Sharing Queries, Grouping Users, and Its Relation To a Decentralized Web Index

Emmanuel Benazera
The Seeks Project
juban@free.fr

Sylvio Drouin
The Seeks Project
sylvio@senseido.net

ABSTRACT

Community views and concerns about the current state of web search technologies in relation to the accuracy, privacy and origin of information are leading to the introduction of novel web search models and algorithms. For example, there is much hope in correcting the corporation's unfair advantage over individual users when deploying content on the world wide web. This paper exposes the long term architectural shape defended by the Seeks project, a fledgeling Open-Source effort toward increasing the capabilities of the network leaves in order to foster distributed collaborative indexing and searching. The presented architecture achieves a locality sensitive mapping from the space of queries to the space of computers in the network. Users that perform similar Web search queries are automatically regrouped, meet and collaborate. Here user queries to search engines are shared with the global search community rather than being recorded by corporate entities.

1. INTRODUCTION

The Internet has relied on an end-to-end architecture, where the power lies in the leaves of the network, such as Web servers or user machines. However, the current topological state of the websphere has suffered the rise of two additional topologies: first, the gateway-like topology where servers gather the network traffic and redistribute it to the leaves, namely the search engines; second, the bag-like topology where the traffic gets trapped within a single set of servers, namely the so-called social web-communities. We would not be concerned with these topological trends if they did not appear as both risky and inefficient in the long term. The most prevalent problem lying in the massive personal and public information collections now being held by the businesses initially responsible for the above mentioned topological changes.

Consider how a Web search query always reveals something about ourselves, our interests, our opinions, etc. In general, such exchanges rely on a chain of trust. In the case of search engines, the interlocutor is not another human being but a set of sophisticated and confidential algorithms that record, re-use and most probably distort all the information accumulated about us.

The Seeks project is both a theoretical and practical initiative motivated by the belief that the Web has reached a point where the chain of trust has broken down. We argue that the process by which centralized entities, with high-traffic capacity, obtain private profile and behavioral data, has become transparent to the point where users are lured into using free, and most of the time essential services; gradually come to rely on them; in the end to be coerced (through clever web authoring) into revealing extensive personal information without ever realizing they are doing so. This is the reason why users need to invent ways to protect themselves, share information, and evaluate this information based on the collective trust of all users rather than based on the results of few corporations greed-influenced algorithms.

It is theoretically understood that peer-to-peer (P2P) Web search lies an order of magnitude of feasibility behind its centralized counterpart [6]. The communication over-cost is a clear consequence of the exact match lookups over the query-to-computer mapping achieved by a distributed hash table (DHT). In fact, the local computation of a globally usable similarity index over ranges allows for dodging the pitfall of distributed set intersection and data clustering. This index encrypts ranges of user Web search queries that are stored on a DHT. A DHT lookup returns a bucket of peer net addresses that are performing similar searches, known as a *search group*. Collaboration and filtering naturally follow from this query-based clustering of users.

Here, we sketch a roadmap of our current and future work. First, we explain how a collaborative Web search that brings together users with similar queries can be set at almost no cost standing on the shoulders of existing engines. Second, we sketch a fame-based marketplace for self-publishing and fair discovery. Third, we give elements for a distributed Web indexing mechanism based on the locality sensitive DHT (LS-DHT) and expose both its advantages and limitations.

2. COLLABORATIVE WEB SEARCHING

2.1 Mapping queries, grouping users

A DHT hash function builds a mapping from the space of queries to the space of computers in the network. Consider the two options for mapping the space of queries onto a real valued set: (i) queries are single keywords; (ii) queries are chains of characters, or an (order or unordered) group of keywords, with an imposed maximum size. Option (i) presents the advantage that looking for keywords is expected to maximize the pool of returned users thus building up relatively large search groups, which in general is a factor of increase of a collaborative filter prediction's accuracy. It also implies that the load over the network is very unbalanced: nodes that correspond to popular keywords are easily over-

.

loaded. Third, complex queries (i.e. made of more than a single keyword...) have to be assembled in the network, which has been proved to be inefficient [6]. In option (ii), queries are chains of characters and are stored as such in the DHT. Now, this presents the obvious disadvantage that search groups are made smaller. However, the load over the network nodes is more well balanced. No more reduction operations are needed to be performed on the network neither.

