Tagging with Queries: How and Why?

Ioannis Antonellis
Computer Science Dept
Stanford University
antonell@cs.stanford.edu

Hector Garcia-Molina
Computer Science Dept
Stanford University
hector@cs.stanford.edu

Jawed Karim
Computer Science Dept
Stanford University
jawed@cs.stanford.edu

ABSTRACT

Web search queries capture the information need of search engine users. Search engines store these queries in their logs and analyze them to guide their search results.

In this work, we argue that not only a search engine can benefit from data stored in these logs, but also the web users. We first show how clickthrough logs can be collected in a distributed fashion using the http referer field in web server access logs. We then perform a set of experiments to study the information value of search engine queries when treated as "tags" or "labels" for the web pages that both appear as a result and the user actually clicks on. We ask how much extra information these query tags provide for web pages by comparing them to tags from the del.icio.us bookmarking site and to the pagetext. We find that query tags can provide substantially many (on average 250 tags per URL), new tags (on average 125 tags per URL are not present in the pagetext) for a large fraction of the Web.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.5 [**Information Storage and Retrieval**]: On-line Information Services; H.1.2 [**Models and Principles**]: User/Machine Systems - *Human information processing*

General Terms

Design, Experimentation, Human Factors, Measurement

Keywords

query tags, query logs, click logs, tagging, web search, web navigation

1. INTRODUCTION

Search engines analyze every piece of information available on the web to guide their search results. Such information comes mainly from two sources: (a) the web page creators and (b) the web users. Web page creators produce most of the *page content* and take part in the creation of the *link structure* on the web. Web users are implicitly providing new knowledge, through the queries they pose and the navigational paths they follow; such information is in part available in a search engine's *query and clickthrough logs*. An additional type of user generated content which is recently becoming available is *URL tags* from social bookmarking sites (e.g., del.icio.us, StumbleUpon).

All web content, except for the search engine logs, is created with the goal of enabling web users to access information. For example, the page content of a news article on cnn.com contains valuable information and is accessible to everybody. Hyperlinks present in that article are also useful for every reader as they enable them to navigate on relevant (according to the article's author) web pages. Finally, del.icio.us tags, such as "interesting, late news, politics" for that article, could contribute to making it easier for users to discover fresh content.

In contrast, search logs are only stored by the search engines and thus are also only available to them. Web site owners can also gain access to a small fraction of these data through a web site analytics service (e.g., google analytics, awstats). However, such services only give access to those parts of the search logs that are particularly related to a single website (referal URLs, search queries, etc). In addition, no other web user, except for the site owner, has access to these data.

We believe that web users can also benefit from knowledge about who accessed what pages and how they got to those pages. In particular, say a user submits a query "a b" to a search engine and then clicks on page p. If we know this fact, we can associate tags a, b (or the single tag "a b") to page p, creating a succinct, user-generated description of p. For example, such a description or summary for Microsoft's main page could be: microsoft, windows, software, bill gates. Thus, search queries can provide a new type of tags for web pages: *query tags*. Query tags tend to be accurate because web users have been trained all these years to formulate queries with the least ambiguous and most meaningful words. Thus, in addition to requesting information, searchers are also providing useful information that can annotate the pages they visit.

Query tags can also be used to annotate a set of web

Table 1: Sample web access logs. The IP address and User-agent fields have been removed.

