# シソーラスを用いた情報間類似性評価手法について

5M−04

後藤 将志　大囿 忠親　新谷 虎松
名古屋工業大学知能情報システム学科
e-mail: {shoji, ozono, tora}@ics.nitech.ac.jp

## 1 はじめに

本論文では，テキストで表現される情報を対象として，シソーラスを用いた情報間の類似性評価手法を提案する．情報や情報源間の類似性評価は，推薦システムや分散情報検索 [1] などにおいて重要な役割を果たす．

テキスト情報の類似性を計算するとき，従来は，主に語の出現頻度を利用したベクトル空間モデルに基づく手法が用いられていた．ベクトル空間モデルに基づく手法では，テキスト情報を多次元空間上のベクトルとして表現し，二つのベクトルを比較することにより類似度を計算する．ベクトルの各次元は索引語が対応しており，各成分の値には，その情報における索引語の重みを割り当てられる．ベクトル空間モデルでは，語の出現頻度のみに基づいて語の重みを決定する．そのため，同一の語であっても異なった意味で出現することがある多義語が存在した場合，計算に誤差が発生することがある．

本論文では情報をシソーラスで表現し，シソーラス間の類似性を評価することにより，情報間の類似性を評価する手法を提案する．シソーラスとは語の関係を表現した辞書である．語の関係は，その語が持つ意味によって異なる．多義語が異なった意味で出現するテキスト情報から構築されたシソーラスには異なった語の関係が現れる．本手法では，語義を他の語との関係によって区別することにより多義語の問題を解消する．

本論文では，まずテキスト情報の構造化のために本提案手法で利用する，代表的なシソーラス自動構築手法について述べ，次に，本論文で提案するシソーラスで表現された情報間の類似性評価アルゴリズムを説明する．そして，類似性評価に関する実験結果を述べ，最後にまとめる．

## 2 シソーラスの構築

テキスト情報からのシソーラスの自動構築に関する研究は現在，活発に行われている．しかし，語の同義関係や上位/下位関係を自動的に判別することは大変困難であり，自動構築されたシソーラスでは，単純な語の関連のみを表現するのみにとどまっている．このため，自動構築されたシソーラスは，木構造ではなく，より複雑なグラフ構造を持つ．現在主に提案されているシソーラスの自動構築手法は共起関係に基づいている．つまり，同一の場所に現れる語は関係があるという仮定に基づいて，シソーラスを自動構築する．共起関係に基づくシソーラス構築では，共起した語の類似尺度によって，さまざまな手法が提案されている．現在提案されている手法は大きく分けて二つに分類される．一つは語の出現頻度に重きをおいたもので，これには Jaccard 類似度，Dice 類似度などが含まれる．これら手法では，低い出現頻度の語に関して信頼に欠けるという data sparseness 問題が起こると指摘されている [2]．もう一つが語の関係に重きをおいた手法である．本手法では，語の関係に重きをおいた手法として，二つの語の出現についての条件付き確率の平均値を利用する [2]．この類似尺度は以下の式で定義できる．

$$s_{xy} = \frac{P(x|y) + P(y|x)}{2} \tag{1}$$

ここで，$P(x|y)$ は，語 $y$ の出現ときに，語 $x$ が出現する条件付き確率である．

## 3 情報間の類似性評価手法

本節では，それぞれのテキスト情報から構築されたシソーラスを元に，情報間の類似性を評価するアルゴリズムを提案する．構築されたシソーラスはグラフ構造を持つため，各行と列にシソーラス中の語を対応させ，各要素に，式 (1) で計算される，その行と列が対応する語同士の関連度が入れることにより，式 (2) のような行列で表現することができる．

$$\begin{array}{c} \\ t_1 \\ t_2 \\ \vdots \\ t_m \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \\ \begin{pmatrix} s_{t_1t_1} & s_{t_1t_2} & \cdots & s_{t_1t_m} \\ s_{t_2t_1} & s_{t_2t_2} & \cdots & s_{t_2t_m} \\ \vdots & \vdots & \ddots & \vdots \\ s_{t_mt_1} & s_{t_mt_2} & \cdots & s_{t_mt_m} \end{pmatrix} \end{array} \tag{2}$$

本提案手法では，テキスト情報から作成されたシソーラス同士の類似性を評価することにより，情報間の類似度を計算する．各列ベクトルは，ある語と他の語の関連度を表すベクトルである．ある語について，関連する語の傾向が二つの情報間で類似しているとき，それぞれの情報で用いられているその語の語義は同一であるといえ，その二つの情報は類似していると判断される．本提案手法では，二つのシソーラスで，同じ列のベクトル間の余弦の値を計算し，その平均値を二つの文書間の類似度とし，類似度計算で利用する．

---

A Method for Assessment of Similarity among Information using Thesaurus
Shoji GOTO
Tadachika OZONO
Toramatsu SHINTANI
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya, 466-8555, JAPAN

