データマイニング入門

第９章 潜在意味解析

10NM733X　林華

1. 潜在意味解析

- 意味：潜在意味解析 (LSA :latent semantic analysis) : 情報検索分野において、言葉の同義性や多義性に対処するための統計的技法
- 形式：文書背後の意味構造を行列の形で表現
- 手法：表現された文書などは多変量解析に適応
- 目的：複数の語句の背後に共通して潜在する意味構造を抽出

2. 語句の検索、意味の検索

- 例１：「病院」を検索
  - 結果：「医院」、「クリニック」が除外される
  - 問題点：同義語判定できていない
- 例２：「カード」を検索
  - 結果：キャッシュカード、クレジットカード、カードゲームなど様々
  - 問題点：多義語により、意図しない結果が得た

3. 意味の定式化：語句 – 文書行列

- 対応法：複数文章の意味内容を数量的に捉える
  → 語句と文書の対応を数値化

- 作法：
  - 全ての文書に現れた語句を行に配する
  - 対象の文書を列に配し、共起行列を作成
  - 第 i 行の第 j 列目の要素は i 番目の文書に j 番目の語句の出現頻度数

4. 語句–文書行列の一例

	医師	医療	病気	患者	診察	治療	収容	施設	校正	送致	裁判所
病院	0	1	0	1	1	1	1	0	0	0	0
医院	0	0	1	0	1	1	0	0	0	0	0
クリニック	2	0	0	0	1	1	0	1	0	0	0
少年院	0	0	0	0	0	0	1	1	1	1	1

5. 特異値分解

- 語句が t 個、文書が d 個ある行列 $A_{t\times d}$ を例に
- LSA には $A_{t\times d} = T_{t\times n} \times S_{n\times n} \times D'_{d\times n}$
  - T ：語句ベクトル
  - S ：特異値で、対角行列かつ非負の値が大きい順
  - n ： Min(t,d)
  - D ：各列は右特異ベクトルで、文書の特徴
  - D' ：文書ベクトル、 $T'T=D'D=I_{n\times n}$
- 少次元で A を近似 $\bar{A}_{t\times d} = T_{t\times k} \times S_{k\times k} \times D'_{d\times k}$
  - 語句の同義性や多義性を縮減→本質意味構造取り出し

6. 類似性の検討

- 相関係数と内積などで類似性を表示

$$\cos(a,\ b)=\frac{\Sigma\, a_i\, b_i}{|\,a\,|\cdot|\,b\,|}$$

- 類似性を検討する観点としては
  - 文書間の類似性
  - 語句間の類似性
  - 語句と文書間の類似性
  - 検索質問文と元の文書間の類似性

7. 特異値分解前後のコサイン

上三角：元文書間コサイン

下三角：特異値分解後文書間コサイン

	病院	医院	クリニック	少年院
病院	–	0.52	0.34	0.20
医院	0.90	–	0.44	0.00
クリニック	0.97	0.98	–	0.17
少年院	0.40	-0.05	0.15	–

8. 検索質問文

- 検索質問文を潜在意味空間上のベクトルに変換
- X=TSD' 転置 X'=DST' → X'T=DS→ D=X'TS^-1
- 検索質問文をｋ次元潜在意味空間上での表現 ~q=q'Tt×kS^{-1}k×k
- 検索文例：「病気を治療する施設」

医師	医療	病気	患者	診察	治療	収容	施設	校正	送致	裁判所
0	0	1	0	0	1	0	1	0	0	0

病院	医院	クリニック	少年院
0. 999	0. 833	0. 945	0. 282

- 高次元潜在意味空間への追加 ~t=t'Dd×kS^{-1}k×k

9. LSA の可能性

- 知識の獲得
- 小論文の自動採点
- 記憶の研究
- 音声・画像の意味的解析
- データベースの改良

10. 語句文書行列の作成 - その 1

- 通常：各セルにある文書における単語の頻度
- LSA ：語句と文書間の関係を重み関数による

$$a(i, j) = L(i, j) \times G(i)$$

- L(i, j) ：セルに適応　G(i) ：語句 (行) に対応
- 局所的な重み関数 L(i, j) の種類
  - L(i, j) = tf(i, j) 文書 j における語句 i の頻度
  - L(i, j) = log(tf(i, j) + 1)　L(i,j)=1(tf>0) or 0(tf=0)

11. 語句文書行列の作成 - その 2

- 大域的重み関数 G(i, j) ：各語句の文書間の影響

$$G(i)=\frac{1}{\sqrt{\sum_j L(i,j)^2}} \qquad G(i)=\frac{gf(i)}{df(i)}$$

$$G(i)=1+\frac{\sum_j p(i,j)\log p(i,j)}{\log ndocs}$$

$$G(i)=1+\log(ndocs\ /df(i))$$