1000-9825/2005/16(03)0392 ©2005 Journal of Software 软 件 学 报 Vol.16, No.3

k-Median 近似计算复杂度与局部搜索近似算法分析*

潘 锐[1+], 朱大铭[1], 马绍汉[1], 肖进杰[2]

[1](山东大学 计算机科学与技术学院,山东 济南 250061)

[2](山东工商学院 信息与电子工程学院,山东 烟台 264005)

Approximated Computational Hardness and Local Search Approximated Algorithm Analysis for *k*-Median Problem

PAN Rui[1+], ZHU Da-Ming[1], MA Shao-Han[1], XIAO Jin-Jie[2]

[1](School of Computer Science and Technology, Shandong University, Ji'nan 250061, China)

[2](School of Information and Electronic Engineering, Shandong Institute of Business and Technology, Yantai 264005, China)

+ Corresponding author: Phn: +86-531-2730931, E-mail: roypanjn@163.com, http://www.sdu.edu.cn

Received 2003-09-12; Accepted 2004-07-06

Pan R, Zhu DM, Ma SH, Xiao JJ. Approximated computational hardness and local search approximated algorithm analysis for *k*-median problem. *Journal of Software*, 2005,16(3):392–399. DOI:10.1360/jos160392

Abstract: Research on the approximated algorithms for *k*-Median problem has been a focus of computer scientists, and most of the existing results are based on the Euclidean and Metric *k*-Median problem. However, results for general distance space *k*-Median has not been found for many years. In general distance space, let $d_{\max}/d_{\min}$ denote the maximum value of the length of the longest edge divided by the length of the shortest edge for one client point. In this paper, it is first proved that there are no polynomial algorithms of approximation ratio $1+\frac{\omega-1}{e}$ for *k*-Median with the condition $d_{\max}/d_{\min}\le\omega+\varepsilon$, unless $NP\subseteq DTIME(n^{O(\log\log n)})$. This result implies there are no polynomial algorithms of approximation ratio $1+\frac{2}{e}$ for Metric *k*-Median unless $NP\subseteq DTIME(n^{O(\log\log n)})$. Then a local search algorithm for *k*-Median is presented. New analysis shows that the local search can achieve a ratio of $1+\frac{\omega-1}{2}$. This result can also be extended to the Metric *k*-Median, and if $\omega\le 5$, the local search algorithm can achieve a ratio less than 3 for the Metric *k*-Median, which is better than the existing best ratio $3+\frac{2}{p}$. Finally, computer verification is used to study the real computational effect and the improved method of the local search algorithm.

Key words: *k*-median; algorithm; local search; approximation ratio; facility; client

摘 要: *k*-Median 问题的近似算法研究一直是计算机科学工作者关注的焦点,现有研究结果大多是关于欧式空间和 Metric 空间的,一般距离空间 *k*-Median 的结果多年来一直未见.考虑一般距离空间 *k*-Median 问题,设 $d_{\max}/d_{\min}$

* Supported by the National Natural Science Foundation of China under Grant Nos.60073042, 60273032 (国家自然科学基金)

作者简介: **潘锐**(1977-),男,江苏洪泽人,博士生,主要研究领域为算法分析与设计;**朱大铭**(1964-),男,博士,教授,博士生导师,主要研究领域为算法与计算复杂性,神经网络;**马绍汉**(1938-),男,教授,博士生导师,主要研究领域为算法与复杂性,人工智能,并行分布处理;**肖进杰**(1979-),男,硕士,主要研究领域为图算法.

表示 k-Median 实例中与客户点邻接的最长边长比最短边长的最大者.首先证明 $d_{\max}/d_{\min}\le\omega+\varepsilon$的 k-Median 问题不存在近似度小于 $1+\frac{\omega-1}{e}$ 的多项式时间近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$,由此推出 Metric k-Median 问题不可近似到 $1+\frac{2}{e}$,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.然后给出 k-Median 问题的一个局部搜索算法,分析表明,若有 $d_{\max}/d_{\min}\le\omega$,则算法的近似度为 $1+\frac{\omega-1}{2}$.该结果亦适用于 Metric k-Median,$\omega\le5$ 时,局部搜索算法求解 Metric k-Median 的近似度为 3,好于现有结果 $3+\frac{2}{p}$.通过计算机实验,进一步研究了 k-Median 局部搜索求解算法的实际计算效果和该算法的改进方法.

