Learning From User Behavior

- Types of logs
- What do the data look like?
- What can we do with them?
  - Queries
  - Documents
  - Users

Modern search engines log *everything*.

- Query
    - Timestamp
    - IP Address (sometimes hashed)
    - User ID?
- Search results
- Click-through data
- Advertisements?

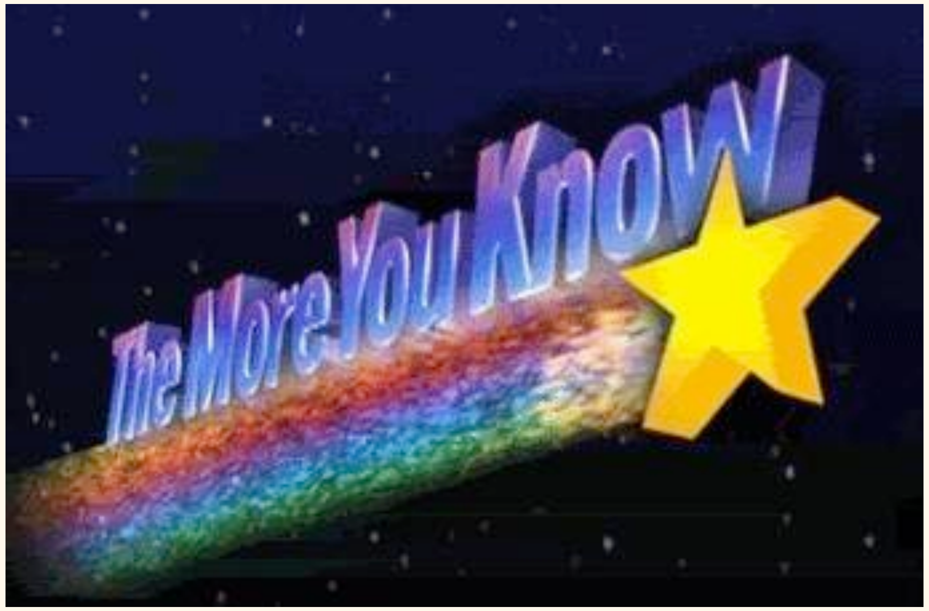
Your browser might also log:

What pages you visit

When you visit them

What you click on

Etc. etc. etc.



What can we learn from this information?

What people search for;

How they look for it;

What they do once they find it;

How they decide they’ve found it;

Where they go next;

Etc. etc. etc.

What can we learn from this information?

We can learn about three main families of things:

- Understanding queries
- Understanding documents
- Understanding users

And use that to guide our system's behavior!

Search log data

Table 1
Snippet from a Web search engine transaction log

User identification	Date	Time	Search_url
ce00160c04c4158087704275d69fbecd	25/Apr/2004	04:08:50	Sphagnum Moss Harvesting+New Jersey+Raking
38f04d74e651137587e9ba3f4f1af315	25/Apr/2004	04:08:50	emailanywhere
fabc953fe31996a0877732a1a970250a	25/Apr/2004	04:08:54	Tailpiece
5010dbbd750256bf4a2c3c77fb7f95c4	25/Apr/2004	04:08:54	1’personalities AND gender AND education’1
25/Apr/2004	**04:08:54**	**dmr panasonic**	
89bf2acc4b64e4570b89190f7694b301	25/Apr/2004	04:08:55	Bawdy poems
	“Mark Twain”	**25/Apr/2004**	
397e056655f01380cf181835dfc39426		**04:08:56**	**gay porn**
a9560248d1d8d7975ffc455fc921cdf6	25/Apr/2004	04:08:58	skin diagnostic
81347ea595323a15b18c08ba5167fbe3	25/Apr/2004	04:08:59	Pink Floyd CD label cover scans
3c5c399d3d7097d3d01aeea064305484	25/Apr/2004	04:09:00	freie stellen dangaard
9dafd20894b6d5f156846b56cd574f8d	25/Apr/2004	04:09:00	Moto.it
415154843dfe18f978ab6c63551f7c86	25/Apr/2004	04:09:00	Capability Maturity Model VS.
c03488704a64d981e263e3e8cf1211ef	25/Apr/2004	04:09:01	ana cleonides paulo fontoura

Note. Intentional errors are shown in boldface.

Jansen, Bernard J. "Search log analysis: What it is, what's been done, how to do it." *Library & information science research* 28.3 (2006): 407-432.

Search log data

Query	Count
facebook	3, 157 K
google	1, 796 K
youtube	1, 162 K
myspace	702 K
facebook com	665 K
yahoo	658 K
yahoo mail	486 K
yahoo com	486 K
ebay	486 K
facebook login	445 K

Fig. 3. An example of query histogram, which consists of queries and their frequencies.

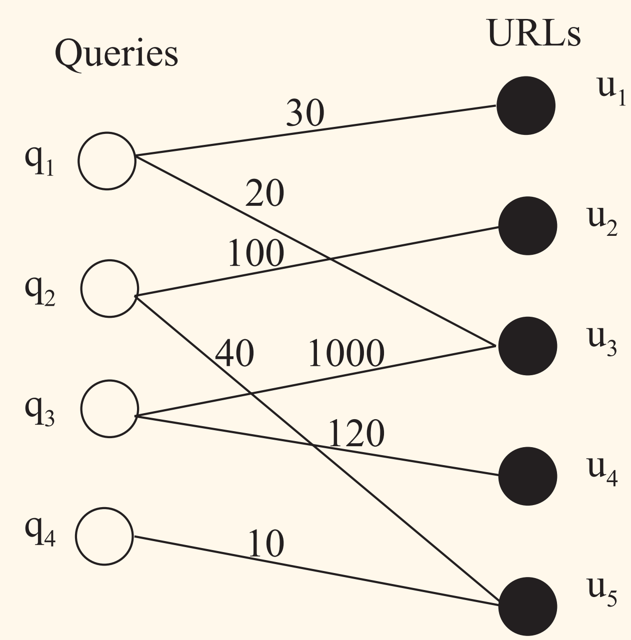


Fig. 4. An example of click-through bipartite graph. In a click-through bipartite graph, nodes represent queries and URLs, and edges represent click relations between queries and URLs.

