Addressing People's Information Needs Directly in a Web Search Result Page

Lydia B. Chilton
University of Washington
Seattle, WA, USA
hmslydia@cs.washington.edu

Jaime Teevan
Microsoft Research
Redmond, WA, USA
teevan@microsoft.com

ABSTRACT
Web search engines have historically focused on connecting people with information resources. For example, if a person wanted to know when their flight to Hyderabad was leaving, a search engine might connect them with the airline where they could find flight status information. However, search engines have recently begun to try to meet people's search needs directly, providing, for example, flight status information in response to queries that include an airline and a flight number. In this paper, we use large scale query log analysis to explore the challenges a search engine faces when trying to meet an information need directly in the search result page. We look at how people's interaction behavior changes when inline content is returned, finding that such content can cannibalize clicks from the algorithmic results. We see that in the absence of interaction behavior, an individual's repeat search behavior can be useful in understanding the content's value. We also discuss some of the ways user behavior can be used to provide insight into when inline answers might better trigger and what types of additional information might be included in the results.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, search process*.

General Terms
Measurement, Human Factors.

Keywords
Answers, question asking, query log analysis, Web search.

1. INTRODUCTION
Most Web search engine users have discovered (perhaps without realizing it) a search feature that we call in this paper *Answers*, where relevant content is provided directly within the search result page. For example, the query "weather" no longer returns a link to http://www.weather.com as the first result. Instead, major search engines like Bing, Google, and Yahoo! use the top result space to answer the user's query directly within the result page, providing a pictorial weather forecast of the user's local weather. Figure 1 shows a snapshot of the Weather Answer returned by Bing for the query "weather Beijing". Answers are a step towards the long-standing goal of Web search engines to directly address their searchers' needs, versus merely linking to relevant content.

WWW 2011, March 28 – April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0632-4/11/03.

The presence of an Answer on a search result page changes the value of the interaction metrics that have traditionally been used to evaluate and improve Web search result quality. When people's needs are met by an Answer, they may not interact with the search result page at all, a signal that is traditionally interpreted as a negative experience. More sophisticated interaction metrics, such as ones based on interaction with other elements on the page or repeat engagement, must be used instead. However, because there are many different ways Answers can address people's needs, there are also many different ways they can influence user interaction.

In this paper, we explore the important factors that influence Answer use, and suggest several new ways to interpret query log data in the presence of Answers. After a discussion of related work, we describe the query log data we analyzed and provide details of the specific Answer types we studied. We then present our findings, including:

- Answers that provide content inline can reduce engagement with the search result page, thus cannibalizing interaction.
- Repeat usage gives us insight into the relevance of Answers, even when clicks are cannibalized.
- People who consistently use the same Answer type over time are often monitoring Answer content. Repeat Answer usage within a session indicates task-based reuse.
- Answers that are triggered with identical queries are often being monitored for new content, while those that are triggered with different queries are exploratory

These findings suggest rich opportunities for search engines as they attempt to directly meet their user's needs. We conclude with a discussion of these promising directions.

Figure 1. An example of a *Weather Answer*. The first result for the query "weather beijing" is a pictorial weather forecast.

2. RELATED WORK

There has been a lot of valuable research done in the area of information retrieval and natural language processing to automatically answer the questions people ask via Web search engines. Much of this research focuses on identifying Web queries that have a question-answering intent or using content from the Web to identify and provide good answers [1,11]. This body of work is valuable for understanding when to trigger Answers and what information to provide when they trigger. People's interactions with search results have been studied by a number of researchers as a way to infer how those people understand the results they are presented with. Kelly and Teevan [9] give an overview of this work. Typically, features such as which results are clicked are used infer which results are relevant and whether the search was an overall success.

Log data can be a useful tool for understanding what real people do in real world situations with real, self-motivated tasks. But observed behavior does not always reflect the most obvious interpretation and the use of implicit data must be done with care. For example, people are strongly biased by position in what they choose to click [8], and clicks on results do not necessarily imply satisfaction with the search [6]. Joachims et al. [8] found that result clicks can be more accurately understood as a relative preference for a result over those surrounding the current result.

When no results are clicked following a query, the query is considered *abandoned*. Abandoned queries are generally understood to be the result of low quality search results. For example, Radlinski [12] found a negative correlation between a query's result quality and its abandonment rate, and Hassan et al. [7] found abandoned search sessions were only 10% likely to be successful. As a result, abandonment has been used to evaluate search engine performance for specific queries. Wang et al. [19] made use of abandonment as a measure of search result quality, and Sarma et al. [13] used skipped results as negative feedback to reduce query abandonment rate.

In this paper we focus on what Web query log data can tell us about people's interactions with search results that contain Answers. As search engines evolve to provide more content directly to their users, versus through links to content, it is important to step back and reflect on how behavioral data can best be used to understand the user's experience. Recently there has been some recognition of the fact that people do not always abandon searches because they are dissatisfied with the results. White and Dumais [1] studied people's motivations for switching search engines, a behavior often proceeded by query abandonment, and found that dissatisfaction and frustration accounted for only about a third of the changes. Feild et al. [5] used abandonment as a baseline for predicting user frustration during search, and found that 43% of the time it did not predict frustration. They were able to predict frustration much better by using richer query log features suggested by White and Dumais [1], like query length and duration of the task. Dupret and Liao [4] developed a model for interpreting post-click log data to understand document relevance that is particularly useful for queries with a high abandonment rate.

Queries without clicks can represent successful searches when the search engine is able to satisfy its user's information need directly within the search result page. Stamou and Efthimiadis [14,15] found that for 27% of the queries they studied, searchers intended to find what they were looking for directly on the result page. Common reasons cited for this included checking the spelling of the query term, monitoring the query for new content, and learning generally about the query term from the search result snippets returned.

