Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

ScaDiPaSi: An Effective Scalable and Distributable MapReduce-Based Method to Find Patient Similarity on Huge Healthcare Networks[☆]

Mohammadhossein Barkhordari[1], Mahdi Niamanesh[1]

Information and Communication Technology Research Center, No. 5, Saeedi alley, College intersection, Enghelab street, Tehran 1599616313, Iran

A R T I C L E   I N F O

Article history:
Received 18 September 2014
Received in revised form 8 January 2015
Accepted 11 February 2015
Available online 17 February 2015

Keywords:
Healthcare
Patient similarity
Big data
MapReduce

A B S T R A C T

Healthcare network information growth follows an exponential pattern, and current database management systems cannot adequately manage this huge amount of data. It is necessary to use a “big data” solution for healthcare problems. One of the most important problems in healthcare is finding *Patient Similarity* (PaSi). Current methods for finding PaSi are not adaptive and do not support all data sources, nor can they fulfill user requirements for a query tool. In this paper, we propose a scalable and distributable method to solve PaSi problems over MapReduce architecture. ScaDiPaSi, supports storage and retrieval of all kinds of data sources in a timely manner. The dynamic nature of the proposed method helps users to define conditions on all entered fields. Our evaluation shows that we can use this method with high confidence and low execution time.

1. Introduction

Nowadays, with huge amounts of information generation, common database management systems cannot adequately support data management and analysis in many fields, including meteorology, scientific instruments, social networks, and medical networks. In these and other fields we need a paradigm shift to solve our problems. Capturing, storing and retrieving information in a timely manner are vital issues in these systems. It is necessary to have distributed and scalable solutions for these kinds of problems because the prevalent single-node and parallel approaches are far from offering a timely solution. On the other hand, scalable and distributable solutions have their own problems, in particular network bottlenecks, low performance of hardware nodes, and requirements for other nodes' information [21,31].

Healthcare is one of the fields that need scalable and distributable solutions, because current solutions cannot properly solve this area's problems. One of the most important problems in this area is identifying patient similarity, or PaSi, defined as the rate of similarity between two or more patients in terms of their symptoms, treatments, personal information, etc. The goal in PaSi is to identify those patients who have the greatest amount of information in common in order to use their treatments for new patients. We have two main issues in PaSi: the huge amount of information per patient; and the fact that most of this data is non-structured, lacking a predefined record structure that is common among all patients. A large number of fields per patient may add complexity to PaSi problems as well. Given these characteristics, we have to use so-called “big data” solutions. One of the methods which can be used for scalable and distributable solutions for big data is MapReduce [30]. MapReduce is used to solve healthcare problems [2,36,31]. But MapReduce and other distributable solutions have problems such as data locality, network bottlenecks, hardware inefficiency etc. [21,31].

In this paper, we propose ScaDiPaSi, a scalable and distributable method for investigating patient similarity. In this method, a MapReduce-based method is used to solve PaSi problems. Unlike other approaches, we do not use structured or semi-structured methods for patient information storage. ScaDiPaSi can use different data sources with different data items. Even the same data source can have different data items for two patients. Rather, ScaDiPaSi uses a dynamic method to store patient information which can be easily distributed over hardware nodes. In the proposed method hardware nodes can execute their tasks simultaneously, and none of the nodes needs information from other nodes which is the main problem of MapReduce-based methods.

The structure of this paper is as follows. Section 2 investigates some preliminaries concerning MapReduce and healthcare problems. In Section 3, PaSi-related literature is discussed. Section 4 focuses on the proposed method. Section 5 presents the evaluation of the proposed method. Section 6 provides the conclusion.

[☆] This article belongs to Comp, bus & health sci.

E-mail addresses: M.H.Barkhordari@ictrc.ir (M. Barkhordari), Niamanesh@ictrc.ir (M. Niamanesh).

[1] Tel.: +982188930155.

http://dx.doi.org/10.1016/j.bdr.2015.02.004
2214-5796/© 2015 Elsevier Inc. All rights reserved.

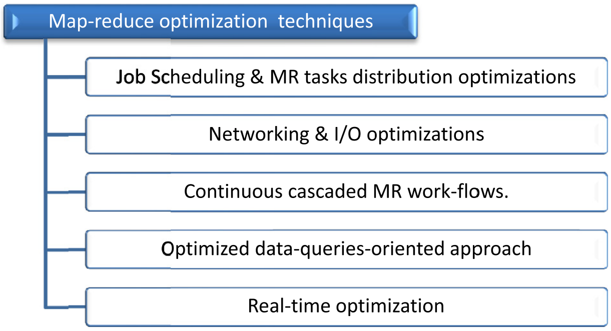
Map-reduce optimization techniques
Job Scheduling & MR tasks distribution optimizations
Networking & I/O optimizations
Continuous cascaded MR work-flows.
Optimized data-queries-oriented approach
Real-time optimization

Fig. 1. MapReduce optimization techniques.

