Parallel Computing 39 (2013) 653–668

Contents lists available at ScienceDirect

Parallel Computing

journal homepage: www.elsevier.com/locate/parco

A map-reduce lagrangian heuristic for multidimensional assignment problems with decomposable costs

Gregory Tauer [a], Rakesh Nagi [b,*]

[a] Department of Industrial and Systems Engineering, State University of New York at Buffalo, 438 Bell Hall, Buffalo, NY 14260, USA
[b] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, 117 Transportation Building, MC-238, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Received 15 February 2012
Received in revised form 20 August 2013
Accepted 27 August 2013
Available online 4 September 2013

Keywords:
Data association
Multidimensional assignment problem
Lagrangian heuristic
Map-reduce framework

ABSTRACT

Data association is the problem of identifying when multiple data sources have observed the same entity. Central to this effort is the multidimensional assignment problem, which is often used to formulate data association problems. The nature of data association problems dictate that solution methods for the multidimensional assignment problem must return results promptly, and work on large data sets. The contribution of this work is to describe a Lagrangian relaxation based heuristic for the multi-dimensional assignment problem with decomposable costs that can be largely implemented in a map-reduce framework and thus easily distributed across a cluster of computers. Distribution allows the heuristic to address run time and large data requirements of data association. The developed algorithm is tested on a synthesized dataset, and shown to achieve an optimality gap ranging from 0.00008% to 0.6% for dense (no filtering) problems having 10,000 observation. Distribution of the algorithm was found to offer a significant reduction in run time on 30,000 observation problems for an 8 node computing cluster with 96 processors over a single node with 12 processors.

1. Introduction

The problem of data association is to take information from multiple (noisy) data sources and identify any overlapping information. Frequently, data sources describe entities, and in this case the task of data association is to decide when observations from different data sources are describing the same entity. Answering this question is advantageous since it allows information from multiple sources to be combined into a more complete representation. In doing so, association is a necessary step when working from many noisy observations of an environment.

An example of data association is the problem of multi-target tracking, or the partitioning of observations into tracks [20–22]. Based on prior domain knowledge of the objects under observation, a process called “scoring” is employed to assign a numerical score to how likely an observation belongs to a specific track. Given these scores, the multi-target tracking problem can be expressed as a multidimensional (K-dimensional) assignment problem which seeks to maximize the sum of within-partition scores.

The K-dimensional assignment problem is known to be NP-Hard for $K \geqslant 3$ [20–22]. This work addresses a specific variant of the multidimensional assignment problem where association scores are decomposable into pairwise scores [4,3,15]. Although it possesses some simplification properties, the multidimensional assignment problem with decomposable costs is also NP-Hard [23]. A more in-depth discussion these two multidimensional assignment problem varieties will be given in Section 2.

* Corresponding author. Tel.: +1 716 645 2427.
E-mail address: nagi@buffalo.edu (R. Nagi).

0167-8191/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.parco.2013.08.012

These problems are commonly used to model multi-target tracking and entity resolution problems. In applications of these problems, the number of objects to associate may be extremely large. The need for timely association on large data has been at conflict with approaches firmly rooted in the theory of combinatorial optimization [20]. This has led to the development of many metaheuristic algorithms, especially local search approaches [3,22], for multidimensional assignment. While effective at producing acceptable solutions quickly, a criticism of many metaheuristic approaches is their lack of a performance guarantee; applications of these algorithms risk ignorantly producing very poor solutions. In contrast to metaheuristics are methods that provide both upper and lower bounds, such as branch and bound or Lagrangian relaxation based approaches.

One way of addressing a metaheuristic's lack of an upper bound to the multidimensional assignment problem is to compute one as necessary on the side [3]. A drawback to this approach is that there is no overlap between these two efforts; time spent computing feasible solutions (lower bounds) by a metaheuristic does not help with computing an upper bound.

