分割問題と素集合データ構造

@t_uda （東北大学材料科学高等研究所 数学連携グループ）

1　分割問題

集合の分割は基礎的な概念であり，分割問題は数学に限らず学問のいたるところに現れる．分割するというのは，数学的には，集合と同値関係が与えられたときにその商集合を求めるだけのことであるが，それを計算機の中で実現しようとすると事はそう単純でない．データ（有限集合）と何らかのグループ分けの基準（同値関係）が与えられたとき，どのようなデータ構造で，またどのようなアルゴリズムで分割を実現するのがよいだろうか？

2　問題の難しさ

分類が難しい・時間がかかるという状況は想像しづらいかもしれないが，本質的な問題点は途中でどこに振り分けるべきかがそもそも分からないというところにある．もし振り分け方が最初から分かっていれば（例えば分類番号が定まっていれば）そもそも簡単で，先生が出席番号で組を作れと指示するようなものである．グループ分けの基準だけあるという前提は，全てのデータが出揃って初めて同じグループか否か判定できるということである．例えば，半径 3m 以内の人同士では情報を伝達できるとし，バケツリレー方式で情報をやり取りできる人々を同じグループとしたいとき，近くはともかく遠くの人が同じグループか否か直ちには分からないのだ[*1]．

3　Union-Find アルゴリズム

答えを述べてしまうと，素集合森というデータ構造と，それに対する Union-Find アルゴリズムが非常に有効であることが知られている．n 個のデータを分類するための操作の償却計算時間は $O(\alpha(n))$ である[*2]．ここで $\alpha(n)$ は逆アッカーマン関数でありその増大は非常に遅い[*3]．これは $\log n$ より緩やかな重複対数 $\log^* n$ よりも更に緩やかである[*4]．つまり n がかなり増えても $\alpha(n)$ はめっちゃ小さい．したがって，n 個のデータを分類するとき 1 データあたりにかかる “平均処理時間” [*5]が極めて短いということであり，大量のデータを分類する場合に適している訳である．

講演ではまずナイーブな方法を示したあと，Union-Find アルゴリズムがそれをどのように改善するかを説明する．

参考文献

[1] R. E. Tarjan, Worst-case Analysis of Set Union Algorithms, JACM, **31**, Issue 2, 1984, pp. 245–281

[*1] 無向グラフを連結成分に分解せよということである．

[*2] 正確にいえば Union と Find という 2 つの操作が相補的でありそれぞれの償却計算時間が $O(\alpha(n))$ という寸法である．これらの手続きの詳細は講演で述べる．

[*3] アッカーマン関数については巨大数に詳しい寿司屋にでも聞いて欲しい．逆寿司屋．

[*4] 歴史的には，アルゴリズムが発見されたのち $O(\log^* n)$ が上界であったが Tarjan の解析で $O(\alpha(n))$ に改善された．

[*5] 単位処理あたりにかかる時間を平均時間と言うは正確ではない．計算量の用語として平均計算時間と償却計算時間は異なる概念であるが，本稿の目的からは逸脱するため説明しやすい代替語として用いている．