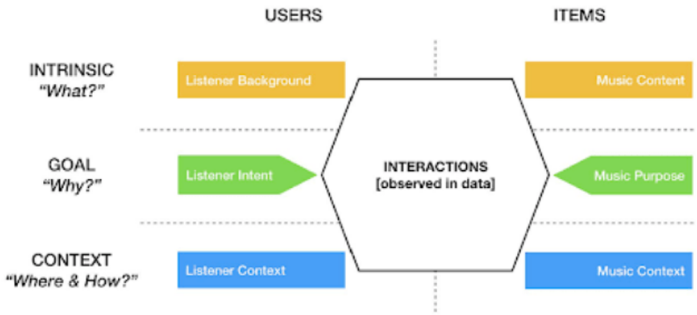
A Music Recommendation System Based on Network

---------- A Network Study of Million Song Dataset

Team: Xinyi(Sarah) Zhao, Yue Ma, Ryan Deng

Motivation



- Popular music is a kind of music with wide appeal that is distributed to large audiences, which is an important cultural expression that captures the listener's attention for years

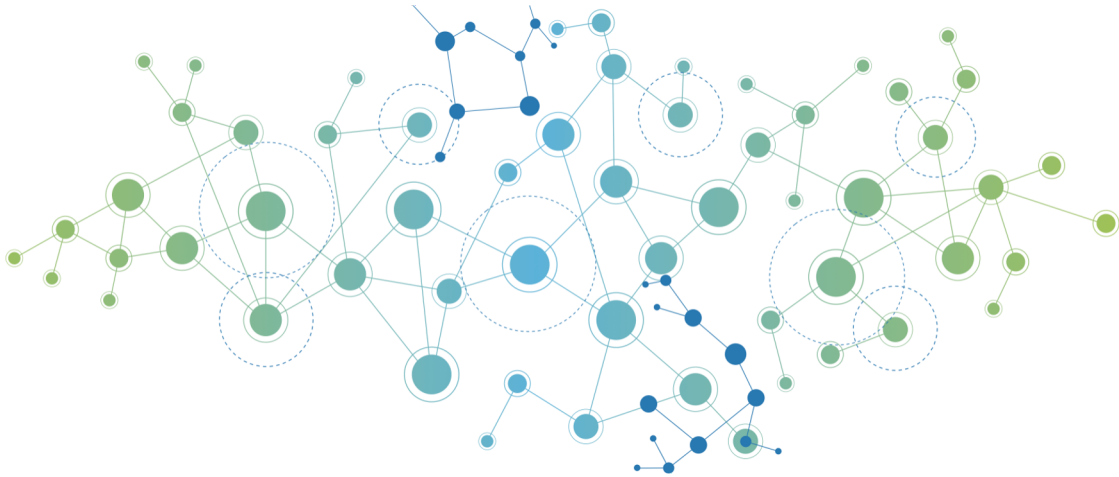
- Popular music has also been greatly developed in this era of the Internet

- We want to set up a music recommendation system based on the song network to help the audience to find more music that suits their taste

Research Questions

- What are the characteristics of the song network based on the user's playing history?
- Can the song network meet the basis to build a recommendation system?

