Procedia Computer Science

Volume 53, 2015, Pages 131–140

2015 INNS Conference on Big Data

An Investigation of parallel road map inference from Big GPS Traces Data

Wiam Elleuch[1], Ali Wali[1], and Adel M. Alimi[1]

REGIM-Lab: Research Groups in Intelligent Machines, National Engineering School of Sfax (ENIS), BP 1173, Sfax, 3038, Tunisia

wiam.elleuch.tn@ieee.org, ali.wali@ieee.org, adel.alimi@ieee.org

Abstract

With the increased use of GPS sensors in several everyday devices, persons trip data are becoming very abundant. Many opportunities for exploration of the wealth GPS data and in this paper, we inferred, the geometry of road maps in Tunisia and the connectivity between them. This phenomenon is known as map generation and also map inference procedure. For that, we gathered big GPS data from about ten thousands of vehicles equipped with GPS receivers and circulating in Tunisia, which does not have a road map like other developing countries. We collected a big database with approximately 100 gigabytes. After preprocessing it ,we were obliged to partition data in order to facilitate handling an unstructured database with a such size. In fact, we used for that K-means with its sequential mode and the parallel mode based on Mapreduce, which is one of the most famous proposed solution to analyse the rapidly growing data. The proposed parallel k-means algorithm was tested with our GPS data and the results are efficient in processing large datasets. It is a parallel data processing tool which is gathering significant importance from industry and academia especially with appearance of a new term to describe massive datasets having large-volume, high-complexity and growing data from different sources, "big data".

Keywords: road generation ,road map, GPS big data, Mapreduce, K-means, clustering, map matching

1 Introduction

The processing of big data becomes the center of interest in science and industry. The sources of these data are countless. We can cite for example social networks, emails, audio and video sequences, pictures, log files, GPS frames, search queries,scientific papers, health records and images, sensors data, mobile devices and all the applications installed. All those data constitute significant values in their domain. Therefore, they should be stored in huge databases which grow continuously, in order to be analysed and managed. The emission of data can be very large in each second especially with the widespread social networks such as Facebook and Twitter. So, the amount of data created nowadays in the age of information increases rapidly than in

Selection and peer-review under responsibility of the Scientific Programme Committee of INNS-BigData2015

doi:10.1016/j.procs.2015.07.287

the previous years. For example, until 2003, information collected reached approximately 5 exabytes. However, today this quantity of data can be easily created in a couple of days [2].

According to the International Data Corporation (IDC)[6], it is predictable that between 2005 and 2020, the digital universes will increase by about 300 times, from 130 exabytes to 40 zettabytes, which means that in 2020 every person even children, will generate more than 5200 gigabytes. Also, it is estimated that between 2012 and 2020, the digital data in the world will approximately double every couple of years. With the explosive growth of global data, industries become concerned in big data, and several government agencies declared some plans to motivate and encourage big data research [1].

When we hear about big data term, we think that the encountered challenge is only the *volume*. However, challenges of big data are several and characterised by four aspects (4V):

- Volume: The first criterion is the Volume. In fact, it is one of the most important contributors to the deep problem of traditional relational database which become unable to manage databases with very large volume.
- Variety: Another issue is to handle and merge data that have several different sources with different structures and types.
- Velocity :This criterion is also very important. Indeed, it represents the speed of generation of data.
- Veracity: It represents the uncertainty of data.

In this paper, we will focus on the transportation field which is also impacted by big data challenges. In fact, in order to overcome big data challenges, industries including transportation companies require sophisticated tools called big data analytics such as Hadoop and Mapreduce. These tools allow to handle oversized data collected from several sources such as sensors existing in the roads or vehicles equipped with Global Positioning System (GPS) or mobile devices, etc... and to efficiently explore them.

