

# Performance Evaluation of Peer-to-Peer Information Retrieval Systems

Jérôme SICARD  
TÉLÉCOM & Management Sudparis  
9 rue Charles Fourier  
91011 Évry Cedex

Bruno DEFUDE  
TÉLÉCOM & Management Sudparis  
9 rue Charles Fourier  
91011 Évry Cedex

**Abstract**—This article describes a methodology for performance evaluation of Peer-to-Peer Information Retrieval (P2P-IR) systems from the family of Gnutella. We break down Gnutella into a set of four mechanisms. Then we evaluate each mechanism in turn, independently. For a reason of simplicity, we will only study one of them - selection mechanism - in this article. To that end, we describe it in terms of *Parameters, Constraints, Cost indicators and Quality indicators*. We don't use simulation, but compute the probability of quality of this mechanism. We believe that this methodology allows a better understanding of P2P-IR systems. It also permits an easier performance analysis and reuse of results, since we analyse the systems component by component. Finally, it opens a new way to an easier comparison of different P2P-IR systems.

## I. SCOPE OF THE STUDY : Gnutella-LIKE SYSTEMS

Let us first define the targeted systems. The definition of Gnutella-like systems might be much like the definition of "pure P2P" systems. But we prefer to use a different vocabulary, to be free to give our own, precise definition. We first need to describe Gnutella, then we can define Gnutella-like systems.

### A. Description of Gnutella

We break down Gnutella into a set of four mechanisms:

1) *Overlay management*: Build and maintain a random overlay of degree 20, thanks to three mechanisms :

a) *Entrance*: Get a contact peer into the Gnutella network.

b) *Join*: Initialize neighborhood.

c) *Maintenace*: Maintain the neighborhood, by replacing failing neighbors.

All these mechanisms are implemented at the network level by querying for peers with IP broadcast queries.

2) *Querying mechanism*: Transmit a query to peers for local evaluation. As suggested in [1], we outline two steps:

a) *Selection*: Select a random set of peers that will evaluate the query. This is done at an overlay network level, with partial broadcast. It defines the sub-corpus the query will be evaluated on.

b) *Query resolution*: Transmit queries to selected peers. This step is implicit in Gnutella, when joining the resolution query to selection queries.

3) *Local Evaluation*: Not specified. We won't include it in further discussions.

4) *Answer construction*: The answer is built locally by the query initiator, from partial answers returned by queried peers. Partial answers are collected by backward routing.

### B. Gnutella-like systems

We define Gnutella-like systems as any system which selection mechanism is based on randomness. This implies that selection is done thanks to a random propagation of queries.

## II. FIRST ANALYSIS

We analyse each component in terms of constraints, parameters, quality indicators and cost indicators. They can be represented in a graph (see graph 1). Note that quality indicators for some mechanisms can be interpreted as constraints for others (see edges in graph 1).

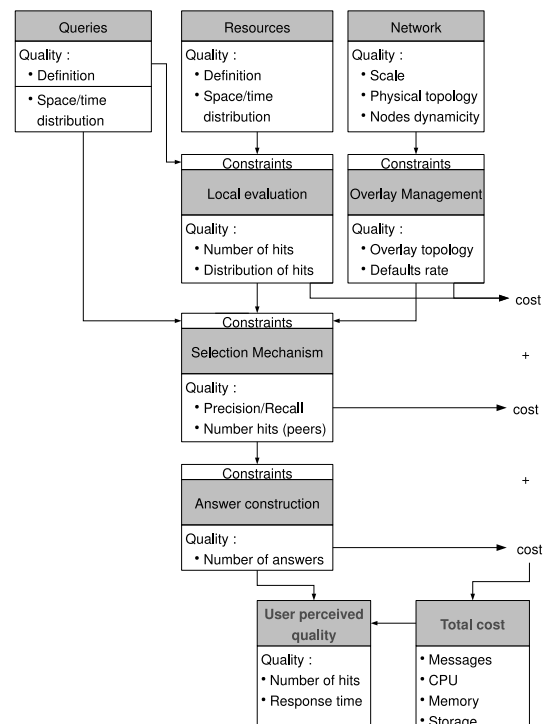


Fig. 1. Graph of gnutella-like systems

Here is the set of variables for selection mechanism:

