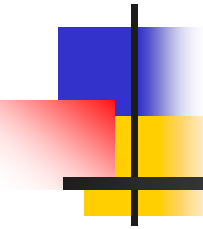


文書クラスタリングの技法ゼミ

B. 階層的クラスタリング

1. 階層的クラスタ分析法の応用
2. 転置索引ファイルの応用
3. Voorheesによる単連結法アルゴリズム

階層的クラスタ分析法の解説

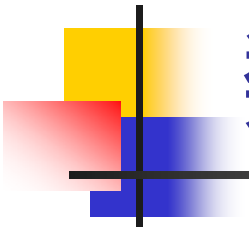
- 文書集合を予めクラスタリングしておいて、検索質問との類似度の高いクラスタを出力する方法。

- 通常　・・・検索質問と各文書との類似度を計算。
- 分析法・・・単一の文書ではなく、クラスタを検索対象として扱う。

・・・基本的には階層型・非階層型でも構わないが、前者の場合には出力文書数の調整が可能。

単連結法の利点・手順

- データを処理する順序が結果に依存しない。
- ①文書間の類似度を定義し、類似度行列を計算。
- ②類似度行列に対して、単連結法アルゴリズムを適用し、樹形図を得る。
- ③樹形図の最上層から出発し、下に降りていく。
- ④各水準で最も検索質問と一致するクラスタを選択する。また、次に進んだときに、もしもその一致度が減少したら処理を終え、一致度の最も高かったクラスタを出力。



類似度の計算方法

- 文書ベクトルの要素を2値として、余弦係数、Dice係数、Jaccard係数、重複係数等が使われる。

- 検索質問とクラスタの一致度も、基本的には同様な方法で計算される。

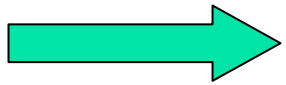
クラスタ仮説

- 2つの文書が酷似している場合、それは同一の検索質問に共に適合している可能性が高く、それに対して、酷似していない場合には、同一の検索質問に適合する可能性は低いという仮説。

- 前述のような階層的クラスタ分析を使った検索の妥当性を保証する。

単連結法について

- 一般に「偏った」出力をしやすい。

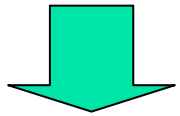


chain effect

- 情報検索の場合でも、そのような状況が観察されており、完全連結法や郡平均法等の適用も試みられている。

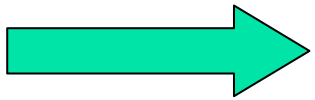
転置索引ファイルの利用

- 階層的クラスタ分析法を大規模な文書集合に適用する場合の最大の問題点は・・・



計算量！！

- 「共有する語が無ければ、それらの文書間の類似度は0になり、しかもそのようなケースが頻発する。」



いくぶん緩和される可能性

階層的クラスタ分析法を実行する方法①

- ①行列内臓法

  ・・・主記憶に類似度行列が収まる場合

- ②データ内蔵法

  ・・・そうでない場合

- ③分類行列法

  ・・・そうでない場合

階層的クラスタ分析法を実行する方法②

- ②データ内蔵法
  - 類似度行列よりも、その元データの方が小さい場合を想定したもの。
  - 元データを主記憶に格納しておき、クラスタリングの途中で類似度を計算する。

階層的クラスタ分析法を実行する方法③

- ③分類行列法
  - 類似度行列を一旦、外部記憶装置に出力して、その上でソートを実行する。
  - 結果を順次読み取って、クラスタリングを実行していく。

アルゴリズム

- ①1件の文書Diのレコードを読み込み、それに含まれる各語について、それぞれ転置索引ファイルを探索する。
- ②転置索引ファイルにはXij及び出現文書数Njの値が記録されているので、文書Diと1語以上を共有する他の文書に関して計算し、この値をファイルに書き足しておく。
- ③転置索引ファイルの探索によって、文書Di自体のノルムも計算出来るので、その値を手順②とは別のファイルに書き足しておく。
- ④上記の手順を全ての文書に対して繰り返す。

Voorheesによる単連結法アルゴリズム①

(a) nn[]: 各文書に最も近い文書の番号を記憶しておく為の配列

(b)sim[]: 各文書に最も近い文書に対する類似度を格納するための配列

(c)InHier[]: 各文書が樹形図に取り込まれたかどうかを記録するための配列

(d)sim2[]: ある1つの文書に対する他の文書の類似度を計算した結果を一時的に保存するための配列

Voorheesによる単連結法アルゴリズム②

(ⅰ)CurrId: 処理対象となる1件の文書の番号を記録するための変数

(ⅱ)MaxSim: 類似度の最大値を見つける為の記録用変数

(ⅲ)NextId: 次の処理対象となる1件の文書の番号を記録するための配列

Voorheesによる単連結法アルゴリズム④

- ①i=2…Nについてsim[i],InHier[I],nn[i]を初期化。また、第１番目の樹形図を取り込むと共にCurrIdをこの番号に初期化。
- ②もし、CurrIdが0ならば、全ての処理を終了する。そうでなければ、CurrIdが樹形図に既に取り込まれていることを記録する。
- ③CurrIdに対するほかの文書の類似度を計算し、その結果をsim2[]に記録する。
- ④MaxSim←0.0,NextId←0,及びi←1として（4-a）に進む。

Voorheesによる単連結法アルゴリズム⑤

- (4-a)もしi番目の文書がまだ樹形図に取り込まれていないならば、(4-b)に進む。そうでなければ(4-d)に進む。
- (4-b)もし、Sim2[i]>sim[i]ならば、nn[i]とsim[i]を更新する。
- (4-c)sim[i]がMaxSimよりも大きければ、MaxSimを更新し、その文書の番号をNextIdに仮に記録しておく。
- (4-d)i←i+1とする。この結果もしiがNを超えれば⑤に進む。そうでなければ(4-a)に戻る。

Voorheesによる単連結法アルゴリズム⑥

- ⑤もしNextId≠0ならば、NextIdの文書を樹形図に追加する。
- ⑥CurrId←NextIdのように更新し、②に戻る。