

Parallel Accurate Localization from Cellular Network

Chao Wu^(✉), Bin Xu, and Qi Li

Department of Computer Science and Technology, Tsinghua University, Beijing, China
ariesnix93@gmail.com, xubin@tsinghua.edu.cn,
zhongguoliqi@163.com

Abstract. Localization from cellular network (LCN) is of great importance for mobile operator. Different from GPS, LCN utilizes operator's network infrastructure to identify the location of mobile devices. The advantage of LCN is that operator can calculate all mobile devices' location according to the mobile big data, without each device reporting its location got from GPS. The challenges of LCN are localization accuracy and computing efficiency. In this paper, we propose three methods of localization by analyzing mobile signal strength data collected by cellular network per 480 milliseconds. Using an optimized propagation model which considers the penetration loss through buildings, we get fairly high localization accuracy. To improve the localization efficiency, a parallel method based on Spark is proposed to process and analyze mobile big data. Through the experiments of tracking people in real cellular network, our methods reach an accuracy of 100 meter (mean error), which exceeds the international standard (125 meter by FCC). The parallel method also gets a significant efficiency promotion, which reduce 97% time cost than serial methods.

1 Introduction

Localization from cellular network (LCN) utilizes operator's network infrastructure to identify the location of mobile devices, which is a kind of network-based localization technology. Network-based and handset-based localization technologies are two main categories of wireless location. Handset-based technology, such as GPS have a relatively high degree of accuracy, but embedding a GPS receiver into mobile devices leads to increased cost, size, and battery consumption. Network-based technology, on the other hand, measure the signals transmitted from a mobile device and relay them to a central site for further processing and data fusion to provide an estimate of the mobile device's location. The significant advantage of network-based technology is that the mobile device is not involved in the location-finding process, which makes it possible to calculate all mobile devices' location according to the mobile big data.

A number of ranging techniques are developed to get the location of a mobile device, including received signal strength (RSS), time of arrival (TOA), angle of arrival (AOA), and so on. In our research, we try to localize people through a kind of RSS data which is measurement report (MR) data. MR data is collected by digital cell mobile communications network sent by user equipment. The first challenge is how to estimate people's real location from MR data accurately. In addition, the MR data is

generated for each mobile device per 480ms, which is huge amount in the whole cellular network. The challenge is how to process and analyze the mobile big data efficiently.

In this paper, we propose three localization methods. The contributions of our study are as follows:

1. We propose two methods which consider the penetration loss through buildings and use more optimized parameters. Through the experiments of tracking people in real cellular network, our methods reach an accuracy of 100 meter (mean error).
2. We propose a parallel method based on Spark [1], which can rapidly process and analyze mobile big data. Comparing with serial methods, it gets a forty times promotion in efficiency.

The rest of paper is organized as follows: In section 2, we describe related work on localization technology. In section 3, we give an overview of the mobile big data, and describe some important characteristics of the data. Section 4 describes the serial methods we propose for localization, and the parallel method we design in detail. Section 5 shows experimental results applied to our dataset, and gives detail analysis on the result. Finally, we conclude this paper in Section 6.

2 Related Work

In this section, we will show some related work on localization technology.

Mobility prediction and localization are tightly coupled with location-based applications, such as, navigation [2], instant messaging [3], emergency rescue [4], and many other safety and security services [5] [6]. Many of the above applications need precise tracking of mobile terminal within cells to provide an adaptable quality of service.

The most important mobile location techniques are propagation time and signal strength techniques. Propagation time-based techniques (e.g. TOA [7], TDOA [8] [9], E-OTD [10]) rely on time measurement. However, the disadvantage of these techniques is that the accuracy of the estimated position depends mainly on the number of measurements and on the geometric configuration, which is even worse in urban environments.

Signal strength-based techniques instead are based on Received Signal Strength Indication (RSSI), which measures signal attenuation, assuming free space propagation and omnidirectional antennas. Exploiting this assumption mobile antenna location problem is reduced to the well-known triangulation position problem which is similar to time-based and angular approaches. The disadvantage of these techniques is that free space propagation assumption does not work because of multipath propagation and shadowing, leading to complex signal shapes. There are four kinds of physical phenomena influencing radio propagation: reflection, diffraction, penetration, and scattering. To solve these problems, many propagation models have been developed, such as cost231-hata model [11] [12], SDR model, cost231-walfish-ikegami model

[13], and so on. How to use those theoretical model to meet the real environment and improve their accuracy, is what we do in our research.

