Application of K-Means Clustering for Mining nearest Wi-fi Hotspots

Proposal for a Project

Presented to

Prof. Sharon Stranahan

Department of Computer Science

San José State University

In Partial Fulfillment

of the Requirements for the Class

CS200W Sec 02 2017

By

Yuvraj Singh Kanwar

October 19, 2017

# Table of Contents

[I. Problem Statement 1](#_Toc497155208)

[II. Research Objective 1](#_Toc497155209)

[III. History and Background 2](#_Toc497155210)

[IV. Technical Approach/Methodology 3](#_Toc497155211)

[V. Requirements 4](#_Toc497155212)

[VI. Progression Timeline 4](#_Toc497155213)

[References 5](#_Toc497155214)

# Problem Statement

Wi-fi is one of the most important technologies being used in the internet era because it allows the users to connect to the internet wirelessly. It has become a necessity just like food and shelter. Many restaurants, malls, cafes and libraries provide free public wi-fi these days. Naturally, it becomes a valid option to have an application that can assist you in searching nearest wi-fi accurately. People do not have efficient solution to identify wi-fi hotspot near them so that they can just reach there and get connected to the internet. In the era of technology that we are living in, life without internet has become unimaginable.

With this project I intend to generate a k-means clustering application to gather wi-fi hotspots near a user and identify the nearest available hotspot. The main area of focus will be creating be a K-means Cluster implementation in java with an uncommon equal option which enforces a constraint of equal cardinality on the clusters and still efficiency being as spatially cohesive as possible. This will ensure user always gets best possible results in the most efficient time. Equal size of cluster will ensure that the user gets more, and accurate number of results close to the user when it queries for wi-fi hotspots near it.

# Research Objective

Contribution of this project is mainly to develop a k-means clustering model to deal with the large dataset of wi-fi hotspots [1], that offer free internet access to people in a city, to generate accurate navigation application that can provide intelligent solution to basic navigation queries of the people like

1. Where can the user find hotspot nearby?

2. How many wi-fi hotspots are present close to user?

3. How far does the user need to travel to get to nearest wi-fi?

This application can further be scaled to many other datasets like restaurants, grocery outlets, etc.

The main object of the project is to be able to achieve a variation of k-means that will ensure that each k-means cluster will have uniform cluster size and each cluster has a uniform number of points. With the project I intend to set the centroid dynamically for the k-means algorithm to provide the user at least five nearest wi-fi hotspots.

The goal is to use machine learning algorithm to locate the nearest wi-fi hotspots by augmenting the current location of the user.

# History and Background

Edwin Diday, in his early phase of career, chose cluster analysis as prime research topic. He wrote a monograph (Diday et al. 1979), alongside twenty-two co-writers, that traces noticeable level of generalization of the crux of k-means and rooted its utility in model-based clustering. ’Principles of numerical taxonomy’, a monograph that Sokal and Sneath wrote in 1963 gained attention across the globe and there was a curiosity in clustering methods among researchers. It became catalyst for the publication of books like ‘Automatische Klassifikation’ written by Bock in 1974, ‘Les bases de la classification automatique’ which was written by Lerman in 1970, and ’Cluster analysis for applications’ which was written by Anderberg in 1973. Consequently, the basic ideas and functions of clustering became familiar over broad scientific communities of data analysis, decision analysis, statistics and particularly, applications.

K-means clustering algorithm is the most famous clustering approach being used today and it is built on sum of squares criteria. If you trace back the origins of the algorithm, several scientists proposed in various forms and under diverse presumptions the same algorithm. Then many scholars like Cox in 1957, Fisher in 1958 and Bock in 1974 explored analytical and functional aspects of the method while looking for ’continuous’ analogues of the SSQ criteria. Hartigan in 1975, Pollard in 1982 and Bock in 1985 investigated the algorithm by exploring asymptotic behavior under arbitrary sampling strategies and stretching the algorithm’s sphere to probabilistic models and new data types. [2]

