自 然 言 語 処 理 152－15
（２００２．１１．１３）

事物間関係の推定における文書内頻度を考慮した補完類似度の性能評価

山本 英子　　内山将夫　　井佐原 均
独立行政法人　通信総合研究所
{eiko, mutiyama, isahara}@crl.go.jp

本研究では，文字認識の分野で用いられている補完類似度をテキストコーパスから事物間の関係を推定する問題に適用する際に，事物が持つ各文書における頻度を考慮した場合を考える．補完類似度は，ベクトルで表された文字の画像パターンの類似度を測ることによって劣化印刷文字を認識するために経験的に開発された尺度である．この扱うベクトルをコーパス中の事物の出現パターンに置き換えると，補完類似度は事物間関係の推定に適用できる．そこで，これまでに二値ベクトルを対象として事物間関係の推定を行った．しかし，二値ベクトルでは，Document Frequency しか考慮しておらず，Term Frequency(文書内頻度)を考慮していない．そこで，Term Frequency を考慮した多値ベクトルを対象とした補完類似度を用いて事物間関係の推定を行った．その結果，Term Frequency を考慮した補完類似度のほうが推定能力が高かったことを報告する．

Estimation of Relationship between Entities by Complementary Similarity Measure Considering Term Frequency

Eiko Yamamoto　Masao Utiyama　Hitoshi Isahara
Communications Research Laboratory

In this paper, we applied CSM (Complementary Similarity Measure) considering term frequency to estimate relationship between entities. Here, term frequency is times that certain entity appears in a document. CSM was developed experientially for robust character recognition. This measures inclusion degree of vectors expressing character image pattern. We have even estimated relationship between entities by replacing the image pattern to occurrence pattern of entity in corpus. However, we have considered only document frequency and have not considered term frequency. From experimental results, we reported that CSM considering term frequency obtained higher performance than original CSM.

1　はじめに

近年，電子化された情報があふれ，そこに表されている事物の関係を推定する研究がこれまでに多くなされている．しかし，これらの研究の多くでは，事物間の関係を暗黙的に一対一関係と想定している．これは，関係を持つ事物は共起する関係にあるという前提に基づいているためである．しかし実際には，事物が一対多関係にある場合があり，この特徴を捉えるために工夫が必要である．ここで，一対多関係にある事物の出現パターンを観ると，パターンは一致するのではなく，包含関係にあることが多く観察できる．そこで，この出現パターン間の包含関係を抽出することができる類似尺度を探し，文字認識の分野で有効であるとされる補完類似度[4]に着目した．この着眼点を基に，これまでに補完類似度を事物間の一対多関係の推定に適用し，文献[2]に記載された一般に知られている尺度に比べ，有効であることを報告した[7]．この報告の際に用いた事物の出現パターンは二値ベクトルで表した，文書内頻度情報を考慮しないものであった．情報検索や情報抽出において，事物の文書内頻度は重要な情報源である．そこで本研究では，事物の文書内頻度情報を考慮した出現パターンを用いた補完類似度が事物間関係を推定する問題に，より有効であったことを報告する．

2　補完類似度の関係推定への適用

本研究で用いる補完類似度とは，文字認識の分野で有効とされている類似尺度である[2]. この尺度は文字を画像特徴として扱い，劣化印刷文字の画像パターンとテンプレート文字の画像パターンとの類似度を測ることによって文字認識を行う補完類似度法に用いられる．この手法は文字の汚れやかすれに強い特長を持ち，かすれにおいては人による認識よりも高い精度を得られることが報告されている[4]．これは，劣化印刷文字の画像パターンがテンプレート文字のパターンに包含される形であれば，文字であると認識できるように考案された類似尺度である．本研究ではこれまでに，二値画像のための補完類似度において二値ベクトルで表現された画像特徴のパターンを単語の出現パターンに置き換え，事物間の一対多関係を推定する問題に適用した[7]. ここで，二値画像のための補完類似度の定義式を示す．

