Evaluating Peer-to-Peer Networking for Information Retrieval within the Context of Meta-Searching

Iraklis A Klampanos, James J Barnes, and Joemon M Jose

Department of Computing Science – University *of* Glasgow
17 Lilybank Gardens, Glasgow G12 8QQ, Scotland
{iraklis, barnesjj, jj}@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/∼iraklis

Abstract. Peer-to-peer (P2P) computing has shown an unexpected growth and development during the recent years. P2P networking is being applied from B2B enterprise solutions to more simple, every-day file-sharing applications like Gnutella clients. In this paper we are investigating the use of the Gnutella P2P protocol for Information Retrieval by means of building and evaluating a general-purpose Web meta-search engine. A similar project, Infrasearch, existed in the past only to demonstrate the usefulness of P2P computing beyond just file-sharing. In this work, a Java-based Gnutella enabled meta-search engine has been developed and used in order to evaluate the suitability and usefulness of P2P networking in the aforementioned context. Our conclusions concerning time efficiency in different networking topologies are presented. Finally, limitations of this approach as well as future research areas and issues on the subject are discussed.

1 Introduction

During the recent years the Peer-to-Peer (P2P) networking paradigm has become very popular among simple users as well as the corporate sector. As its name suggests it defines equal participation among the nodes in a network. This comes in contrast to the well-known Client-Server networking model which is predominant both on the Internet and on Local Area Networks (LANs).

P2P networking is a subset of distributed computing and it can be seen as one more layer on top of the layers of the various networking protocols like TCP/IP. The fact that it is software based makes P2P solutions very flexible with numerous possible applications. Unfortunately, because of various historical turns, and especially after the Napster incident, most people consider P2P systems simply as file-sharing applications while companies treat them as a tools for enterprise solutions. However, despite Information Retrieval (IR) has been projected as a major application of P2P protocols, no proper investigation has been made into this issue.

There are countless potential information sources on the Internet nowadays; each of those with its own characteristics, policies, hardware and software architectures. We believe that P2P solutions could help use the Internet in a way never seen before, by building active and collaborating information pools that exchange information regardless of underlying network characteristics and retrieval policies. By distribution of knowledge and technical resources, the retrieval process could also be greatly aided.

In this paper, we will investigate the effects of different networking topologies of a P2P network in retrieval efficiency. This paper is organised as follows. In the next Section we will introduce P2P networking as well as some specifics of the Gnutella protocol. In Section 3 we will briefly describe the experimental metasearching tool we used and some of the policies we have followed while designing and implementing it. Following that, in Section 4, we will describe the evaluation strategy we used for the aforementioned purpose and finally, in Section 5 we will discuss conclusions and future possibilities of the P2P approach for IR.

2 Peer-to-peer Networking and the Gnutella Protocol

2.1 Defining Peer-to-Peer

> “[...] Instead, machines in the home and on the desktop are connecting to each other directly, forming groups and collaborating to become user-created search engines, virtual supercomputers, and filesystems.[...]” [1, page 3]

As the name “Peer-to-Peer” implies, a P2P network comprises of nodes that are communicating with each other in a predefined framework of equality; the equality having to do with the services each node is *capable* to provide. Because of the various flaws that this description has, it is vital that we precisely define a P2P network before proceeding any further. In order to do so, we first have to define the building blocks of such networks, the *peers* and those aforementioned services will drive our definition.

Definition 1. Peers *are the living processes running on the participating machines in the network that are* potentially capable of *providing and using remote services in an equal qualitative and quantitative manner. Peers, though, may not exhibit equal participation levels. The latter are proportional to the peers’* willingness *to participate which can be dictated by hardware or other fixed circumstances (eg. limited bandwidth etc.). Peers provide both server and client functionality hence why they are often called servents (**Serv**er - Cli**ent**).*

We can therefore define a P2P network as follows:

Definition 2. *A* Peer-to-Peer network *is a directed graph whose nodes are represented by* peers *and its edges are represented by abstract communication channels. In such a network, the equality of peers is defined by their potential capabilities while their participation level is analogous to their willingness to participate, as defined above.*

2.2 The Gnutella Protocol

The protocol used for developing our experimental meta-search engine is the Gnutella protocol[1, 7]. Gnutella was the first completely decentralised, genuine P2P protocol, that was born, matured and evolved in the public Internet. Gnutella is a message-based protocol, typically implemented as a distributed P2P application for file-sharing. Some popular Gnutella clients, for instance, are: Limewire[1], Bearshare[2], and GTK-Gnutella[3].

Infrasearch In the past, the Gnutella protocol, by being simple and elegant, gave rise to new ideas concerning Information Retrieval in conjunction with true decentralised, distributed environments. A P2P meta-search capacity, intended as a demonstration of the adaptability of the protocol, has been developed before.