关键词: k 中间点;算法;局部搜索;近似度;设备;客户
中图法分类号: TP301 **文献标识码**: A

给定设备集合 F 和客户集合 C,$d(F_i,C_j)\in Z^+$表示设备 F_i 为客户 C_j 提供服务的费用,欲在 F 中选择 k 个设备为 C 中客户服务,并使总服务费用最小,即所谓 k 中间点问题,简称 k-Median.

距离空间对 k-Median 的计算复杂度具有显著影响,欧氏空间和 Metric 空间的 k-Median 近似算法研究近年来取得引人瞩目的进展.关于欧氏空间 k-Median,1998 年 Arora 等人[1]利用随机分割技术首次给出多项式时间近似方案;2002 年,Badoiu[2]等人在研究 Cluster 问题时,利用核心集技术给出了欧式距离 k-Median 的新多项式时间近似方案.Metric 空间中,Lin 和 Vitter[3]在 1992 年利用 Filter 技术给出了选取$(1+\varepsilon)k$ 个中心近似度是 $2\left(1+\frac{1}{\varepsilon}\right)$的结果;Charikar[4]等人于 1999 年利用线性规划技术设计了第 1 个 k-Median 近似度 $6\frac{2}{3}$的近似算法;Jain 和 Vazirani[5]于 1999 年利用线性规划和原始对偶方法求解 k-Median 获得近似度为 6 的结果;Charikar 和 Guha[6]进一步利用贪心技术将近似度改进到 4;Vijay Arya 等人[7]在 2001 年利用启发式局部搜索技术给出近似度为 $3+\frac{2}{p}$ 的局部搜索近似算法,这也是截至到目前为止 Metric 空间 k-Median 最好的结果.

1992 年,Lin 和 Vitter[3,8]曾考虑两点间距离任意赋值的 k-Median,并给出一个多项式时间算法,对于任意$\varepsilon>0$,算法选取$(1+1/\varepsilon)(\ln n+1)k$ 个设备,所得解值至多是最优解的$(1+\varepsilon)$倍.此后未见国内外有关于一般距离空间 k-Median 算法与复杂度研究结果的相关报道.考虑一个客户被设备服务的费用,记:$d_{\max}(C_k)=\max\{d(F_i,C_k)|1\le i\le n\}$,$d_{\min}(C_k)=\min\{d(F_i,C_k)|1\le i\le n\}$,往往 $d_{\max}(C_k)$与 $d_{\min}(C_k)$之间并无很大差异,且距离并不满足三角不等式.针对这一实际应用背景,本文讨论一般距离空间 k-Median 的求解复杂度和求解算法,给出如下结果.

(1) 将集合覆盖问题归约为一般 k-Median,证明 $\max\{d_{\max}(C_k)/d_{\min}(C_k)|k=1,\ldots,m\}\le\omega+\varepsilon$的 k-Median 不存在近似性能比小于 $1+\frac{\omega-1}{e}$ 的近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.进而推出,Metric 空间 k-Median 不存在 $1+\frac{2}{e}$常数近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.关于 Metric 空间 k-Median 的反面结果多年来一直未曾见到.Lin 和 Vitter[8]曾推测求解任意距离取值 k-Median 的难度与集合覆盖问题是一样的,并提出证据表明 k-Median 的结果极可能通过对集合覆盖问题进行归约得出,但未能给出严格证明.

(2) 给出了针对一般距离空间 k-Median 局部搜索算法新的分析方法.可以证明,若 $\max\{d_{\max}(C_k)/d_{\min}(C_k)|k=1,\ldots,m\}\le\omega$,则局部搜索算法的近似度不超过 $\frac{1+\omega}{2}$,算法时间复杂性是 $O(n)$.进而推出,对于 Metric 空间中的 k-Median,在$\omega\le5$ 时,算法近似性能比不超过 3.对于 Metric k-Median,Vijay Arya 等人[7]在 2001 年给出的 $3+\frac{2}{p}$ 局部搜索近似算法仍是目前所知最好的结果,但其分析方法对于一般距离空间 k-Median 已不适用,本文给出的算法求解$\omega\le5$ 的 Metric k-Median 近似度保证小于 3,该结果好于文献[7]中的 $3+\frac{2}{p}$ 的结果.

(3) 编程实现了局部搜索算法,利用实验研究了 k 和ω不同取值对算法求解质量的影响,并通过实验进一步研究了算法的改进方法.值得注意的是,实验中 88%的实例可以求得全局最优解,这进一步验证了该算法的实用

价值.