2. Preliminaries

In this part, both MapReduce and the relationship between healthcare and big data are delineated.

2.1. MapReduce

In this section, the literature related to MapReduce design is discussed. According to [27], a decomposable algorithm, partitionable data, and sufficient small data partition are the main characteristics required for effective use of MapReduce. In [23], classic MapReduce was optimized to decrease the data transformation load. In the method described in [23], a shared area for information was considered. This type of design is suitable for solving problems, such as k-nn and top k queries. In [26], MPI (Message passing interface) was used for message passing in a MapReduce structure. The goal of that paper was to decrease the amount of data transferred in the MapReduce network. In [28], a method was developed for tackling workloads in hierarchical MapReduce architectures. HadUP was presented in [29]. HadUP is a modified version of Hadoop and uses a deduplication-based snapshot differential algorithm (D-SD) and update propagation. Haloop [25] is another type of MapReduce structure suitable for iterative problems. iMapreduce [24] also supports iterative processes. In [20], HDFS (Hadoop file system) was substituted with a concurrency-optimized data storage layer based on the BlobSeer data management service. In [22], a model was presented to estimate I/O behavior of MapReduce applications. In [21], optimization over MapReduce structure was divided into five groups. Fig. 1 shows these groups

2.2. Healthcare and big data

In this section, healthcare and its relation to big data are investigated. These days, patients' information is generated at an exponential rate. This information has different formats and standards. According to [19], there are various standard data sources, as shown in Fig. 2.

As shown in Fig. 2, huge *Volume* of information is generated in *Various* formats with high *Velocity*; therefore, we have three Vs of *Big data* in healthcare networks. With PaSi there is an additional challenge, namely *Veracity*, meaning that for many patients we typically have dubious or uncertain information. Healthcare problems manifest all of the V's, and therefore it is inevitable that we will use big data solutions to solve them. Nevertheless, according to [19], existing big data technologies do not adequately address the full spectrum of healthcare problems, so it is necessary to customize them for our purposes. According to high volume of information in healthcare big data is necessary for data analysis [32]. Also costs are reduced by using big data analytics in healthcare [33]. In [34] a patient-centered framework is proposed that can personalize healthcare with a big data driven approach. In [35] big data is used to solve problems like the selection of appropriate treatment paths or improvement of healthcare systems. AITION [37] proposed a scalable knowledge data discovery platform for big data Healthcare.

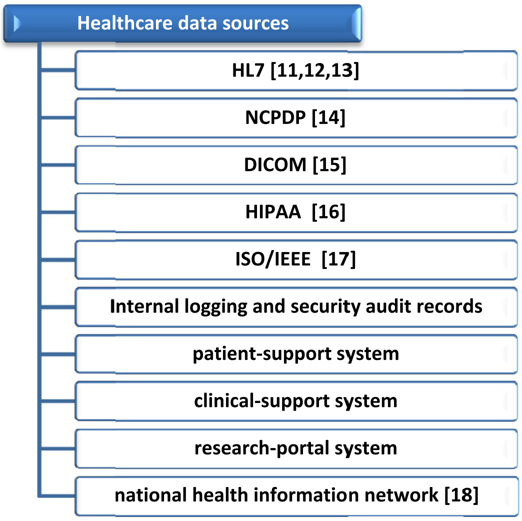

Fig. 2. Standard data sources in healthcare.

3. Literature on PaSi

In this section, literature specifically concerned with PaSi is investigated.

According to [1], finding PaSi solutions can be divided into two parts. Fig. 3 shows this categorization.

The first category is solutions that identify PaSi relationships by machine learning algorithms [3–5]. These types of solutions are offline and they require a long time for the machine learning to take place. Also there are data mining methods which work on streaming data and they can be considered as online data mining methods. These methods can only work on a part of data. In other word they have methods like sliding window, sampling, synopsis etc. over stream data; therefore, this method is not appropriate for PaSi problem because we need to analyze all data items [40]. The second category uses information retrieval techniques. Some techniques use simple search [6,7]; however, searching over limited keywords within a predefined structure may have severe limitations. Another information retrieval solution involves using Entity-Relationship Graphs (ERG) to investigate similarities between defined entities [8,9]. These types of solutions are expensive, and some are not online [8,9]. Some methods try to improve the ERG solution by unified search [10,11]. In [2] MapReduce is used to solve the problem. They tried to reduce algorithm execution time by distributing computation on hardware nodes. PARAMO [36] is a method which uses MapReduce to develop a predictive modeling platform in the healthcare analytics domain. Some methods used LSH [39] (Locality-Sensitive Hashing) for finding similarities [31]. In [31] LSH and MapReduce are used to extract patient similarity. LSH is not suitable for PaSi problem because it works with predefined data structure and with ever changing data sources accuracy will reduced dramatically.

According to our investigation, none of the above-mentioned methods are fully effective for solving PaSi problems, because of the following considerations:

- PaSi requires a dynamic structure to store patients' information. Different patients have different data items, and thus