The wi-fi dates back a while ago now and its history is fascinating. ALOHA System wanted to and managed to link the Hawaiian Islands with a UHF wireless packet network in 1971. Before IEEE 802.11, the forerunners to Ethernet were ALOHA System and the ALOHA protocol. Vic Hayes is considered the “father of Wi-Fi.” His reseach on wi-fi technology started in 1974 when he worked with subsidiary of Agere Systems, NCR Corp. Agere Systems makes components for semiconductors. U.S. Federal Communications Commission made a ruling in 1985 which released the frequencies of 2.4GHz ISM band for unlicensed use. These bands are subject to interference and are equivalent to ones used in microwave ovens. AT&T Corporation along with NCR Corporation developed the forerunner to 802.11 in 1991. They invented what we now call wi-fi. The first wireless products were under the name WaveLAN. [3]

# Technical Approach/Methodology

Clustering algorithms are generally used in an unsupervised way and a set of data instances must be presented after grouping them with a concept of similarity. The function has admittance only to few features that describe every object i.e. algorithm is not given any data about where each instance should be placed within the partition. [4] The Κ-means clustering algorithm produces an outcome using iterative refinement. The input provided to K-means algorithm is a data set along with the K number of clusters. Κ centroids is taken as an initial estimate at the beginning of the algorithm, which can either be randomly generated or randomly selected from the data set. [5]

The implementation of this would be done following the below steps:

1. Find the data set of wi-fi hotspots located in New York city.
2. Clean the data to fill in missing values and remove any noise from the data.
3. Identify and remove outlier data that will not impact the accuracy of algorithm in the dataset.
4. Perform Principal Component Analysis data reduction technique on data set to reduce unnecessary dimensionality or numericity.
5. Partition dataset into k nonempty subsets and then compute seeds as the centroids of the clusters of the current partition.
6. Enable a special equal option to ensure K-means algorithm variation will have equal cluster size.
7. Devise clustering procedure where each cluster has an equal number of points. The algorithm, in a reasonable amount of time, possess ability to do its best to regard the equal cardinality constraint, but one needs to be very careful when this option is enabled because sometimes the algorithm might generate clusters that are not as spatially cohesive and thus equally efficient like the original K-means algorithm (NP-hard problem).
8. Assign each object to the cluster with the nearest seed point.
9. Find the cluster nearest to the user based on its location and calculate the nearest wi-fi hotspots.

# Requirements

To conduct the research, the resources that I would require is a computer with a multicore processor, preferably i7, running Linux with software as below:

1. Hadoop
2. Weka
3. Oracle Virtual Machine
4. Eclipse JEE Oxygen
5. Java JDK 1.7

These are all easily available over the internet and are open source. I will also be using a large dataset of New York City Wi-Fi Hotspots for the project [1]which is also publicly available over the internet.

# Progression Timeline

|  |  |
| --- | --- |
| Week 1: September 4 – September 8 | Select Project Topic and identify Issues |
| Week 2: September 11 – September 15 | Identify Solutions to problems and tools that are needed |
| Week 3: September 18 – September 22 | Create Proposal and get feedback for the idea |
| Week 4: September 25 – September 29 | Deliverable 1: 1st Proposal |
| Week 5: October 2 – October 6 | Literature Review begins |
| Week 6: October 9 – October 13 | Literature Review Continues |
| Week 7: October 16 – October 20 | Deliverable 2: Literature Review |
| Week 8: October 23 – October 27 | Choose base paper and base algorithm |
| Week 9: October 30 – November 3 | Develop on the experiments and begin working on report |
| Week 10: November 6 – November 10 | Continue Development |
| Week 11: November 13 – November 17 | Create Presentation |
| Week 12: November 20 – November 24 | Deliverable 3: Final Presentation and Report |

# References

1. City of New York (2017) New York City wi-fi Hotspots Available: https://www.kaggle.com/new-york-city/nyc-public-wifi/version/1/data
2. H. H. Bock (2007) Clustering Methods: A History of k-Means Algorithms. In: P. Brito, G. Cucumel, P. Bertrand, F. de Carvalho Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg
3. Cablefree (2017) The History of wi-fi: 1971 to Today, Wireless Excellence Limited, Oxford Science Park, Oxford, OX4 4GA, UK Available: http://www.cablefree.net/wireless-technology/history-of-wifi-technology/
4. K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl - ICML, 2001, Constrained k-means clustering with background knowledge, Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577–584.
5. A. Trevino... Introduction to K-means Clustering. Available: https://www.datascience.com/blog/k-means-clustering. June 12, 2016.