Although search engines have for years tried to provide as much relevant information as possible about a result within the result page summary [3,17], search results have recently begun to actively provide relevant content separately from search results. These Answers represent an important shift in how search engines respond to information requests, but people's interactions in the presence of Answers are poorly understood. Li et al. [10] analyzed query logs to estimate the potential impact of Answers on abandonment, arguing that any abandoned query for which an Answer was shown probably represents a type of *good abandonment*. They found nearly half of all abandoned queries returned an Answer that could have satisfied the information need without requiring additional interactions. Their initial exploration into the types of queries with good abandonment suggests that behavior varies greatly by entry point and country of origin.

In this paper, we look more closely at the implicit user interaction behavior that surrounds Answers. We build on the work described above to understand how positive search interactions can be observed in Web search logs for items that do not require interaction to provide benefit. We use different types of implicit feedback than have previously been explored. Bailey et al. [2] present a way to understand the relevance of a search page as a whole based on a complete picture of a person's Web page interactions. In a related manner, one approach we take is to look at how behavior with other parts of a result page can tell us about the part we are interested in, namely the Answer. Building on the notion that clicks on some results provide feedback about the unclicked results [8], we introduce the notion of *cannibalism*. When the click through rate on search result content that is traditionally clicked a lot is reduced, or cannibalized, it helps inform us of the usefulness of search result content that does not receive a lot of clicks. We also look at using search behavior over time to understand a user's satisfaction with the search results returned for repeat or similar queries. Teevan et al. found that people repeat search engine queries regularly [16,18]. Here we show that this repeat behavior can us be used as a way to better understand the quality of the user's experience with the query.

3. METHODS

Our analysis of Answer interaction is based on a one week sample of Bing Web search engine query data from August 23, 2010 to August 29, 2010. Associated with each query, the query logs include the query text, an anonymized user identifier, a timestamp, a list of the Web search results returned, their position on the search result page, and whether or not each result was clicked. The query logs also contain information about any Answers that are displayed to the user following a query, and the users' interactions with those Answers.

Although Answers can appear anywhere within the search result list, we focus in our analysis only on Answers shown at the top of the result list. For some queries (e.g., "william shatner") Answers are aggregated into a group (e.g., pictures of William Shatner, tweets posted by William Shatner, and a list of movies William Shatner has been in). We exclude these instances from our analysis so as to focus on individual Answers.

Users in our sample were selected to have issued at least ten queries spanning at least 24 hours to ensure they had a baseline level of activity. Users are identified by an anonymous ID associated with a user account on a particular computer. As is the case with most log analyses, if a user has more than one computer, that user will have multiple IDs. Conversely, if more than one person uses the same account on a computer, they are amalgamated into a single user. We only look at queries issued in English from within the United States. We exclude IP addresses from within Microsoft, and exclude users with extremely atypical behavior, including any user with more than 400 queries in the seven day window or 100 queries in a single session. (A session is defined as the set of queries issued by an individual with less than 30 minutes between sequential queries.) The sample was further filtered to remove spam and processed so that pagination and back button clicks were treated as the same query.

The resulting sample represents over 200 million queries from 8 million users. Almost a hundred different Answer types triggered as a result of those queries, 42 of which appear in the top position 10,000 times or more (more than 0.005% of the time). We focus in this paper on a subset of 15 Answer types that 1) occur 10,000 times or more, and 2) illustrate interesting properties of Answer interaction. They are:

1. ***Attractions Answer*** A list of attractions in a location.
2. ***Currency Answer*** Currency conversion information.
3. ***Dictionary Answer*** A dictionary definition of a query term.
4. ***Finance Answer*** Financial information for a company mentioned in the query.
5. ***Flight Status Answer*** The status of the flight number indicated in the query.
6. ***Golf Answer*** Information related to professional golf.
7. ***Horoscope Answer*** A list of links to horoscope readings for all Zodiac signs.
8. ***News Answer*** Top news headlines related to the query.
9. ***Newspaper Answer*** A list of newspapers in a location.
10. ***Phonebook Answer*** Contact information for local people and businesses.
11. ***Reference Answer*** Inline factual information.
12. ***Time Zone Answer*** The local time in a specified time zone.
13. ***Translate Answer*** Direct translation for query terms, or a link to the Bing Translator.
14. ***Twitter Answer*** Twitter updates from a verified celebrity or company Twitter account related to the query.
15. ***Weather Answer*** A multi-day weather forecast related to the user's location or a location mentioned in the query.

Screenshots of the 15 different Answer types can be found in Figures 1 and 2, and example queries can be found in Table 2. Note that there are other Answer types that are similar to the above 15 that are returned for different (but related) queries or circumstances. For example, the Horoscope Answer displays a list of horoscope readings following the query *horoscope*, but a different Answer appears for the query *virgo horoscope*, displaying a reading specific to the Virgo Zodiac sign. In the next section, we analyze the log data to build a rich picture of how the 15 Answers we studied are used.

4. ANALYSIS

After a brief overview of Answer occurrence, we look at how people engage with Answers and with the associated Web search results. We then discuss how Answer use over time can tell us more than interaction behavior alone, diving deeply into the consistency of the queries people use to trigger them and their use within a session versus across multiple sessions.

4.1 Answer Occurrence

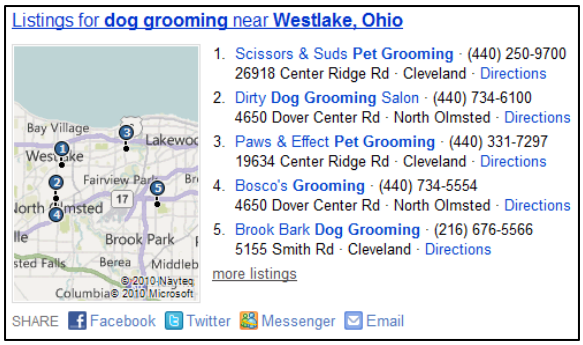
Given the great variety of queries that get issued to a Web search engine, most Answer types appear only rarely. In our data, the most frequent Answer type was the Phonebook Answer, and the least frequent was the Time Zone Answer. The disparity in occurrence rates was great; the Phonebook Answer appeared almost 500 times as often as the Time Zone Answer. As a result, just a few Answer types accounted for a large portion of the Answer volume. The two most popular Answers we studied (Phonebook and News) were responsible for 89% of the Answer query volume in our sample.

