English Series | NUMERICAL MATHEMATICS A Journal of Chinese Universities | Vol.14, No.2 May 2005

IMPROVED RANDOMIZED ALGORITHM FOR THE EQUIVALENT 2-CATALOG SEGMENTATION PROBLEM*

Yuan Yubo (袁玉波) Xu Chengxian(徐成贤)

Abstract *An improved randomized algorithm of the equivalent 2-catalog segmentation problem is presented. The result obtained in this paper makes some progress to answer the open problem by analyze this algorithm with performance guarantee. A 0.6378-approximation for the equivalent 2-catalog segmentation problem is obtained.*

Key words *combinatorial optimization, data mining, approximation algorithm, semidefinite programming.*

AMS(2000)subject classifications $O179, TP311$

1 Introduction

Jon Kleinberg, et al in [1] introduced the segmentation problem and developed an approximation algorithm for the catalog segmentation problem. This work is motivated by several applications to data mining and clustering operations. Several open problems are presented (see in [1]). One of them is how to obtain good approximation algorithms and to improve the trivial 0.5-approximation algorithm for the 2-catalog segmentation problem.

In this paper, we introduce an equivalent 2-catalog segmentation problem and then relax it to a semidefinite programming problem. By solving the semidefinite programming problem, an improved randomized algorithm is proposed. By employing this algorithm, we get the approximation optimal solution of the equivalent 2-catalog segmentation problem with a performance guarantee-0.6378.

* This work is supported by National Natural Key product Foundations of China 10231060.
This work is supported by the Younth Key Fundation of UESTC: JX04042.
Received: Mar. 27, 2003.

The paper is organized as follows: In section 2, we present a concrete example and introduce the equivalent 2-catalog segmentation problem. In section 3, we present the improved randomized algorithm for the semidefinite programming. In section 4, we analyze the algorithm with it's performance guarantee.

2 The Equivalent 2-Catalog Segmentation Problem

In this section we give the equivalent 2-catalog segmentation problem by introduing a concrete example.

Consider a company who has enough information about n customers, denoted by $C = \{c_1, c_2, \cdots, c_n\}$, and has a set of possible marketing strategies $D = \{d_1, d_2, \cdots, d_m\}$ (m is even). Any customer $c_i \in C$ likes certain marketing strategies and fail to attract others, that is, there exist a family of subsets $T = \{t_1, t_2, \cdots, t_n\}$, here t_i is a subset of D which denotes the set of all marketing strategies that the customer c_i likes. For many reasons, customers can not read the marketing strategies one by one. The company wish to create two equivalent cardinality catalogs of marketing strategies in order to maximize the sum of marketing strategies that customers like by reading one of the two catalogs. This can be stated as the following the equivalent 2-catalog segmentation problem.

The Equivalent 2-Catalog Segmentation Problem:

Given a set $D = \{d_1, d_2, \cdots, d_m\}$ (m is even) and n subsets $t_1, t_2, \cdots, t_n$ of D, find two subsets D_1 and D_2 of D with the same cardinality, so that

$$\sum_{i=1}^{n} \max\{|t_i \bigcap D_1|,\ |t_i \bigcap D_2|\}$$

is maximized, and $D_1 \bigcap D_2 = \emptyset$.

In Kleinberg's paper [1],a trivial 0.5-approximation algorithm for the equivalent 2-catalog segmentation problem is proposed and it is pointed out that it is an open problem that how to improve the trivial 0.5-approximation algorithm for the 2-catalog segmentation problem. In the next section,an improved randomized algorithm is proposed for the equivalent 2-catalog segmentation problem.

3 Improved Randomized Algorithm

In this section, we transfer the equivalent 2-catalog segmentation problem into a max-cut problem by constructing a bipartite graph. The max-cut problem is then relaxed as a semidefinite programming problem. Finally a randomized algorithm is proposed for the solution of resulting semidefinite programming problem.

The bipartite graph $G = (C, D, E)$ is constructed by the set of customer nodes C, the set of marketing strategies nodes D, and the set of edges E. If customer c_i likes marketing strategy