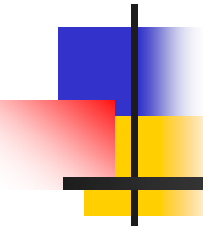
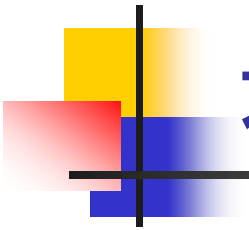


文書クラスタリングの技法

12. 遺伝的アルゴリズムの応用
13. Lightweghtアルゴリズム

遺伝的アルゴリズムの応用



- クラスタリング手法は遺伝的アルゴリズムを使って、ある基準に基づく「適切な」クラスタリングを近似的に求めようとするもの。

クラスタリングの手法

1．文書ベクトルは2値とする。

文書d_iに語が含まれる→$w_{ij}=1$

含まれない→$w_{ij}=0$

2．ある1つのクラスタのベクトルをそのクラスタ中の一定数以上の文書に出現する語によって定義する。

クラスタリングの手法2

3. クラスタに含まれる各文書のベクトルとクラスタベクトルとの類似度を定義する。

＊クラスタが適したものかどうかはそのクラスタ内での式の平均で測定される。

4. クラスタリングの成功の程度を測定する基準として使い、遺伝的アルゴリズムを適用する。

クラスタリングの手法3

5. 各文書がクラスタに属すれば1、そうでなければ0とする。

あとは、標準的な遺伝的アルゴリズムを使って最適な分割を近似的に求める。

＊パラメータとして遺伝子の数、交配率などを与える必要がある。

Lightweightアルゴリズムとは？

- CBCアルゴリズムと同様に、文書$d_i, \cdots d_N$ それぞれに対して類似度の高い順にn件の文書からなる集合を作成する。

  この文書の集合を $\tilde{D}_i$ とし、

$$\tilde{D}_i = \{d_{i(1)}, d_{i(2)}, \ldots, d_{i(n)}\}$$

と表記す。

Lightweightアルゴリズム2

- 文書d_iに対してそれが属するクラスタの番号を返す関数を$g(d_i)$と定義する。もしd_iがどのクラスタにも属さない場合にはこの関数は0を返すものとする。

実際のアルゴリズム1

（1）閾値θをきめ、i←1と設定して、文書d_iを読む。h←1とする。

（2）文書$d_{i(h)}$を読み、もし類似度$s(d_i,d_{i(h)})$の値が閾値θを超えていれば（2－a）に進む。そうでなければ（3）へ跳ぶ。

（2－a）もしd_iと$d_{i(h)}$の両方がどのクラスタにも属していなければこれらを合わせて1つのクラスタを構成し、（3）へ跳ぶ。そうでなければ（2－c）に跳ぶ。

実際のアルゴリズム2

（2－b）もし$d_{i(h)}$のみがどのクラスタにも属していなければ、d_iの属するクラスタに$d_{i(h)}$を割り当てた後、（3）へ跳ぶ。そうでなければ（2－c）に進む。

（2－c）もしd_iと$d_{i(h)}$とが同一クラスタに属していれば何もせずに（3）へ跳ぶ。そうでなければ（2－d）に進む。

（2－d）d_iの属するクラスタと$d_{i(h)}$が属するクラスタとを併合し（3）へ進む。

実際のアルゴリズム3

（3）文書集合 $\tilde{D}_i$ の中の次の文書に処理を移すために、h←h+1とする。もしh<=nならば処理する文書が残っているので（2）に戻る。

そうでなければ（4）に進む。

（4）次の文書に処理を移すために、i←i+1とする。もしi<=Nならば処理する文書が残っているのでh←1として（2）に戻る。そうでなければクラスタリング処理を終了する。

Lightweightアルゴリズムの特徴

- 基本的に文書の各組についてのN×Nの類似度行列が出発点となっている。

  この類似度行列から文書ごとに一部のデータ(n個の類似度)を取り出して、断片的に使用していく点に特徴がある。