テキスト情報から構築されたシソーラスでは，構築に利用した情報によって行と列に対応する語の種類が異なる．そこで，まず行と列が対応する語を統一する．ここで，二つのテキスト情報 $d_1, d_2$ について考える．それぞれのテキスト情報から得られる語の集合を $W_1, W_2$ としたとき，$W_1 \cup W_2 = \{w_1, \cdots, w_n\}$ とする．このとき，$d_1, d_2$ から構築されたシソーラスを元に，$W_1 \cup W_2$ の要素を各行と列に対応させた行列を作成する．それらを $\boldsymbol{T}_1, \boldsymbol{T}_2$ とすると，それぞれ式 (3)，(5) のようになる．

$$\boldsymbol{T}_1 = \langle \boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_n \rangle \tag{3}$$

$$\boldsymbol{a}_i = \begin{cases} \langle \mathrm{tf}^1_{w_i}, \mathrm{s}^1_{w_1 w_i}, \cdots, \mathrm{s}^1_{w_n w_i} \rangle^T & (w_i \in W_1) \\ \langle 0, \cdots, 0 \rangle^T & (w_i \notin W_1) \end{cases} \tag{4}$$

$$\boldsymbol{T}_2 = \langle \boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_n \rangle \tag{5}$$

$$\boldsymbol{b}_i = \begin{cases} \langle \mathrm{tf}^2_{w_i}, \mathrm{s}^2_{w_1 w_i}, \cdots, \mathrm{s}^2_{w_n w_i} \rangle^T & (w_i \in W_2) \\ \langle 0, \cdots, 0 \rangle^T & (w_i \notin W_2) \end{cases} \tag{6}$$

ここで，$\mathrm{s}^k_{w_i w_j}$ は，$d_k (k = 0, 1)$ から構築されたシソーラスにおける語 $w_i, w_j$ の関連度である．$\mathrm{tf}^k_{w_i}$ は $d_k$ における，語 $w_i$ の出現頻度 (term frequency) である．この作業では，$d_1$ から構築されたシソーラスに対し，$d_1$ 中で出現しなかった語に対応する列に $\mathbf{0}$ ベクトルを挿入したものが $\boldsymbol{T}_1$ になる．これにより，二つの文書から構築されたそれぞれのシソーラスを表現した行列を元に，行と列に対応する語が同一の行列が得られる．

次に，この二つのシソーラス $\boldsymbol{T}_1$ と $\boldsymbol{T}_2$ を利用して情報 $d_1, d_2$ の類似度を計算する．そこで，関連する語の傾向の類似度を計算するため，列ベクトルの類似度としてベクトル間の余弦を計算する．$d_1$ と $d_2$ の類似度 similarity$(d_1, d_2)$ は式 (7) で計算される．

$$\mathrm{similarity}(d_1, d_2) = \sqrt{2f(d_1, d_2) - f^2(d_1, d_2)} \tag{7}$$

$$f(d_1, d_2) = \frac{\sum_{i=1}^{|W_1 \cup W_2|} \frac{(\boldsymbol{a}_i, \boldsymbol{b}_i)}{||\boldsymbol{a}_i|| \cdot ||\boldsymbol{b}_i||}}{|W_1 \cup W_2|} \tag{8}$$

ここで，$(\boldsymbol{a}_i, \boldsymbol{b}_i)$ はベクトル $\boldsymbol{a}_i$ と $\boldsymbol{b}_i$ の内積，$||\boldsymbol{a}_i||$ はベクトル $\boldsymbol{a}_i$ のノルム，$|W_1 \cup W_2|$ は語の集合 $W_1 \cup W_2$ の要素数を表す．

## 4 実験結果

本節では，本論文で提案した類似性評価手法に関する実験結果を示す．本実験では，Text categorization のテストコレクションとして公開されている，Reuters-21578[3] のうちの 2000 文書を利用し，それらの文書の組合わせに関して類似度を計算した．本実験では，ある文書対が同一カテゴリに属している場合，その文書対はお互い類似しているとし，本アルゴリズムにより計算された類似度の n-best の内，同一カテゴリに属する文書対の割合を適合率として評価を行った．

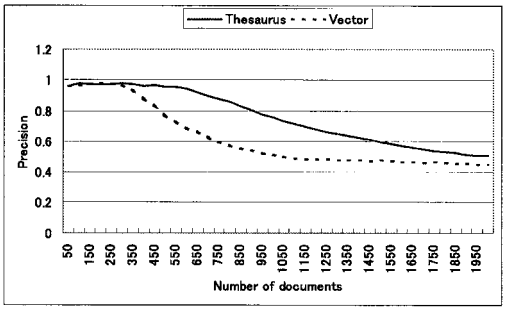


図 1: 実験結果

本実験では，比較対象として，従来のベクトル空間モデルにおける余弦を利用した類似度計算手法を利用し，本手法と比較した．実験結果を図 1 に示す．図 1 において，横軸は n-best の n の値，つまり，選択した文書数，縦軸が適合率を表す．本実験結果では，前半部は本手法とベクトル空間モデルに基づく手法とで，大きな差が無いが，文書数が多くなるにつれて，適合率に大きな差が現れた．この結果から，本提案手法ではベクトル空間モデルに基づく手法に比べ，高い適合率を得ることができることを示すことができた．

## 5 おわりに

本論文では，テキスト情報間の類似性の手法として，シソーラスを利用した計算手法を提案した．従来の，ベクトル空間モデルでは多義語の問題があったが，本手法では，シソーラスを利用することにより，語の意味を，他の語との関連によって区別することにより，多義語の問題に対処した．本研究では，Text categorization における同一カテゴリ内の文書の類似性を評価する実験を行い，ベクトル空間モデルより優れた結果が得られることを示した．

## 参考文献

[1] Shoji Goto, *et al.*: "A Method for Information Source Selection using Thesaurus for Distributed Information Retrieval", Proc. of the Pacific Asian Conference on Intelligent Systems 2001, pp.272–277, 2001.

[2] Young C. P. *et al.*: "Automatic thesaurus construction using Bayesian networks", Proc. of the International Conference on information and knowledge management, pp.212–217, 1995.

[3] David D. L.: "Reuters-21578 text categorization test collection", `http://www.research.att.com/~lewis/reuters21578.html`, 1997.