ベクトル空間モデルの概要

佐々木稔

ベクトル空間モデル

- 文書情報の数理モデルのひとつ
- 文書や検索質問をベクトル空間内の「点」と表現

$$\mathbf{d} = (t_1, t_2, \cdots, t_n)$$

- 通常、文書は大量に存在するので行列表現となる

$$\mathbf{d} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{bmatrix}$$

ベクトル空間モデルの例(1/4)

D1:ソフトバンク、ボーダフォンと新会社
D2:ソフトバンク、ボーダーフォングループと合弁会社
D3:ボーダフォンは「ソフトバンクモバイル」に
D4:ソフトバンク、ボーダフォンと合弁・資本金最大110億円
D5:ソフトバンク、ボーダフォンを社名変更「ソフトバンクモバイル」に

- 索引語
  - t1 :ソフトバンク
  - t2 :ボーダフォン
  - t3 :会社
  - t4 :合弁
  - t5 :モバイル
  - t6 :資本金
  - t7 :最大
  - t8 :110億円
  - t9 :社名
  - t10:変更

ベクトル空間モデルの例(2/4)

- 文書集合の行列表現

$$\begin{bmatrix} D1 \\ D2 \\ D3 \\ D4 \\ D5 \end{bmatrix} = \begin{bmatrix} 1110000000 \\ 1111000000 \\ 1100100000 \\ 1101011100 \\ 2100100011 \end{bmatrix}$$

ベクトル空間モデルの例(3/4)

- 検索質問:「ソフトバンクモバイル」

$$q = (1, 0, 0, 0, 1, 0, 0, 0, 0, 0)$$

- 類似度計算(cosine)

$$\cos(D1, q) = 0.408$$
$$\cos(D2, q) = 0.354$$
$$\cos(D3, q) = 0.816$$
$$\cos(D4, q) = 0.289$$
$$\cos(D5, q) = 0.750$$

ベクトル空間モデルの例(4/4)

- 類似度の高い順に並び替える

順位	文書	類似度
1	D3	0.816
2	D5	0.750
3	D1	0.408
4	D2	0.354
5	D4	0.289

索引語(キーワード)の抽出

- 単語
  - 文書内容を特徴付ける単語を抽出
    - 名詞、動詞、形容詞など
    - 助詞、助動詞、冠詞、前置詞などは無関係
- 単語の抽出方法
  - 英語、フランス語は区切りが明白で抽出が容易
  - 日本語、中国語は分かち書きされないので抽出が困難
- 形態素解析
  - 日本語、中国語の文を解析する技術
  - 単語分割、品詞や語形変化などの情報を与える

単語以外の索引語

- 形態素解析の利用
  - 文字列を細かい単語に分割
  - 「自然言語」→「自然」、「言語」
- 他の索引語
  - *N*-gram ： *N* 文字単位の文字列を索引語とする
  - ユニグラム ：「自」、「然」、「言」、「語」
  - バイグラム ：「自然」、「然言」、「言語」
  - トライグラム ：「自然言」、「然言語」

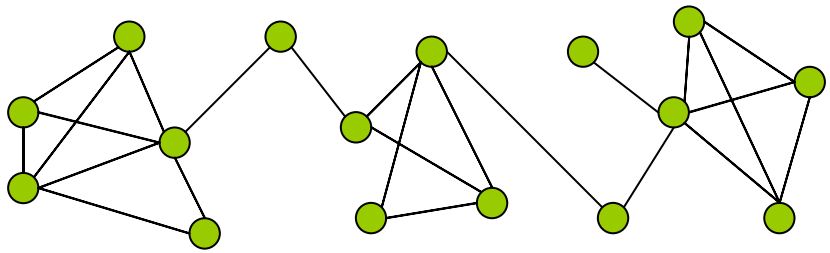
不要語リスト

- 索引語として適当ではない単語を除去
  - 機能語 ： 単語間の関係を表す単語
  - 助詞（「は」、「の」など）、冠詞（”the”, “a” etc.）
- 不要語リスト(stop words)
  - 不要語とする単語の一覧

- **内容語を残す**

グラフによるキーワード抽出

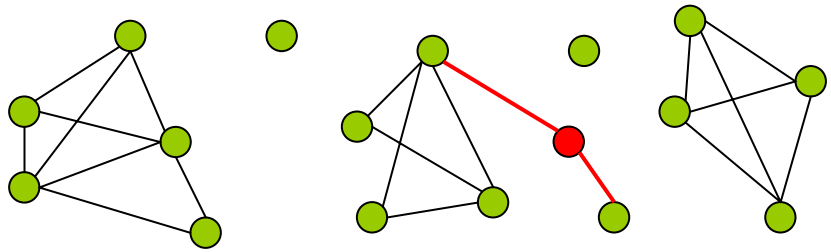
- 文書内での単語
  - 出現頻度、共起関係によりグラフ化
- 土台の形成
  - 出現頻度が上位の単語集合を抽出
    - この要素をグラフのノードとする
  - 共起頻度の高い単語対を枝で結ぶ
  - 極大連結部分グラフのみを残す
    - 枝 e_{ij} を除いてもノード i から j に到達できる枝を残す



グラフによるキーワード抽出

- 主張の抽出
  - 文書中の単語 *w* が土台に支えられる力を算出

- 主張 key(w) の算出方法
  1. 単語 w と土台 PG の共起頻度 co(w, PG) を計算
  2. 土台 PG と co(w, PG) の合計を key(w) とする



索引語の重み付け

- 抽出した索引語
  - 文書内容との関連に差が存在
  - 重み付け　：　索引語の文書内容に対する重要度
- 重み付けの種類
  - 局所的重み L_{ij}
    - ひとつの文書 D_j 内における索引語 w_i の重要度
  - 大域的重み G_i
    - 文書集合全体に対する索引語 w_i の重要度
  - 正規化 N_j
    - 文書の長さ（索引語数）による影響をなくす

局所的重み

- バイナリ重み（f_{ij} は文書 j での索引語 i の頻度）

$$L_{ij} = \begin{cases} 1 & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases}$$

- 頻度

$$L_{ij} = f_{ij}$$

- 対数頻度

$$L_{ij} = \log\left(f_{ij} + 1\right)$$

- 拡大正規化頻度

$$L_{ij} = \begin{cases} 0.5 + 0.5 \log \dfrac{f_{ij}}{\max\limits_{k} f_{kj}} & (f_{ij} > 0) \\ 0 & (f_{ij} = 0) \end{cases}$$

大域的重み

- IDF (Inverse Document Frequency)

$$G_i = \log \frac{N}{n_i}$$

- 確率的 IDF

$$G_i = \log \frac{N - n_i}{n_i}$$

- 大域的頻度IDF（F_i は文書全体の索引語 i の頻度）

$$G_i = \frac{F_i}{n_i}$$

- エントロピー

$$G_i = 1 + \frac{1}{\log n} \sum_{j=1}^{n} \frac{f_{ij}}{F_i} \log \frac{f_{ij}}{F_i}$$

正規化

- コサイン正規化

$$n_j = \sqrt{\sum_{i=1}^{m} \left(L_{ij} G_i \right)^2}$$

索引語の重み

- 文書 j での索引語 i の重要度

$$d_{ij} = \frac{L_{ij} G_i}{n_j}$$