Final Paper

Ayan Patel, Robert Hensley

Introduction

Dataset: Gowalla (Stanford SNAP Data)
https://snap.stanford.edu/data/loc-Gowalla.html

Gowalla is a location-based social networking website where users share their locations by checking-in. The friendship network is undirected and was collected using their public API, and consists of 196,591 nodes and 950,327 edges. We have collected a total of 6,442,890 check-ins of these users over the period of Feb. 2009 - Oct. 2010.

Using the user friendship table and the total check ins table we are able to generate some new data and create a graph representation that we can use to answer some questions. Our questions have three main categories: Popularity, Location, and Time.

Popularity

- Who is the most influential / popular person?
- Given a user, which other users are friends with them?
- How many groups, or clusters, are present that are strongly connected?

Location

- What is the most popular location?
- How are locations connected? Where do similar groups of people meet up? Are certain locations clustered?

Time

- What time is a popular time for people to meet?
    - Check in terms of hours
- Are most of the people that checkin at a given place and time friends?
- What time do people meet at different locations?

The article tied to the dataset focuses on how long distance travel influences social networks, specifically what motivates a user to travel. Their choice of extracting data from Gowalla makes sense because Gowalla users were active in several different countries and many users use Gowalla in several countries (perhaps to document where they went on vacation). Our research will focus more on the centrality of social network groups as well as the centrality of popular locations. However our research will be similar to the paper's research in the sense that we are going to be observing social networks on a global scale.

Graph Outline

The Gowalla data needed to be cleaned into a readable format before it could be used. This includes creating proper datetime objects for time analysis and storing the data in pandas dataframes to more easily query important information.

Additionally, we extracted additional information from the dataset to generate more data and graphs. We used OpenStreetMap (OSM), a map API to get more information about the locations using the latitude and longitude coordinates to query information about the location such as the city and country the location is located in. To do this, we had to iterate through the totalcheckins table and combine it with the user table to make assumptions about locations and friend groups. We used the top 10k places, as the amount of data was worldwide.

Graph Algorithms

We were able to use our graph to answer our questions in the following ways:

Popularity

In terms of popularity we can use the user's friends to identify what user is connected to the most number of users. We can compare this to using the PageRank algorithm on the user friendship graph. We can also identify the user's friends and compare their connections. We can identify groups of friends using clustering algorithms.

Location

We can use the locations to identify the most popular locations based on the number of users that visited a particular location. We can also use our graph to determine if a user has visited multiple locations and we can check the type of location to infer what places a user likes. This can show us if a group of people prefer a certain type of location to meet up at. Since the total checkin table was large, we used the top 10k places, using the OpenStreetMap API to identify the locations. With the information gathered from the map API, we found which cities and which countries had the most social influence using the pagerank algorithm

Time

We can use the timestamp to identify at what time of day, or even what season, different users prefer to go to different locations. We can also compare if users that are friends go to the same locations at the same time of day or at different times of day.

Results

All of our code and explanations are in the attached python jupyter notebook, where we load the data, pre-process the data, create dataframes, create graphs, and run graph algorithms to obtain answers to our questions. Some of our questions could be answered by filtering and querying the dataframes. Some required the use of the graph algorithms. Some could be derived from both methods so it was also interesting to see the performance differences.

Conclusion

Overall, we were able to build graphs from location based social network data and run algorithms to answer our questions. Most of our time went into getting information about a location using a map API as well as pre-processing data into dataframes we could use to obtain useful information and create graphs. We learned how to manipulate data and how to run different algorithms for different use cases and successfully answered our questions.

A lot of improvements come down to technical cost. Many of the algorithms we ran were too time intensive given the limited hardware we had at our disposal. One way to improve this would be to run our algorithms on a more sophisticated cloud computing system. The book *Graph Algorithms* by Needham & Hodler uses Apache Spark to run their example algorithms and queries a graph Database in Neo4J. Using a cloud computing service like AWS with tools like Apache Spark and Neo4J would undoubtedly improve the speed of running algorithms on massive amounts of data like the data presented in our project.

We additionally could use a better data source than OpenStreetMap for our datasource. The information provided in the JSON file is pretty sparse and because the content is community generated can often be prone to errors or misspellings. A paid API service such as the *Yelp Fusion API* might be better for this project because it provides more accurate information and more importantly introduces context to the locations. This context in particular would be descriptors for the locations, such as if the location is a restaurant, park, movie theater, etc. and would give more valuable insights into how places are connected.

Sources

- https://snap.stanford.edu/data/loc-gowalla.html
- https://cs.stanford.edu/people/jure/pubs/mobile-kdd11.pdf
- Graph Algorithms 2nd Edition (O'REILLY Media, Needham & Hodler)
- https://www.openstreetmap.org/#map=18/35.29290/-120.66709
- https://www.yelp.com/fusion
- https://igraph.org/python/