PeerCrawl
A Decentralized Peer-to-Peer Architecture for Crawling the World Wide Web

VAIBHAV J. PADLIYA
MS CS Project (Fall 2005 – Spring 2006)

PeerCrawl is a distributed and decentralized P2P web crawler based on Gnutella protocol. The crawler is built upon Phex, open source P2P file sharing client, which provides the network layer. The crawler is separable and loosely coupled with Phex so that it can be ported to other P2P clients in the future. The two layers communicate via the interfaces provided by Phex. This prototype uses the Phex version 2.8.2.92 downloaded from http://sourceforge.net/projects/phex.

Source installation instructions
This package contains a folder PeerCrawl which includes the source code for the Phex client as well as the crawler. The crawler uses a library inetfactory.jar and is included in this package. The external libraries used by Phex client can be downloaded from http://svn.sourceforge.net/viewcvs.cgi/phex/.

Libraries to download:

- commons-httpclient-3.0.jar
- commons-logging.jar
- forms-1.0.6.jar
- jaxb.jar
- junit.jar
- log4j-1.2.12.jar
- looks-2.0.1.jar
- MRJ141Stubs.jar
- MRJAdapter.jar
- MRJToolkitStubs.jar

This prototype requires JAVA 1.4 or higher and was primarily developed using Eclipse in windows environment. When executed, an installation folder named PeerCrawl is created in the user home directory which contains the configuration files for Phex. In windows the "user home" is C:\Documents and Settings\user\". The crawler uses this folder for dumping the log files too.

CHANGES TO THE PHEX CLIENT

The following changes were made to the standard Phex client for implementation purposes.

File "PeerCrawl.hosts" replaces the file phex.hosts.

- Contains the IP addresses of other peers in the network that a peer can connect to on startup. The file acts as local host cache.

File “PeerCrawl_gwebcache.cfg” replaces the file gwebcache.cfg.

- Contains the URLs of web caches which hold the URLs of other web caches and IP addresses of peers in the network. This acts as the global web cache.

File phex.commons.Environment.java

- “*PRIVATE_NETWORK*” is set to “PeerCrawl”.
  (to form the overlay network for crawling)

- “*mListeningPort*” is set to 8000
  It’s the listening port for incoming connections. It can be set to any user defined value.

File phex.connection.ConnectionEngine.java

- “*handleQuery*” routine
  This routine handles the incoming query message broadcasted by a peer. The query message contains the URL as its payload. Hence to process the query message a call is made to the processQuery routine in phex.Peer.java instead of processing it according to the standard Gnutella protocol.

File phex.query.Search.java

- “*startSearching*” routine
  This routine is invoked by the crawler for broadcasting the URLs that the current peer is not responsible for.

File phex.common.HorizonTracker.java

- “*trackPong*” routine
  This routine computes the number of peers on horizon for the current peer. This count is used for dynamically re-computing the crawl range for the peer.