2.2 Enhancing the overlay net with locality

Traditional DHT lookups query the network for exact matches. However, we can expect very similar Web search queries to return highly correlated results, and therefore be of interest to the same users. Thus a search group is a set of users clustered by Web search query similarities. A locality sensitive hash function [3, 2] controls the collision rate to produce real valued keys that are the same for groups of queries. A preliminary analysis of an LS-DHT in a different context can be found in [4].

2.3 Distributed collaborative filtering

Our application builds on the cumulative experience of users searching through related parts of the websphere, and naturally relies on a user-based rating of visited webpages. Importantly, the above LSH-based clustering of users naturally offers the reference user pool for solving the distributed k nearest neighbors problem that underlies the collaborative filtering schemes [5].

2.4 Building on top of existing engines

Queries to existing search engines are channeled through the LS-DHT. Their results are re-sorted and enhanced with the information fetched from the DHT, such as ratings and similar queries, thus a low-cost extension.

3. SELF-PUBLISHING AND FAME

3.1 Self-publishing

This second step introduces what we believe are Seeks most beneficial features. It proposes a self-publishing mechanism accessible to anybody with a browser and an Internet connection. Instead of relying on a search engine for linking keywords to web contents, Seeks will let the users register any URL using their own set of keywords. Users querying the DHT will thus be recommended web content without using any existing search engine. This operation is the combination of a DHT lookup plus a local selection and does not cost much more than a P2P lookup.

3.2 Virtual fame-based marketplaces

Users registering their personal web content or that of others under unsatisfactory keywords or queries would see their keyword associations naturally rated down by other users, in a move that we believe should lead to a better match of keywords and queries to the true content of a web page. Finally, and nonetheless, we are in the process of defining the setup of virtual marketplaces over keywords for publishing web content, at each of the DHT peers. These marketplaces would not rely on money but on fame instead, understood as a measure of a user attachment to truth. Thus any hot content recommended to Seeks users would come from a fair collective pre-selection among bidding users. Interesting problems arise such as the automated fostering of new publishers and increased diversification of content through the market parameters.

4. AN LSH-BASED DECENTRALIZED WEB INDEXING SCHEME

We describe a decentralized web information index to gradually re-capture public information currently stored in private corporate facilities. We propose to implement small software extensions to common Web servers, such as Apache and IIS. These extensions would allow Web servers to locally index their webpages and share the locality sensitive indexes with other web servers and users, in a decentralized manner, on the LS-DHT. The consequence is that over time, such an architecture would evolve a parallel search engine, processing queries against a decentralized database of information rated by the community for the community.

5. DISCUSSION

First, while this reports on an on-going effort, we understand how a distributed web search is both theoretically feasible and practical through an LS-DHT. However, as a possible architecture, the decentralized index is only feasible through the collaboration of all servers and the spread of local indexing algorithms.

We reckon the practical difficulties in properly setting the local root similarity measures. However, we have already implemented scaling schemes with successful results. The main drawback of [1] that is to rely on several DHTs is circumvented by tuning the index locally.

Also, we understand the difficulties and opposition users may (and certainly will) have to the sharing of queries, but we also believe that the benefits far outweighs the risks. Most notably by allowing a take-back of the Web publishing opportunities.

6. REFERENCES

[1] M. Bawa, T. Condie, and P. Ganesan. Self-tuning indexes for similarity search. In *WWW 2005*, 2005.

[2] M. Datar, P. Indyk, N. Immorlica, and V. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *In Proceedings of the Symposium on Computational Geometry*, 2004.

[3] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In *In Proceedings of the 25th VLDB Conference, Edinburgh, Scotland*, 1999.

[4] A. Gupta, D. Agrawal, and A. El Abbadi. Approximate range selection queries in peer-to-peer systems. In *Proceedings of the First Biennial Conference on Innovative Data Systems Research*, January 2003.

[5] P. Han, F. Yang, and R. Shen. A novel distributed collaborative filtering algorithm and its implementation on p2p overlay network. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia*, 2004.

[6] J. Li, B. Loo, J. Hellerstein, F. Kaashoek, D. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *In 2nd International Workshop on Peer-to-Peer Systems.*, 2003.