Researchers have explored how to identify queries that might best be addressed via a direct response (e.g., [20]). In this paper, we do not focus on how Answers are triggered, but rather look at user behavior when an Answer appears. However, to do this it is necessary to understand the range of ways Answers are triggered. Examples of the queries that trigger each Answer type can be found in Table 2.

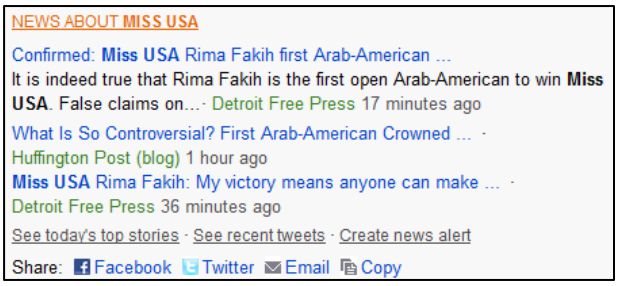
Some Answer types trigger for many different queries, and others for only very few. The Horoscope Answer, for example, is only triggered by queries closely related to the term "horoscope", while the News Answer triggers for a wide variety of queries including (during the week of our analysis) "miss usa", "justin bieber sick", and "ghost train hunter killed".

Sometimes Answers trigger for queries that have clear interpretations, while others trigger for queries that are harder to interpret. The Flight Status Answer for "alaska 600", the Currency Answer for "500 cny", and the Dictionary Answer for "define retrench", all shown in Figure 2, are examples of Answer-query pairs for which the Answer is most likely relevant to the searcher's need and able to fully satisfy it. In contrast, when the Golf Answer triggers for "pga", the News Answer triggers for "miss usa", or the Finance Answer triggers for "jedi mind inc", it is less likely that the Answer will fully satisfy the user's information need, or even address the user's specific need at all. A person searching for "miss usa", for example, could want to find an application to be in the next pageant and may not be interested in the news items.

Answers do not always trigger every time they could be useful. For example, while the queries "weather", "weather boston", and "tomorrow's weather" all trigger the Weather Answer, the query "what's the weather?" does not. Of course, even a seemingly straightforward query like "weather" can be ambiguous. It could express a desire to learn about what causes weather patterns. Moreover, if the user wants to see a 10-day forecast, the short-term Weather Answer forecast is relevant but not sufficient.

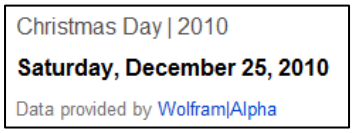


Phonebook Answer ("dog grooming westlake ohio")



News Answer ("miss usa")

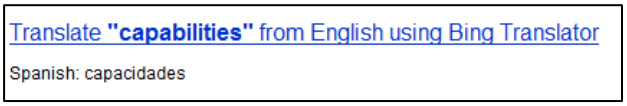
Currency Answer ("500 cny")



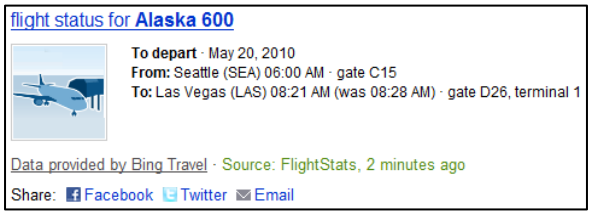
Reference Answer ("what day is christmas?")

Newspaper Answer ("arkansas newspaper")

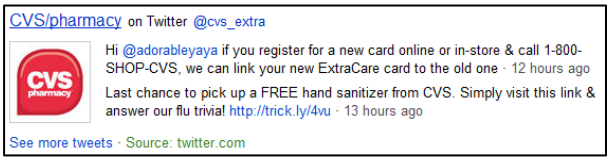
Attractions Answer ("bozeman montana attractions")



Translate Answer ("translate capabilities into spanish")

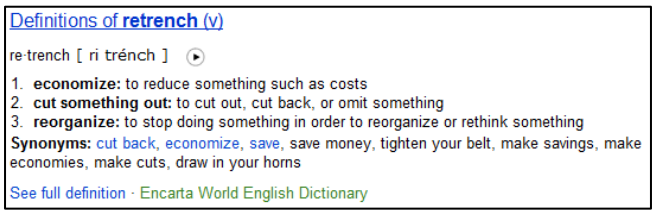


Flight Status Answer ("alaska 600")

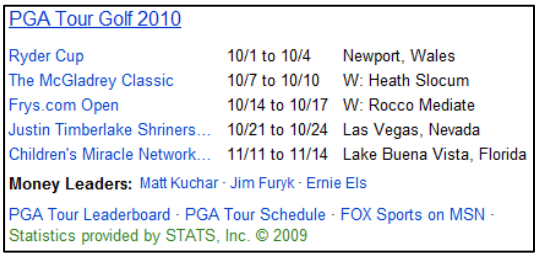


Twitter Answer ("cvs")

Time Zone Answer ("time beijing")



Dictionary Answer ("define retrench")



Golf Answer ("pga")

Finance Answer ("jedi mind inc")

Horoscope Answer ("horoscope")

Figure 2. Examples of 14 of the Answer types we studied. The fifteenth (the Weather Answer) is shown in Figure 1.

Table 1. User engagement with Answers and the corresponding search results, broken down by how often the engagement was with only the Answer, only the search results, or both. Values are normalized relative to the average engagement with all 15 Answer types. The highest and lowest two values for each column have a gray background.

