An Architecture for Peer-to-Peer Information Retrieval

Iraklis A Klampanos
Department of Computing Science
University of Glasgow, Scotland
iraklis@dcs.gla.ac.uk

Joemon M Jose
Department of Computing Science
University of Glasgow, Scotland
jj@dcs.gla.ac.uk

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous—*Information Retrieval*; C.2.4 [**Computer-Communication Networks**]: Distributed Systems

General Terms

Management, Measurement, Design

Keywords

P2P IR, Peer Clustering, Query Routing

1. INTRODUCTION

P2P networking is one of the most rapidly developing areas of modern computing. By the utilisation of the exponentially increasing Internet nodes (users) as well as the ever powerful home computer systems and mobile devices, the P2P paradigm attempts to create open and collaborative networks of the most diverse functionality nature.

In this study we propose an architecture for IR over large semi-collaborating P2P networks based on clustering. By the term "semi-collaborating" we mean networks where, although peers have to collaborate in order to achieve overall effectiveness, they do not have to share any proprietary information with the rest of the network, nor do they have to be consistent with respect to the IR systems they use. Also, we reason toward the usefulness of clustering in open P2P networks by relying on two basic assumptions (introduced in Section 3.1).

2. BACKGROUND

After the explosion of file-sharing protocols like Gnutella, the Infrasearch project [1, page 100] demonstrated the potential of IR over open P2P networks. However, potential problems were also revealed. The problem caused by the naive approach of query flooding is twofold: the scaling of the network becomes impossible and also the quality of the returned results is limited. Therefore the most immediate challenge is that of the effective discovery of potentially useful peers and of the efficient routing of queries to those peers only.

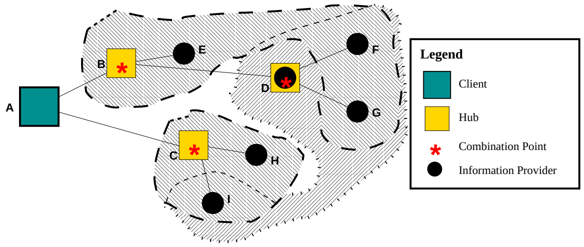

Figure 1: A sample P2P network.

Approaches so far are more or less divided between distributed hash table (DHT) approaches and document description advertisement approaches. The major disadvantage of DHT approaches is that they are unable to cope sufficiently well with keyword searching. The content description approaches are more promising in terms of IR, but they usually do not scale up beyond some thousands of peers.

3. A PROPOSED ARCHITECTURE

Recognising the nice properties of division of labour within a P2P network as proposed by JXTASearch in [4], we adopted a similar model without deviating from the definitions we proposed in [3]. In our model the peers may choose to implement one or more of the following services. The *Client Service* provides the user-end interfacing with the network; the *Information Provider Service* denotes willingness and ability to share documents (for our experimental purposes, textual); the *Combination Service* handles the fusion of results on behalf of other weaker peers; and finally the *Hub Service* performs various management activities in the network. Similarly to JXTASearch, *Hub*s are the only entities that are allowed to interconnect with each-other as well as with other peers, thus forming network topologies. For our experiments, we restricted the results combination to happen on *Hub*-enabled peers only. An example P2P network is depicted in Fig. 1.

3.1 The Assumptions

We base our work, by taking into consideration existing P2P file-sharing applications, on the following assumptions.

1. *Individual peers will tend to hold information relevant to a small number of queries. That is, the user's information provision area will not be unlimited nor random.*

2. *Documents that are outliers to some peers will have a high probability to also reside into peers where they will be part of the information bulk.*

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.
ACM 1-58113-646-3/03/0007.

3.2 Clustering

To avoid query flooding, we chose to deploy simple clustering techniques. Within the individual peers' collections we cluster the documents using a simple form of hierarchic clustering. The descriptor of each document is simply its term frequency (tf) vector. For each of the internal clusters, we also compute two statistics to aid us in the further clustering of peers.

The first metric is the average standard deviation $\overline{\sigma}$ of the tf components among the respective documents within each cluster and the second one is the participation level of a particular cluster $\mathcal{P}$, which is calculated as $\mathcal{P} = \frac{\#\text{docs within cluster}}{\#\text{docs within peer}}$. Therefore, each cluster within a peer is expressed in terms of its centroid document D^*, $\overline{\sigma}$ and $\mathcal{P}$.

