From Skimming to Reading:

A Two-stage Examination Model for Web Search

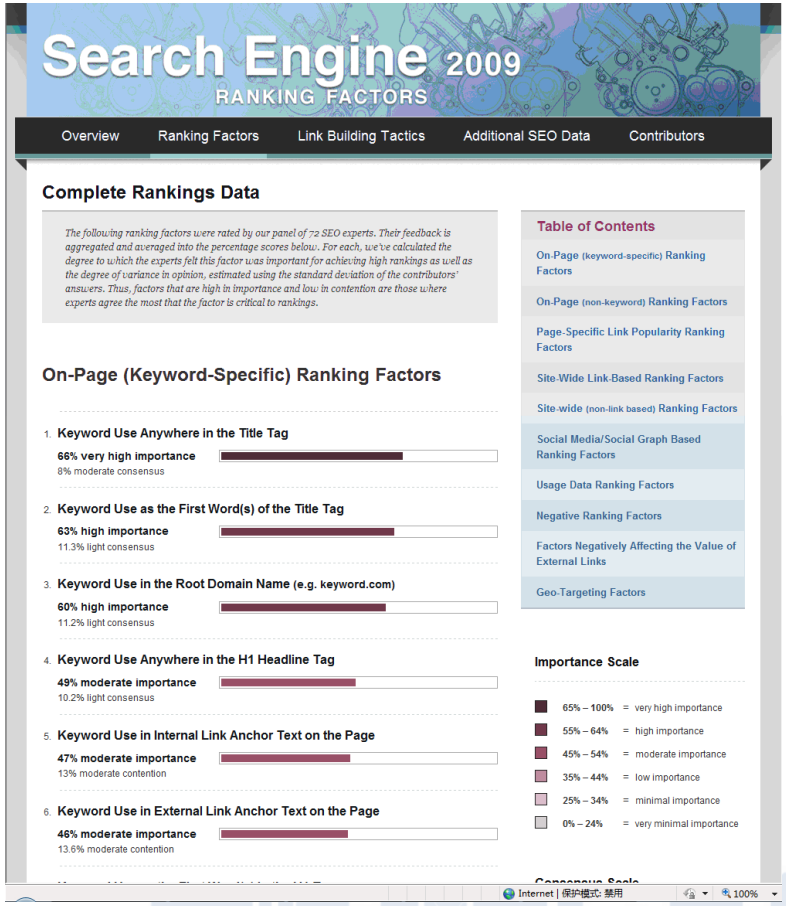
Yiqun Liu, Chao Wang, Ke Zhou,
Jianyun Nie, Min Zhang, Shaoping Ma

Tsinghua University; Yahoo Labs; Université de Montréal

November, 2014

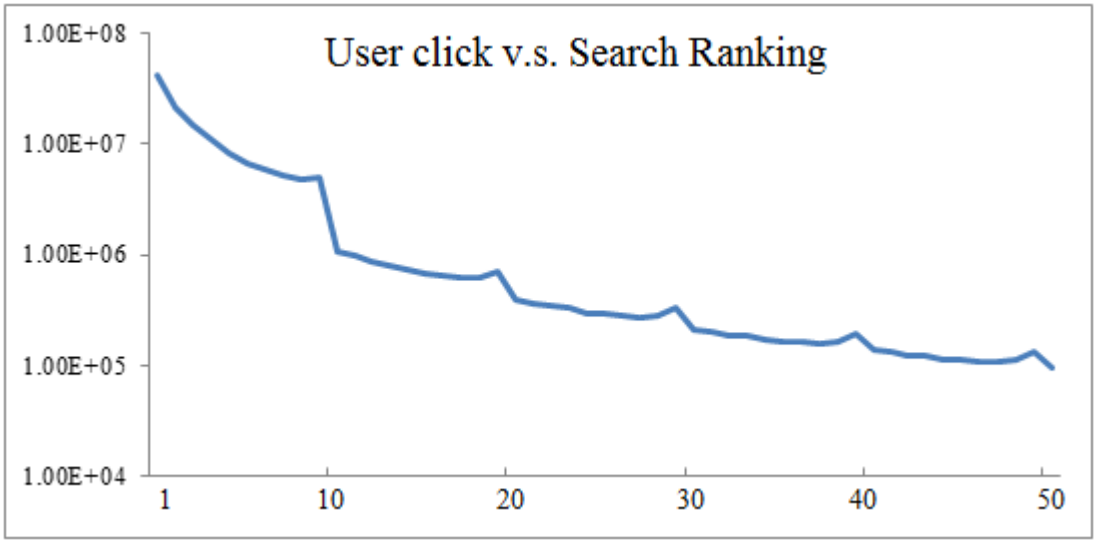
Background: Search Engine Result Ranking

- SEO report: 100+ signals
- Yahoo LTR task: 700+ signals
  - Hyperlink, Content relevance, **User behavior**, Page structure, Freshness, Service stability, ……
- Basic assumption
  - Results that are clicked more tend to be more relevant



Background: User Implicit Feedback

- **A simple solution: user click = relevance voting**
  - Tsinghua University => www.tsinghua.edu.cn
- **Problem: position bias**
  - Users tend to click more on higher-ranked results



Courtesy of http://hubdesignsmagazine.com/2011/03/27/its-good-to-be-on-the-first-page-of-google/

Background: Examination Hypothesis

- The likelihood that a user will click on a search result is influenced by
  - Whether the user examined the search result
  - Whether the result is attractive/relevant

$$C_i = 1 \rightarrow \quad E_i = 1, R_i = 1$$

  - Examination: user has comprehended (part of) the result and made a decision on whether to click.

