E-41

出現頻度を考慮した補完類似度による事象間関係の推定

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

山本 英子† 井佐原 均†
Eiko Yamamoto Hitoshi Isahara

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象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、事象間の関係を暗黙的に一対一関係と想定している。これは、関係を持つ事象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、事象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象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象間の一対多関係の推定に適用し、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[1]。この報告の際に用いた事象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、事象の文書内頻度は重要な情報源である。そこで本研究では、事象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象間関係を推定する問題により有効があったことを報告する。

2 補完類似度

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。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[2]。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、事象間の一対多関係を推定する問題に適用した。ここで、二値画像のための補完類似度の定義式を示す。

† 独立行政法人 通信総合研究所,
Communications Research Laboratory

二つの事象が $\boldsymbol{F} = \{f_1,f_2,\ldots,f_n\}$ $(f_i=0 \text{ or } 1)$, $\boldsymbol{T} = \{t_1,t_2,\ldots t_n\}$ $(t_i=0 \text{ or } 1)$ のとき、補完類似度 $S_c(\boldsymbol{F},\boldsymbol{T})$は次のように定義される。

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

ただし、

$$a=\sum_{i=1}^{n} f_i\times t_i, \quad b=\sum_{i=1}^{n}(1-f_i)\times t_i,$$
$$c=\sum_{i=1}^{n} f_i\times(1-t_i), \quad d=\sum_{i=1}^{n}(1-f_i)\times(1-t_i),$$
$$a+b+c+d=n, \quad T=\sum_{i=1}^{n} t_i.$$

本研究では、ベクトルの次元数 n を対象とした文書の総数とし、事象 F または T が文書 i に出現するなら 1、出現しなければ 0 を置き、各事象の出現パターンをベクトル化する。そして二つのベクトルの包含状態を類似度として計ることによって、事象間の一対多関係を推定できる。しかし、この補完類似度による一対多関係の推定には、文書内頻度情報が利用されていない。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。そこで、対象とする事象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象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[3]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。$\boldsymbol{F_g} = \{f_{g1},f_{g2},\ldots,f_{gn}\}$ $(f_{gi}=0.0 \text{ through } 1.0)$, $\boldsymbol{T_g} = \{t_{g1},t_{g2},\ldots t_{gn}\}$ $(t_{gi}=0 \text{ through } 1)$ のとき、補完類似度 $S_g(\boldsymbol{F_g},\boldsymbol{T_g})$は次のように定義される。

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

ただし、

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

この定義式は、$\boldsymbol{F_g},\boldsymbol{T_g}$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 $S_c(\boldsymbol{F},\boldsymbol{T})$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。

3 文書内頻度による重みの決定

F_g, 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象の文書内頻度に基づいて行う。まず、対象となる事象の文書内頻度を調査した。毎日新聞記事データ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象とした。その結果、一文書に 3 回以上出現する事象は少ないことがわかった。重みの決定においては、すべての事象について、各事象が出現する文書数を df、1 回だけ出現する文書数を df_1、二回出現する文書数を df_2 とすると、出現しない場合は 0、1 回出現する場合は df_1/df、2 回出現する場合は $df_1/df+df_2/df$、3 回以上出現する場合は 1 として重みを計算し、頻度ごとに重みの合計をとり正規化することによって、コーパスにおける事象すべてに用いる 4 段階の重みとした。これらの重みを文書 i において事象が持つ重みとして f_{gi}, t_{gi} に与える。

4 実験

4.1 概要

実験では、読売新聞シソーラス・ヨミダス用語辞書に収録されている用語を事象とし、1991 年から 2001 年までの毎日新聞記事が収録されている 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。用語の重みはコーパスを調査した結果の平均値 0, 0.84, 0.95, 1 を用いる。評価はヨミダスに収録されている用語間の関係を正解とし、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。

4.2 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 18000 語程度の用語が現れる。したがって、各コーパスには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 1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、その値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、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。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、事象の文書内頻度情報を利用することは事象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同じ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ことが必要である。

図 1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、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。

表 1 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
91	44.4	<	44.5	98a	47.1	<	47.3
92	48.3	<	48.4	98b	45.7	<	45.8
93	50.9	>	50.7	99a	47.3	>	46.8
94	47.8	<	48.1	99b	48.8	<	49.1
95	40.3	<	40.5	2000a	46.2	>	46.1
96	47.3	=	47.3	2000b	46.7	<	46.9
97	43.0	<	43.6	2001	44.6	<	45.2

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	

図 1 新聞記事から関係を推定された用語対の例

5 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象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象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象間関係の推定能力が向上したことを示した。

参考文献

[1] 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.
[2] 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.
[3] 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.