At the networking level, peers get clustered into what we will, for clarity reasons, refer to as *Content-Aware Groups* (CAGs). For this clustering procedure we use a one-pass algorithm, but we also take into consideration the two metrics described in the previous paragraph; peers get clustered according to their content differences as well as in terms of $\overline{\sigma}$ and $\mathcal{P}$. Peers can belong to more than one CAGs depending on their internal clusters. The latter are thought to represent the internal information content areas of the peers. The network, effectively the *Hub* layer, gets informed about groups of peers and their content and $\overline{\sigma}$ and $\mathcal{P}$ characteristics (which for a CAG are calculated by averaging its members' corresponding figures).

3.3 Query Routing

Upon receiving a query, a *Hub* ranks the CAGs of the network according to their characteristics by assigning a score to each of them. We calculate this score S as $\alpha\text{CosDiff} + \beta(1-\overline{\sigma})+\gamma\mathcal{P}$[1], where CosDiff is the cosine difference between the incoming query and each CAG's centroid document. A satisfactory set of values for α, β and γ were derived experimentally as 0.8, 0.15 and 0.05 respectively.

For a number of requested results n, we contact the top CAGs with the highest participation levels until n results have been collected. The justification behind the appropriateness of using $\mathcal{P}$ for getting the top results lies within the two assumptions of Section 3.1.

Finally, the results are combined incrementally at the combination points, as they are routed back to the client, by using the Dempster-Shafer (D-S) theory of evidence combination, described in detail in [2].

4. EVALUATION

For our evaluation purposes we used the AdHoc TREC collection and the relevance assessments from TREC 6 and 7 (100 topics in total). The collections we used comprised of 556,077 documents of various lengths. Our experimental setup simulated 1,500 peers. In order to approximate the first assumption of Section 3.1, we assigned the relevant documents to different queries into separate peers in a way that they constituted the bulk of documents within each peer's collection. The rest of the peers were assigned documents at random, which contradicts the same assumption and affected heavily, in a negative way, our results. Finally we issued the 100 queries and calculated the corresponding P-R values. We also evaluated P-R values for the global collection (as a single, centralised database). The results obtained, for 20, 30, 50 and 100 retrieved documents, can be seen in Fig. 2. In this figure, *NZP* and *NZR* are the average P-R values without taking into consideration 0.0 P-R resulting queries (justification is given in the next paragraph); *P* and *R* are the average P-R values and *GlobalP* and *GlobalR* are the average P-R values taken from the global collection. For our evaluation we used the MG system.

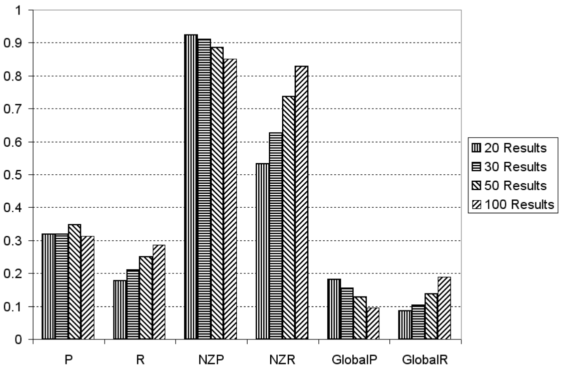

Figure 2: Precision - Recall Results.

For some result sets we got P-R values of 0.0, meaning that the query had not been routed to the relevant peers. However, bearing in mind that the majority of the peers contained randomly allocated documents, their centroid vectors (which don't bear any significance) might, by pure chance, have been closer to particular queries than the centroids of the actual relevant peers. This is close to the worst case for our network and that is why we also provide the average P-R values without taking into consideration those cases.

5. CONCLUSIONS

We consider our results to be significantly better than the centralised corpus in terms of P-R. We believe that if our assumptions were reflected by the test collection there would be a clear advantage of using such a distribution. However further analysis of the obtained results is still taking place.

6. ACKNOWLEDGEMENTS

The authors would like to thank Dr. Peter Dickman for his valuable discussions.

7. REFERENCES

[1] *PEER-TO-PEER: Harnessing the Power of Disruptive Technologies.* O'Reilly & Associates, Inc., March 2001.

[2] J. M. Jose. *An Integrated Approach for Multimedia Information Retrieval.* PhD thesis, The Robert Gordon University, April 1998.

[3] I. A. Klampanos, J. J. Barnes, and J. M. Jose. Evaluating peer-to-peer networking for information retrieval within the context of meta-searching. In *LNCS 2633*, pages 528–536, Pisa, Italy, April 2003. ECIR'03.

[4] S. Waterhouse. Jxta search: Distributed search for distributed networks. http://search.jxta.org/JXTAsearch.pdf, May 2001.

[1] $(1-\overline{\sigma})$ makes sense since it is calculated upon normalised tf vectors and therefore $\overline{\sigma} \in [0, 1)$