Answer	Only Answer	Only results	With both
Horoscope	2.24	0.78	2.60
Golf	1.90	0.78	0.35
Translate	1.73	1.23	1.42
Flight Status	1.67	0.80	0.68
News	1.36	0.79	1.30
Phonebook	1.34	1.07	1.38
Weather	1.22	0.97	0.86
Attractions	1.09	0.98	2.85
Currency	0.76	0.66	0.37
Newspaper	0.73	1.64	0.48
Dictionary	0.31	0.30	0.54
Finance	0.29	0.73	0.24
Twitter	0.26	1.67	1.60
Reference	0.09	1.30	0.32
Time Zone	0.00	1.32	0.00

4.2 Interaction Behavior

Interaction behavior can tell us something about how useful an Answer is. Interaction data must be interpreted carefully because different Answers provide different amounts of information directly in the search result page and support different levels of engagement. For example, as can be seen in Figure 2, the horoscopes provided by the Horoscope Answer are useless without first clicking on the desired Zodiac sign, whereas the Phonebook Answer displays phones numbers and addresses directly on the search result page. Additionally, some Answers provide more opportunity for interaction than others. The Phonebook Answer in Figure 2 displays a map with the locations of five local businesses and lists contact information for each of the businesses. There are many links within this Answer: the map, the business URLs, and links to directions to name a few. In contrast, the Time Zone Answer (also shown in Figure 2) offers no opportunity for interaction.

In this section we look more deeply at how people engage with Answers. We begin by looking at people's level of engagement with each Answer types when it appears. We then explore how engagement with the algorithmically generated search results interacts with Answer engagement to better understand what this can tell us about the value of the Answer.

4.2.1 Engagement with the Answer

To measure a user's overall engagement with an Answer type, we look at how often the user clicks on some component of the Answer given the Answer was shown. Table 1 shows the relative percentage of clicks each Answer type received, as compared with the level of engagement we see averaged across all 15 Answer types. The Attractions Answer displayed the most average level of engagement, and was interacted with 1.09 times as often as the average Answer type.

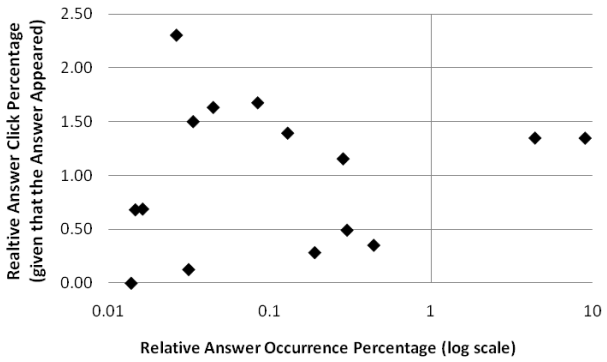


Figure 3. How often each of our 15 Answer types occurred, as compared to how often each Answer was clicked.

Engagement varied substantially by type. Some Answers were engaged with a lot. For example, the Horoscope Answer was clicked 2.24 times as often as the average Answer. The Golf, Translate, and Flight Status Answers were all interacted with more than 50% as often as the average Answer. These Answers all provide ample opportunity for the user to interact with them, and some of them, such as the Horoscope and Translate Answers, require user interaction before presenting content.

Other Answers types were engaged with very rarely. Lack of interaction with a search result page is typically taken to be a negative sign, but for some Answer types interaction is not even possible. The Time Zone Answer is only text and contains no links for the user to click. Other Answers provide some opportunity, but still are interacted with rarely. The Dictionary, Finance, Twitter, and Reference Answers are all interacted with less than half as much as the average Answer. All potentially provide a full response to the searcher's information need directly. Most Answers with low engagement also display relatively few links as compared with other Answers. An exception is the Finance Answer, which provides several links to additional financial information and related stock quotes. The lack of interaction may be because it also provides a lot of direct content. In contrast, the Translate Answer has high engagement despite only having one link. Although translated content is sometimes presented in the text of the Answer (see Figure 2 for an example), that content is visually overwhelmed by the link to the translator.

As can be seen in Figure 3, there is no correlation between how often an Answer type appeared and how often users engaged with that type. Answers that appeared often were not necessarily clicked a high percentage of the time they were presented.

4.2.2 Cannibalization of Algorithmic Clicks

In addition to studying Answer engagement, we looked at how the presence of an Answer affected people's engagement with the algorithmically provided search result links. Given a user interacted with either the Answer or the results for a query with an Answer, Table 1 shows the percentage of time just the Answer was clicked, just an algorithmic search result was clicked, and both results and the Answer were clicked. Table 2 shows several examples of queries that triggered each Answer, broken down by whether the user clicked only on the Answer following the query or clicked only on a search result following the query.

Table 2. Example queries that triggered each Answer type, given a click only on the Answer or only on the results.

Answer	Answer Click Only	Results Click Only
Attractions	baseball hall of fame; kansas attractions; chinatown san francisco	key west attractions; places to go in georgia;
Currency	1 usd in pkr; 37 euros; convert 100000 indonesia rupiah to sgd	euro exchange rate;turkish lira;
Dictionary	synonym for dispute; define whole numbers; the meaning of compassion	what is a green card;christian name meanings; what is death
Finance	msft; goog; aapl; bac quote; sara lee stock	amok stock; dsny; christian quotes;
Flight Status	aa 154; airtran airways flight schedule	qantas 93;flight status
Golf	golf tiger woods; us open schedule; pga	golf scores;pga
Horoscope	horoscope	horoscope
News	2010 emmy awards; movies in theaters;	big brother 12; mariah carey pregnant?
Newspaper	tennessee newspapers; sri lanka newspapers; louisanna newspaper	fayetteville, ar. newspapers;vermont newspaper;
Phonebook	mini-golf in olney; keg steakhouse; chevrolet dealers;	bank of america; lucky draw tattoo in phoenix; language classes seattle
Reference	hawaii capital; russell crowe height; homer simpson quotes	superman returns sequel; evel knievel death; iron man 2 release date
Time Zone	[No clicks possible]	what time is it in italy;germany time;
Translate	Translator; free translation; translate capabilities to spanish	english to chinese pronunciation; english to italian phrases
Twitter	snooki; youtube; spirit airlines; obama twitter	american airlines; youtube;facebook;cvs;
Weather	weather 92808; weather in vegas; weather;	weather; tucson az weather;

We saw many examples where the same query led to either a click on just the returned Answer or on just the search results. Examples of such queries include "weather", "germany time", "horoscope", "pga" and "flight status". Users can have many intents for the same query, and an Answer may not always address every user's intent. For highly engaging and interactive Answers, engagement with the Answer or the result can be indicative of what the user is seeking. In these cases, engagement with the search results instead of an Answer can indicates imperfect triggering. For example, the Finance Answer is probably inappropriate for the query "christian quotes", as is the Dictionary Answer for the query "what is death". And most people looking for the television show "big brother 12" probably do not want the News Answer. Search result content may also be preferred to Answer content when it is presented in a more appealing manner. The search result snippets for specific newspapers, for example, contain much more information than the corresponding links shown in the Newspaper Answer.