- **How do users examine search results?**
- **How to estimate the probability of examination?**

M. Richardson, et al. Predicting clicks: estimating the click-through rate for new ads. WWW 2007, pp. 521-530.

Result Examination: Click Models

- Examination Hypothesis

$$C_i = 1 \rightarrow \quad E_i = 1, R_i = 1$$

- Estimating examination with search user **behavior assumption**
  - Cascade model: $P(E_{i+1} = 1 | E_i = 1, C_i) = 1 - C_i$
  - Dependent click model (DCM):

$$P(E_{i+1} = 1 | E_i = 1, C_i = 0) = 1$$
$$P(E_{i+1} = 1 | E_i = 1, C_i = 1) = \lambda_i$$

  - User browsing model (UBM):

$$P(E_i = 1 | C_{1 \ldots i-1}) = \lambda_{r_i, d_i}$$

  - Other models: DBM, DBN, CCM, ...

Result Examination: Eye-tracking

- ***Strong Eye-mind Hypothesis***
  - There is no appreciable lag between what is fixated on and what is processed (Just et al., 1980).
  - Most existing studies: Inferring Examination with a threshold in fixation (200-500ms)

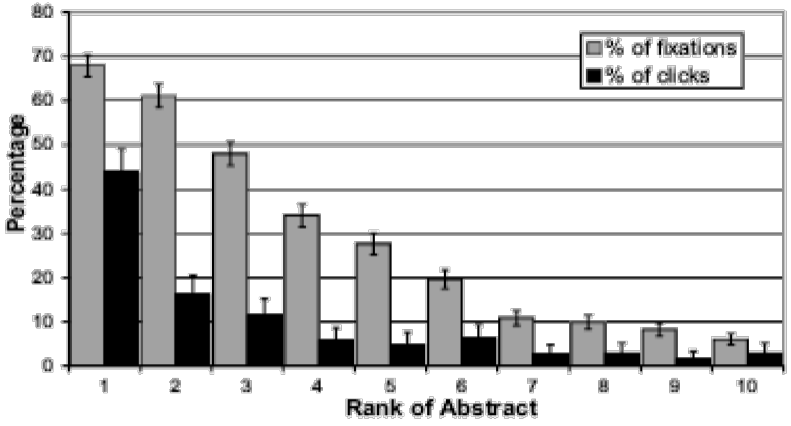


Figure 1: Percentage of times an abstract was viewed/clicked depending on the rank of the result.

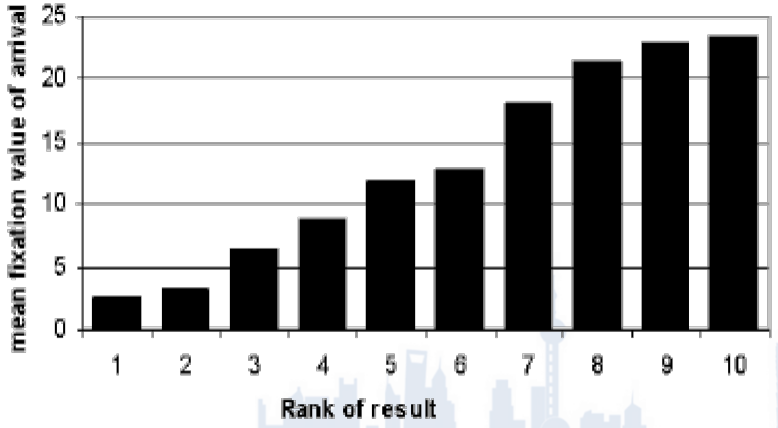
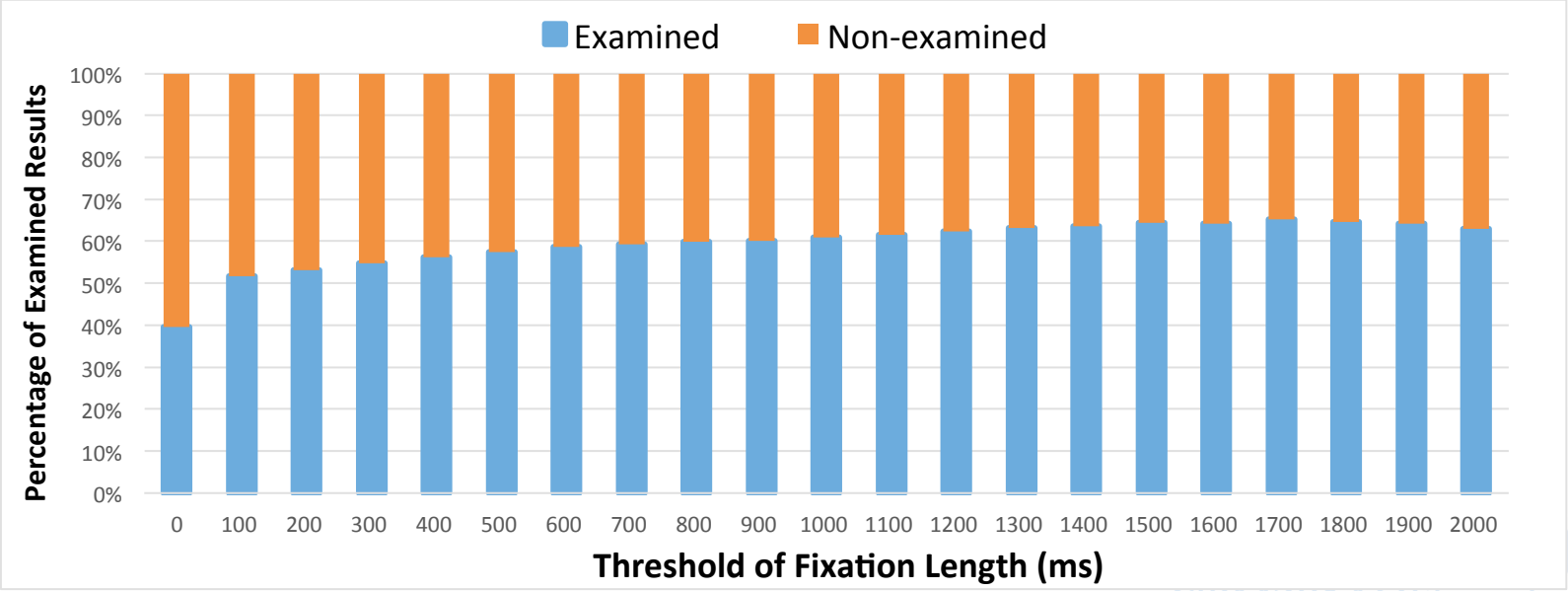


