Finding Our Way on the Web: Exploring the Role of Waypoints in Search Interaction

Ryen W. White
Microsoft Research
Redmond, WA 98052 USA
ryenw@microsoft.com

Adish Singla
Microsoft Bing
Bellevue, WA 98004 USA
adishs@microsoft.com

ABSTRACT

Information needs are rarely satisfied directly on search engine result pages. Searchers usually need to click through to search results (landing pages) and follow search trails beyond those pages to fulfill information needs. We use the term *waypoints* to describe pages visited by searchers between the trail origin (the landing page) and the trail destination. The role that waypoints play in search interaction is poorly understood yet can be vital in determining search success. In this poster we analyze log data to determine the arrangement and function of waypoints, and study how these are affected by variations in information goals. Our findings have implications for understanding search behavior and for the design of interactive search support based on waypoints.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, selection process.*

General Terms

Measurement, Experimentation, Human Factors

Keywords

Waypoints, search trails

1. INTRODUCTION

Web search engines retrieve ranked lists of Web pages based on user queries. However, the contents of these ranked lists represent only starting points from which users can explore retrieved search results. Research in areas such as orienteering [4], berrypicking [1], and information foraging [3], has shown that trails followed by searchers after they leave the search engine are important in determining search success [8]. Pages that users visit on these trails, so-called *waypoints*, play an important role in helping users reach their information seeking goals. Despite the importance that they appear to have, little is known about the role that waypoints play in the search process. In this poster we present the findings of a log-based study into the role of waypoints in search interaction.

2. TRAILS AND WAYPOINTS

The primary source of data for this study was the anonymized logs of URLs visited by users who consented to provide interaction data through a widely-distributed browser plugin. Log entries include a unique user identifier, a timestamp for each page view, an identifier for each browser instance, and the URL of the Web page visited. Revisits to pages made through the browser "back" button were also captured. To remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. The results described here are from a sample of the URL visits from May 2010 to July 2010, totaling billions of URLs from millions of Web searchers. From these logs, we mined search trails [7], hereafter referred to as T_A. Each trail is represented as a temporally-ordered URL sequence connected by hyperlink clicks, followed by a single user. Trails start with a search engine query followed by a click on a search engine results (trail *origin*). Trails terminate with another query, once the tab/browser is closed, or following a period of user inactivity of 30 or more minutes, suggesting that the active task has been terminated. The terminal page is referred to as the trail *destination*. Waypoints reside on trails between origins and destinations. Across all of T_A, 62% of trails have a waypoint and 12% of waypoint visits come from within-trail revisitation.

3. CHARACTERIZING WAYPOINTS

Through manual labeling of a sample of five thousand randomly-selected search trails with waypoints (T_L) (each with a unique query) we aim to characterize the arrangement and function of waypoints in trails. During labeling one of the authors reviewed trail URLs and dwell times, and labeled how the waypoints were arranged in the trail (e.g., in a hub-and-spoke formation) and the function that the waypoint appeared to have (e.g., navigational).

Three main arrangements of waypoints within search trails were observed: (i) linear (no backtracking, 28 % of trails), (ii) hub-and-spoke (22%), and (iii) branched (20%). Figure 1 presents behavior graphs illustrating these arrangements. Other arrangements such as undirected (general browsing with occasional backtracking, 13%) and hybrid (mixtures of arrangements, e.g., linear then branched, 17%) were also observed. However, due to space constraints we focus analysis on the three primary arrangements.

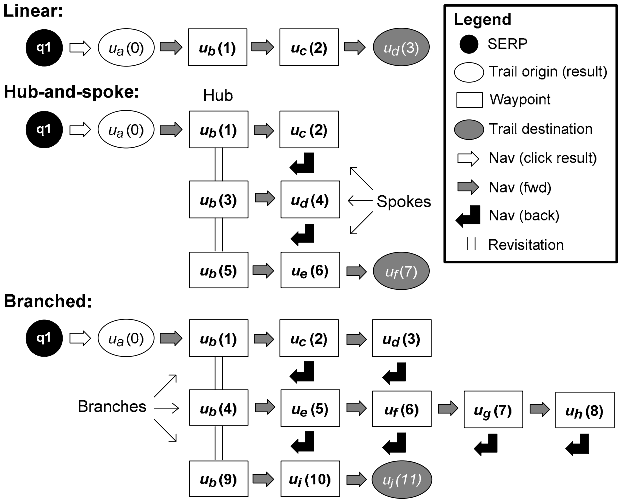

Figure 1. Linear, hub-and-spoke, and branched arrangements.

Linear: Waypoints were visited in a linear sequence with no backtracking. The functions of the waypoints that were arranged in this way were navigational (21%) (i.e., pages that users passed through, usually briefly, en route to a destination), bottlenecks

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

(16%) (i.e., pages that users *had to visit* to get to a target resource such as a download)[1], pagination (14%) (i.e., visiting search result pages in a sequence defined by the website (page1, page2, etc.)), topic exploration (18%) (i.e., visiting multiple sites on a given topic within domain or across domains), site-search refinement (15%) or requisite steps for task completion (14%) (i.e., page visits in a logical sequence as defined by the site or by the user). Linearly-arranged waypoints were also visited unintentionally due to auto-redirects from emails and logins (2%).