Several Answers that have very low Answer engagement also have very low engagement with the search results. For example, the Dictionary Answer is interacted with only 0.31 times as often as the average Answer type, but people are only likely to click on a search result 0.30 times as well. Similar behavior can be observed for the Currency Answer. In these cases, the presence of the Answer appears to be *cannibalizing* clicks from the search results. The absence of interaction with traditional search results can be seen as an indication that the Answer is meeting the user's need, even in the absence of direct interaction. Other Answers, like the Twitter Answer, have low Answer engagement but high search result engagement. In these cases the Twitter content is probably supplemental or not relevant to the user's direct task.

The Answer types that are most likely to receive both Answer and search result clicks are the Horoscope, Attractions and Dictionary Answers. This might indicate these Answers are sometimes relevant to the query that triggers them, but not sufficient to fully satisfy the searcher's need.

4.3 Repeat Usage of Answer Types

Although many Answer types receive very little direct user interaction during any given query, we found that repeated use by the same user of the exact same Answer (e.g., the Weather Answer for weather in Boston, or the Finance Answer for a specific company's stock price) or the same Answer type (e.g., a lookup of a phone number or a flight's status) could suggest that the user was finding value in the specific Answer or Answer type. In this section, we explore what repeat usage can tell us about Answers. In the subsequent section, we look more closely at whether people re-used particular Answer types to return to a class of information or specific content by looking at whether they usually triggered an Answer with the same query or different queries.

Earlier we discussed how different Answers trigger with different probabilities. The Phonebook Answer, for example, is very popular and appears hundreds of times more than less popular Answer types like the Time Zone or Reference Answer. We now look at how often an individual triggers a query over time, rather than how often the query triggers as a whole. For example, a frequent air traveler may search for Flight Status for "ua 600" many times in one session looking for status updates, and "alaska 240" later in the week. We show that some Answer types were used occasionally by all people, while others were used repeatedly by the same user.

If Answer behavior were independently and identically distributed across queries, the number of times we would expect to see an individual trigger the same Answer would follow a multinomial distribution. Not surprisingly, however, none of the Answer types we studied appeared to match a multinomial distribution. All of our Answer types were clustered by user in a way that would be almost impossible if Answer triggering were i.i.d.

Although the Answer types co-occur much more than expected, we would like to know which co-occur more than others. This is challenging because popular Answers will inherently cluster to some degree. For example, given the Phonebook and News Answers triggers fairly often, it is not surprising when we observe an individual issuing queries that return those Answer types two or three times. However, the Time Zone Answer triggers rarely, so it is surprising when the Answer is observed being triggered by

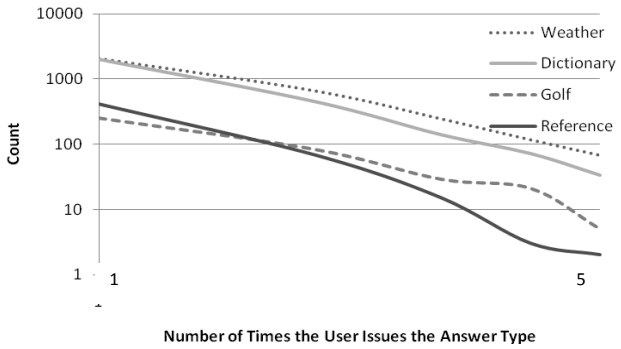


Figure 4. How often four different Answer types are issued multiple times by the same person, shown on a log-log scale.

the same person two or three times. To understand what information is contained in a particular Answer type being used by the same individual multiple times, we must account for the triggering frequency of the Answer.

Figure 4 shows the distribution of Answer co-occurrence for four of our Answer types. The curves show the number of users who issued exactly 26 queries and triggered each Answer type exactly *x* times. The Weather and Dictionary Answers are higher than the Golf and Reference Answers because they are more popular Answers, and thus are more likely to be triggered in the course of 26 queries. However, it is also evident that the Weather Answer has a heavier tail than the Dictionary Answer, meaning the same user is more likely to issue the Weather Answer multiple times than the same user is to issue the Dictionary Answer. Similarly, the Golf Answer has a heavier tail than the Reference Answer, meaning the same user is likely to issue multiple Golf Answers compared to the Reference Answer.

The curves in Figure 4 are shown on a log-log scale, and appear fairly straight, suggesting that a power-law distribution would be a much better fit for the data than a multinomial distribution. Power-law distributions can be represented as:

$$f(x) = ax^k$$

where *a* and *k* are constants. The value of *k* represents the slope of the specific distribution. What is nice about power-law distributions is that they exhibit scale-invariance, meaning that rescaling the function's argument preserves the shape of the function. Queries that trigger Answers very often can have the same slope and be either very popular or unpopular. If we use *k* to represent the "heaviness" of the tail, we can compare tail weight across Answers of different popularity.

For this reason, we fitted a power-law distribution to each Answer type. Because uncommon Answers were unlikely to co-occur many times together despite our large number of observations, we used only the first five points in each curve to do the fit.

The value of *k*, or the slope or heaviness of the tail for each Answer type, is shown in Table 3. The higher the number, the less steep the slope and the heavier the tail. We can see, for example, that, as expected given Figure 4, the Weather Answer (*k*=-2.15) has a heavier tail than the Dictionary Answer (*k*=-2.52), and the Golf Answer (*k*=-2.25) has a heavier tail than the Reference Answer (*k*=-3.45).

Table 3. The tail strength of each Answer type. Answers with a high tail strength are repeated more by the same user than answers with a low tail strength are.