Many recent works are aimed at increasing the localization accuracy of RSSI triangulation. One interesting work [14] treats the problem of mobility in ATM network. It develops a hierarchical user mobility model that closely represents the movement behavior of a mobile user, and uses pattern matching and Kalman filtering, yielding to an accurate location prediction algorithm. Another work [15] proposes two algorithms for real-time tracking, location, and dynamic motion of a mobile station in a cellular network. This method is based on pre-filtering and two Kalman filters. [16] [17] propose a mobility model built on a dynamic linear system driven by a discrete command process that was originally developed for tracking maneuvering targets in tactical weapons system. Different from these works, our research base on RSSI data from cellular network, which has a larger data scale and wider data features.

3 Data Description

Measurement report (MR) data is a type of RSS data user equipment (UE) send to TD-LTE digital cell mobile communications network every 480ms, which can be used to evaluate and optimize the cellular network.

MR data contains many features. Each feature has its special meaning. Table 1 lists those features we use in our research.

Table 1. Features of MR data

Feature	Description
CellId	Unique id of the serving cellular.
MmeCode	Triple generated from Imsi, represents a unique people.
MmeGroupId	
MmeUeSlapId	
TimeStamp	The generation time of this record.
ScRSRP	Corresponds to the signal strength that UE receives from the serving cellular.
NcRSRP	Corresponds to the signal strength that UE receives from a neighborhood cellular of the serving cellular.
ScTadv	The signal propagation time from the serving cellular to UE.
ScEarfcn	The carrier service number of the serving cellular.
ScPci	The physical identification code of the serving cellular.
NcEarfcn	The carrier service number of a neighborhood cellular of the serving cellular.
NcPci	The physical identification code of a neighborhood cellular of the serving cellular.

4 Localization Methods

In this section, we describe how we use MR data defined in section 3 to calculate people's real location. We propose three methods: baseline method, optimized method and parallel method.

4.1 Data Preprocessing

The original MR data contains a huge number of gzip files. Each gzip file is a compression of one xml file. In addition, an xml contains lots of features, while the useful features listed in table 1 are only a quarter. It is very meaningful and important to extract useful features from original MR data in a short time.

We use multiple threads to decompress and parse data. Each thread will extract useful features from the compressed file, and parse them into a buffered data queue. The buffered data queue will store and flush batch data into disk regularly. The workflow is shown in figure 1.

Each UE (user equipment) is identified by his id (MmeCode, MmeGroupId, MmeUeSlapId). Combining with the geographic information of cellular towers, we preprocess data into a specified format. The formatted data is shown in figure 2. From this data, how to calculate the location?