Figure 2: Mean time of arrival (in number of previous fixations) depending on the rank of the result.

Joachims et al., Eye-tracking analysis of user behavior in www search. SIGIR 2005

Result Examination: beyond Eye-tracking

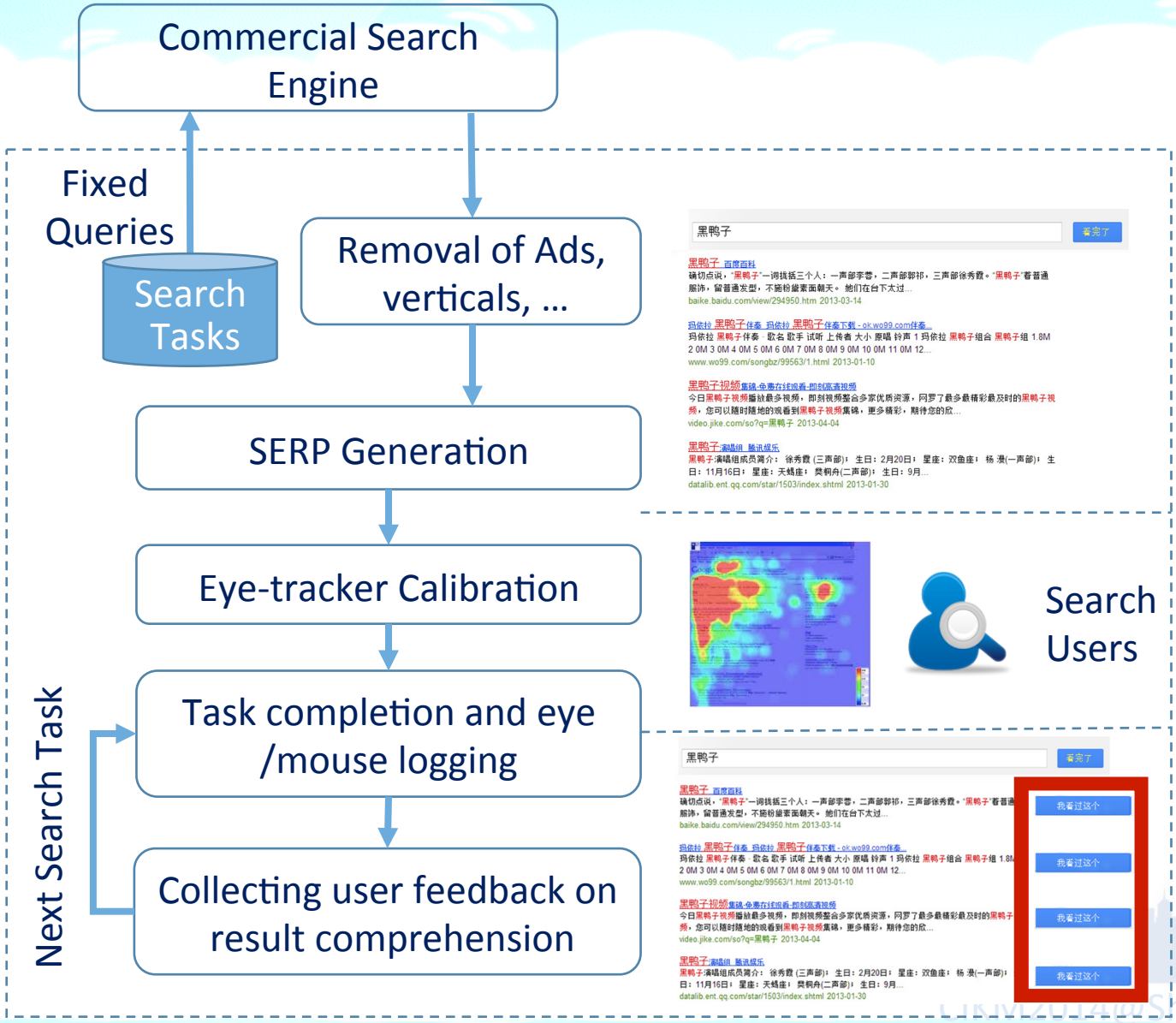
- **Problems with *Strong Eye-mind Hypothesis***
  - While the duration of the gaze is closely related to the duration of cognitive processes, *the two durations are not necessarily identical*. (Just & Carpenter, 1980)