Jiang D, Pei J, Li H. Mining Search and Browse Logs for Web Search: A Survey. ACM Trans Intell Syst Technol. 2013;4(4):57:1–57:37.

Search log data

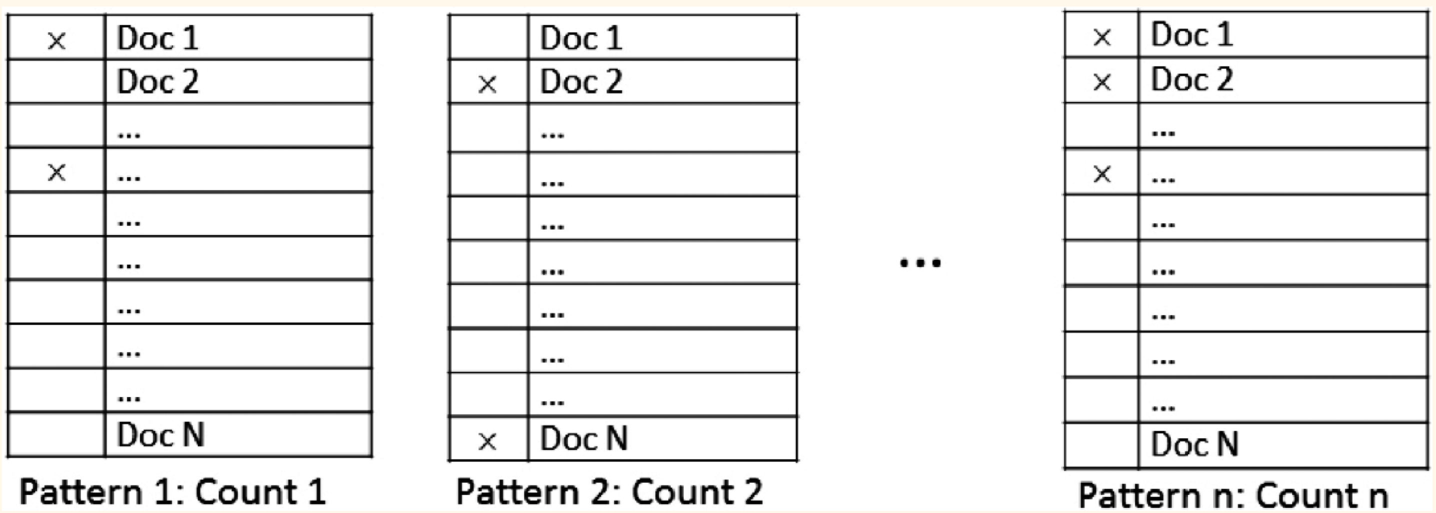


Fig. 5. An illustration of click patterns. Click patterns summarize positions of clicked URLs in search results of queries.

Jiang D, Pei J, Li H. Mining Search and Browse Logs for Web Search: A Survey. ACM Trans Intell Syst Technol. 2013;4(4):57:1–57:37.

Understanding queries

Query tasks:

- Describing & quantifying
- Classifying (by search goal, semantic class, etc.)
- Transforming (spelling correction, suggestion, etc.)
- Segmentation
- Entity recognition

Understanding queries: Describing & quantifying

Queries tend to be very short (Jansen et al. says 1.66-2.6 words)

Queries strongly follow power law distributions

Typical search session involves 2-3 queries

Understanding queries: Describing & quantifying

Major topical categories (according to Jiang et al.):

- People & place
- Commerce
- Health
- Entertainment
- Internet & Computer
- Pornography

Major linguistic structures:

- Noun phrases
- Compositions of noun phrases
- Titles
- Natural language

Understanding queries: Classifying

Search goal (navigational vs. informational)

Can be inferred for more common queries by looking at click-through data

Semantic class

Using click-through data, can classify based on text of target URLs...

... can also cluster based on click-through bipartite.

Location sensitivity

Does the query co-occur with location names?

Temporal sensitivity

Use time-stamp info, compare likelihoods within time windows

Understanding queries: Transforming

Change a less effective query to a more effective query ("ny times" -> "new york times"; spelling correction, etc.

Idea: Use click-through bipartite to identify similar queries

Pearson correlation; agglomerative clustering; etc.

The challenge: click-through graph can get *very* large...

Another approach: Use session data

Intuition: Users often issue similar queries in the same session, as part of "natural" reformulation.

Can use likelihood ratio of two queries w/in a session, etc. to identify "similar" pairs.

Understanding queries: Transforming

Change a less effective query to a more effective query ("ny times" -> "new york times"; spelling correction, etc.

Model-based transformation:

If we know that "sign on hotmail" and "sign up hotmail" are similar..a

... generalize to learn that "sign on *X*" and "sign up *X*" are similar.

Understanding queries: Segmenting

"new york times square" could be:

"new york" AND "times square", or

"new york times" AND "square"

We can use query frequency data!

