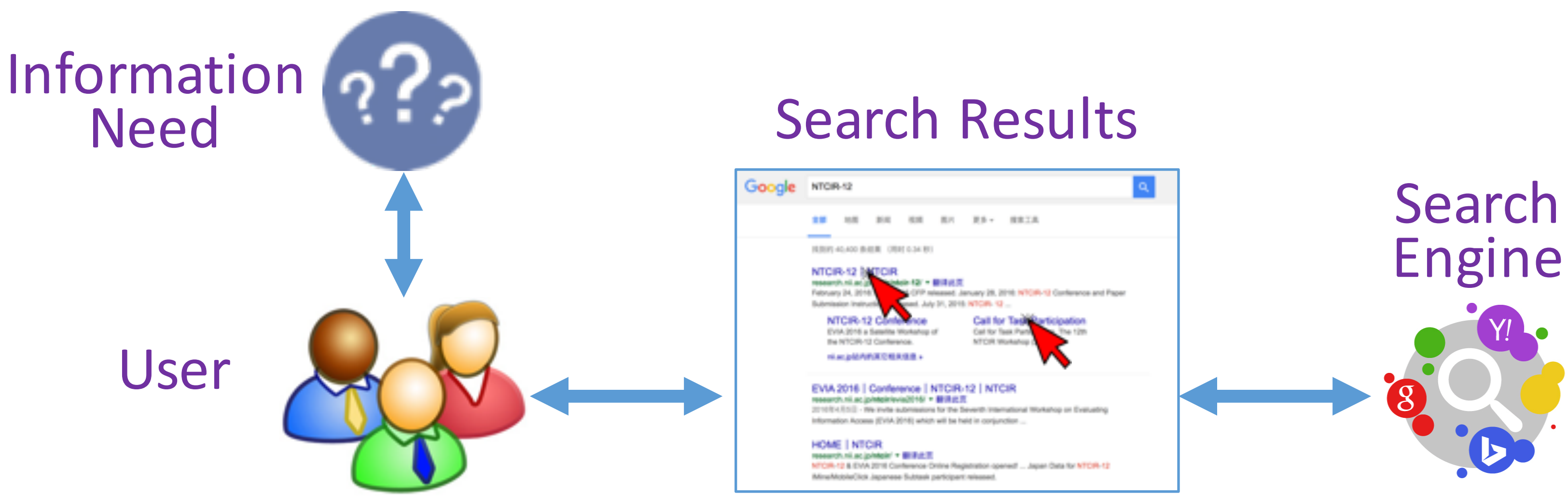
Understanding Web Search Satisfaction in a Heterogeneous Environment

Yiqun LIU

Department of Computer Science and Technology

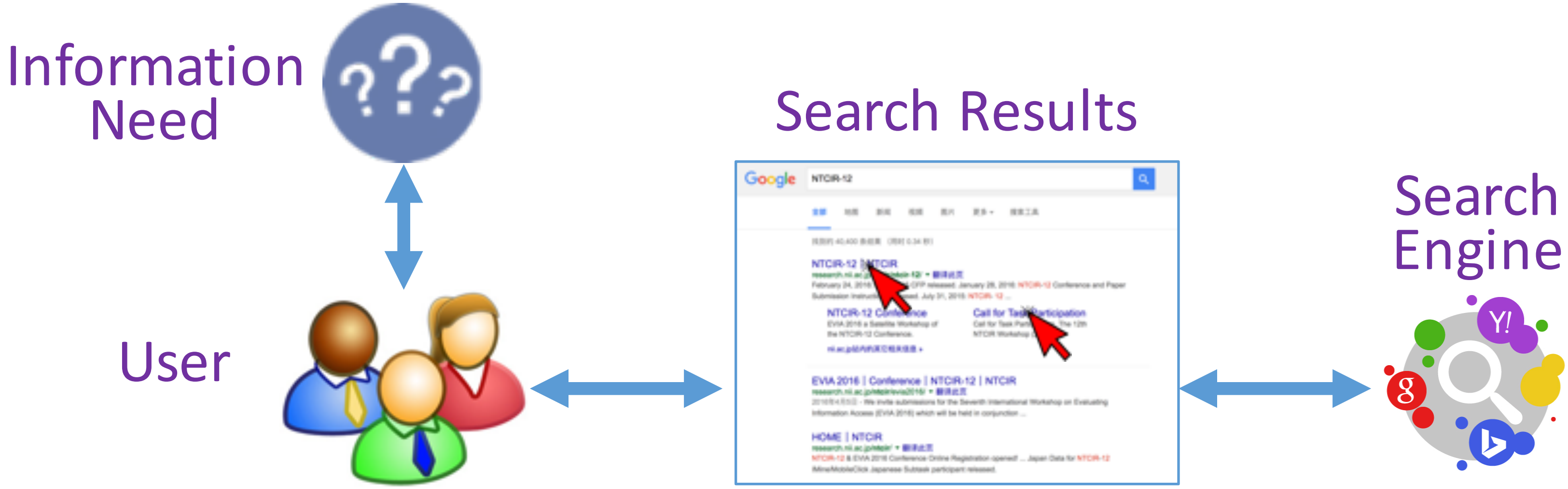
Tsinghua University, China

What’s the Gold Standard in Web Search



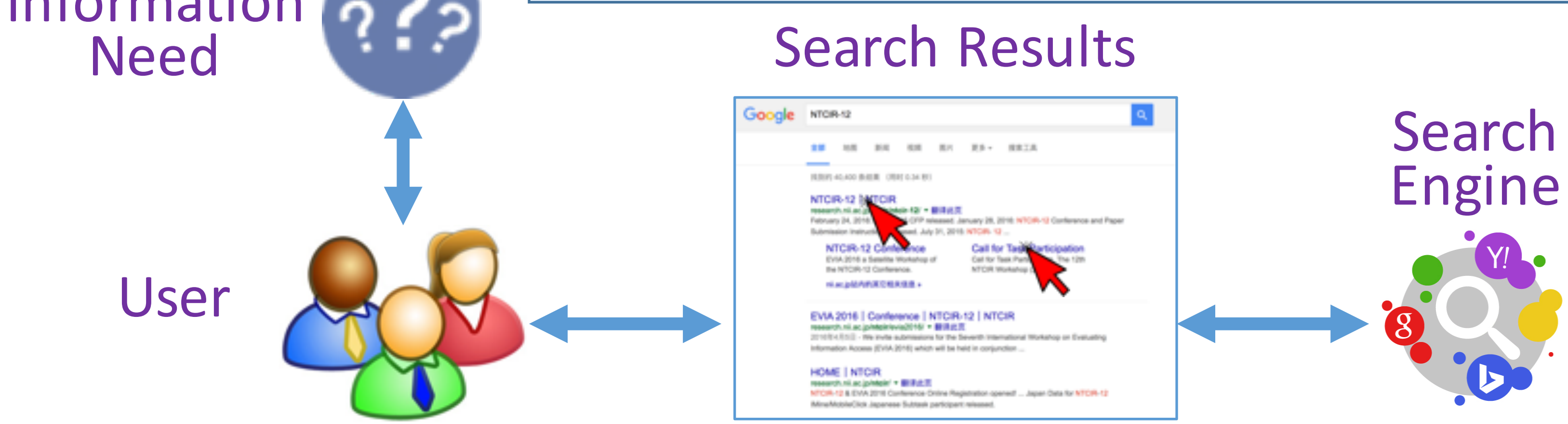
What's the Gold Standard in Web Search

- Is the information need SATISFIED OR NOT?
  - Questionnaire, Quiz, Concept Map (Egusa et. al., 2010), etc.
  - Problem: Efforts? User Experiences?



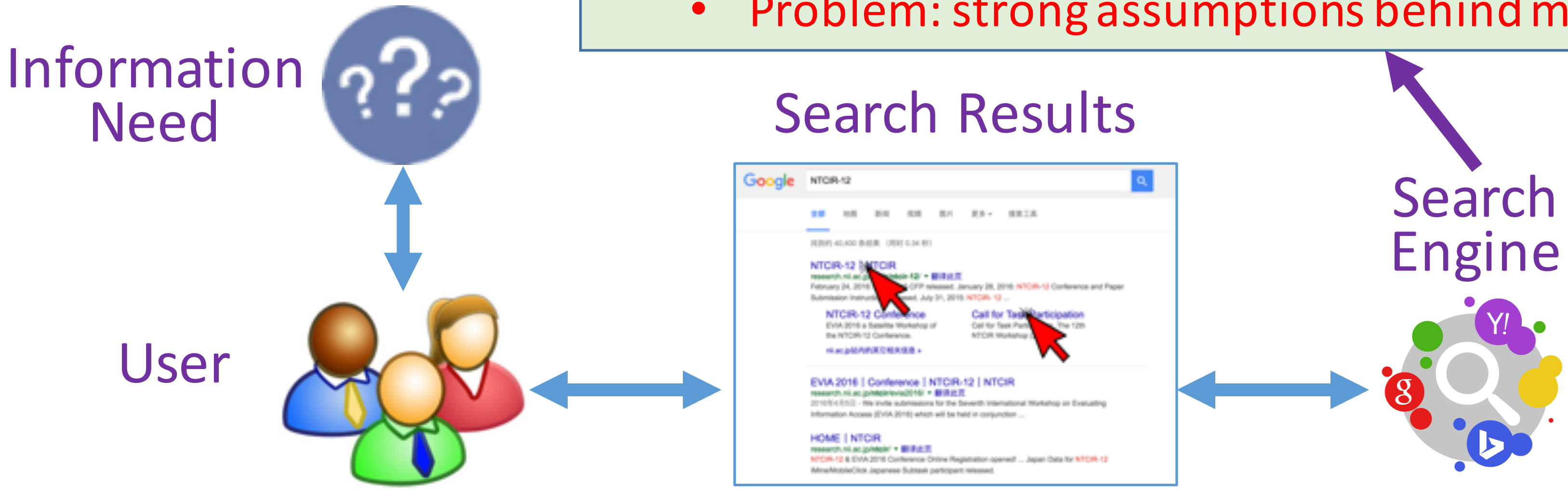
What's the Gold Standard in Web Search

- Are results RELEVANT WITH the user query?
  - Cranfield-like approach, Relevance judgement, evaluation metrics (nDCG, ERR, TBG, etc.)
  - Problem: behavior assumptions behind metrics