1) *Parameters:*

- **W** width of the propagation tree
- **TTL** depth of the propagation tree

2) *Constraints:*

a) *from overlay management:*

- **Scale** Number of peers in the system.
- **Overlay topology** Thus we can measure the impact of overlay topology on querying mechanism.
- **Peers dynamicity** Involves losses of neighbors.

b) *from local evaluation:*

- **Percentage of hits** Percentage of nodes that can answer the query. Depends on the level of data replication.
- **Distribution of hits** Non-uniform distribution of hits might involve non-uniform selections.

c) *from queries:*

- **Queries space/time distribution**

3) *Quality indicators:*

- **Number of hits** We define the quality of a selection as the number of hits in the corresponding corpus.
- **Computation time** Good indicator of the effect of selection mechanism on overall computation time.

4) *Cost indicators:*

- **Number of messages**
- **Memory**
- **CPU**
- **Storage**

### III. PERFORMANCE EVALUATION OF SELECTION MECHANISM

Let us explain our methodology on selection mechanism. To get more precise results, **we don't simulate**. We prefer to **run each possibility** : for a fixed  $W$  and  $TTL$ , we compute every possible propagation, from each peer of the network. This gives the exact **probability of quality** of the algorithm, given a set of constraints.

### IV. REDUCING THE SET OF VARIABLES

- **Assumption 1** : *computation time*  $\equiv$  *depth of the tree*
- **Theorem 1 (peers dynamicity)** Let  $x$  be the rate of defaults in routing tables,  $w$  the width of the propagation algorithm and  $RT$  be the size of routing tables. The average actual width of propagation trees  $W$  is defined as :

$$W = \sum_{i=0}^{xRT} (w - i) \frac{C_{xRT}^i P_w^i P_{RT(1-x)}^{w-i}}{P_{RT}^w}$$

- **Assumption 2** : *queries space/time distribution*  $\equiv$  *hits distribution*

Remark that the definition of the topology might not correspond to only one variable, but to a set of variables.

[2] make the assumption that the purpose of random propagation algorithms is to make a random selection of peers. We want to compare the quality of a selection from a random propagation and a random selection - picking up

peers randomly from the whole network. To that end, we also compute the size of the selections - number of peers selected.

Finally, our set of variables becomes :

a) *Parameters:* (W, TTL)

b) *Constraints:* (Scale, Overlay topology, Percentage of hits, Distribution of hits)

c) *Quality/cost indicators:* (Number of hits, Number of messages, Number of peers contacted)

### V. METHODOLOGY

Finally, we can define a performance evaluation methodology for the selection mechanism of Gnutella-like systems

```
for each scale
  for each overlay topology
    for each percentage of hits
      for each distribution of hits
```

```
    for each width
      for each depth
```

```
        for each possible selection, compute
          number of peers selected
          number of messages
          number of hits
```

This evaluation allows to choose the better set of parameters (*width, depth*), given a characterization of constraints and the quality one wants to reach, e.g maximize the number of hits or minimize the number of messages.

We programmed a test module under PeerSim. PeerSim is a P2P simulator, but we didn't use it for simulation (we don't simulate here) : we chose it because it provides interesting modules for overlay construction and management.

The last problem we have to solve is the complexity of this evaluation. Let  $R_t$  be the size of routing tables (characterizes overlay topology),  $W$  be the width of propagation tree and  $TTL$  its depth. The maximum number of possible propagation trees is :

$$\left( \frac{R_t!}{(R_t - W)!} \right)^{\sum_{i=1}^{TTL} i}$$

### VI. CONCLUSION

To conclude, this study proposes a methodology for performance evaluation of selection mechanism in Gnutella-like systems. To that end, we broke down Gnutella into a set of four mechanisms and analysed interdependencies between them. So we could abstract the quality of overlay management and local evaluation and the querying scheme as constraints on selection mechanism. We proceeded to a reduction of the set of variables. Finally, we proposed an algorithm for performance evaluation of selection mechanism.

### REFERENCES

- [1] M. del Pilar Villamil Giraldo, "Service de localisation de données pour les systèmes p2p," PhD Thesis, Institut National Polytechnique de Grenoble, June 2006.
- [2] C. Gkantsidis, M. Mihail, and A. Saberi, "Random walks in peer-to-peer networks," 2004.