Automatic Query Type Identification Based on Click Through Information

Yiqun LIU[1], Min ZHANG[1], Liyun RU[2] and Shaoping MA[1]
liuyiqun03@mails.tsinghua.edu.cn, z-m@tsinghua.edu.cn
1: State Key Lab of Intelligent Tech. & Sys., Tsinghua University; 2: R&D center, Sogou Corporation

RESEARCH BACKGROUND

Observe User Behavior from a Search Engine's Prospect

- One dimension world: Query stream & click stream.
- User request behind the query: Users who have different search requests may share a same query.
  - Example: War craft (Site visiting, Software download, Information overview…)
- Identifying user's information need behind query is necessary for Web search

Query Type Identification according to user's information need

- Proposed by Broder et al (Broder, 2002) and Rose et al (Rose, 2004)
- Navigational search: queries with fixed target pages.
- Informational & Transactional search: queries with no fixed target pages.

Retrieval Algorithm Improvement based on Query Type Identification

- Retrieval algorithms have different performance with different query types.
- Navigational: exact match, anchor text based ranking, …
- Informational & Transactional: link analysis based ranking, standard IR model, …

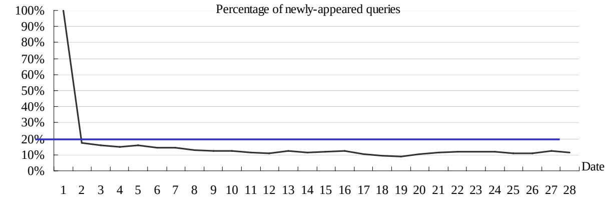
ANALYSIS INTO SEARCH ENGINE LOGS

Sogou Query Logs

- Collected from *http://www.sogou.com* from 2006/02/01 to 2006/02/28
- 86538613 non-empty queries, 4345557 unique ones, 26255952 query sessions
- Including click-through information

Possibility of predicting query type using click through information and other features

- Anchor text information (Lee et al 2005 and Kang et al 2003)
  - Only 15% queries have a matched web page anchor
- Features collected from query content
- Features extracted from click-through data (**our key idea**)
  - More than 80% queries have past click-through information

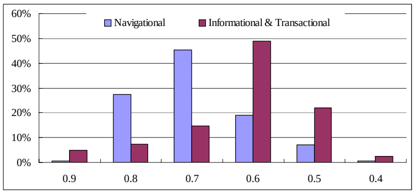


CLICK-THROUGH BASED EVIDENCES

Less Effort Assumption & N Clicks Satisfied (nCS) Evidence

- While performing a navigational type search request, user tend to click a small number of URLs in the result list.

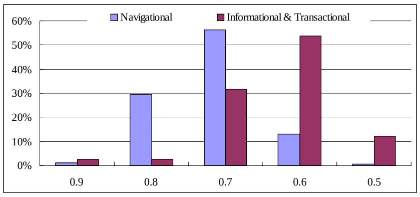
$$nCS(Query\ \boldsymbol{q}) = \frac{\#(Session\ of\ \boldsymbol{q}\ that\ involves\ less\ than\ \boldsymbol{n}\ clicks)}{\#(Session\ of\ \boldsymbol{q})}$$



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

- While performing a navigational type search request, user tend to click only the first few URLs in the result list.

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



Click Distribution Evidence

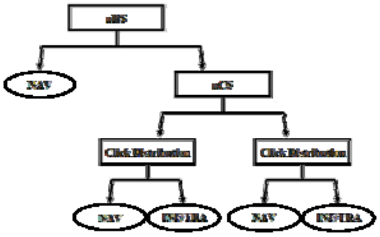
- Proposed by Lee (Lee, 2005). Also based on click-through information.

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

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

Learning Based Query Type Identification Algorithm

- Based on two new proposed nCS, nRS evidences and Click-distribution evidence.
- Adopt decision tree learning algorithm because of the small number of evidences.



EXPERITMENTAL RESULTS

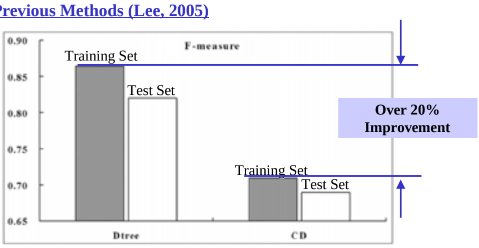
Training Set, Test Set and Evaluation Measures

- Training set: Chinese, 45 informational / transactional, 153 navigational, collected from Sogou.com
- Test set: Chinese, 81 informational / transactional, 152 navigational, collected from TianWang.com and hao123.com
- Evaluation measures: Precision/Recall/F-measure

Query Type Identification 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

Compared with Previous Methods (Lee, 2005)



Over 20% Improvement

CONCLUSIONS AND FUTURE WORKS

Query Type Identification Can be finished via Click-through Data Analysis

- Over 80% queries can be classified with the help of click-through information.
- Two new evidences (nCS, nRS) are proposed.
- A learning based identification algorithm is used to combine evidences.
- Over 80% queries can be correctly classified according to experimental results.
- Over 21% performance improvement is made compared to previous click-through based methods

Possible Future Works

- Automatic Web Resource Finding
- Automatic Search Engine Evaluation

BIBLIOGRAPHY

1. Andrei Broder: A taxonomy of web search. SIGIR Forum(2002), Volume 36(2):3-10, 2002.
2. Daniel E. Rose and Danny Levinson, Understanding User Goals in Web Search. In proceedings of the 13th World-Wide Web Conference, 2004.
3. Uichin Lee, Zhenyu Liu and Junghoo Cho, Automatic Identification of User Goals in Web Search. In proceedings of the 14th World-Wide Web Conference, 2005.
4. I. Kang and G. Kim. Query type classication for web document retrieval. In Proceedings of ACM SIGIR '03, 2003.