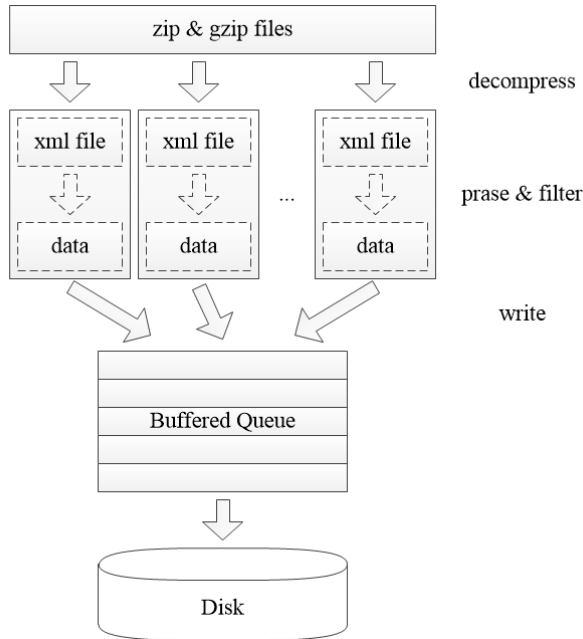


Fig. 1. Workflow of extracting features

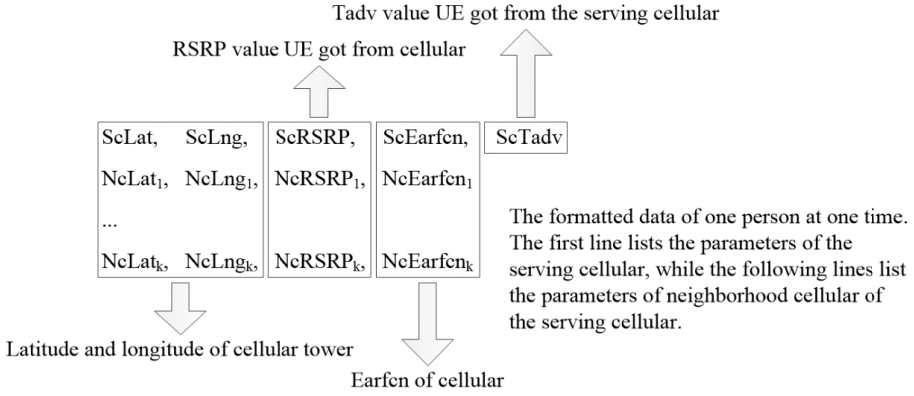


Fig. 2. Formatted data of one person

4.2 Calculating Distance

At first, we calculate the distance between UE and cellular tower.

As described in section 3, RSRP corresponds to the signal strength, while the signal strength of electromagnetic wave is losing energy through its propagation from cellular tower to UE. This energy loss can be calculated by the following formula:

$$\text{loss} = \text{prs} + \text{gain} - \text{rsrp} - \text{shadowLoss} \quad (1)$$

In this formula, prs is the transmit signal strength of cellular tower, while gain is the signal gain from the antenna. When electromagnetic wave passes through buildings, it losses energy. This is a variable determined by the distribution of buildings surround the cellular tower. We use the Gaussian distribution to represent it.

The energy loss is related by the distance between cellular tower and UE, the relationship is described by the following equations:

$$\text{loss} = L_u - a(H_{ue}) + C_m \quad (2)$$

$$L_u = 46.3 + 33.9 * \lg(f) - 13.82 * \lg(H_{eNB}) + (44.9 - 6.55 * \lg(H_{eNB})) * \lg(d) \quad (3)$$

$$a(H_{ue}) = (1.1 * \lg(f) - 0.7) * H_{UE} - (1.56 * \lg(f) - 0.8) \quad (4)$$

In these equations, f is the operating frequency of cellular, which can be calculated by Earfcn. Table 2 lists the calculation of Earfcn. H_{eNB} is the height of cellular tower, H_{UE} is the height of UE, C_m is an experience value depending on urban area type, d is the distance from cellular tower to UE, which need calculation.

Table 2. The calculation of f by Earfcn

Start frequency	Offset	Range of Earfcn	f
2570	37750	37750 - 38249	$f = \text{startFrequency} + 0.1 * (\text{earfcn} - \text{offset})$
1880	38250	38250 - 38649	
2300	38650	38650 - 39649	

Through equations (1), (2), (3) and (4), we calculate several distances from the formatted data. These distances are shown in figure 3.

T_{adv} also corresponds to the distance between serving cellular and UE. We calculate it by:

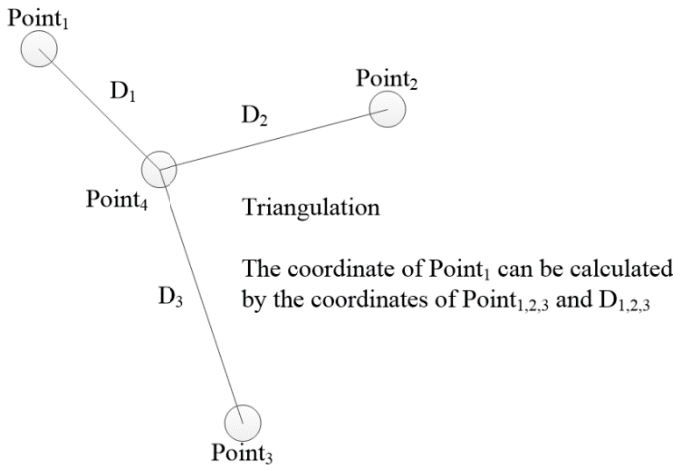
$$D_{tadv} = (scTadv + 0.5) * 78.12$$

D_{ScRSRP}	ScLat, ScLng, ScRSRP, ScEarfcn, ScTadv
$D_{NcRSRP1}$	NcLat ₁ , NcLng ₁ , NcRSRP ₁ , NcEarfcn ₁
...	...
$D_{NcRSRPk}$	NcLat _k , NcLng _k , NcRSRP _k , NcEarfcn _k

