Available online at www.sciencedirect.com

ScienceDirect

AASRI Procedia 5 (2013) 156 – 163

AASRI Procedia

www.elsevier.com/locate/procedia

2013 AASRI Conference on Parallel and Distributed Computing and Systems

Large Distributed Arabic Handwriting Recognition System Based on the Combination of FastDTW Algorithm and Map-reduce Programming Model via Cloud Computing Technologies

Hamdi Hassen[a,b*], Maher Khemakhem[a]

[a]*Mir@cl Lab, FSEGS University of Sfax BP 1088, 3018 Sfax, Tunisia*
[b] *Computer science departement, College Of Science And arts at Al-Ola, Taibah University, KSA.*

Abstract

This paper proposes a robust, efficient and scalable distributed_Arabic handwriting OCR system based on a parallel FastDTW algorithm via cloud computing technologies. The three techniques Hadoop, MapReduce and Cascading are used to implement the parallel FastDTW algorithm. The experiments were deployed on Amazon EC2 Elastic Map Reduce and Amazon Simple Storage Service (S3) using a large scaled dataset built from the IFN/ENIT database.

Keywords : Large OCR system, Fast DTW, MapReduce, Cloud computing ;

1. Introduction

Today there are many OCR systems in use based on different approaches and algorithms. All of the popular OCR systems support high accuracy and most high speed especially those dedicated for printed characters and high quality documents. Unfortunately, this is not the case especially for the Arabic handwriting characters

[*] Corresponding author. Tel: +966-563598400 ; fax: + 216 74 278 777
E-mail address hhassen2006@yahoo.fr

doi:10.1016/j.aasri.2013.10.072

where OCR systems are limited to recognize small and rarely medium quantity of documents for some specific purposes.

OCR systems that treat large amounts of documents are very limited and not powerfully enough such as the Australian Newspaper Digitization Project [1], OCRGrid [2], Kirtas[3] and OCRopus [4].

Conducted experiments and evaluations on several Arabic handwriting OCR systems show and confirm that : in the first hand, the euclidean distance technique is used for classification. However, this technique is less robustness and more fragile [5]. In the second hand, the Dynamic Time Warp (DTW) algorithm stands among the best techniques for such a mission [6].

The major problem of the DTW is the slowness of its response time because of the enormous amount of computation to achieve [7]. Distributed system, such as cloud computing technologies, provides viable framework to speed up the time of the OCR system based on DTW algorithm. Cloud computing is primarily used to deliver many services such as Infrastructure (I), Platform (P) and Software (S) as services. All these services are available to consumers as registration based services in a pay-as-you-consume model [8].

This paper is organized as follows: an overview on the DTW algorithm and especially the FastDTW and the use of them in Arabic character recognition, is presented in section 2. Hadoop, MapReduce and Cascading models are presented in section 3.The proposed approach is explained in section 4. Experimental and results are presented and discussed in section 5. Conclusion and future work are presented in the last section.

2. Dynamic Time Warp (DTW)

2.1. Dynamic Time Warping Algorithm

The Dynamic Time Warping (DTW) is a technique intended to compute similarities between two different sequences of patterns even when they are not aligned in time or in space [9].

Let's consider A and *B* two different sequences.

*n :*is the feature vectors of the sequence *A*. *m:* is the feature vectors of the sequence *B*.

$$A= a_{1,} a_2,a_3,\ldots..a_n \quad (1)$$

$$B= b_{1,} b_2,b_3,\ldots..bn \quad (2)$$

D [*n , m]* : the distance matrix.

Cell (i, j) represents the distance between the i^{th} element of the sequence *A* and the j^{th} element of the sequence *B* (Fig1).