定義 1　二値画像のための補完類似度

二つの画像 *Pic1, Pic2* がそれぞれ二値ベクトル $\vec{F}=\{f_1,f_2,...,f_n\}(f_i=0 \quad or \quad 1)$ $\vec{T}=\{t_1,t_2,...,t_n\}(t_i=0 \quad or \quad 1)$ で表されるとき，補完類似度 $S_c(\vec{F},\vec{T})$ は次のように定義される．

$$S_c(\vec{F},\vec{T})=\frac{a\times d-b\times c}{\sqrt{T\times(n-T)}}$$

ただし、

$$a=\sum_{i=1}^{n} f_i\times t_i, \qquad b=\sum_{i=1}^{n}(1-f_i)\times t_i,$$

$$c=\sum_{i=1}^{n} f_i\times(1-t_i), d=\sum_{i=1}^{n}(1-f_i)\times(1-t_i),$$

$$a+b+c+d=n, \qquad T=\sum_{i=1}^{n} t_i.$$

本研究ではこの定義において，ベクトルの次元数 *n* を対象とした文書の総数とし，画像 *Pic1, Pic2* を事物 *Thg1*, *Thg2* に対応させ，事物が文書 *i* に出現するなら 1，出現しなければ 0 を置き，各事物の出現パターンを二値ベクトル化する．定義に現れるパラメータ *a,b,c,d* はコーパスにおける事物間の情報として，*a,d* は一致情報を表し，*b,c* は不一致情報を表す．*a,b,c,d* はそれぞれ次のような数である．

- *a* : *Thg1*, *Thg2* がどちらとも出現する文書数.
- *b* : *Thg1* は出現しないが，*Thg2* は出現する文書数.
- *c* : *Thg1* は出現するが，*Thg2* は出現しない文書数.
- *d* : *Thg1*, *Thg2* がどちらとも出現しない文書数.

このように対応させて作成した二つのベクトルの包含状態を類似度として測ることによって，事物間の関係を推定する．**図 1** に

本研究で用いた二値ベクトルで表した出現パターンのイメージを示す.

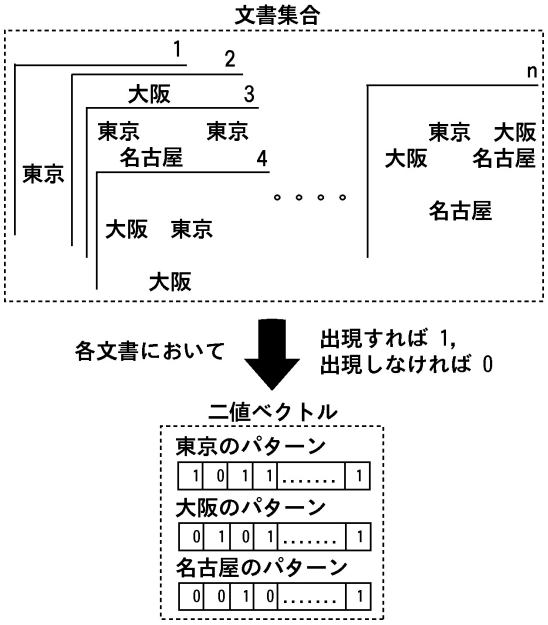

図 1　二値ベクトルで表した出現パターン

3　文書内頻度を考慮した補完類似度

前節に示した二値ベクトルを対象とした補完類似度には，コーパスにおける出現情報として，事物が出現する文書数，すなわち文書頻度(Document Frequency)を利用しているが，文書内頻度(Term Frequency)を利用していない．一般に，文書の主題となるようなキーワードはその文書に繰り返し現れる傾向にある．このことより，情報検索の分野では質問文と文書の類似度を測るために用いる基本的なスコアとして，文書頻度に関する情報量を表す IDF(Inverse Document Frequency)と，文書内頻度(TF)の内積 TF・IDF がある[2]．そこで，対象とする事物に文書内頻度に沿った重みを付け，文書内頻度情報を利用することによって，文書において主要な事物に関する関係を優先的に得られるかを検討する．重み付けにより，出現パターンは 0 か 1 ではなく，文書での出現状態によって重みの要素は多値となる．そこで，多値画像のための補完類似度[5]を利用することを考えた．次に，多値画像のための補完類似度の定義式を示す.