Related Work

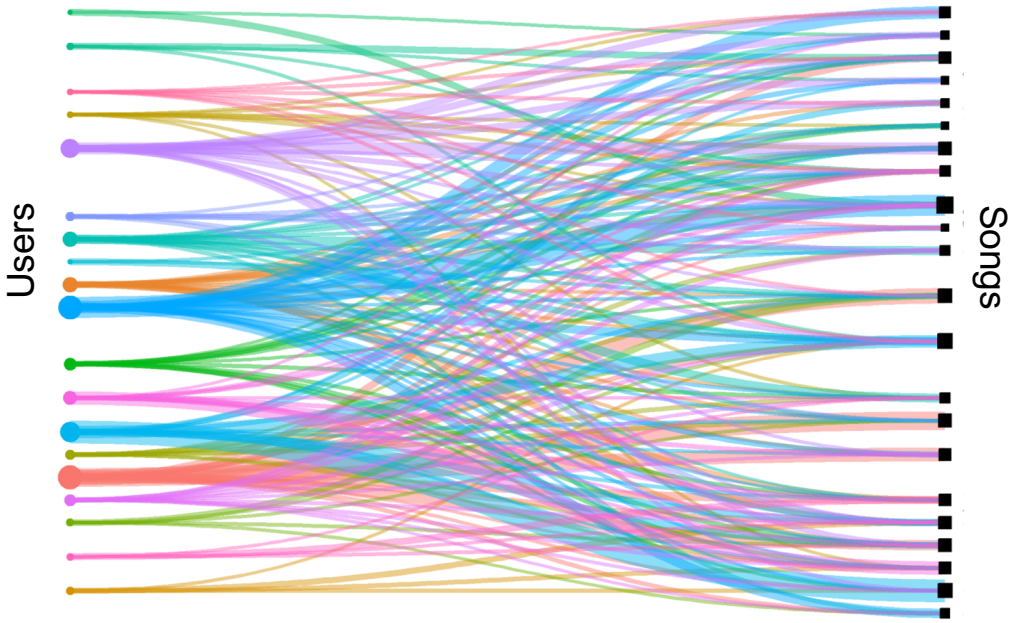


Vinothini Kasinathan's team: proposed a music recommendation system based on **fuzzy logic**, which makes decisions on music recommendation based on users' music listening habits, music genres, and their impact on human beings (Kasinathan et al., 2019).

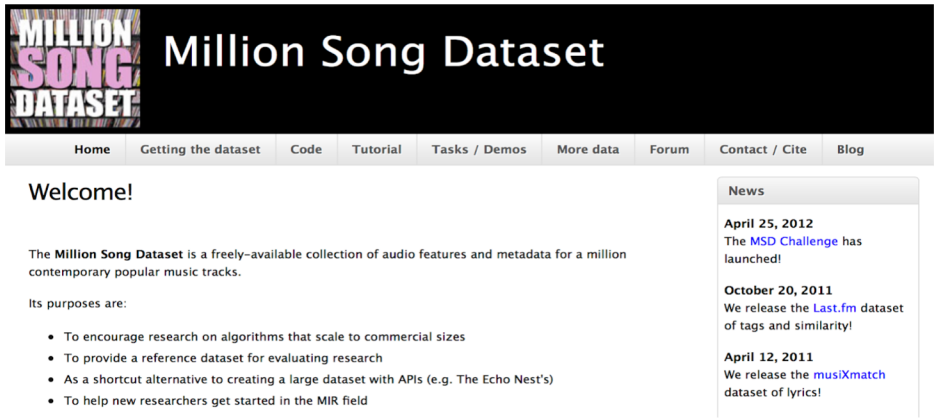
Byeong-jun Han's team: proposed a **context-based** music recommendation (COMUS) ontology to model users' music preferences and contexts, and to support reasoning on users' expected emotions and preferences (Han et al., 2009).

Analysis Approaches

Our Aim: we consider the connection between users and songs as **a bipartite network**, to find the connection between users. Our strategy is to recommend new songs to users based on the preferences of other users who have played the same songs, and the songs that would be selected for recommendation will be evaluated by the **HITS** or **PageRank** or **Degree Centrality** in the network.



Datasets



1. Song Dataset (subset-compiled.csv)

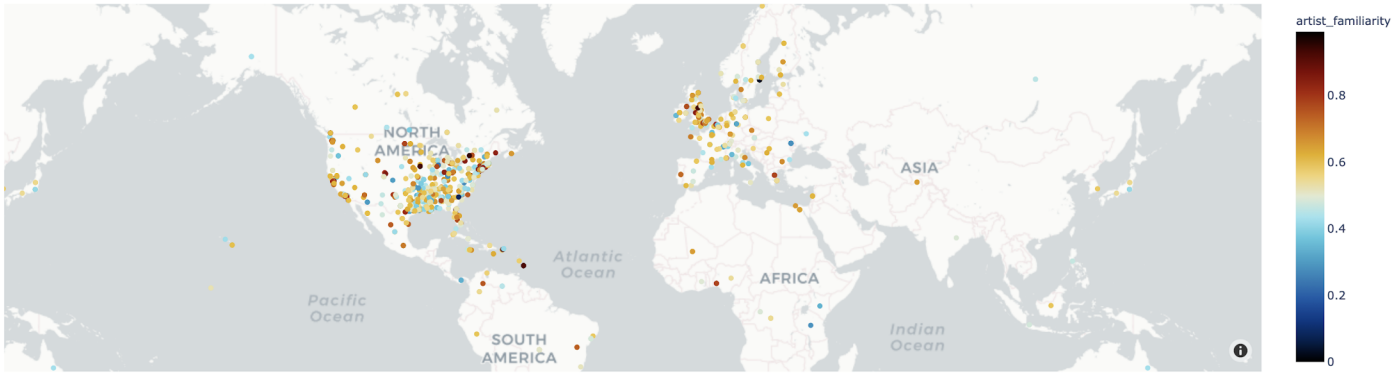
A subset of Million Song Dataset, containing 10,000 songs, about 1% selected at random from the original data. (source: https://github.com/subha5gemini/MillionSongDataset/blob/master/subset-compiled.csv)

