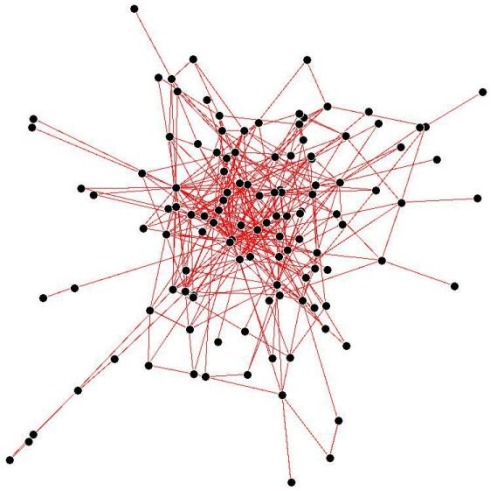
UNIVERSITY OF JYVÄSKYLÄ



Resource Discovery in A Social Network

Presentation for P2P content and distribution seminar of ICTSHOK Future Internet 18.6.2010

Mikko Vapa, mikko.vapa@jyu.fi

P2P Research Group

Department of Mathematical Information Technology

University of Jyväskylä

research.jyu.fi/p2pgroup

Background

- Mobile devices can now run webservers
- However, the content inside the mobile devices is currently not accessible from outside (and cannot be indexed by Google)
- The content needs to be made available and flexibly searchable from outside depending on the access rights of the searcher
- Lots of research has been done how to index content and make it searchable on servers
- But research how to search real-time from multiple servers organized into a social network is still very young
- Our work advances social network search area and possibly generates a global scale search service

Results

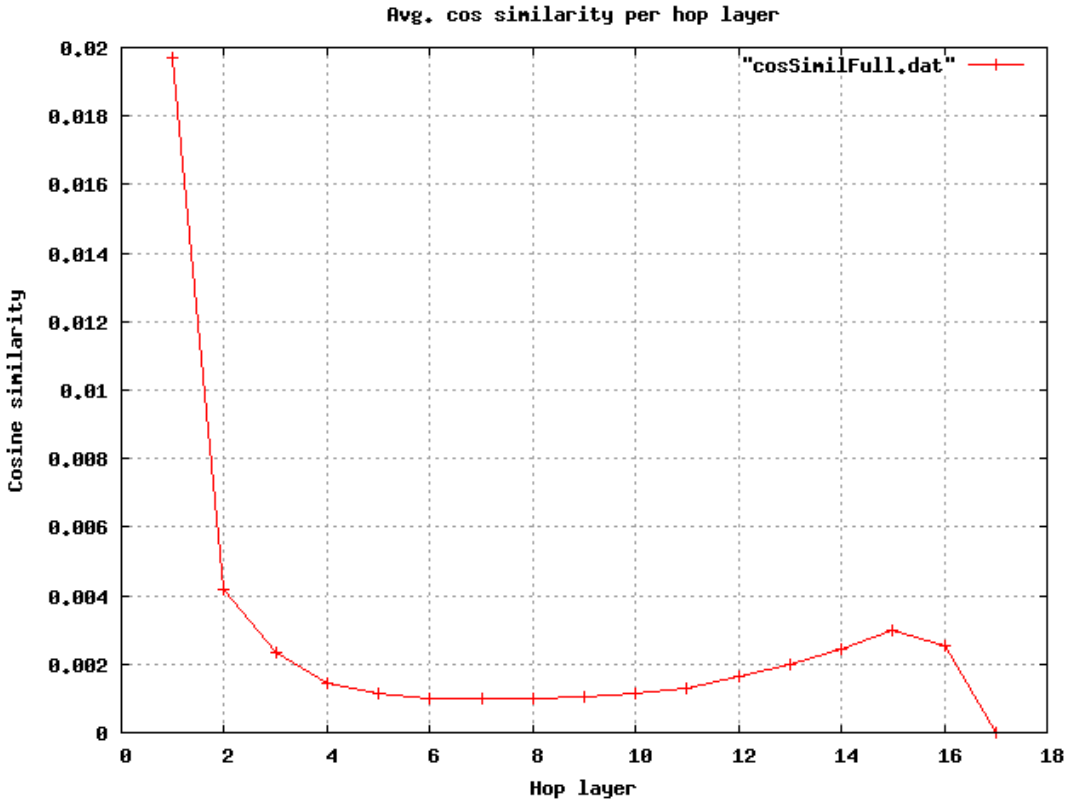
- Last.fm social network site has been crawled with data of 177000 social network users in 2008
  - Last.fm social network site was selected as a source of social network simulation data to get information how a music community is organized
  - The crawl includes the top songs the users are listening, the associated albums and social connections between users
- Topology awareness algorithm has been specified and implemented to P2PRealm network simulator in 2009
  - P2PRealm was chosen as a simulation platform since we are not interested in simulating the network behavior (i.e. connection delays, packet loss, data corruption...), but we just need to emulate data connection between the peers
  - Low-level simulation is also a significant performance bottleneck when simulating large graphs