1 一般 *k*-Median 的计算复杂度

k-Median 形式化描述如下:

实例:给定设备集合 $F=\{F_1,F_2,\ldots,F_n\}$,客户集合 $C=\{C_1,C_2,\ldots,C_m\}$,正整数 $k,0<k\le|F|$.其中任意两点 $v_i,v_j\in F\cup C$ 之间距离 $d(v_i,v_j)>0$.

目标:求 $S\subseteq F,|S|\le k$,使得:$\min\sum_{v_j\in C}d(v_{\delta(j)},v_j)$,其中 $v_{\delta(j)}\in S$.

k-Median 实例中,设 $d_{\max}(C_k)=\max\{d(F_i,C_k)|1\le i\le n\}$,$d_{\min}(C_k)=\min\{d(F_i,C_k)|1\le i\le n\}$, $\frac{d_{\max}}{d_{\min}}=\max\{d_{\max}(C_k)/d_{\min}(C_k)|k=1,\ldots,m\}$.

Guha 与 Khuller[9]于 1998 年利用集合覆盖问题证明了 Metric Facility-Location 问题不存在近似性能比小于 1.463 的近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.下面利用类似方法讨论一般 k-Median 的计算复杂度.

将集合覆盖问题归约到一般 k-Median,可以证明对满足 $\frac{d_{\max}}{d_{\min}}\le\omega+\varepsilon$的 k-Median,如果存在一个近似度小于 $1+\frac{\omega-1}{e}$ 的多项式时间近似算法,那么必存在一个求解集合覆盖问题近似性能比是 $c\ln|X|,c<1$ 的多项式时间近似算法.而若后者成立,依据 Feige[10]的研究结果,则表明 $NP\subseteq DTIME(n^{O(\log\log n)})$.

集合覆盖问题形式化描述如下:

实例:给定一个集合 $X=\{x_1,x_2,\ldots,x_n\}$,X 子集构成的集合 $Y=\{Y_1,Y_2,\ldots,Y_m\}$,其中任意 $Y_i\subseteq X$.

目标:选取一个 $Y'\subseteq Y$,使 $\bigcup_{Y_k\in Y'}Y_k=X$,并最小化$|Y'|$.

下面证明中,总假设 Y'是给定集合覆盖实例的最优解,并设 $k_y=|Y'|$.由于 $1\le k_y\le m$,可以使用任意一种集合覆盖算法对每一可能 k_y 值进行计算,因此证明过程中假设 k_y 是已知的.

Feige 引理[10]. 如果存在$\varepsilon>0$,使得某个多项式时间算法可以将集合覆盖问题近似到$(1-\varepsilon)\ln n$,其中 n 是欲覆盖的元素数,那么 $NP\subseteq DTIME(n^{O(\log\log n)})$.

引理 1. 假设存在一个集合覆盖问题的实例(X,Y),其最优解是 $Y',k_y=|Y'|$,那么如果存在一个多项式时间算法 B 可以从 Y 中选取 k_y 个元素覆盖 $c'|X|$个 X 中的元素,且 $c'>\left(1-\frac{1}{e}\right)$,则 $NP\subseteq DTIME(n^{O(\log\log n)})$.

证明:对给定实例首先调用算法 B 求得一个可覆盖部分 X 元素的解,然后删除 X 中已被覆盖的元素,同时从 Y 中删除已包含在解里的元素,再次调用算法 B 寻找可以覆盖 X 中剩余元素的解.如此循环计算,直至最终 $X=\varnothing$.

假设第 i 次迭代开始时有 n_i 个元素未被覆盖,此次迭代后算法 B 选取了 k_y 个集合覆盖了 c_in_i 个元素,其中 $c_i>\left(1-\frac{1}{e}\right)$.显然有 $n_{i+1}=n_i(1-c_i)$,且 $n_1=|X|$.设在迭代到第 l 次时,$n_{l+1}=1$.此时总共选取了 $l\cdot k_y$ 个集合用于覆盖 X 中元素,即近似度是 l.显然下述关系成立:

$$n_{l+1}=1=|X|\prod_{i=1}^{l}(1-c_i).$$

于是得到:

$$\ln|X|=\sum_{i=1}^{l}\ln\frac{1}{1-c_i}.$$

依条件 $c_i>\left(1-\frac{1}{e}\right)$,可得 $1<\ln\frac{1}{1-c_i}$.因此,存在 $c<1$ 使得 $1\le c\cdot\ln\frac{1}{1-c_i}$,最终有 $l\le c\sum_{i=1}^{l}\ln\frac{1}{1-c_i}=c\ln|X|,c<1$.而根据 Feige 引理[10],这表明 $NP\subseteq DTIME(n^{O(\log\log n)})$. □

引理 1 表明,集合覆盖问题不存在多项式时间算法可以选出 k_y 个集合覆盖多于$\left(1-\frac{1}{e}\right)|X|$个 X 中元素.