Hagen et al.'s unsupervised approach:

$$score(S) = \begin{cases} \sum_{s \in S} |s|^{|s|} freq(s), & \forall s,\ freq(s) > 0,\ |s| \geq 2, \\ -1, & \text{otherwise}, \end{cases}$$

S is a given segmentation, and *s* is an n-gram in *S*.

Intuition: longer, more common sub-sequences should be rewarded.

Understanding documents

Document tasks:

- Representing
  - Queries & Clicks as annotations
- Determining relative importance
  - Queries & Clicks as endorsements
  - Browse time as endorsement
- Ranking search results
  - Queries & Clicks as endorsements
  - Preference pairs (direct ranking)

Understanding documents: Representation

Intuition: If a user clicks on a page in response to a query, the page is probably useful/relevant

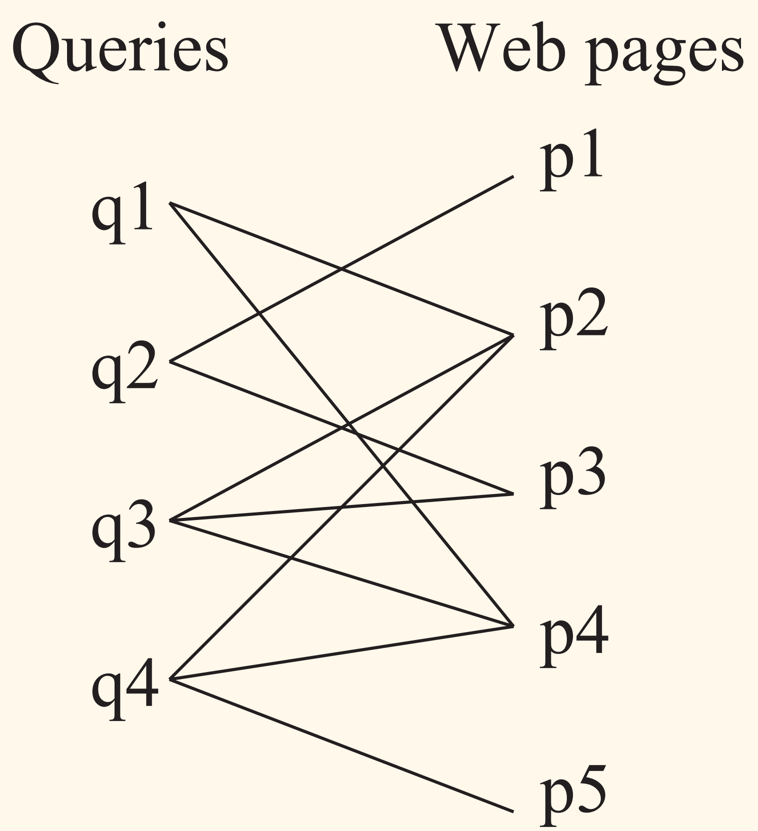
Simplest idea: use query terms as additional index terms on clicked document; weight accordingly

- Advantage: Simplicity, works surprisingly well
- Disadvantage: Assumes query term independence, click-through data is very sparse (many pages have zero clicks)

Understanding documents: Representation

Intuition: If a user clicks on a page in response to a query, the page is probably useful/relevant

More robust idea: Use click-through bipartite to identify *similar pages and queries*; use queries from similar pages.



Understanding documents: Importance

We've talked about PageRank, HITS, etc.

One drawback: those methods only represent the point of view of site authors.

By analyzing user browsing behavior, we can identify pages that users actually spend time on!

This is helpful for dealing with link spam.

Understanding documents: Ranking

We can use click-through information to improve ranking.

A user clicks on document #2…

... then, a minute later, clicks on #5.

Possible interpretation: #2 was insufficiently relevant.

Possible interpretation: #5 was more relevant than #3 and #4.

Problem: Position bias!!

Understanding users

User tasks:

Personalization

User *A* might want different results than user *B*.

Contextualization

Task *A* might need different results than task *B*. What task is the user performing?

Evaluation of satisfaction (or performance, behavior, etc.)

Understanding users: Personalization

Observation: Users often repeat a query and click on the same result each time.

Click-based personalization up-ranks page *p* for query *q* and user *u* if there is reason to think that this is a common query/selection.

$$S(q, p, u) = \frac{click(q, p, u)}{click(q, \cdot, u) + \beta}$$ Dou et al.

Problem: sparsity; doesn't work for new queries.

$$S(q, p, u) = \frac{\sum_{u_s} sim(u_s, u) click(q, p, u_s)}{\beta + \sum_{u_s} click(q, \cdot, u_s)}$$ Dou et al.

Solution: Find similar users, and use their data!

Problem: Calculating $sim(u_s, u)$ can be challenging!

Understanding users: Personalization

Another approach: term-based personalization

Using records of pages visited, queries issued, etc., build a probabilistic profile of the user, and integrate into search scoring.

$$S^u(q,d) = \sum_{t_i \in q} \frac{tf_i(k_1+1)}{k_1+tf_i} w_i^u$$

Teevan et al.

Or build a language model based on the user's search history:

$$p\left(t|\theta_i^u\right) = \lambda_i p\left(t|\theta_i\right) + (1-\lambda_i) p\left(t|\theta_i^h\right)$$

Tan et al.

Also can do topic-modeling, etc., to handle novel queries.

Jones R, Klinkner KL. Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs. Proceedings of CIKM '08. pp. 699–708.

Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs

Rosie Jones
Yahoo! Research
3333 Empire Ave
Burbank, CA 91504
jonesr@yahoo-inc.com

Kristina Lisa Klinkner*
Dept of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
klinkner@cmu.edu

ABSTRACT

Most analysis of web search relevance and performance takes a single query as the unit of search engine interaction. When studies attempt to group queries together by task or session, a timeout is typically used to identify the boundary. However, users query search engines in order to accomplish tasks at a variety of granularities, issuing multiple queries as they attempt to accomplish tasks. In this work we study real sessions manually labeled into hierarchical tasks, and show that timeouts, whatever their length, are of limited utility in identifying task boundaries, achieving a maximum precision of only 70%. We report on properties of this search task hierarchy, as seen in a random sample of user interactions from a major web search engine's log, annotated by human editors, learning that 17% of tasks are interleaved, and 20% are hierarchically organized. No previous work has analyzed or addressed automatic identification of interleaved and hierarchically organized search tasks. We propose and evaluate a method for the automated segmentation of users' query streams into hierarchical units. Our classifiers can improve on timeout segmentation, as well as other previously published approaches, bringing the accuracy up to 92% for identifying fine-grained task boundaries, and 89-97% for identifying pairs of queries from the same task when tasks are interleaved hierarchically. This is the first work to identify, measure and automatically segment sequences of user queries into their hierarchical structure. The ability to perform this kind of segmentation paves the way for evaluating search engines in terms of user task completion.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Query formulation

*This work was conducted while this author was at Yahoo! Inc

General Terms

Algorithms, Experimentation, Measurement

Keywords

query log segmentation, query session, query session boundary detection, search goal

1. INTRODUCTION

Web search engines attempt to satisfy users' information needs by ranking web pages with respect to queries. But the reality of web search is that it is often a process of querying, learning, and reformulating. A series of interactions between user and search engine can be necessary to satisfy a single information need [18].

To understand the way users accomplish tasks and subtasks using multiple search queries, we exhaustively annotated 3-day long query sequences for 312 web searchers. We limited the duration to three days to allow complete annotation of every query sequence, with an extremely thorough approach. These spans of time allowed us to identify tasks which result in queries placed over multiple days, as well as multiple tasks which may occur over several days. We manually annotated these query sequences with tasks and subtasks (which we will define as *missions* and *goals*), finding that many tasks contained subtasks, and many different tasks and subtasks were interleaved. While previous work has examined the way users interleave tasks [9], no previous work has examined the way tasks contain subtasks.

If we are able to accurately identify sets of queries with the same (or related) information-seeking intent, then we will be in a better position to evaluate the performance of a web search engine from the user's point of view. For example, standard metrics of user involvement with a search engine or portal emphasize *visits* or *time spent* [1]. However, each page view can constitute small pieces of the same information need and each visit could encompass some larger task. If we could instead quantify the number of information needs or tasks which a user addresses via a website, we would have a clearer picture of the importance of the site to that user. In particular, we could evaluate user effort in terms of queries issued or time spent on a task, as the user attempts to satisfy an information need or fulfill a more complex objective.

To this end, we built classifiers to identify task and subtasks boundaries, as well as pairs of queries which correspond to the same task, despite being interleaved with queries from other tasks.

Agichtein E, Brill E, Dumais S. Improving Web Search Ranking by Incorporating User Behavior Information. SIGIR '06. pp. 19–26.

Improving Web Search Ranking by Incorporating User Behavior Information

Eugene Agichtein
Microsoft Research
eugeneag@microsoft.com

Eric Brill
Microsoft Research
brill@microsoft.com

Susan Dumais
Microsoft Research
sdumais@microsoft.com

ABSTRACT
We show that incorporating user behavior data can significantly improve ordering of top results in real web search setting. We examine alternatives for incorporating feedback into the ranking process and explore the contributions of user feedback compared to other common web search features. We report results of a large scale evaluation over 3,000 queries and 12 million user interactions with a popular web search engine. We show that incorporating implicit feedback can augment other features, improving the accuracy of a competitive web search ranking algorithms by as much as 31% relative to the original performance.

Categories and Subject Descriptors
H.3.3 Information Search and Retrieval – *Relevance feedback, search process*; H.3.5 Online Information Services – *Web-based services*.

General Terms
Algorithms, Measurement, Experimentation

Keywords
Web search, implicit relevance feedback, web search ranking.

1. INTRODUCTION
Millions of users interact with search engines daily. They issue queries, follow some of the links in the results, click on ads, spend time on pages, reformulate their queries, and perform other actions. These interactions can serve as a valuable source of information for tuning and improving web search result ranking and can compliment more costly explicit judgments.

Implicit relevance feedback for ranking and personalization has become an active area of research. Recent work by Joachims and others exploring implicit feedback in controlled environments have shown the value of incorporating implicit feedback into the ranking process. Our motivation for this work is to understand how implicit feedback can be used in a large-scale operational environment to improve retrieval. How does it compare to and compliment evidence from page content, anchor text, or link-based features such as inlinks or PageRank? While it is intuitive that user interactions with the web search engine should reveal at least *some* information that could be used for ranking, estimating user preferences in real web search settings is a challenging problem, since real user interactions tend to be more "noisy" than commonly assumed in the controlled settings of previous studies.

Our paper explores whether implicit feedback can be helpful in realistic environments, where user feedback can be noisy (or adversarial) and a web search engine already uses hundreds of features and is heavily tuned. To this end, we explore different approaches for ranking web search results using real user behavior obtained as part of normal interactions with the web search engine.

The specific contributions of this paper include:

- Analysis of alternatives for incorporating user behavior into web search ranking (Section 3).
- An application of a robust implicit feedback model derived from mining millions of user interactions with a major web search engine (Section 4).
- A large scale evaluation over real user queries and search results, showing significant improvements derived from incorporating user feedback (Section 6).

