Trustworthy Information Distribution and Retrieval

Isaí Michel Lombera, Yung-Ting Chuang, P. M. Melliar-Smith, L. E. Moser

Department of Electrical & Computer Engineering, University of California, Santa Barbara, CA 93106-9560

Introduction

- Modern society and commerce depend on access to information over the Internet.
- Information is accessed over the Internet using centralized search engines and search indexes, for efficiency and scalability.
- We cannot assume that centralized search engines will always deliver the information we seek, uncensored and unbiased.
- iTrust is a system for publishing, searching for, and retrieving information over the Internet that provides trustworthy access to information.

Overview of iTrust

Main objectives are to:

- Provide users with information they need.
- Publish information for other people to access.
- Search for, and retrieve, published information.
- Detect that the system is under attack.
- Increase the probability of distribution and retrieval when the system is under attack.

Basic elements of iTrust:

- P*articipating nodes:* Form the *membership* of iTrust.
- *Source node:* A participating node that produces metadata describing the information, which it makes available to other participating nodes chosen at random.
- R*equesting node:* A participating node that generates requests and distributes the requests to a subset of participating nodes chosen at random.

How iTrust works:

- When a participating node receives a request,
  first it compares the metadata in the request with the metadata it holds
  if it finds a match, it returns the URL of the associated information to the requesting node.
- The requesting node then uses the URL to retrieve the information from the source node.

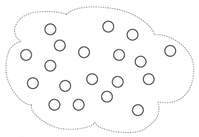


Figure 1. A network with participating nodes.

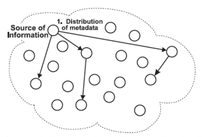


Figure 2. A node (a source node) distributes metadata, describing its information, to randomly selected nodes in the network.

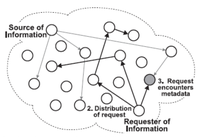


Figure 3. A node (a requesting node) distributes its request to randomly selected nodes in the network. One of the nodes has both the metadata and the request and, thus, an encounter occurs.

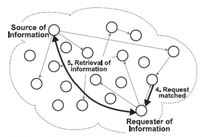
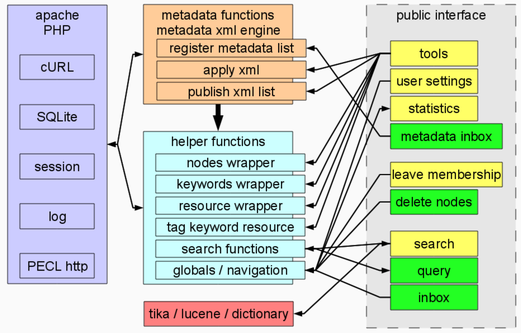


Figure 4. The node matches the metadata and the request and reports the match to the requester, which then retrieves the information from the source.

Implementation of iTrust

The iTrust implementation is based on HTTP.

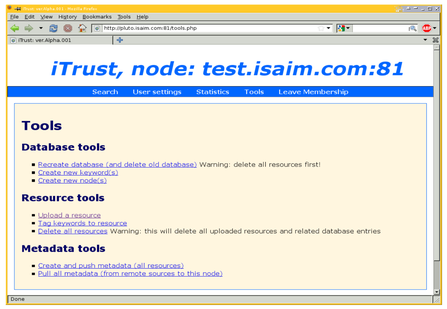


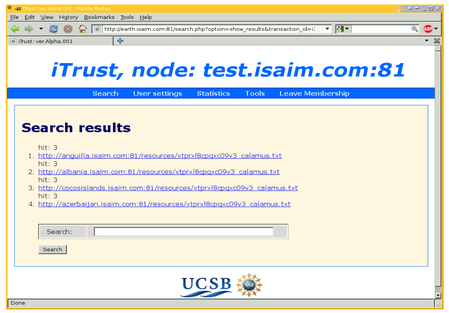
(a) Web server foundation: Uses Apache Web server with PHP extensions. cURL for inter-node communication and SQLite for node / keyword / resource tracking.

(b) Application infrastructure: XML component generates resource lists; helper functions manage node and resource information; uses Tika / Lucene / WordNet JARs to add tagging and synonyms.

(c) Public interface. Machine interface (green) used for inter-node communication (membership, query, etc). Human interface (yellow) used for node searching, administration, and uploading resources.