定義 2　多値画像のための補完類似度

$$\vec{F}_g = \{f_{g1}, f_{g2}, ..., f_{gn}\}(f_{gi} = 0.0 through 1.0)$$
$$\vec{T}_g = \{t_{g1}, t_{g2}, ..., t_{gn}\}(t_{gi} = 0.0 through 1.0)$$

のとき，補完類似度 $S_g(\vec{F}_g, \vec{T}_g)$ は次のように定義される.

$$S_g(\vec{F}_g, \vec{T}_g) = \frac{a_g \times d_g - b_g \times c_g}{\sqrt{n \times T_{g2} - T_g^2}} = \frac{n \times a_g - F_g \times T_g}{\sqrt{n \times T_{g2} - T_g^2}}$$

ただし、

$$a_g = \sum_{i=1}^{n} f_{gi} \times t_{gi}, \qquad b_g = \sum_{i=1}^{n} (1 - f_{gi}) \times t_{gi},$$
$$c_g = \sum_{i=1}^{n} f_{gi} \times (1 - t_{gi}), d_g = \sum_{i=1}^{n} (1 - f_{gi}) \times (1 - t_{gi}),$$
$$F_g = \sum_{i=1}^{n} f_{gi}, \qquad T_g = \sum_{i=1}^{n} t_{qi}, \qquad T_{g2} = \sum_{i=1}^{n} t_{gi}^2.$$

この定義式は，f_{gi}, t_{gi} がとる重み要素を 0,1 だけにすると，二値画像のための補完類似度 $S_c(\vec{F}, \vec{T})$ になる.

本研究では，二値画像のための補完類似度を単に補完類似度と呼び，多値画像のための補完類似度を**重み付き補完類似度**と呼ぶ.

図 2 に多値ベクトルで表した出現パターンのイメージを示す.

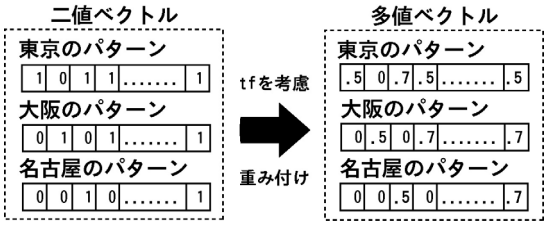

図 2 多値ベクトルで表した出現パターン

4　文書内頻度による重みの決定

$\vec{F}_g, \vec{T}_g$ の要素となる重みの決定は関係推定の対象となる事物の文書内頻度 *tf* (Term Frequency)に基づいて行う．まず，対象となる事物の文書内頻度を調査した. CD 毎日新聞 1991-2001 年版をコーパスとして用い，各年版に含まれる固有名詞や一般名詞を対象とする事物とした．その結果，一文書に 3 回以上出現する事物は少ないことがわか

った．このことから，毎日新聞記事データにおける重みは文書内頻度が 0, 1, 2, 3 以上の四段階とした．

四段階の重みは，各事物 *Thgj* が出現する文書数を *df* (*Thgj*) (Document Frequency)，1 回だけ出現する文書数を *df1(Thgj)*，2 回出現する文書数を *df2(Thgj)*としたとき，すべての事物 *wj (1≦j≦m)*について *df, df1, df2* を数え上げ，次の式で求めた値をすべての事物において *tf* に応じて共通に与える *weigh(tf)* と決定した．

定義 3 文書内頻度による重みの決定法

- *tf* = 0 のとき，*weight(tf)* = 0
- *tf* = 1 のとき，

$$weight(tf) = \frac{\sum_{j=1}^{m} df1(w_j)/df(w_j)}{m}$$

（1 回出現する文書の割合の平均値）

- *tf* = 2 のとき，

$$weight(tf) = \frac{\sum_{j=1}^{m} (df1(w_j) + df2(w_j))/df(w_j)}{m}$$

（1,2 回出現する文書の割合の平均値）

- *tf* ≧ 3 のとき，

$$weight(tf) = \frac{\sum_{j=1}^{m} df(w_j)/df(w_j)}{m} = 1$$

（1 回でも出現する文書の割合の平均値）

この定義式を用いて，関係推定の対象とする事物について各年版の毎日新聞について調査した結果，次の値を得た．

- *tf* = 0 のとき，　*weigh(tf)* = 0
- *tf* = 1 のとき，　*weigh(tf)* = 0.84
- *tf* = 2 のとき，　*weigh(tf)* = 0.95
- *tf* ≧ 3 のとき，*weigh(tf)* = 1