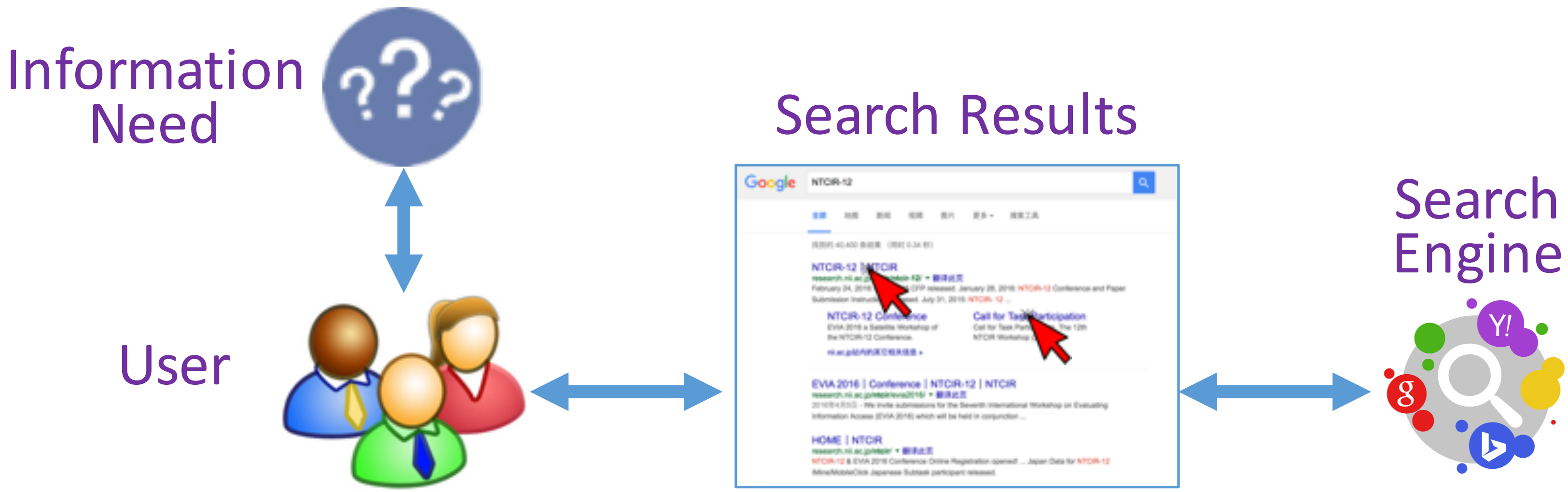


What’s the Gold Standard in Web Search

- Can we keep the boss HAPPY?
  - Various on-line metrics: CTR, SAT Click, interleaving, etc.
  - Problem: strong assumptions behind metrics



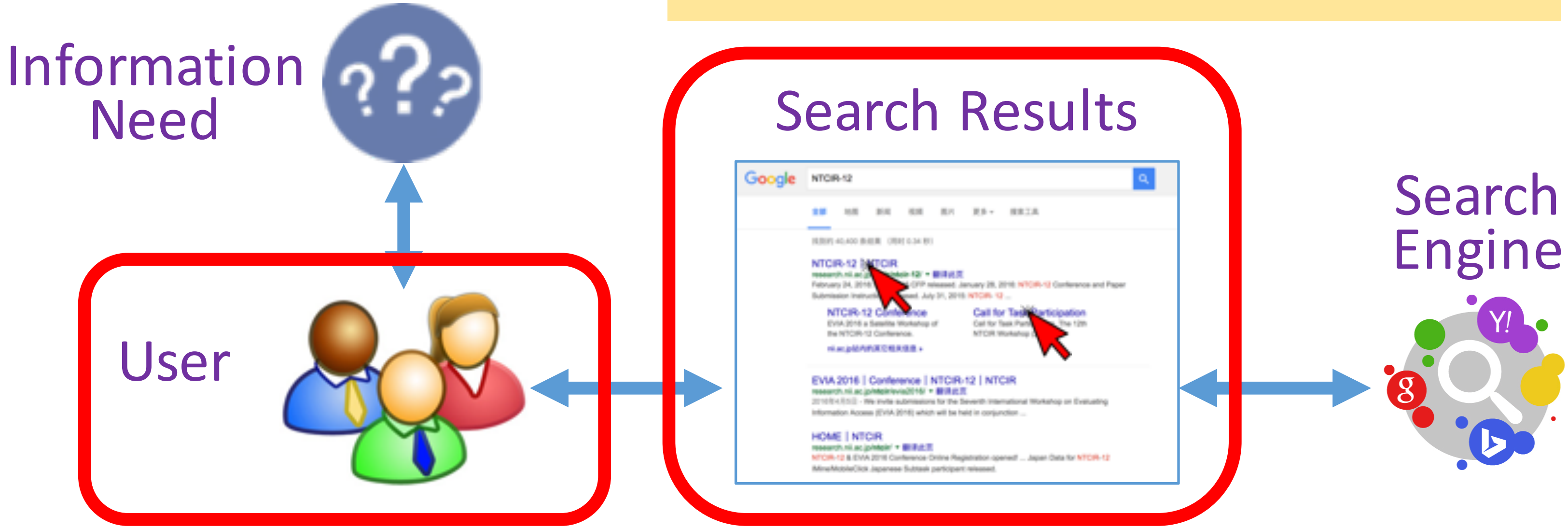
What’s the Gold Standard in Web Search



- Is the user SATISFIED OR NOT?
  - Post-search questionnaire; annotation by assessors (Huffman et. al., 2007)
  - Implicit feedback signals: satisfaction prediction (Jiang et. al., 2015)
  - Physiological signals: skin conductance response (SCR), facial muscle movement (EMG-CS) (Ángeles et. al., 2015).

Satisfaction Perception of Search User

RQ2: How heterogeneous results affect user satisfaction



RQ1: Satisfaction perception v.s. Relevance judgment

RQ3: Satisfaction prediction with interaction features

Outline

- ***Satisfaction v.s. Relevance judgment***
  Can we use relevance scores to infer satisfaction?
- ***Satisfaction v.s. Heterogeneous results***
  Do vertical results help improve user satisfaction?
- ***Satisfaction v.s. User interaction***
  Can we predict satisfaction with implicit signals?

Relevance

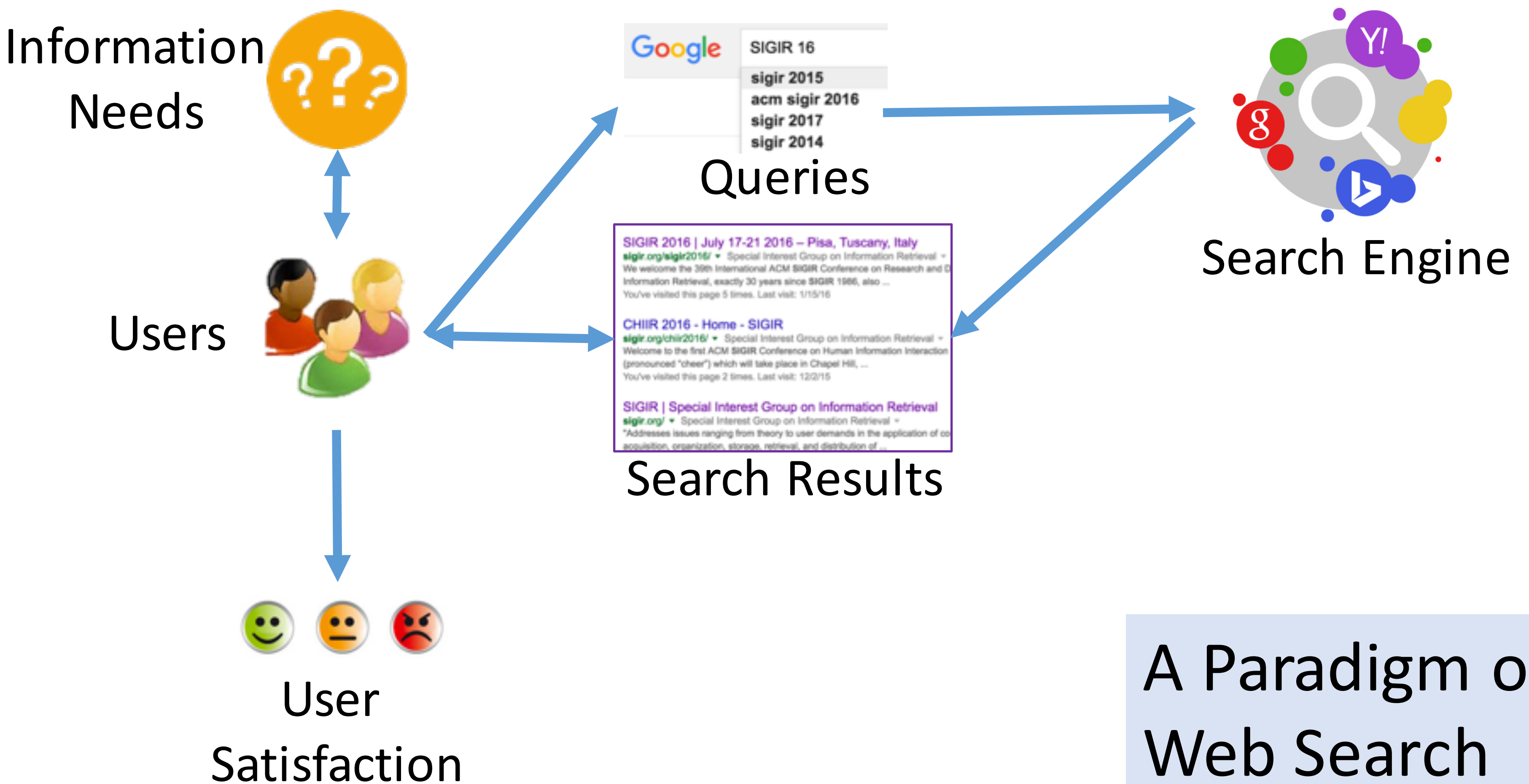
- **A central concept in information retrieval (IR)**