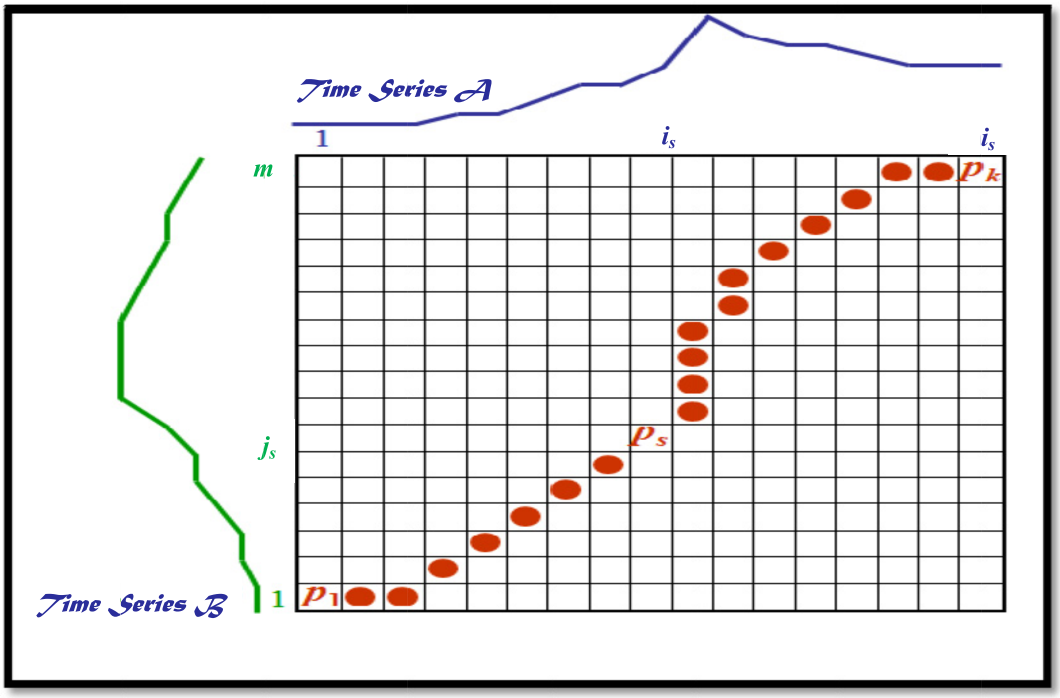

Fig 1. The DTW mechanism.

To find the best alignment between *A* and *B*, we need to find the path through the grid.

$$P = p_1 \ldots p_s \, , \; \ldots \, , p_k \tag{3}$$

$$p_s = (i_s \, , \, j_s \,) \tag{4}$$

This minimizes the total distance between them.

P is called a warping function.

To calculate the length path, is simply, just we sum all the cells that were visited along this path[10].

$$D\ (A,\ B) = \left[\frac{\sum_{s=1}^{k} d(p_s)\cdot w_s}{\sum_{s=1}^{k} w_s} \right] \tag{5}$$

D (ps): the distance between i_s *and* j_s*;* $w_s > 0$: weighting coefficient.

$P_{0:}$ The best alignment path between A and B:

$$P_0 = \arg_p \min\ (D\ (A\ ,\ B\ )). \tag{6}$$

2.2. Fast DTW

DTW presents many disadvantages such as the time and space complexity which are exponential. This model is practical only for small and medium data sets *(<3,000)* and time series are often very long [11]. The FastDTW algorithm can be a solution to solve this problem. FastDTW is based on the multi-resolution approach inspired by the multi-level graph bisection algorithm [12].

2.3. Fast DTW for Arabic Handwriting recognition system

FastDTW for Arabic handwriting recognition system consists to prepare a reference database of *R* trained Arabic alphabet and number in some given scripter, and presented by C_i, $i = 1,2, \dots R$. Our approach consists on using the FastDTW pattern algorithm to classify the character to recognize against the template library. Thus the input character is classified as the best character that gives the optimal time alignment *p* among all R characters.

This technique is based on three steps: While the original resolution is not reached,

1) Set the resolution of the character to recognize in order to be the coarsest.
2) Research and find the initial path with DTW algorithm

   $P_k = min \{P_r\}$ where $1 <= r <= R$
3) Repeat the
   - Double the resolution,
   - Project the path on to the finer resolution,
   - Find a path through the projected area.

3. MapReduce, Hadoop an Cascading: an overview.

3.1. MapReduce technology

MapReduce [13] is a tools using to parallelize problems that process large datasets with different computers (nodes) (distributed architecture) like a cluster or a grid computing . Amazon Elastic MapReduce provides the option to analyze vast amounts of data. This advantage is offered by distributing the computational work across a cluster of virtual servers running in the Amazon cloud. All clusters are managed using an open-source framework called Hadoop.

3.2. Hadoop

Hadoop [14] is a distributed infrastructure for processing large-scale data. This infrastructure can be used for single machine. The real power of this architecture lies on the ability to use hundreds or thousands of nodes, each with different processor cores. The Hadoop model is also used to share efficiently huge work across different machines". Hadoop does this by the storage layer that manage large amounts of data, and the running layer that parallelize the execution of the user application using coordinated data subsets.

3.3. Cascading

Cascading website [15] define Cascading "is a framework written with Java language that helps typical developers to easily and quickly develop Data Analytics and Data Management system that can be deployed and managed by a variety of computing environments." This model is based on a metaphor of data streams called pipes and data operations called filters. Thus, the Cascading API allows the developer to regroup pipe assemblies that do many actions such as the split, the merge,.. of data while applying operations to the different data record.

4. The proposed approach

To solve the problem of the OCR systems that can treat large amount of documents, we have opted for the cloud computing paradigm that promises to reduce substantially investments by eliminating the necessity of managing huge computing capacity by institutes and organizations.