We summarize our findings and discuss extensions to the current work in Section 7, which concludes the paper.

2. BACKGROUND AND RELATED WORK
Ranking search results is a fundamental problem in information retrieval. Most common approaches primarily focus on similarity of query and a page, as well as the overall page quality [3,4,24]. However, with increasing popularity of search engines, implicit feedback (i.e., the actions users take when interacting with the search engine) can be used to improve the rankings.

Implicit relevance measures have been studied by several research groups. An overview of implicit measures is compiled in Kelly and Teevan [14]. This research, while developing valuable insights into implicit relevance measures, was not applied to improve the ranking of web search results in realistic settings.

Closely related to our work, Joachims [11] collected implicit measures in place of explicit measures, introducing a technique based entirely on clickthrough data to learn ranking functions. Fox et al. [8] explored the relationship between implicit and explicit measures in Web search, and developed Bayesian models to

Lagun D, Hsieh C-H, Webster D, Navalpakkam V. Towards Better Measurement of Attention and Satisfaction in Mobile Search. Proceedings of SIGIR '14. pp. 113–22.

Towards Better Measurement of Attention and Satisfaction in Mobile Search

Dmitry Lagun
Emory University
dlagun@emory.edu

Chih-Hung Hsieh
Google Inc.
chh@google.com

Dale Webster
Google Inc.
drw@google.com

Vidhya Navalpakkam
Google Inc.
vidhyan@google.com

ABSTRACT

Web Search has seen two big changes recently: rapid growth in mobile search traffic, and an increasing trend towards providing answer-like results for relatively simple information needs (e.g., [weather today]). Such results display the answer or relevant information on the search page itself without requiring a user to click. While clicks on *organic* search results have been used extensively to infer result relevance and search satisfaction, clicks on answer-like results are often rare (or meaningless), making it challenging to evaluate answer quality. Together, these call for better measurement and understanding of search satisfaction on mobile devices. In this paper, we studied whether tracking the browser *viewport* (visible portion of a web page) on mobile phones could enable accurate measurement of user attention at scale, and provide good measurement of search satisfaction in the absence of clicks. Focusing on answer-like results in web search, we designed a lab study to systematically vary answer presence and relevance (to the user's information need), obtained satisfaction ratings from users, and simultaneously recorded eye gaze and viewport data as users performed search tasks. Using this ground truth, we identified increased scrolling past answer and increased time below answer as clear, measurable signals of user dissatisfaction with answers. While the viewport may contain three to four results at any given time, we found strong correlations between gaze duration and viewport duration on a per result basis, and that the average user attention is focused on the top half of the phone screen, suggesting that we may be able to scalably and reliably identify which specific result the user is looking at, from viewport data alone.

Keywords

Search on mobile phone; user attention and satisfaction; viewport logging.

1. INTRODUCTION

Recent years have witnessed a rapid explosion in the usage of mobile devices on the web. According to recent surveys, web browsing on mobile devices increased five fold from 5.2% three years ago to 25% in April 2014[26]; and a significant amount of

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609631.