Tefko Saracevic
Former president of ASIS
SIGIR Gerard Salton Award in 1997
ASIS Award of Merit in 1995

"It (relevance) expresses a criterion for assessing *effectiveness in retrieval of information*, or to be more precise, of *objects* (texts, images, sounds ...) *potentially conveying information*."
[Saracevic, 1996]

Relevance judgment in Web search

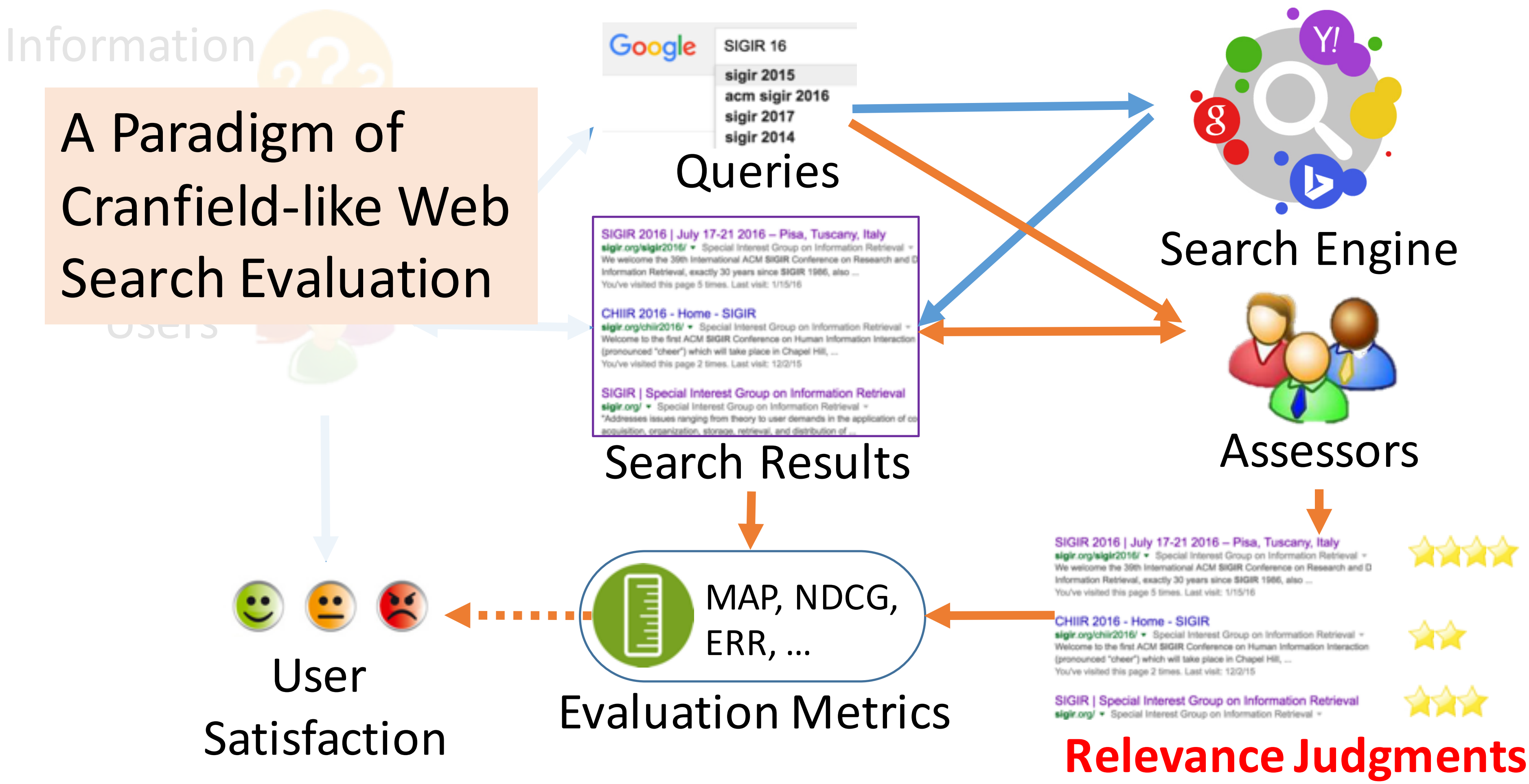
- **The role of *Relevance* in IR evaluation**



A Paradigm of Web Search

Relevance judgment in Web search

- **The role of *Relevance* in IR evaluation**



Relevance judgment in Web search

Idea (first-tier annotation):
Relevance is expected to represent users' opinions about whether a retrieved document **meet their needs** [Voorhees and Harman, 2001].

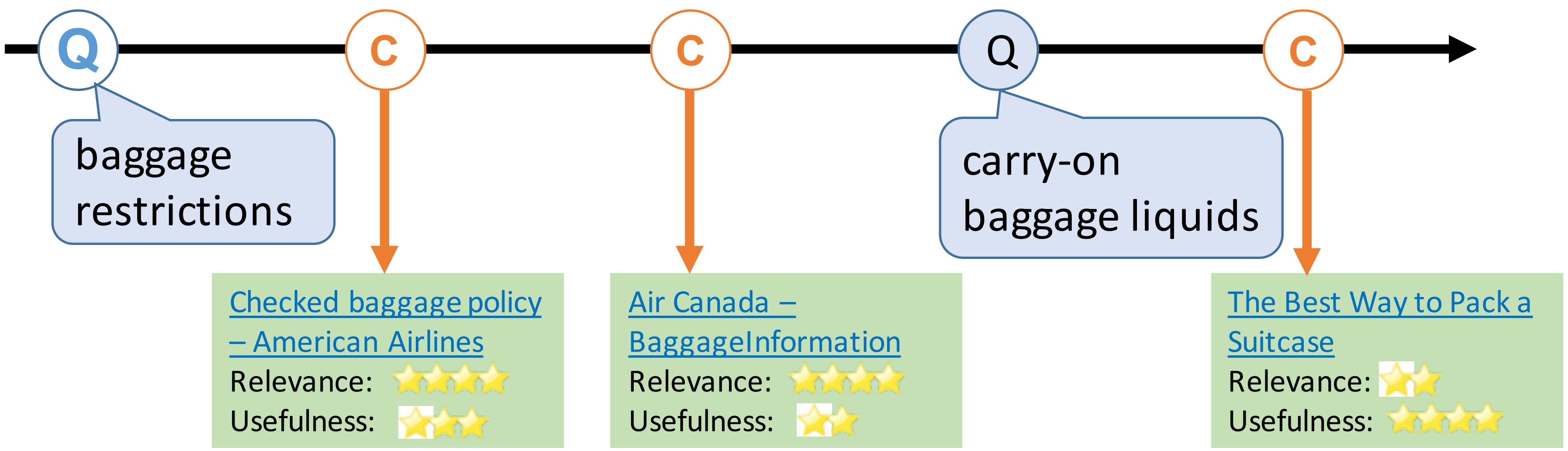
Practice (second-tier annotation):
Relevance is made by external assessors who do not:

- originate or fully understand the **information needs**
- have access to **search context**

Relevance judgments are often limited to the topical aspect, and different from **user-perceived usefulness**.

Example: Relevance v.s. Usefulness

You are going to US by air and want to know restrictions for both checked and carry-on baggage during air travel.



Relevance judgments ≠ perceived usefulness

Research Questions

Satisfaction

- Gold standard
- User feedback
- Query or session level

Relevance

- Assessor annotated
- W/o session context
- Document level (query-doc pair)

Usefulness

- User feedback
- With session context
- Document level (information need v.s. doc)

Research Questions

- **RQ1.1 Difference between annotated relevance and perceived usefulness**

Satisfaction

- Gold standard
- User feedback
- Query or session level

Relevance ⟷ ? ⟷ **Usefulness**

Relevance:

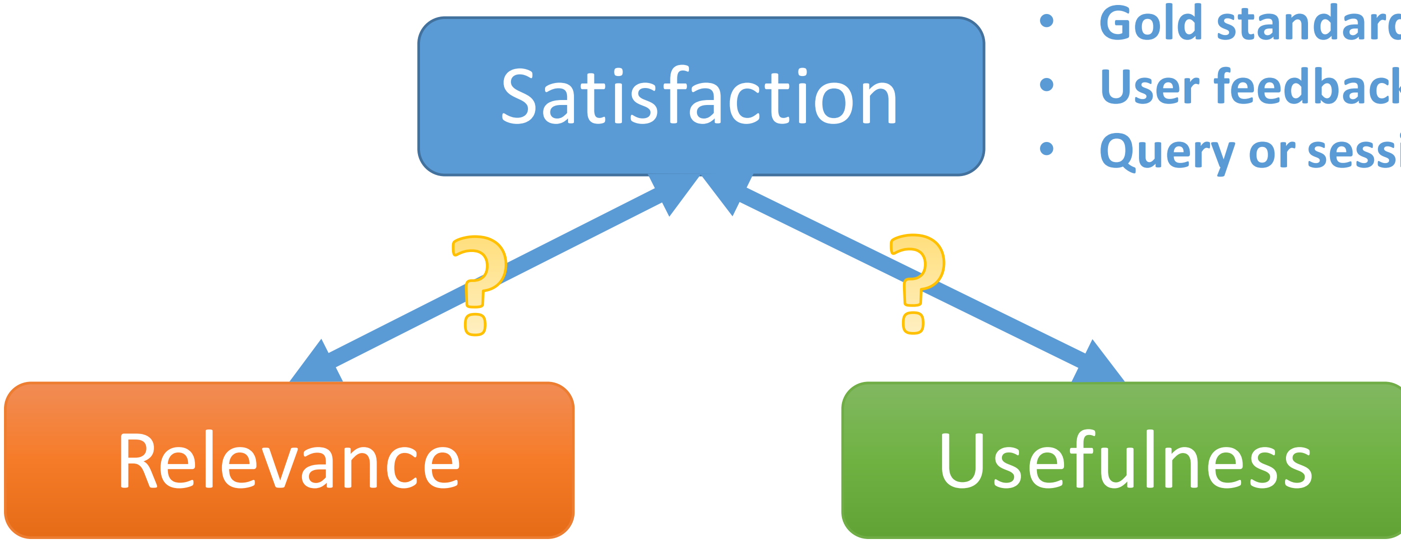
- Assessor annotated
- W/o session context
- Document level (query-doc pair)

