Linking and Navigating Data in a P2P File-Sharing Network

Alan Davoust
Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada
adavoust@sce.carleton.ca

Babak Esfandiari
Carleton University
1125 Colonel By Drive
Ottawa, Ontario, Canada
babak@sce.carleton.ca

ABSTRACT

We demonstrate a tool for publishing and navigating linked data over the highly dynamic infrastructure of a P2P file-sharing network. Our links are based on a URI scheme which allows unambiguous designation of replicated data items, regardless of their location in the network. In a true decentralized P2P spirit, users publish and distribute the links just like other data items.

Categories and Subject Descriptors

H.5.4 [**Hypertext/Hypermedia**]: Navigation

General Terms

Keywords

Peer-to-peer, linked data, URI, demo

1. INTRODUCTION

The tool that we demonstrate is a framework for publishing linked data in a P2P *File-Sharing*-type infrastructure. By the term *File-Sharing* we refer to the typical properties of popular file-sharing networks, (e.g. Napster[1], Limewire[2], etc.) in terms of churn, decentralized control, and data circulation. More specifically, we assume a highly dynamic and unstructured network, with peers frequently appearing and disappearing, where each peer is in full control of its local storage space, and where data is propagated and replicated by downloads.

We contrast this model with other P2P applications such as *Peer Data Management Systems* (PDMS), (e.g. Piazza [3]) which are essentially distributed database systems, and Distributed Hash Tables (DHT, e.g. CHORD [5]), which are rather distributed storage facilities. PDMS are meant to answer expressive database queries and usually avoid data replication, and DHT are meant to optimize access to specific data items based on system-generated keys.

In a File-Sharing infrastructure, popular files are more available due to the high number of copies stored in different peers. We generalize the concept of a *file* - identified by no more than a name and some binary content - to complex *data items*, structured according to an arbitrary schema, with a number of binary and non-binary fields.

[1]no longer exists in its original form
[2]http://www.limewire.com

LDOW2009, April 20, 2009, Madrid, Spain.

Our prototype application U-P2P (for "Universal P2P") was first created (and presented in [4]) as a tool to share data with arbitrary schemas over a P2P network, and has since then been extended to combine simple queries and downloads with a more exploratory *navigation* feature across a graph of linked data items.

U-P2P defines the notion of a *community* as the abstraction of a data schema, and some presentation tools specific to that schema, and supports a URI scheme to uniquely identify a data item, based on the two-level hierarchy *community / data item*. This URI scheme allows us to create *links* which point to *any copy* of a replicated data item, taking advantage of data replication to optimize its availability, in true P2P spirit.

Our demonstration will show the simplicity of publishing linked data in the U-P2P framework, and navigating the resulting graph.

In this short description we discuss the *data items* that are shared in our P2P network, and the P2P-specific URI scheme that supports linking and navigation, then finally we outline some perspectives for future development.

2. DATA AND DATA PRESENTATION

The data items of interest to us are structured according to some schema, possibly combining simple elements such as numbers or short text fields, with binary elements, or "blobs" in database jargon. We will hereafter use the term "attachments" after the idea of email attachments.

Tuples conforming to such schemas can be seen as an abstraction for files with meta-data attributes, or for entries in databases.

Most of the data found on the web has a semantics that end-users can best access using customized interfaces; for example, geographical data is best viewed on a map, and complex molecule descriptions are certainly best-viewed in a custom browser plugin that can produce fancy 3-D renderings of the molecules.

Current tools such as the Tabulator [1] still only offer a limited number of "views" on the data (such as a map view and a calendar view), that are hard-coded into the browser itself. The Fresnel [2] language is a presentation language for RDF data that goes in the direction of providing such customized views of data.

The data model of U-P2P allows us to bundle a data schema along with some user-friendly "views" as a single data item with several *attachments*. Such a data item defines the reified concept of a *community*. A peer who downloads a *community data item* obtains by this means the community

schema, and a set of presentation templates to search and view data structured according to the schema. Downloading such a data item thus becomes a way of *joining* a community of peers sharing a particular schema, such as a molecule description schema. A single U-P2P client can simultaneously manage any number of *communities*.

