Exploring the Use of Labels to Shortcut Search Trails

Ryen W. White and Raman Chandrasekar
Microsoft Research
One Microsoft Way, Redmond, WA 98052 USA
{ryenw, ramanc}@microsoft.com

ABSTRACT

Search trails comprising queries and Web page views are created as searchers engage in information-seeking activity online. During known-item search (where the objective may be to locate a target Web page), searchers may waste valuable time repeatedly reformulating queries as they attempt to locate an elusive page. Trail *shortcuts* help users bypass unnecessary queries and get them to their desired destination faster. In this poster we present a comparative oracle study of techniques to shortcut sub-optimal search trails using labels derived from social bookmarking, anchor text, query logs, and a human-computation game. We show that labels can help users reach target pages efficiently, that the label sources perform differently, and that shortcuts are potentially most useful when the target is challenging to find.

Categories and Subject Descriptors

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

General Terms

Measurement, Design, Experimentation, Human Factors

Keywords

Trails, labels, click graph, anchor text, social bookmarks

1. INTRODUCTION

In recent years, there has been significant interest in using social media and community behavior to improve Web search. Annotations from social bookmarking sites [2], anchor text garnered from Web crawls [3], and query logs from search engines [7] (collectively referred to here as *labels*) have been included as additional content for result ranking. However, labels may also be useful for other purposes such as query recommendations [1,5] to help users experiencing difficulty in formulating queries required to find specific pages, perhaps due to vocabulary mismatches [4].

In this poster, we present a log-based study on the effectiveness of labels as navigational *shortcuts* to reduce the average number of query refinements in search trails and to help users get to a destination page in fewer queries. Previous work has shown that popular destinations can help users search more effectively [8]. We extend that research to support searchers in situations where they may struggle to find a destination page and perform a large number of query refinements. A log-based methodology allowed us to carefully compare methods for shortcutting such sub-optimal trails using large numbers of real searching episodes providing evidence of users targeting specific Web pages. Our study answers two questions: (i) can labels reduce the average number of query refinements to reach a desired destination page? and (ii) if labels do help, how do the label sources compare?

2. LABEL SOURCES

Trail shortcuts use labels assigned to target Web pages to help users bypass unnecessary intermediate queries appearing between the first and the last queries in the trail. There are many sources from which we could derive page labels. We selected four representative sources for comparison in this study: (i) query data from a human computation game, (ii) queries from a commercial Web search engine click-graph, (iii) anchor text extracted from a commercial search engine index, and (iv) social bookmarks. We now describe each of the label sources in more detail.

Page Hunt Queries: Page Hunt [6] is a human computation game operating like search in reverse. In the game, the player is shown a random Web page, and challenged to construct a query that would return this page as one of the top results on a search engine. In effect, the player is identifying labels for the Web page shown to her. The Page Hunt query dataset has information about 577 URLs (henceforth referred to as *Page Hunt URLs*) and 23,807 queries related to these URLs, where the user successfully found the target page. As part of that work, the authors defined a metric called *findability*, which measures how easy it is for users to find a given page – if everyone can hunt down a page, it has 100% findability and if no one is able to find it, it has 0% findability.

Click-graph Queries: A click graph is a bipartite graph of search engine users' queries and the result URLs they clicked on, represented as a set of triples $< q, u, c >$, where URL u was clicked c times by users when they issued a query q. From a click graph collected from 18 months of Web search engine logs, where each URL had at least five clicks, we extracted all click data available for the Page Hunt URLs. This allowed us to create the Click-graph dataset consisting of 546 URLs and 491,689 queries and counts associated with these URLs.

Anchor Text: Anchor text refers to the visible, clickable text (often underlined) in a hyperlink on a Web page. Anchor text is used by Web site authors to provide a contextually relevant description or label for the Web page (landing page) it is linked to. Anchor text is used by search engines as additional metadata to rank landing pages. For our experiment, we identified the anchor text for the Page Hunt URLs from crawls performed by a major web search engine. We found label and count data for 512 of these URLs giving us 77,663 rows of anchor text data.

Social Bookmarks: Social bookmarking services allow users to store bookmarks for Web sites, and share and discover other bookmarks. One example of such a service is delicious.com. Users of the service can bookmark any site with tags and get to the site using those tags. Users can send bookmarks to others, keep track of users and tags, and view popular tags and sites. In this experiment, we downloaded all the tags available from delicious.com for the Page Hunt URLs using their programmatic interface; this gave us 6,787 labels corresponding to 364 URLs.

3. EXPERIMENT

We performed an oracle study to determine if query shortcuts created from label sources help users reach their goal faster.

In our study we use search trails comprising queries and pages defined as in [9]. Trails were mined from six months of log data from consenting users of a widely-distributed browser toolbar. The information in these log entries includes a unique user identifier, a timestamp for each page view, and the URL of the web page visited. Intranet and secure (https) URL visits were excluded at source. Only entries generated in the English speaking United States locale were included. From these data, we extract millions of search trails where the Page Hunt URLs were visited.

