Automated Communities in Social Networks Using Kohonen SOM

By

Dinesh Gadge

Parthasarathi Roy

Motivation

- Virtual World

- Many social networks : Orkut, Gazzag, Linked In, Multiply, Facebook, MySpace

- Finding "like-minded" people

State of the art

- Social Network Analysis : Communities
- Kohonen SOM : Clustering
- Weblog Mapping

Social Network Analysis

- Case Study : Orkut
- Interests, Activities, Sports, Music, Movies
- Communities
- "Like-minded"

Orkut Snapshot

passions:	rock music, dancing, friends, fast bikes ...
sports:	cricket, soccer, badminton, F1 ..
activities:	hmm .. partying, dancing, sleeping and occasionally hiking ..
books:	no time .. the closest I get to books is newspapers
music:	Green day, Metallica, Oasis, Limbizkit, Nirvana, GooGoo Dolls, Cranberries, Def Leppard .. long list
tv shows:	Friends, Prison Break, Grey's Anatomy, Scrubs, The 70's show, One Tree Hill ...
movies:	Motorcycle Diaries .. recently
cuisines:	Punjabi, chinese, thai, italian, mexican, portugese, carribean, french ... out of the ones ive treid

- Source : http://www.orkut.com/Profile.aspx?uid=17785808993583780837

Kohonen SOM

- **Clustering**

- **Winner : neuron with minimum distance**

- **Update rule :**

  - Online : $w_i(t+1) = w_i + h_{ck}[x(t) - w_i(t)]$

  - Batch : $w_k = \dfrac{\sum_{t'=t_0}^{t_f} h'_{ck}(t')x(t')}{\sum_{t'=t_0}^{t_f} h'_{ck}(t')}$

  - Neighbourhood : $h_{ck} = \alpha(t) \times exp(-||r_c - r_k||^2/(2\sigma^2 \times (t)))$

Main Results

- Kohonen SOM : effective method for clustering this type of data (?)
- Challenges : Data Collection and Standardization.

Challenge : Data Collection

- Need for customized Web-Crawler : Orkut pages are session-managed, so some approach is required to maintain sessions while crawling Orkut to collect data.

- Where should the data be collected from ?
  - Network of friends
  - Existing communities

Challenge : Data Standardization

- Data needs to be structured : Initially the data in terms of interests would tend to be very sparse.

- Ideas : Use tuples. Restrain the number of parameters. Apply "genres" to movies. Ignore semantic-analysis.

- <Profile ID, Movie-related items, Music related items>

- It needs to be seen what kind of attributes can be given in Movie-related items and Music related items so that good results are obtained from Kohonen SOM.

Challenge : Distance function

- Use Euclidean distance.
- But standardize data accordingly so that this distance can be used.
- This would require numerical data to be stored in the tuples.
- So tuples can contain `count' of movies, music, tv shows etc. of different kinds.

Another Tangential Application

- Matrimonial and Dating websites
- Train Kohonen SOM on "features" of individuals e.g. age, height, education etc.
- Test using a query for "ideal-match."
- Kohonen SOM should give a cluster of "best-matches"

Use of Kohonen SOM in SNA

- Visualization

- Clustering as a means to find communities / like-minded people

Visualization

- **Humans cannot visualize high dimensional data**
  - Eg. 10 dimensional data
- **Technique needed to understand high dimensional data**
- **Kohonen SOM is one such technique**

Visualization

- Kohonen SOM produces **map** of high dimensional data to 2 dimensions

- This 2-D map is useful for seeing features of higher dimensional data
    - Eg. Cluster tendencies of data

- Topology of higher dimensional data preserved in 2-D map

Visualization
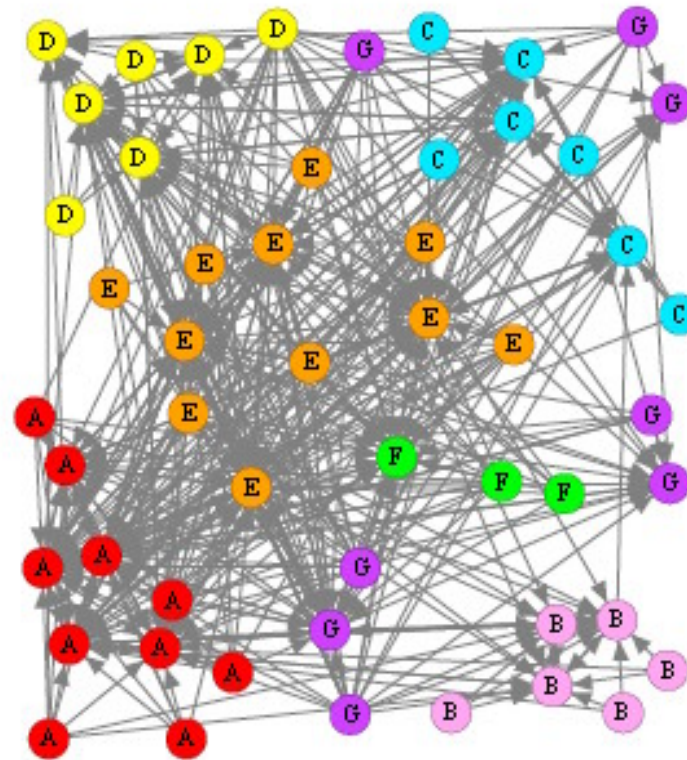
- High dimensional data mapped to 2 dimensions [3]

Future Work

- Fuzzy Kohonen Clustering to take care of a node being a member of many communities

- Other heuristics to remove dependence of output on input-sequence

Conclusions

- Kohonen SOM can be used in SNA (specially Orkut-like networks) to group members with similar interests

- Communities can be generated automatically

- Suggestion system can be implemented using this approach

- Another similar network was analyzed (dating/matrimonial profiles)

References

1. Amalendu Roy, A Survey on Data Clustering Using Self-Organizing Maps, 2000. http://www.cs.ndsu.nodak.edu/~amroy/courses.html
2. Merelo J.J., Prieto A., Prieto B., Romero G., Castillo P., Clustering Web-based Communities Using Self-Organizing Maps, Submitted to IADIS conference on Web Based Communities, 2004.
3. Visualisation of Social Networks using CAVALIER, Anthony Dekker, Australian Symposium on Information Visualisation, (invis.au 2001)
4. S. Wasserman and K. Faust. Social Network Analysis: Methods & Applications. Cambridge University Press, Cambridge, UK, 1994.