In our case, the model master–slave and the SPMD (Single Process, Multiple Data) architecture are used on the distribution of the memory .This model is applied to the FastDTW algorithm as the parallelization technique. The distributed FastDTW approach consist on running each copy of the single program on independently processors and Hadoup is the responsible of the communication between processors .

The big amount of document to recognize is better to split it into small parts *(D1, D2, D3 ...Dn)* and assign each one to a slave node number to achieve the recognition task. The FastDTW algorithm will be implemented by a jobs flow for each node.

Amazon Elastic MapReduce execute automatically the Hadoop program of the OCR application on different Amazon EC2 instances. First, the map function is applied, this function consist on sub-dividing the huge amount of documents in a job flow into smaller process so that they can be processed in parallel. Second, the reduce function that consist on merging the processed data into the final output is applied (Fig2). The Amazon Simple Storage Service (S3) is in the first hand, the source for the data to process and in the second hand, is the output destination.

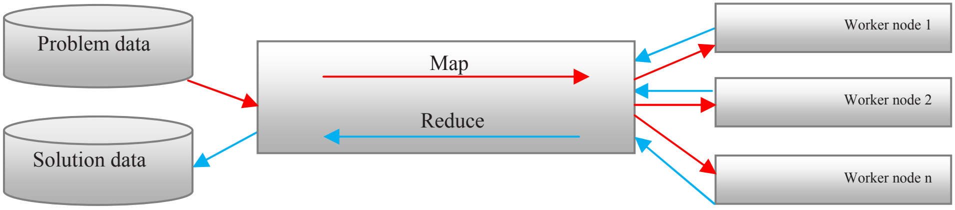

Fig2. The MapReduce mechanism

5. The experimental study

5.1. Datasets

To examine the proposed idea, a corpus with 16000 pages (370 characters/page) and a reference database formed of 345 shapes representing approximately the different Arabic alphabet randomly chosen from the Arabic handwritten word images dataset IFN/ENIT[16] are used. For the preprocessing image, the IFN/ENIT dataset was already normalized [16]. Wavelet transform [17] is used as a features extraction technique.

5.2. Experimental environment

The experiments were conducted on a local Intel Core 2 Duo desktop having the configuration: 3.00 GHz *2, 2 GB of RAM executing a Windows XP operating system, Cygwin [18] is the shell to run Linux command. We used java as a programming language and JDK 1.6 was installed. Eclipse 3.4 was used to program and built our application. 100 MG bits/s was the network capacity.

Based on the state art of the cloud computing technologies [19], Amazon Elastic Computing Cloud have selected for the implementation of our approach. In order to verify that distributed FastDTW functions correctly in cloud technologies, we created six running Jobs flow on the Amazon Elastic Computing Cloud service. We have allocated 100 cores using the three Standard Amazon EC2 Instances. First, the “small” instances each with 1.7 GB of memory, 160 GB of instance storage, and 32-bit platform. Second, the Large

Instance 7.5 GB of memory, 850 GB of instance storage, 64-bit platform and finally the Extra Large Instance 15 GB of memory, 1690 GB of instance storage and 64-bit platform. S3 [20] is used to manage the input and output data.

5.3. Results and analysis

To prove the efficiency of the cloud computing technologies on the execution time of the proposed FastDTW based approach, we created six running Jobs flow in cascading on Amazon Elastic MapReduce Cloud and two experiments are conducted in order to compare the DTW and FastDTW speedup in three instances of Amazon Elastic Computing Cloud service.

The table below illustrates the execution time using DTW and Fast DTW with the different instances.

Table1. DTW and FastDTW time execution

Instance	Number of core	DTW(H)	FastDTW(H)
Small instance of Amazon Elastic Computing	25	0.563	0.444
	50	0.300	0.258
	75	0.205	0.174
	100	0.150	0.123
Medium instance of Amazon Elastic Computing	25	0.500	0.421
	50	0.273	0.229
	75	0.191	0.160
	100	0.143	0.119
Large instance of Amazon Elastic Computing	25	0.450	0.364
	50	0.257	0.216
	75	0.180	0.157
	100	0.136	0.118

The average duration of the test time in a sequential mode (a single computer) using DTW and FastDTW are approximately and respectively 9 hours and 8 hours and the average test time in distributed mode and for 100 computers are 0.136 hours and 0.118 Hours. These show that the sequential mode allows recognizing only 18 and 21 characters per second for the two algorithms described above. However, the results in the distrusted mode, illustrated in figures 3and 4, in particular :

- The speedup factor which is the ratio of the execution time in a sequential manner using a single processor to the execution time using multiple processors, increases simultaneously with the number of cores used and with the different Standard Amazon EC2 Instances.
- If we use 100 cores with large instance of Amazon Elastic Computing, then the execution time reaches the value 489 and 425 seconds and the speedup factor reaches the values 66 for DTW and 68 for FastDTW. These results are very important since our distributed OCR pattern can recognize more than 1200 and 1400 (characters /second) for DTW and FastDTW respectively.