The remaining part of this paper is organized as follows: Section 2 presents an overview of the proposed system which contains the process of gathering big GPS data from vehicles, clustering them with K-means in sequential mode. Also, in the same section, we introduce our proposed parallel K-means based on Mapreduce and we give an overview of the results and finally the generation of a road map and the map-matching with Google Maps[8].

2 Proposed System Architecture

Modern transportation services and management agencies rely mainly on the vehicle location, especially with the increasing and wide use of GPS and mobile devices. In fact, in our work, we used vehicles equipped with GPS devices. Those vehicles are travelling along the Tunisian road network. This section describes the proposed architecture as it is indicated in the Figure1. Each step is detailed in the following subsections.

2.1 Gathering GPS data

Traffic information are the key of contribution in Intelligent Transportation Systems(ITS) [13]. The traditional strategies to collect traffic data are based on the fixation of sensors along the adequate road. Whereas, the cost can be very high and cannot supply enough traffic information especially when the sensors number is limited in order to decrease the cost of this step in a

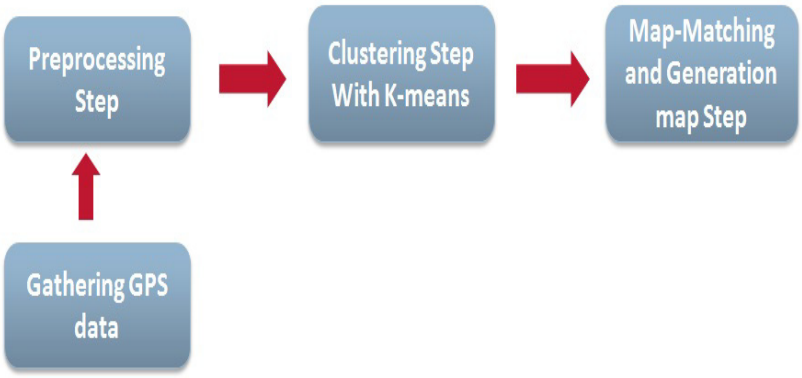

Figure 1: Proposed System Architecture

project. Some other works used for collection of such information some sophisticated cars to accomplish this task[9, 7] . However, this method does not allow researches to collect the details of some streets because vehicles will focus on the principle roads, else the cost will be very high as the traditional methods. Also, some previous works are satisfied with latitude and longitude coordinates and neglect the timestamp data and the velocity of the vehicle. In our work, we rely on all those data thanks to the importance of the information which can be utilized in many ITS fields, such as tracking vehicle services,traffic congestion detection, bus arrival time prediction, etc ...

In our work, we used a large number of fleet of vehicles which exceeds 10 thousands of vehicles to generate the road network in Tunisia. All vehicles are equipped with GPS devices which allow to capture their movements in each second. GPS receiver receives signals from the satellites. After that, it calculates the positions with triangulation formula. To ensure a precision of the vehicle position, four satellites are indispensable. The GPS receiver calculates the distances between its position (the vehicle position) and each satellite's position and we get the (x,y,z) coordinates and the time of signal arrival. Next, it sends geographical information coupled with timestamp information to the servers. The communication is insured thanks to the General Packet Radio Service (GPRS) protocol between GPS and servers. While the number of vehicles is very big and it increases more and more, we were obliged to use eight servers until now to save the data. After that, we implement an algorithm which extracts the National Marine Electronics Association (NMEA) sentences, emitted by the GPS and transferred to the servers. The Figure 2 shows some samples of our NMEA sentences emitted by a GPS receiver implemented in a vehicle. As it is shown in the Figure 2, the NMEA sentences emitted by this GPS receiver contains two different samples of GPRMC sentences used in our project. The first subset of GPS frames is the useful one. In fact, they contain the needed components such as the latitude, longitude and their orientation, the speed, time and date, etc... However, we can often find some sentences like the second subset. It is unused part. In fact, those generated sentences are due to the errors of GPS. One of the major factors leading to the emission of some wrong measurements is error in calculation of triangulation formula. In fact, this is caused

