PPDN — A Framework for Peer-to-peer Collaborative Research Network

Vlado Kešelj and Nick Cercone

Faculty of Computer Science, Dalhousie University
E-mail: {vlado,nick}@cs.dal.ca

Abstract

PPDN, Push-Pull Distribution Network, — a proposal for a novel framework for peer-to-peer collaborative research network is presented. Some requirements not addressed by the currently proposed systems are discussed, and we show how these issues are addressed in our framework. The framework is based on a distributed approach and the concept of semantic web. The collaborative network is represented as a graph, with push and pull edges. The nodes can act as autonomous or semi-autonomous agents, implementing different policies.

1 Introduction

The Internet is continuously maturing from its early years of exciting but somewhat mechanical and static applications and protocols toward a more flexible and more intelligent network. Although the traditional means of communication and information sharing on Internet, such as e-mail, WWW, or Usenet, still require further research to address the problems such as spam, authentication, and information privacy, we can say that their scope and usage are well-understood. Under this umbrella of traditional methods, we could add search engines, database interfaces, e-mail lists, and web-based forums. The new level of integration and collaboration includes the so-called groupware applications, peer-to-peer systems, and similar kinds of distributed systems.

From a vast area of different application domains we limit our domain to the web-based research support systems. To give a motivation for such system, we list some of the activities from the life of a typical researcher X that are not well supported currently:

- easy access to relevant publications and to corresponding meta-data (e.g., BibTeX entry),
- keeping track of X's publications, in X's own database, using it to generate a Web list, in her/his CV, grant applications etc.,
- passing publications or their metadata to the research group(s) web sites, selectively, to co-author, collaborators, organizational web site, wider research community, research search engines, and similar,
- receiving information about new publications, conference announcements, calls for papers (CFPs), software releases, books, and similar.

While an obvious item of exchange described above is a publication, there are several different types of information that require similar kind of dissemination:

- publications and publication metadata,
- software and software metadata,
- conference calls for papers (CFP), and
- links, web resources, and web services.

Additionally, in order for our application to be useful and to be used, the experience has shown that the following requirements also need to be satisfied:

Low maintenance: The researchers are usually happy to share their contributions, but they refuse to put any significant work into preparing meta-data and system maintenance [6].

Non-centralized: Non-centralized solutions do not scale very well. They also represent a one-size-fits-all approach, which hardly fit in a wide domain such as scientific research.

Flexible: Setting elaborate and rigid standard and frameworks in advance would be premature. It is hard to predict future requirements, and complex standards require time to learn and train. It is desirable for a standard or framework to be learnable incrementally—learn only as much as you need. Such flexibility would provide an environment for emerging standards and solutions.

Under flexibility, we also assume connection flexibility. Instead of a rigid distributed system depending on real-time communication among peers, we put forward a network for information dissemination using push and pull communication links.

2 Related Work

We divide the related work into two groups: the centralized repositories and peer-to-peer (P2P) systems.

Centralized repositories. The centralized research repositories are available to the scientific community for several years now. Some of them are CiteSeer[1] since 1999, DBLP[2], CS BibTeX[3], CompuScience[4], CoRR The Computing Research Repository or arXiv[5], NZ-DL[6], Zentralblatt MATH[7], and MathSciNet[8].

While they have proved to be invaluable to the research community, showing that they do scale up to certain non-trivial amount of publications[9], these centralized sites also confirmed weaknesses of the centralized approach. They are limited in scalability, for user connection as well as for submissions. The user completely depends on the connection to the site: so if the site is too busy, or simply down, the system is unusable. A user is not provided with software to maintain his own database of publications.

A new solution is needed, but still we would like to make the centralized repositories part of it. One step in this direction is CiteSeer's compliance with the OAI—the Open Archives Initiative protocol for metadata harvesting[10].

Peer-to-peer systems. Several P2P projects to support research are described recently.