The main contribution of this work is the development of a novel Lagrangian heuristic for the multidimensional assignment problem with decomposable costs that can be distributed in a map-reduce framework. This directly addresses both the run time and large data requirements of multidimensional assignment by allowing the work required for the subgradient search procedure on this problem to be split across a large number of computers. More broadly, the approach adopted here is applicable to a variety of Lagrangian relaxation based approaches that rely on both decomposition into subproblems and the subgradient search method. To the authors' knowledge, this is the first work to apply the map-reduce framework to the solution of a Lagrangian relaxation. Distributed data association is not itself novel [14,8,18,19], however none of this work directly addresses distribution of the multidimensional assignment problem as is addressed here.

Map-reduce is a popular framework for distributed computing based on two fundamental steps: a map step and a reduce step [25,17,9]. This framework is especially appealing as it provides a simple interface that improves convenience and robustness [9]. Map-reduce has become a standard of parallel computing. Many large organizations have deployed large computing clusters for the purpose of running map-reduce algorithms [9,25,2], and because of this wide availability it is important to investigate algorithms that can utilize map-reduce computing resources. The application of map-reduce to a subgradient-optimization based Lagrangian heuristic may seem counterintuitive at first, since gradient descent methods are widely recognized as a major class of problems poorly suited for map-reduce [16]. However, as will be discussed in Section 5, specific properties of the Lagrangian heuristic presented here make it amenable to a map-reduce approach.

The remainder of this paper is organized as follows. Section 2 discusses the multidimensional assignment problem, the multidimensional assignment problem with decomposable costs, and summarizes trade-offs between these formulations. Section 3 describes a Lagrangian relaxation to the formulation for multidimensional assignment with decomposable costs, and also describes the repairing step necessary to transform the Lagrangian relaxation solutions to be feasible to the multidimensional assignment problem with decomposable costs. The map-reduce framework is briefly described in Section 4. Section 5 contains the main contribution of this work, where it is shown how the relaxation from Section 3 can be solved in a map-reduce framework. This process is tested in Section 6, which contains a summary of the resulting algorithm's performance. Finally, Section 7 contains a description of possible future work and conclusions.

2. Problem formulation

This section discusses the multidimensional assignment problem for data association as expressed by an integer programming formulation. This problem applies when K data sources each report on N objects that need to be partitioned into groups by their true identities. For the remainder of this paper, it will be assumed that every data source in a problem will contain the same number of observations, that the probability of detection is 1 and there are no false detections. This assumption implies that each of the N objects appears exactly once in each of the K data sources, but this can be generalized by introducing dummy objects. In each data source $p \in \{1, \ldots, K\}$, the individual observations are defined by v_i^p. Each v_i^p represents an observation of some real world object, such as target i from scan p in a multi-target tracking example. The set of all observations from source p is $V_p = \{v_1^p, \ldots, v_N^p\}$. The set of all data sources is $F = \{V_1, \ldots, V_K\}$. A partition $X_j \in X$ of observations is defined as a set of observations, one from each of the K data sources. Based on the observations, a reward $r(X_i)$ is assigned for each partition. The multidimensional assignment problem is to select partitions with maximum reward so that every observation is assigned to a partition and no two observations from the same data source appear in a given partition. This problem is NP-Hard, and its formulation is more thoroughly discussed by [20,21,3].

One difficulty with the multidimensional assignment problem formulation is the large number of decision variables required. In an effort to reduce the quantity of decision variables, applications of data association commonly reduce the total number of considered partitions through a process called "gating" where only the more highly scored tracks are kept [20].

In addition to gating, other approaches have been taken to improve the tractability of this problem. For instance, the reward function $r(X_i)$ can be decomposed [4,3,15]. This paper is specifically interested in reward functions that are pairwise decomposable by sum cost, which is the case when $r(X_i)$ is defined as in Eq. (1).

$$r(X_i) = \sum_{\{V_m, V_n\} \in \binom{F}{2}} \sum_{v_k^m \in V_m \cap X_i} \sum_{v_j^n \in V_n \cap X_i} r(v_k^m, v_j^n) \tag{1}$$