by a limitation of satellites number (less than four). This case can mainly appear when the vehicle is in basement parking or it is travelling through a tunnel... This mixture of having different types of GPS frame, different number of items in a one and sometimes appearance of symbols like "?" provide one of the challenges of big data that we should overcome, which is *unstructured data*.The phase of gathering data allows us to collect a big database. In fact, it is very huge and its size exceeds 100 GB. The vehicles were circulating in approximately all the Tunisian territory. Figure3 shows an overview of the traces of the vehicles and its distribution in the map. Therefore, we encountered the second challenge of the big data which is the *huge volume*. That's why big data analytics are required in this project such MapReduce framework developed and discussed in section 2.3.2.

2.2 Preprocessing of our big database

In this part, we preprocessed our database in order to be after analysed. In deed, some treatments are necessary to convert GPS data into useful data. Conversion step affects many components of GPS frame such as timestamp format (date and time),position coordinates to decimal degrees and velocity from mile/hour to kilometre/hour.

2.3 K-means Clustering of GPS data

As we dispose of an avalanche of data, it is obligatory to use sophisticated tools in knowledge discovery. In fact,data mining methods are very well-known as an efficient knowledge discovery techniques for those purposes [15, 16, 12]. Clustering is one of those methods. Its principal idea consists of dividing data into subsets so that objects composed each subset have more similarity compared to other objects existing in other subsets [4]. Data clustering is used in different fields of computer science and related areas. In spite of the widespread use of clustering in data mining, it is immensely utilized in other study fields such as machine learning, pattern recognition, networking and transportation. Thus, several research works have been carried out

Figure 2: Structure of our NMEA senetences emitted by the GPS receiver

[5, 10, 17, 14]. In this part, the clustering is an interesting part because it facilitates handling each part of the big GPS database. All the GPS nodes which are in the same zone are treated together.

2.3.1 Overview of K-means sequential clustering

The K-Means clustering is the most famous technique of data mining. It is mainly efficient in several field including the field of transport data management[3]. It belongs to the group of partitional clustering methods.Besides the simplicity of this method, it has several utilities such as its easiness ,simplicity to be comprehensible and to be implemented. Moreover, K-means is very efficient [11]. K-means algorithm allows the concerned objects to be divided into several clusters. In fact, it arranges them without any prior knowledge or learning step. It is defined as unsupervised classification algorithm.

K-means sequential algorithm: The main purpose of K-means clustering use is the automatic partition of a data set into k subsets. In our work, the input data are the GPS data, and K is the number of clusters required. Let P_1=(lat_1,$long_1$),P_2=(lat_2,$long_2$)... P_n=(lat_n,$long_n$) be the set of n GPS data, where P_i=(lat_i,$long_i$) is the i^{th} GPS point with lat_i its latitude and $long_i$ its longitude, extracted from the 100000 of vehicles discussed in previous sections. Since the GPS data are arranged along latitudes and longitudes and had two dimensions coordinates. Let C_1=(lat_1,$long_1$),C_2=(lat_2,$long_2$)... C_K=(lat_K,$long_K$) be the K clusters which is the input of K-means algorithm.

The considered K-Means algorithm contains the following steps:

- Step1: It begins with the generation of k initial clusters within the GPS data domain,
- Step2: The creation of the k clusters is done by joining every P_i to the closest centroid C_i after calculation of Euclidean distance between all the P_i and C_i,