Answer	Strength (*k*)	Answer	Strength (*k*)
Phonebook	-1.12	Dictionary	-2.52
Flight Status	-1.70	Newspaper	-2.54
News	-1.81	Finance	-2.58
Weather	-2.15	Currency	-2.89
Golf	-2.25	Attractions	-2.93
Twitter	-2.28	Time Zone	-3.04
Translate	-2.31	Reference	-3.45
Horoscope	-2.46		

In general, the Phonebook Answer, the Flight Status Answer, and the News Answer are very likely to cluster by user. News and Phonebook are seen by many of the users, and also have a strong set of users who use them heavily. However, the Flight Status and Weather Answers are seen by relatively few users, but within the set of users are large contingent that use them often. Assuming the users are intentionally triggering these Answers, their behavior is a signal that the users who trigger them find value in them.

On the other hand, the Reference Answer, the Time Zone Answer, and the Attraction Answer are relatively less likely to cluster by user. It is possible that fewer users, once discovering the Answer, have strong use cases that would merit repeated use.

4.4 Query Similarity in Answer Triggering

When an Answer type is triggered more than once by an individual, it can be useful to know the relationship between the queries that individual used to trigger it. In this section we look at whether repeat Answer types are triggered by repeat queries, related queries, or entirely different queries.

To do this, for each Answer type we identified all users who triggered the Answer type exactly four times (5% of all users). We took the four queries they issued that triggered the Answer, and calculated three values to express the similarity of the set. We disregarded the order in which the queries were issued, and generated a list of the six possible query pairs from the four queries. Each pair can be either:

- ***Repeat*** The two queries in the query pair are identical.
- ***Overlapping*** The two queries overlap by at least one term, but are not identical.
- ***Different*** The terms in the two queries are completely different.

For example, in the following set of queries that trigger the Weather Answer {"weather", "weather", "weather boston", "wether"}, there is one repeat pair, two overlapping pairs, and three pairs that are completely different. We sum across all users to determine how the query pairs of each Answer type relate to each other. The results can be seen in Figure 5.

Queries can, of course, sometimes appear different when they are related or intended to be the same. For example, "horoscope" gets misspelled often, but still triggers the Horoscope Answer. For the Phonebook Answer, the pair "big timber campground" and "seeley lake cabins" are likely related. For the News Answer, "body left in hearse 9 days" and "las vegas hoarder found dead"

are related in that they are both examples of morbid news items. But we find these metrics to be a valuable approximation.

4.4.1 Answers Triggered with Repeat Queries

The Answer types with the highest percentage of identical query terms are: Twitter, Newspaper, Horoscope, Finance and Weather. When Twitter and Newspaper queries were repeated, the intent often appeared to be navigational. For example, the Newspaper Answer triggers repeatedly for "tulsa newspaper," and the Twitter Answer triggers repeatedly for "american airlines", "pbs kids" and "youtube." Navigational queries have been shown to be repeated commonly [18], and this use is supported by what we observed in the interaction data for these queries. In these cases, the Answer information is probably supplemental at best.

It makes sense that users are fairly consistent in the way they trigger the Weather Answer; users are probably most interested in their local weather, and issue the same query that they know will trigger the Answer. The Finance Answer is similar; people are often interested in a limited set of stocks, or even one stock in particular that they want to check in this manner. The Horoscope Answer has limited triggers, so we are not surprised that users use the same query to trigger it.

The Answer types with the lowest percentage of identical queries are the Dictionary, News, and Phonebook Answers. People most likely do not want to find the same word definition, news item, or local business multiple times. When queries with these Answers are repeated, users are sometimes looking for updated information. For example, Phonebook-triggering query "jobs in 08210" is issued multiple times, as is the News-triggering query "lottery ticket". The Dictionary Answer does not typically update. It is sometimes used repeatedly for seemingly normal definitions such as "definition of nuance." Other repeated Dictionary queries do not appear to be intended to trigger the Answer. Examples include "what is time", "the apocrypha definition", or "what is the self".

The Reference Answer gives fairly definitive answers to queries such as "israeli currency" "stand by me director" and "aa milne birthday" and yet they are queried multiple times. A surprisingly high 43% of users' Reference query pairs are identical. Some of this is probably due to repeat interest (consider the query "hannah montana songs").

4.4.2 Answers Triggered by Overlapping Queries

The Answers that have a large number of query pairs that overlap are the Dictionary, Reference, Weather, Attractions and Phonebook Answers. The Dictionary, Weather and to some degree, Attractions Answers are usually triggered by a smallest of words that are shared between many of the queries, both within a user and across users. Almost all Dictionary Answer queries contain the word "define" or "meaning." Many Weather Answer queries contain the word "weather." Finance Answers are likely to contain "stock", and Phonebook Answers sometimes contain a city, for example "spca Cincinnati" and "channel 12 cincinnati". Such queries are unrelated in any way other than by the word that triggers the Answer.

However, the Reference Answer is more interesting. Users seem to issue several queries of the same kind of response (e.g., "john wayne death" and "gary cooper death", or "number the stars author" and "wrinkle in time author", or "barack obama's birthday" and "bill clinton's birthday"). This suggests that the users are consciously triggering the Answer. The Reference Answer can be hard to discover, and it may be that when a user stumbles upon a particular way to trigger it, they later seek to use it again.

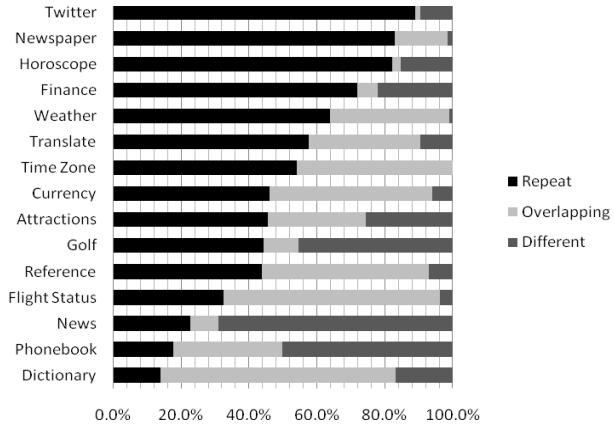


Figure 5. Percentage of users' query pairs that are repeat, overlapping, and different.