Known as Infrasearch[1, page 100], the search engine that was pioneered, used the Gnutella protocol to operate over a private Gnutella network, the interface being a standard Web browser and response results being expressed as HTML rendered by the browser. In Infrasearch, each node interpreted the query and supplied results according to its knowledge resources only if it could provide any relevant information.

3 A Gnutella Meta-Search Engine

In order to evaluate the suitability of P2P techniques for IR we developed a distributed meta-search engine. Such an engine that uses the Gnutella protocol is described in this section.

3.1 Overview

The main purpose for building such a meta-searching tool was to try to identify the limits of a P2P approach in a well known IR area. The steps of meta-searching are found to be computationally expensive[5], always depending on the algorithms at hand. It would, therefore, be important to make those steps more manageable by distributing the CPU and memory requirements among a set of processing units.

The meta-search engine components, as defined in [5], divide up naturally in a P2P context like Gnutella as it can be seen in Fig. 1. The *Document Selector* component is not applicable for our purposes since we used commercial, general-purpose search engines to retrieve sets of relevant documents, whose most important characteristics and policies are proprietary.

Perhaps the most important feature of such a solution is the distribution of the *Results Merger* component. The reason for that is because, depending on

[1] `http://www.limewire.com`

[2] `http://www.bearshare.com`

[3] `http://gtk-gnutella.sourceforge.net/`

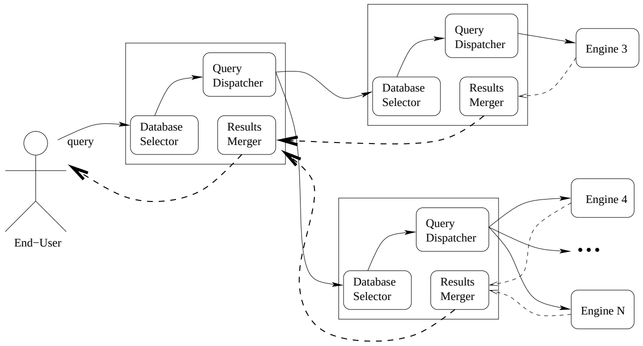


Fig. 1. An example P2P meta-searching network.

the re-ranking strategy that we follow and on the desired number of results, its complexity is usually non-trivial. Another reason why such a distribution is an interesting feature is that depending on the different information sources, we would, ideally, be able to apply different recombination procedures at different nodes of the network (stages of the retrieval process).

3.2 Distribution of Meta-Search Engine Components

Database Selector and Query Dispatcher The *Query Dispatcher* component is responsible for the modification and routing of the queries in the network. It perceives the rest of the network as a black box. The interface of their interaction is the nodes that a potential query can reach from the current peer.In our prototype, the *Database Selector* was just granting permission to route in an un-informed manner.

Result Merger Probably the most important component concerning the particular experimental system was the *Result Merger*. This component receives results from the nodes its peer has previously sent queries to, re-combines them and routes them back. The re-combination policy followed is discussed in Section 3.4.

3.3 The Components of the System

For the implementation of the system we used the JTella [4] Gnutella library. JTella is written in Java and implements the Gnutella protocol in an intuitive way although it is still at an early development stage. JTella also implements all the Gnutella descriptors described in [7], which constituted an additional convenience factor for information retrieval.

The web meta-searching system developed can be divided into four (4) major components:

`PeerFrame` This is the main frame of the program. This meta-searcher is a client-based one, even though disadvantages of this kind of systems have been noted

[6]. This approach was decided since the system was aimed for experimental purposes and its easy configuration and adaptation was desired.

`GUI_Connection` This is the main means for the `PeerFrame` to communicate with the JTella classes and the rest of the logic in the application.

`Receiver` This is actually the peer's listener. It listens for messages arriving at a predefined port and, depending on the kind of message, it performs the analogous actions.

`SearchEngineParser` This component is responsible for managing the search engines that its peer is responsible for.

3.4 Combination and Re-ranking

For the combination of the results in the system, we used the Dempster-Shafer (D-S) Theory of Evidence Combination as presented in [2].

Suppose that a peer is connected to k information sources and one of those sources, j, returns n_j results in total. Then, the initial score of each of j's returned results would be:

$$S_{i_j} = \frac{[n_j - (p_{i_j} - 1)]}{R_j} \tag{1}$$

where p_i is the proposed, by the information source, rank of the result i (i.e. 1 for the most relevant result, 2 for the next and so on) and R_j is given by $R_j = \sum_{\iota=1}^{n_j} \iota$, which acts as a normalising factor.