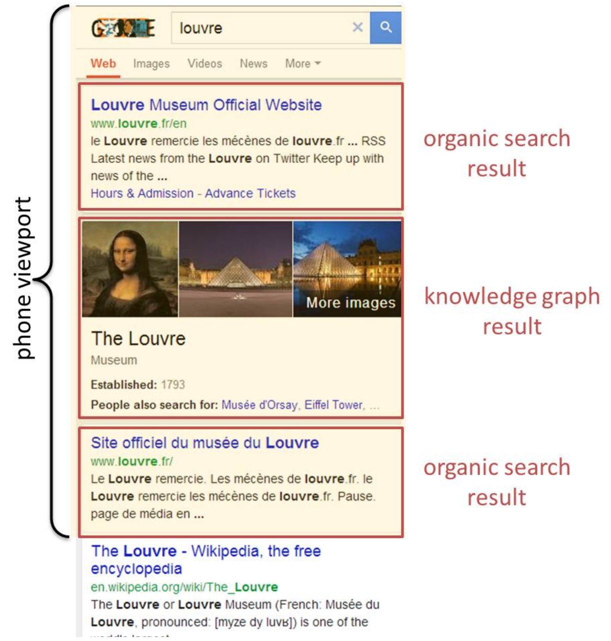
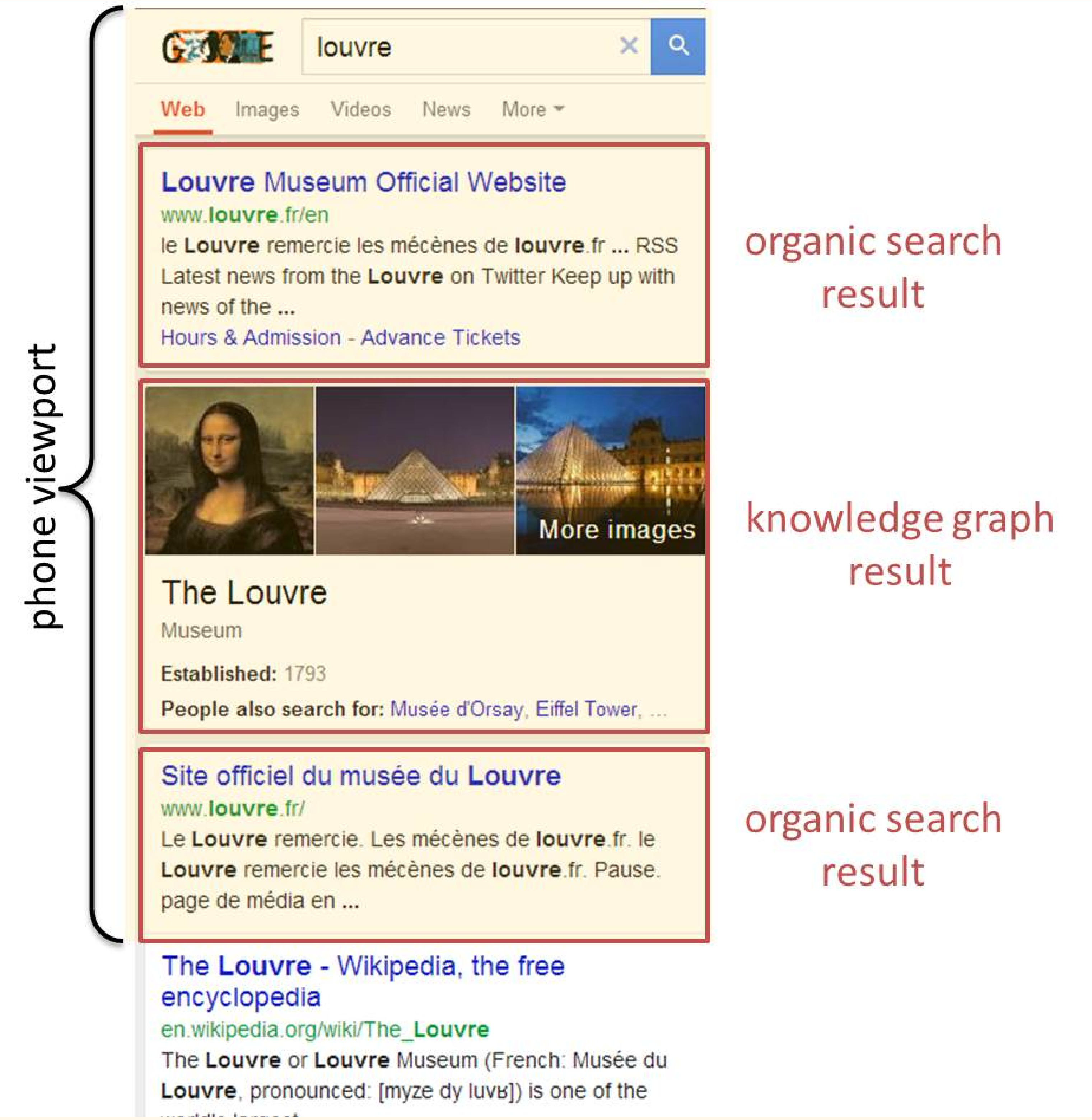


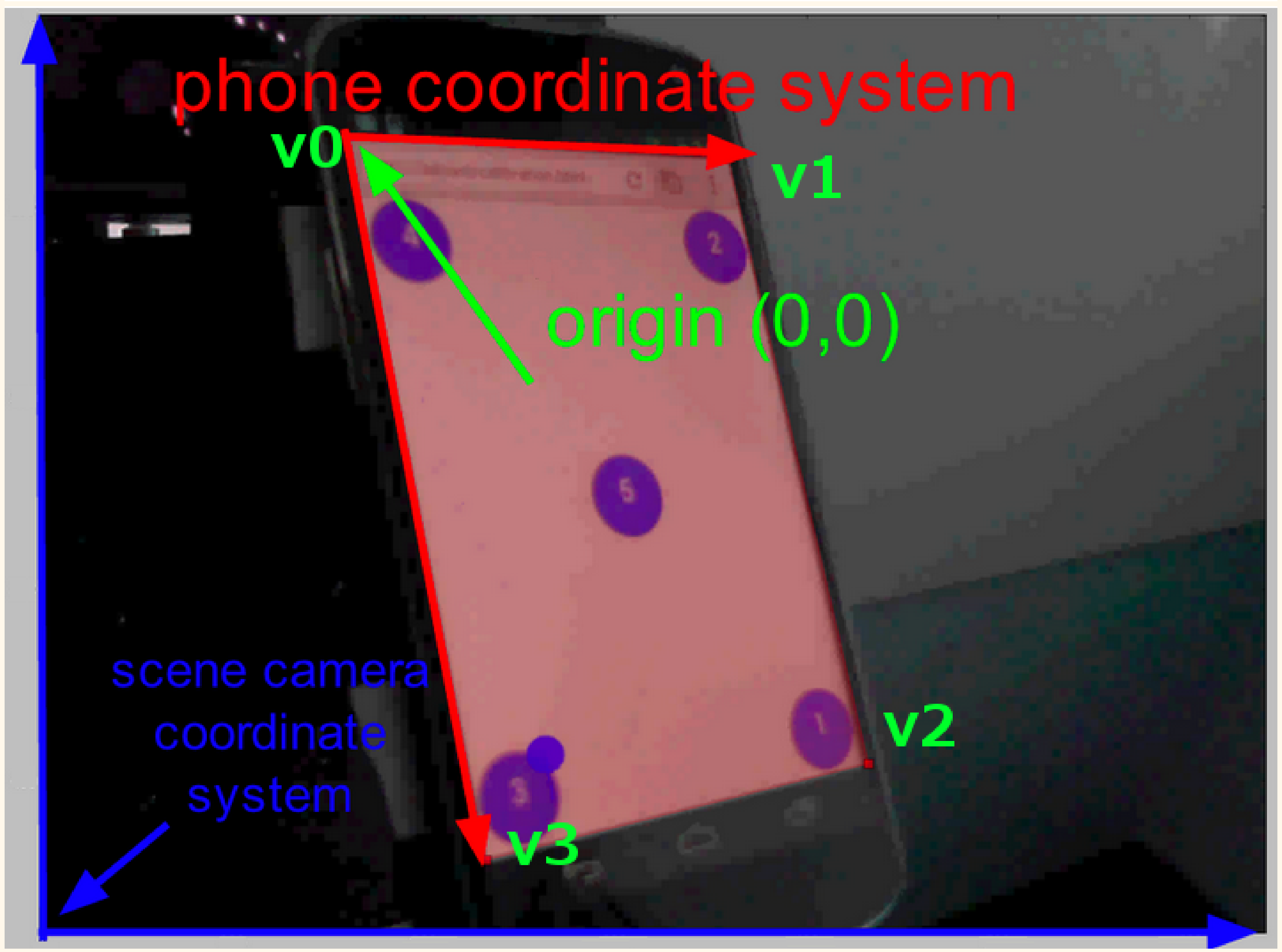
Figure 1: An example of the search results page showing Knowledge Graph result. The yellow area indicates current position of the browser's viewport (visible portion of the page).