定理 1. 对于 $\frac{d_{\max}}{d_{\min}} \le \omega+\varepsilon$ 的一般 k-Median,如果存在一个求解其近似度小于 $1+\frac{\omega-1}{e}$ 的多项式时间近似算法 A,那么 $NP \subseteq DTIME(n^{O(\log\log n)})$.

证明:首先将集合覆盖问题归约到一般 k-Median.设集合覆盖问题实例为 (X,Y),$X=\{x_1,x_2,\dots,x_n\}$,$Y=\{Y_1,Y_2,\dots,Y_m\}$.k-Median 实例如下构造:

(1) 取 $C=X,F=Y$.

(2) $d(F_j,C_i)=\begin{cases}1, & C_i \leftrightarrow x_i \in Y_j \leftrightarrow F_j \\ \omega+\varepsilon, & \text{其他}\end{cases}$,其中 $\omega>1,\varepsilon\ge 0$.

显然,上述 k-Median 实例不一定满足三角不等式,但满足 $\frac{d_{\max}}{d_{\min}} \le \omega+\varepsilon$.假设集合覆盖问题实例最优解是 Y',$k_y=|Y'|$,那么令归约后与其对应的一般 k-Median 中 $k=k_y$.设 $A(C,F,k)$为求解 k-Median 的算法.

设计求解集合覆盖问题的算法如下:

(1) 将给定集合覆盖问题实例(X,Y)变换为一般 k-Median 实例 I.设(X,Y)的最优解是 Y',$k_y=|Y'|$,令变换得到的一般 k-Median 实例中 $k=k_y$.

(2) while $C\ne\varnothing$,do

a) $S=A(C,F,k)$.

b) $C'\subseteq C$ 是被 $S\subseteq F$ 覆盖的客户集合,任给一个客户 $C_i\in C'$,存在一个 $F_j\in S$,使得 $d(F_j,C_i)=1$.

c) $F=F-S,C=C-C'$;

(3) 算法结束.

在上述算法求解过程中,若第 j 次迭代时$|C|=n_j$,因此时 k-Median 实例所对应集合覆盖实例的最优解解值是 k_y,所以 k-Median 实例最优解的解值应是 n_j,即 n_j 个客户中的每一个在求得的最终设备集合中可以找到一个设备,他们之间的距离是 $d=1$.如果求解一般 k-Median 算法 A 的近似度为α,那么利用算法 A 可以得出一个解,其解值至多是αn_j.

现考虑在迭代到第 j 次时,$|C|=n_j$ 个客户中有 c_jn_j 个在算法 A 求得的设备集合 S 中可以找到某个设备,两者之间的距离是 $d=1$.在剩余$(n_j-c_jn_j)$个的客户中,任意一个到 S 中设备的距离至少是 $d\ge\omega$,$\omega>1$.那么,算法 A 所求设备集合 S 的总服务费用:$Cost(S)\ge 1c_jn_j+\omega(n_j-c_jn_j)$.依 $\alpha n_j\ge Cost(S)$,可得$(\alpha-1)\ge(\omega-1)(1-c_j)$,由引理 1,$c_j\le\left(1-\frac{1}{e}\right)$,有 $\frac{1}{e}\le 1-c_j$.进而得出 $\frac{1}{e}\le\frac{\alpha-1}{\omega-1}$,即:$\alpha\ge 1+\frac{\omega-1}{e}$.

如果在任意第 j 次迭代中,$c_j>\left(1-\frac{1}{e}\right)$,则表明存在一个算法可以每次选取 k_y 个集合,覆盖 c_jn_j 个元素,而由引理 1 有 $NP \subseteq DTIME(n^{O(\log\log n)})$.

综上所述,证明了一般 k-Median 不存在近似性能比小于 $1+\frac{\omega-1}{e}$ 的近似算法. □

推论 1. Metric k-Median 不存在近似性能比小于 $1+\frac{2}{e}$ 的近似算法,除非 $NP \subseteq DTIME(n^{O(\log\log n)})$.

