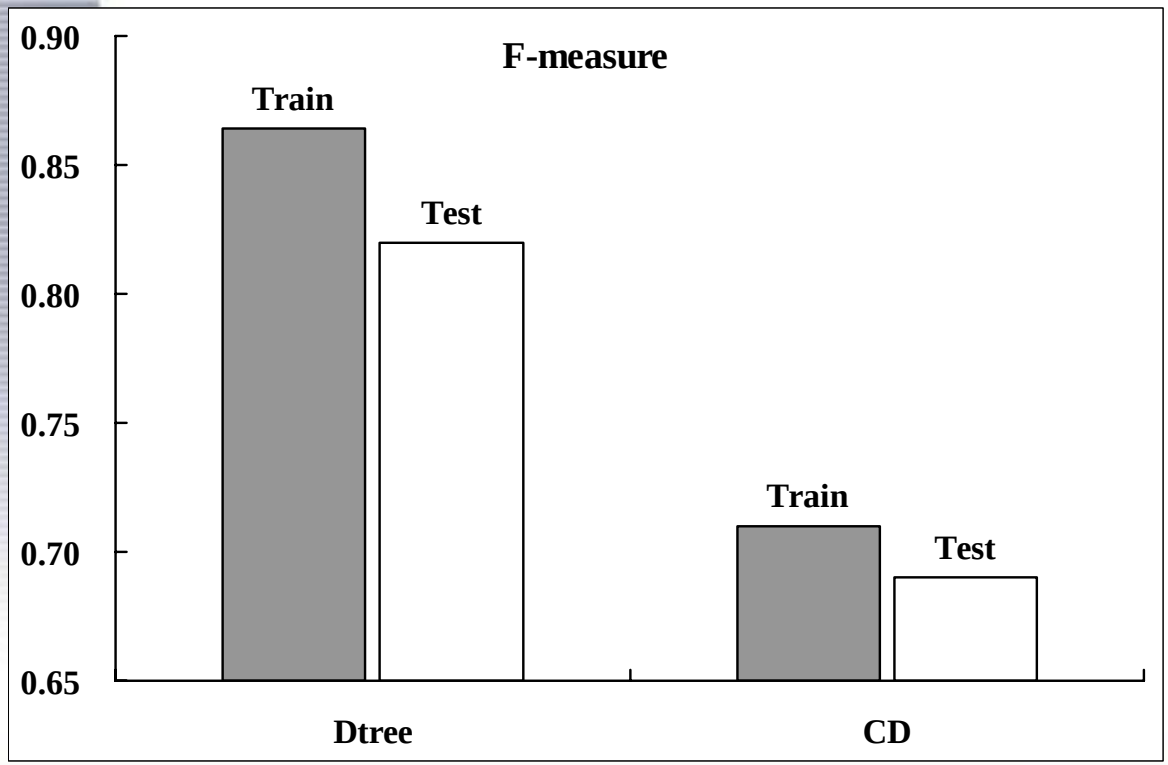
实验结论与应用方式讨论

清华大学
Tsinghua University

- Test set
  - Completely different from the training set
    - Different annotation methods:
    - Obtain informational type queries from a Chinese search engine performance contest organized by TianWang.com
    - Obtain navigational type queries from a famous Chinese Web directory (Hao123.com)
    - 200+ test queries

实验结论与应用方式讨论

- Experimental results
  - Our method outperforms previous Click-Distribution based method. (+30% in training, +19% in testing)



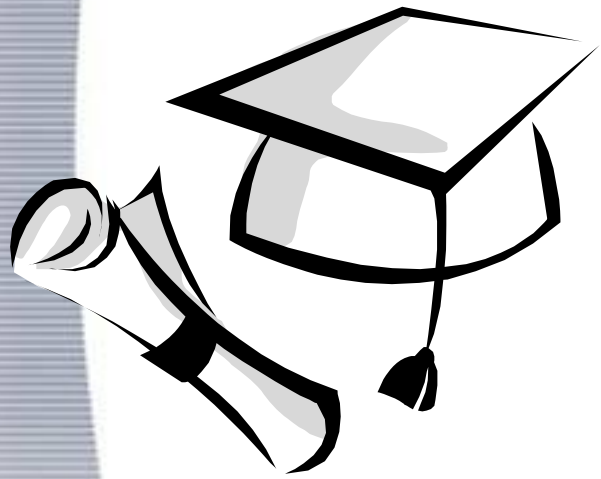
实验结论与应用方式讨论

清华大学 Tsinghua University

- Experimental results

	Training set			Test set		
	INF/TRA	NAV	Mixed	INF/TRA	NAV	Mixed
Precision	76.00%	91.07%	87.65%	73.74%	85.62%	81.49%
Recall	66.67%	90.71%	85.25%	72.84%	86.18%	81.54%
F-measure	0.71	0.91	0.86	0.73	0.85	0.81

  - Over 80% queries are correctly classified both in training and testing sets



Thank you!

Questions or comments?