A P2P REcommender system based on Gossip Overlays (PREGO)

Ranieri Baraglia, Patrizio Dazzi, Matteo Mordacchini
ISTI, CNR, Pisa, Italy

Laura Ricci
University of Pisa, Italy
PRESENTATION OUTLINE

• Recommender Systems: general characteristics
• Our Approach: PREGO
• Gossip Protocols: basic issues
• PREGO
  – User profile definition
  – Similarity metrics
  – Architecture of the P2P overlays
  – Recommendations spreading
• Experimental Results
• Conclusions
RECOMMENDER SYSTEMS: CHARACTERISTICS

• Recommender systems (recommendation systems):
  – information filtering systems attempting to recommend information items that are likely to be of interest to the user
  – compare the user profile to some reference characteristics to predict the 'rating' that a user would give to an item he does not know yet

• Content Based Approaches
  – The user profile is compared with some characteristics of the item

• Collaborative Filtering (Social Information Filtering)
  – Exploits the social environment of the user
  – Automates the 'word-of-mouth' process: items are recommended to a user by other people with similar tastes
RECOMMENDER SYSTEMS: PREGO

• A P2P system exploiting the collaborative exchange of information between users in order to build a system able to
  - cluster users with similar interests
  - spread useful suggestions among them

• PREGO distinguishing features:
  - builds a user profile exploiting content accessed from the user (purchase items, visited pages, ...). Based on the fact that a user will show interest in the future for content accessed in the past
  - exploits a set of gossip protocols to cluster users with similar interests
  - defines a distributed recommendation service
    - no centralized authority storing users profiles and ratings: preserve user privacy
    - no centralized/controlled recommendations: a 'democratic' suggestion system
GOSSIP PROTOCOLS: CENTRALIZED APPROACHES

- **Gossip (epidemic) protocols**: originally introduced in 1987 for propagating updates in loosely replicated databases to maintain replica consistency.

- The original gossip protocol: each node
  - has a complete view of the nodes of the network
  - periodically picks a random node $N$ from the view and exchange some data with $N$

- **Main characteristics of the approach**
  - information known by any one node of the network is spread to all other nodes with very high probability.
  - requires a complete knowledge of the network at each node and this is infeasible for large scale, highly dynamic networks, like P2P systems.
A P2P RANDOM SAMPLING SERVICE

- A P2P gossip protocol requires the definition of a distributed service that returns to each peer a random sample of the nodes.

- A distributed Random Sample Service implemented through a gossip approach: each peer P
  - knows a small, continuously changing set of other peers (its neighbours) defining its view
  - periodically chooses a random node Q from its view
  - P and Q exchange a set of neighbours randomly picked from their views.

- The continuous mixing of neighbours generates random overlays.

- The random sampling service is exploited by higher level gossiping protocols to epidemically spread information over the network.
A P2P RANDOM SAMPLING SERVICE

• When a peer P (e.g., 2) chooses a peer Q (e.g., 9) to gossip, the neighbour relation between P and Q is reversed.
• P sends its descriptor to Q, and deletes the descriptor of Q from its view.

Before swapping:
- 2 → 9 : {2, 0, 6}
- 2 ← 9 : {0, 5, 7}

After swapping:
CYCLON: AN ENHANCED P2P SAMPLING SERVICE

- **Cyclon [Voulgaris 2006]:** A P2P random sampling service improving the quality of the overlay in terms of randomness

- View = set of node descriptors including
  - the contact information of a peer
  - a numeric age field: number of gossip cycles since the moment the descriptor was created (descriptor age)

- Enhanced protocol: A peer P
  - select the neighbour Q with the highest age
  - sends its 'fresh descriptor' to Q, i.e. a descriptor of age 0

- With respect to the basic gossip protocol, Cyclon
  - prevents links to dead nodes from lingering indefinitely
  - imposes a predictable lifetime to each descriptor
    - this implies an even spreading of links across the overlay
OUR PROPOSAL

- Extract a profile for each peer according to the past interests of the user. The profile may be constructed by considering:
  - resources recently accessed
  - items recently purchased
  - pages visited
- Exploit a gossip protocol to define clusters of peers characterized by similar interests:
  - definition of a proper similarity metrics
  - exploits the distributed sampling service defined by Cyclon
- Spread recommendations of new acquired content to proper cluster of peers
A peer may be characterized by a set of interests
An overlay for each interest (sport, music, films ..)

within a single overlay clusters correspond to highly similar peers, for instance:
• an overlay for peers interested in films
• different clusters for different film genres (action, westerns,..)
THE DEFINITION OF THE USER PROFILE