実験では，これらの重みを文書 i において事物が持つ重みとして f_{gi}, t_{gi} に与えることとした．たとえば，三つの文書 x, y, z があり，文書 x に「大阪」は 2 回，「東京」は 1 回現れ，文書 y に「大阪」は 0，「東京」は 2 回現れ，文書 z に「大阪」は 1 回，「東京」は 4 回現れたとする．このとき，文書 x, y, z において「大阪」と「東京」の各ベクトル要素 f_{gx}, f_{gy}, f_{gz} (または t_{gx}, t_{gy}, t_{gz}) には**表 1** に示す重みが与えられる．

表 1　ベクトルの要素となる重み付けの例

文書	X	y	z
大阪	0.95	0	0.84
東京	0.84	0.95	1

5　実験

5.1　概要

実験では，ヨミダス用語辞書に収録されている用語を事物とし，

表 2 に示す 1991 年から 2001 年までの毎日新聞 14 種類のテキストデータをそれぞれコーパスとして用いる．一記事を文書単位とし，記事数をベクトルの次元数とする．各文書における用語の重みは前節に示す *weigh(tf)* を用いる．評価はヨミダス用語辞書に収録されている用語間に何らかの関係がある場合を正解とし，類似度の高い順に 1000 対を見た場合の適合率によって，二つの補完類似度の性能を比較し，頻度情報の貢献を測る．

5.2　結果・考察

ヨミダス用語辞書に収録されている用語は 54186 語である．そのうち，各コーパスには約 18000 語の用語が現れる．したがって，辞書に収録されている用語がすべて現れないので，正解とした用語間の関係をすべて再現することはできない．

表 2 に各コーパスにおける二つの補完類似度の適合率を示す．これは各コーパスにおいて二つの用語間の類似度をそれぞれ補完類似度と重み付き補完類似度を用いて計算し，類似度が高いものから上位 1000 件について正解判定を行った場合の適合率である．不等号はどちらの補完類似度の適合率が高かったかを示す．

この結果において，14 個のコーパスのうち適合率が同じであったものが 1 個あり，残る 13 個のうち重み付き補完類似度の適合率が高かったものが 10 個ある．ここで，「二つの補完類似度の推定能力がすべてのコーパスにおいて等しい．」という仮説を立て，

表 2　適合率

コーパス	補完類似度		重み付き補完類似度	コーパス	補完類似度		重み付き補完類似度
1991	44.4	<	44.5	1998a	47.1	<	47.3
1992	48.3	<	48.4	1998b	45.7	<	45.8
1993	50.9	>	50.7	1999a	47.3	>	46.8
1994	47.8	<	48.1	1999b	48.8	<	49.1
1995	40.3	<	40.5	2000a	46.2	>	46.1
1996	47.3	=	47.3	2000b	46.7	<	46.9
1997	43.0	<	43.6	2001	44.6	<	45.2

表 3　新聞記事から関係を推定された用語対の例

類似度	事物 1	事物 2	正解・不正解
11291.890	同時多発テロ	アフガン	*****
11124.555	小泉純一郎	小泉首相	
9310.221	選挙	参院選	
8587.430	官房長官	福田康夫	*****
7042.404	同時多発テロ	ウサマ・ビンラディン	*****
6733.389	選挙	選挙区	
6615.725	ファクス	Ｅメール	
6343.695	財務相	塩川正十郎	*****
5948.384	選挙	比例代表	
5869.183	選挙	投開票	
5769.533	株式市場	平均株価	
5726.595	訴訟	損害賠償	
5563.195	同時多発テロ	報復	*****
5559.984	ウサマ・ビンラディン	ビンラディン	
5455.257	米大リーグ	ア・リーグ	
5318.772	選挙	立候補	
5237.116	経済財政担当相	竹中平蔵	*****
5215.294	同時多発テロ	空爆	*****
5176.856	厚生労働省	厚労省	
5176.104	選挙	当選	
5155.707	ＴＯＰＩＸ	東証株価指数	
5048.065	米大リーグ	ナ・リーグ	
4985.615	狂牛病	肉骨粉	
4787.469	扇千景	国土交通相	*****
4775.770	経済産業相	平沼赳夫	*****
4583.608	選挙	市長選	