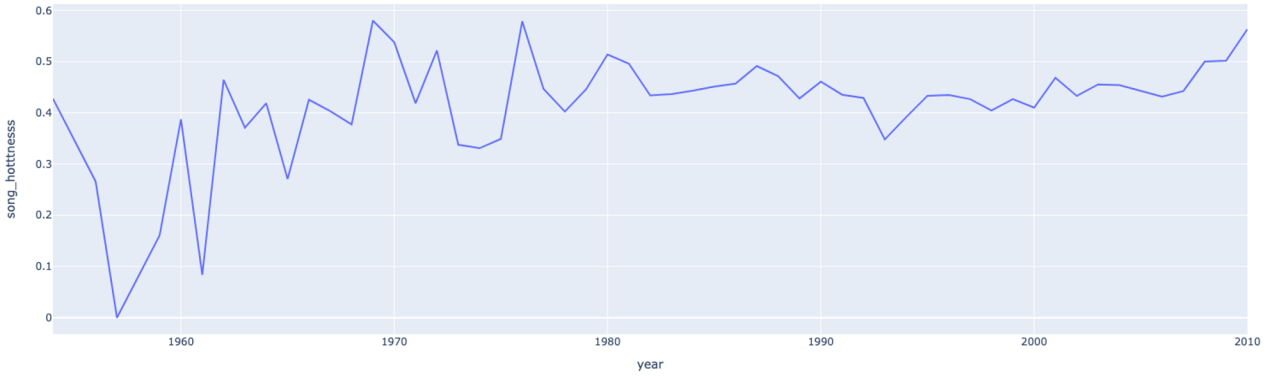
1. Taste Profile Dataset (Train_triplets.txt)

A taste profile data, containing 1,019,318 unique users, 384,546 unique songs, and 48,373,586 user-song-play_count triplets. (source: http://millionsongdataset.com/sites/default/files/challenge/train_triplets.txt.zip)

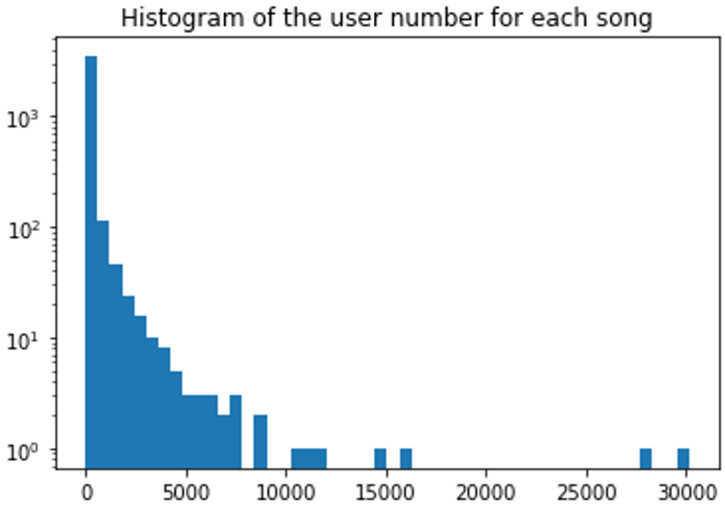
Sample Set: as the taste profile dataset is too large, we only use the songs in both datasets, resulting in 772,661 user-song-play_count triplets