The Phonebook and Attractions Answers are also often triggered with related queries, such as "oak alley plantation" and "nottoway plantation" (Attractions) and "medford oregon hotels" and "hotels in westley ca" (Phonebook). These queries look like they are being used within a session to accomplish a task. To a lesser extent, related queries that seem to be part of a larger task also exist for the News Answer ("brandy sued" and "jersey shore sued", or "miss mexico 2010" and "mexico drug war") and the Currency Answer ("50 peso" and "20 peso").

4.4.3 Answers Triggered with Different Queries

Overwhelmingly, the Answers with the highest percentage of different query pairs are Phonebook and News. This makes sense; a large number of needs that trigger these Answers are fairly unique. The Golf Answer has a large fraction of query pairs with no shared terms, usually the name of tournaments.

4.5 Session versus Cross-Session Answer Use

In the previous two sections, we have looked at how individuals reuse Answers. In this section, we look more closely at how Answers are reused within a single session as compared to across all of the queries an individual issues.

The first column of Table 4 shows the average number of times each Answer type was triggered in a session of any length, given that the Answer was triggered at least once. For example, the Dictionary Answer was often used many times in a single session, while the Twitter Answer was typically used just once in a session. The second column of Table 4 how often users repeated queries to trigger the Answer, as presented in Figure 5. We compare this to the third column, which reports what percentage of query pairs were repeated within all sessions of length four. We see, for example, that the queries that triggered the News Answer were relatively similar within a single session, and relatively dissimilar across all of a user's sessions.

The Dictionary and Flight Status Answers are both triggered multiple times per session. We observe that people often look for

Table 4. The average number of times the Answer is observed by an individual (given it is seen at least once) compared to the percentage of query pairs that are repeated within a user or session.

Answer	Occ. per User	Repeat Query Pairs	
		Within Users	Within Sessions
Dictionary	2.64	14%	13%
Flight Status	1.84	33%	26%
Phonebook	1.48	18%	24%
Golf	1.47	44%	32%
News	1.30	23%	40%
Attractions	1.29	45%	39%
Newspaper	1.29	83%	78%
Translate	1.28	58%	47%
Weather	1.26	64%	44%
Currency	1.25	46%	30%
Finance	1.24	72%	42%
Reference	1.22	44%	45%
Horoscope	1.21	82%	84%
Time Zone	1.20	54%	56%
Twitter	1.17	89%	89%

many definitions at the same time. The common repeat use of the Flight Status Answer within a session may represent individuals researching flights or monitoring an upcoming flight.

The Answers that on average trigger the least number of times per session are Reference, Horoscope, Time Zone, and Twitter. It is unsurprising that Horoscopes are queried for, on average, just once per session (and probably once per day). Since many of the queries that trigger the Twitter Answer are navigational, it is also not surprising that the Twitter Answer appears once per session, most of the time.

When the percentage of repeat query pairs differs between within a user and within a session, it provides a strong indication that the Answer is being used differently in sessions. The Finance Answer has a very pronounced difference; query pairs are more likely to be similar within a user than within a session. This is consistent with the behavior of having a constant set of stocks that one person checks consistently over time, but each session in which that person checks, a great variety of stocks are seen. A similar story applies to the Currency and Weather.

The opposite is true of the News and Phonebook Answers. Here the query pairs are more likely to be identical within the same session. This suggests users are searching for more information on the same topic at any particular time (e.g., a breaking story or a particular local business type), but a greater variety over time.

5. DISCUSSION

We have looked at which Answers are triggered the most, interacted with the most, and used over and over again by the same person the most. We have seen that the relevance of an Answer to a query must be judged according to the nature of the Answer type. Some Answer are engaging, providing links to valuable information or even requiring the user to click to retrieve the desired information. For these Answers, interaction data can useful. Many of these Answer types provide a list of curated or structured examples. For example, the Attractions Answer shows a set of attractions in an area, and the News Answer shows a set of relevant news articles. These often complement the algorithmic results, and click data with these results can be understood in a similar way to how click data is understood for search results.

Other Answers types, like the Time Zone and Currency Answers, provide a clear inline Answers to unambiguous queries. In these cases, user interaction is not expected. Instead, we have seen that it may be possible to use clicks on the algorithmic search results (rather than the Answer) as a negative sign of satisfaction. It can be particularly difficult to assess the satisfaction users derive from Answers that provide diversity in results, especially for Answers with low interaction. For example, if a person wants to learn about a celebrity, they may not engage with the Twitter Answer returned and may choose to click on a search result, but still receive value from seeing Tweets by that celebrity embedded in their search result page. We believe that repeat triggering of an Answer by the same user can be a sign of satisfaction. Repeat engagement could be useful for understanding any search system where direct interaction with the search results is difficult to capture (e.g., sentence retrieval engines or Twitter search engines).

However, Answer interaction data is valuable for determining which specific queries appropriately trigger the Answer and which did not. In some cases, Answers with limited interaction could be designed to encourage greater interaction. The Translate Answer, for example, gets significant click through in part because of the prominence of the link to the translator in the Answer. Interaction does not need to come in the form of links. For example, if the Phonebook Answer allowed the user to directly place a call by clicking on a telephone number, this interaction could be captured by the search provider.

The links that are clicked for Answers that receive significant interaction may suggest additional content that could be displayed directly on the search result page. For example, the Weather Answer is clicked a lot compared with other Answers that similarly provide content directly on the result page. The content found following these clicks could be pulled up into the Answer, so that if everyone clicks on the hourly forecast link, the Answer could be modified to provide hourly weather information. Potential modifications to the content displayed by an Answer could be tested by including it via links and then measuring the relative engagement. Similarly, the search results that are clicked in conjunction with Answers can provide clues as to what additional information might be useful to include in an Answer. Although the Movie Showtimes Answer is not discussed in this paper, we observed that people who engage with it are very likely to click on Fandango.com as well. This suggests that the Answer could provide additional value to users if it were augmented to support movie ticket purchases.