证明:同样将集合覆盖问题归约到 Metric k-Median,并采用定理 1 的证明方法.因 Metric k-Median 中距离是满足三角不等式的,因此在构造中,当 $C_i\leftrightarrow x_i\notin Y_j\leftrightarrow F_j$ 时,$d(F_j,C_i)$距离使用最短路方法取值,否则 $d(F_j,C_i)=1$.例如,若 $x_1,x_2\in Y_1,x_2\in Y_2$,则根据最短路方法,有 $x_1\rightarrow Y_1\rightarrow x_2\rightarrow Y_2$,因此 $d(F_2,C_1)=3$.当 $C_i\leftrightarrow x_i\notin Y_j\leftrightarrow F_j$ 时,$d(F_j,C_i)=3$ 是保证归约所得 k-Median 实例距离满足三角不等式的最小取值.设计的求解集合覆盖问题算法同定理 1.设在第 j 次迭代中,$|C|=n_j$ 个客户中有 c_jn_j 个在算法求得设备集合 S 中可以找到某个设备,他们之间的距离 $d=1$,剩余$(n_j-c_jn_j)$个客户中任意一个到 S 中设备的距离至少是 $d=3$.因此算法的解值至少是 $1\cdot c_jn_j+3(n_j-c_jn_j)$.设算法近似度是$\alpha$,由于$\alpha n_j\ge 1\cdot c_jn_j+3(n_j-c_jn_j)$,因此可得$(\alpha-1)\ge 2(1-c_j)$,由引理 1,$c_j\le\left(1-\frac{1}{e}\right)$,有 $\frac{1}{e}\le 1-c_j$.进而得出 $\alpha\ge 1+\frac{2}{e}$.同理,若在任意第 j

次迭代中,$c_j>\left(1-\frac{1}{e}\right)$,则由引理 1 有 $NP\subseteq DTIME(n^{O(\log\log n)})$. □

推论 2. 一般 k-Median 不存在任意常数近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.

证明:假设 k-Median 存在常数 K 的近似算法 A.在定理 1 的证明中令 $\omega=eK-e+1$,则根据定理 1,不存在近似性能比小于 $\left(1+\frac{\omega-1}{e}\right)=K$ 的近似算法.两者矛盾. □

2 一般 *k*-Median 的局部搜索近似算法设计与分析

利用局部搜索技术求解 k-Median 主要是对当前所选设备集合反复施加某种操作以改善解值,直到不能进一步优化为止.Korupolu 等人[11]1998 年给出一个求解 Metric k-Median 的局部搜索算法,反复对当前设备集合增加、删除和交换设备,最终输出一个包含 $k(1+\varepsilon)$个设备的集合,其局部最优解至多是全局最优解的 $3+5/\varepsilon$倍.目前所知求解 Metric k-Median 最好的算法也是由 Vijay Arya 等人[7]于 2001 年利用局部搜索技术给出的,该算法每次交换当前设备集合中 p 个设备,其算法近似度达到 $3+\frac{2}{p}$.

对给定的一般 k-Median 实例 I,设局部搜索算法输出的解是 S;I 的全局最优解为 S^*.$C_j\in C$ 由 $F_{\delta(j)}\in S$ 提供服务,服务费用 $d(F_{\delta(j)},C_j)=\min\{d(F_i,C_j)|F_i\in S\}$,即:每个 C 中客户由 S 中距离其最近的设备来提供服务.总服务费用定义为 $Cost(S)=\sum_{C_j\in C}d(F_{\delta(j)},C_j)$,$F_{\delta(j)}\in S$ 如前所述.局部搜索算法描述如下:

(1) $S\leftarrow$任意一个具有 k 个设备的子集.

(2) While 存在一个操作 op,使得 $Cost(op(S))<Cost(S)$,do

$S\leftarrow op(S)$.

(3) 返回 S.

上述算法中,$op(S)$表示对当前设备子集 S 施加某种操作后形成的新设备子集.如 $Cost(op(S))<Cost(S)$,则称操作 op 是可行的.算法中第 2 步的执行,至多有多项式数量的 op 操作被进行检查测试.此处定义 op 操作是将 S 中 p 个设备同 $F-S$ 中 p 个设备进行交换,构成新的设备子集 S'.下面证明,对于前面满足限制条件的一般 k-Median,$p=1$ 时使用此操作最终可以得到近似性能比不超过$\frac{1+\omega}{2}$的结果.

定理 2. 给定一般 k-Median 实例:F,C,以及距离 $d(F_i,C_j)>0$, $\frac{d_{\max}}{d_{\min}}\le\omega$,$op$ 操作为每次交换 1 个设备,设局部搜索算法求得解为 S,实例对应的全局最优解为 S^*,则 $\frac{Cost(S)}{Cost(S^*)}\le\frac{1+\omega}{2}$.