Usefulness:

- User feedback
- With session context
- Document level (information need v.s. doc)

Research Questions

- **RQ1.2 Correlation relations between satisfaction and relevance/usefulness**



- Gold standard
- User feedback
- Query or session level

Relevance:

- Assessor annotated
- W/o session context
- Document level (query-doc pair)

Usefulness:

- User feedback
- With session context
- Document level (information need v.s. doc)

Research Questions

- **RQ1.3 Can perceived usefulness be annotated by external assessors?**

Satisfaction

- Gold standard
- User feedback
- Query or session level

Relevance

- Assessor annotated
- W/o session context
- Document level (query-doc pair)

Usefulness

- Assessor annotated
- With session context
- Document level (information need v.s. doc)

?

Research Questions

- **RQ1.4 Can perceived usefulness be predicted with relevance judgment?**

Satisfaction

- Gold standard
- User feedback
- Query or session level

Relevance

- Assessor annotated
- W/o session context
- Document level (query-doc pair)

Usefulness

- Automatic Prediction ?
- With session context
- Document level (information need v.s. doc)

Collecting Data

- **I. User Study:**
  - 29 participants
    - 15 female, 14 male
    - Undergraduate students from different majors
  - 12 search tasks
    - From TREC session track
  - Collect:
    - Users’ behavior logs
    - Users’ explicit feedbacks for usefulness and satisfaction

- **II. Data Annotation:**
  - 24 assessors
    - Graduate or senior undergraduate students
    - 9 assessors assigned to label document relevance
    - 15 assessors assigned to label usefulness and satisfaction
  - Collect:
    - Relevance annotations
    - Usefulness annotations
    - Satisfaction annotations

User Study Process

- I.1 Pre-experiment Training
- I.2 Task Description Reading and Rehearsal
- I.3 Task Completion with the Experimental Search Engine
- I.4 Satisfaction and Usefulness Feedback
- I.5 Post-experiment Question

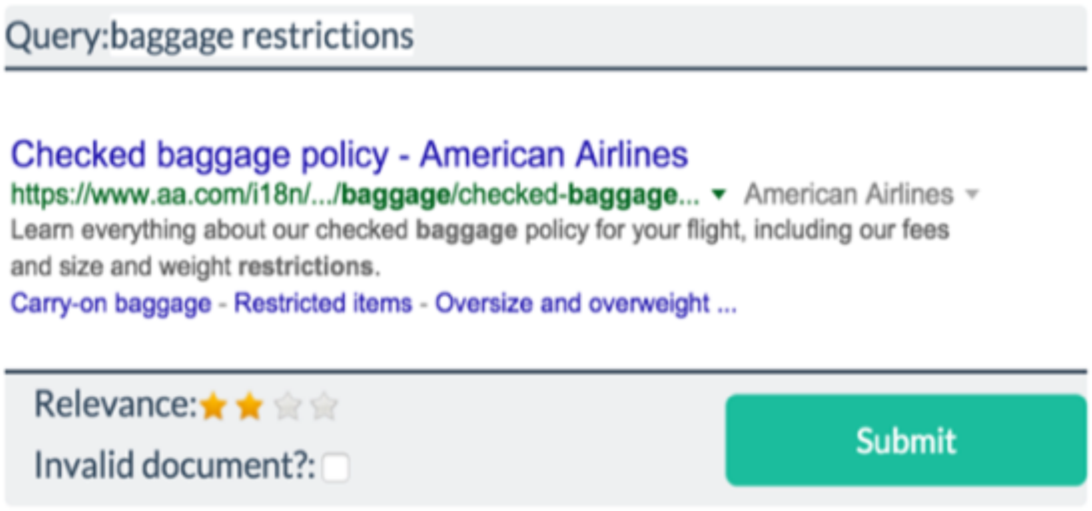
Query-level satisfaction feedbacks: $QSAT_u$

Usefulness feedbacks: U_u

We also collect task-level satisfaction feedbacks: $TSAT_u$

Data Annotation Process

- **Relevance annotation** (R)
  - Four-level relevance score
  - For all clicked documents and top-5 documents
  - Only query and document are shown to assessors
  - Each query-doc pair is judged by 3 assessors



Data Annotation Process

- **Usefulness and satisfaction annotations**
  - Each search session is judged by 3 assessors

Annotation Instructions:

Search Task: You are going to US by air, so you want to know what restrictions there are for both checked and carry-on baggage during air travel.

The left part shows the issued queries and clicked documents when a user is doing the search task via a search engine, you need to complete the following 3-step annotation:

STEP1: Annotate the usefulness of each clicked document for accomplishing the search task:

1 star: Not useful at all;
2 stars: Somewhat useful;
3 stars: Fairly useful;
4 stars: Very useful.

STEP2: Annotate query-level satisfaction for each query

(1 star: Most unsatisfied - 5 stars: Most satisfied)

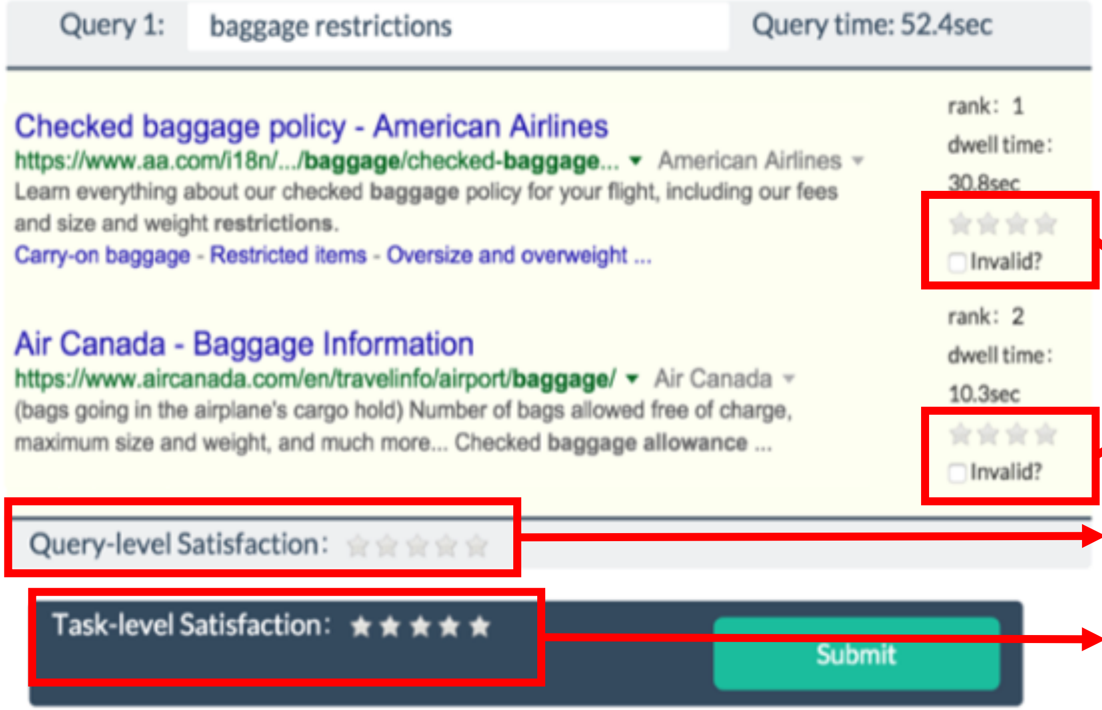
STEP3: Finally, please annotate the task-level satisfaction

(1 star: Most unsatisfied - 5 stars: Most satisfied)

Completed units/all units ： 0/29

II. Data Annotation

- **Usefulness and satisfaction annotations**
  - Each search session is judged by 3 assessors