Werlen 2003 [6] presents the DFN—the German Research Network, which is a non-profit organization that provides research infrastructure in Germany. The focus of the project is on the search capability in indexing and gathering scientific information. It is a peer-to-peer network that uses JXTA[11] open search protocol. An important fact noted in [6] that the messages in the network are much more efficiently exchanged if the network is organized around 'super-peers' or hubs, so that the small-world phenomenon can be exploited, i.e., the routes from peer to any other peer are always short. Another important observations are that researchers do not want to invest any significant amount of time to prepare data, and the open networks are prone to spam data, i.e., material inappropriate for the network.

Haase and Siebes 2004 [4] discuss peer selection in peer-to-peer networks with semantic topologies. The focus of this paper is on finding a peer in a peer-to-peer network that has relevant information for our query. Instead of the traditional approach where a query is broadcasted to all peers, they propose that peers advertise their expertise, which is organized into a semantic network. The approach resembles multi-agent systems proposed for distributed information retrieval about ten years ago, e.g., see [5].

The focus of our approach is different being focused on an approach of selective information dissemination instead of active peer querying, however an important commonality with [4] is the domain of application. Haase and Siebes [4] consider the case study of bibliographic metadata about publications, which is included in our target domains. The common ontology used in [4] is the Semantic Web Research Community Ontology (SWRC) [2].

Ahlborn *et al.* [**?**] discuss how an existing peer-to-peer system Edutella could be reused to provide OAI repositories with search capability.

Very recently, BIBSTER[12]—an open source P2P system for managing, searching and sharing bibliographic metadata from BibTeX files was announced [3]. The system is implemented in java on top of the JXTA platform. It provides search capability by routing the query to peers. Bibster is an application based on technology that combined Semantic Web and P2P technologies. It does not have centralized control.

3 Problem Specification

As we saw in the previous section, we could roughly divide the existing approaches to the problem of web-based research support into two groups: (1) centralized publication repositories like CiteSeer and arXiv, and (2) new distributed approaches such as Bibster, which somewhat resemble the multi-agent systems for information retrieval being proposed several years ago [5]. While we find the both of these approaches useful, we would like to offer a new peer-to-peer approach called *Push-Pull Distribution Network* (PPDN) to address certain applicative approach. The PPDN framework is designed to address the following issues:

- The weaknesses of centralized sites are well-known [3]: centralized server, which can be a single point of failure, it does not scale well with the number of users nor data items, complete dependence on direct network connection to the server and on its bandwidth and delay.

[1] http://citeseer.ist.psu.edu/
[2] http://dblp.uni-trier.de/
[3] http://liinwww.ira.uka.de/bibliography/index.html
[4] http://www.zblmath.fiz-karlsruhe.de/COMP/quick.html
[5] http://arxiv.org/archive/cs/intro.html
[6] http://www.nzdl.org/
[7] http://www.emis.de/ZMATH/
[8] http://www.ams.org/mathscinet/search
[9] DBLP announced recently that they reached 520,000 papers.
[10] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[11] http://www.jxta.org
[12] http://bibster.semanticweb.org

The existing P2P approaches have other issues:

- They typically require significant effort from users involved w.r.t. maintaining the peer system and keeping it on-line.
- While in a centralized system we rely on one host, which is normally reliable, in a P2P system we rely each time on different hosts, with expected higher probability that one of them may be off-line.
- P2P system relies on distributed querying in search for relevant publications. Routing, dividing, and merging such queries is a complex problem, it wastes bandwidth due to a lot of redundancies, and may have longer delay than the centralized approach. It is reported that such systems may produce too many queries if the network topology is not carefully designed [4].

These issues are addressed in the PPDN approach in the following way:

