УДК 004
МРНТИ 20.53
https://doi.org/10.55452/1998-6688-2023-20-1-45-53

Imed Eddine Semassel,*[1] Sadok Ben Yahia[2]

[1]Department of Computer Science, Faculty of Sciences of Tunis, El Manar University, Tunis, Tunisia
[2]Tallinn Univeristy of Technology, Tallinn, Estonia
*E-mail: imededdine.semassel@fst.utm.tn

MOBILITY EMBEDDING FROM CALL DATA RECORD USING WORD2VEC TO SUPPORT NETWORK WITH UNMANNED AERIAL VEHICLE

Abstract. Call Detail Records (CDRs) are records that provide information about phone conversations and text messages. CDR data has been proved in several studies to give useful information on people's mobility patterns and associations with fine-grained temporal and geographical characteristics. This paper proposes to embed the traces recorded in the CDRs to extract meaningful information. These latter provide insights about the location that may need support to cover or recover the network. After embedding the users' trajectories step, we use the embedding results to recommend the antennas with coordinates and support demand needed to a fleet of Unmanned Aerial Vehicle. Finally, we ended up with a capacitated vehicle routing problem that we solved using a Google open-source software named OR-Tools.

Key words: Mobility embedding, Word embedding, Word2Vec, CDR data.

Имед Эддин Семассель,*[1] Садок Бен Яхиа[2]

[1]Тунис жаратылыстану факультеті, Эль-Манар университетінің информатика бөлімі
[2]Таллин технологиялық университеті, Таллин қ., Эстония
*E-mail: imededdine.semassel@fst.utm.tn

ДРОН ЖЕЛІСІН ҚОЛДАУ ҮШІН WORD2VEC КӨМЕГІМЕН ҚОҢЫРАУ ДЕРЕКТЕРІН ЖАЗУДАН ҰТҚЫРЛЫҚТЫ ЕНГІЗУ

Аңдатпа. Қоңыраулар туралы мәліметтер жазбалары (CDR) телефон сөйлесулері мен мәтіндік хабарлар туралы ақпаратты беретін жазбалар болып табылады. CDR деректері адамдардың ұтқырлық үлгілері мен ұсақ түйіршікті уақытша және географиялық сипаттамалары бар ассоциациялар туралы пайдалы ақпарат беру үшін бірнеше зерттеулерде дәлелденді. Бұл құжат маңызды ақпаратты алу үшін CDR-де жазылған іздерді енгізуді ұсынады. Бұл соңғылар желіні жабу немесе қалпына келтіру үшін қолдау қажет болуы мүмкін орын туралы түсінік береді. Пайдаланушылардың траекториясының қадамын енгізгеннен кейін біз енгізу нәтижелерін координаттары бар антенналарды ұсыну және ұшқышсыз ұшу аппараты флотына қажетті қолдау көрсету үшін пайдаланамыз. Бұл, мақалада Google-дың OR-Tools деп аталатын ашық бастапқы бағдарламалық құралын пайдаланып, көлік құралын бағыттау мәселесін шештік.

Тірек сөздер: ұтқырлықты ендіру, Word ендіру, Word2Vec, CDR деректері.

Имед Еддине Семассел,*[1] Садок Бен Яхиа[2]

[1]Департамент компьютерных наук Тунисского факультета естественных наук,
Университет Эль-Манар, Тунис
[2]Таллинский университет технологий, г. Таллин, Эстония
*E-mail: imededdine.semassel@fst.utm.tn

ВСТРАИВАНИЕ МОБИЛЬНОСТИ ИЗ ЗАПИСИ ДАННЫХ О ВЫЗОВАХ С ИСПОЛЬЗОВАНИЕМ WORD2VEC ДЛЯ ПОДДЕРЖКИ СЕТИ С БЕСПИЛОТНЫМ ЛЕТАТЕЛЬНЫМ АППАРАТОМ

