Automatic Search Engine Evaluation with Click-through Data Analysis

Yiqun Liu

State Key Lab of Intelligent Tech. & Sys

Jun. 3th, 2007

Yiqun Liu

Ph.D student from Tsinghua University, Beijing, China.
liuyiqun03@gmail.com

Recent work:

- Using query log and click-through data analysis to:
  - identify search engine users' information need types
  - evaluate search engine performance automatically
  - separate key resource pages from others
  - estimate Web page quality

Our Lab:

清華大學 Tsinghua University

Beijing 2008 搜狐 SOHU.com 北京2008年奥运会赞助商

- A joint lab
- R&D Support to a widely-used Chinese Search Engine Sogou.com, platform to get research results realized.

Yiqun Liu

Ph.D student from Tsinghua University, Beijing, China.

liuyiqun03@gmail.com

- Web Data Cleansing
  - Using query-Independent features and ML algorithms
  - 5% web pages can meet >90% user's search needs
- Query type identification
  - Identify the type of user's information need
  - Over 80% queries are correctly classified
- Search engine performance evaluation
  - Construct query topic set and answer set Automatically .
  - Obtain similar evaluation results with manual based methods, and cost far less time and labor.

Introduction

- Lots of search engines offer services on the Web
- Search Engine Performance Evaluation
  - Web Users
    - over 120 million users in mainland
  - Search Advertisers
    - spending 5.6 billion RMBs in 2007
  - Search engineers and researchers

Introduction

清华大学
Tsinghua University

- **Evaluation is a key issue in IR research**
  - Evaluation became central to R&D in IR to such an extent that new designs and proposals and their evaluation became one. (Saracevic, 1995)
- **Cranfield-like evaluation methodology**
  - Proposed by Cleverdon et al in 1966.
  - A set of query topics, their corresponding answers (usually called qrels) and evaluation metrics.
  - Adopted by IR workshops such as TREC and NTCIR.

Introduction

清华大学
Tsinghua University

- **Problems with Web IR evaluation**
  - 9 people months are required to judge one topic for a collection of 8 million documents. (Voorhees, 2001)
  - Search engines (Yahoo!, Google) index over 10 billion Web documents.
  - Almost Impossible to use human-assessed query and qrel sets in Web IR system evaluation.

Related works

清华大学
Tsinghua University

- Efforts in automatic search engine performance evaluation (Cranfield-like)
  - Considering pseudo feedback documents as correct answers
    (Soboroff, 2001; Nuray, 2003)
  - Adopting query topics and qrels extracted from Web page directories such as open directory project (ODP)
    (Chowdhury, 2002; Beitzel, 2003)

Related works

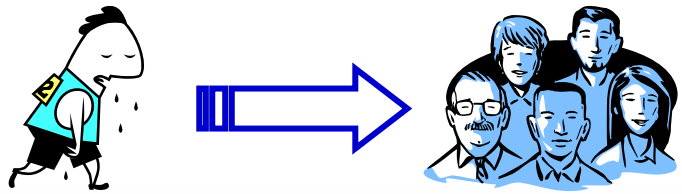
清华大学
Tsinghua University

- Efforts in automatic search engine performance evaluation (other evaluation approaches)
  - *Term Relevance Sets (Trels)* method. Define a pre-specified list of terms relevant and irrelevant to these queries. (Amitay, 2004)
  - The use of click-through data. Construct a unified meta search interface to collect users' behaviour information. (Joachims, 2002)

Our method

清华大学
Tsinghua University

- A cranfield-like approach
  - Accepted by major IR research efforts
  - Difficulty: annotating all correct answers automatically
- Click-through behavior analysis
  - Single user may be cheated by search spams or SEOs.
  - User group's behavior information is more reliable.