符号検定を片側検定で行うと，仮説は 5%水準で棄却される[1]．このことから，本実験において，重み付き補完類似度は補完類似度より推定能力が高いと言える．したがって，事物の文書内頻度情報を利用することは事物間関係の推定に有効であると言える．残念ながら，この統計的有意は本実験では「顕著」ではない．これは新聞記事においては用語が繰り返し使われることが少ないためと推定される．今後，用語の繰返しが多い学術論文等で評価を行う必要がある．

表 3 に重み付き補完類似度を用いて毎日新聞記事データ 2001 年版から得た結果の一部を示す．これらの用語対は上位 50 件から選んだ興味深いものである．行に現れる数値は類似度である．また，*****は不正解と判定した印であるが，これらは実際には関係があり，本手法により役職と人名など既存の用語辞書には含まれない最新の関係が抽出される．

6　応用実験

前節では，補完類似度を用いて新聞記事に現れる用語について関係を推定した．そして実験結果から，文書内頻度の情報を考慮した重み付き補完類似度は推定能力が向上することがわかった．この実験では，日本語で書かれた単一のコーパスを対象としていた．そこで，本研究では，補完類似度の応用として，日英パラレルコーパスを対象とした場合，補完類似度は訳語間の関係を推定することを試みた．

表 4　補完類似度による訳語間の推定結果

英単語	日単語	英単語	日単語
Yen	円	Had	た
And	や	economic	経済
percent	%	Not	ない
Was	た	economy	経済
And	など	To	に
Also	も	And	と
should	べき	Or	や
such	など	To	ため
No	ない	As	など
economy	景気	And	・
other	など	is	ある
As	として	To	こと

実験に用いたパラレルコーパスは 1989 年から 2001 年までの読売新聞と The Daily Yomiuri とから文対応を得た 15 万文対の日英パラレルコーパスである[6]．実験では，単純に，日文を「茶筌」[3]によって形態素解析し得た形態素を各単語とし，各日文に含まれるすべての単語と，その文に対応した英文に含まれるすべての単語との重み付き補完類似度を求めた．**表 4** に高い類似度を得た訳語対の一部を示す．結果を見ると，正しいと思われる訳語対が上位に多く現れた．適用法や性能評価は今後の課題とする．

7　おわりに

本研究では，出現頻度情報を考慮した補完類似度を事物間の関係を推定することに適用した．補完類似度はこれまでに事物間の一対多関係を推定することに有効であると報告したが，出現頻度情報を考慮していなかった．新聞記事データにおける実験結果を比較した結果，出現頻度を考慮することによって事物間関係の推定能力が向上したことを示した．

参考文献

[1] 池田央，統計ガイドブック，新曜社，1989.

[2] Christopher D. Manning and Hinrich Schutze, Foundations of Statistical Natural Language Processing, The MIT Press, Cambridge MA, 1999.

[3] 松本裕治 北内啓 山下達雄 平野善隆 今一修 今村友明，日本語形態素解析システム「茶筌」.

[4] 澤木美奈子 萩田紀博，補完類似度による劣化印刷文字認識，電子情報通信学会信学技報 PRU95-106, pp.19-24, 1995.

[5] Minako Sawaki, Norihiro Hagita, and Kenichiro Ishii, Robust Character Recognition of Gray-Scaled Images with Graphical Designs and Noise, Proc. of ICDAR, Ulm, Germany, August 18-20, pp.491-494, 1997.

[6] 内山将夫 井佐原均，日英新聞記事の対応付けと精度評価，情報処理学会 NL-151-3(FI-68-3), pp.15-22, 2002.

[7] 山本英子 梅村恭司，コーパス中の一対多関係を推定する問題における類似尺度，自然言語処理 Vol.9 No.2 pp.45-75, 2002.