# 巨大なデータから目立つものを見つける

情報学プリンシプル研究系　宇野 毅明

## どんな問題？

**巨大データの調査は人間の目では無理なので、コンピュータにさせたい**

- 何を見つけましょうか？　何を計算しましょうか？
  - ➔ 特徴は何か知りたいので、たくさんある構造を見つけよう

- 牛乳、パン、菓子
- お茶、弁当、味噌汁
- おにぎり、雑誌、お茶
- はさみ、のり、テープ
- おにぎり、お茶
- コップ、皿、箸
- 弁当、おにぎり、お茶
- 弁当、おにぎり

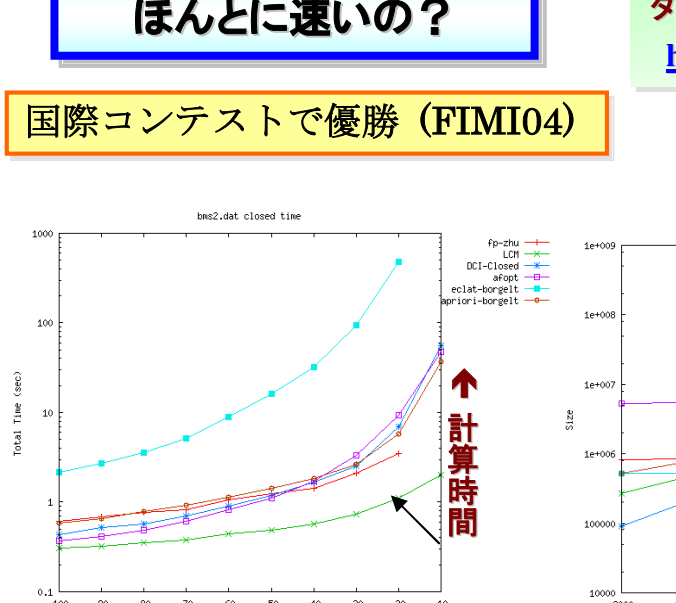
. . .

{牛乳、パン}
{お茶、弁当}
{おにぎり、雑誌、お茶}
{はさみ、のり}
. . .

売上げデータから、よく売れる組合せを見つける

- 生活習慣と病気の関係
- 客の種類と売上げの関係
- 文書の分類
- 共通の趣味を持つグループ

・・・



関係性を持つ問題設定も可能

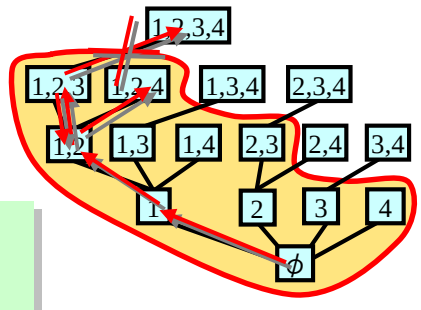
- ある反応をする化合物の特徴
- 組織図の特徴
- XMLデータの解析
- Webリンクの解析
- 企業間取引の解析

・・・

## 何が難しいの？

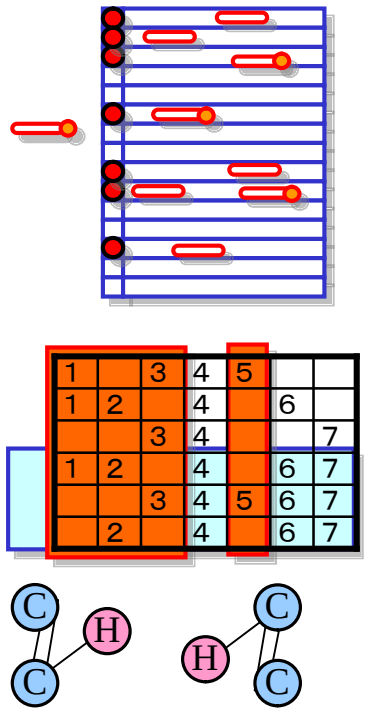
① 組合せを全て調べると計算が終わらない
  ← 50個のものの組合せは1000兆通り！
② 適当に探すと見つけ損ないが出るし、同じ物が何度も出てくる
  ← 全部メモリに取っておくと、手間がかかる
③ たくさん現れるかどうかを調べるのも手間
  ← 1つにつき 1-10秒かかる。1万個なら1日！