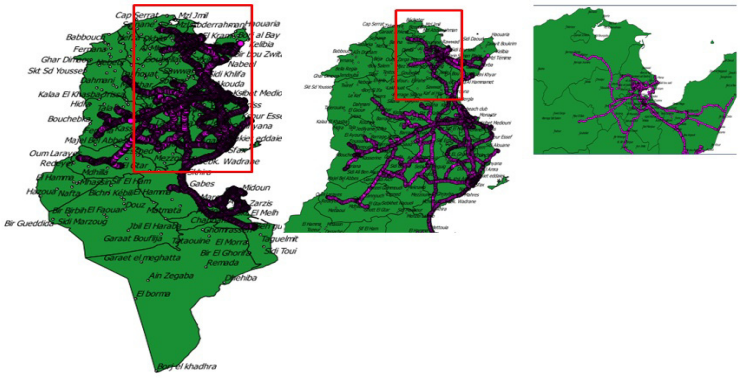

Figure 3: Overview of the GPS database in Tunisian map

- Step3: Each center C_i is adapted to a new position with taking into consideration the new configuration of the GPS data clusters.
- Step4: Repeat step2 and step3 until the all the centers C_i reach effectively the best positions.

In this section, we tested K-means algorithm with our big dataset of GPS data. The Table1 contains the number of GPS data points used in the clustering, the number of clusters inserted and the time cost for the processing.

Table 1: Time cost of clustering GPS data

Number of GPS data points	Number of clusters "K"	Time cost (s)
10 000	2	1.461060
10 000	5	2.482969
100 000	2	2.310787
100 000	5	13.654104
1 000 000	2	16.428530
1 000 000	5	26.942392
2 000 000	2	33.855862
2 000 000	5	53.787785
5 000 000	2	79.135197
5 000 000	5	118.505268
10 000 000	2	154.440376
10 000 000	5	292.221043

K-Means in its sequential version has several advantages ,however, a considerable execution time and memory consumption deeply increases while the size of the inputs and "K",the number of clusters increase as it is shown in the Table1 . Some ideas can be revealed to ameliorate the efficiency of K-means. The first one is a software solution to optimize it and the other one is hardware to use a parallel environment. We preferred in this paper to perform the second solution which may be resolve those problems.

2.3.2 Parallel K-means process based on Mapreduce framework

In this section, we investigate a solution to reduce the time cost of K-means clustering. In fact, traditional relational database become unable to manage databases with very large volume of data. Some complex issues that can embarrass of traditional databases are mainly the cost, waste of time for waiting the answers to the queries, and their disability to manage unstructured or semi-structured data. For that, we used Apache Hadoop and Mapreduce. Concerning Apache Hadoop, it is an open source framework written in Java. Its structure which is composed of computer clusters of commodity hardware allows to handle big datasets. In fact, it is divided into two parts which are (1) distributed storage component: Hadoop Distributed File System (HDFS), and (2)Processing component : the MapReduce programming model to process huge data.

2.3.3 Parallel algorithm

In this section, we will focus on the K-means algorithm based on MapReduce. In fact, the K-means algorithm in its sequential mode has a heavy and tiring computation of distances

between GPS data and the centroids. Indeed, the problem becomes deeper when the number of GPS nodes and clusters increase. Suppose that we have n ,number of GPS data, and k ,the number of clusters, K-means should process (n*k) distance operations in each iteration. Here, the parallelism can be created in our K-means in this level. However, the update of the centroids should be serially performed.

To achieve our goal, the GPS data were stored on the HDFS. They were represented as a file which contains $< key, value >$ pairs. Each pair represents (lat,long) coordinates. Concerning the *key*, it constitutes the pair offset into the file. However, the pair value is stored in the *value*. Our developed algorithm is based on three main functions:

- Map function: Its role is to partition the input and send them to the map functions. Each one reads the information related to the center and computes the distance between GPS data and the adequate center. The outputs of each mapper in this phase are the value and the index of the closest center.

- Intermediate function: Its role is to merge the outputs of each mapper and make the summation of the GPS data values which belong to the same cluster in each mapper.The outputs of this step are sum of GPS data belonging to the same cluster and the number of the samples.

- Reducer function: Its role is the summation of all the GPS data as well as computation of number of GPS data belonging to each cluster. Finally, the output is the centers of the clusters.