Each web search engine has been assigned a positive real number β_j, an un-trust coefficient, where $0 \geq \beta_j \geq 1$. Normally, this comes from a trust coefficient which is provided by the user or calculated by a relevance feedback process; on this prototype though each search engine was assigned a constant un-trust coefficient. β_j is assigned to the entire set of documents and is interpreted as the uncertainty of the source of informal evidence. By using β_j and Equation 1, we can evaluate the mass function for each result:

$$m_j(\{d_i\}) = S_{i_j} \times (1 - \beta_j) \tag{2}$$

Finally, the results coming from different information sources (different evidence) can be calculated by applying the Dempster-Shafer theory of Combination of Evidence as follows. For each two information sources 1 and 2, the new mass function of each result d_i of the information provider 1 is given by:

$$\begin{aligned} m'(\{d_i\}) &= m_1(\{d_i\}) \otimes m_2(\{d_i\}) \\ &= m_1(\{d_i\}) \times m_2(\{d_i\}) + m_1(\{d_i\}) \times m_2(\Theta) + m_2(\{d_i\}) \times m_1(\Theta) \end{aligned} \tag{3}$$

where Θ is the global set of documents.

Additionally, the new un-trust coefficient $m'(\Theta)$ of the combination can be obtained from

$$m'(\Theta) = m_1(\Theta) \times m_2(\Theta) \tag{4}$$

Any new set of results, from a third information source, can be combined further by re-using Equations 3 and 4. This is a simplified form of D-S theory for IR purposes whose details can be found in [2, 3].

4 Evaluation and Comparisons

The retrieval efficiency of a tightly controlled Gnutella P2P network depends highly on the peer topology used. In this section we describe the evaluation procedure we followed in order to measure the effect of various topologies on efficiency. Followed to that, we will briefly present and discuss over the acquired results.

4.1 Evaluation Method and Results

The different topologies that were evaluated can be seen on Fig. 2. For each of these topologies a fixed three-term query was issued to the network from the initiating peer[4]. Also, for each experiment, a standard number of results was required by each participating search engine.

We ran this experiment requesting ten (10), twenty (20) and fifty (50) results from each web search engine in order to observe the difference in retrieval times depending on the number of requested results.

After executing the experiment for each of the topologies of Fig. 2 and for each of the different number of expected results, we obtained the average times presented in Fig. 3.

As it can be seen from Fig. 3, the centralised approach (Fig. 2(a)) was proven to be worse than all the distributed ones. We believe that this is a strong indication of the potential of distributed P2P IR systems from the retrieval efficiency point of view.

An interesting outcome is that the linear setup with a repeating search engine (Fig. 2(d)) was closely as inefficient as the centralised one (Fig. 2(a)).

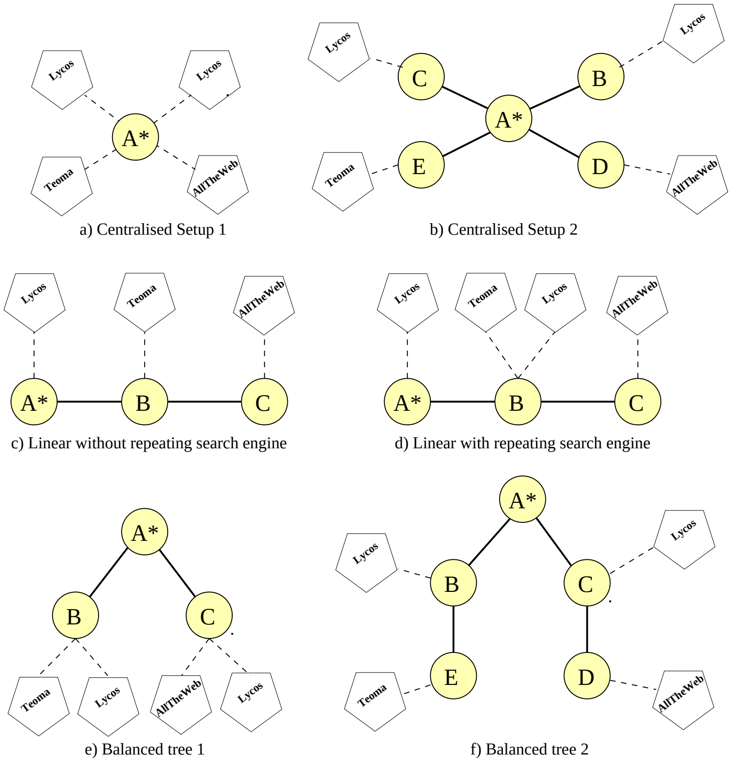
The tree structures provided us with useful insight. Firstly, the fact that Balanced Tree 2 (Fig. 2(f)) was significantly more efficient than Balanced Tree 2 (Fig. 2(e)), clearly depicts the advantage of well thought-out distributed systems over the more centralised ones for non-trivial tasks.