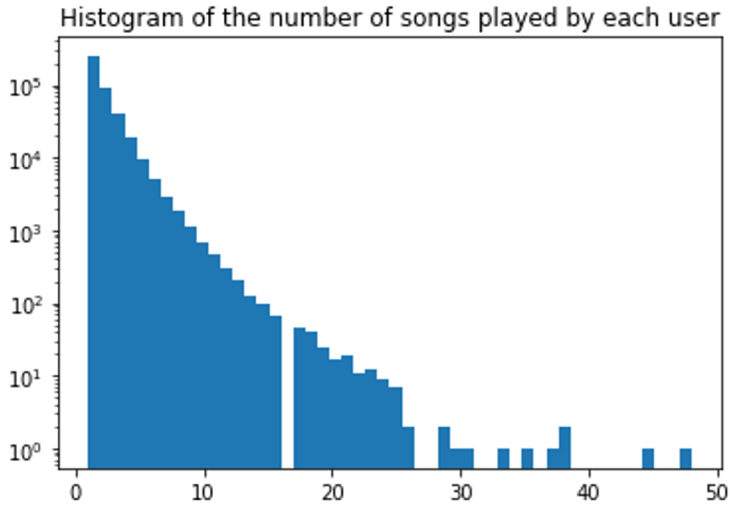
Global distribution of music artists and song hotness by years





Exploration Analysis





The song with the max number of users has 30,117 users.

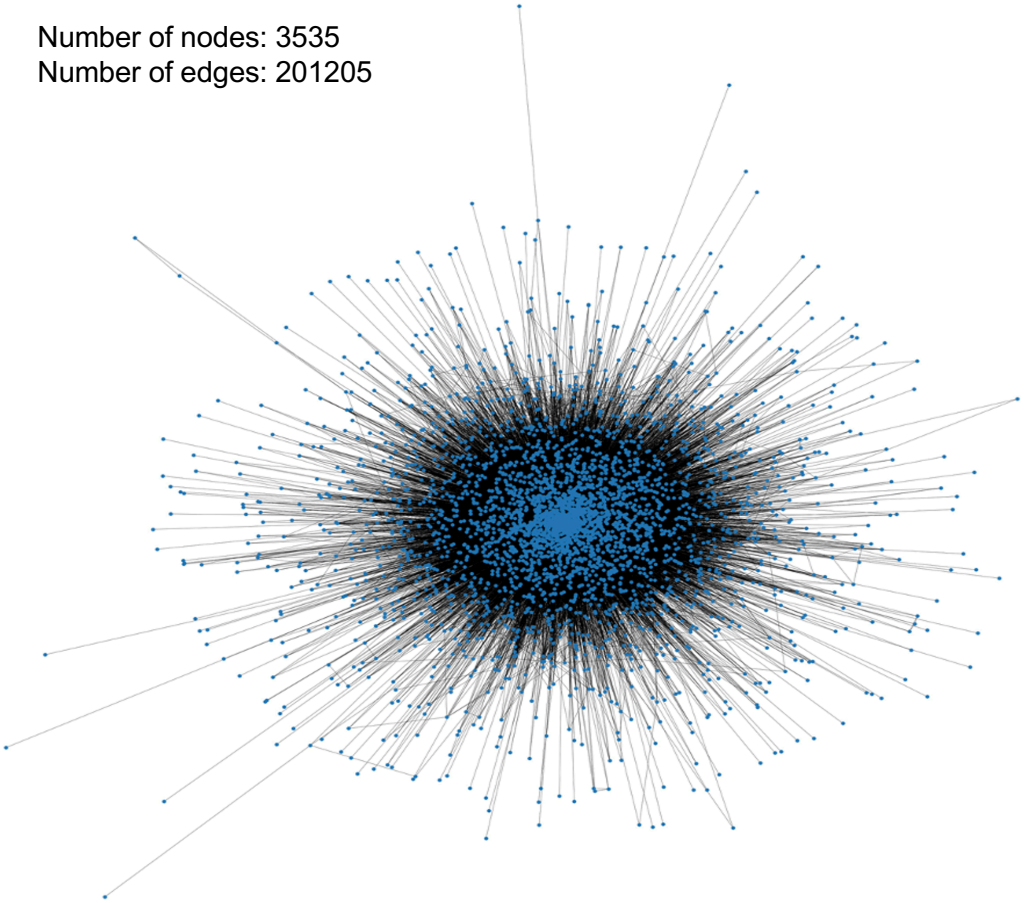
The user played the max number of songs played 48 songs.