Fig. 3. Distances calculated from the formatted data

4.3 Baseline – Triangulation Localization

Since we know the location of cellular towers, after we calculate the distances between UE and cellular towers, we use triangulation to calculate UE's location as a baseline. The schematic diagram of triangulation is shown in figure 4. Formula 1 is the detail calculation method, the (Lat, Lng) pair which makes formula 1 minimal is the location of UE.

**Fig. 4.** The schematic diagram of triangulation

$$\left\| \begin{array}{l} \sqrt{(Lat - scLat)^2 + (Lng - scLng)^2 - D_{adv}} \\ \sqrt{(Lat - scLat)^2 + (Lng - scLng)^2 - D_{scRSRP}} \\ \sqrt{(Lat - ncLat_1)^2 + (Lng - ncLng_1)^2 - D_{ncRSRP1}} \\ \dots \\ \sqrt{(Lat - ncLat_k)^2 + (Lng - ncLng_k)^2 - D_{ncRSRPk}} \end{array} \right\|$$

Formula 1.

4.4 LocatePlus – Optimized Localization

In this subsection, we propose a method named LocatePlus to improve the baseline.

To let our localization result more accurate, we use past N consecutive points to optimize current time point.

People's trace is made by a set of consecutive points, if the time interval between adjacent points is very short, e.g. 5 seconds. We can use the past N points' location to predict $N + 1$ point's location.

The result of baseline in section 4.3 can be described as:

$$(T_0, S_0), (T_1, S_1), \dots, (T_n, S_n)$$

T is time, S is location (latitude & longitude pair).

We can calculate the speed in each time:

$$V_i = \frac{S_i - S_{i-1}}{T_i - T_{i-1}}$$

The statistics value of V :

$$\begin{aligned} \hat{\mu} &= \frac{1}{N} \sum_{i=1}^N V_i \\ \hat{\sigma}^2 &= \frac{1}{N-1} \sum_{i=1}^N (V_i - \hat{\mu})^2 \\ \hat{\sigma}_1^2 &= \frac{1}{N-1} \sum_{i=1}^{N-1} (V_i - \hat{\mu})(V_{i+1} - \hat{\mu}) \end{aligned}$$

$$\hat{\alpha} = \begin{cases} 1 & \text{if } \hat{\sigma} = 0 \\ \max \left\{ 0, \frac{\hat{\sigma}_1^2}{\hat{\sigma}^2} \right\} & \text{otherwise} \end{cases}$$

If we get four statistics values from point 1 to point N, and the speed on T_n , then the speed on T_{n+1} is:

$$V_{n+1} = \hat{\alpha}V_n + (1 - \hat{\alpha})\hat{\mu} + \hat{\alpha}\sqrt{1 - \hat{\alpha}^2}N$$

So the distance between point N and point N+1:

$$D_v = V_{n+1}(T_{n+1} - T_n)$$

Formula 2 is the detail calculation method of LocatePlus. The (Lat, Lng) pair which makes formula 2 minimal is the location of UE.

$$\left\| \begin{array}{c} \sqrt{(Lat - Lat_n)^2 + (Lng - Lng_n)^2 - D_v} \\ \sqrt{(Lat - scLat)^2 + (Lng - scLng)^2 - D_{tadv}} \\ \sqrt{(Lat - scLat)^2 + (Lng - scLng)^2 - D_{scRSRP}} \\ \sqrt{(Lat - ncLat_1)^2 + (Lng - ncLng_1)^2 - D_{ncRSRP1}} \\ \dots \\ \sqrt{(Lat - ncLat_k)^2 + (Lng - ncLng_k)^2 - D_{ncRSRPk}} \end{array} \right\|$$

Formula 2.

4.5 LocatePara - Parallelization

To improve the processing speed on large scale data, we implement a parallel localization algorithm on Spark, which is a fast and general cluster computing system. Because LocatePlus is very suitable for parallelization, we encapsulate them in two functions, which can both be used for parallel computing and serial computing.