Lastly, the most effective approach was the Centralised setup 2 (Fig. 2(b)). We believe that this setup was better than the others because of the small number of search engines used as well as of the fast internal LAN that these experiments were executed over.

5 Conclusion and Future Work

In this paper it was shown that distributed P2P solutions have a clear potential for more efficient IR; in particular that networking topology plays an important role in retrieval effectiveness. This was investigated within the meta-search engine context but, in fact, IR can benefit from P2P solutions in a much more direct and whole way. The utter aim of the P2P IR approach is that information

[4] denoted by an asterisk in Fig. 2



a) Centralised Setup 1

b) Centralised Setup 2

c) Linear without repeating search engine

d) Linear with repeating search engine

e) Balanced tree 1

f) Balanced tree 2

Fig. 2. The topologies used during the benchmarking of the P2P meta-search engine.

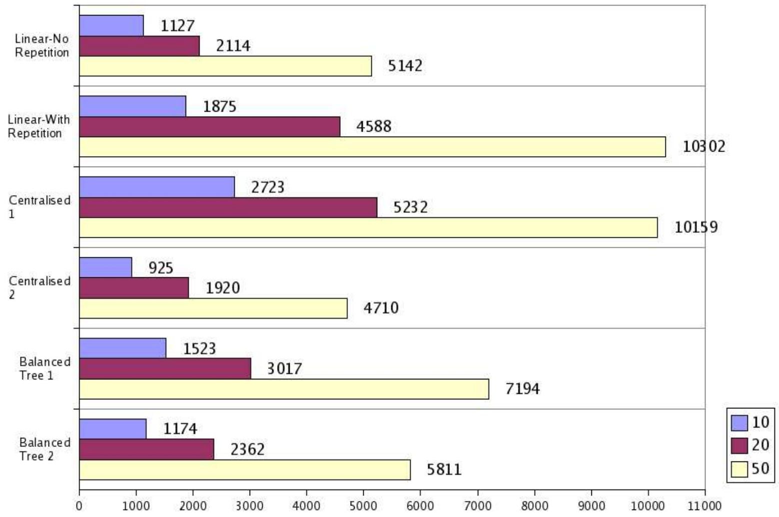


Fig. 3. Comparison of retrieval effectiveness.

should be retrieved from a wide variety of information sources, domains and underlying topologies and architectures, seamlessly and transparently to the end user.

We believe there are numerous possibilities for future work in this field so that both efficiency and effectiveness of IR can be significantly increased. Firstly, concerning this particular piece of work, a large-scale evaluation, incorporating more than one queries, of the P2P meta-search system described in Section 3 is being considered to take place in the near future. Additionally, of high importance would be the systematic exploration of suitable meta-data for IR over P2P networks. Another potential research field is the investigation of ways of finding willing and suitable information sources in a dynamically changing network. Such research would aid the IR process in larger and loosely controlled environments such as the Internet.

6 Acknowledgements

The authors would like to thank Dr. Peter Dickman of the University of Glasgow for his valuable help on practical matters concerning the development of distributed and in particular P2P systems.

The first author would also like to thank the Department of Computing Science of the University of Glasgow for the summer internship opportunity that it provided, thus allowing him to investigate further the field of P2P IR.

References

1. *PEER-TO-PEER: Harnessing the Power of Disruptive Technologies.* O'Reilly & Associates, Inc., 101 Morris Street, Sebastopol, CA 95472, March 2001.
2. Joemon Jose and David J Harper. Retrieval mechanism for semi-structured photographic collections. DEXA'97, pages 276–292, Toulouse, France, 1997. Springer.
3. Joemon M Jose. *An Integrated Approach for Multimedia Information Retrieval.* PhD thesis, The Robert Gordon University, April 1998.
4. Ken McCrary. The gnutella file-sharing network and java. *http://www.javaworld.com/javaworld/jw-10-2000/jw-1006-fileshare.html*, as viewed on November 20th 2002.
5. Weiyi Meng, Clement T. Yu, and King-Lup Liu. Building efficient and effective metasearch engines. *ACM Computing Surveys*, 34(1):48–89, 2002.
6. Wolfgang Sander-Beuermann and Mario Schomburg. Internet information retrieval - the further development of meta-search engine technology. In *Proceedings of the Internet Summit*, Genf, July 22-24 1998. Internet Society.
7. Clip2 Distributed Search Services. The gnutella protocol specification v0.4. *http://www.gnutella.co.uk/library/pdf/gnutella_protocol_0.4.pdf*, as viewed on November 3rd 2002.