A-011

共起ネットワークを用いた電子掲示板からの情報抽出

Information Extraction in Bulletin Board System using Co-occurrence Network

田中 航介† 鈴木 育男‡ 山本 雅人† 古川 正志†
Kousuke TANAKA Ikuo SUZUKI Masahito YAMAMOTO Masashi Furukawa

1. はじめに

スレッドフロート型掲示板は電子掲示板の形態の一つであり，スレッドと呼ばれる話題及び投稿の集合を，最終投稿時間順に表示する方式である．現在の電子掲示板はこの形態が主流であり，Web 上でユーザが情報を発信，収集，及びコミュニケーションといった様々な活動を行う場として幅広く利用されている．その情報量は利用者数，または利用頻度の増加や荒らし行為，スパム等の要因により有意性に関わらず増大している．したがって掲示板の内容とは無関係な投稿も多く存在し，本来語られるべき話題と乖離しているスレッドも少なくなく，結果としてユーザが目的のスレッドを探索する負担の増大に繋がる．

本研究では，あるカテゴリ内に存在するスレッドを投稿内容から分類し，各スレッドの話題の類似性の視覚化，または興味のあるスレッドに類似するスレッドの抽出[1]による前述の負担の軽減を目的とする．そのための手段として投稿内容を元に単語の共起関係を抽出，共起ネットワークを生成し，これに類似度を適用してスレッドを分類する手法を提案，及びその評価を行う．

2. スレッドからの投稿内容の抽出

スレッドを共起ネットワーク化するにあたり，単語の共起関係及びその頻度を導出する必要があり，共起関係を導出するにはスレッドに対し形態素解析を行う必要がある．本研究ではスレッドに投稿された本文のみを対象とし，MeCab[2]により形態素解析を行う．

2.1 共起関係と共起頻度

共起とはある集合内で事象 X，Y が同時に発生するような状態であり，この時 X と Y は共起関係にある．この関係は主に集合間の類似度を算出する場合，特に自然言語処理の分野等で用いられる．本研究の場合ではスレッド内の文章に単語 X，Y が同時に出現する場合に X，Y を共起関係とし，これを用いて 3 章で述べる手法によりスレッド間の類似度を算出する．また，全ての共起関係を共起ネットワークに落とし込もうとすると，解析するスレッドの文章量に比例して爆発的にリンクが増加する．そのため計算時間が膨張するので，共起頻度と呼ばれる尺度により共起関係の絞り込みを行う．

共起頻度は得られた共起関係の出現率を表し，X，Y の出現数$|X|$，$|Y|$を用いた幾つかの指標により求められる[3]．式(1)は Simpson 係数により共起頻度を算出する式となる．

$$S(X,Y)=\frac{|X\cap Y|}{\min(|X|,|Y|)} \tag{1}$$

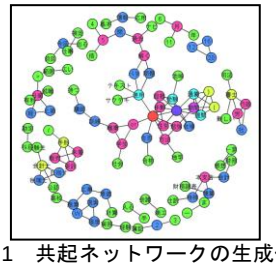

図1　共起ネットワークの生成例

Simpson 係数は他の指標のように$|X|>|Y|$かつ両者に大きな差がある場合でも共起関係が低く出ない長所を持ち，代わりに出現数の非常に少ない単語との共起頻度が非常に高く出る欠点を持つ．そこで式(2)のように Simpson 係数の分母に閾値 k を設け，これを解消する．

$$S(X,Y)=\begin{cases}\dfrac{|X\cap Y|}{\min(|X|,|Y|)} & \left(\min(|X|,|Y|)\geq k\right)\\ 0 & \left(\min(|X|,|Y|)<k\right)\end{cases} \tag{2}$$

本研究では式(2)により共起頻度を算出する．

2.2 共起ネットワーク

図 1 のように共起関係における単語をノード，単語間の共起頻度を重み付きリンクとし生成されるネットワークを共起ネットワークと呼ぶ．共起に前後関係を持たせた場合はリンクが向きを持ち，有向グラフとなる．共起ネットワークの構造は，同一の文章でも共起関係の取り方により異なる．

3. 共起ネットワーク間の類似度の導出

スレッドから生成した共起ネットワーク A，B の類似度を導出するために，以下の手順を行う．

1. A のノード i と単語の一致するノード j を B より探索する．存在するならばステップ 2 へ進む．
2. 同様にノード i の近傍ノードの個数を C，ノード j の近傍ノードの個数を N，両近傍におけるノードの一致数を R とする．
3. 式(3)～(5)よりそれぞれ適合率，再現率，F 値を算出し，F 値をノード i 近傍の類似度とする．

$$precision = \frac{R}{N} \tag{3}$$

$$recall = \frac{R}{C} \tag{4}$$

$$F=\frac{2\times precision\times recall}{precision+recall} \tag{5}$$

4. 全てのノードに対しステップ 1～3 を繰り返し，F 値の平均を共起ネットワーク A，B の類似度とする．

† 北海道大学　大学院情報科学研究科
‡ 北見工業大学　情報システム工学科

F 値は予測結果の集合 N がどれだけ正解の集合 C に近いかを評価する指標であり，予測結果の正解率である適合率と予測された正解の割合である再現率の調和平均により求められる．適合率と再現率は二律背反の関係にあるため，F 値が高い程予測の性能が高く，これは本手法において共起ネットワーク A，B の類似度が高いとみなせる．