Аннотация. Записи сведений о вызовах (CDR) – это записи, содержащие информацию о телефонных разговорах и текстовых сообщениях. Некоторые исследования доказали, что данные CDR дают полезную информацию о моделях мобильности людей и связях с точными временными и географическими характеристиками. В данной статье предлагается встраивать трассировки, записанные в CDR, для извлечения значимой информации. Трассировки предоставляют информацию о местоположении, для которого может потребоваться поддержка для покрытия или восстановления сети. После внедрения траекторий пользователей мы используем обработанные результаты, для того чтобы рекомендовать антенны с координатами и запросом на поддержку, необходимые для парка беспилотных летательных аппаратов. В данной статье мы столкнулись с проблемой маршрутизации транспортных средств с вместимостью, которую мы решили с помощью программного обеспечения Google с открытым исходным кодом под названием OR-Tools.

Ключевые слова: встраивание мобильности, встраивание Word, Word2Vec, данные CDR.

Introduction

Mobile phones are used by over ninety percent of individuals in their everyday lives. These are moving, arguably leaving traces of their movements, which may create a lot of data and information. A Telecom Service Provider records the data of telephone calls or Short Message Services (SMS) that flows via such devices, referred to as Call Data Record in the rest of this document (CDR). The latter is a data structure that stores details on a certain telephonic activity.

Mobile phones are often regarded as the most popular and convenient mode of communication [1]. Researchers can investigate topics that primarily rely on CDR data to gain insights into the location of distinct populations and their evolution over time using mathematical modeling methodologies [7,10].

By nature, CDRs are generated in large volumes. One of their main advantages is that they can be viewed as a wide-area sensor network as long as they provide a statistically valid representation of the distribution of individuals in a given region and can be combined with other data sources to track extensive and diverse groups of people [2].

Furthermore, and despite this, they may be used to supplement self-reported data from interviews and questionnaires, which are time-consuming, labor-intensive, and difficult to predict dynamic changes. CDRs are one type of passively gathered data that is increasingly being used in research alongside other big data types [5].

Telecom providers collect a large number of CDRs regularly, from which it is feasible to extract additional information at little cost and develop valuable datasets. We may acquire helpful information from the study of this data [8], such as city planning, user profiling, disease spread patterns, natural disasters, and the occurrence and influence of social events.

The ability to predict population movement insights based on just a single mobile source, such as CDRs, is a very challenging task.

We investigate whether valuable information is encoded in embeddings that a sequence of locations includes. Our motivation comes from the Word2vec approaches, which recommends the following words based on the assumption that similar words appear in similar contexts. Traces recorded by the telecom providers can be modeled similarly and represented by the embedded vectors to predict the next population movement and, thus, predict the activity load in the locations.

Our proposition embeds the traces recorded in the CDRs to extract meaningful information. The latter provides insights about the location that may need support to cover or recover the network. After embedding

the users' trajectories, we recommend the antennas with coordinates and support demand needed for a fleet of Unmanned Aerial Vehicles (aka drones). We end up with a capacitated vehicle routing problem that we solve using a Google open-source tool named OR-Tools.

We structure the remainder of the paper as follows. First, in section 2, we present scrutiny of the related works that used the embedding technology. Then, Section 3 describes the CDRs dataset we have used. Then Section 4 describes our proposition, starting with the main idea and embedding step, passing by evaluating our work with a bunch of algorithms, arriving at a use case that uses a fleet of unmanned aerial vehicles to solve the obtained routing problem. Finally, we present some concluding remarks and future work that can improve results in Section 6.

Related work

In Natural Language Processing (NLP), word embedding technology refers to language models and feature learning methods. It is frequently utilized in text and seeks to learn the vectorized representation of words. Recently word embedding methods have been widely employed to learn dense vector representation of locations in mobility data. Many methods rely on the word2vec model [9] aiming to learn trajectory embedding vectors.

Zhu et al. [14] trained a skip-gram model word2vec [9] to create location embeddings, which were then used to comprehend the human mobility between urban places. Location2vec was proposed, which uses mobile cell stations as words taking advantage of the word2vec to learn embedding vectors considering the interaction between locations and moving objects.