- PPDN is a distributed approach that does not require a centralized server. The users can with little effort keep their PPDN nodes, connect and disconnect them in a flexible way without disrupting significantly the system as a whole.
- We delegate the issue of searching and querying to reliable and high-performance servers, which are part of the PPDN network. These are equivalent to 'superpeers,' or hubs, as called recently. The long experience with information retrieval on the Internet provides arguments that a very distributed approach to information retrieval would not perform favorably compared to strong and reliable single-site search engines.
- The issue of relying on some peers to be on-line in a typical peer-to-peer system in a moment when we need information is addressed in PPDN by using the push and pull transfer of information. Rather than waiting for the moment when we need information, we focus on information dissemination, so that by the time we need information, it is available either locally or it is stored in a search engine repository. Thus, the system reliability is improved. In our prototype system, we rely on the e-mail protocol, SMTP, as the transport protocol, which further improves system robustness, since SMTP transfer can be performed over relays, not requiring that a sender and a recipient are on-line in the same time. The search and retrieval task is left to a centralized repositories which are part of PPDN.
- The network is semi-autonomous, allowing users by creating forwarding policies to create sub-networks, networks of trust, and to avoid spam.

```
X-DBWorld-Message-Type: conference/announcement
X-DBWorld-Name: iiWAS2004
X-DBWorld-Start-Date: 27-Sep-2004
X-DBWorld-Location: Jakarta; Indonesia; Asia
X-DBWorld-Deadline: 23-Jul-2004
X-DBWorld-Call-For: papers, demos, reports,
X-DBWorld-Web-Page: http://www.iiwas.org/conf...
```

Figure 1. DBWorld Example

PPDN Description. A PPDN is a network of nodes, i.e., a directed graph with two kinds of edges: *push* and *pull*. Each site is a semi-autonomous site since it can automatically forward or store received data items, or it can be moderated by a user. The data items are transfered through the edges using a transport protocol. In our prototype, we are use SMTP and CGI as the transport protocols. In a *push* connection, the sender initiates data transfer; e.g., buy sending an e-mail through a link; while in a *pull* connection, the receiver initiates transfer, for example, by accessing a web site, running a CGI script, or sending a query by e-mail. In our prototype, the pull connections are implemented using the CGI protocol.

4 Examples

The PPDN framework is not a completely new idea: it is more a matter of gluing and merging existing pieces than designing and launching a new paradigm from the scratch.

Example 1. The members of the DBWorld mailing list[13] may have noticed that the information about conference announcements is encoded using the RFC 822 standard into headers of the list e-mail messages. This represents an elegant example of a transition from natural-language-only to semantic web style of informing. An example is given in Figure 1

Example 2. The second example is taken from the arXiv mailing list [14] and it is shown in Figure 2. This is example of an e-mail list used to disseminate publication information in a well-formated, but still user-readable style. The formatting is similar to Example 1, following the style of the RFC 822 headers. The arXiv mailing list is integrated with the CoRR repository of the publications with a search interface.

Example 3. The third example presents an e-mail message used to disseminate information about new links available at the ACL NLP/CL Universe[15]. The ACL Universe

[13] http://www.cs.wisc.edu/dbworld/
[14] http://arXiv.org
[15] http://perun.si.umich.edu/˜radev/u/db/acl/

```
--------------------------------------------...
 Submissions to:
Computational Complexity

 received from  Thu  1 May 03 20:00:02 GMT  t...
--------------------------------------------...
\\
Paper: cs.CC/0305035
Date: Mon, 19 May 2003 16:02:54 GMT   (3kb)

Title: P is not equal to NP
Authors: Craig Alan Feinstein
Comments: The body is less than 2 pages and e...
  recently submitted to the SIAM Journal of D...
Subj-class: Computational Complexity
ACM-class: F.1.3
\\
  The question of whether the class of decisi...
solved by deterministic polynomial-time algor...
the class of decision problems that can be so...
polynomial-time algorithms (\textit{NP}) has ...
first formulated by Cook, Karp, and Levin in ...
prove that they are not equal by showing that...
solves the SUBSET-SUM problem must perform at...
\lfloor\frac{n}{2} \rfloor}$ computations for...
${\rm O}(n^2)$, where $n$ is the size of the ...
\\ ( http://arXiv.org/abs/cs/0305035 ,  3kb)
```

Figure 2. The arXiv Example

web site contains a hierarchy of links with descriptions, which is browseable as well as searchable. The hierarchy is encoded using attributes 'cat1,' 'cat2,' 'cat3,' and 'cat4.' The format is similar to previous examples, being text-based and having attributes and values paired at each line.

