CS5232 Project Proposal

Antoine Francois Pascal Creux (A0123427M)

March 18th, 2014

1 Concept Study

Goodreads is a website launched in early 2007, which lets "people find and share the books they like and improve the process of reading and learning throughout the world." It is the worlds largest site for readers and book recommendations with a user base of about 30 million members along with 34 million reviews from 900 million books as recorded in 2015[1].

Goodreads provides a multitude of features to its users. It goes beyond the traditional rating and reviewing of books by allowing users to make friends and join and form reading groups based on their literary tastes. Users can not only see what their friends have read, but they can also meet new people with similar reading interests. They can make recommendations to friends , follow authors, track the books they are currently reading, have read and want to read. In addition, goodreads provides personalized recommendations to book readers by analyzing the user data.

To capture the notion of similar interest among the users of goodreads the concept of one-mode analyses of bipartite network data(also known as affiliation network in sociology terminology) can be used [2].Such analyses use matrices derived from the affiliation matrix, $\mathcal{A}$, wherein actors are represented in rows and the events are represented in columns.
Co-membership of actors i and j for an event k is identified if the rows a_i and a_j both have 1 in the column ak . Hence, $a_{ik} = 1$ and $a_{jk} = 1$ indicates that both actors participated in the event k. So, the total number of co-memberships for the two actors can be computed from the number of times that $a_{ik} = 1$ and $a_{jk} = 1$ where k takes all possible values. The number of events with which both actors i and j are associated will vary from 0(if both have no affiliated events in common) to h(if both have all affiliated events in common).

2 Problem Description

In this project, we aim to probe the groups of users within goodreads based on their reading interests. As stated earlier, goodreads is more than just a book reviewing website. Users and authors on goodreads can connect amongst themselves forming a rich social network.

We do not want to observe the groups/clusters within such a social network, which already exists in goodreads, since this network may not necessarily capture the reading interests of the users in it. We use the information about the books to capture the ties amongst the users.

Let U and B be the set of users and the set of books on goodreads respectively. We will construct an undirected bipartite graph $G(V, E)$ using these two sets where $V = U \cup B$. An edge exists betweem a vertex $u_i \in U$ and a vertex $b_j \in B$ if a user u_i has read book b_j.

We consider the rating given by a user u_i to a book b_j to assign weight to the edge between them. The weight to this edge plays an important role in the semantics of the networks. Consider a scenario wherein a book b_k is read by both the users u_i and u_j. Suppose user u_i has liked the book and hence has positively rated the book whereas user u_j has not liked the book and so has given a low rating to the book. If we do not take this fact into the account then these users might end up being placed in a same reading group. So as to prevent situation, we assign weight to the edges in G. Let r_{ij} be the rating given by user u_i to book b_j. Let $w \to E \times \{1, -1\}$ be the weight runction. For every $(i, j) \in E$,

$$w((i,j)) = \begin{cases} 1 & r_{ij} \geq \alpha \\ -1 & r_{ij} < \alpha \end{cases}$$

where α is some threshold which can be tuned during the experimentation.

Let $\mathcal{M}$ be a $|U| \times |B|$ matrix which represents the underlying graph G. If there exists an edge between user u_i and book b_j then $\mathcal{M}_{ij}^{th}$ entry in the matrix is set to the weight of the corresponding edge otherwise the entry is set to zero. We obtain $\mathcal{U}$, matrix encoding the relationship amongst the users, using $\mathcal{M}$ by calculating $\mathcal{M}\mathcal{M}^T$. The graph encoded by $\mathcal{U}$ is the graph of the interest for finding the user groups. We propose to explore the interesting user groups by using graph partitioning algorithms like Girvan-Newmann algorithm.

3 Underlying Assumptions

4 Requirements

Goodreads data is not available online, we have to acquire all data through their APIs. Stanford students have already studied Goodreads platform by building a product recommendation. [**?**]. However, they have not published the data they gathered and we think that their data acquisition is wrongly implemented based on erroneous assumptions.

5 Project Objectives

6 Proposed Contributions

7 Success Measures

8 Project Plan

9 Deliverables

References

[1] goodreads - About us. `https://www.goodreads.com/about/us`, 2012.

[2] S. Wasserman. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.