- Let $\mathcal{I}$ be the set of items of the system and $\mathcal{I}_p \subseteq \mathcal{I}$ the subset of items belonging to a peer $p$

- The profile $\pi_p$ is defined as follows:
  $$\pi_p = \{ (i, C(i), R(i)) \mid i \in \mathcal{I}_p \}$$
  - $C(i)$ is the content associated with the item $i$
  - $R(i)$ is the rating given by $p$ to $i$

- The set of items $\mathcal{I}_p$ can be partitioned according to the set of interests of the peer $p$
  - if the peer $p$ has a set $\mathcal{I}_p^p = \{I_{p,1}, \ldots, I_{p,k}\}$ of interests, $\pi_p(I_{p,j})$ is the set of items associated with the interest $I_{p,j}$

- Different overlays associated to different interests
THE DEFINITION OF A SIMILARITY METRICS

- Jaccard similarity. Given a pair of peer $p_1$ and $p_2$ and a pair of interests

$$\text{sim}(p_1, p_2) = \frac{|\pi_{p_1}(I_{s}^{p_1}) \cap \pi_{p_2}(I_{t}^{p_2})|}{|\pi_{p_1}(I_{s}^{p_1}) \cup \pi_{p_2}(I_{t}^{p_2})|}$$

The vector represents which URLs are shared by $\alpha_i$ & $\beta_i$ and their frequencies.
THE GOSSIP PROTOCOLS

- Each interest based P2P overlay is constructed through a gossip protocol
- A two-layered gossip framework

```
+----------------+       +----------------+
|                |       |                |
|       PREGO     |       |    PEER SAMPLING |
|                |       |                |
```

- The Sampling Layer is responsible for feeding the PREGO-layer with nodes uniformly randomly selected from the whole overlay
- The PREGO Layer in charge of discovering peers characterized by similar interests
A peer $P$ gossip active process for each interest

Let $N(P)$ be the set of $P$'s actual neighbors

for all $P_i \in N(P)$ do

Get from $P_i$ a set $NewPeers$ from its neighborhood

for all $P' \in NewPeers$ do

if $P' \notin N(P)$ then

connect with $P'$

if $Sim(P, P') \geq \min_{P_j \in N(P)} Sim(P, P_j)$ then

add $P'$ to $N(P)$

end if

end if

end for

end for
A P2P REcommender system based on Gossip Overlays (PREGO)

R.Baraglia, P.Dazzi
M.Mordacchini, L.Ricci
SPREADING RECOMMENDATIONS

• When an overlay is stabilized, a peer can consider its neighbours as the representatives of a community of 'friends' with similar interests.

• When a peer p of the community becomes acquainted with a new item h:
  – p discovers the interest group IG corresponding to h
  – p forwards a recommendation for h on the overlay corresponding to IG
  – p exploits the similarity function to evaluate the interest of its neighbours for the item h

• Recommendation is forwarded only iff the similarity between h and the neighbour is above a given threshold.
  – For instance, a new film is recommended to a neighbour iff the intersection among the genres of the film and the profile of the neighbour is above a given threshold.
EXPERIMENTAL ENVIRONMENT

• A prototype implemented by the Overlay Weaver framework
  – rapid prototyping
  – high scalability
  – experiments replication
  – possibility of monitoring the status of the network

• Experiments exploit the Movie-lens Data Set
  – A DataBase of movies used for user recommendations
  – Includes 1 millions ratings for 3900 movies by 6040 users
  – Each user rates a few movies
    • a very sparse matrix of preferences
    • a complex scenario to build communities
  – User profile: movies seen by the corresponding user
GOSSIP PROTOCOLS COMPARISON

• Vicinity, TwinFinder two similar gossip algorithm based on interest similarity
• Reports average similarity (Jaccard) between a peer and its neighbourhood
GOSSIP PROTOCOLS COMPARISON

Twinfinder obtains the smallest deviation, i.e. a more uniform overlay with respect to similarity.
• Goal of the test: measuring the ability of gossip protocols to provide recommendations to interested communities

Coverage: percentage of film genres which may be recommended to a user by its neighbours
Coverage Evaluation

- Coverage measure for different network sizes
- Best result for networks including high number of nodes
CONCLUSIONS

• Definition of a gossip algorithm for the clustering of peers characterized by similar interests
• Creation of communities characterized by similar interests
• Spreading of recommendations
• Future work:
  – definition of a distributed algorithm to detect a profile characterizing a community of peers
    • distributed voting algorithms to detect the 'most representative peer' of a community
  – investigating the structured overlays (DHT) approach to register community profiles
  – LSH (Locality Sensitive Hashing) to support the search of community with profiles similar to that of a given user