**計算方法を工夫して、困難を乗り越える**




- {お茶、弁当、おにぎり}を買った人は、{お茶、弁当}の組合せも買っている
  ➔ {お茶、弁当、おにぎり}がよく現れるなら、{お茶、弁当}もよく現れる
- ①② 何も無いところから、ものを1つずつ追加して、よく現れるものだけを探索できる
- ② 小さい物から順に追加するようにすると、同じ物を2回見つけることが無くなる
- ③ {お茶、弁当、おにぎり}を含む項目を調べるときは、{お茶、弁当}を含む項目だけ見ればよい
- ③ さらに1つアイテムを追加するたびにデータから不要な部分を削ると、より深い探索が速くなる

- ② 隣接性があるデータ（グラフ）は、パターンを唯一的にコード化し、同じコードを持つ物を2回見つけないようにする （木のコード化）



## ほんとに速いの？

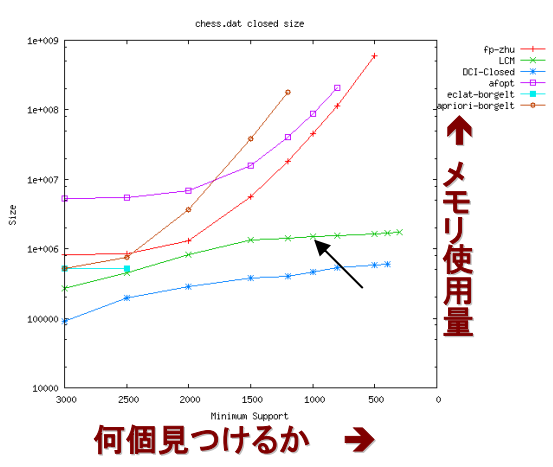
国際コンテストで優勝（FIMI04）

ダウンロード
http://research.nii.ac.jp/~uno




計算時間 ↑

何個見つけるか ➔




メモリ使用量 ↑

何個見つけるか ➔

- 100万項目を超えるようなデータでも、パソコン1台で数分で可能
- 他の方法に比べ、解が増えても時間がさほど伸びず、メモリ使用量も一定
- 多くの他分野の研究で使われている（参照件数50件以上）

| データ種別 | POS | クリック | Web閲覧 | 顧客 | 単語 |
|---|---|---|---|---|---|
| 項目数 | 51 万 | 99 万 | 7.7 万 | 8.8 万 | 6 万 |
| データサイズ | 330 万 | 800 万 | 31 万 | 90 万 | 23 万 |
| パターン数 | 460 万 | 110 万 | 53 万 | 37 万 | 100 万 |
| 計算時間 | 80 秒 | 34 秒 | 3 秒 | 3 秒 | 6 秒 |

**事実上、パソコンに乗るデータなら短時間で計算できると思っていいだろう**

## 使い心地はどうなの？

- 面白い解を見つけるには、解をたくさん見つけないといけない
  - ➔大きな解だけを出力
  - ➔含まれる項目が同じものは同一視（飽和パターン）
    どちらも効率良く見つけられる

- 「牛乳とパンを買う人はヨーグルトを買うことが多い」のような、データを説明するルールを見つけたい（アソシエーションルール）
  - ➔ほぼ同じ時間で見つけられる

- 正常なデータと異常なデータを区別するルールが知りたい（病気の人と健康な人の、習慣・遺伝子などの違い、など）
  - ➔異常なデータに多く含まれ、正常なデータにはあまり含まれないパターンを見つける

### ■ 応用研究 ■

- 遺伝子を、働いている場面が共通している物で自動分類する研究
- 遺伝子の典型的な相互作用の仕方を解析する研究
- 画像の特徴的な要素に注目して画像を自動分類する研究
- 遺伝的な病気を持つマウスの、原因となる遺伝子のパターンを見つける研究
- 遺伝病のデータベースから、典型的な遺伝子異常の組合せを見つける研究