Content-based and Topological Friend Recommendation System

Kiruba Dhayalan, Mayank Saddi

Problem Overview:

Comparison of recommendation of friends on twitter using the social network topology and by clustering users interests. In this project we aim to design a system to suggest user with most similar members who are within n hops from the user.

To do this we plan to use and evaluate link prediction methods based on social network topology and also use user's tweets to analyze his areas of interests and use this to recommend most similar user.

Network topology methods use the structure of the sub graph around the user. It uses properties of the network like the number of common neighbors between the users and ranks users based on their similarity. Using just the topology of the network may not be enough metric to determine if a user connects with another. If we consider the user's attributes as well to make such recommendations it provides better results. As people in the real world do not just connect with one another based on how many common neighbors each have but they become friends with people based on how similar they are. In order to do this we mine for user's personality traits using his tweets/retweets data to obtain areas of the user's interest. Once we have the user's personality information we check for traits for the members who are under n hops from the user. Then compare rank them based on how similar they are with the user.

Once we have ranked users for each trait we can aggregate the values and obtain top x users then compare with the top x users recommended by the topological method. As an extension we would like to apply topological methods on just similar users and see if they improve results.

Data:

To acquire data we are making use of Twitter streaming API and pick a user and sub graph of users within n hops from the user and save their user objects.

Then we collect tweets/retweets and other related data of all users in that sub graph. It may be a large task especially with twitter rate limits it can take a long time to accumulate so we limit them to 100 to 200 tweets per user. Since most users may not be as active we only take the users who have made at least 50 tweets.

As an alternative a twitter circles data set can be used for preliminary analysis. https://snap.stanford.edu/data/egonets-Twitter.html.

It consists of a small ego network which can be used to evaluate and test the topological approaches.

Another problem we can anticipate is the segmentation of tweets of users. Tweets include several terms which are not very grammatically accurate. They contain acronyms, hashtags, emojis and mentions. For such characters it becomes difficult to segment them and have to be handled separately by extracting and saving mentions and hashtags and eliminating emojis.

Method:

Once we have user objects of all nodes n hops away from the main user and tweets from each of those users. First we do profile extraction of the main user. We do this by taking all of his tweets and segmenting them into words filtering out emojis, hashtags and acronyms. Then we can use TF/IDF to determine what the user tweets the most about and create a general profile of user's interests.

Then we do the same for all the other users' tweets and extract the top x users who have a profile similar to the main user. We now compare these top x users with the predictions that the topological methods have made and compare how the rankings vary.

Additionally we want to apply the topological methods discussed in class on only the similar users and see if we obtain any improvement in performance.

For evaluating our results we take the original subgraph and remove x number of edges from the graph. While removing the edges we only remove the last x edges added to the graph. We can do this since twitter stores the list of friends in the order in which they were added.

We then perform our predictions and see how the predictions compare to the true edges. We can create a confusion matrix and use metrics like accuracy, precision and recall to compare and evaluate the topological, user interests based and combined results.

Intermediate/Preliminary Experiments & Results:

We are in the process of streaming twitter objects from Twitter API. Parallely we are working on profile extraction with existing Twitter objects.

Related work:

User recommendation had been approached previously using the social network's topological properties. There has also been some work using user profile extraction and behavior.

In our project we are planning to implement both approaches and compare. And also see how the results may vary if they are used in tandem to provide user recommendations.

Who does what and Timeline:

No	Work Item	Deliverables	Assignee	Due Date
1	Preparation and Planning	Project proposal and plan	Kiruba, Mayank	October 30
1.1.1	Topic Decision	Target problem	Kiruba, Mayank	September 28
1.1.3	Write preliminary proposal	Preliminary proposal	Kiruba	September 30
1.1.4	Write proposal	Project Proposal	Mayank	October 2
1.2.1	Discussion	Related Work	Kiruba, Mayank	October 12
1.2.2	Write Plan	Initial edition of plan	Kiruba	October 26
1.2.3	Submit Milestone	Milestone	Mayank	October 30
2	Data Collecting and Preparation	Target Data	Kiruba, Mayank	November 9
2.1.1	Data Download	Data sets	Mayank	November 2
2.1.2	Build SQL Database	Database	Kiruba	November 2
2.1.3	Discussion	Resource Exchange	Kiruba, Mayank	November 3
2.2	Data Cleaning	Target data	Mayank	November 4
2.2.1	Missing Values Detection	Missing Plot	Mayank	November 5
2.3	Descriptive Analysis	Variables Selection	Kiruba	November 6

2.3.1	Variables Summary	Matrices plots	Kiruba	November 7
2.4	Binding Data	Target dataset in database	Kiruba, Mayank	November 9
3	Feature Extraction		Kiruba, Mayank	November 12
3.1	Building Feature	Feature Matrix	Kiruba	November 10
3.2	Profile Feature Extraction	Feature Matrix	Mayank	November 12
4	Models		Kiruba, Mayank	November 14
4.1.1	Model 1		Kiruba	November 13
4.1.2	Model Validation	Validation plot	Kiruba	November 14
4.2.1	Model 2		Mayank	November 13
4.2.2	Model Validation	Validation plot	Mayank	November 14
5	Deployment		Kiruba, Mayank	November 17
5.1.1	Model Evaluation	Performance Report	Mayank	November 15
5.1.2	Code organization	Code files and blocks	Kiruba	November 15
5.1.3	Submission of Report		Mayank	November 17
6	Presentation	Slides	Kiruba, Mayank	November 20
6.1	Making Slides	Slides	Kiruba	November 19
6.2	Attend Presentation	Presentation	Kiruba, Mayank	November 20

References:

1. Wang, J., Gao, S., Wang, L., & Yu, Z. (2018). Micro-Blog Friend-Recommendation Based on Topic Analysis and Circle Found. *2018 IEEE Fourth International Conference on Big Data Computing Service and Applications (BigDataService)*. doi: 10.1109/bigdataservice.2018.00033
2. Tasgave, P., & Dani, A. (2015). Friend-space: Cluster-based users similar post friend recommendation technique in social networks. *2015 International Conference on Information Processing (ICIP)*. doi: 10.1109/infop.2015.7489465

3. Srilatha, P., & Manjula, R. (2016). User behavior based link prediction in online social networks. *2016 International Conference on Inventive Computation Technologies (ICICT)*. doi: 10.1109/inventive.2016.7823266
4. Ahmed, C., & Elkorany, A. (2015). Enhancing Link Prediction in Twitter using Semantic User Attributes. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM 15*. doi: 10.1145/2808797.2810056
5. Lee, W.-J., Oh, K.-J., Lim, C.-G., & Choi, H.-J. (2014). User profile extraction from Twitter for personalized news recommendation. 16th International Conference on Advanced Communication Technology. doi: 10.1109/icact.2014.6779068
6. Moreno, D. R. J., Gomez, J. C., Almanza-Ojeda, D.-L., & Ibarra-Manzano, M.-A. (2019). Prediction of Personality Traits in Twitter Users with Latent Features. *2019 International Conference on Electronics, Communications and Computers (CONIELECOMP)*. doi: 10.1109/conielecomp.2019.8673242
7. Volkova, S., Bachrach, Y., & Durme, B. V. (2016). Mining User Interests to Predict Perceived Psycho-Demographic Traits on Twitter. *2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService)*. doi: 10.1109/bigdataservice.2016.28