D2-27

表からの類義語・同義語抽出

関 恒仁† 嶋田和孝‡ 遠藤 勉‡

† 九州工業大学大学院情報工学研究科 ‡ 九州工業大学情報工学部

1 はじめに

Web の情報を検索したり，利用する場合の問題点として，テキスト中の表記の揺れや類義語等の存在が挙げられる．この問題の一般的な解決法は，同義語や類義語をまとめたシソーラスを利用する事である．しかしながら，人手によって構築された既存のシソーラスには，日々生成されている新しい語や，専門性の高い語などは，殆ど記載されていないという問題がある．したがって，単に類義語をまとめただけのシソーラスだけではなく，新語や専門用語についてのシソーラスを自動構築する必要がある．

本稿では，多くの情報を保持し，更新が速い情報源である Web 上の情報の中で，表をコーパスとして利用する事により，シソーラスの自動構築を目指して，類義語・同義語の抽出を行う手法を提案する．

2 対象語のベクトル化

一般に，類義語・同義語抽出は，コーパス中での語間の共起情報を利用して，語と語の関連度を定量的に評価する手法が多く用いられている．本稿では，表形式のデータを処理対象とする．図 1 に示す様に属性と属性値の関係を用いて，ベクトル空間を作成し，類義語・同義語の抽出を行う．属性部分の語を類義語・同義語を抽出するための対象語とし，それに対応する属性値を用いて，その対象語をベクトルで表現する．

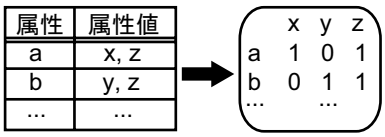

図 1: 表からのベクトル空間生成

具体的な処理の流れとしては，属性値に現れる文字列を形態素解析し，不要語削除処理を行い，ベクトル空間構成要素となる索引語を生成する．索引語を要素とする対象語ベクトルを作成するとき，各対象語に共起する索引語の頻度情報に統計的な重みを加え，その数値をベクトルの要素とする．なお，数字は全て同じものとして扱い，数字を持つか持たないかの判別のみを行う．ここでの重み付けは，局所的重み L_{ij} を対象語 w_j に対する索引語 t_i の重み，大域的重み G_i を全データにおける索引語 t_i の重みとすると，それぞれ以下の様に表す．

$$\begin{cases} L_{ij} = log(1 + f_{ij}) \\ G_i = \frac{F_i}{n_i} \end{cases} \tag{1}$$

ここで，f_{ij} は対象語 w_j に共起する索引語 t_i の頻度，F_i は全データ集合における索引語 t_i の頻度，n_i は索引語 t_i と共起する対象語数を表す．これにより，対象語 w_j を表すベクトルにおける索引語 t_i を表す要素は，L_{ij} と G_i の積で求められる．

3 対象語・索引語行列の潜在的意味解析

潜在的意味解析は，以下の様な行列の特異値分解に基づき，ベクトルの次元を圧縮する技術である [1]．次元を圧縮する事によりデータ中のノイズを取り除き，より本質的な情報を捉える事ができる．

前節における処理を行う事により，n 個の対象語と m 個の索引語から，$m \times n$ の対象語・索引語行列 D が得られる．このとき，行列 D の階数が r であるとすると，D の特異値分解は式 (2) で定義される．

$$D = U\Sigma V^T \tag{2}$$

ここで，U と V は $U^TU = V^TV = I$(単位行列) を満たす直交行列，Σ は $m \times n$ 行列であり，$\Sigma = diag(\sigma_1, \ldots, \sigma_n)$, $(\sigma_1 \geq \cdots \geq \sigma_r \geq \sigma_{r+1} = \sigma_n = 0)$ を満たす対角行列で，この σ_i $(i = 1 \ldots r)$ は D の特異値と呼ばれる．ここで，k 個の大きな特異値 $\sigma_1, \ldots, \sigma_k (k < r)$ だけを使って，式 (3) で示す近似行列 D_k を得る事が出来る．

$$D_k = U_k \Sigma_k V_k^T \tag{3}$$

4 対象語のクラスタリング

これまでの処理で得られた対象語ベクトルについて，類義語・同義語となる語をいくつかのクラスタに自動分類を行う．クラスタリングアルゴリズムとしては，球面 k 平均アルゴリズム [2] を利用する．球面 k 平均アルゴリズムは，ユークリッド空間内における概念ベクトルと対象語ベクトル間の内積を類似度として，多次元空間の単位円を分割する事によりクラスタリングを行うものである．ここで，あるクラスタ π_j の概念ベクトル $\mathbf{c}_j$ は，π_j に属する対象語ベクトルの重心を正規化したものである．

本アルゴリズムは各クラスタ π_j $(1 \leq j \leq k)$ の密度を式 (4) で評価する．

$$\sum_{\mathbf{x} \in \pi_j} \mathbf{x}^T \mathbf{c}_j \tag{4}$$

クラスタの結合密度の総和を目的関数として，式 (5) に示す目的関数 Q が局所的に最大になるまで，クラスタリングが繰り返される．

$$Q = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \pi_j} \mathbf{x}^T \mathbf{c}_j \tag{5}$$

5 実験

異なるメーカ 7 社のパソコンのスペックに関する表 20 個を用いて，類義語・同義語抽出の実験をおこなった．実験データに特異値分解を行った結果，218 個の特異値が得られた．その内上位 170 個の特異値を用いて近似行列を作成した．最後に，球面 k 平均アルゴリズムを用いて，対象語のクラスタリングを行った．$k = 55$ とし，55 個のクラスタに対象語を分類した．類義語・同義語の多くが同じクラスタに分類されており，良好な結果が得られた．

6 おわりに

表の構造を利用して，類義語・同義語の抽出を行った．実験結果から提案手法の有効性が確認された．しかしながら，1 つのクラスタに複数の類義語・同義語グループが存在するという問題もあった．これは，クラスタリング時のクラスタ数の決定が原因であると考えられる．このクラスタ数の決定は非常に困難であり，本実験においてもクラスタリング結果を参考に人手で調整を行った．

今後の課題としては，様々な表への適用を目指すためにクラスタリングアルゴリズムにおいて，情報基準を取り入れる等してクラスタ数の自動決定を行う必要があると考えられる．

参考文献

[1] 北研二, 津田和彦, 獅々堀正幹.“ 情報検索アルゴリズム”. 共立出版,2002.

[2] I. S. Dhillon and D. S. Modha,“ Concept decompositions for large sparse text data using clustering” Technical report, IBM Almaden Research Center,1999.