In our algorithm, the word **op** means executing one of the following operations:

map, apply a function to all elements.

groupByKey, group the values for each key into a single sequence.

Input: Formatted data

Output: The location in each time

function LocatePara

1. **op** map:
 2. **Input:** rsrp, height, earfcn, tadv, longitude, latitude, timeStamp, userId.
 3. $k1 = (\text{userId}, \text{timeStamp})$
 4. $v1 = \text{array}(\text{rsrp}, \text{height}, \text{earfcn}, \text{tadv}, \text{longitude}, \text{latitude})$
 5. **Output:** ($k1$, $v1$)
 - 6.
 7. **op** groupByKey:
 8. **Input:** ($k1$, $v1$)
 9. **Output:** ($k1$, iterator ($v1$))
 - 10.
 11. **op** map:
 12. **Input:** ($k1$, iterator ($v1$))
 13. $k2 = k1.\text{userId}$
 14. $v2 = (\text{iterator}(\text{v1}), \text{Baseline}(\text{iterator}(\text{v1})), k1.\text{timeStamp})$
 15. **Output:** ($k2$, $v2$)
 - 16.
 17. **op** groupByKey:
 18. **Input:** ($k2$, $v2$)
 19. **Output:** ($k2$, iterator ($v2$))
 - 20.
 21. **op** map:
 22. **Input:** ($k2$, iterator ($v2$))
 23. sort (iterator ($v2$)) by timestamp
 24. $v3 = \text{locatePlus}(\text{iterator}(\text{v2}))$
 25. **Output:** ($k2$, $v3$)
-

Input: Formatted data at time T_i

Output: The location in T_i

function BaseLine

1. **for** Each line in the formatted data except the first
 2. Compute $D_{ncRSRPj}$
 3. **end**
 4. Compute Lat and Lng
 5. Define (Lat, Lng) pair S_i
 6. Return (T_i, S_i)
-

Input: Formatted data at time T_{i+11} , Past 11 results of Baseline

Output: The location in T_{i+11}

function LocatePlus

1. **for** Each two adjacent points (T_{i+j-1}, S_{i+j-1}) and (T_{i+j}, S_{i+j})
 2. Compute speed V_{i+j}
 3. **end**
 4. Compute $\hat{\alpha}$ and $\hat{\mu}$ by $V_{i+1}, V_{i+2}, \dots, V_{i+10}$
 5. Compute D_v by $\hat{\alpha}, \hat{\mu}, T_{i+10}$ and T_{i+11}
 6. Compute D_{scRSRP} and D_{tadv} by the parameters of the first line in formatted data
 7. **for** Each line in the formatted data except the first
 8. Compute $D_{ncRSRPj}$
 9. **end**
 10. Compute Lat and Lng
 11. Define (Lat, Lng) pair S_{i+11}
 12. Return (T_{i+11}, S_{i+11})
-

5 Experiments

In this section, we discuss the accuracy and efficiency of our localization methods.

5.1 Data Preprocessing

The original MR data we have is a 50GB zip file compressed by gzip files. The environment we deployed for data preprocessing is a computer which contains a 3.4GHz quad-core CPU with eight threads, 16GB RAM and a 7200rpm hard disk. The relationship between time consumption and the number of threads is shown in figure 5.

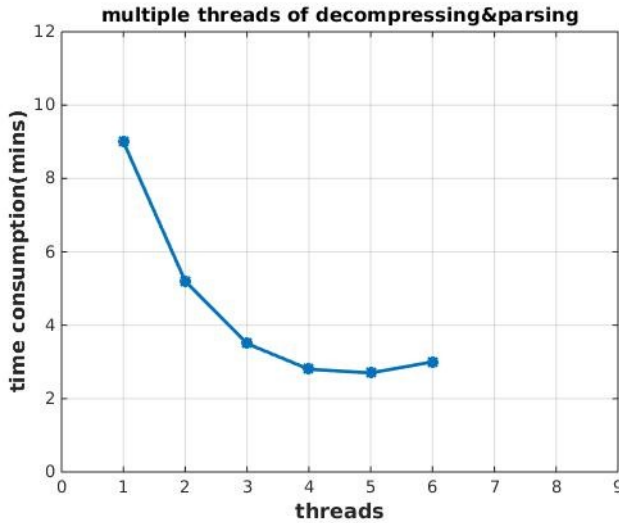


Fig. 5. The relationship between time consumption and the number of threads

It is clear that our preprocessing part achieves huge efficiency improvement with the increasing of thread, but more than four threads bring no more contribution because the reading speed of disk is low when concurrent accessing and little files extracting.