To test each of the label sources, we first selected trails: (i) that had at least three queries, offering some scope for shortcuts; (ii) where all consecutive queries shared at least one term (signifying information need consistency in a similar way to *query chains* [7]); (iii) where queries did not contain spelling errors (to avoid situations where shortcuts may not help); (iv) that had no page visit until the last query (signaling potential dissatisfaction with all but the last query), and (v) where the last query led immediately to a visit to a Page Hunt URL. This gave us a test set of trails around 5% the size of the original sample. The average number of queries in these trails was 3.61 (median=3). Thus there was an opportunity to save users at least one query on average and perhaps more queries for pages that were more challenging to find. From the filtered trail set we created ten samples of ten thousand randomly-selected trails. Within each set we do the following:

For each label source L, for a destination URL u:

a. Select the top-20 most frequent labels for u from L.
b. Identify T, the set of trails that have u as destination URL.
c. Extract Q, the set of queries on each trail in T that is not the last query in the trail.
d. For each query in Q (starting with the first query and proceeding in temporal order), check if a query obtained by appending a top-20 label from L or one of the top-20 labels from L by itself matches the last query in the trail.
e. If there is a match, compute the distance saved in terms of the number of unsuccessful queries skipped.

The number of steps saved is averaged across all trails and runs.

4. FINDINGS

We present findings on the number of trail queries saved over all label sources and then broken down by source and by the findability of the destination page. Table 1 shows performance metrics for each labeling source and a combination (*Best*) that picks the most performant source for each trail based on the number of queries each source saves (randomly selecting a source in the case of ties). We report the average number of queries saved over trails where at least one label source helps (which is 27.3% of all trails sampled), as well as source coverage over those sampled trails. Also shown is the fraction of the ideal number of queries that could be saved in each trail (e.g., if a trail has four queries, then the ideal number saved is two, jumping from first to last).

The findings suggest that if presented as shortcuts, labels could shorten sub-optimal search trails by around two queries for almost 20% of such trails (bottom row of Table 1). The findings also show that anchor text saved a greater fraction of possible queries and that the click-graph covered more of the sampled trails. Similar trends in the results were observed across all trails (not just those where at least one source helped) and for those trails where all sources offered a shortcut. Additional analysis was conducted for destinations with high ($> 40\%$) and low findability scores. The results, summarized in Table 1, show that for *Best*, on average, 2.65 steps were saved for pages with low findability, versus 1.48 steps for pages with high findability; a trend mirrored by the sources individually. Thus trail shortcuts may help users more when they seek hard-to-find Web pages. Given the large sample sizes, all differences between sources were significant at $p < .01$ with ANOVA and Tukey post-hoc testing where appropriate.

Table 1. Number of queries saved and percentage of ideal for all saved trails and trails with high/low destination findability.

Source	All destinations		High find.	Low find.
	Avg. num. queries saved (% ideal)	% cov.	Avg. num. queries saved (% ideal)	Avg. num. queries saved (% ideal)
Page Hunt	1.58 (90.9)	10.7	1.27 (90.5)	2.01 (91.6)
Click graph	1.83 (89.9)	15.8	1.43 (89.3)	2.39 (90.7)
Anchor text	1.65 (93.1)	12.0	1.20 (93.0)	2.26 (93.2)
Bookmarks	1.54 (81.4)	5.6	1.19 (81.2)	2.02 (81.6)
Best (of all)	1.97 (95.2)	19.6	1.48 (94.8)	2.65 (95.4)

5. CONCLUSIONS AND FUTURE WORK

We have presented a study of using labels to shortcut search trails, in particular sub-optimal trails typified by multiple query reformulations. Findings show that labels could help users search more efficiently, especially when the target page is hard to find. When a search engine receives a query, the most frequent label assigned to pages in the result set could be shown as a shortcut on the result page. Future work will perform a more extensive analysis of the reported differences, study when adding a new term or substituting the query is more appropriate, use frequently-followed query chains extracted from log data for shortcut generation, and investigate the use of shortcuts for tasks beyond known-item search.

REFERENCES

[1] Anick, P.G. & Tipirneni, S. (1999). The paraphrase search assistant: terminological feedback for iterative information seeking. *SIGIR*, 153-159.

[2] Bao, S., Xue, G., Wu, X., Yu, Y., Fei, B. & Su, Z (2007). Optimizing web search using social annotations. *WWW*, 501-510.

[3] Craswell, N., Hawking, D. & Robertson, S.E. (2001). Effective site finding using link information. *SIGIR*, 250-257.

[4] Furnas, G.W., Landauer, T.K., Gomez, L.M. & Dumais, S.T. (1987). The vocabulary problem in human-system communication. *CACM*, 30(11): 964-971.

[5] Kraft, R. & Zien, J.Y. (2004). Mining anchor text for query refinement. *WWW*, 666-674.

[6] Ma, H., Chandrasekar, R., Quirk, C. & Gupta, A. (2009). Improving search engines using human computation games. *CIKM*, 275-284.

[7] Radlinski, F. & Joachims, T. (2005). Query chains: learning to rank from implicit feedback. *SIGKDD*, 239-248.

[8] White, R.W., Cucerzan, S. & Bilenko, M. (2007). Studying the use of popular destinations to enhance web search. *SIGIR*, 159-166.

[9] White, R.W. & Drucker, S. (2007). Investigating behavioral variability in web search. *WWW*, 21-30.