Lessons from a Gnutella-Web Gateway

Brian D. Davison, Wei Zhang and Baoning Wu[*]
Department of Computer Science & Engineering, Lehigh University
19 Memorial Dr. West, Bethlehem, PA 18015 USA
{davison,wei,baw4}@lehigh.edu

ABSTRACT

We present a gateway between the WWW and the Gnutella peer-to-peer network that permits searchers on one side to be able to search and retrieve files on the other side of the gateway. This work improves the accessibility of files across different delivery platforms, making it possible to use a single search modality. We outline our design and implementation, present access statistics from a test deployment and discuss lessons learned.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*distributed systems*

General Terms

Measurement, Design

Keywords

Peer-to-peer, Gnutella, search engine, World Wide Web

1. INTRODUCTION

The search for information can take many forms. The availability of multiple kinds of search tools both helps and hurts this process — it can narrow the search to a smaller collection (when knowledge of the collection contents is available), but it may also require the searcher to perform multiple steps. Today, for example, it is common to provide access to files by putting them on a Web server, or by sharing them within a peer-to-peer network. Some files may only be available on one platform or the other, and so searchers are thus required to use multiple tools to do an exhaustive search.

Our work is designed to improve the accessibility of files across different delivery platforms, making it possible to use a single search modality. We propose a gateway between the WWW and the Gnutella peer-to-peer network [1] that permits searchers on one side to be able to search and access files on the other.

While operating a P2P-Web gateway may consume substantial resources, there are many motivations for doing so, including: improving the accessibility of information within an organization, as in an intranet, and, increasing the accessibility of specific information (such as product literature), such as via a paid search service.

*The authors were supported in part by NSF grant ANI 9903052.

WWW2004, May 17–22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. BACKGROUND

Peer-to-peer file sharing provides an important channel to allow users to share their own files and retrieve information from remote machines. Peer-to-peer systems are decentralized, and nodes will freely join and leave the network on a frequent basis. Thus, a central concern is how to locate available content. In Gnutella, queries are distributed using a naive message flooding approach. When a query arrives at a Gnutella node, it searches locally and passes the query to its neighbors (where the process is repeated until the query has traveled a maximum number of hops). If files match the query, a query hit message will be generated and passed back.

A gateway enables communication between networks that use different protocols. This is exactly the purpose of our system, to convert between the WWW and Gnutella, within the context of search and retrieval. Our gateway operates both as a Gnutella client and server and as a Web server and client. It captures Gnutella queries that it receives as a member of the network, and forwards them to a search engine. The gateway takes the results from the search engine responses and forms Gnutella messages to transfer them back to the searcher via the P2P network. Similarly, when the gateway receives a query through its Web interface, it distributes the query to all of its Gnutella neighbors. As results are collected from the network, they are presented to the Web searcher.

In addition to searching files, the gateway also helps searchers retrieve the files from the alternate network. Although Gnutella uses HTTP for file transfer, additional constraints on the URL specification make a Gnutella client unable to retrieve an arbitrary URL from the Web. In the other direction, a Web browser is able to generate a correct Gnutella retrieval request, but most Gnutella implementations will refuse such requests from browsers to discourage users that are not participating in the network. Thus, it is necessary to relay data from the Web to the Gnutella network (and vice versa) to accommodate acceptable request formats. By providing bi-directional search and retrieval services, the gateway provides significant benefits for both Gnutella users and Web surfers.

In past years, there have been Gnutella search services (e.g., [2]) that provided a Web interface for users to enter a query, send that query into the Gnutella network, and return results via the Web. Links on the result page would go directly to the Gnutella systems, allowing the searcher to retrieve files directly. This type of Web search no longer works — most Gnutella software will reject browser clients with a message asking users to contribute to the network by running Gnutella software.

3. IMPLEMENTATION

The operation of our gateway is illustrated in Figure 1. A Gnutella node broadcasts a query message (1), which is received, translated, and forwarded by the gateway to a Web search engine

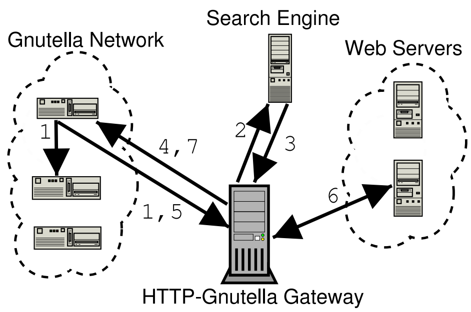

Figure 1: Operation of the Gnutella-Web gateway.