5.2 Accuracy

To get the test data, a user with a mobile phone is asked to walk around in the city, while the cellular towers record his MR data. At the same time, his mobile phone records his trace through GPS, which is the ground truth of the trace.

In subsection 4.2, the aforementioned RSRP corresponds to signal strength, this relationship is shown in table 3.

We use Baseline and LocatePlus to calculate locations, and compare with ground truth location. The statistical result is shown in table 4. The distribution is shown in table 5 and figure 5. We display the locations on google map, shown in figure 6.

Enhanced 911 is a system used in North America that links emergency callers with the appropriate public resources. E911 rules require wireless service providers to provide the latitude and longitude of the caller which must be accurate to within 50 to 300 meters depending upon the type of location technology used. [18]

From table 4, we can see that the accuracy of our method is about 100 meters, which is a big improvement in cellular network localization area. The calculated trace can almost reflect the real moving tendency.

From table 5 and figure 5, we see that we can estimate people's location with deviation in 300m, and estimate people's location with deviation in 150m with 75% confidence.

Table 3. The relationship between RSRP and real signal strength

RSRP	Range of signal strength (dBm)	Value we use (average)
0	$RSRP < -140$	-140
1	$-140 < RSRP < -139$	-139.5
2	$-139 < RSRP < -138$	-138.5
...	...	...
95	$-46 < RSRP < -45$	-45.5
96	$-45 < RSRP < -44$	-44.5
97	> -44	-44

Table 4. The statistical result of accuracy

Method	Points	Mean deviation	Standard deviation
Baseline	639	109.2m	73.0m
LocatePlus		97.7m	71.3m

Table 5. The distribution of accuracy

Accuracy	$\leq 50m$	$\leq 100m$	$\leq 150m$	$\leq 200m$	$\leq 250m$	$\leq 300m$
Points	164	309	469	549	606	623
Percentage	26%	49%	75%	87%	96%	99%