In [4], the authors generated vector representations of locations, constructed based on how people move between them, their algorithm named motion-to-vector(Mot2vec) has two steps. First, the trajectories are preprocessed and converted into sequences of locations. Then, the location embeddings are constructed using a Skip-gram Word2vec model. The meaningful representation provided by Mot2vec defines a metric of similarity for evaluating locations connectivity and identifying mobility patterns among different categories of people.

The study of [11] offered an inference technique that employs embeddings to represent GPS trajectories using the Word2vec approach, which is then combined with multiple classification models to infer attributes such as gender, age, marital status, and if the user has children. The results of this study confirmed the efficacy of Word2vec outside of the NLP domain, demonstrating that the approach can predict demographics with accuracy.

In [12] they trained a GCN aided skip-gram model named GCN-L2V to learn representations that embed context information in human mobility as flow graph and spatial adjacency as a spatial graph. The approach could capture relationships among locations and provide a notion of semantic similarity in a spatial environment. Quantitative experiments and case studies demonstrated that the representations learned by GCN-L2V are effective and may be imported as features for down-streaming tasks.

In addition to these works, there exist other works as [13] that focus on the recommending POI (point of interest) with embedding vectors using word2vec.

We also rely on word2vec in our approach as all the mentioned works. The difference resides in the use of only the CDR data users as inputs, which is less informative. These latter are embedded to predict and recommend antennas that need support to cover or recover the charge in case saturation or antennas fail due to a disaster. These recommended antennas are fed into a vehicle routing system that manages the unmanned aerial vehicle fleet.

CDR Datasets

The anonymous CDRs data contains the time and location where every individual sends or receives a call and/or a text message. In this regard, as part of the D4D challenge in 2015, Sonatel-Orange Telecom made CDRs available to the research community. The idea is to use data from mobile phone calls to accelerate a country's socio-economic growth. Researchers can explore numerous areas directly influencing development variables using anonymized CDR data sets.

The considered data includes the time and location where users make a phone call are briefly described below:

Dataset 1: it represents the traffic per pair of antennas for the $1,666$ antennas (*aka* sites) hourly. The dataset contains monthly voice traffic between the sites, and data about monthly traffic text messages between the antennas (sites).

Dataset 2: It represents fine mobility data spread per user over an interval of two weeks for an entire year.

These data are unique at the individual level for about $300,000$ users randomly sampled.

Dataset 3: It summarizes the rounded one-year mobility data volume (country district level) at the individual level for approximately $150,000$ randomly sampled users.

Methodology

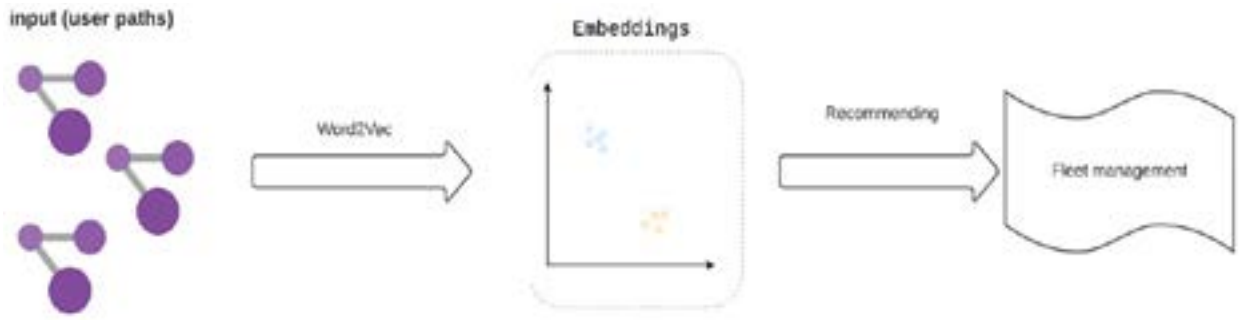
Our approach aims to automatically recommend antennas to a fleet of unmanned aerial vehicles based only on CDRs data. The latter contains records that include users' mobility. We leveraged the user's fine mobility data spread over two weeks to generate the user's mobility paths from it. Then we use these paths to determine antennas that may need support. Finally, a routing algorithm manages a fleet of unmanned aerial vehicles according to their capacities and the needs of the antennas.