[28/Feb/2007:23:57:57] /group/ICS/html/alumni.html "http://www.google.com.sg/search?hl=en&q=**Anindya**+**Bakrie**+&meta"
[28/Feb/2007:23:58:38] /group/stanfordbirds/text/species/Golden_Eagle.html "http://www .google.fi/search?q=**golden**%20**eagle**%20**diving**"
[28/Feb/2007:23:58:36] /class/cs193c/handouts/h06-ajax.pdf "http://www.google.com/search?q=**requesting**+**file**+**transfer**+**with**+**AJAX**"
[28/Feb/2007:23:58:52] / "http://www.google.com/search?hl=en&q=**stanford**+**university**+**bookstore**"

pages (e.g., pages in the domain infolab.stanford.edu, or pages linked by the infolab.stanford.edu pages) rather than individual web pages. For example, the term "pagerank" can be considered as a tag for the main Infolab web page because all the following three are true: (i) there exist queries that contain this term (e.g., "pagerank idea"), (ii) pages linked by the main Infolab page are contained in the search results for those queries (e.g., The Anatomy of a Search Engine [1]) and (iii) people that issued those queries actually clicked on those pages.

In this paper, we first illustrate how it is feasible to collect query tags using (a) the http_referer field from web server access logs and (b) embedded javascript code on web pages. Second, we look at a query tags dataset we collected from the stanford.edu domain and we ask whether query tags provide extra information for the pages they tag. We compare query tags for web pages with the actual pagetext and with tags from the del.icio.us bookmarking system. We find that query tags can provide substantially many, new tags for a large fraction of the Web.

2. QUERY TAGS COLLECTION

The underlying observation that allows query tags to be collected, is that web servers store the search engine queries in the http referer field of requests that originate from a search engine. Table 1 contains a small sample of the web access log from the main stanford web server. Requests that originate from a search engine contain the search query (in bold on Table 1) along with the requested page. The requested web pages correspond to the second field of Table 1 and are relative to the stanford domain (e.g., the string "/" on the second field of the last line corresponds to the URL www.stanford.edu/). For example, these logs indicate that someone searched for "golden eagle diving" and clicked on a page from the stanfordbirds group. Also, someone searched for "stanford university bookstore" and clicked on the official stanford webpage (indicated as "/" in the access logs). We have built a system that extracts query tags (1-grams) from a web server's access log and we used it to collect such data from the stanford.edu domain. We are also currently working on building a dataset with all possible n-grams as query tags.

Query tags can also be inferred without using web access logs. The idea is to embed Javascript code in each page of interest. When the code is activated, it detects whether the referer field of the http request comes from a search engine. If this is the case it extracts the query used. We have implemented the necessary Javascript code and are currently collecting data from pages in the CS department at Stanford. We do not use this data for this paper.

3. DATASETS

[1] `http://infolab.stanford.edu/~backrub/google.html`

Using the method described above, we collected a dataset from the stanford.edu domain. This dataset (Dataset Q) consists of all queried (from the three major search engines) web pages that appeared in the access logs of the main stanford web server during a period of 12 months beginning March 2007. Each entry in the dataset is a pair $<x, y>$ where x is a URL and y is a query tag (1-gram extracted from a query). The dataset consists of 359,749 unique URLs, 10,997,818 unique queries out of which we extracted 937,075 unique query tags (1-grams). Although all possible n-grams extracted from a query can be candidate query tags, we limit our experiments in this work only on 1-grams.

We also used a subset of the del.icio.us dataset used in [2], by keeping only URLs from the www.stanford.edu domain. Each entry in this subset is a a pair $<x, y>$ where x is a URL and y is a tag for that URL by some delicious user. The collection period of this dataset was March-April of 2007 and as a result it has temporal overlap with Dataset Q. The subset we extracted (Dataset D), consists of 2,965 unique URLs and 5,670 unique tags.

Both the distribution of delicious tags and query tags per URL seem to follow a power law (Figure 1(a)). As we can see, the top 10 most query tagged URLs have more than 10,000 query tags each, while the top 10 most delicious tagged URLS have around 100 del.icio.us tags each. Also, as Figure 1(a) shows, the frequency of query tags follows a power distribution with a much shorter tail. The tail becomes shorter when we look at the distribution of tags per URL for the common URLs of datasets Q and D Figure 1(b).