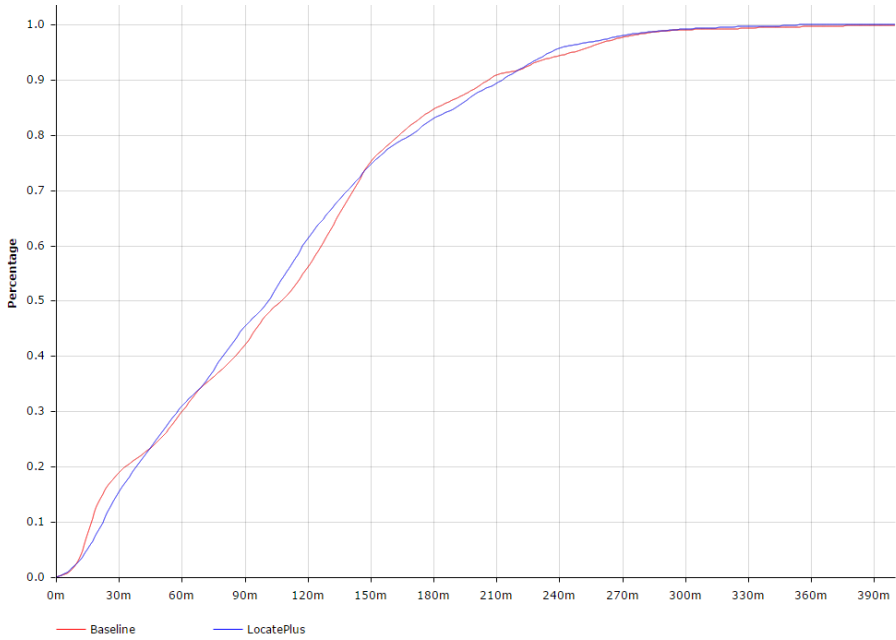


Fig. 6. The distribution of accuracy

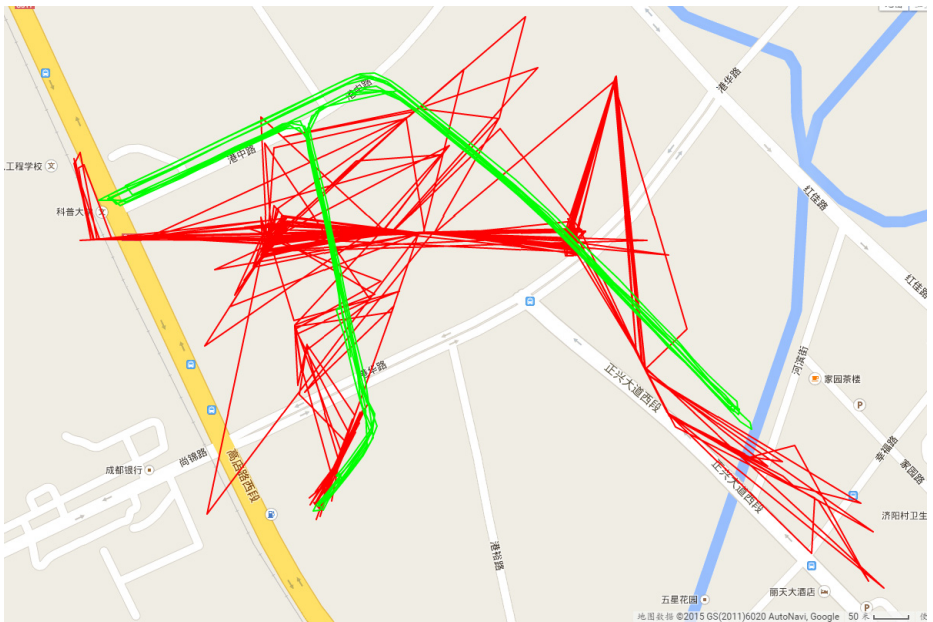


Fig. 7. Trace on google map: Green line shows the real locations, red line shows the locations we calculate

5.3 Efficiency

After we verify the accuracy of our methods, we turn to verify the efficiency.

The parallel platform we built is a cluster with six nodes. Each node owns a 3.4GHz quad-core CPU with eight threads and 32GB RAM. The version of Hadoop we deployed is 2.4.1. The version of Spark we deployed is 1.2.0, and the deployed mode of Spark is Spark-on-YARN.

We do experiment on different number of records. The time cost is shown in table 6 and figure 7.

The speedup of LocatePara is 42 when the number of records is 10 million, which means the parallel algorithm reduce 97% time cost than serial algorithm. It is a near linear relationship between time and records when using LocatePara. When the number of records is more than 20 million, the execution time of LocatePlus is more than 24 hours. LocatePlus runs so long time that we can't wait for its termination. This is why the character "-" appears in the table.

The result shows that it is feasible to use a cluster with enough number of nodes to do large-scale computing of localization within a relatively short time.

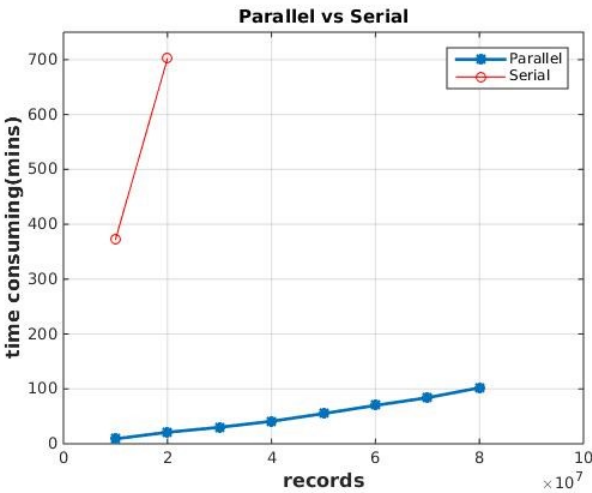


Fig. 8. Time cost comparison

Table 6. Time cost comparison of LocatePara and LocatePlus

Records (M for million)	LocatePara	LocatePlus
10M	9mins	6.2hours
20M	21mins	11.7hours
30M	30mins	-
40M	41mins	-
50M	55mins	-
60M	70mins	-
70M	84mins	-
80M	102mins	-

6 Conclusions

In this paper, we present three localization methods using mobile signal strength data in cellular network. Our methods consider the penetration loss through buildings and optimize the propagation model. Furthermore, we propose a parallel method based on Spark to improve the localization efficiency. We applied our methods through tracking people in real cellular network, the accuracy of our method reaches 100 meter (mean error), which exceeds the international standard (125 meter by FCC). The parallel method also gets a significant efficiency promotion, which reduce 95% time cost than before.