The user number for each song and the song number for each user both follow power law distribution.

Network Setup

The relationship between users and songs can be considered as a bipartite network with users group and songs group, so we built a one-mode projection graph of the songs for our sample set. The weight of each edge that connects a pair of songs was created by the number of users who played the pair of songs.

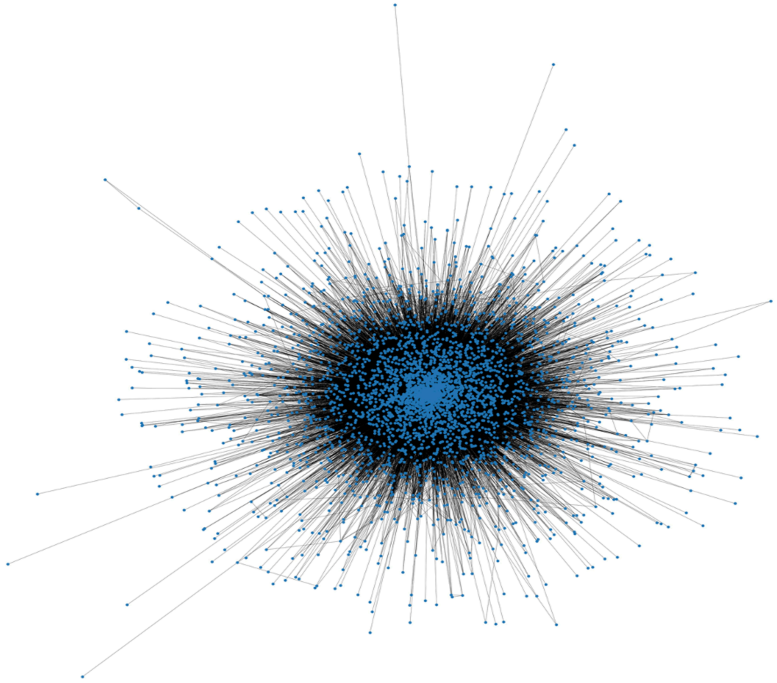
Number of nodes: 3535
Number of edges: 201205



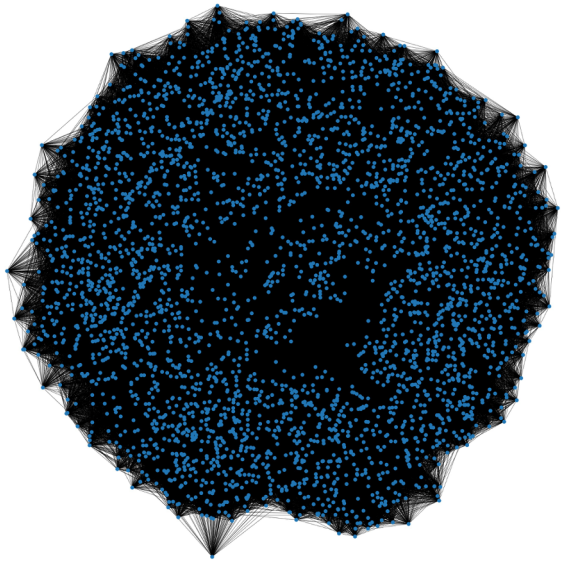
Network Graph of Sample Set

This graph fulfills the conditions of a Small-World Network

Number of nodes: 3535
Number of edges: 201205
Average degree: 113.8359
Average shortest path length: 2.2898
Average clustering coefficient: 0.6072