Finally, for our experiments, we also collected a crawl (Dataset C) of all pages (duplicates eliminated) whose URLs appear in in either Dataset Q or Dataset D. The crawl was performed on early September 2008 and thus all pagetext for URLs in Dataset C corresponds to the web pages' version available online at that point. We also found that 12,611 URLs were no longer available online. We did not consider those URLs in our experiments that involved comparing query tags with the pagetext.

4. EXPERIMENTS

In the following experiments, we address the following questions: Are queries useful for generating tags for web pages? How often are these tags "non-obvious"? How these tags compare to tags from social bookmarking sites? How do they overlap with the text of the web page they label? What coverage do these tags have of the web?

Due to lack of space we only give a short high level summary of the results of each experiment and a simple conclusion based on the outcome.

4.1 URLs

Result 1: Query logs provide tags for approximately 350,000 URLs in the stanford.edu domain, whereas del.icio.us

Table 2: Sample query tags. Tags that do not appear in the URL's pagetext are in bold.

URL (QT/common)	Sample query tags
www.stanford.edu (7650/47)	stanford, university, **standford**, ca, univ, **usa**, school, california, **bookstore**, **palo**, **alto**, graduate, **universidad**, campus, **address**, **fireworks**, medical, **church**, **leland**, **memorial**, **museum**, arts, uc, **location**, 94305, **president**, **store**, admission, **admissions**, **pictures**, **summer**, law, **tower**, jobs, student, **history**, **websites**, **email**, **bookshop**, **masters**, **photography**, **undergrads**, football, engineering, **service**, **santa**, **clara**, **monterey**, press, **american**, **athletic**, **classes**, **music**, phycology, **management**, **teaching**, **menlo**, **park**, **cardinal**, visitors, **statistics**, **relations**, jobs, **economics**, **game**, **computing**, center, **escondido**, **mysql**, enviroment, **physics**, **trustees**, **sandhill**, **provost**, maps, **space**, **redwood**, **volunteer**, **infolab**, **professional**, **distance**, **rentals**, **marine**, **los**, **altos**, children, **public**, **paloalto**, **dining**, clinics, **institution**, **director**, **apartment**, **computer**, **review**, parents, **fellowship**, **professors**, **theory**, **training**, **stock**, **books**, **union**
infolab.stanford.edu (99/5)	stanford, infolab, **database**, **db**, **university**, **research**, **pagerank**, lab, info, **databases**, **google**, **standford**, **db.stanford.edu**, hector, garcia-molina, **univ**, garcia, molina, **i.stanford.edu**, **p2p**, **change**, **dbgroup**, **computer**, **science**, department, **biosource**, page, **digital**, **media**, **personalized**, **larry**, **facilities**, **backrub**, **jennifer**, **widom**, **cs**, **stream**, **hotwire**, blog, **search**, **universities**

Table 3: Sample delicious. Tags that do not appear in the URL's pagetext are in bold.

URL (Del.icio.us/common)	Sample delicious tags
www.stanford.edu (80/47)	stanford, design, humanities, science, **imported**, research, **American**, school, education, law, web, business, **compsci**, **reference**, california, **inspiration**, **webdesign**, University, work, **USA**, **college**, **thesis**, **study**, us, engineering, academic, **eLearning**, **future**, Graduate, e-**Learning**, **mba**, edu, **universities**, CA, **homepage**, **sample**, **BayArea**, medicine, undregraduate, top, alumni, Sciences, Earth, **Old**, **abroad**, **universidad**, **March**, **adventures**, **StanfordSitestoExplore**, **My_Colleges**, **Uni.US.WestCoast**, **collegeWebsite**, **Organismes/Etats-Unis**
infolab.stanford.edu (21/5)	stanford, research, **search**, **information**, **Misc**, **california**, **researcher**, **library**, **technology**, **University**, **papers**, **Phd**, **management**, **learningToCode**, **researchgroup**, **oracle_sites**, **DatabaseGroup**, **dbgroup**

covers only 2,965 stanford URLs. There are 357,164 URLs that only query logs provide tags for.