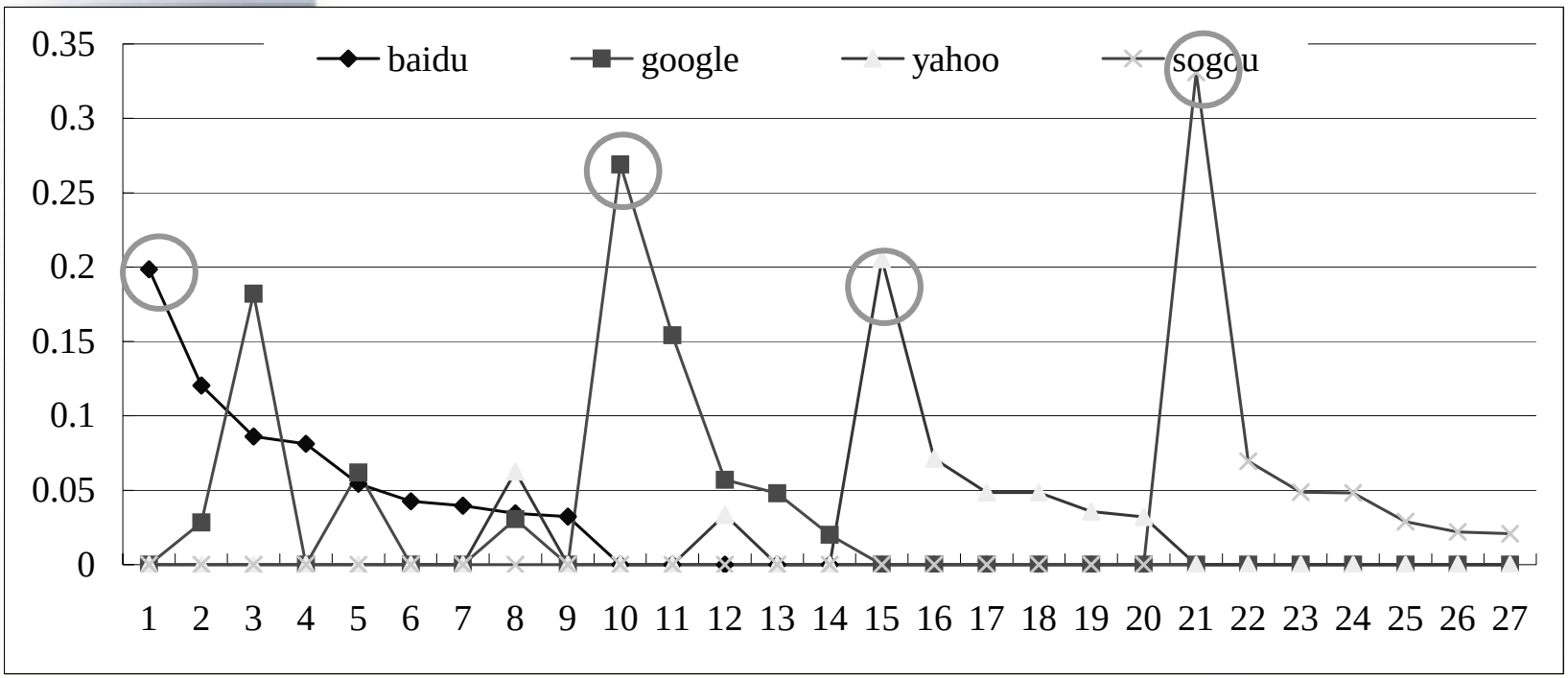
Automatic Evaluation Process

清华大学 Tsinghua University

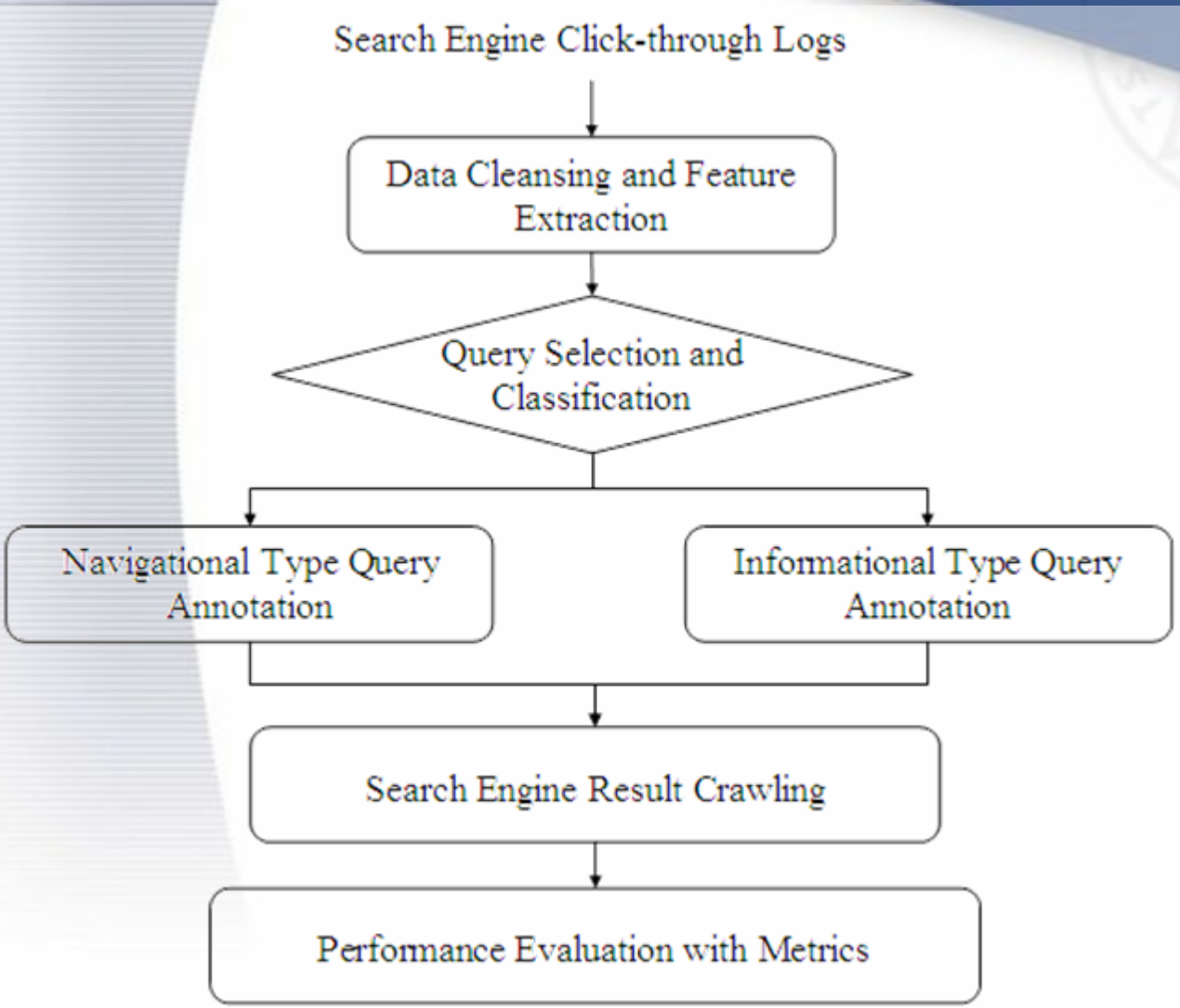
- **Information need behind user queries**
  - **Proposed by Broder (2003)**
  - **Navigational type:**
    One query have only one correct answer.
  - **Informational type:**
    One query may have several correct answers.
- Different behavior over different types of information needs

Information needs and Evaluation

- Informational queries cannot be annotated
  - People click different answers while using different search engines.



