DEWS: A Decentralized Engine for Web Search

Reaz Ahmed*, Md. Faizul Bari*, Rakibul Haque*, Raouf Boutaba*, and Bertrand Mathieu†
*David R. Cheriton School of Computer Science, University of Waterloo
{r5ahmed | mfbari | m9haque | rboutaba}@uwaterloo.ca
†Orange Labs, Lannion, France
bertrand2.mathieu@orange-ftgroup.com

***Abstract*—Contemporary Web search is governed by centrally controlled search engines, which is not healthy for our online freedom and privacy. A better solution is to enable the Web to index itself in a decentralized manner. In this work we propose a decentralized Web search mechanism, named DEWS, which enables existing webservers to collaborate with each other to build a distributed index of the Web. DEWS can rank search results based on query keyword relevance and relative importance of webpages. DEWS also supports approximate matching of query keywords in web documents. Simulation results show that the ranking accuracy of DEWS is very close to the centralized case, while network overhead for collaborative search and indexing is logarithmic on network size.**

I. Introduction

Internet is the largest knowledge base that mankind has ever created. Autonomous hosting infrastructure and voluntary contributions from millions of Internet users have given the Internet its edge. However, contemporary Web search services are governed by centrally controlled search engines, which is not healthy for our online freedom due to the following reasons. A Web search service provider can be compromised to evict certain websites from the search results, which can reduce the websites' visibility. Relative ranking of websites in search results can be biased according to the service providers' preference. Moreover, a service provider can record its users' search history for targeted advertisements or spying. For example, the recent PRISM scandal surfaced the secret role of the major service providers in continuously tracking our web search and browsing history.

A decentralized Web search service can subside these problems by distributing the control over a large number of network nodes. No single authority will control the search result. It will be computed by combining partial results from multiple nodes. Thus a large number of nodes have to be compromised to bias a search result. Moreover, a user's queries will be resolved by different nodes. All of these nodes have to be compromised to accumulate the user's search history.

A number of research works ([1], [2], [3]) and implementations (YacY$^{www.yacy.net}$, Faroo$^{www.faroo.com}$) have focused on distributed Web search and ranking in peer-to-peer (P2P) networks. These approaches have two potential problems in common: (a) *lookup overhead*: number of network messages required for index/peer lookup is much higher in P2P networks compared to a centralized alternative, (b) *churn*: maintaining a consistent index in presence of high peer churn is not feasible. Thus, those solutions have issues with performance and accuracy requirements.

In this paper we take a very different approach to decentralized web indexing and ranking. Instead of relying on an overlay of regular Internet users, we build an overlay between webservers. We exploit the stability in webserver overlay to heavily cache links (network addresses) that we use as routing shortcuts. Thus we achieve faster lookup, lower messaging overhead, and higher ranking accuracy in search results.

The rest of this paper is organized as follows. First we present the DEWS architecture in §II. Then we validate the concepts presented in this work through extensive simulations and present the results in §III. We present and compare with the related works in §IV. Finally, we conclude with future research directions in §V.

II. System Architecture

We have used Plexus protocol [4] to build an overlay network between the webservers participating in DEWS. Since the webserver overlay is fairly stable, each webserver caches links (network addresses) to other servers. This link caching reduces network overhead during the indexing and routing processes. On top of the overlay topology we maintain two indexes for distributed ranking (§II-A1) and keyword search (§II-A2), respectively. Functionally DEWS is similar to a centralized search engine. It generates ranked results for keyword search (§II-B1). From now on we use the terms webserver and node interchangeably.

Plexus is a unique Distributed Hash Table (DHT) technique with built-in support for approximate matching, which is not easily achievable by other DHT techniques. Plexus routing scales logarithmically with network size. Plexus delivers a high level of fault-resilience by using systematic replication and redundant routing paths. Because of these advantages we have used Plexus protocol to build the webserver overlay. Here, we summarize the basic concepts in Plexus followed by the proposed extensions.

A. Indexing Architecture

Metrics used for ranking web search results can be broadly classified into two categories: a) hyperlink structure of the webpages, and b) keyword to document relevance. Techniques from Information Retrieval (IR) literature are used for measuring relevance ranks. While link structure analysis algorithms like PageRank [7] is used for computing weights or relative