For future work, one possibility to extend the work is to continue improving localization accuracy. The other possibility is to extend 2D localization into 3D, to find the floor that people is in combining with building information.

Acknowledgement. This work is supported by China National Science Foundation under grant No.61170212, China National High-Tech Project (863) under grant No.2013AA01A607, Ministry of Education-China Mobile Research Fund under grant No.MCM20130381, and Tsinghua University Initiative Scientific Research Program (No. 20131089190). Beijing Key Lab of Networked Multimedia also supports our research work.

References

1. Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. In: HotCloud 2010, June 2010
2. Bejuri, W., Yaakob, W.M., et al.: Ubiquitous Positioning: A Taxonomy for Location Determination on Mobile Navigation System. *Signal & Image Processing: An International Journal (SIPIJ)* **2**(1), 24–34 (2011)
3. Zhao, S., Xia, F., et al.: MobiMsg: a resource-efficient location-based mobile instant messaging system. In: 2012 Second International Conference on Cloud and Green Computing, pp. 466–471 (2012)
4. Villemaud, G., Decroze, C., et al.: Dual-band printed dipole antenna array for an emergency rescue system based on cellular-phone localization. *Microwave and Optical Technology Letters* **42**(3), 249–253 (2004)
5. Marmasse, N., Schmandt, C.: User-centered location awareness. *IEEE Computer* **37**(10), 110–111 (2004)
6. Ardagna, C.A., Cremonini, M., et al.: Supporting location-based conditions in access control policies. In: Proc. of the ACM Symposium on Information, Computer and Communications Security (ASIACCS 2006), Taipei, Taiwan, March 2006
7. Drane, C., Macnaughtan, M., Scott, C.: Positioning gsm telephones. *IEEE Comm. Mag.* **46**, 46–55 (1998)
8. Caffery, J.J., Stüber, G.L.: Overview of radiolocation in cdma cellular systems. *IEEE Communications Magazine*, April 1998
9. Xu, B., Yu, R., Sun, G., Yang, Z.: Whistle: synchronization-free TDOA for localization. In: 2011 31st International Conference on Distributed Computing Systems, pp 760–769
10. Spirito, M.: On the accuracy of cellular mobile station location estimation. *IEEE Trans. Veh. Technol.* **50**, 3674–3685 (2001)

11. Okumura, Y., Ohmori, E., et al.: Okumura-hata propagation prediction model for uhf range, in the prediction methods for the terrestrial land mobile service in the vhf and uhf bands. ITU-R Recommendation P. 529–2. ITU, Geneva (1995)
12. Damosso, E.: Cost 231: (digital mobile radio towards future generation systems). Final report, European Comission, Bruxelles (1999)
13. Walfisch, J., Bertoni, H.L.: A theoretical model of UHF propagation in urban environments. *IEEE Trans. on Antennas and Propagation* **36**(12), 1788–1796 (1988)
14. Anisetti, M., Bellandi, V., Damiani, E., Reale, S.: Accurate localization and tracking of mobile terminals. In: 2th International Conference on Wireless Communications, Networking and Mobile Computing, Whuan, Cina (2006)
15. Anisetti, M., Bellandi, V., Damiani, E., Reale, S.: Localization and tracking of mobile antenna in urban environment. In: International Symposium on Telecommunications, pp. 353–358, September 2005
16. Singer, R.A.: Estimating optical tracking filter performance for manned maneuvering targets. *IEEE Trans. Aerosp. Electron. Syst.* (1970)
17. Moose, R.L., Vanlandingham, H.F.: Modeling and estimation for tracking maneuvering targets. *IEEE Trans. Aerosp. Electron. Syst.* (1979)
18. Federal Communications Commission. 911 Wireless Services, December 5, 2014. <http://www.fcc.gov/guides/wireless-911-services>