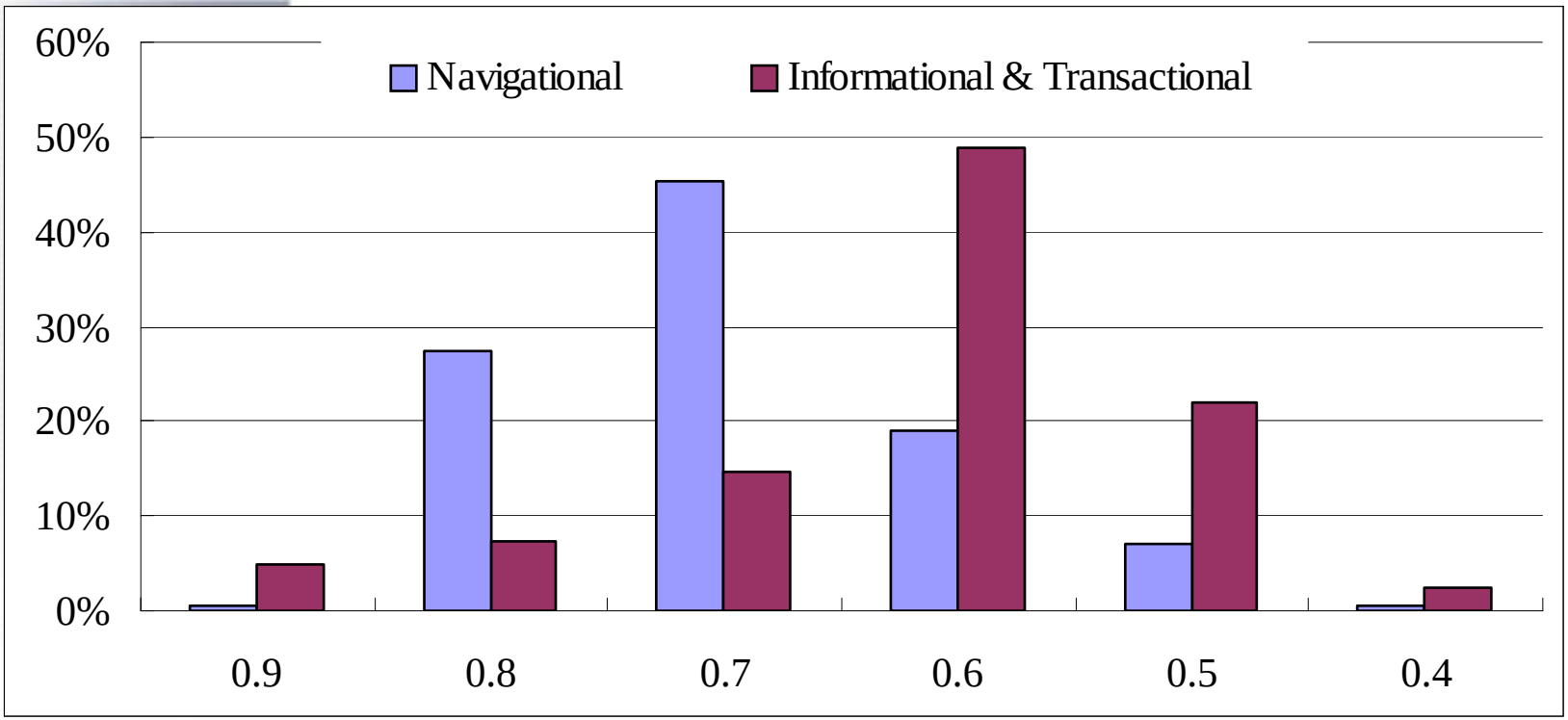
Automatic Evaluation Process



Query Set Classification

清华大学
Tsinghua University

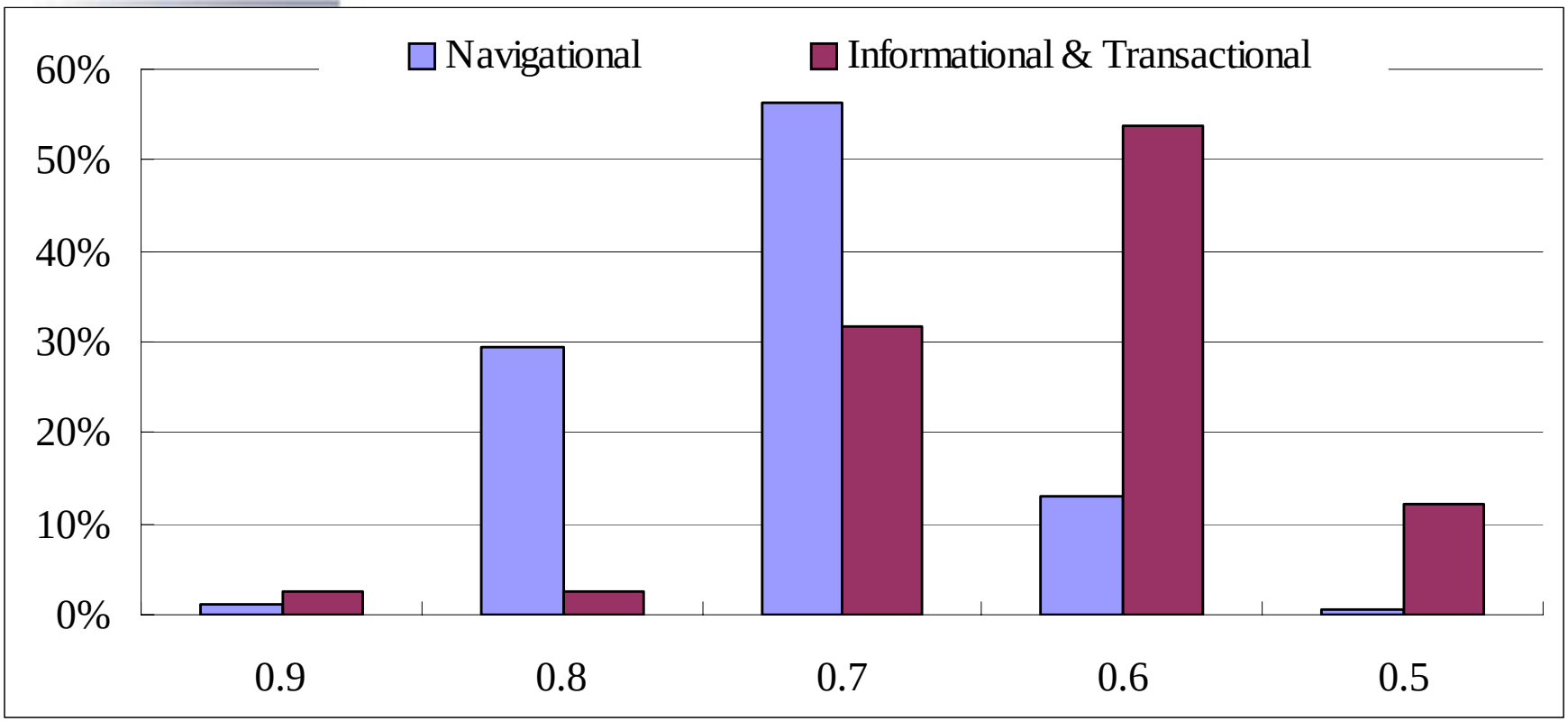
- **Less Effort Assumption & N Clicks Satisfied (nCS) Evidence**



Query Set Classification

清华大学
Tsinghua University

- **Cover Page Assumption and Top N Results Satisfied (nRS) Evidence**



Query Set Classification

清华大学
Tsinghua University

- **Click Distribution Evidence**
  - Proposed by Lee (Lee, 2005). Also based on click-through information.
  - Users tend to click the same result while proposing a same navigational type query