Conclusion: Query logs can provide tags for up to 110 times more URLs than a social bookmarking site.

Result 2: Query logs (Dataset Q) provide tags for 2,582 (87%) of the 2,965 stanford URLs in del.icio.us (Dataset D). There are 383 (13%) URLs that only del.icio.us provides tags for.

Conclusion: Social bookmarking sites discover and provide tags for URLs never searched for by anyone. This can be seen both as a strength and as a weakness of a social bookmarking site. Also, assuming that social bookmarking sites discover URLs with fresh content, we can conclude that query logs can also eventually discover those URLs.

4.2 Tags

Result 3: Query logs (Dataset Q) provide on average 42 tags for each web page, while del.icio.us (Dataset D) gives on average 3 tags per URL.

Conclusion: Query logs can provide on average 14 times more tags per URL than a social bookmarking site does.

Result 4: Looking only at URLs from the stanford.edu domain posted on del.icio.us (URLs in Dataset D), query logs (Dataset Q) provide on average 250 tags per page, while del.icio.us (Dataset D) gives on average 4.6 tags per URL on the same set of URLs.

Conclusion: Query logs can provide on average 55 times more tags per URL than a social bookmarking site does for web pages from a popular (high pagerank) domain. Combining this with Result 2 we see that for each URL, query logs provide on average 14 tags not present in the pagetext. Delicious provides on average 1.5 tag not present in the pagetext for each URL.

Result 5: Looking only at URLs from the stanford.edu domain posted on del.icio.us (URLs in Dataset D), query logs (Dataset Q) provide more tags than del.icio.us for almost all URLs (Figure 1(c)). However, there are URLs in the tail of Dataset Q that get more tags from del.icio.us (the red circles above the the blue line in Figure 1(c)). These are URLs that get the average number of tags from del.icio.us (URLs that correspond to the blue crosses that are below the red line on Figure 1(d)).

Conclusion: There exist urls that get more del.icio.us tags than query tags.

Result 6: For each web page, 1 out of 3 query tags is not present in the pagetext, while 1 out of 2 del.icio.us tags is not present in the pagetext. Looking only at URLs from the stanford.edu domain posted on del.icio.us, 1 out of 2 (49.45%) tags from query logs is not present in the pagetext.

Conclusion: In contrast to common thought that all terms in a search query appear in the pagetext, tags from query logs contain valuable new information. Tables 2, 3 give examples of such terms. As we can see, query tags provide many interesting tags that do not appear in the pagetext. There are two possible explanations for this: (a) given a query, some search engines display results that do not contain all search keywords (e.g., Hector's website appears as the result for the query "Hector Garcia Molina infolab mexican" although the term "Mexican" does not appear