证明:设 $N_S(F_i)$表示接受 $F_i\in S$ 服务的客户集合;$N_{S^*}(F'_j)$表示接受 $F'_j\in S^*$服务的客户集合;$\langle F_i,F_j\rangle$表示 op 操作:"F_i 与 F_j 交换",其中 $F_i\in S,F_j\in F-S$.由于 S 是局部最优解,因此对于任意 $F_i\in S,F'_j\in S^*$,操作$\langle F_i,F'_j\rangle$导致如下不等式成立:

$$Cost(S-\{F_i\}+\{F'_j\})-Cost(S)\ge 0,\forall F_i\in S,\forall F'_j\in S^* \tag{1}$$

为了比较 $Cost(S)$和 $Cost(S^*)$,考虑 $S=\{F_1,F_2,\dots,F_k\}$中的每个设备 F_i 与 $S^*=\{F'_1,F'_2,\dots,F'_k\}$中对应设备 F'_i 进行交换操作 $op=\langle F_i,F'_i\rangle$,$1\le i\le k$,$op(S)=S'=\{F_1,F_2,\dots,F_{i-1},F'_i,F_{i+1},\dots,F_k\}$.下述不等式成立:

$$\begin{gathered}Cost(S-\{F_1\}+\{F'_1\})-Cost(S)\ge 0,\\ Cost(S-\{F_2\}+\{F'_2\})-Cost(S)\ge 0,\\ \dots,\\ Cost(S-\{F_k\}+\{F'_k\})-Cost(S)\ge 0.\end{gathered}$$

对于每次交换$\langle F_i,F'_i\rangle$后形成的具有 k 个设备的新集合 S',需要重新为客户集合 C 中的每个客户 C_j 指定一个提供服务的设备.对于 $N_S(F_i)\cup N_{S^*}(F'_i)$中的客户指定全部由 F'_i 来提供服务.而其他所有 $C-(N_S(F_i)\cup N_{S^*}(F'_i))$中的客户,依旧由交换发生前 S 中的原设备为他们提供服务.现在,交换$\langle F_i,F'_i\rangle$发生后的总服务费用变化如下:

$$Cost(S-\{F_i\}+\{F_i'\})-Cost(S)=\sum_{C_j\in N_{s^*}(F_i')}(d(F_i',C_j)-d(F_{\delta(j)},C_j))+\sum_{\substack{C_j\in N_S(F_i)\\ C_j\notin N_{s^*}(F_i')}}(d(F_i',C_j)-d(F_i,C_j))\geq 0 \tag{2}$$

对于客户 $C_j\in N_S(F_i),C_j\notin N_{S^*}(F_i')$,设在 S^*中是由 F_j' 提供服务的, $F_j'\neq F_i'$,根据限制条件,有:

$$d_{\min}(C_j)\leq d(F_j',C_j)\leq d_{\max}(C_j)\leq\omega\cdot d_{\min}(C_j)\leq\omega\cdot d(F_j',C_j),$$

$$d_{\min}(C_j)\leq d(F_i',C_j)\leq d_{\max}(C_j).$$

所以 $d(F_i',C_j)\leq\omega\cdot d(F_j',C_j)$,结合不等式(2),可以推出:

$$\sum_{C_j\in N_{s^*}(F_i')}(d(F_i',C_j)-d(F_{\delta(j)},C_j))+\sum_{\substack{C_j\in N_S(F_i)\\ C_j\notin N_{s^*}(F_j')}}(\omega\cdot d(F_j',C_j)-d(F_i,C_j))\geq 0 \tag{3}$$

对 k 个形如式(2)的不等式进行上述变换可得到 k 个形如式(3)的不等式.将此 k 个不等式进行累加,其中式(3)中的第 1 项 k 次累加后正好是 $Cost(S^*)-Cost(S)$;对于式(3)中的第 2 项,在 k 次累加后的结果至多是 $\sum_{C_j\in C}(\omega\cdot d(F_j',C_j)-d(F_{\delta(j)},C_j))$,即 $\omega\cdot Cost(S^*)-Cost(S)$.将两项累加结合后可得如下结果:

$$Cost(S^*)-Cost(S)+\omega\cdot Cost(S^*)-Cost(S)\geq 0.$$

进而推出 $Cost(S)\leq\frac{1+\omega}{2}Cost(S^*)$. □

对于上述分析,若任意 $C_k\in C$ 到 F 中设备的最长距离是 $d_{\max}(C_k)$,最短距离为 $d_{\min}(C_k)$, $\frac{d_{\max}(C_k)}{d_{\min}(C_k)}\leq\omega$,则一般距离空间 k-Median 利用上述局部搜索技术,借助交换当前设备子集 S 与 $F-S$ 中设备的操作,可以得到 $\frac{1+\omega}{2}$ 的近似度,而且每次只需交换一个设备即可,时间复杂度是 $O(n)$.