Research Questions

- RQ1: How do users examine results on SERPs
- RQ2: How do behavior biases happen in user's examination process
- RQ3: How can we identify examination behavior

Collecting Examination Information



Collecting Examination Behavior on SERPs

- Search task details
  - Data Collected: click-through, mouse movement, eye movement, explicit feedback on examination.
  - 37 participants, 25 queries (INF:TRA:NAV = 2:2:1)

Examination Behavior Analysis

- Examination v.s. Fixation: *Eye fixation on a search result is a prerequisite for examining this result*

	Fixation=0	Fixation=1
Examine=0	31.61%	28.81%
Examine=1	5.49%	34.09%

Why don't you annotate the fixed results as examined?

Proportion	Answers
48%	Take a glance at the result without thinking about it.
26%	Take a glance at the result and feel unattractive to read it
16%	Feel that the result is not relevant.
10%	Cannot tell clear reason.

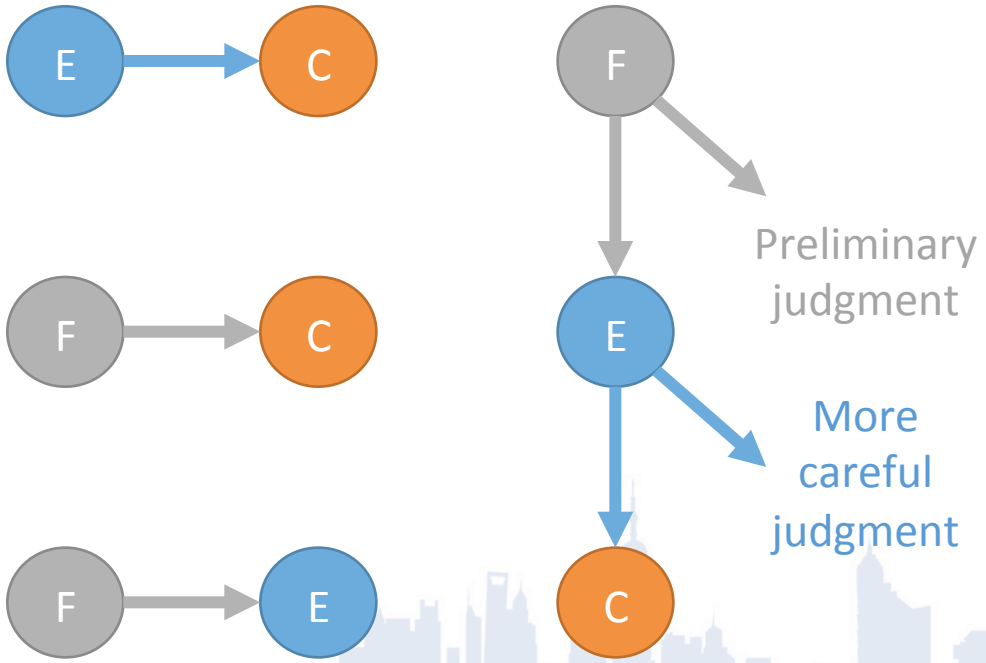
Examination Behavior Analysis

- Examination v.s. Click: *Examining a search result is a prerequisite for clicking on the result.*