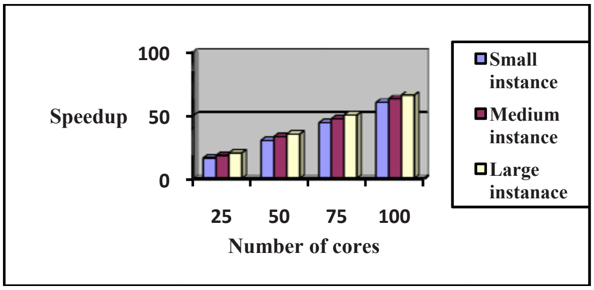

Fig 3. DTW speedup with Amason elastic Mapreduce

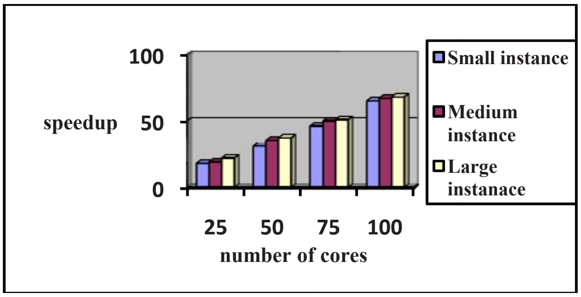

Fig4. FastDTW speedup with Amason elastic Mapreduce

Our model presents several advantages since cloud computing provides almost the entire required tools in order to ease the distribution of any greedy algorithm or application such as data management, task scheduling, hosting machine failures, facilitating communication between machine and consequently everything is completely transparent to the programmer/analyst/user.

6. Conclusion and perspectives

Performance evaluation of the proposed model confirms that Mapreduce, hadoop and cascading technologies provide an adequate platform to accelerate the Arabic handwritten recognition process. Moreover such platform allows building much more powerful OCR systems compared to the existing system [21] since it provides enough computing and storage powers in addition to some useful tools and facilities.

In future work, we examine how to deploy our OCR application on a multi-cloud infrastructure and how to improve the recognition rate of our OCR system by the integration (combination) of some strong complementary approaches (e.g. HMM, SVM).

References

[1] Available at:http://www.nla.gov.au/
[2] Available at: http://www.ocrgrid.org/
[3] Available at: http://www.kirtas.com/
[4] Available at: http://code.google.com/p/ocropus/
[5] Y.Jun Weng and Z. Ying, “Time Series clustering based on shape dynamic warping using cloud models” Proceedings of the Second International Conference on Machine Learning and Cybernetics, Xi'an, 2-5

November 2003

[6] M. Khemakhem and A. Belghith. A P2P grid architecture for distributed Arabic OCR based on the DTW algorithm, IJCA-ACTA press, V.31, N1, 2009.

[7] Maher Khemakhem, Abdelfettah Belghith Towards distributed Cursive writing OCR systems based on the combination of complementary approaches, springer chapter, 2012

[8] Sushil Bhardwaj1, Leena Jain1, Sandeep Jain2 cloud computing: a study of infrastruture as a service (IAAS), International Journal of Engineering and Information Technology Vol 2 , No. 1 IJEIT 2010

[9] Berndt D J, Clifford J. Finding patterns in timeseries: a dynamic programming approach. In Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 1996, 229–248.

[10] M. Khemakhem et al. , Reconnaissance de Caractères Imprimés par Comparaison Dynamique, Proc. AFCET, Antibes, Sept. 1987.

[11] Stan Salvador and Philip Chan, “FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space” Workshop on Mining Temporal and Sequential Data, pp. 70-80, 2004.

[12] Stan Salvador and Philip Chan, “FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space” Workshop on Mining Temporal and Sequential Data, pp. 70-80, 2004

[13] Available at: http://aws.amazon.com/elasticmapreduce/

[14] Available at: http://hadoop.apache.org/

[15] Available at: http://www.cascading.org/

[16] M. Pechwitz, S. S. Maddouri, V. Mrgner, N. Ellouze, and H. Amiri. Ifn/enit - database of handwritten arabic words. In In Proc. of CIFED 2002, pages 129– 136, 2002.

[17] Hassen Hamdi, Maher Khemakhem, A Comparative study of Arabic handwritten characters invariant feature. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 12, 2011

[18] Available at: http://www.cygwin.com/

[19] R. Prodan and S. Ostermann, “A Survey and Taxonomy of Infrastructure as a Service and Web Hosting Cloud Providers,” Proc. Int’l Conf. Grid Computing, pp. 1-10, 2009.

[20] Available at: http://aws.amazon.com/s3/

[21] M.Khemakhem and A. Belghith. Towards A Distributed Arabic OCR Based on the DTW Algorithm: Performance Analysis The International Arab Journal of Information Technology, Vol. 6, No. 2, April 2009.