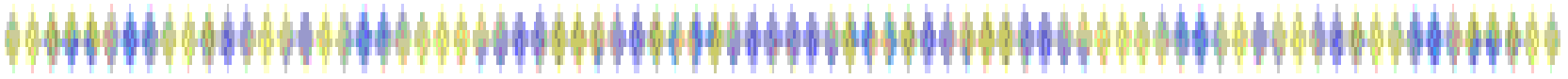
文書クラスタリングの技法ゼミ



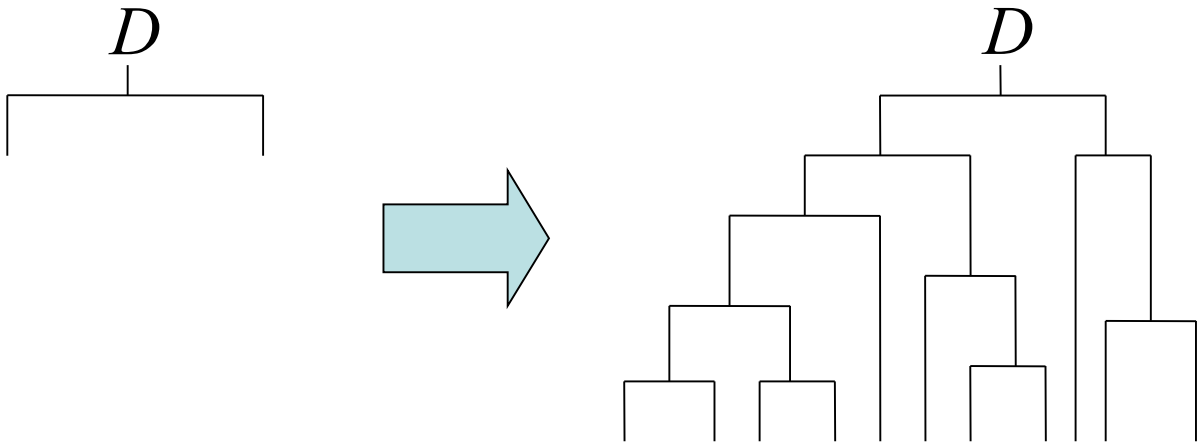
7. 分割型の階層的クラスタリング
8. 制限された凝集型クラスタリング

発表日：2007/5/16

発表者：阿部 竜之介

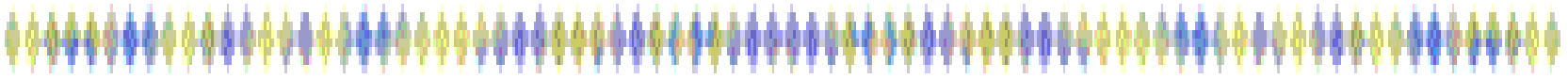
分割型の階層的クラスタリング

■文書集合全体 D から出発して、それを逐次的に分割していく



2つ以下のクラスタに分けていった場合（2分木）

bisecting k-means法



bisecting k-means法とは

- k-means法を用いた各段階での文書集合の分割法

計算量

- 単純な凝集型に比べて、減少することが期待できる
- N個の対象を2分木に構成する場合、この2分木が完全に平衡（木構造の任意の節においてその配下の節点の数が等しい）ならば、各段階でのk-means法で走査される文書数の総計は $N \log_2 N$　すなわち、計算量は $O(N \log N)$程度

bisecting k-means法の手順

(1) 分割対象の文書集合をHと表記。初期設定として、$H = D$とおく。

(2) Hの中から何らかの方法（単純無作為抽出等）によって2つの文書を選び、それを種子点とする。

(3) 2つの種子点を使って、k-means法を実行し、新たに2つのクラスタを生成。

(4) その時点でまだ分割されていないクラスタの中から、何らかの基準を使って、1つのクラスタを選び（例えば最大のクラスタを選択）、それをHとして（2）に戻る。1つもクラスタが選ばれない場合（基準を満たすクラスタがない、終了条件到達時）には処理終了。