The overall system model presented in our approach comprises three stages, and its workflow is flagged in Figure 1:

Corpus extractions step,

Embedding step, and

Recommending step.



We proceed as node2vec algorithm, it aims to generate vector representations of nodes on a graph through random walks to form a sequence of nodes as sentences in natural language processing (NLP), then pass it to the word2vec algorithm to produce the embedding, and capture some of the semantics of the nodes by placing semantically similar nodes close together in the embedding. First, we use the user's movements instead of random walks in our proposed approach. Then we pass these walks as input to the word2vec to get the embedding of the nodes. Finally, we use embedding to predict the next moves of users. This latter helps make early actions to support the antennas in case of antenna saturation or failure.

We applied our work to a subset of the data, which contains records of one district. This latter includes 71 antenna spots with high activity on average compared to the other districts of the DAKAR region.

Node2vec applies a random walks exploration on a given graph to generate many sequences of nodes. These latter are considered sentences, the inputs for the word2vec model that embeds the nodes. The word2vec model groups word with high context similarity in near neighborhoods. This method results in vector representations of nodes.

In our approach, we apply node2vec starting with random walks. The latter is replaced with the user's movements. We extract the path for each user, which results in a corpus of paths that can be interpreted as sentences. Indeed since this corpus is produced from users' movements, it preserves contextual information of areas. In other words, the corpus can be viewed as the description of the district through the user's movements.

After the corpus extraction, the word2vec model generates embeddings for each antenna. The embeddings are vector representations of the antennas that capture the semantics produced on the basis of the user's movements. Similar vectors are placed close together in the embedding space. Figure 2 shows the embeddings for each antenna.