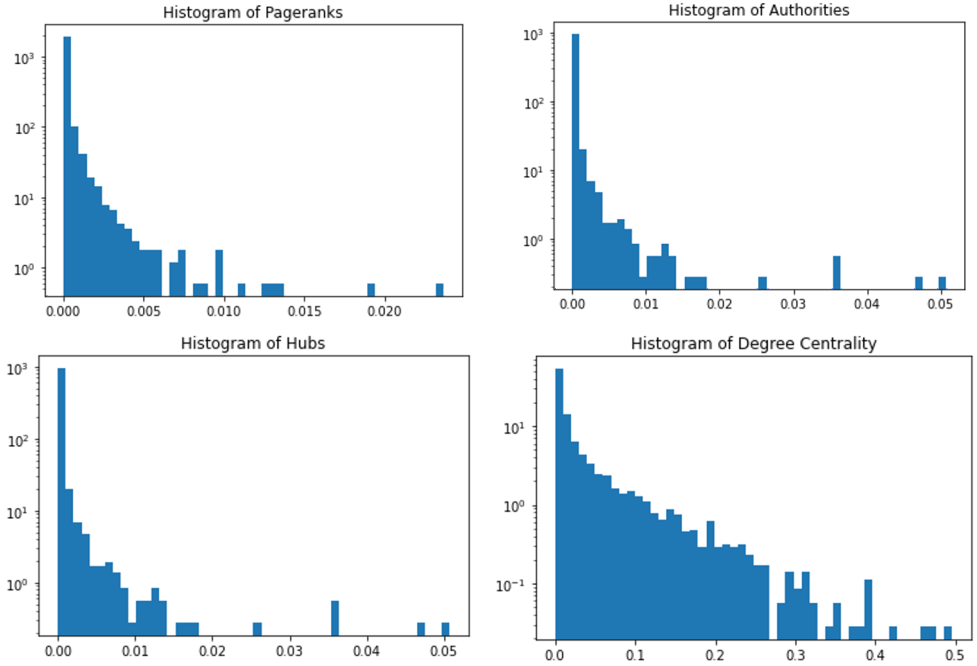


Number of nodes: 3535
Number of edges: 201495
Average degree: 114
Average shortest path length: 2.2050
Average clustering coefficient: 0.3888



Network Graph of a Watts Strogatz network
(`nx.watts_strogatz_graph(3535,114,0.2)`)

PageRank and Hitting-Time



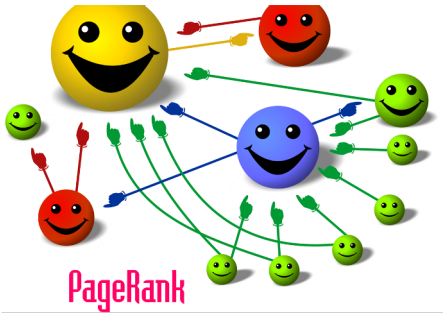
Pagerank, HITS and Degree Centrality for each node are calculated for the songs network. All of them follow the power-law distribution, which means that a small fraction of nodes plays more important roles in the network.

HITS computes two numbers for a node: Authorities estimate the node value based on the incoming links, Hubs estimate the node value based on the outgoing links

PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links.

Degree Centrality is a metric to analyze the centrality of nodes.

Compare



- Two features have been used as reference to evaluate the accuracy of the algorithms, song play count and song hotness.
- Approach: sort the songs with the top 50 highest values of total play count, song hotness, and also sort the songs with the top 50 highest values of pagerank, authorities, hubs and degree centrality, then compare the similarity of the songs between PageRank|Authorities|Hubs|Degree Centrality group with the play_count|song_hotness group by jaccard_similarity_score.
- Result: referring to play_count is better than referring to song_hotness, PageRank has a higher similarity score, and the similarity score of Authorities, Hubs and Degree Centrality are much lower.
- So, we use PageRank to do the recommendation.

Jaccard Similarity	PageRank	Authority	Hub	Degree Centrality
Play Count	0.1	0.06	0.06	0.04
Song Hotness	0.02	0.0	0.0	0.02

