Mining Web Activity Logs

Thesis Proposal: Ioannis Antonellis * Thesis Advisor: Hector Garcia-Molina †

February 7, 2009

Web mining has been a major research topic during the last decade. Traditional sources of information on the Web include the text of web pages, the structure of the pages as organized by the web page creators, as well as the link structure of the Web. Another emerging source of information is the tags for web pages as collected and organized by collaborative tagging systems (e.g. del.icio.us, StumbleUpon). Search engines are interested in organizing all the information provided by these sources so that they are able to respond accurately to different search tasks.

However, the ample use of web search engines by billions of people worldwide has created another interesting source of information. The search and browsing activity of these users implicitly captures human knowledge and reveals the most interesting places on the web. By looking at the search queries people issue, we can extract a succinct, user-generated summary for pages appearing in the search results. For example, such a summary for Google's main page could be: google, search engine, larry page, pagerank. These summaries tend to be accurate given the fact that web users have been trained all these years to formulate queries with an emphasis on selecting the least ambiguous and most meaningful words. In a way, in addition to requesting new information, search users are also revealing new knowledge (that perhaps does not even exist yet on the web) through the keywords they select for their queries and the navigational paths they follow in response. The paths users follow are also implicit votes on the interesting parts on the Web.

Search engines are the first to record the users' search history in order to be able to analyze it and improve their services in the future. These query logs contain the terms users issue in a search engine along with the search results people click on in response. In addition, through browser toolbars, search engines can record web activity logs that contain the history of pages a user visits. Although web activity and query logs contain invaluable information for the search engines, it is also the web users who can explicitly benefit from such data. However, we have to identify meaningful ways to mine the web activity logs and make them available back to the web users. The primary goal of my research has been to investigate novel uses of web activity logs in formulating and solving novel problems related to the Web. The research on query and web activity logs that we have performed, with collaborators from Yahoo! and Microsoft Research, apply to problems related to sponsored search, recommendation systems, web information extraction systems, and web navigation systems. Section 1 outlines the research I have performed to date, while Section 2 describes my current research and plans for the future.

1 Results so far

1.1 Query rewriting for sponsored search In sponsored search, paid advertisements (ads) relevant to a user's query are shown above or along-side traditional web search results. The placement of these ads is in general related to a ranking score which is a function of the semantic relevance to the query and the advertiser's bid. For many queries, there are not enough direct bids, so the sponsored search system has to find other ads that may be of interest to the user who submitted the query.

In our work [1, 2], we tackle this problem by periodically generating query rewrites based on the recent history of ads displayed and clicked on. For the query rewrites generation, we exploit a historical click graph that records the clicks that were generated by ads when a user inputs a given query. The click graph is a weighted bipartite graph, with queries on one side and ads on the other. The schemes we present analyze the connections in the click

*Computer Science Dept, Stanford University. Email: antonell@cs.stanford.edu
†Computer Science Dept, Stanford University. Email: hector@cs.stanford.edu

graph to identify rewrites that may be useful. Our techniques identify not only queries that are directly connected by an ad (e.g., users that submit either "mp3" or "i-tunes" click on an ad for "iPod.") but also queries that are more indirectly related (e.g. "camera" and "batteries"). Our techniques are based on the notion of SimRank [6], which can compute query similarity based on the connections in a bipartite click-graph. However, in our case we need to extend SimRank to take into account the specifics of our sponsored search application. In [1, 2] we presented Simrank++ as a query rewriting technique and we compared it with existing query rewriting techniques. Simrank++ generates higher quality rewrites for more queries than existing techniques. The additional queries can have a significant potential impact on revenue for a search engine.

Since the query rewriting problem falls into the category of collaborative filtering problems, Simrank++ is essentially a new collaborative filtering technique and can be also used in traditional recommendation systems. In section 2.1 I discuss our current work on Simrank++ and future plans in this direction.

1.2 Scalable entity-document relationship extraction using click logs In my recent work ([4]) with Microsoft Research, we looked at ways to identify relationships between a given set of entities that belong to the same class (e.g. products, CS researchers, US cities), attributes of this class (e.g. price, university name, conference name) and the web documents that mention this relationship. For example, our techniques can extract the fact that Hector Garcia-Molina (entity) is related to the VLDB conference (attribute) as mentioned in the publications page of the Stanford Infolab (web page that explains the type of the relationship). Our extraction mechanism relies on the observation that when a user poses a search query and clicks on some results, he essentially tags the web documents he clicked on with the query terms. As a result, search queries can be treated as tags for web documents. By analyzing a tags-document graph that is generated from click logs we are able to extract and name relationships between entities and web documents.

1.3 Queries as tags for web pages Search queries capture the information need of a search engine user. Given an information need, a user tries to express it by forming a query that is composed of a small number of keywords. He can then inspect the search engine results, navigate on some and finally refine or reformulate his original query to start over again.