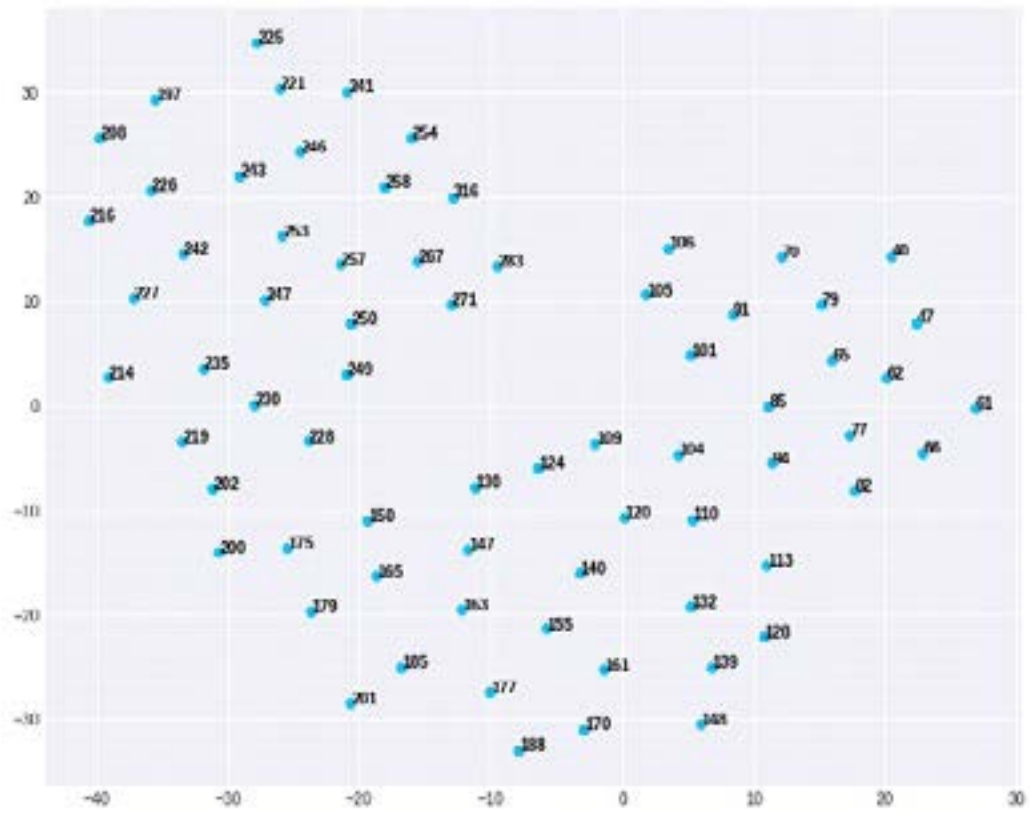


Figure 2 – The embeddings for each antenna

Finally, we use the embeddings results to recommend unmanned aerial vehicle fleet antennas. This latter supports the antennas to ensure the reliability of the communication services in case of saturation or help rescue operations in case of disasters. First, we compute the mean distances between antennas according to the user's trajectory. We considered the user mobility rate for each antenna visited since it can be viewed as how far the user moves away from the antenna. Then we train a recommending algorithm with the resulted dataset, and we predict the users' next moves. By counting this latter, we gain insights about the charge the antennas will face as presented in Figure 3.

Approach evaluation

To evaluate our approach, we used three accuracy metrics named MAE (Mean Absolute Error), MSE (Mean Squared Error), and RMSE (Root Mean Squared Error), with a list of recommendation algorithms presented in Table 1, and Table 2.

MAE represents the average of the absolute difference between the actual and predicted values. MSE represents the average of the squared difference between the original and predicted values. RMSE measures the standard deviation and it is the square root of MSE.

The algorithms showed better results with our approach, where we used the distances between embedded vectors of the antennas instead of using the geographical distances between antennas. Our approach outperforms all algorithms in both accuracy and time. Statistics shown in Tables 1, and Table 2 are sorted according to the RMSE metric.

Table 1 – The accuracy metrics for recommendation algorithms using the embedded vectors of the antennas

Algorithm	RMSE	MSE	MAE	fit_time	test_time
KNNBaseline	0.147769	0.021836	0.110582	9.917238	11.370682
KNNBasic	0.151011	0.022804	0.111156	7.823353	11.032049
SVDpp	0.176208	0.031049	0.142616	7.016650	0.456572

BaselineOnly	0.178941	0.032020	0.143688	0.231516	0.098518
SVD	0.180857	0.032709	0.146929	3.538417	0.300407
KNNWithMeans	0.182629	0.033353	0.137166	8.787129	11.479792
KNNWithZScore	0.183184	0.033556	0.137315	8.676554	10.777294
SlopeOne	0.193250	0.037345	0.147194	0.187910	0.202224
NormalPredictor	0.271933	0.073948	0.218916	0.0718180	0.166444
CoClustering	0.474721	0.225360	0.438905	2.116327	0.129506

Table 2 – The accuracy metrics for recommendation algorithms without using the embedded vectors of the antennas

Algorithm	RMSE	MSE	MAE	fit_time	test_time
KNNBasic	0.404790	0.163878	0.273863	14.272777	17.845253
KNNBaseline	0.4157248	0.172828	0.278298	14.757514	18.107223
SVDpp	0.464500	0.215766	0.335290	6.434196	0.632079
KNNWithMeans	0.535915	0.287218	0.369874	13.830763	17.630687
KNNWithZScore	0.571756	0.326909	0.376868	16.480092	20.197913
SlopeOne	0.621707	0.386525	0.425297	0.241426	0.317950
SVD	0.663268	0.439982	0.469696	2.962695	0.226513
BaselineOnly	0.677504	0.459033	0.500647	0.395443	0.199885
CoClustering	0.817951	0.669534	0.620654	3.089043	0.200401
NormalPredictor	1.060172	1.123993	0.823223	0.118616	0.236111

Next, since KNNBaseline has shown the best accuracy according to the metrics, we used it to predict users' next moves by counting all users' top-five rated antennas. The latter gave us insights about the charge the antennas will face, as presented in Figure 3. Finally, we use these obtained insights in a vehicle routing system that manages the fleet of unmanned aerial vehicles. The system considers each antenna's capacity to cover or recover the charge if saturation or antennas fail due to a disaster.

