

# IR-Wire: Gnutella Data Crawler And Analyzer



Database and Information Retrieval Lab  
Illinois Institute of Technology  
10 W 31<sup>st</sup> St,  
Chicago, IL 60616

Website: <http://ir.iit.edu/~waigen/proj/pirs/>

Wai Gen Yee [waigen@ir.iit.edu](mailto:waigen@ir.iit.edu)  
Ophir Frieder [ophir@ir.iit.edu](mailto:ophir@ir.iit.edu)  
Linh Thai Nguyen [nguylin@iit.edu](mailto:nguylin@iit.edu)  
Dongmei Jia [jiadong@iit.edu](mailto:jiadong@iit.edu)

## USER MANUAL

### I. LimeWire Data Crawler:

Data crawler collects information from Gnutella networks and stores it in MySQL database. The information collected by IR-Wire includes:

- User queries
- Peer information
- Peer content information

IR-Wire data crawler is built on top of LimeWire client, thus it supports all functionalities of LimeWire. In addition, there is an additional menu added to LimeWire by IRWire for data crawling control. The IR-Wire menu includes the following items:

- Log Incoming Queries
- Crawl Shared Content
- Query the Network
- Collect Peer Information

Each menu item starts/stops a process when selected. The “Log Incoming Queries” menu item starts a process to collect user queries on the network. In order to collect as many user queries as possible, IR-Wire set all bits of its LimeWire’s query routing table to 1s before sending out to the super peer. This makes IR-Wire appeared to be a peer that can answer all queries, and thus all queries at the super peer will be forwarded to it. In addition, IR-Wire increases the number of super peer connections to 10 instead of 3, the default number in LimeWire, hence more queries from a broader range can be collected.

The “Crawl Shared Content” menu item starts crawling the shared content of peers available on the network, who support “Host Browsing” functionality. It uses LimeWire’s host browsing function to collect information about files shared on other

peers, which includes file name, file hash value and file's metadata. This function does not actually download any file from the network.

In order to crawl the content of a peer, IR-Wire needs its IP address. The "Query the Network" and "Collect Peer Information (from PONGs)" menu items are for this purpose. IP addresses of peers can be discovered in several ways. A peer may include its IP address in its outgoing queries so other peers can directly send answers to it, thus IP addresses can be extract from incoming queries. By sending out PINGs messages and getting back PONGs answers, a peer can also get IP addresses of other peers. IR-Wire uses both these methods to collect peers' addresses. However, these two methods cannot collect IP addresses of peers which are not included in PONG messages, or which do not include their IP addresses in their outgoing queries. To discover those peers, IR-Wire periodically sends out queries to the network and listens for the answers. IP addresses of peers who response can be extracted from their answers.

## **II. LimeWire Query Result Ranker:**

IR-Wire has the ability to rank query results using several Information Retrieval ranking functions. Three ranking functions were implemented in IR-Wire. *Term-frequency* ranks query results based on the number of terms, i.e. words, that overlap with query terms. The more query terms a result contains, the higher it is ranked. *Precision* is similar to *Term-frequency*, but uses the ratio of number of terms overlap and total number of terms in a result instead. *Cosine* ranks results based on their cosine-similarity with the query.

In the result panel, IR-Wire adds one column for each ranking function. The ranking score of each result is displayed, and selecting and ranking results using a ranking function is done by simply clicking on the appropriated column. Ranking functionality is useful, especially for queries which return too many results, or in the situation when there are many spam in the results. Other ranking functions can be easily added to IR-Wire.

## **III. LimeWire Data Analyzer:**

Data collected from Gnutella network are stored in a MySQL database and can be accessed by other programs. IR-Wire Data Analyzer is a tool that analyzes collected data. The implemented analyses include:

- General statistics
- Top ranked analysis
- Term correlation analysis
- Temporal analysis

IR-Wire requires some inputs before it can analyze data. First, time range of input data has to be specified. Second, parameters for term correlation analysis are required.

Data collected by IR-Wire are time-stamped, thus a subset of the dataset can be selected to analyze by specifying its time range. Queries and shared files whose time stamp fall in the range are analyzed.

Term correlation analysis is expensive. IR-Wire may use a lot of memory and takes a long time to accomplish if it has to analyze a large volume of data. To alleviate this, IR-Wire allows user to specify the smaller dataset used for term correlation analysis. In general, it includes specifying how many terms among the most frequent ones and how many data items (queries or files) among the most popular ones IR-Wire has to consider. Term correlation analysis is conducted based on this dataset only.

Results for each analysis are displayed in one panel. For example, “Statistics” panel displays statistics about the dataset, such as the type distribution and the length distribution of queries and shared files. “Term correlation” panel displays the highly correlated pairs of terms, based on their Pearson coefficient, together with their Chi-squared values. Results displayed in each panel are self-explanatory. For temporal analysis, the dataset is divided into a number of subsets, and statistics of each subset is compared with the first subset.

IR-Wire automatically saves the analysis results to files on disk so that you can load and redisplay the results in the future without having to rerun the analyses. In addition, results for different categories can be displayed separately. The file name format is “IRWireFyyyymmdd\_yyyyymmdd.irw” for results of shared file analysis and “IRWireQyyyymmdd\_yyyyymmdd.irw” for results of query analysis, where “yyyymmdd\_yyyyymmdd” is the date input range.

## Reference

1. Shefali Sharma, Linh Thai Nguyen, Dongmei Jia. IR-Wire: A Research Tool for P2P Information Retrieval. In Proceeding of SIGIR Open Source Information Retrieval Workshop, Seattle, USA, 2006.
2. Linh Thai Nguyen, Wai Gen Yee, Dongmei Jia, Ophir Frieder. A Tool for Information Retrieval Research in Peer-to-Peer File Sharing Systems. Demo paper. In Proceeding of International Conference on Data Engineering (ICDE'07), Istanbul, Turkey, 2007.
3. Linh Thai Nguyen, Dongmei Jia, Wai Gen Yee, Ophir Frieder. Analysis of Query Logs in Gnutella Peer-to-Peer Network. Accepted poster paper. Conference on Research and Development on Information Retrieval (SIGIR'07), Amsterdam, Netherlands, 2007.