User Interface





Probabilistic Analysis

Variables:

- Membership contains n participating nodes.
- Metadata are distributed to m nodes.
- Requests are distributed to r nodes.
- Proportion x of the participating nodes are operational.

Formulas:

- Probability q of no match on r trials:

$$q = \frac{n-mx}{n}\frac{n-mx-1}{n-1}\frac{n-mx-2}{n-2}\cdots\frac{n-mx-r+1}{n-r+1}$$

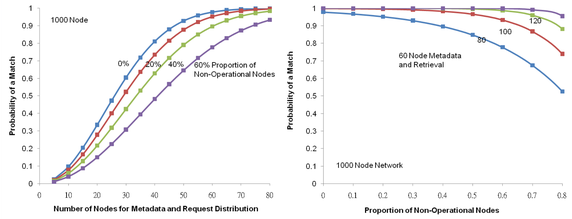
- Probability p of a match on r trials:

$$p = 1-q = 1-\frac{n-mx}{n}\frac{n-mx-1}{n-1}\frac{n-mx-2}{n-2}\cdots\frac{n-mx-r+1}{n-r+1}$$

for $n \geq mx + r$.

If $m = r = \lceil\sqrt{n}\rceil$ and $x = 1$, then $p > 0.6321$.
If $m = r = \lceil\sqrt{2n}\rceil$ and $x = 1$, then $p > 0.8647$.
If $m = r = \lceil 2\sqrt{n}\rceil$ and $x = 1$, then $p > 0.9817$.

- Probability of a match:



Note: If $mx + r > n$ then $p = 1$
In this case, the subset of nodes to which the metadata are distributed and the subset of nodes to which the requests are distributed intersect in at least one node.

Simulation of iTrust

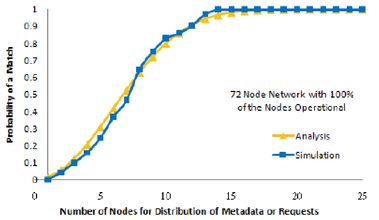
The iTrust simulation is based on the implementation of iTrust and collects the match probabilities using libCURL.

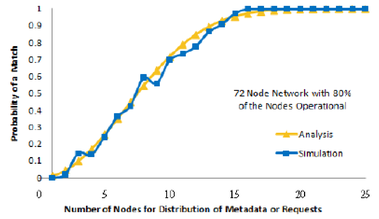
Simulation steps:

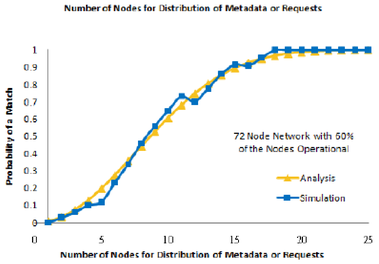
1) First we delete all resources and data from the SQLite databases.
2) The program then adds all nodes to the membership.
3) Once all the nodes are added to the membership, we supply the number of nodes for distribution of metadata and requests, and the proportion of operational nodes, to the simulation program.
4) Next, we call the source node to upload a file and the program then creates the corresponding metadata.
5) Then, the program randomly selects the nodes for metadata distribution and distributes the metadata to those nodes.
6) Next, the program randomly selects the nodes for the requests and distributes the requests. If one or more nodes returns a response, there is a match and the simulation program returns 1; otherwise, there is no match and the simulation program returns 0.

Simulation Results

Match probability versus number of nodes for distribution of metadata and requests in a network with 72 nodes where 100%, 80%, and 60% of the nodes are operational:







Conclusions and Future Work

- iTrust is a trustworthy information distribution and retrieval network with no centralized mechanisms and no centralized control.
- iTrust involves distribution of metadata and requests, matching of requests and metadata, and retrieval of information corresponding to the metadata.
- With iTrust, the probability of matching a query is high for 100%, 80%, and 60% operational nodes, given a reasonable number of participating nodes.
- We plan to do experimental evaluations of the prototype implementation using PlanetLab.
- We are investigating other implementations of iTrust based on SMS, Twitter, etc.
- We plan to make the iTrust source code, tools, documentation, *etc.* freely available for all to use.

Acknowledgments

This research is supported in part by NSF CNS 10-16193.
.