4-level usefulness annotation: U_a

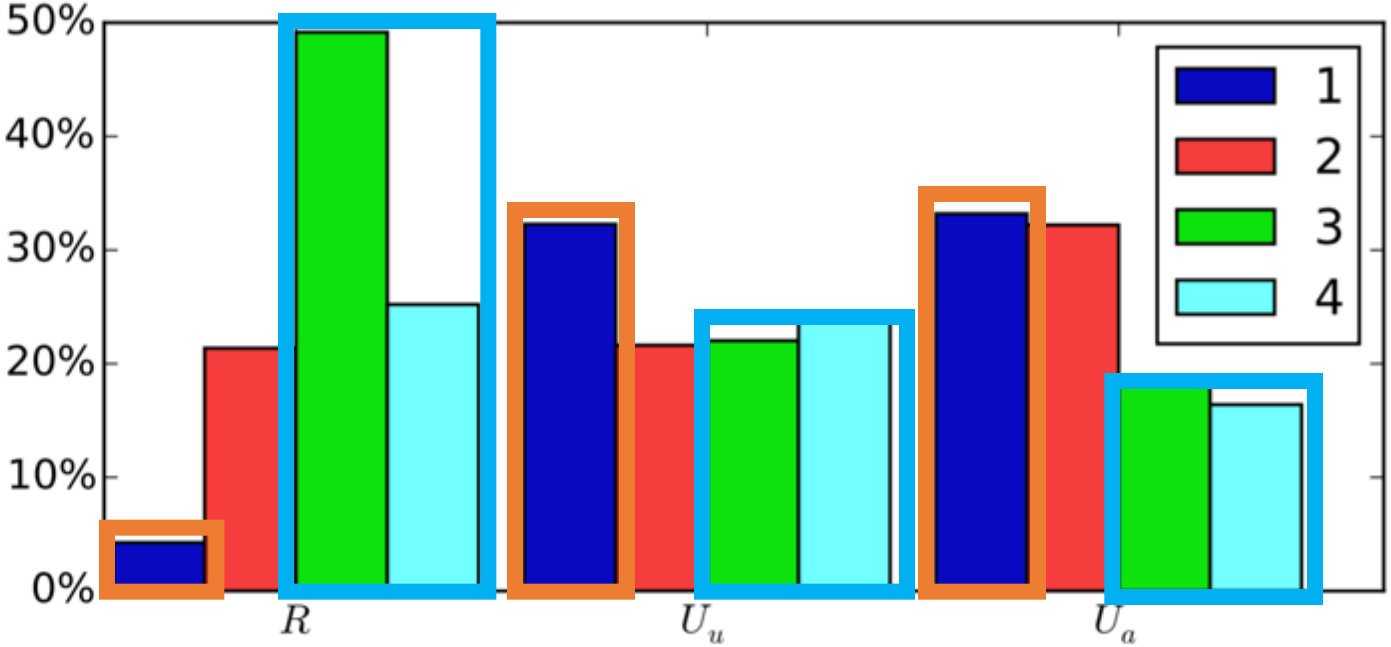
5-level query satisfaction annotation: $QSAT_a$

5-level task satisfaction annotation: $TSAT_a$

RQ1.1. Usefulness v.s. Relevance

- **Relevance (assessor, *R*) / Usefulness (user, U_u) / Usefulness (assessor, U_a)**

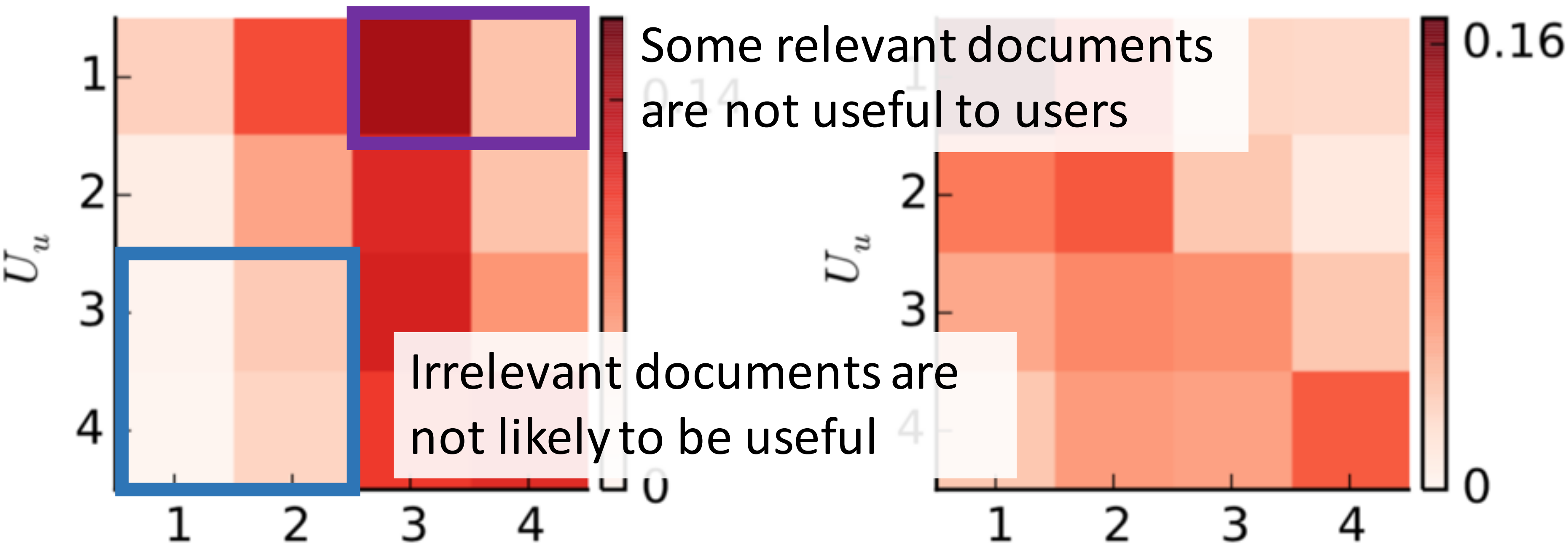
Finding #2: A large part of docs are relevant, much fewer are useful



Finding#1 : Only a few docs are not relevant, much more are not useful

RQ1.1. Usefulness vs. Relevance

- **Joint distribution of *R*, U_u and U_a**
  - Positive correlation (Pearson's r: 0.332, Weighted κ: 0.209) between *R* and U_u



Finding: Relevance is **necessary but not sufficient** for usefulness

RQ1.2. Correlation with Satisfaction

- **Correlation with query-level satisfaction *$QSAT_u$***
  - Offline metrics (based on relevance annotation *R*)
    - Results are ranked by original positions
    - MAP@5, DCG@5, ERR@5, weighted relevance
  - Online metrics (based on *R* or usefulness *U_u*)
    - Results are ranked by click behavior sequences

$$cCG(CS,M) = \sum_{i=1}^{|CS|} M(d_i) \qquad cDCG(CS,M) = \sum_{i=1}^{|CS|} \frac{M(d_i)}{log_2(i+1)}$$

$$cMAX(CS,M) = \max(M(d_1), M(d_2), \ldots, M(d_{|CS|}))$$

RQ1.2. Correlation with Satisfaction

- **Correlation with query-level satisfaction $QSAT_u$**

All correlations (measured in Pearson's r) are significant at $p < 0.001$. *(or **) indicates the difference is significant at $p < 0.05 (p < 0.01)$, comparing to the same metric based on relevance annotation R

	All Queries ($df = 933$)		Queries with only top 5 clicks ($df = 635$)	
	U_u	R	U_u	R
cCG	0.572**	0.425	0.647**	0.499
cDCG	0.724**	0.498	0.747**	0.535
cMAX	0.751**	0.563	0.759**	0.599
[illegible]	0.733**	0.551	0.751**	0.587
MAP@5	-	0.192	-	0.255
DCG@5	-	0.295	-	0.363
ERR@5	-	0.258	-	0.332
Weighted Rel. [20]	-	0.229	-	0.273

Metrics based on U_u correlate better with $QSAT_u$ than R.

Click sequence based metrics are better than rank based ones

RQ1.2. Correlation with Satisfaction

- **Correlation with task-level satisfaction $TSAT_u$**
  - Online metrics (based on R or usefulness U_u)

$$sCG(M) = \sum_{j=1}^{n} gain(q_j) = \sum_{j=1}^{n} cCG(CS_j, M)$$

$$sDCG(M) = \sum_{j=1}^{n} \frac{gain(q_j)}{1 + log(j)} = \sum_{j=1}^{n} \frac{cCG(CS_j, M)}{1 + log(j)}$$

	U_u	R
sCG	0.110	-0.046
sCG/#query	0.437	0.330
sCG/#click	0.525	0.320
sDCG	0.317	0.142

Metrics based on U_u correlate better with $TSAT_u$ than R.

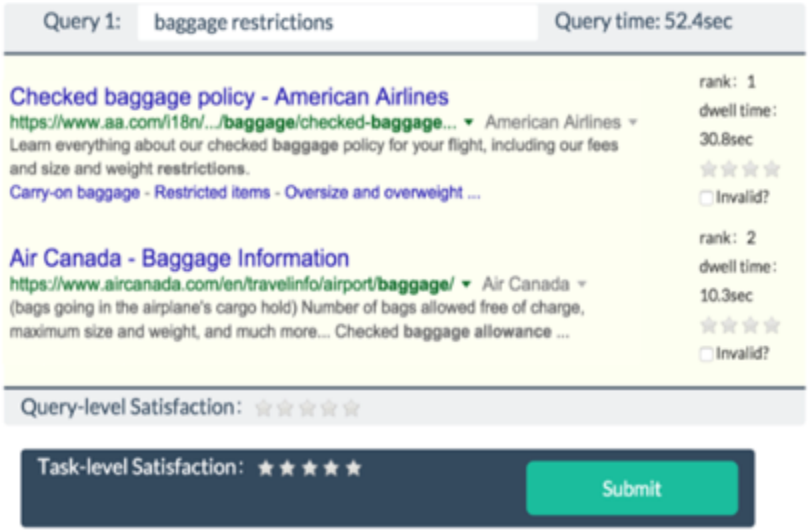
RQ1.2. Major Findings

1. Metrics based on **usefulness feedbacks** are strongly correlated with $QSAT_u$ and moderately correlated with $TSAT_u$
2. The **click-sequence-based metrics** correlates better with satisfaction than the rank-position-based ones
3. **Usefulness** has a stronger correlation with satisfaction than **relevance** in all metrics

RQ 1.3. Collecting Usefulness Labels

- **NOT practical to collect usefulness labels from users => collected from external assessors?**
  - An augmented search log for assessors

