P2P MetaData Search Layers

Sam Joseph
Strategic Software Division
Graduate School of Information Science and Technology
University of Tokyo 7-3-1, Hongo, Bunkyo-ku
Tokyo 113-8656 Japan
sam@mtl.t.u-tokyo.ac.jp, sam@neurogrid.com

Abstract. Distributed Hashtables (DHTs) provide a scalable method of associating file-hashes with a particular location in a distributed network environment. Modifying DHTs directly to support meta-data is difficult, and meta-data search systems such as flooding tend to scale poorly. However, a number of more scalable distributed meta-data search systems have recently been developed that could be deployed in tandem with DHTs, and several are discussed here along with some novel simulation results that concern the scalability and resource limitations of a meta-data search layer that employs semantic routing. Semantic routing is a method of pruning a flooding search such that queries are preferentially forwarded to nodes that can answer those queries. Previous simulations [1] showed that under certain circumstances semantic routing leads to a reduction in search path length. This paper presents further simulation results indicating that the scalability of this effect is a function of the query distribution of individual user search activity.

1 Introduction

Given a distributed network environment we can break down the process of obtaining a file or document into three distinct stages:

1. WHAT: Identify which file you want from some meta-data criteria
2. WHERE: Work out where it is (potentially multiple locations/pieces)
3. HOW: Download it (from one or multiple locations)

These stages are merged in some systems, and arguably one could add a stage 0 in which the user specified what kind of meta-data schema they would like to be able to search over [2]. The Freenet system [3] merges the WHERE and HOW stages and relies upon a separate search layer to implement the WHAT stage. Merging stages may be necessary in some cases, however to the extent that they are separable they can be implemented by entirely different systems. For example, one might use FASD [1] to identify a file from keyword meta-data; use Chord [5] to work out the location of the file itself, and then BitTorrent [6] to actually download it. Distributed Hashtables such as Chord and CAN [7], with their bounds on path-length and/or connectivity as network size increases, seem...
well suited to providing the functionality of the WHERE stage as opposed to the WHAT stage. Storing file meta-data directly in DHTs, such as by splitting up filenames into n-grams \[8\], generates routing hotspots where certain nodes end up with heavy routing traffic because they are responsible for a particularly popular meta-data field or n-gram chunk \[9\].

Among the variety of recent developments in meta-data search layer techniques targeted at the WHAT stage, is "Semantic Routing"; a method of pruning a flooding search such that queries are preferentially forwarded to nodes that can answer those queries. Existing work \[1\] has shown that under certain circumstances semantic routing leads to a reduction in search path length. This paper describes some new semantic routing simulations and considers results indicating that the scalability of this effect is a function of the query distribution of individual user search activity.

The rest of the paper is structured as follows: in section 2 we review some of the recent developments in scalable meta-data search layers that could be deployed in tandem with DHTs. Section 3 describes a semantic routing meta-data search layer simulation, and includes some analysis of the search query distributions that will be used to make some scalability predictions. In section 4 we look at the simulation results, comparing them with the predictions of section 3 and discuss the scalability implications. Finally in section 5 we discuss the issues arising from trying to develop scalable meta-data search layers, and simulation of p2p systems in general.

2 Related Work

As mentioned in the introduction Freenet \[3\] merges the WHERE and HOW stages and relies on other systems to implement the WHAT stage. Freenet forwards queries according to beliefs about the contents of other nodes and considers file similarity in terms of closeness in a "key-space" generated by a cryptographic hash. A file’s key is used to retrieve/insert files from/into particular locations. Combined with aggressive caching activity the arrangement of files ends up reflecting that of the key-space and the relative demand for different documents. Gnutella \[10\] relies on individual users sharing files stored locally, using broadcast search to identify file locations from partial search of plain text filenames, and thus merges the WHAT and WHERE stages. The HOW stage is farmed out to HTTP. Other systems have emerged that attempt to deal with the issues of distributed file storage, such as MojoNation \[11\] (now Mnet), which provides all three stages in one package.

Recently a number of systems have been developed that try to provide sophisticated meta-data search. One possible P2P meta-data approach is to try and use Chord to store keyword-document relations. Kronfol \[4\] suggests that under this scheme popular query terms would drive excessive traffic to certain nodes. As an alternative Kronfol describes and simulates FASD, which adds keyword searching to the Freenet system by inserting meta-data keys that include the TFIDF (Term Frequency Inverse Document Frequency \[12\]) rankings of key-