5 PPDN Framework

PPDN. Push-Pull Distribution Network (PPDN) is directed graph with two kinds of edges: *push* and *pull* edges. For two vertexes a and b, there may exist two edges (a, b), one push and one pull edge. The nodes can be regarded as information repositories. The direction of edges describe information flow. In a push edges, the information transfer is initiated by the source node, while in a pull edge the transfer is initiated by the destination node.

The transfers are either triggered by an event, they are invoked periodically, or they are invoked manually.

Node structure. The structure of a node in the network is shown in Figure 4. The information is received through in-edges and disseminated through out-edges. The edges are grouped into channels. For example, a channel is a list of e-mail addresses to be informed about new items. In the prototype we use plain text files for site archives, but one could use any database engine as well. The suggested organization of the archive is into the personal archives and cache archives.

```
--------------------------------------------...
Additions to the ACL NLP/CL Universe:
June 7 - October 20, 2003
--------------------------------------------...

link_id L000002988
url     http://cf.hum.uva.nl/computerlinguist...
title   Amstelogue\'99 - Workshop on Dialogue
author
cat1    CONFERENCE
cat2    1999
cat3    5
cat4
email
annotation      May 7-9, 1999, University of ...
date_added      Wed Jul 23 12:26:50 EDT 2003
date_indexed
```

Figure 3. The ACL NLP/CL Universe Example

The moderating module acts as a semi-autonomous agent. It can be configure to automatically store or forward received data items, wait for user approval, or drop them, based on a set of rules. The rules depend on the incoming or outcoming channel, but they can also be arbitrary regular expression-based rules on data items.

A typical scenario is the following: A researcher X would set up his own PPDN site. A department, research groups, projects, and collaborators would also have defined sites. A site typically would have a Web interface to produce a list of items.

Communication Issues. There are several communication issues that needs to be addressed in a PPDN network:

access control: If our site is source pull site, we may need a protocol to restrict access to the site based on channel. This can be solved in various ways based on the transport protocol. In our case, the CGI access is regulated through htaccess method.

authentication: If we are the receiver push site, we need a way to authenticate the sender. Since the transport protocol for push edges is SMTP, we use GPG (or PGP) public key signatures for authentication.

encryption: If we need encrypted transport, so that a third party cannot observe data transfer, in case of SMTP a GPG/PGP-based encryption is used, and in case of CGI, HTTPS protocol is used.

Transport protocol. In the prototype we use SMTP transport protocol for push, and CGI for pull edges. However, a whole slew of alternative transport protocols is available: SOAP, web services, scp and ssh, ftp, being among them.

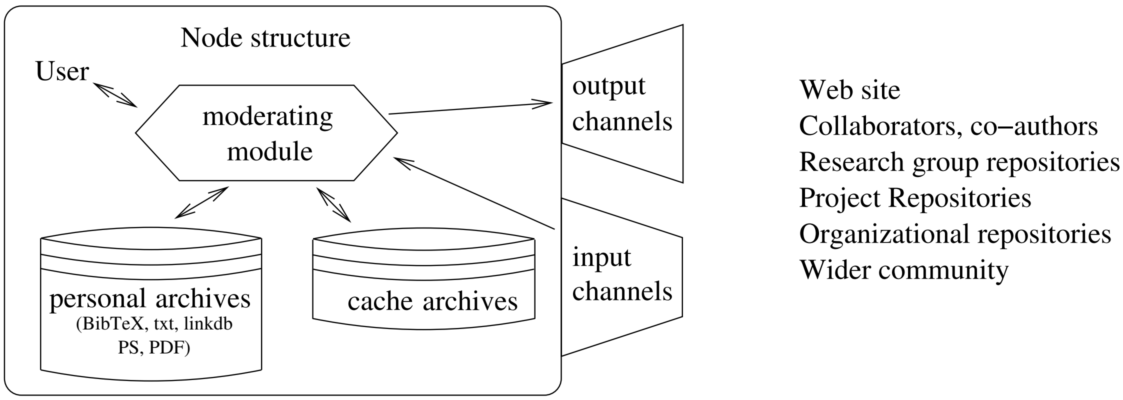

Figure 4. Node structure

Data items. The following data items are exchanged in our prototype:

- publication metadata,
- conference CFPs and announcements,
- software metadata,
- links (e.g., resources, web services),
- e-mail list metadata,
- publications, and
- software.