Hub-and-spoke: Waypoints were also arranged as hubs and spokes where searchers navigated to a single page (the hub) and then out to another page (the spoke), and then back to the hub, and then out to another spoke, and so on. As an example, hub waypoints might be a site search result page and spoke waypoints might be the results within the site that users navigate to.

Branched: Waypoints were also arranged as branches within a trail. A branch is a linear sequence of page visits (as defined in [7]) and branching is effectively a combination of linear and the hub-and-spoke. Users may still have one or more pages that they revisit frequently in the trail, but the hops away from those hubs contain multiple steps, with the waypoints on each branch often having characteristics similar to a linear arrangement.

Information goals have been shown to dramatically affect search behavior [2,7,8]. We next studied the impact of varying goals on the arrangement and function of waypoints.

4. INFORMATION GOAL EFFECTS

To represent different types of information goal we selected subsets of T_L trails with varying origin and destination entropies. Query-click entropy has been used to capture variations in the number of unique search engine result page (SERP) URLs clicked for a query [5]. We define entropy for the origins and the destinations across all trails in T_A for each query (q) in T_L as:

$$Entropy(q) = -\sum_{URL\,u} p(c_u|q) \cdot log_2(p(c_u|q))$$

Where u is the origin or the destination depending on whether we are computing origin or destination entropy, and c_u is a click to u.

Queries with high origin entropy will have many landing pages, suggesting that the query may be informational or ambiguous [5]. In contrast, low origin entropy suggests that the query might be unambiguous and associated with navigation or a known-item searching. Queries with a high destination entropy will have many possible trail end points, suggesting that the overall goal may be ambiguous [2], and those with low destination entropy have clear goals and may benefit from teleportation to the destination [6].

We analyzed the role of waypoints given different origin and destination entropies. By varying *both* origin and destination entropies as part of the study we aimed to more closely control for different types of information goal than would be obtained by just varying either type alone. From T_L we selected four trail subsets:

1. **Origin clear + Destination clear:** Trails in lowest 10% (least entropic) of origin entropy, lowest 10% of destination entropy.
2. **Origin clear + Destination unclear:** Trails in lowest 10% of origin entropy and in highest 10% of destination entropy.
3. **Origin unclear + Destination clear:** Trails in highest 10% of origin entropy (most entropic), lowest 10% destination entropy.
4. **Origin unclear + Destination unclear:** Trails in highest 10% of origin entropy and in highest 10% of destination entropy.

We used the human labels assigned to trail waypoints in T_L, and computed the most popular trail arrangements and functions for each of the four subsets. Table 1 shows the most popular two waypoint arrangements and functions for each group and their associated percentages (of the trails in each group).

Table 1. Waypoint arrangement(function) given entropic variations.

Origin	Destination	
	Clear	*Unclear*

Clear	Linear(navigation) (63%)
	Linear(task steps, site) (26%)
		Hub-and-spoke (47%)
		Linear(pagination) (27%)

Unclear	Linear(bottlenecks) (43%)
	Linear(task steps, user) (35%)
		Branched (59%)
		Linear(exploration) (24%)

The findings show marked differences in the role of the waypoints with different goals. When both origins and destinations are clear, waypoints appear to be arranged linearly, with users primarily navigating to a resource on their own or being guided through steps by a website. When the origin is unclear but the destination clear, the waypoint arrangement is still mostly linear; despite multiple starting points, users may encounter bottlenecks through which they may need to pass. When the origin is clear and the destination is unclear, waypoints are often arranged in a hub-and-spoke formation or linearly as users paginate through options. When both the query and the destination are unclear, waypoints are arranged in branches or linearly as searchers explore the topic.

5. CONCLUSIONS

We have presented a study characterizing the role of waypoints in search interaction. We illustrated some arrangements of waypoints and the functions that they can have. Our findings can help inform waypoint selection algorithms that can be used to present waypoints on SERPs. Observed differences in the arrangement and function of waypoints based on the nature of the information goal suggests that we should consider goals when selecting query waypoints. In future work we will explore the development of waypoint recommendation algorithms and conduct user studies on integrating waypoints directly into search engine result pages.

REFERENCES

1. Bates, M.J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5): 407-424.
2. Downey, D., Dumais, S., Liebling, D. and Horvitz, E. (2008). Understanding the relationship between searchers' queries and information goals. *Proc. CIKM*, 449-458.
3. Pirolli, P. and Card, S.K. (1999). Information foraging. *Psychological Review*, 106(4): 643-675.
4. O'Day, V. and Jeffries, R. (1993). Orienteering in an information landscape: how information seekers get from here to there. *Proc. INTERCHI*, 438-445.
5. Teevan, J., Dumais, S. and Liebling, D. (2008). To personalize or not to personalize: modeling queries with variation in user intent. *Proc. SIGIR*, 163-170.
6. White, R.W., Bilenko, M. and Cucerzan, S. (2007). Studying the use of popular destinations to enhance web search interaction. *Proc. SIGIR*, 159-166.
7. White, R.W. and Drucker, S.M. (2007). Investigating behavioral variability in web search. *Proc. WWW*, 21-30.
8. White, R.W. and Huang, J. (2010). Assessing the scenic route: measuring the value of search trails in web logs. *Proc. SIGIR*, 587-594.

[1] We distinguished bottlenecks from other navigation by visiting cached versions of the pages in question (from the same timeframe as T_A) to ensure that the outlink click was necessary to follow the trail.