Annotation Instructions:
Search Task: You are going to US by air, so you want to know what restrictions there are for both checked and carry-on baggage during air travel.
The left part shows the issued queries and clicked documents when a user is doing the search task via a search engine, you need to complete the following 3-step annotation:
STEP1: Annotate the usefulness of each clicked document for accomplishing the search task:
1 star: Not useful at all;
2 stars: Somewhat useful;
3 stars: Fairly useful;
4 stars: Very useful.
STEP2: Annotate query-level satisfaction for each query
(1 star: Most unsatisfied - 5 stars: Most satisfied)
STEP3: Finally, please annotate the task-level satisfaction
(1 star: Most unsatisfied - 5 stars: Most satisfied)
Completed units/all units : 0/29



	R_{nc}	R_c	U_a	$QSAT_a$	$TSAT_a$
#Annotations	1,944	1,161	1,512	935	225
Weighted κ	0.344	0.413	0.530	0.535	0.274

RQ 1.3. Collecting Usefulness Labels

- **Comparing U_a and U_u; $QSAT_u$ and $QSAT_a$**
  - Gold standard: satisfaction annotated by user, $QSAT_u$

Finding #2: U_a is not as good as user feedback, but still better than R

	Pearson's $r(df = 933)$			Pref. agreement ratio		
	U_a	U_u	R	U_a	U_u	R
cCG	$.466^{\blacktriangledown/*}$	.572	.425	$.701^{\blacktriangledown/**}$	.751	.669
$cDCG$	$.518^{\blacktriangledown/*}$	.724	.498	$.742^{\blacktriangledown/**}$	.826	.698
$cMAX$	$.580^{\blacktriangledown/*}$	.751	.563	$.681^{\blacktriangledown/**}$	.779	.632
$cCG/\#clicks$	$.548^{\blacktriangledown}$	.733	.551	$.716^{\blacktriangledown/*}$	.807	.689
$QSAT_a$	.508			.584		

Finding #1: Satisfaction annotation is not as good as metrics with U_a

RQ 1.4. Predicting Usefulness Labels

- **Prediction method: user behavior signals**
  - Search context and behavior Features: Query features (*Q*); Session features (*S*); User features (*U*)
  - Annotations: Metrics based on relevance annotation (*R*) or Usefulness annotation (*A*)

Query features(Q)	
rank	The rank of clicked document in result list
#clicks	The number of clicks in the query
query length	The length of the query, in words and in characters
click position	Whether the click is the first/last/intermediate click in a query with more than one click, and whether the query has only one click
dwell time	click dwell time and query dwell time
Session features(S)	
#queries	The number of queries in the search session
#queries w/o click	The number of queries without click in session
query position	Whether the query is the first/last/intermediate query in a session with more than one query, and whether the session has only one query
time to completion	The total time spent on this search session
query reformulation	Whether the query is generated from a specification/ generalization/ parallel reformulation, and whether the query leads to a specification/ generalization/ parallel reformulation
User features(U)	
user #clicks	The average/max/min/standard deviation of #clicks per query of the user
user #queries	The average/max/min/standard deviation of #queries per session of the user
user #dwell time	The average/max/min/standard deviation of query/click dwell time of the user

RQ 1.4. Predicting Usefulness Labels

- **Results: with user feedback U_u as gold standard**

Finding #2: Search context and behavior features can help enhance assessors' annotations, especially the relevance annotation *R*

	Pearson's *r*	*MSE*	*MAE*
U_Q	0.398^{*}	1.198^{**}	0.894^{**}
U_{Q+S}	0.410^{**}	1.186^{**}	0.889^{**}
U_{All}	0.461^{**}	1.103^{**}	0.851^{**}
U_{All+A}	0.467^{**}	1.105^{**}	0.845^{**}
U_{All+R}	0.519^{**}	1.021^{**}	0.815^{**}
$U_{All+A+R}$	0.521^{**}	1.023^{**}	0.803^{**}
U_a	0.413	1.512	0.852
R	0.332	1.786	1.020

Finding #1: Prediction results U_{All} is comparable or better than U_a and *R*

RQ 1.4. Predicting Usefulness Labels

- **Results: for prediction of user satisfaction**

Finding #3: Context and behavior features can improve annotations.

Finding #4: Metrics based on predicted usefulness are better than direct prediction or users' direct annotation of satisfaction

	U_{All}	$U_{\mathrm{All+A+R}}$	U_a	U_u
cCG	$0.459^{\blacktriangledown}$	$0.490^{**/\blacktriangledown}$	0.466	0.572
cDCG	$0.580^{**/\blacktriangledown}$	$0.612^{**/\blacktriangledown}$	0.518	0.724
cMAX	$0.601^{\blacktriangledown}$	$0.635^{**/\blacktriangledown}$	0.580	0.751
cCG/#clicks	$0.571^{\blacktriangledown}$	$0.608^{**/\blacktriangledown}$	0.548	0.733
$QSAT_a$	0.508			
Jiang et al.	0.539			

Finding #1: Prediction results are not as good as users' feedback

Finding #2: Prediction results are better than assessors' annotations

Take-Home Messages

- **Why should we use usefulness labels**
  - Relevance is necessary but not sufficient for usefulness
  - Click-sequence-based metrics with usefulness scores strongly correlate with user satisfaction
  - Usefulness annotation is more consistent than relevance annotation among assessors
- **How to collect usefulness labels:**
  - External assessors can make reliable and valid usefulness labels when context information is provided
  - We can automatically generate valid usefulness labels

Limitations and Discussions

- **Relevance annotation cannot be replaced with usefulness annotation**
  - Reusability: usefulness annotation cannot be reused to evaluate previously unseen systems
  - Efficiency: more information and more effort is required for usefulness annotation
- **A possible evaluation paradigm**
  - Generating usefulness scores with relevance judgment and context/behavior information
  - Evaluation results with click-sequence-based metrics

Outline

- *Satisfaction v.s. Relevance judgment*
  Can we use relevance scores to infer satisfaction?
- ***Satisfaction v.s. Heterogeneous results***
  Do vertical results help improve user satisfaction?
- ***Satisfaction v.s. User interaction***
  Can we predict satisfaction with implicit signals?

Heterogeneous Search Results

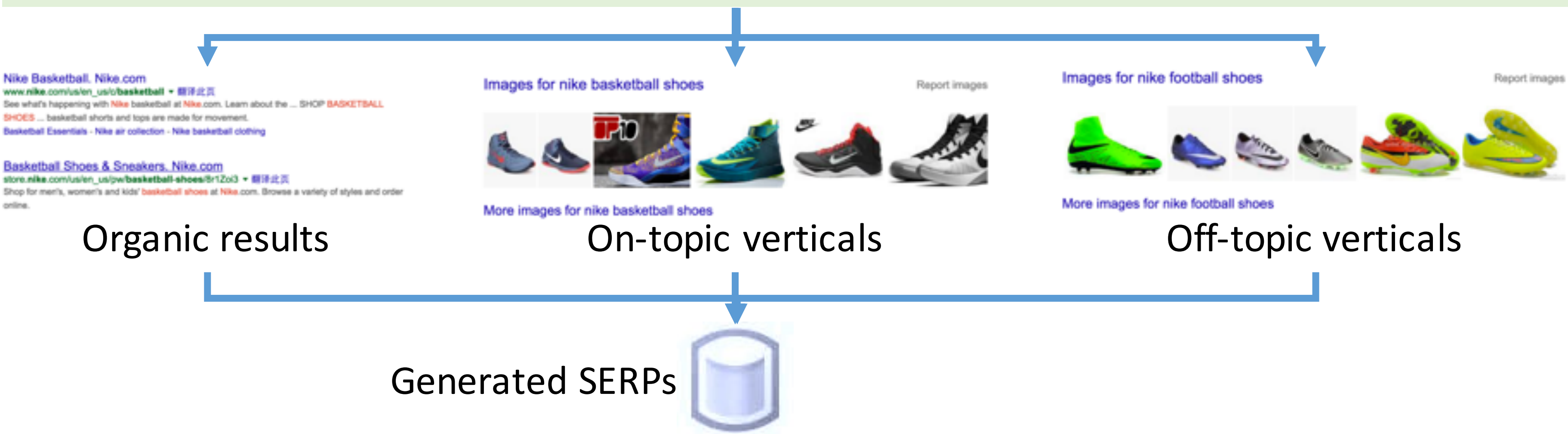
- Vertical results are everywhere (over 80% SERPs)

RQ2: How do vertical results affect users' search satisfaction?