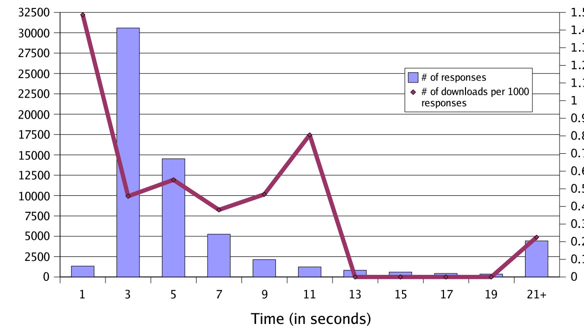

Figure 2: Query response time versus downloads.

(2), which generates a set of query results and returns them (3). The gateway translates the results into Gnutella formats and then forwards them to the sending node (4), such that results are available through the gateway. If the user sends a download request to the gateway (5), the gateway will fetch the data from the Web server (6) and return the data to the user (7).

In addition to serving Gnutella clients, our gateway also permits Web users to query and retrieve resources in the Gnutella network. In this direction, a Web user submits a query via a Web interface, which is transmitted to the Gnutella systems connected to the gateway. As results are received from the Gnutella nodes, the gateway compiles the results and presents a hit list to the Web searcher, specifying the gateway as the source of those files. If the Web searcher attempts to download a file, the gateway will extract the original Gnutella system's address and filename and contact that node to download the file, passing the contents back to the Web browser.

We modified an existing open-source Gnutella servent, Gtk-Gnutella[1], version 0.91.1, to implement our gateway system. Gtk-Gnutella supports both versions 0.4 and 0.6[2] of the Gnutella protocol. Instead of searching a local filesystem when a query is received, our system sends the query to a Web search engine.

A downloading request from a Gnutella client requires an index number as well a file name to display to the user. We compose a file name from the title of the Web page, generate a unique index value, and record the actual URL along with the index value into a local database. Thus, when a retrieval request arrives, we are able to retrieve the URL from the database via using the index provided in the Gnutella client's request.

Since our gateway employs intranet or Web search engines for back-end query processing, recklessly forwarding many P2P queries may affect their service and even lead to a perceived Denial-of-Service attack. Therefore, we enforce a minimum delay between query transmissions (placing delayed queries into a queue).

While Web browsers don't need the file size before downloading a file, the Gnutella client must know it when a query hit occurs. However, the gateway is unable to accurately know the file size when Web links are returned from search engine, and an incorrect file size will lead to downloading failure. To compensate, we used an arbitrary file size of 52572 bytes, padding the returned (HTML) file to match the advertised size.

[1] http://gtk-gnutella.sourceforge.net/

[2] http://rfc-gnutella.sourceforge.net/src/rfc-0_6-draft.html

No. of received queries	1,281,565
No. of served queries	166,462
No. of query responses	631,037
No. of successful downloads (Lehigh/all)	608 / 621

Table 1: Gateway usage statistics.

4. EXPERIMENTS

We operated a test gateway connected to our university LAN in October and November 2003. The gateway provides Gnutella nodes with HTML page search and download for files found via the Lehigh University search engine. Usage statistics are shown in Table 1. The number of served requests is much lower than that of received requests because we filter out some queries (e.g., those specifically for filetypes mp3, mpg, etc.) and ignore many queries so as to not overwhelm the Lehigh search engine. Most downloads are from internal Lehigh users, suggesting that such users may be interested in retrieving information hosted within the intranet, even though they are not intentionally using a tool to do so.

Since our prototype implementation queues P2P requests that cannot be immediately satisfied, the overall response time of a query can vary. We analyzed the relationship between the response time and the number of downloads performed for results with a particular response time (that is, the time between when we received the request and when we sent out the first response). Figure 2 plots the distribution of response times, and a normalized value of the popularity of results at each response time. It demonstrates that users clearly prefer to download files from responses that arrive quickly, particularly within 12 seconds.

5. SUMMARY

We presented a novel gateway system to extend the search and retrieval abilities of Gnutella participants to the Web domain, and vice versa. Our prototype has been tested to demonstrate the potential of our approach.

In the process, we have provided evidence that: 1) Gnutella users are interested in retrieving content available on the Web; 2) university students were able to automatically discover our server (even though we did not advertise it, and most Gnutella clients ignore network locality) and to download university-hosted Web content; and, 3) Gnutella queries have a lifetime of 12 seconds, after which results are ignored, and generally that faster responses are more likely to be useful.

6. REFERENCES

[1] G. Kan. Gnutella. In A. Oram, editor, *Peer to Peer: Harnessing the Benefits of Disruptive Technologies*, chapter 8, pages 94–122. O'Reilly, Sebastopol, CA, Mar. 2001.

[2] K. C. Sia, C. H. Ng, C. H. Chan, S. K. Chan, and L. Y. Ho. Bridging the P2P and WWW divide with DISCOVIR – DIStributed COntent-based Visual Information Retrieval. In *Poster Proc. of the* 12^{th} *WWW Conf.*, Budapest, May 2003.