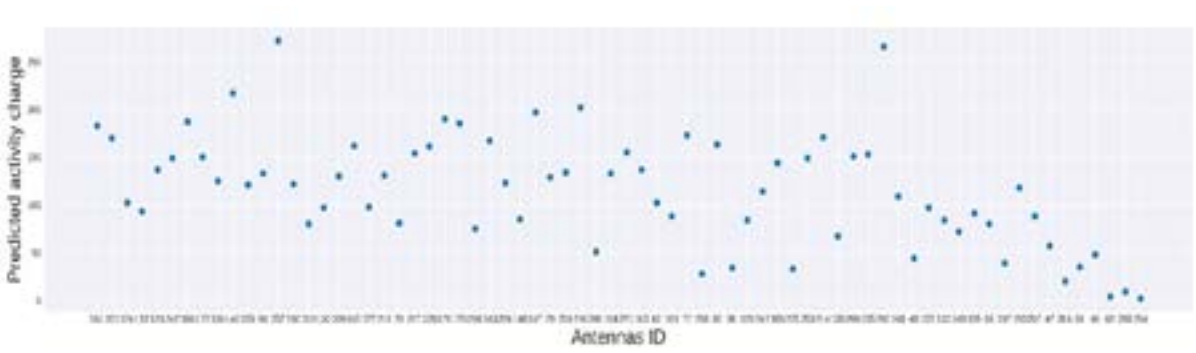


Figure 3 – Insights about the charge the antennas will face according to the users movements

Indeed we end up with a capacitated vehicle routing problem to solve this latter. We use the OR-Tools [6] of the google developers, which provides operations research software libraries and APIs for constraint optimization, linear optimization, and flow and graph algorithms.

The capacitated vehicle routing problem (CVRP) is a vehicle routing problem in which vehicles with a limited carrying capacity must pick up or deliver things at many locations. The items have a quantity, such as their weight or volume, while the vehicles have a carrying capacity. The challenge is to pick up or deliver the

products for the least cost while never exceeding the vehicles' capacity. In our case, the vehicles are the fleet of unmanned aerial vehicles. Each vehicle has a limited covering capacity, and the network's antennas are the locations that need support.

Before applying the solver to our example, we performed a constrained clustering [3] to divide the antennas into groups where the sum of the distances between each antenna and the centroid of its cluster are minimized, in addition to some constraints on cluster sizes. The cluster centroid is located where a base station for the fleet can be mounted. Figure 4 shows the antennas with a blue marker and the centroid of the antenna clusters with a red marker. In this example, we have applied clustering with a size of 10 antennas each. This latter can be changed according to the size of the fleet.