Results

- Features of future information network has been briefly sketched in 2009-2010
- Similarity graphs of Last.fm have been obtained in 2010
  - Hops-to-Similarity diagram shows how similar two nodes are based on their distance on a social network
  - Similarity was measured using cosine and cardinality similarity measures
- Next step is to implement a social network search algorithm
  - Provides the first simulation results on Last.fm data

Similarity of Nodes in A Social Network

- Last.fm contains homophily properties i.e. neighbors are more similar than others



Similarity of Nodes in A Social Network

- Because neighbors are more similar than others it seems that social network organizes data and thus makes it searchable
  - It is assumed that locating just one node with a matching resources will make it easier to find other nodes with matching resources
    - A social network search algorithm can concentrate on that part of the social network where resources have been found
- Last.fm social network is sparse and therefore similarity properties could be stronger in more developed social network sites like Facebook

Social Network Search Software

- Contents of a mobile device needs to be searchable depending on the access rights of the person issuing the search
- With a single click the user should be able to get new popular and socially relevant content
  - Searching is a feature that should be actively done only by certain users and providing relevant information automatically to other inactive searchers would increase the speed of information flow significantly
- Additionally the user might type keywords and other criterias to tailor the search
- Consists of a mobile/non-mobile search engine and social network search components

Resource Discovery Algorithm

- Search space consists of thousands of users
- Resource discovery algorithm needs to be developed to find new and socially relevant data from the mobile users
- As a benchmark algorithm for social network search Expanding Ring (Lv et al. 2002) can be used
- Implementation and evaluation on P2PRealm simulator and further implementation to a social network search prototype

Possible Search Cases for Social Search

1. ”What should I know at the moment?” searching or popular URLs within social neighborhood searching or alert of interesting new information (essentially these are same things)
2. Multiple keyword searching
3. Friend, hobby group or work group location searching or general location restricted searching
4. Searching for deep expert information which might not be found from closeby social groups

A Search Scenario to Simulate

- ~170000 nodes (or users) have been crawled from Last.fm
- Each node contains a set of keywords defined as resources
- The keywords consists of music songs and the associated band and album
- Each node/user has connections to its neighbors
- A randomly selected node starts a query
- The query keyword is selected from the node's set of keywords to ensure that nodes with somewhat similar content should be found (this utilizes homophily properties of Last.fm data)
- There might be multiple keywords in a single query, so selecting two keywords is also worth simulating

Search Parameters

- Nodes will be ranked for querying based on different parameters
- Similarity might be utilized via such kind of parameters:
  - RepliesOnNeighbors (0,1,2,...) is the number of queried node's neighboring nodes which have returned replies to this query (nearby replying nodes can be an indication that this part of social network has matching replies)
  - RepliesOnPath (0,1,2,...) is the number of nodes which have returned replies to this query on the shortest path between querier and the queried node in the social network (replying nodes on the shortest path can be an indication that this part of social network has matching replies)
  - Reply% ([0,1]) is the amount of queries the node has replied divided by the total number of queries queried from the node (nodes which replied earlier to queries might be more likely to answer in the future)

