文書クラスタリングの技法

C.次元縮約法に基づくクラスタリング

1.PDDPアルゴリズム

2.LSIの応用とIRRアルゴリズム

2007/05/23　名雪　勲

PDDPアルゴリズムとは？

- 主成分分析を利用して、文書集合を逐次的に分割していくアルゴリズムである。
- 2分木としてクラスタが階層的に構成されることになるのでPDDPは分割型の階層的アルゴリズムの一種であるといえる。

アルゴリズムの手順（1）

1．行列Wの各要素を以下で定義し、各文書の長さを1に標準化する。

$$w_{ij} = \frac{x_{ij}}{\sqrt{\sum_{j=1}^{M} x_{ij}^2}}$$

2．Wに対する転置行列の共分散行列

$$S = (W^T - ue^T)^T (W^T - ue^T)$$

を対角化する行列 $G^T SG = \Lambda$ を求める。

＊ここでuはM次元平均ベクトルを意味する。

アルゴリズムの手順（2）

3. 行列G中のN次元列ベクトルをg_1と表記する。

　これは語×文書の重み行列に対して主成分分析を実行し、「寄与率」の最も大きな第1主成分を求めたことになる。

4. 固有ベクトルg_1の第i番目の要素は、文書ベクトルd_iを第1主成分へ射影した値なので、この値が正であるグループと負である2つのグループに文書集合を分割できる。

アルゴリズムの手順（3）

5. 分割された2つの文書集合のそれぞれに対して同様な処理を反復的に繰り返す。

以上の手順で文書集合に関する2分木を構成することができる。

LSIとは？

情報検索の向上のために行列Wに対して特異値分解（SVD）を施し、文書空間の次元を圧縮する手法である。

語×文書の行列 W^T に対する特異値分解は

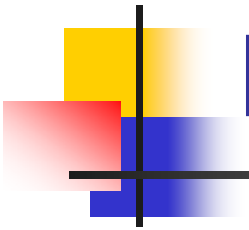
$$W^T = UQV^T$$

とかける。

変数解説

変数名	意味
U	M×Nの直交行列
Q	N×Nの対角行列
V	N×Nの直交行列

＊WのランクをrとするとQにおけるN-r個の対角要素は0である。



LSIの問題点

LSIを文書クラスタリングに応用する際は、規模の小さなクラスタがうまく抽出されない。

理由は本来LSIは全体的に影響力の小さい語を排除することによって検索機能の向上を図っている。しかしこれは同時に他の文書とあまり類似しないような文書が除かれる傾向にあるから。

IRRアルゴリズムの概念

- LSIの問題点を補うために考案された。
- 元の行列から次元を逐次的に抽出することにより、それより前の段階で取り出された次元に関連しない部分を説明するような次元の選択を可能にしている。
- 前段階で取り出された次元を差し引いた剰余を計算し、そこから次の次元を抽出する。

＊行列をここではRと表記し書く列をr_iと書く。

アルゴリズム（1）

（1）R=W^Tとする。またqをある値に定め、k←1とする。さらに前もってクラスタ個数Lを決めておく。

（2）R_sを求める（式30参照）

（3）R_s R_s^Tの固有ベクトルを計算し、そのうち最大の固有地に相当するベクトルをb_kとする。

（4）R←R－b_k b_k^T としてRを更新する。

アルゴリズム（2）

（5）k←k+1とする。k>Lならば L個のベクトルbkが求められたことになるので処理を終了する。

そうでなければ（2）に戻る。

この結果、文書ベクトルbiを

$$\overline{d}_i = (b_1, \ldots, b_L)^T d_i$$

のように次元の少ないL次元ベクトルに変換できる。このベクトルを使えばクラスタリングや特徴抽出を実行することが可能となる。