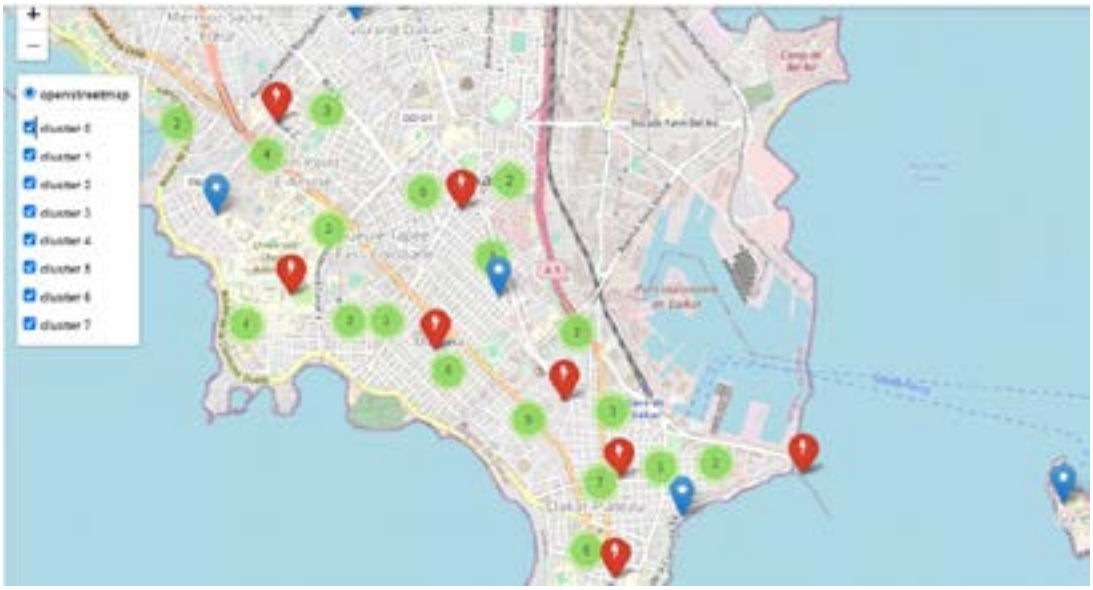


Figure 4 – Map with the clustered antennas and the centroid points where a base station for the fleet can be mounted

After the clustering step, we performed a simulation example for one of the clusters. We set four as the number of vehicles of the fleet, each vehicle has 15 as a maximum capacity, and the antennas demand as follows (140:2, 185:1, 124:1, 150:2, 147:4, 120:2, 163:4, 165:8, 179:8, 130:1). Zero stands for the Id of the base station. Figure 5 shows the results of the simulation and path each vehicle must follow to satisfy the antenna's demands.

```
Route for vehicle 0:
 0 Load(0) ->  179 Load(8) ->  185 Load(9) ->  163 Load(13) ->  0 Load(13)
Distance of the route: 3048m
Load of the route: 13

Route for vehicle 1:
 0 Load(0) ->  165 Load(8) ->  0 Load(8)
Distance of the route: 338m
Load of the route: 8

Route for vehicle 2:
 0 Load(0) ->  140 Load(2) ->  120 Load(4) ->  124 Load(5) ->  130 Load(6) ->  150 Load(8) ->  147 Load(12) ->  0 Load(12)
Distance of the route: 2483m
Load of the route: 12

Route for vehicle 3:
 0 Load(0) ->  0 Load(0)
Distance of the route: 0m
Load of the route: 0

Total distance of all routes: 4869m
Total load of all routes: 33
```

Figure 5 – The OR-TOOLS simulation results showing paths for each vehicle to satisfy the antennas requests

Conclusion

This paper presented a method that uses word2vec to embed Call Data Records traces. The key idea is to use the users movements as random walks and embed these latter with word2vec to get the embedding of the locations. Results of the embedding provided movement insights of users, which were helpful to manage a fleet of unmanned aerial vehicles. With this approach, CDR data can help make early actions to support the antennas in case of saturation or failure due to crowd events or disasters.

We can increase the precision of this work and add time feature, and we can use multiple data sources. For example, we can analyze traffic between antennas to determine the relationship between places based on the connectivity and the site's frequent users.

Furthermore, it can also be expanded by including events aspects to determine their relationship with user movement patterns.

REFERENCES

Association C. T. (2017, July). *How mobile phones are changing the developing world.* Retrieved from https://www.cta.tech/News/Blog/Articles/2015/July/How-Mobile-Phones-Are-Changing-the-Developing-Worl.aspx.

Bianchi F. M., Scardapane, S., Uncini, A., Rizzi, A., & Sadeghian, A. (2015). Prediction of telephone calls load using Echo State Network with exogenous variables. *Neural Networks*, *71*, 204–213. https://doi.org/https://doi.org/10.1016/j.neunet.2015.08.010.

Bradley P.S., Bennett K.P. & Demiriz A. (2000). Constrained k-means clustering. *Microsoft Research, Redmond*, *20*(0), 0.