The explicit schemas and data presentation tools of U-P2P, based on XML and the XSLT language, provide an opportunity for end-users to easily contribute new schemas and data using a simple interface of HTML forms. This feature is very much in line with the current trends of the so-called Web 2.0, giving the end-user a producing role.

3. LINKS IN A DYNAMIC P2P NETWORK

We have presented the basic data model of U-P2P, which extends traditional P2P file-sharing systems by supporting multiple, arbitrary schemas.

The next step, which we introduce in this demo, is to support links and navigation between data items.

Links on the WWW are based on a hierarchical addressing system that relies on the assumption that documents are statically stored by a single server, and hence are tied to the location of the document on that server.

A central characteristic of P2P file-sharing systems is that whereas a given peer is as likely as not to be available at any given time, documents (in our model, data items) tend to be replicated and stored by many peers, and hence be more likely available somewhere in the network.

We introduce a URI scheme based on the following assumptions:

1. Peers only store data of a given *community* if they also store the community data item containing the schema and view(s) defining this community;
2. Within a given community, a data item can be uniquely identified (e.g. by a one-way hash);
3. Communities can be uniquely identified within the greater set of communities, in the same way as any data item.

The URI scheme supported by U-P2P is hence a sequence of two MD5 hashes, one for the data item itself, and one for the community that it was based on, e.g. `up2p:h3214/h3487`.

The dereferencing of a URI is then a limited flooding search within the peer neighborhood: the peer maintains a community-specific list of neighbors which share the community, and sends a query message requesting the data item identified by the hash.

The simplest navigation is then a one-way navigation following the model of the WWW, but we have found it interesting to provide two-way navigation of links. A naive implementation of this would require searching all existing data for links pointing to the current data item, obviously a costly solution.

Inspired by the concept of RDF links, we have created a community with the simple data schema of a triple $(URI_1, label, URI_2)$ where URI_1 and URI_2 identify two data items. By storing these triples in a single community, which peers are free to join, the search for "reverse" links is limited to a single community.

Users can thus simply search and navigate links to and from any data item that they are viewing. As members of the "links" community, they can also publish links between arbitrary data items, which then become visible to all their neighbors, and may be propagated and replicated across the network. Furthermore, on searching for linked data, the results can be sorted by popularity, since the number of copies retrieved for a link and for a document are direct indicators of their popularity.

4. CONCLUSION AND FUTURE WORK

Our framework U-P2P develops the ideas of Linked Data on a P2P file-sharing infrastructure, using a URI scheme and a dereferencing mechanism that takes advantage of data replication, a specific property of file-sharing networks, to mitigate the possibly low availability of peers.

In this implementation, links are shared in a specific community modeled on the simple triple format of RDF links. Our application supports *navigation* of such RDF links, and we believe it would be a relatively small step to extend the interface to integrate a SPARQL query engine.

The use of labels, and possibly a richer range of properties - based on our flexible framework - could allow for future semantic-web oriented extensions, where new links between data items could be inferred, for example, from the presence of two existing links and the known transitivity of a particular *link type*.

5. REFERENCES

[1] T. Berners-lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing linked data on the semantic web. In *In Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.

[2] C. Bizer, E. Pietriga, D. Karger, and R. Lee. Fresnel: A browser-independent presentation vocabulary for RDF. 5th International Semantic Web Conference, November 2006.

[3] A. Y. Halevy, Z. G. Ives, J. Madhavan, P. Mork, D. Suciu, and I. Tatarinov. The piazza peer data management system. *IEEE Transactions on Knowledge and Data Engineering*, 16(7):787–798, 2004.

[4] A. Mukherjee, B. Esfandiari, and N. Arthorne. U-P2P: A peer-to-peer system for description and discovery of resource-sharing communities. In *ICDCS Workshops*, pages 701–705. IEEE Computer Society, 2002.

[5] I. Stoica, R. Morris, D. Liben-Nowell, D. R. Karger, M. F. Kaashoek, F. Dabek, and H. Balakrishnan. Chord: a scalable peer-to-peer lookup protocol for internet applications. *Networking, IEEE/ACM Transactions on*, 11(1):17–32, 2003.