

P2P-4-DL: Digital Library over Peer-to-Peer

James Walkerdine, Paul Rayson
Computing Department, Lancaster University, Lancaster, UK
{walkerdi, paul}@comp.lanc.ac.uk

1. Introduction

Digital Libraries (DL) are collections of digital objects that are accessible to users via digital/electronic interface technologies (such as web browsers). A typical DL may store documents, images, sounds and video media.

Existing DL systems are usually centralised, with digital objects being stored on a server where they can be checked in and out by the clients. Although this approach can provide advantages, such as the ease of versioning control (if required) and security regulation, there are also negative aspects such as the large network load around the server and the inefficient use of resources around the network. In fully client-server systems a large proportion of the resources that are available on the network (such as storage space and network bandwidth) remain un-used.

The P2P-4-DL project aims to investigate and build a DL system that would operate over a P2P structure. Rather than storing digital objects centrally they remain the responsibility of the individual peers that provide them. This allows the system to utilise network resources more efficiently as well as providing users with a greater sense of control over the digital objects they share. Our prototype also draws upon Natural Language Processing (NLP) techniques in an attempt to increase the usability of the system. Other related work within this area includes EDUTELLA[1], a RDF based P2P infrastructure that can support the development of DL's.

The P2P-4-DL project is ongoing with an initial version of the prototype having been produced. This paper summarises the features that have been provided so far and our intentions for further development. The current release is publicly available to download.

2. Building the P2P-4-DL prototype

In order to focus on the development of a P2P based DL we decided to build our prototype on top of Lancaster's existing P2P Application Framework. The framework has been discussed in detail elsewhere [2], but essentially it can be viewed as an abstract layer

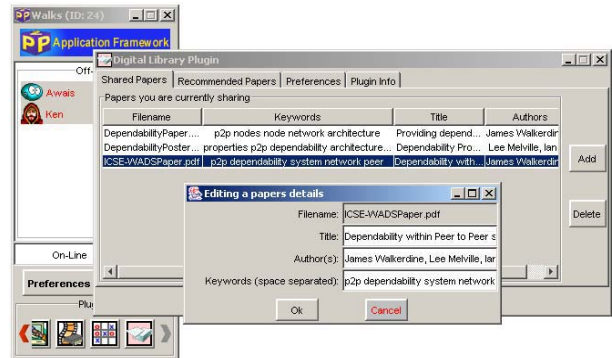


Figure 1 - The P2P-4-DL prototype

geared specifically for P2P application development. The framework reduces the burden on developers to understand the underlying P2P technology by instead providing them with a set of generic application orientated services. The framework operates over a semi-centralised network structure where an index peer exists that supports the running of the network.

Our DL prototype is a plug-in within the framework structure and makes use of its search, file sharing and awareness services. Figure 1 provides a screenshot of the prototype in use.

The current release of the prototype focuses primarily on document objects, however it can also support other types of digital object. The prototype provides a number of features:

Sharing of documents on the P2P network - The DL plug-in allows users to make their documents (resources) publicly available to other users on the network. However, unlike traditional DL systems where documents are typically stored centrally, in our P2P based version a user's documents are stored on their PC. References to these documents (based on title, authors and keywords) are then registered with the index peer, allowing for a Napster style search system. Obviously when a user disconnects from the network his/her documents are also removed. Currently there exists no redundancy within the plug-in, although this is an area that we intend to explore in the future.

Automatic document keyword identification - Our plug-in also draws upon our existing NLP research within the department¹ to allow for the automatic generation of document keywords if desired. A combination of linguistic annotation for part-of-speech and semantic fields, and word frequency profile analysis results in the generation of a set of keywords for a document. This analysis can be either simple word frequency or semantically based and is calculated by comparing the document's frequency profile to a standard frequency profile generated from a much larger corpus of natural language [3]. Currently the prototype is able to handle Microsoft Word, PDF and plain text documents. Document processing itself is a CPU intensive task and so the plug-in performs this in the background [4].

Searching of documents on the network - Shared documents that have been made available as part of the DL can be searched in a manner similar to that used by applications such as Napster. Using the P2P Application Framework users can interrogate the index peer and search for documents by providing keywords, or on a per user basis (those providing the documents).

Recommendation of documents - It was also desirable to make the DL plug-in pro-active, in that it would also suggest potentially relevant documents to users. To achieve this users are able to maintain an interest profile comprised of area keywords (for example, P2P, dependability, etc). Periodically the plug-in on each peer performs a transparent search for documents on the network that satisfy these keywords. The returned results are stored and are made accessible to the user upon demand. It was important to make sure these recommendations were not intrusive and so, although the searches are being performed in the background, it is left to the user to actually view the results.

Retrieving the documents - The P2P Application Framework carries out the transferral of documents between peers transparently to the DL plug-in. The name of the document and its source is passed to the framework which then negotiates the document's transfer.

3. Future work

The initial release of the DL prototype has been made publicly available and has been used and tested within the Computing department as a means to share academic related documents, for example academic papers. Although feedback has been positive, there are still a number of areas that we intend to further develop (aside from supporting more types of digital objects).

Redundancy - because peers can join and leave a network at will there is the issue of being able to

guarantee the availability of objects within the DL. To achieve this we intend to investigate the possibility of incorporating redundancy into the system, where digital objects will be duplicated on many peers. This will draw upon the P2P Applications Framework network monitoring facilities, in order to ensure a fairly high level of availability for digital objects (for example, by monitoring a peer's typical on-line time).

Improved NLP support - currently the prototype can only generate keywords for a document. A beneficial extension to this would be to automatically extract the title and authors. This is more complicated than standard keyword generation as it involves additional structural parsing and name identification.

Additional DL features - the current version of the prototype only provides very general features. We wish to extend the prototype to include *social* support in the form of user - user communication and social recommendations of shared digital objects. Versioning support is another feature we also intend to examine.

4. Conclusions

This paper has presented our initial work on the P2P-4-DL project, to investigate and build a prototype Digital Library system that operates over a P2P network. A P2P based DL system provides numerous advantages in particular more efficiently using available resources. Our initial prototype development has been presented, highlighting the features it currently provides and our future work. The prototype has been utilised within an academic environment and initial feedback has been positive.

The DL prototype along with the P2P Application Framework is available for download, from our departmental P2P website - <http://polo.lancs.ac.uk/p2p>

5. Acknowledgements

This work has been part-funded by the European Commission within the P2P ARCHITECT project (IST-2001-32708).

6. References

- [1] Nejdil, W., Wolf, B., Qu, C., Decker, S., Sintek, M., Naeve, A., Nilsson, M., Palmer, M., Risch, T., EDUTELLA: A P2P Networking Infrastructure based on RDF. In *proceedings of WWW 2002*, May 2002, USA.
- [2] Walkerdine, J., Melville, I., Sommerville, I., A Framework for P2P Application Development, *Technical Report COMP-004-2004*, Computing Department, Lancaster University, 2004.
- [3] Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In *proceedings of the workshop on Comparing Corpora, held in conjunction ACL 2000*. October 2000, Hong Kong, pp. 1 - 6.
- [4] Sawyer, P., Rayson, P., and Garside, R. (2002) REVERE: support for requirements synthesis from documents. *Information Systems Frontiers Journal*. 4, (3), Kluwer, Netherlands, pp. 343 - 353.

¹ <http://www.comp.lancs.ac.uk/ucrel/>