search engines' traffic (about one in every five searches) is generated by mobile devices[25]. Another recent change in search is the increasing trend towards providing answer-like results for simple information needs that are popular on mobile (e.g., [weather today], [pizza hut hours]). Such results display the answer or relevant information on the search page itself without requiring the user to click. Instant information is desirable on mobile devices, but poses a challenge – while clicks on *organic* search results have been extensively used to infer result relevance and search satisfaction [5, 6], answer-like results often do not receive clicks, which makes it difficult to evaluate answer quality and search satisfaction. Together, the rapid growth in mobile traffic and answer-like results in Search warrants better understanding of user attention and satisfaction in search on mobile devices.

Search behavior on mobile devices can be different than on desktop for several reasons. Unlike traditional desktop computers with large displays and mouse-keyboard interactions, touch enabled mobile devices have small displays and offer a variety of touch interactions, including touching, swiping and zooming. As a result, user



phone viewport
louvre
Web Images Videos News More
Louvre Museum Official Website
www.louvre.fr/en
le Louvre remercie les mécènes de louvre.fr ... RSS Latest news from the Louvre on Twitter Keep up with news of the ...
Hours & Admission - Advance Tickets
More images
The Louvre
Museum
Established: 1793
People also search for: Musée d'Orsay, Eiffel Tower, ...
Site officiel du musée du Louvre
www.louvre.fr/
Le Louvre remercie. Les mécènes de louvre.fr. le Louvre remercie les mécènes de louvre.fr. Pause. page de média en ...
The Louvre - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/The_Louvre
The Louvre or Louvre Museum (French: Musée du Louvre, pronounced: [myze dy luvʁ]) is one of the
organic search result
knowledge graph result
organic search result

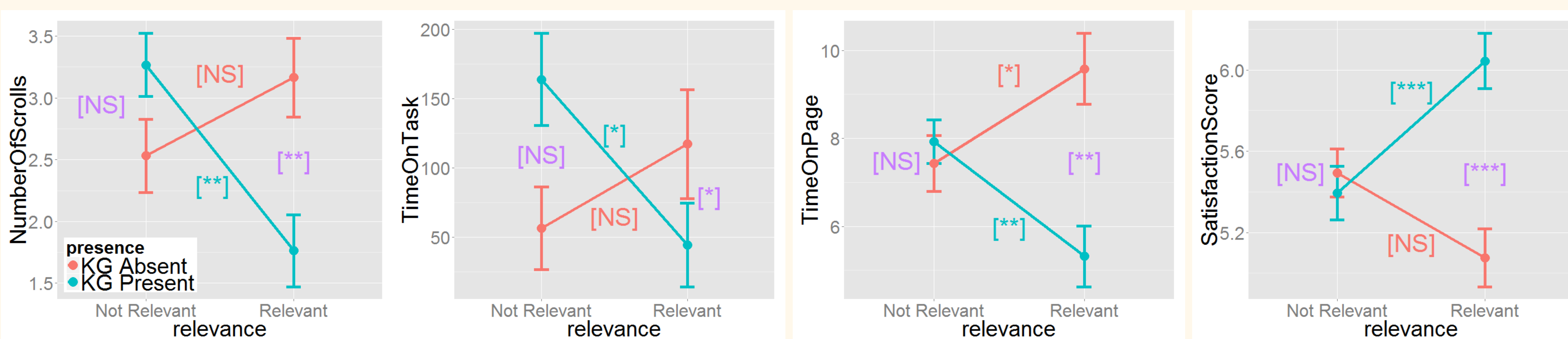

phone coordinate system
v0
v1
origin (0,0)
scene camera
coordinate
system
v2
v3

Query	**Task Description**	
	KG Relevant	**KG Not Relevant**
university of cambridge	What was the enrollment of the University of Cambridge in 2012?	Find the rank of University of Cambridge in academic rankings.
golden gate bridge	What is the length of the Golden Gate Bridge?	Find information regarding tolling and transit through the Golden Gate Bridge.
the avengers movie	Who was director of the Avengers movie?	Find a link to watch the Avengers movie trailer.
	IA Relevant	**IA Not Relevant**
sfo to atl price	Find the ticket price of the Delta flight from San Francisco (SFO) to Atlanta (ATL).	Find a website to compare different prices for flights from San Francisco (SFO) to Atlanta (ATL).
aapl earnings	What is the current stock price of Apple Inc.?	Find Apple Inc. earnings in second quarter of 2013.
world cup 2014	When does the FIFA 2014 world cup start?	Find a website to buy tickets for the FIFA 2014 world cup.