推论 3. Metric k-Median 中若 $\frac{d_{\max}}{d_{\min}}\leq\omega=3$,则存在求解其近似性能比是 2 的近似算法.若 $\frac{d_{\max}}{d_{\min}}\leq\omega=5$,则存在求解其近似性能比是 3 的近似算法.

推论 3 是很显然的.推论 1 指出 Metric k-Median 不存在 $1+\frac{2}{e}$ 的近似算法,实际上,由其证明可知, $\frac{d_{\max}}{d_{\min}}\leq\omega=3$ 的 Metric k-Median 就不存在 $1+\frac{2}{e}$ 的近似算法.目前所知 Metric k-Median 最好的近似算法近似度为 $3+\frac{2}{p}$,推论 3 说明,对于 $\frac{d_{\max}}{d_{\min}}\leq\omega=5$ 的 Metric k-Median 子问题,局部搜索算法的近似度可达到 3 以内,且此时的局部搜索算法时间复杂度比 $3+\frac{2}{p}$ 近似度算法的时间复杂度要小得多.

3 一般 *k*-Median 局部搜索算法实验分析

实验过程允许 $p\geq 1$ 的交换设备操作,目的是通过实验结果研究交换多个设备操作对最终解值的改进情况.

初始设备集合由随机选取的 k 个设备构成,局部搜索首先从 $p=1$ 的操作开始,若此操作不能改进解值,则令 $p=2$,继续局部搜索,若在此过程中出现可以改进解值的情况,那么当全部搜索完 $p=2$ 的情况后,重新从 $p=1$ 的操作开始搜索.如此循环,直至最终任何操作不能改善解值且 p 到达设置上限为止.

从表 1 和图 1 可以看出不同取值 k 和 ω 对算法求解质量的影响,k 值固定时,随着 ω 取值的逐渐增加,近似性能比呈现出上升趋势,这与定理 2 结论相符.表 1 的“Instance”列含义是:ω-客户数-设备数-k.

局部搜索算法中使用 $p>1$ 操作显然可以获得更好的结果,虽然目前不能给出 $p>1$ 情况下的具体近似度理论结果,但可以给出一个 $\omega=2$ 的一般 k-Median 实例,如图 2 所示,它同时也是一个 Metric k-Median 实例,因为距离满足三角不等式.它满足:使用 $p=2$ 操作求得的最终解优于 $p=1$ 所得结果,且该结果恰好是全局最优解.

在实验中,平均 88%的实例利用 $p=1$ 的操作求得了全局最优解;94.75%和 99.25%的实例分别利用 $p=2$ 和 $p=3$ 的操作也求得了全局最优解.这表明算法的有效性和实用性都非常高.虽然 $p>1$ 操作可以提高局部最优解的

质量,但随着 p 值的增加,这种对最终解质量的提高幅度越来越小,且算法执行时间也在增加,其时间复杂度是 $O(n^p)$.

Table 1 Result of local search algorithm

表 1 局部搜索算法结果

Instance	Local search result	Optimum result	Ratio
5-200-30-10	43 256	42 825	1.010 064
4-200-30-10	49 331	49 137	1.003 948
3-200-30-10	58 606	58 425	1.003 115
2-200-30-10	74 875	74 826	1.000 655
5-200-30-8	45 681	45 243	1.009 681
4-200-30-8	52 019	51 742	1.005 353
3-200-30-8	57 535	57 276	1.004 522
2-200-30-8	69 189	68 887	1.004 384
5-200-30-5	52 400	52 030	1.007 111
4-200-30-5	56 035	55 800	1.004 211
3-200-30-5	66 881	66 621	1.003 891
2-200-30-5	77 671	77 388	1.003 656
5-200-30-3	60 640	60 296	1.005 705
4-200-30-3	68 166	67 736	1.006 348
3-200-30-3	73 197	73 112	1.001 162
2-200-30-3	82 624	82 563	1.000 734