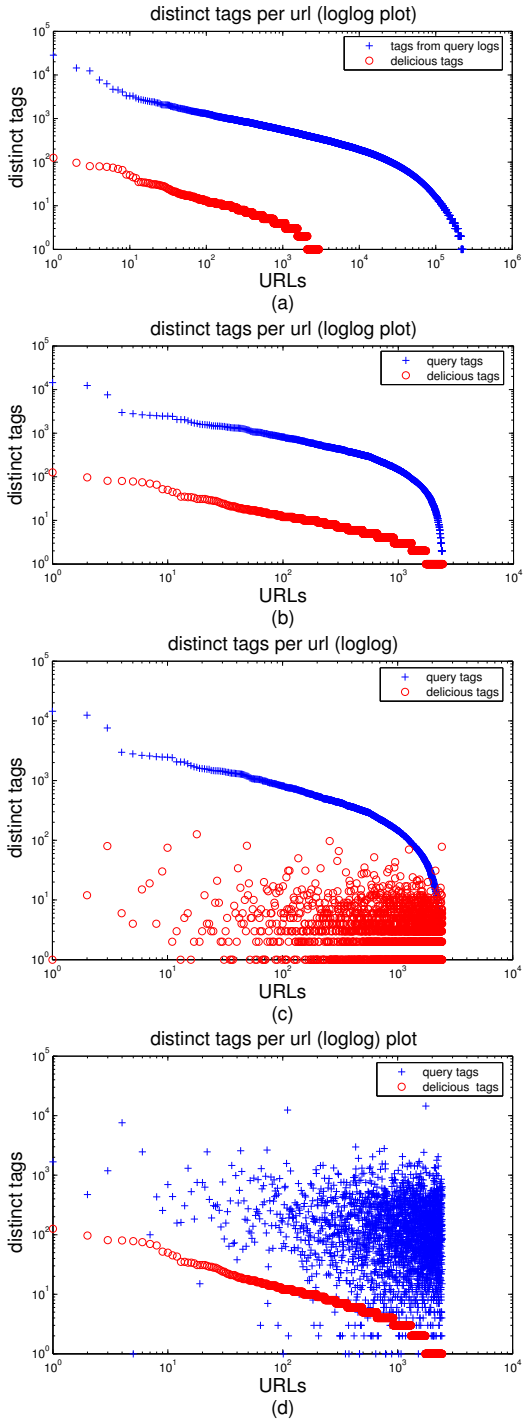

Figure 1: (a) Loglog distribution of tags per URL for Datasets Q and D. (b) Loglog distribution of tags per URL for the common URLs of Datasets Q and D. (c) Loglog distribution of tags per URL for Datasets Q and D; URLs in x-axis are ordered according to the distribution of query tags. (d) Loglog distribution of tags per URL for Datasets Q and D; URLs in x-axis are ordered according to the distribution of delicious tags.

in his webpage), (b) Many websites frequently change their content and as a result keywords in older queries no longer exist in the pagetext. Combining this with Result 4, we conclude that for each URL in the stanford.edu domain, query logs provide on average 125 tags not present in the pagetext. Del.icio.us provides on average 2 such tags per URL.

Result 7: Looking only at URLs from the stanford.edu domain posted on del.icio.us (URLs in Dataset D), 1 out of 5 (21.34%) of the common query and del.icio.us tags is not present in the pagetext.

Conclusion: Tags that are coming both from query logs and social bookmarking sites are the most "obvious" ones.

Result 8: Looking only at URLs from the stanford.edu domain posted on del.icio.us (common URLs in Datasest Q and D), 1.8 out of 2 (83.88%) of the tags appearing only in del.icio.us (and not in the query logs) is not present in the pagetext.

Conclusion: There exist tags from social bookmarking sites that do not appear in the pagetext. However, almost all these tags do not appear in an English dictionary; they are artificial concatenations of English words (e.g., ComputerScience, Stanford_University, to_read).

5. CONCLUSIONS

We looked at how search queries can be collected in a distributed fashion using the referer field of the http protocol. We illustrated that by collecting query tags for web pages we can get many tags (on average 250 tags per URL) for a large fraction of the web. In addition, we saw that in contrast to common thought that all terms in a search query appear in the page text of found pages, query tags often do not occur in the text; on average 125 query tags per URL do not appear in the pagetext.

Our results suggest that query tags can be a promising new source of information. Although previous work has looked at how query logs can be utilized by a search engine (e.g., [3], [1]), our work illustrates that query logs could be useful for web users as well. The main two questions that further arise are: How can query tags be used to improve navigation on the web, and how do we give incentives to site owners to share their query tags? For example, we are currently experimenting with a browser plugin that enables users to navigate through the query tags for the pages they visit.

6. REFERENCES

[1] Carlos Castillo, Claudio Corsi, Debora Donato, Paolo Ferragina, and Aristides Gionis. Query-log mining for detecting spam. In *AIRWeb 2008*.

[2] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *WSDM 2008*.

[3] Barbara Poblete and Ricardo Baeza-Yates. Query-sets: using implicit feedback and query patterns to organize web documents. In *WWW 2008*.