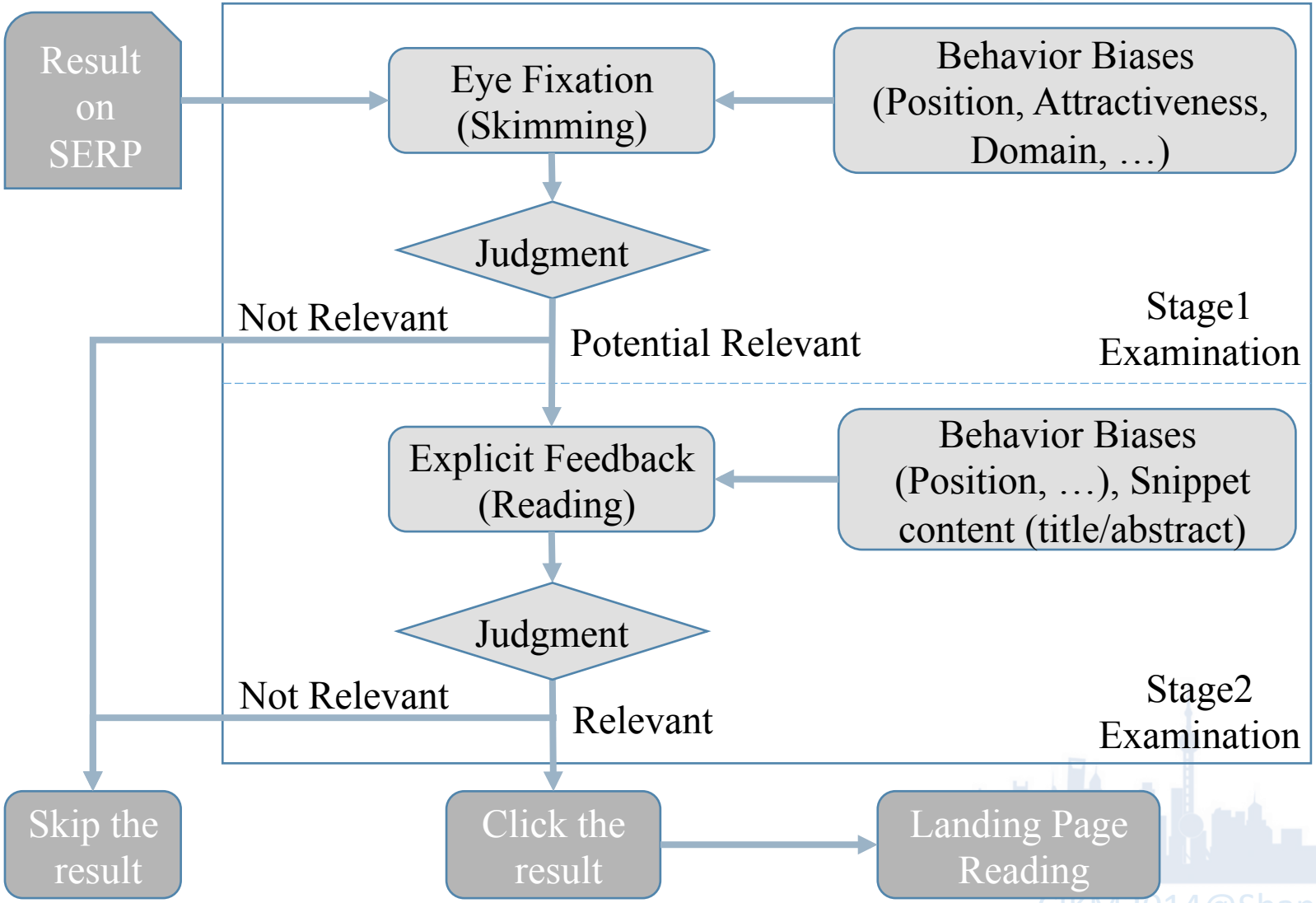
	Examine=0	Examine=1
Click=0	59.24%	17.57%
Click=1	1.18%	22.01%

	Fixation=0	Fixation=1
Click=0	34.96%	41.85%
Click=1	2.15%	21.04%

	Fixation=0	Fixation=1
Examine=0	31.61%	28.81%
Examine=1	5.49%	34.09%



A Two-Stage Examination Model



A Two-Stage Examination Model

- **Answer to RQ1: Users examine results with a two-stage model**
  - Stage1: skimming process, careful reading or not
  - Stage2: reading process, clicking or not
- Relationship with *information triage*
  - *the process of determining the priority of processing*
- Relationship with *selective attention*
  - *the process whereby the brain selectively filters out large amounts of sensory information to focus*