4. スレッドの分類実験

提案手法の性能評価を行うため，実際のスレッドを幾つか用いて提案手法と共に，共起関係のみによる類似度を算出し，スレッドを分類し結果を考察する．共起関係のみによる類似度は C 及び N をスレッドの共起関係の個数，R をその一致率とし，F 値を導出したものである．

4.1 実験条件

実験に使用するスレッドとして，2 ちゃんねる[4]の資格全般板より表 1 に示す 10 つのスレッドを取得する．スレッドはそれぞれ 1〜4，5〜8，9，10 と関連性を想起させるものを選択，いずれも投稿数は最大の 1000 である．共起関係は Simpson 係数の閾値 k を 10 とし，出現数が 10 未満の単語を除外する．更に，共起頻度にも閾値として 0.2 を設け，ノード及びリンクの本数をある程度絞り込む．

表2　実験に用いるスレッドの一覧

	スレッド名(一部省略)	ノード数	リンク数
1	日商簿記 2 級　part371	201	477
2	日商簿記 2 級　Part372	176	384
3	日商簿記 1 級 Part80	219	452
4	日商簿記 3 級 Part198	130	173
5	基本情報技術者試験 Part351	160	374
6	基本情報技術者試験 Part352	169	367
7	応用情報技術者 Part102	175	504
8	IT パスポート試験 Part54	100	139
9	環境計量士・一般計量士　Part30	131	204
10	漢検 1 級・準 1 級専用スレッド 18	127	290

4.2 実験結果・考察

各々のスレッドに対し，共起関係のみによる類似度と提案手法による類似度をそれぞれ算出した結果，表 2 のようになった．上端及び左端はスレッドの番号に対応，セルの上部の値が提案手法，下部の値が共起関係のみによる F 値の算出結果となっている．これに多次元尺度構成法を適用しその位置関係をプロット，類似度に基づいた距離による分類を行ったところ，それぞれ図 2 のようになった．

分類されたスレッドは座標値が異なるが，どちらの手法もほぼ同じ位置関係を示している．また，スレッド 4 及び 8 が関連性の高いと予想されたスレッド 1〜3 及び 5〜7 と距離が大きく離れており，スレッド 9 や 10 と近い位置を取った．このような結果が生じる理由としては

- 共起関係たるリンク数(C 及び N)の差が大きいと，適合率と再現率の差も大きくなり，結果として F 値が小さく出やすい．
- 両共起ネットワーク内で一致した単語の近傍における F 値の和が，共起関係のみにおける F 値に対し全体的におおよそ 2〜3 倍である場合が多く，分類結果を決定的に分ける差が現れなかった．
- 得られた共起関係の単語にスレッドの内容とはあまり関わらない一般的な単語が多く，それらにより F 値が高く(低く)出ている．

が考えられる．

表 2　スレッド間の類似度(上：提案手法，下：共起頻度のみ，適合率及び再現率は省略)

	2	3	4	5	6	7	8	9	10
1	0.341 0.260	0.164 0.103	0.156 0.080	0.254 0.139	0.184 0.130	0.140 0.084	0.147 0.042	0.164 0.044	0.125 0.050
2		0.190 0.120	0.160 0.083	0.271 0.148	0.225 0.162	0.185 0.104	0.105 0.038	0.162 0.054	0.133 0.059
3			0.149 0.074	0.144 0.063	0.112 0.073	0.082 0.048	0.101 0.034	0.123 0.058	0.120 0.057
4				0.141 0.051	0.126 0.059	0.170 0.062	0.139 .051	0.178 0.064	0.193 0.078
5					0.363 0.372	0.266 0.196	0.168 0.058	0.191 0.076	0.151 0.051
6						0.274 0.250	0.092 0.051	0.141 0.070	0.149 0.064
7							0.131 0.056	0.141 0.051	0.094 0.045
8								0.179 0.052	0.140 0.056
9									0.226 0.073

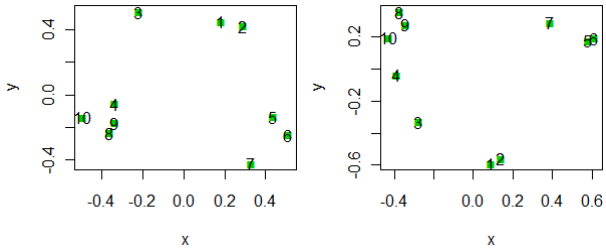

図 2　類似度に基づいた距離によるスレッドの分類結果(左：提案手法，右：共起頻度のみ)

5. おわりに

本研究ではスレッドフロート型掲示板のスレッドを，投稿内容から共起ネットワーク化し分類する手法を提案した．しかしスレッドの分類実験は結果に大きな差が現れず，また実験自体も小規模で差を比較するには条件設定が不十分であると考えられる．

今後は投稿内容のみならず，スレッドの持つ特徴も交えた手法の検証やより大規模なスレッドの分類実験，共起関係の抽出方法の見直しを行い，スレッドの分類精度の向上をさせる事が課題として挙げられる．

参考文献

[1]深谷 雅志, 倉本 到, 渋谷 雄, 辻野 嘉宏, “電子掲示板における行動履歴を用いたユーザ にとって興味あるスレッドの推薦手法”, 電子情報通信学会技術研究報告, Vol.106, No.410, pp.149-154(2006).
[2]MeCab (和布蕪), http://mecab.sourceforge.net/
[3]相澤 彰子, “共起に基づく類似性尺度”, オペレーションズ・リサーチ: 経営の科学, Vol.52, No.11, pp.706–712(2007).
[4]2 ちゃんねる，http://www.2ch.net/