Great Title or Something

Authors Name/s per 1st Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: Email: name@xyz.com

Authors Name/s per 2nd Affiliation (Author)
line 1 (of Affiliation): dept. name of organization
line 2: name of organization, acronyms acceptable
line 3: City, Country
line 4: Email: name@xyz.com

***Abstract*—**

***Keywords*-**

I. Introduction

II. Previous Work

III. System Design

IV. Experiments

V. Introduction

A. Motivation

The motivation for our work comes from the analysis of a set of cell-tower data records (CDR) provided by the nation of Andorra. Andorra is a small (approx. 468 km^2), mountainous nation in Central Europe located between France and Spain [1]. The economy of Andorra is primarily centered around tourism, retail, and finance, with approximately three-quarters of the nation's $3.163B economy coming from these three sectors [1]. With the goal of improving the responsiveness of both industry and government to tourism, the government of Andorra provided the entirety of the CDR records, which we summarize as follows:

User Info.	ID (sender/receiver), Nationality (sender)
Call Length	start time, end time
Location	cell tower (sender and receiver)
Device ID	IMEI, IMSI, and TAC Codes [1]

B. Limitations

As alluded to above, a variety of research has explored the problem of path prediction from a variety of contexts. In particular, much of the literature has focused on developing path prediction for individuals using the concepts of recommender systems. Generally, the first approach is characterized by generating some characterization of points of interest in multiple areas and then generating path predictions based on a combination of the user's history and a recommender system for similar points of interest [2], [3]. Alternatively, other systems utilize a context-aware approach to group users by preference limited to a specific geographic area [4], [5].

While these methods would provide a great solution to the given problem, our data provides a certain set of limitations. In particular, our data lacks the prior knowledge assumed in [2], [3] and lacks a way of connecting our information to the sentiment knowledge assumed in [4], [5]. Therefore, we now move to a discussion of the methods we utilized to approximate the methods above that utilize more rich datasets.

VI. Methods

As described above, the set of CDR data provided is relatively limited with respect to the relevance of the information provided by the dataset. To this end, we utilized two separate methods aimed at providing path predictions from different analyses. Specifically, these models can be broken into a method for characterizing movement at the individual level (Section II-A) and at the national level (Section II-B). These models, however, are limited with respect to their predictive accuracy at different scale. Therefore, we conclude by developing a hybrid model (Section II-C) that combines the benefits of each individual model.

A. Individual-Level Prediction

The first model of movement developed is derived by collaboratively filtering individual paths with respect to paths previously observed in the data. First, this algorithm assumes a set of reference paths, which we will call $Reference$, with each record of the form $urec = (nationality, path)$. Then, given a new user record $urec' = (nationality, path')$ where $path'$ represents some part of the given path, we choose the k-Nearest Neighbors to the given path using the Levenshtein Distance (LD) as a metric of path similarity and then take the k most frequent elements of these paths not in the initial path as the prediction.

B. National-Level Prediction

The second model of movement developed is derived from generating a comparison of the different towers in the region based on features derived from the original data. From these

[1]These codes are universal codes assigned by manufacturers identifying their products. Of particular interest are the TAC codes which identify the type of phone the user has.

features, we then clustered these regions to generate a list of the towers most similar to each other.

1) Features: From the features described in Section I-A, we derived four distance metrics that compared regions based on user-dependant (connectivity, Jaccard index) and context-dependant (nationality, phone cost) features.

(a) **Connectivity**: Given a region, we derive a vector C_r that is comprised of a summation over all regional records of the cell-tower of the receiver for the 33 towers in the region. From this, we then defined the distance between two regions, A and B, to be norm of the difference between the two normalized vectors:

$$\text{dist}_C(A, B) = ||\frac{C_A}{||C_A||} - \frac{C_B}{||C_B||}||$$

(b) **Jaccard Index**: Next, for each region we derive a second pairwise distance metric as the Jaccard index of two regions based on users in the region. Specifically, for two regions, A and B, let $U(A)$ be the set of users of A and likewise for B. We define the distance between A and B to be the Jaccard index:

$$\text{dist}_J(A, B) = \frac{|U(A) \cap U(B)|}{|U(A) \cup U(B)|}$$

(c) **Nationality**: Similar to the definition of connectivity, we define the nationality distance between two regions to be the norm of the difference of the normalized vectors summing the regions composition per-nationality. Specifically, if we let N_r to be the sum of each individual nationality, the distance between two towers A and B is defined as:

$$\text{dist}_N(A, B) = ||\frac{N_A}{||N_A||} - \frac{N_B}{||N_B||}||$$

(d) **Phone Cost**: Finally, we define the distance between the phone cost of two regions to simply be the absolute value of the difference in mean phone cost for each region. Specifically, given a function $c(i)$ which maps user i to the value of user i's phone, the distance between two regions A and B is defined as:

$$\text{dist}_P(A, B) = |\frac{1}{|B|}\sum_{i \in A} c(i) - \frac{1}{|B|}\sum_{i \in B} c(i)$$

2) Distance Metrics:

3) Clustering:

C. Hybrid Model

VII. Experiments

VIII. Conclusion

Acknowledgment

References

[1] The CIA World Factbook: Andorra, *The World Factbook*. Web Access, February 11, 2016.

[2] V.W. Zheng, Y. Zheng, Q. Yang *Collaborative Location and Activity Recommendations with GPS History Data* ACM, 2012.

[3] Y. Zheng, L. Zhang, X. Xie, W.Y. Ma *Mining Interesting Locations and Travel Sequences from GPS Trajectories* ACM, 2009.

[4] J. Bao, Y. Zheng, M. Mokbel *Location-Based and Preference-Aware Using Sparse Geo-Social Networking Data* ACM-SIGSPATIAL, 2012.

[5] J. Levandoski, M. Sarwat, A. Eldawy, M. Mokbel *LARS: A Location Aware Recommender System* ICDE, 2012.