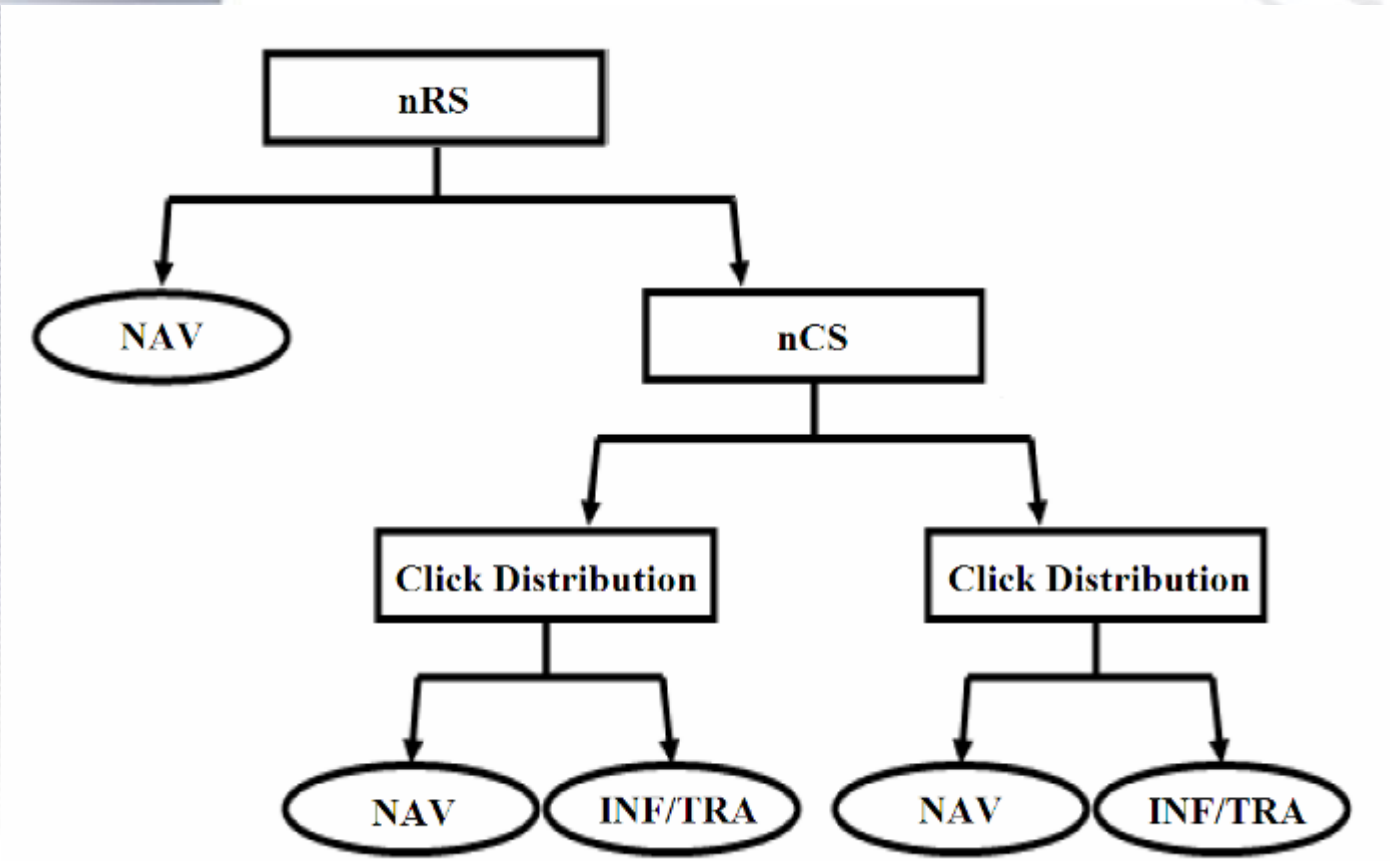
$$CD(Query\ q) = \frac{\#(Session\ of\ q\ that\ involves\ clicks\ on\ the\ most\ frequently\ clicked\ results)}{\#(Session\ of\ q)}$$

  - Less than 5% informational / Transactional queries' CD value is over ½, while 51% navigational queries' corresponding value is more than ½.

Query Set Classification

清华大学
Tsinghua University

- A decision tree algorithm



Answer Annotation

- Navigational type query annotation
  - Define: Click focus

$$ClickFocus(Query\ q, Result\ r) = \frac{\#(Session\ of\ \ q\ that\ clicks\ r)}{\#(Session\ of\ \ q)}$$

  - Annotate *q* with the result *r* whose *ClickFocus* value is the largest.

Answer Annotation

清华大学
Tsinghua University

- Annotation Algorithm

For a given Query *Q* in the Query Set and its clicked result list *r1*, *r2*, …, *rM* :

IF *Q* is navigational
    Find *R* in *r1*, *r2*, …, *rM*, *ClickFocus*(*Q*,*R*) = *ClickDistribution*(*Q*);
IF CD(*Q*) > *T1*
        Annotate *Q* with *R*;
        EXIT;
ELSE
        Q cannot be annotated;
END IF
ELSE   //Q is informational
    Q cannot be annotated;
END IF