Research Questions

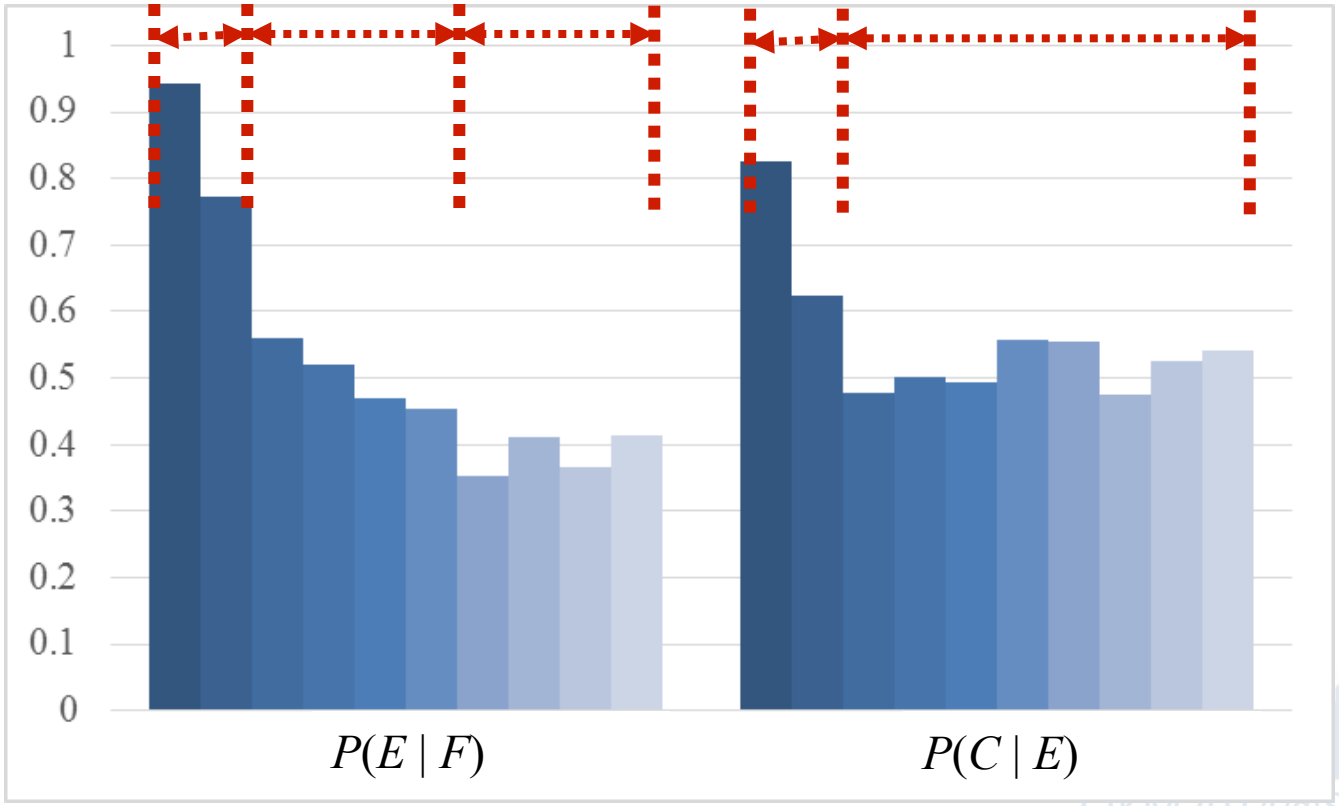
- RQ1: How do users examine results on SERPs
- RQ2: How do behavior biases happen in user's examination process
- RQ3: How can we identify examination behavior

Behavior Biases in Two-Stage Model

- **Behavior biases in Web search environment**
  - ***Position bias*:** Higher-ranked results receive more user attention (Craswell et al. 2008)
  - ***Attractiveness bias*:** attractiveness in result titles and abstracts affects user judgment(Bar-Ilan et al. 2009), multimedia vertical results draws much user attentions (Wang et al. 2013)
  - ***Trust bias*:** Results from trust-worthy Web domains are preferred by users (Ieong et al. 2012)

Position Bias

- **User judgments (for relevant results) in two stages are both affected by positions**