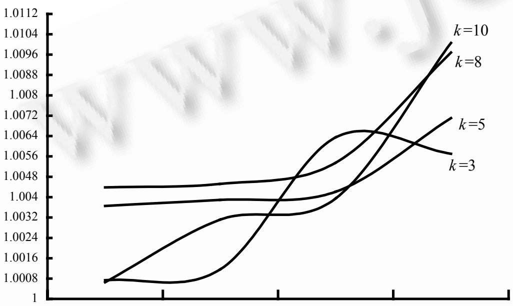

Fig.1 Approximated ratio of local search algorithm

图 1 局部搜索算法结果的近似性能比图示

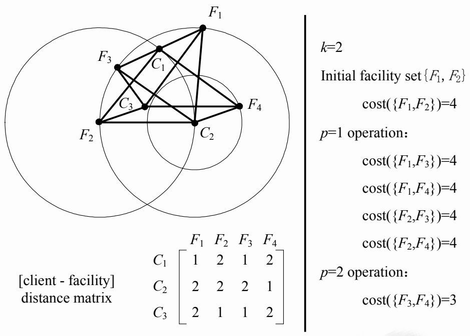

Fig.2 Special case of swapping two facilities

图 2 交换两个设备的特例

4 结束语

本文针对一般距离空间 k-Median 研究求解复杂度和求解算法.限定 $\max\{d_{\max}(C_k)/d_{\min}(C_k)|k=1,\ldots,m\}\le\omega+\varepsilon$,证明了 k-Median 不存在近似度 $1+\dfrac{\omega-1}{e}$ 的近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.进而推出,Metric k-Median 不存在近似性能比小于 $1+\dfrac{2}{e}$ 的近似算法,除非 $NP\subseteq DTIME(n^{O(\log\log n)})$.给出一个 $\max\{d_{\max}(C_k)/d_{\min}(C_k)|k=1,\ldots,m\}\le\omega$的一般 k-Median 近似度 $\dfrac{1+\omega}{2}$ 的局部搜索算法.利用该算法求解 Metric k-Median 满足$\omega\le5$ 的子问题可以保证近似度不超过 3.几年来,寻找 Metric k-Median 近似度不超过 3 的多项式近似算法一直是许多算法研究者关注的焦点,但始终未见任何进展,本文恰好找到满足上述条件的一类 Metric k-Median 子问题的多项式近似算法.实验结果进一步表明本文算法具有一定价值.

致谢 在此,我们向对本文提出宝贵意见的评审老师表示感谢.

References:

[1] Arora S, Raghavan P, Rao S. Approximation schemes for euclidean k-Medians and related problems. In: Jeffrey V, ed. Proc. of the 30th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 1998. 106−113.

[2] Badoiu M, Har-Peled S, Indyk P. Approximate clustering via core-sets. In: John R, ed. Proc. of the 34th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 2002. 250−257.

[3] Lin JH, Vitter JS. Approximation algorithms for geometric Median problems. Information Processing Letters, 1992,44(5):245−249.

[4] Charikar M, Guha S, Tardos E, Shmoys D. A constant-factor approximation algorithm for the *k*-Median problem (Extended Abstract). In: Jeffrey V, ed. Proc. of the 31th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 1999. 1−10.

[5] Jain K, Vazirani V. Primal-Dual approximation algorithms for metric facility location and *k*-Median problems. In: Alok A, ed. Proc. of the 40th Annual Symp. on Foundations of Computer Science. Washington: IEEE Computer Society, 1999. 2−13.

[6] Charikar M, Guha S. Improved combinatorial algorithms for the facility location and *k*-Median problems. In: Alok A, ed. Proc. of the 40th Annual Symp. on Foundations of Computer Science. Washington: IEEE Computer Society, 1999. 378−388.

[7] Arya V, Garg N, Khandekar R, Meyerson A, Munagala K, Pandit V. Local search heuristics for *k*-Median and facility location problems. In: Jeffrey V, ed. Proc. of the 33rd Annual ACM Symp. on Theory of Computing. New York: ACM Press, 2001. 21−29.

[8] Lin JH, Vitter JS. ε-Approximations with minimum packing constraint violation. In: Rao K, ed. Proc. of the 24th Annual ACM Symp. on Theory of Computing. New York: ACM Press, 1992. 771−782.

[9] Guha S, Khuller S. Greedy strikes back: Improved facility location algorithms. In: Howard K, ed. Proc. of the 9th Annual ACM-SIAM Symp. on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 1998. 649−657.

[10] Feige U. A threshold of ln*n* for approximating set-cover. Journal of the ACM, 1998,45(4):634−652.

[11] Korupolu M, Plaxton C, Rajaraman R. Analysis of a local search heuristic for facility location problems. In: Howard K, ed. Proc. of the 9th Annual ACM-SIAM Symp. on Discrete Algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 1998. 1−10.