2.3.4 Experiments and discussion

In this subsection, our algorithm performance is evaluated with respect to the time cost of the K-means computation. In fact, to process a big data GPS, we used a cluster of computers. Each one has two cores with 2.8 GHz cores and 4GB of RAM ans Hadoop version 2.2.0. The Figure4 contains the results of our experiments. In fact, as we increase the number of nodes, the processing of big data becomes faster and efficient and all problems encountered in the previous section are resolved. So, we can generate our network map .

2.4 Generation of Map Network

Route planning relies on the road network representation . In fact, users require accurate representations in order to guarantee a safe trip. For that, road network should be geometrically precise to allow users to discover routes and have clear recommendations to drive easily between turns. In this paper, we investigate an accurate representation of the road map in Tunisia beginning with our big database of GPS traces. After preprocessing the GPS data and clustering them, our purpose in this section is to infer a road map of Tunisia which answers route planning requirements. We tried to map-matching our results to the Google maps to compare the accuracy of our resulted roads.

The figure 5 shows about 5 000 000 GPS data in the Tunisian map classified into 5 clusters in 5(a). We presented also some areas to see the accuracy of our map in 5(b). As it is shown, the trace of the vehicle (represented as blue stars) coincide with the roads presented in Google maps.

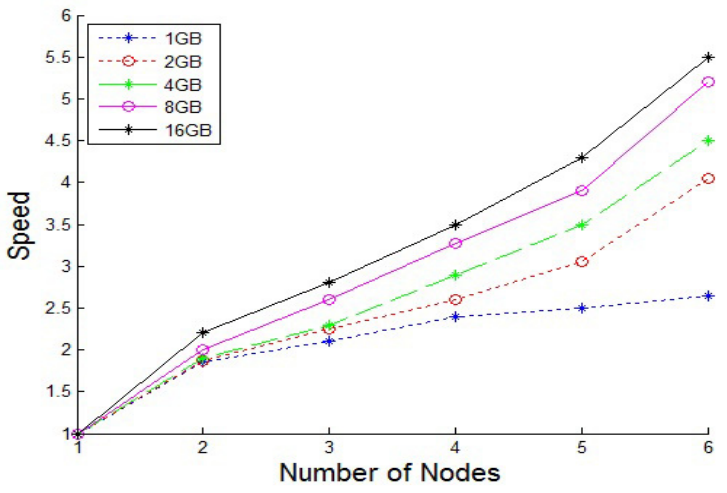

Figure 4: Evaluation of speed of computation with variation of nodes number and samples size

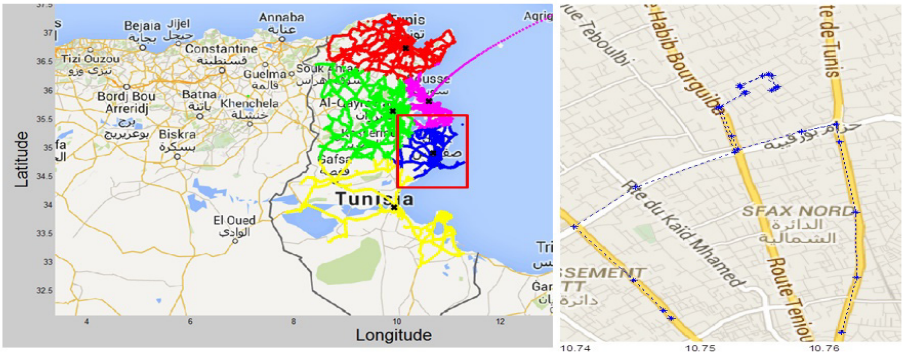

Figure 5: (a)Generation of 5000000 GPS data classified in 5 clusters(b)Map-Matching of a sample trajectory of a vehicle with Google Maps

3 Conclusion

In this paper, we collected a large database of GPS data from more than ten thousands of vehicles equipped with GPS receivers circulating in Tunisia. One of the exploration of this wealth is to generate a road network to be useful for people in this country. For that, several steps were done such as the preprocessing of data and its classification to facilitate its treatment. However, the size of the database was a big problem. For that, we developed a parallel K-means algorithm based on Mapreduce and Hadoop to accelerate the processing. The evaluation of this

method shows good results.