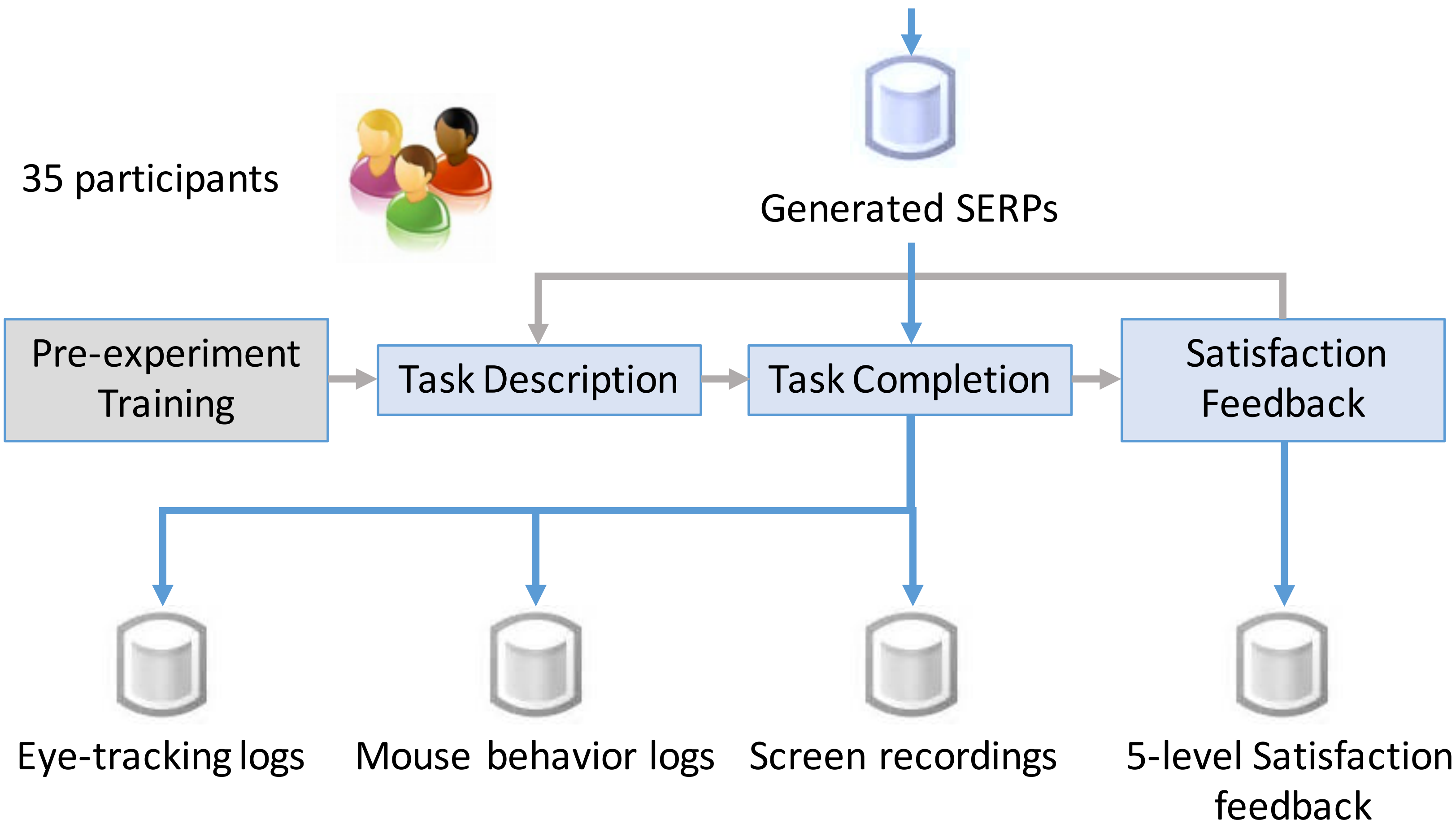
User study: SERP Preparation

User study: SERP Preparation

- **Controlled Variables:**
  - **Vertical relevance**: on-topic or off-topic
  - **Presentation style**: Textual, Encyclopedia, Image, Download, and News
  - **Presentation position**: rank 1, 3, 5, and without vertical

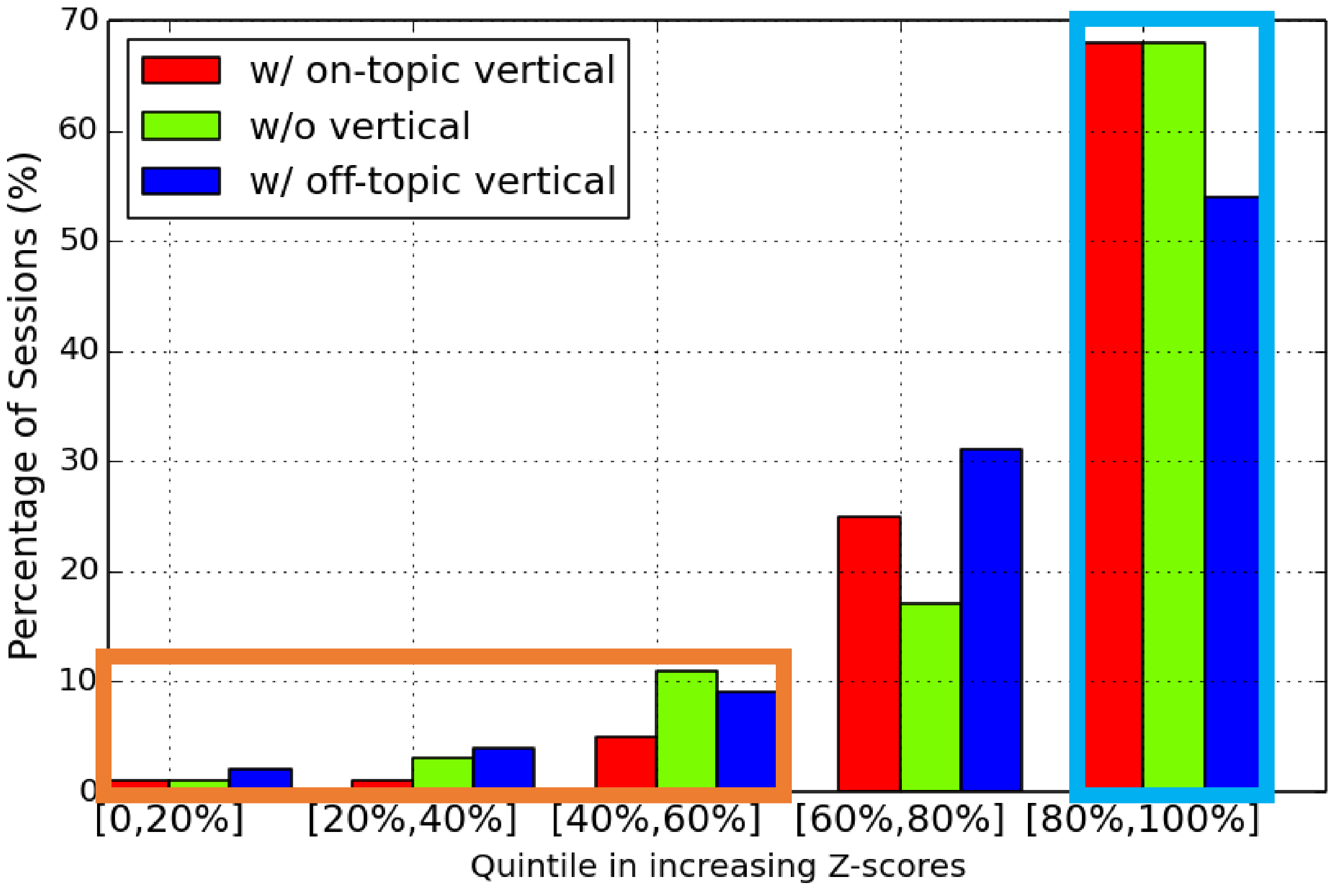


User study: Procedure and Data Collecting



Results: Effect of Vertical Relevance

Finding #1: Users are less satisfied with SERPs with off-topic verticals



Finding #2: users are less likely to be unsatisfied with on-topic verticals

Results: Effect of Presentation Style

Finding #1: *Some kinds of* on-topic verticals help improve satisfaction

Finding #2: *Some kinds of* off-topic verticals hurt user satisfaction

	w/o vertical	w/ on-topic vertical	w/ off-topic vertical	on-off difference
Users' Satisfaction Feedback				
Textual	5.15	5.10 (-0.05)	4.95 **(-0.20**)**	+0.15*
Image & Textual	4.46	4.99 **(+0.53**)**	4.67 (+0.21)	+0.32**
Image	5.17	5.07 (-0.10)	4.58 **(-0.59**)**	+0.49**
Download	4.75	5.25 **(+0.50**)**	4.60 (-0.15)	+0.65**
News	4.43	4.34 (-0.09)	4.38 (-0.05)	-0.04

Finding #3: News verticals have no strong impact in user satifaction

Results: Effect of Result Position

Finding #1: On-topic verticals ranked at 1st help improve satisfaction

Finding #2: Off-topic verticals ranked at 1st hurt user satisfaction

	w/o vertical	w/ on-topic vertical	w/ off-topic vertical	on-off difference
Users' Satisfaction Feedback				
Rank 1	4.79	5.06 **(+0.27**)**	4.43 **(-0.36**)**	+0.63**
Rank 3	4.79	4.93 (+0.14)	4.63 (-0.16)	+0.29**
Rank 5	4.79	4.87 **(+0.08*)**	4.85 (+0.06)	+0.02

Finding #3: Lower-ranked verticals have no strong impact in user satisfaction

Take-Home Messages

- **Vertical results will affect users' satisfaction**
  - On-topic Encyclopedia and Download verticals will bring more satisfaction to users
  - Relevant Image verticals have limited positive effect, and irrelevant Image verticals bring negative influence to satisfaction
  - News verticals have no significant effect on satisfaction
  - Vertical results have larger effect when presented at higher positions

Outline

- ***Satisfaction v.s. Relevance judgment***
  Can we use relevance scores to infer satisfaction?

- ***Satisfaction v.s. Heterogeneous results***
  Do vertical results help improve user satisfaction?