Table 1: Example task descriptions used in the user study.

	Metric	KG Present		KG Absent		p-value [3]
		Relevant	Not Relevant	Relevant	Not Relevant	
Gaze	TimeOnKG (s)	0.64 ± 0.20	0.62 ± 0.09			p=0.067
	% TimeOnKG	34 ± 5	39 ± 4			p=0.179
	TimeBelowKG (s)	1.19 ± 0.32	0.73 ± 0.12			p=0.380
	% TimeBelowKG	24 ± 4	28 ± 3			p=0.279
Viewport	TimeOnKG (s)	3.96 ± 0.42	5.38 ± 0.34			p<0.001
	% TimeOnKG	25 ± 2	20 ± 1			p=0.029
	TimeBelowKG (s)	11.28 ± 2.18	12.83 ± 1.26			p=0.001
	% TimeBelowKG	16 ± 2	26 ± 2			p<0.001
Page	NumberOfScrolls	1.77 ± 0.28	3.32 ± 0.25	3.2 ± 0.33	2.52 ± 0.29	p=0.003
	TimeOnPage (s)	5.37 ± 0.65	7.98 ± 0.47	9.80 ± 0.85	7.42 ± 0.65	p<0.001
	TimeOnTask (s)	48.30 ± 30.06	163.82 ± 33.12	115.89 ± 39.31	64.13 ± 29.81	p<0.001
	SatisfactionScore	6.03 ± 0.13	5.39 ± 0.13	5.0 ± 6.15	5.51 ± 0.11	p=0.002

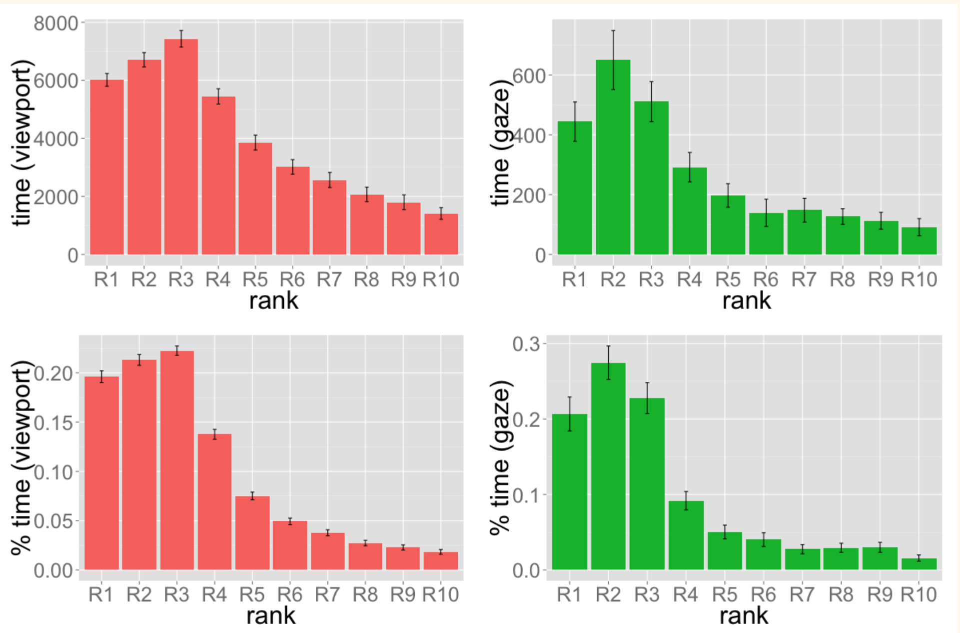
Table 2: Gaze, Viewport and Page metrics summarized for each experiment condition ($M \pm SE$).



(a) (b) (c) (d)

Metric	IA Relevant	IA Not Relevant	p-value
Gaze			
TimeOnIA (s)	0.55 ± 0.09	0.74 ± 0.11	p=0.812
% TimeOnIA	45 ± 5	38 ± 3	p=0.237
TimeBelowIA (s)	1.21 ± 0.23	1.41 ± 0.17	p=0.298
% TimeBelowIA	55 ± 5	62 ± 3	p=0.343
Viewport			
TimeOnIA (s)	1.96 ± 0.24	3.64 ± 0.26	p<0.001
% TimeOnIA	11 ± 1	16 ± 1	p<0.001
TimeBelowIA (s)	11.74 ± 1.59	19.02 ± 1.30	p<0.001
% TimeBelowIA	32 ± 3	56 ± 2	p<0.001
NumberOfScrolls	1.33 ± 0.17	2.96 ± 0.20	p<0.001
NumberOfEvents	6.12 ± 0.39	9.93 ± 0.38	p<0.001
TimeOnPage (s)	3.89 ± 0.43	7.17 ± 0.41	p<0.001
TimeOnTask (s)	90.7 ± 1.65	102.82 ± 1.73	p<0.001
SatisfactionScore	6.25 ± 0.09	5.08 ± 0.11	p<0.001

Table 3: Summary of Gaze, Viewport and Page (M $\pm$ SE) for “IA Relevant” and “IA Not Relevant” experiment conditions. Time related metrics are measured in seconds.



(a) KG Relevant (b) KG Not Relevant

Figure 4: Attention heatmaps for *KG Relevant* and *KG Not Relevant* conditions. This figure shows that on average, across all users in the study, there is increased gaze activity below KG when it is irrelevant than relevant.