bisecting k-means法　基準関数 1/2

■段階(3)におけるk-means法に適用

■単に大きなクラスタを選択すると、本来的に大きなクラスタを無理に分割してしまうが、基準関数を用いることで避けることが出来る

1.クラスタ内の基準関数：以下のJ_1 とJ_2 のいずれかを最大にする

$$J_1 = \sum_{k=1}^{L} \tilde{n}_k \left(\frac{1}{\tilde{n}_k^2} \sum_{d_i, d_h \in C_k} s(\mathrm{d}_i, \mathrm{d}_h) \right)$$

$$J_2 = \sum_{k=1}^{L} \sum_{d_i \in C_k} s(\mathrm{d}_i, \mathrm{m}_k)$$

bisecting k-means法　基準関数 2/2

2.クラスタ間の基準関数：以下の J_3 を最小にする

$$J_3 = \|\mathrm{D}\| \sum_{k=1}^{L} \tilde{n}_k s(\mathrm{m}_k, \mathrm{m})$$

（ D は文書集合全体をクラスタとして考えた場合のベクトル。m はその重心）

3.混合的な基準関数：以下の J_4 と J_5 のいずれかを最大にする

$$J_4 = \frac{J_1}{J_3} \qquad \text{または} \qquad J_5 = \frac{J_2}{J_3}$$

制限された凝集型クラスタリング

制限された凝集型クラスタリングとは

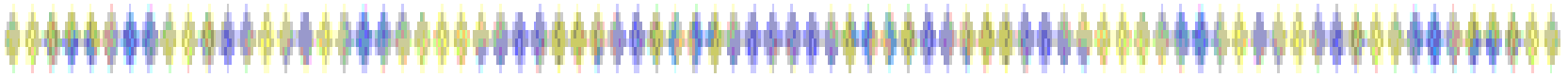
■第1段階として文書集合をいくつかのグループに分割し、各グループに対して凝集型アルゴリズムを適用する方法

第1段階の分割の具体例

■k-means法を適用するやり方（Karypisら）

■各文書を検索質問とみなして、それを除いた文書集合に対して検索を実行し、上位n 件の部分集合を1つのグループとするやり方（Smeatonら）

制限された凝集型クラスタリング（ Karypis ）1/2



考え方

通常の凝集型

最初の段階で不当な
併合が生じる

影響

分割されるべきクラスタが
まとまってしまう可能性

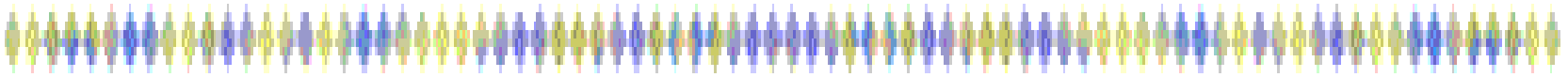
Karypisのやり方

（k-means法で）ある程度
均質なグループにまとめる

影響の波及を防ぐ

クラスタリングの質が向上する！

制限された凝集型クラスタリング（ Karypis ）2/2



利点

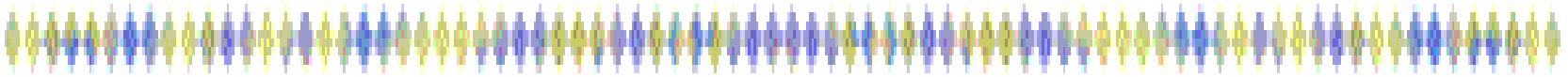
■文書集合全体に対して凝集型のアルゴリズムを適用するよりも計算量が減少する

第1段階のk-means法によって、L' 個のグループが生成され、それぞれ、$\tilde{n}_1,\ldots,\tilde{n}_{L'}$ 件の文書が含まれるとすれば、

$$O(N^2) > \sum_{k=1}^{L'} O(\tilde{n}_k^2)$$

となることが期待できる。

制限された凝集型クラスタリング（ Smeaton ）



内容補足

■第1段階として N 回の検索が繰り返され、N 個のグループが設定される（上位 n 件が1グループ）

■これらのグループは重複する可能性がある

特徴

■上位 n 件の n については、$n = 30, n = 40$ などと設定されるため、これらに対する凝集型クラスタリングにはそれほど時間がかからない

■第1段階の検索実行にはかなり時間を要する