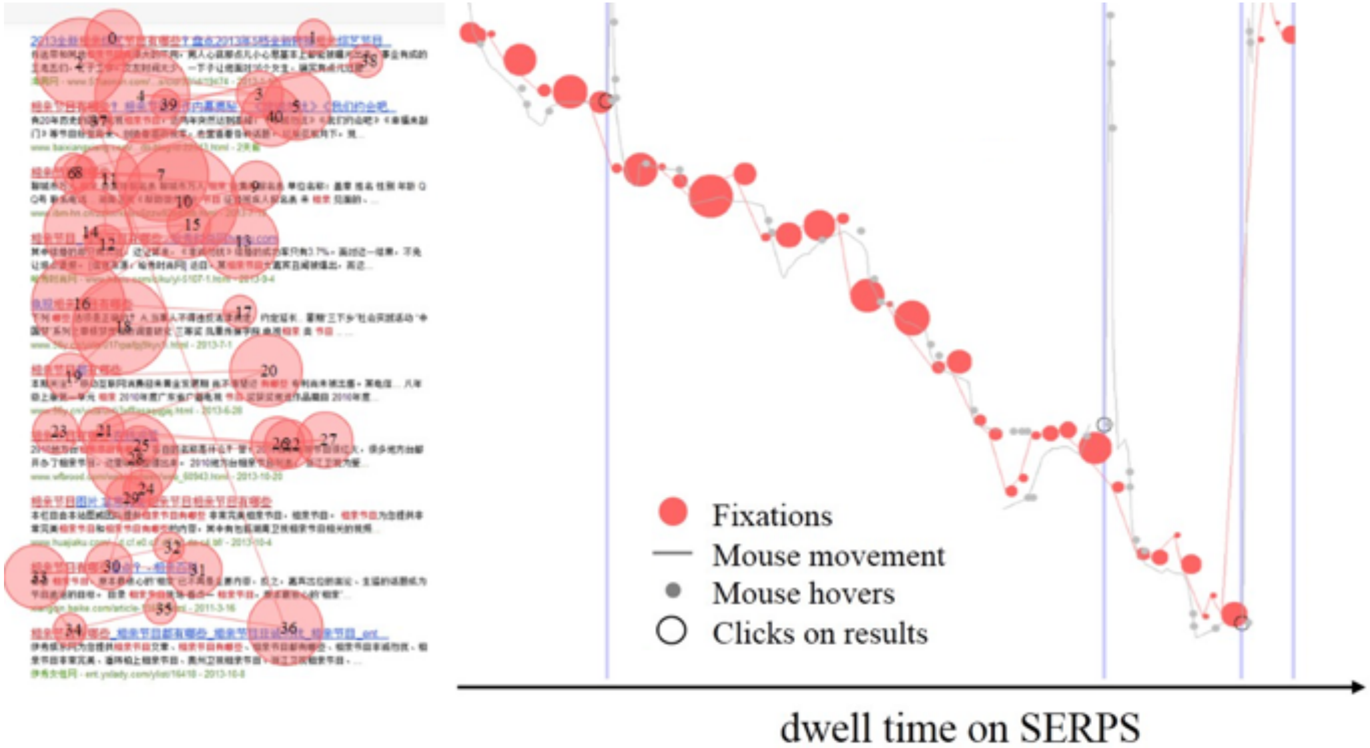
- ***Satisfaction v.s. User interaction***
  Can we predict satisfaction with implicit signals?

Satisfaction Prediction

- **Based on coarse-grained features**
  - Click-through on SERP components [Guo et. al, 2010]
- **Based on fine-grained features**
  - Cursor positions, scrolling speeds, mouse hovers, etc. [Guo et al., 2012]
- **Based on benefit-cost framework**
  - Benefit: information gain measured by NDCG, MAP, etc.
  - Cost: time/effort spent. [Jiang et al., 2015]
- **RQ1.4: satisfaction prediction is possible with context, behavior signals and relevance judgment**

Satisfaction Prediction

- **A new information source: Mouse Movement**
  - Surrogate for eye-tracking data (Poor's eye tracker)
  - Applicable: Collected at a large scale with low cost



Motif Extraction

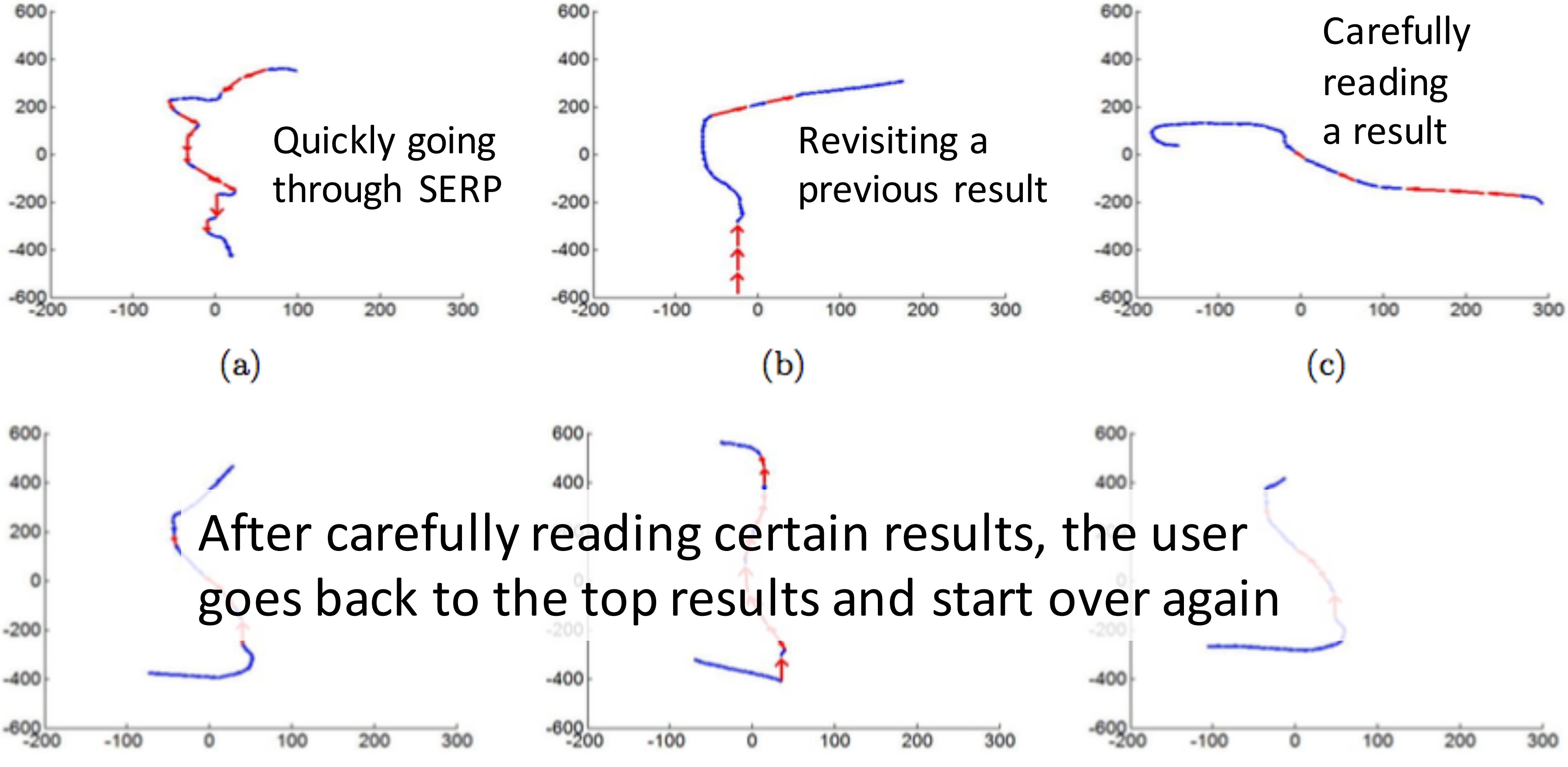
- **Motif: Frequently-appeared sequence of mouse positions [Lagun et al., 2014]**
  - Extraction of motifs from mouse data: sliding window + dynamic time wrapping [Sakoe and Chiba, 1978]

Satisfied User Session

Unsatisfied User Session

Motif Selection

- **Examples of predictive motifs**



(a) (b) (c) (d) (e) (f)

Satisfaction Prediction based on Motif

- **Prediction power of motifs across users/queries**

Finding #2: Motif information can be used to improve existing prediction frameworks which haven't used mouse movement info.

	Annotation & Sampling strategy	Guo et al. [3]	Jiang et al. [8]	motif	motif + Guo et al. [3]	motif + Jiang et al. [8]
Multi-vertical tasks	User annotation & random sample	0.855	0.828	0.853	0.864 (+1.05%**)	0.858 (+3.62%**)
	User annotation & sample by user	0.848	0.821	0.830	0.866 (+2.12%**)	0.852 (+3.78%**)
	User annotation & sample by query	0.848	0.801	0.826	0.861 (+1.53%)	0.842 (+5.12%**)
Single-vertical tasks	User annotation & random sample	0.668	0.643	0.693	0.706 (+5.8%**)	0.707 (+10.0%**)
	User annotation & sample by user	0.629	0.639	0.688	0.715 (+13.7%**)	0.690 (+8.0%**)
	User annotation & sample by query	0.685	0.637	0.709	0.714 (+4.2%)	0.712 (+11.8%**)

Finding #1: Motif feature works as good as other behavior features

Take-Home Messages

- **RQ1.** ***Satisfaction v.s. Relevance judgment***
  - A new evaluation paradigm based usefulness annotation/prediction may better represent user satisfaction (gold standard for Web search)
- **RQ2.** ***Satisfaction v.s. Heterogeneous results***
  - User satisfaction is affected by vertical results
- **RQ3.** ***Satisfaction v.s. User interaction***
  - User satisfaction can be predicted with implicit behavior features, e.g. mouse movement patterns

References

- (***RQ1***) Jiaxin Mao, *Yiqun Liu*, Ke Zhou, Jian-Yun Nie, et. al. When does Relevance Mean Usefulness and User Satisfaction in Web Search? **The 39th ACM SIGIR conference (SIGIR 2016)**
- **(*RQ2*)** Ye Chen, *Yiqun Liu*, Ke Zhou, et. al. Does Vertical Bring more Satisfaction? Predicting Search Satisfaction in a Heterogeneous Environment. **The 24th ACM CIKM conference (CIKM2015**)
- **(*RQ3*)** *Yiqun Liu*, Ye Chen, Jinhui Tang, Jiashen Sun, Min Zhang, Shaoping Ma, Xuan Zhu, Different users, Different Opinions: Predicting Search Satisfaction with Mouse Movement Information. **The 38th ACM SIGIR conference (SIGIR2015**)

- **Data/Codes are available on http://www.thuir.cn/group/~yqliu**

Thank you

Dataset is available for academic use:
Eye fixations, mouse movement features, clicks, relevance annotation, examination feedback, …

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