Result - Make the Recommendation

	user_id	song_id
218033	bc987c193110cd7d1233bf40035a1920f13d4c6c	[SOFSIKY12AF72A0807, SORTEBZ12A6D4FC1FC, SOHXO...
159131	89d56e27df5e808168c338af60f586e60eb71de2	[SOXMGFH12A6701FF31, SOGIACM12AF72A1BFD, SOQKI...
270112	e9691cacc69a0328807c18b7594f9999bfc9c7e2	[SOWCKVR12A8C142411, SOQGVCS12AF72A078D, SOFCP...
193717	a7b27cb9ce8ccd040c9a307e88f203d3dc8ddef6	[SOAXGDH12A8C13F8A1, SOSXLTC12AF72A7F54, SONYK...
160155	8abffe830bd921f06aad7698197f00fd01ece56f	[SOMYXWV12A8C14232E, SOPNLBX12A8C1377D4, SOZVS...
140973	7a228fa1af26c07e864be042debffd81108da7e6	[SOBOUPA12A6D4F81F1, SOWKQYL12AB0183B15, SOARU...
2540	023e8868f2fbb29a91ea88c39101aba9f87c6afd	[SOFRQTD12A81C233C0, SOPTLQL12AB018D56F, SOJYB...
179167	9b068e57a5d4f9d17b281f1c35b2815bc2669d3f	[SOFRQTD12A81C233C0, SOAXGDH12A8C13F8A1, SOAUW...
258238	df3b8db0a93c0bf8c6b6c3abf83d7080b773cf8c	[SOSXLTC12AF72A7F54, SONYKOW12AB01849C9, SONNS...
43129	257fd72f953bf605c2d51302f32f23a3784d7507	[SOPQLBY12A6310E992, SOKUPAO12AB018D576, SOLLN...

For each user

1. Find the neighbor songs in the graph based on the songs that have already been played by the user
2. Recommend the top 10 songs with the highest PageRank values

Evaluate the quality of the recommendation by network

- The playcount values of the user history were utilized as the baseline.
- The commonly used recommendation system “Collaborative Filtering” was implemented to build the recommendation engine, in order to make a comparison with network method.
- SVPpp algorithm was used for Collaborative Filtering method, the accuracy of this model has been calculated by RMSE(0.5079) and the MAE(0.2033), indicating an accurate prediction.
- 5000 users were randomly selected from the dataset and the average play count value of the songs recommended to each user was calculated in three groups, the recommendation results from Network method (14,664), that from Collaborative Filtering(36,792) and from the user history (the baseline)(14,447).
- The result show that there are significant differences between network group and user history group, as well as collaborative filtering. But the difference between collaborative filtering group and user history group is not significant. It indicated that the network recommendation method works well for this dataset.

Challenge

201 users (0.05% of the sample set) only played one song, and these songs were only played by these users, so these songs are not in our sample set network, causing us can’t make a recommendation for these users.

The reason is that our sample set limits a range of songs, and some users have extra songs that are not on our song list. If we expand the song list, we will have more user-song-play_count triplets, so that this problem can be solved.

Conclusion

- The taste profile dataset (train_triplets.txt) can be considered as a bipartite network with users group and songs group. The one-mode projection of the songs is built as a network graph.
- The structure and features of the song graph indicate that it is a Small-World Network.
- We calculated Pagerank and HITS (authorities and hubs), and degree centrality values for each node of the songs network, which will be used for recommendation.
- The accuracy of Pagerank, HITS and degree centrality are evaluated referring to songs' play count and hotness, and PageRank is a more accurate algorithm than all the others.
- The songs recommendation is implemented by looking for the connected songs in the network graph to the songs that has been played by each user, then the connected songs with the highest pagerank value will be recommended to the user.
- The play count values of the user history has been used to evaluate the network recommendation, and the commonly used Collaborative Filtering recommendation system has been used to comparison. The network recommendation result is better than the baseline and works better than Collaborative Filtering in this case.

References

Kasinathan, V., Mustapha, A., Tong, T.S., Rani, M.F., & Rahman, N.A. (2019). Heartbeats: music recommendation system with fuzzy inference engine. *Indonesian Journal of Electrical Engineering and Computer Science, 16*, 275-282.

Han, B., Rho, S., Jun, S., & Hwang, E. (2009). Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications, 47*(3), 433–460. https://doi.org/10.1007/s11042-009-0332-6

Serrà, J., Corral, Á., Boguñá, M., Haro, M., & Arcos, J. L. (2012). Measuring the Evolution of Contemporary Western Popular Music. *Scientific Reports, 2*(1), 521. https://doi.org/10.1038/srep00521

Bertin-Mahieux, T., Ellis, D., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR 2011)*