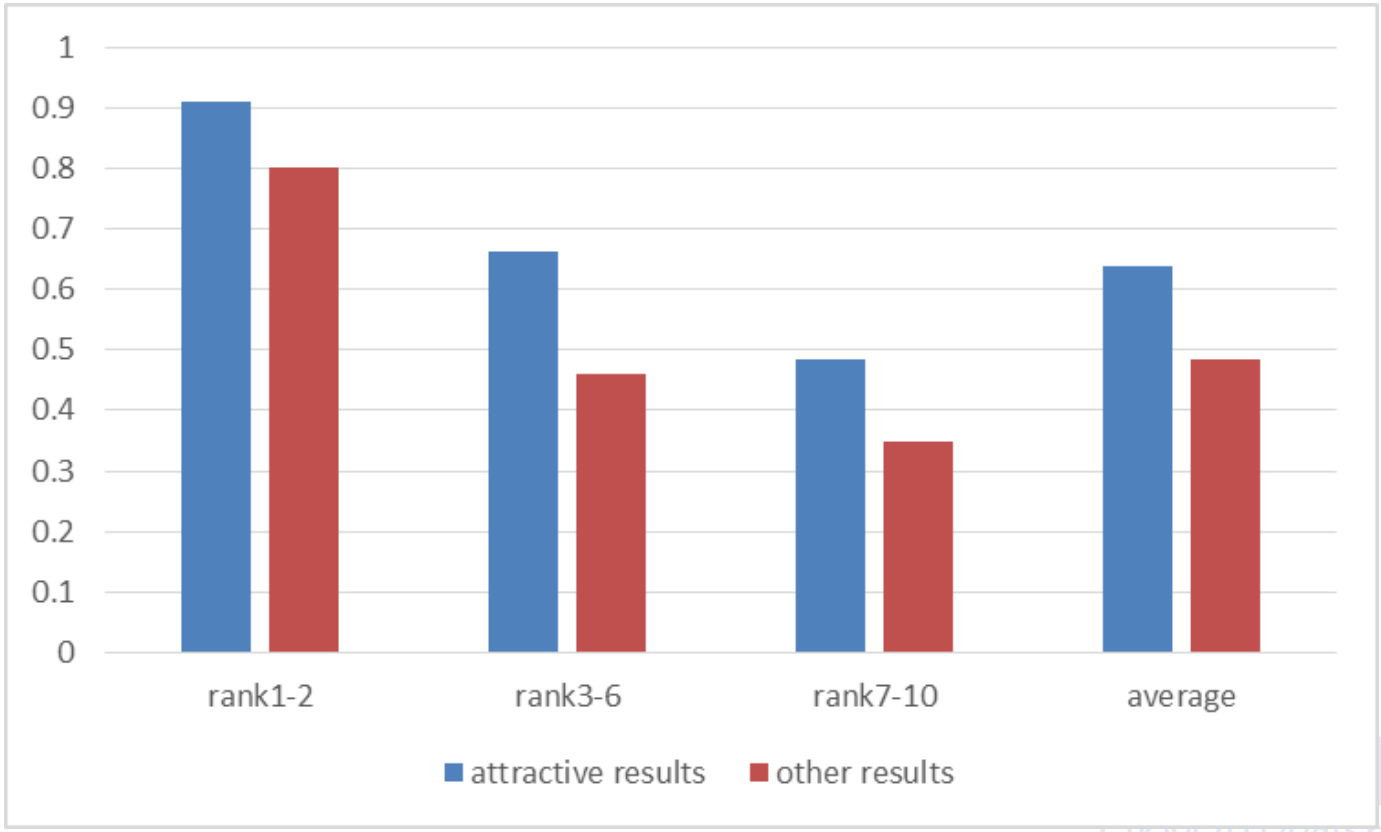
Attractiveness Bias

- **Attractive results draws significantly more attention in Stage 1 while doesn't affect the judgment in Stage 2.**
  - Attractive results: Results with the longest title and abstract exact match on SERPs

		Attractive results	Other results
P(E|F)	Average	0.637301	0.484615
	Variance	0.058769	0.066037
	p-value	0.005788	
P(C|E)	Average	0.57775	0.472463
	Variance	0.122599	0.082748
	p-value	0.158477	

Attractiveness Bias v.s. Position Bias

- **Attractiveness bias happens in all result positions for judgments in Stage 1.**