The set of items and their ontology is not rigidly defined, so in this prototype stage, the network can be used even as a distributed e-mail list—where there is no a centralized server, but each user can decide to have his own redistribution list.

There is an issue of infinite forward loops, which is resolved by keeping MD5 digests of passed data items in the cache archives, and dropping the ones that are repeated.

Encoding The standard encoding schemes used in similar semantic networks are XML and OIL. While we intend to provide a compliant translation into these standards, the prototype is based on a simpler encoding scheme, similar to the RFC 822 e-mail headers standard and YAML standard. An example of encoding of a CFP is given in figure 5. Several data items are separated by blank lines. Within a data item, each line starts with an attribute ending with a colon (:). A line may be continued by starting the next line with space or tab, or by ending the the line with backslash (\).[16] If a binary data needs to be encoded, an encoding standard, such as BASE64 is used. This is needed usually when large data items, such as papers or software is passed.

[16]The difference is that a line ending with backslash, the backslash will be removed and this is a way to encode a new-line character within an attribute value. In a line continued only by space or tab in the next line, the new-line character is removed.

```
Type: conference/announcement/cfp
Name: WSS'04
Full-name: The Second International Workshop on
            Web-based Support Systems
Comments: In conjunction with 2003 IEEE/WIC/ACM
          International Conference on Web Intelligence
Location: Beijing, China
URL: http://www2.cs.uregina.ca/˜wss/wss04/
Due: 20-Jul-2004
Start-Date: 20-Sep-2004
```

Figure 5. CFP Example

Policies. There are four kinds of policies defined for a moderating module in a node:

receiving policy: defining whether a data item will be received at all from a channel,

storing policy: defining whether a received data item will be stored in the cache archive,

sending policy: defining whether a new data item in the archive will be sent to a channel, and

forwarding policy: defining whether a received data item will be forwarded (even if not stored in the archive.).

The policies are rule based, taking into account the receiving channel, sending channel, and based on regular expression matching on data items. There are three policy results: (1) free, i.e., passing the data item, (2) blocked, i.e., dropping (deleting) the data item, and (3) moderated, i.e,, storing a data item in a waiting queue, waiting for users decision.

6 Conclusion

We presented design and a prototype implementation of the PPDN—Push-Pull Distribution Network—framework for peer-to-peer research collaboration support. The current systems were discussed and it is demonstrated how PPDN addresses a new problem specification. The framework prototype is being implemented in Perl and it will be made open-source.

Future Work. The future work includes beta testing with a group of collaborators and network of PPDN sites. A potential issue with a PPDN network is that if the nodes only periodically do forwarding and the time period is very long, or if the forwarding policy is moderating and the users do not attend their moderating duty frequently, then a significant delay in information dissemination could be experienced. This could be explored by running simulation experiments, which is a part of our future plans.

References

[1] The ACM topic hierarchy. WWW. Accessed Jul 2004 http://www.acm.org/class/1998.

[2] The semantic web research community ontology (SWRC). WWW, 2004. Accessed Jul 2004 http://ontobroker.semanticweb.org/ontos/swrc.html.

[3] P. Haase, J. Broekstra, M.Ehrig, M. Menken, P.Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, and C. Tempich. Bibster — A semantics-based bibliographic peer-to-peer system, November 2004.

[4] P. Haase, R. Siebes, and F. Harmelen. Peer selection in peer-to-peer networks with semantic topologies. In *International Conference on Semantics of a Networked World: Semantics for Grid Databases*, Paris, 2004. http://bibster.semanticweb.org/publications/haase_04_peer.pdf.

[5] V. Kešelj. Multi-agent systems for Internet information retrieval using natural language processing. Technical Report CS-98-24, University of Waterloo, 1998. ftp://cs-archive.uwaterloo.ca/cs-archive/CS-98-24/.

[6] R. Werlen. DFN science-to-science: Peer-to-peer scientific research. In *Proceedings of the Terena Networking Conference, TNC 2003*, Zagreb, Croatia, 2003.