Result Ranking

- The list of results obtained from a social network can be large
- Ranking mechanism needs to be developed to sort the relevant results on top of the list
- As a benchmark algorithm Pedro Tiago's naive result interleaving and Page Rank (Brin and Page 1998) could be used
- Implementation and evaluation on P2PRealm simulator and further implementation to social network prototype
- Also social network user behavior can be monitored and collected to centralized P2PStudio server
  - Data analysis methods could be used to find out which attributes characterize the most clicked information and this might affect result ranking algorithm development

Topology Management

- Social network evolves when mobile users add and remove social network links from their addressbook or when users add and remove friends in Facebook etc. services
- Search algorithm should have access to latest social network neighborhood topology information
- Topology management algorithm needs to be developed to have an up-to-date view on the social network
- As benchmark algorithms Pedro Tiago's naive topology querying algorithm can be used and also routing indices (Crespo and Garcia-Molina, 2002)
- Implementation and evaluation on P2PRealm simulator and further implementation to a social network prototype

Topology Awareness Algorithm

- To study how different search algorithms work on a community of music listeners we need a distributed algorithm for discovering local social network topology of each user
- Topology awareness algorithm is in charge of building the social topology of the network and is a sub algorithm for topology management
  - Such an algorithm has now been specified and implemented to P2PRealm
  - Allows studying algorithms in a static network, but needs to be further developed to support dynamic network conditions

Topology Awareness Algorithm

- Every peer node keeps information about the other peers it knows (from address book or from earlier replies got for queries)
- Nodes are saved in two lists: node list and topology list
- Node list is a general list of nodes' addresses/identifiers of the nodes
- The topology list has a fixed size (which can be varied in simulations to find out the suitable size) and it is tried to keep full. The topology list contains those nodes which has some calculated parameters, like social distance
- The topology list is used when node needs to search resources

Topology Awareness Algorithm

- First the known nodes (e.g. contacts in the address book) are copied into the topology list
  - If there are more contacts than space in the topology list, the rest are saved into the node list
  - If there are too few contacts, the querier tries to discover more nodes in the network until the topology list is full
- Nodes in the topology list are ordered by their social distance
- Social distance can be defined as number of mobile phone calls, real-life social distance (hops) or LinkedHops
- In the first version of simulation LinkedHops is used: LinkedHops = hops-1+(1.0/links), where hops is the distance between nodes and links is the number of different shortest paths between nodes + (number of clustered neighbors for evaluated node-1)*0.25

Architectural Choices

- Prototype for mobile and non-mobile social network search could be implemented on Personal-Apache-MySQL-PHP (PAMP) webserver stack running on Symbian phones and on desktop computers as xAMP
- With the prototype it would be possible to demonstrate real-time searching of relevant new content from mobile phones or computers
- Social network data could be obtained from Facebook and Facebook users could install software enabling correct IP address and port to be contacted for queries
- Desktop computers could install Google Desktop Search (or similar indexing system) for indexing the contents of the computer and a piece of software enabling outside queries with access rights

Plans on Autumn 2010

- Simulation of resource discovery algorithms in mobile social P2P network scenario based on Last.fm crawl
  - NeuroSearch framework provides variety of test cases and the purpose is to find a set of input parameters useful for searching in a social network
  - Several publications will be written on the search scenario
- The social P2P network prototype will be implemented if further funding of the project can be obtained after Autumn 2010
- The expected outcomes in the future consists of a social network search prototype, algorithms for topology management, resource discovery and result ranking for a social network, live demonstration of social network search with P2PStudio monitoring tool, user behavior analysis of social network search and several publications on algorithms and the prototype