4 Acknowledgement

This work was carried out within the framework of a PhD MOBIDOC (PASRI program), funded by the EU and managed by the ANPR. The authors would like also to acknowledge the partial financial support of this work by grants from General Direction of Scientific Research (DGRST), Tunisia, under the ARUB program.

References

[1] Fact sheet: Big data across the federal government. [online], 2012. `http://www.whitehouse.gov/sites/default/files/microsites/ostpbigdatafactsheet3292012.pdf`.

[2] Intel IT Center. Planning guide: Getting started with hadoop, steps it managers can take to move forward with big data analytics, June 2012. last viewed February 2015.

[3] W. Elleuch, A. Wali, and A.M. Alimi. Mining road map from big database of gps data. In *Hybrid Intelligent Systems (HIS), 2014 14th International Conference on*, pages 193–198, Dec 2014.

[4] W. Elleuch, A. Wali, and A.M. Alimi. Collection and exploration of gps based vehicle traces database. In *4th International Conference on Advanced Logistics and Transport(ICALT'15)*, May 2015.

[5] Jason Ernst, Gerard J. Nau, and Ziv Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168, 2005.

[6] Reinsel D Gantz J. Digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east, IDC iView, pp 1-16, December 2012. last viewed February 2015.

[7] Hector Gonzalez, Jiawei Han, Xiaolei Li, Margaret Myslinska, and John Paul Sondag. Adaptive fastest path computation on a road network: A traffic mining approach. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, VLDB '07, pages 794–805. VLDB Endowment, 2007.

[8] Google. last viewed February 2015.

[9] Bret Hull, Vladimir Bychkovsky, Yang Zhang, Kevin Chen, Michel Goraczko, Allen Miu, Eugene Shih, Hari Balakrishnan, and Samuel Madden. Cartel: A distributed mobile sensor computing system. In *Proceedings of the 4th International Conference on Embedded Networked Sensor Systems*, SenSys '06, pages 125–138, New York, NY, USA, 2006. ACM.

[10] Flix Iglesias and Wolfgang Kastner. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies*, 6(2):579, 2013.

[11] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 19th International Conference in Pattern Recognition (ICPR).

[12] Terence Kwok, Kate Smith, Sebastian Lozano, and David Taniar. Parallel fuzzy c- means clustering for large data sets. In Burkhard Monien and Rainer Feldmann, editors, *Euro-Par 2002 Parallel Processing*, volume 2400 of *Lecture Notes in Computer Science*, pages 365–374. Springer Berlin Heidelberg, 2002.

[13] Xiaolei Li, Jiawei Han, Jae-Gil Lee, and Hector Gonzalez. Traffic density-based discovery of hot routes in road networks. In Dimitris Papadias, Donghui Zhang, and George Kollios, editors, *Advances in Spatial and Temporal Databases*, volume 4605 of *Lecture Notes in Computer Science*, pages 441–459. Springer Berlin Heidelberg, 2007.

[14] Francois G. Meyer and Jatuporn Chinrungrueng. Spatiotemporal clustering of fmri time series in the spectral domain. *Medical Image Analysis*, 9(1):51 – 68, 2005.

[15] C. Soares P. Williams, , and J. E. Gilbert. A clustering rule based approach for classifi-cation problems. In *Int. J. Data Warehouse*, pages vol. 8, no. 1, pp. 123, 2012.

[16] R. V. Priya and A. Vadivel. Behaviour pattern mining from weblog. In *Int. J. Data Warehouse*, pages Min., vol. 8, no. 2, pp. 122, 2012.

[17] Ying Zhao and George Karypis. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331, 2004.