In our recent work ([3]), we are interested in studying 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. For example, consider a user that submits the query "Barack" and clicks on a blog that discusses the presidential candidates. Then, the word "Barack" can be considered as a tag for that blog. One major question is whether and how such data can be used to augment web search. Sample forms this question can take are:

- Is it the case that all the query keywords (e.g., "Barack") actually appear on the text of all clicked pages (e.g., the blog post)?
- How do query keywords overlap with the anchor text of backlinks to the clicked pages?
- How do query keywords compare to tags for web pages that come from collaborative tagging systems (e.g., del.icio.us, StumbleUpon)?
- Can we attach negative tags (labels) to intermediate web pages that are clicked while the user navigates on the search results (e.g., the blog post is *not* about "McCain")?
- How can we generalize query-based tagging patterns for unseen web pages that belong to specific web page categories (e.g., homepages of DB researchers or product pages)?

To answer these and other related questions we used a click logs dataset from the stanford.edu domain collected as described in Section 2.2 and the Stanford Tag Crawl 2007 dataset [5] based on the del.icio.us bookmarking site. 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, 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 (see Section 2.3), and how do we give incentives to site owners to share their query tags (see Section 2.2)? For example, we are currently experimenting with a browser plugin that enables users to navigate through the query tags for the pages they visit.

2 Research Plan

My research is focused on exploiting query logs and access histories to improve and unify the web search and browsing experience. The following are projects I am currently working on or plan to work in the near future.

2.1 Simrank++ and recommendation systems Simrank++ is essentially a new collaborative filtering technique and can be also used in traditional recommendation systems. The analog to the click graph in a recommendation system is a bipartite graph with user nodes on one side, item nodes on the other and an edge between a user and an item node if the user likes the item. In such a setting, we are looking at the following questions:

- How does Simrank++ perform in traditional collaborative filtering tasks (e.g., netflix challenge)?
- How can we incrementally compute Simrank++ scores when the input data are changing without recomputing all scores from scratch?
- How can we parallelize and speedup the Simrank++ computation so that it can scale to large datasets? We are currently working on a Map/Reduce implementation and a graph summarization technique for fast main memory computation.

2.2 Distributed collection of activity logs A major problem when trying to do research that involves search engine logs, is that these are not publicly available. However, we believe that web users can benefit from a direct access to these logs. For example, while today every web site owner is able to monitor the keywords people issue to get to his website (through a web site analytics service), he does not have access to the same information for any other website. We believe though that this information is useful to everybody and we are working on techniques that will allow web site operators to collect data on their own.

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. Thus, this same data that a search engine has is also distributed among all the web servers. We have already built a system that extracts activity logs from a web server's access log and we have used it to collect such data from the stanford.edu domain.

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 (available through the Stanford Tags Project web site: `http://tags.stanford.edu`) and are currently collecting data from pages in the CS department at Stanford.

Important research issues we are we are trying to tackle are:

- Since there is sensitive information in this kind of logs how can we deal with the resulting privacy issues?
- How do we provide incentives to web server administrators and web site owners to agree on sharing such data? For instance, we working on an enhanced web site analytics service, that provides to the site owner information about the "web neighborhood" of his page.

2.3 Web navigation using click logs By expanding on the idea of treating queries as tags for web pages, we are investigating different ways to provide navigation aid to web users. The driving force of this work is the observation that by aggregating the tags for each web page, we get a succinct, user-generated summary of the web page's content. For example, from data we have already collected, such a summary for the Stanford Infolab is: stanford, infolab, database, research, pagerank, garcia-molina, widom, ullman.

Thus, we are currently building and experimenting with a browser plugin that enhances the web navigation by displaying to the user tags related to the web page he is currently visiting. In addition, as the user navigates

the web, he can explicitly contribute more tags and/or vote in favor or against existing tags for the pages he is visiting.

Currently, web users can navigate the web mostly through hyperlinks and through their interaction with a search engine. Hyperlinks are created by the web page owner and are in general static. However, we believe that tag-based automatically created hyperlinks can substantially improve web navigation.

References

[1] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. Simrank++: Query rewriting through link analysis of the click graph. In *VLDB '08: Proceedings of the 34th International Conference on Very Large Data Bases, Auckland, New Zealand, 2008.*

[2] Ioannis Antonellis, Hector Garcia-Molina, and Chi-Chao Chang. Simrank++: Query rewriting through link analysis of the click graph (poster). In *WWW '08: Proceedings of the 17th International Conference on the World Wide Web, Beijing, China, 2008.*

[3] Ioannis Antonellis, Hector Garcia-Molina, and Jawed Karim. Tagging with queries: How and why? In *WSDM '09: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, Late Breaking Results Session, 2009.*

[4] Ioannis Antonellis, Arnd Christian Konig, and Venky Ganti. Scalable entity-document relationship extraction using click logs. In *MSR Technical Report, TR 09/2008, Data Management Exploration and Mining (DMX) Group, Microsoft Research, Redmond, WA, 2008.*

[5] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social-bookmarking improve web search? In *Proc. WSDM '08.*

[6] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *KDD '02: Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002.*