We observed that some Answer types (and some specific Answer-query pairs) were used over and over again, either by the same user or in the same session. We believe that there is an opportunity here to personalize the user's experience with the Answer. For Answers with high engagement, Answer click data could be used as a guide as to the content that might best be included in the Answer. A user who always engages with the list of attractions provided by the Attractions Answer might be provided with a longer list of attractions, or a user that always clicks on the Virgo Zodiac sign of the Horoscope Answer might later be shown the Virgo reading directly in the context of the search result page. Answers that people sometimes monitor over

time for the same query, like the Twitter, Weather, and Finance Answer could also provide information about what has changed since the last time the user issued the query.

For users that trigger or engage with a particular type of Answer frequently, it may make sense to raise the rate of occurrence of that Answer for that user. For example, somebody who regularly looks up stock quotes may want the query "bank of america" to trigger the Finance Answer, even though it does not for most people. Additionally, some Answers are often used in tandem to complete a task (the Phonebook and Weather Answers are an example). There may be ways enable sets of Answers that are used to complete tasks work together.

A big challenge for Answers is discoverability. Many of the Answers we studied were seen by only a small percentage of our users during the one-week study period. While some of these are probably of interest to small subsets of the population (e.g., fans of professional golf), others are more broadly applicable but hard to discover. Did you know that if you entered a package tracking number into a search engine, you would be told directly where your package is? Or that if you entered your flight number you would learn whether your flight has been delayed? Or that the query "william shatner height" will inform you that he is 5' 9" tall? Answers work well to address the needs that people currently express, but do not work well for needs that people do not yet know to ask about. Discoverability could be improved by suggesting hard-to-discover Answer types when a user issues a related-but-common query (e.g., a user who issues a travel-related query could be informed of the existence of the Flight Status Answer), or by connecting Answer information to existing user content (e.g., emails often contain flight information and package tracking numbers).

6. CONCLUSION

In this paper, we have used large scale query log analysis to explore the challenges a search engine faces when trying to meet an information need within the context of the search result page. We looked at how people's interaction behavior with search result Answers differs from their interactions with typical search results, and found that inline Answers can cannibalize clicks from the algorithmic results. We saw that in the absence of interaction behavior, an individual's repeat search behavior can be useful in understanding the information's value. We also looked at some of the ways user behavior might provide insight into when inline Answers might better trigger and what types of additional information might be included in the results.

The analysis here focused on observable behavior. However, to really understand how users interact with Answers, we need to know what the user is thinking when they trigger an Answer. We plan to conduct a user study informed by what we have learned via log analysis to understand, for example, whether Answer use is intentional and when the presence of an Answer provides negative or positive peripheral value to the search results. It is our hope that a rich understanding of Answers will be useful as Web search engines move towards the long-standing goal of directly addressing searchers' needs.

7. ACKNOWLEDGMENTS

We appreciate the valuable feedback and technical assistance from Dan Liebling, Gheorghe Muresan, Reid Andersen, Alec Berntson, James McCaffrey, Filip Radlinski, and Shane Williams.

8. REFERENCES

[1] Agichtein, E., S. Lawrence, and L. Gravano. Learning to find answers to questions on the Web. *TOIT* 4 (2), 2004.

[2] Bailey, P., N. Craswell, R. W. White, L. Chen, A. Satyanarayana, S. M. M. Tahaghoghi. Evaluating search systems using result page context. IIiX 2010, 105-114.

[3] Cutrell, E. and Z. Guan. What are you looking for?: An eye-tracking study of information usage in Web search. CHI 2007, 407-416.

[4] Dupret, G. and C. Liao. A model to estimate intrinsic document relevance from clickthrough logs of a Web search engine. WSDM 2010, 181-190.

[5] Feild, H. A., J. Allan, and R. Jones. Predicting searcher frustration. SIGIR 2010, 34-41.

[6] Fox, S., K. Karnawat, M. Mydland, S. Dumais, and T. White. Evaluating implicit measures to improve Web search. *TOIS*, 23(2), 2005, 147-168.

[7] Hassan, A., R. Jones, and K. L. Klinkner. Beyond DCG: User behavior as a predictor of a successful search. WSDM 2010, 221-230.

[8] Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *TOIS* 25(2), 2007.

[9] Kelly, D. and J. Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum* 37 (2), 2003, 18-28.

[10] Li, J., S. B. Huffman, and A. Tokuda. Good abandonment in mobile and PC internet search. SIGIR 2009, 43-50.

[11] Lin, J. An exploration of the principles underlying redundancy-based factoid question answering. *TOIS* 25 (2), 2007.

[12] Radlinski, F., M. Kurup, and T. Joachims. How does clickthrough data reflect retrieval quality? CIKM 2008, 43-52.

[13] Sarma, A., S. Gollapudi, and S. Ieong. Bypass rates: Reducing query abandonment using negative inferences. KDD 2008, 177-185.

[14] Stamou, S. and E. N. Efthimiadis. Queries without clicks: Successful or failed searches? Workshop on the Future of IR Evaluation, SIGIR 2009.

[15] Stamou, S. and E. N. Efthimiadis. Interpreting user inactivity on search results. ECIR 2010, 100-113.

[16] Teevan, J., E. Adar, R. Jones, and M. Potts. Information re-retrieval: Repeat queries in Yahoo's logs. SIGIR 2007, 151-158.

[17] Tombros, A. and M. Sanderson. Advantages of query biases summaries in information retrieval. SIGIR 1998, 2-10.

[18] Tyler, S. K. and J. Teevan. Large scale query log analysis of re-finding. WSDM 2010, 191-200.

[19] Wang, K., T. Walker, and Z. Zheng. PSkip: Estimating relevance ranking quality from Web search clickthrough data. KDD 2009, 1355-1364.

[20] Wen, J.-R., J.-Y. Nie and H.-J. Zhang. Query clustering using user logs. *TOIS* 20 (1), 2002.

[21] White, R. W. and S. T. Dumais. Characterizing and predicting search engine switching behavior. CIKM 2009.