Crivellari A. & Beinat E. (2019). From motion activity to geo-embeddings: Generating and exploring vector representations of locations, traces and visitors through large-scale mobility data. *ISPRS International Journal of Geo-Information*, *8*(3), 134.

Cuzzocrea A., Ferri F. & Grifoni P. (2018). Intelligent Sensor Data Fusion for Supporting Advanced Smart Health Processes. In L. Barolli & O. Terzo (Eds.), *Complex, Intelligent, and Software Intensive Systems* (Vol. 611, pp. 361–370). https://doi.org/10.1007/978-3-319-61566-0_33

OR-tools. Retrieved from https://developers.google.com/optimization

Gore R., Wozny P., Dignum F. P. M., Shults F. L. van Burken C. B. & Royakkers, L. (2019). A Value Sensitive ABM of the Refugee Crisis in the Netherlands. *Proceeding 2019 Spring Simulation Conference (SpringSim)*, 1–12.

Louail T., Lenormand M., Ros O.G. C., Picornell M., Herranz R., Frias-Martinez E., … Barthelemy M. (2015). From mobile phone data to the spatial structure of cities. *Scientific Reports*, *4*. https://doi.org/https://doi.org/10.1038/srep05276.

Mikolov T., Chen K., Corrado G. & Dean J. (2013). Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301.3781*.

Mobile policy handbook: an insider's guide to the issues. (2017). Retrieved from https://www.gsma.com/mena/wp-content/uploads/2018/10/Mobile_Policy_Handbook_2017_EN.pdf

Solomon A., Bar A., Yanai C., Shapira B. & Rokach, L. (2018). Predict demographic information using word2vec on spatial trajectories. *Proceedings of the 26th Conference on User Modeling, Adaptation and Personalization*, 331–339.

Tian C., Zhang Y. & Weng Z. (2021). Learning Large-scale Location Embedding From Human Mobility Trajectories with Graphs. *CoRR*, *abs/2103.00483*. Retrieved from https://arxiv.org/abs/2103.00483

Zhou N., Zhao W. X., Zhang X., Wen J.-R. & Wang S. (2016). A general multi-context embedding model for mining human trajectory data. *IEEE Transactions on Knowledge and Data Engineering*, *28*(8), 1945–1958.

Zhu M., Chen W., Xia J., Ma Y., Zhang Y., Luo Y. … Liu L. (2019). Location2vec: a situation-aware representation for visual exploration of urban locations. *IEEE Transactions on Intelligent Transportation Systems*, *20*(10), 3981–3990.

Information about authors

Imed Eddine Semassel

PhD student, Department of Computer Science, Faculty of Sciences of Tunis, El Manar University, Tunis, Tunisia

ORCID ID: 0000-0002-9119-6867

E-mail: imededdine.semassel@fst.utm.tn

Sadok Ben Yahia
Professor, Tallinn Univeristy of Technology, Tallinn, Estonia
ORCID ID: 0000-0001-8939-8948
E-mail: sadok.ben@taltech.ee

Авторлар туралы мәліметтер

Имед Еддине Семассел
Докторант, Тунис жаратылыстану факультеті, Эль-Манар университетінің информатика бөлімі, Тунис
ORCID ID: 0000-0002-9119-6867
E-mail: imededdine.semassel@fst.utm.tn

Садок Бен Яхиа
Профессор, Таллин технологиялық университеті, Таллин қ., Эстония
ORCID ID: 0000-0001-8939-8948
E-mail: sadok.ben@taltech.ee

Информация об авторах

Имед Еддине Семассел
Докторант, Департамент компьютерных наук Тунисского факультета естественных наук, Университет Эль-Манар, Тунис
ORCID ID: 0000-0002-9119-6867
E-mail: imededdine.semassel@fst.utm.tn

Садок Бен Яхиа
Профессор**,** Таллинский университет технологий, г. Таллин, Эстония
ORCID ID: 0000-0001-8939-8948
E-mail: sadok.ben@taltech.ee