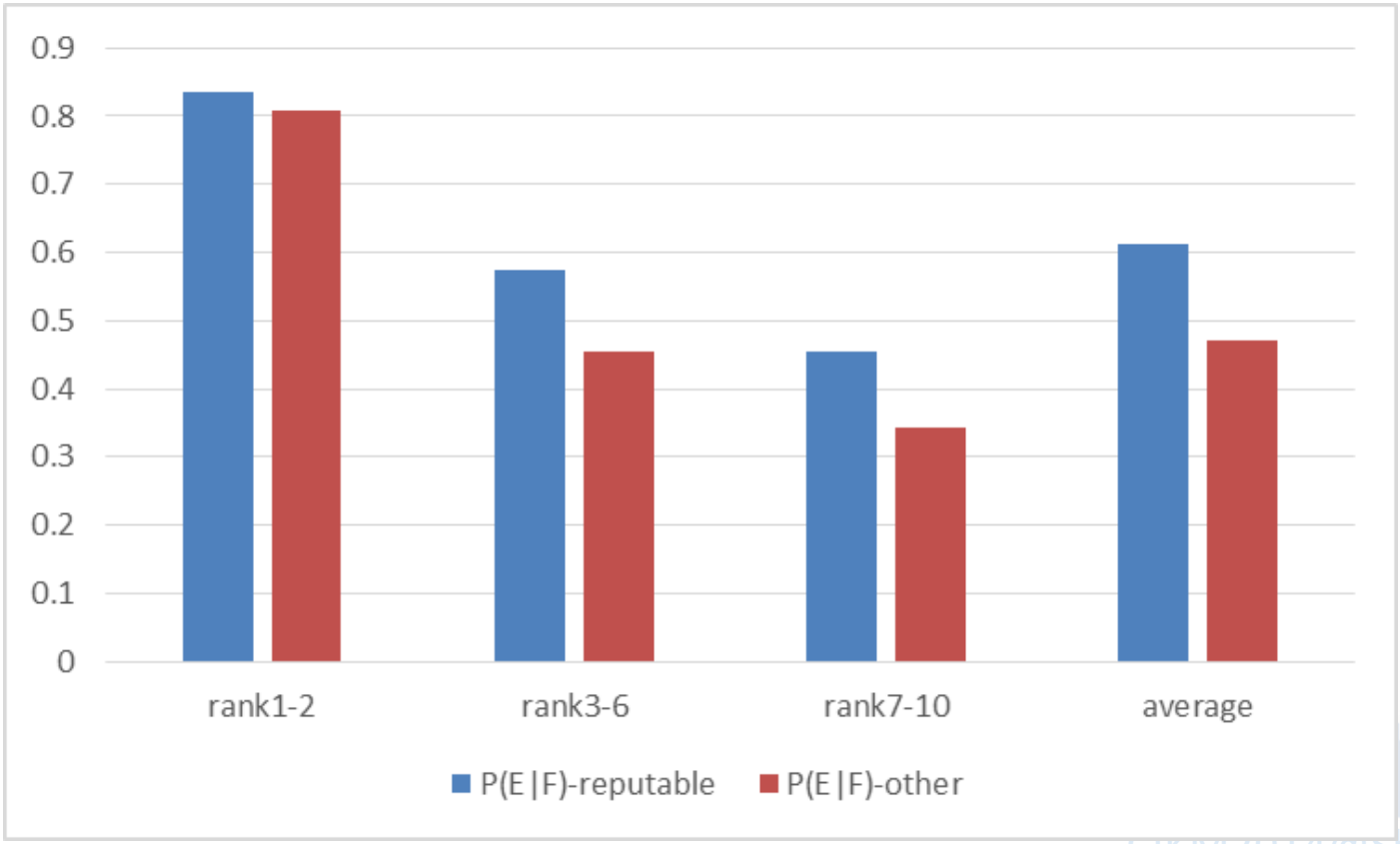
Trust Bias

- **Reputable results draws significantly more attention in Stage 1 while doesn't affect the judgment in Stage 2.**
  - Reputable results: results from Alexa.com top 100 popular sites in China

		Attractive results	Other results
P(E|F)	Average	0.613371	0.519443
	Variance	0.065817	0.079853
	p-value	0.000656	
P(C|E)	Average	0.470799	0.473674
	Variance	0.063693	0.089271
	p-value	0.311937	

Trust Bias v.s. Position Bias

- **Trust bias happens in relatively lower result positions for judgments in Stage 1.**



Effectiveness of Judgments in Two Stages

- **User examines more results in Stage 1, but the effectiveness of judgments in Stage 2 is higher**
  - Relevance judgment in Stage 1: entering Stage 2
  - Relevance judgment in Stage 2: result clicking

	Stage 1	Stage 2	Comparison
Number of examined results	5598/8900	3034/5598	-45.80%
Number of results judged as relevant	3034/5598	1779/3034	-38.27%
Precision	0.5968	0.6738	+11.43%
Recall	0.6040	0.6755	+10.58%
F-measure	0.6004	0.6747	+11.01%
AUC/ROC	0.6523	0.7169	+9.011%

Research Questions

- RQ1: How do users examine results on SERPs
- RQ2: How do behavior biases happen in user's examination process
- RQ3: How can we identify examination behavior

Identifying Examination Behavior

- Existing solution: identification with eye fixation
  - Cognitive studies: McConkie 1975; Just et al. 1980;
  - Web search: Cutrell et al, 2007; Buscher et al, 2012;
- Problems:
  - Equipment is too expensive
  - Users are required to calibrate
  - Fixed threshold setting is not reasonable

Identifying Examination Behavior

- Alternative solution: Mouse movement behavior
  - Mouse movement information could be collected at large scale without interrupting users
- Existing studies on fixation prediction
  - Eye-mouse coordination: Rodden, 2008; Huang, 2012
  - Fixation prediction: Guo, 2010; Huang, 2012
- Problem:
  - Previous target: predict the whole fixation sequence
  - New target: predict whether a result is examined or not

Mouse Movement Features

- Distance features:
  - User's total leftwards/rightwards/upwards/downwards movement distances in the result zone
- Position features:
  - The leftmost/rightmost/upmost/bottommost position cursor ever reaches in the result zone
- Duration features
  - Total mouse dwell time on a result/SERP/search task

Examination Prediction and Relevance Estimation

- Actual v.s. Predicted user behavior

Method	Accuracy	KAPPA	F-measure		
			E0	E1	E2
GBRT	0.6393	0.4519	0.7531	0.4479	0.6754
LogisticRegression	0.6310	0.4389	0.7517	0.4251	0.6668
SVM	0.6191	0.4163	0.7369	0.2853	0.6862
RandomForest	0.6151	0.4167	0.7332	0.4581	0.6286
Naïve Bayes	0.6056	0.3972	0.7236	0.3279	0.6564

	Actual User Behavior (incl. eye movement, user feedback on reading)		Predicted Behavior (mouse movement information only)	
	Two-stage model	Single-stage model	Two-stage model	Single-stage model
Accuracy	0.6440	0.5760	0.6400	0.5720
Precision	0.6910	0.8221	0.6872	0.8155
Recall	0.6970	0.3356	0.6941	0.3345
F-measure	0.6865	0.4747	0.6799	0.4693

Take-Home Messages

- RQ1: How do users examine results on SERPs
  - Two-stage examination: from skimming to reading
  - Information triage / selective attention in Web search
- RQ2: How do behavior biases happen in user's examination process
  - Users rely on different signals in different stages
- RQ3: How can we identify examination behavior
  - Supervised learning with mouse movement features

Thank you

Welcome to visit my homepage

http://www.thuir.cn/group/~YQLiu/