Experiment Results

清华大学
Tsinghua University

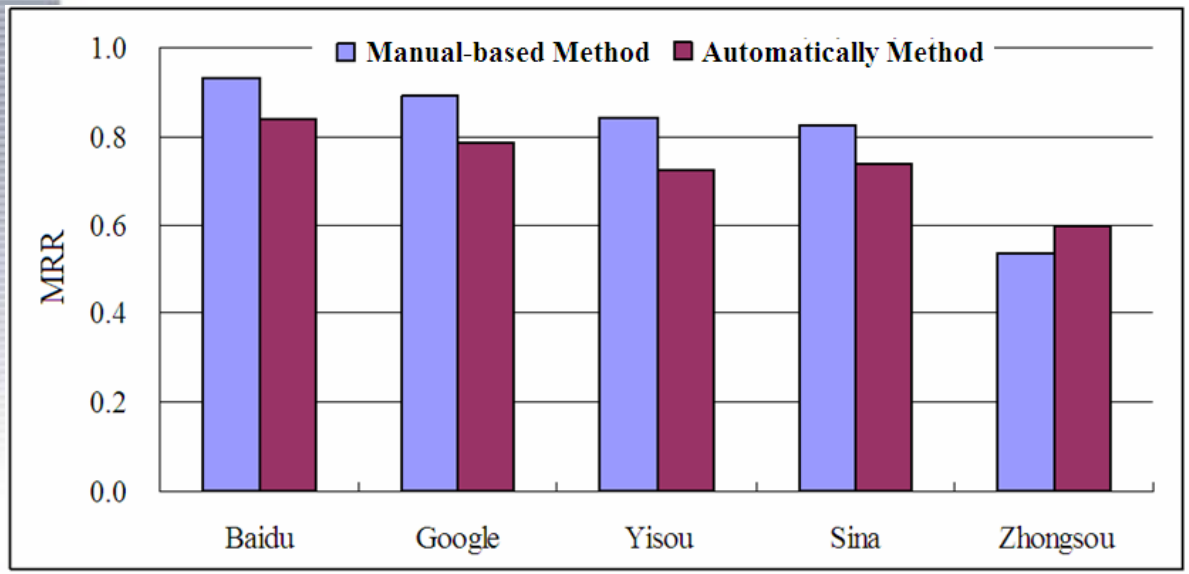
- Experiment data
  - Collected by Sogou.com from Jun 2006 to Jan 2007.
  - Over 700 million querying or clicking events totally.
- Annotation experiment results
  - 5% of all results are checked mannually.

	#(Annotated queries)	#(Checked sample set)	Accuracy
Jun. 06 - Aug. 06	13,902	695	98.13%
Sept.06 - Nov. 06	13,884	694	97.41%
Dec. 06 - Jan. 07	11,296	565	96.64%

Experiment Results

清华大学
Tsinghua University

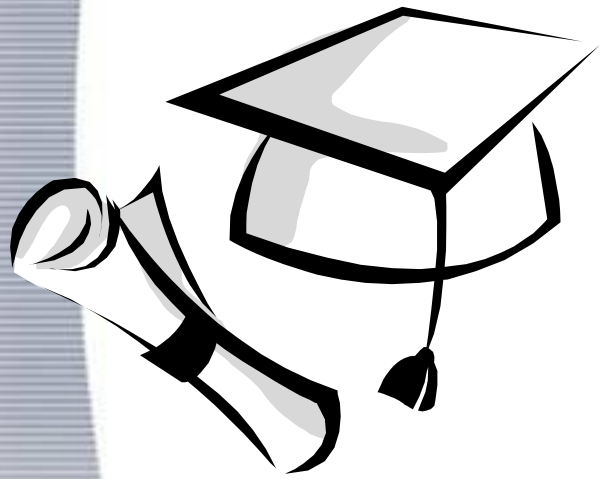
- Performance evaluation experiment
  - 320 manual-developed queries and corresponding answers are used in the evaluation experiment.
  - Correlation value between MRRs of the manual and the automatically methods is 0.965.



Applications and Future works

清华大学 Tsinghua University

- Choosing the correct search portal
  - Overall performance
  - Performance for queries in a certain field
- Search engine monitoring
  - Complicated computer cluster systems are used in modern search engines
  - To notify the engineers when the search engine fails. (performance going down)



Thank you!

Questions or comments?