Small World構造を利用したキーワード抽出による記事分類

對崎　宏也　　三浦　孝夫

法政大学工学部情報電気電子工学科

1．前書き

キーワード抽出は、情報検索、テキストマイニングにおいて重要な手法である。キーワードを手がかりにすれば、膨大な量の文章の中から目的の文章を選択しやすくなる。ニュースコーパスのように、記事内容の意図が比較的明確に生じやすいものでは、内容の相互関連が個別の事件よりもその種別（トピックス）に生じやすい。

Small World構造は、あるノードからほかの任意のノードにたどり着くのに、少数の中継ノードを経由するだけでよいという性質をもつ。本研究ではこの性質に着目し、Small world構造に基づくキーワード抽出法を用いてニュースコーパスからキーワードを抽出し、そのキーワードを基に、記事のトピックスを分類する手法を提案する。

2．Small World構造に基づくキーワード抽出

Small World構造は、ノードがクラスタ状に集まったネットワークであり、ある一定の割合で遠方のノードに繋がったリンクが存在する。このリンクの存在により、ネットワーク全体のリンクの平均パス長が短くなるという特徴をもつ。この特徴を示す指標としとC(clustering coefficient)とL(characteristic path length)の2つの特徴量があげられる。

松尾ら[1]は、論文がSmall World構造であることを示し、キーワード抽出法により文章から抽出された単語は、出現頻度が高い単語と出現頻度は低いものの文章中で重要な役割を果たす単語であることが示されている。

文章構造がSmall World構造であるとすれば、抽出したいくつかのノードは、リンクの平均パス長を短くするのに大きく貢献しているはずである。このような語は、共起関係の薄いノードのクラスタ同士を繋いでいるので、文章において重要な位置を担ったノードであると考えられる。

貢献度(contribution) CBは、次の式を用いて求めることができる。

L'_v：あるノードvはグラフに接続されているが平均には含めないで平均を取る。

L'_{G_v}：あるノードvとvを含むリンクはグラフから除外して平均を取る。

L'はLの定義を非連結グラフに拡張したもの。

$$CB_v = L'_{G_v} - L'_v$$

L'_{G_v}とL'_vの差を取ることで、ノードvがLの減少にどれぐらい影響を与えているかが求まる。
この値が大きいノード程、離れたクラスタをつなぐ重要なノードであると考えられる。

2.3　ニュースコーパスにおけるSmall World構造に基づくキーワード抽出法の検証

松尾ら[1]は、論文を対象にSmall world構造の検証を行っている。本研究では、共起関係の薄いノードのクラスタ同士を繋いでいる重要度の高いノードの存在に着目する。ニュースコーパスのように短い記事で、いかに読者に内容を伝えるかが推敲されている文章は、コストの最小化と連結性の最大化の両方の点で、Small World構造に近いものになると考えられる。

3．実験

英語文章のロイター記事を対象に、ニュース記事にもSmall World構造に基づくキーワード抽出法が利用できること、および抽出したキーワードを用いて記事分類ができることを実験で検証する。

本研究では、規定回数(2回)以上出現する単語が3つに満たない記事は除く。また、共起関係の薄いノードを結ぶノードの存在にのみ着目し、ニュース記事がSmall World構造を構成しているかどうかは判定しばい。このため、Lの定義を非連結グラフに拡張せず単純な扱いをする。

記事のトピックス分類をするために用いたロイタコーパス(Reuter Corpus)を使用する。表1がその一部である。

cocoa	income	orange
coconut	Jobs	Rand
coffee	money-fx	trade
Grain	Oilseed	Yen

表1．ロイターコーパストピックス例

ロイターコーパスから記事3718件を抽出し、出現数2回の単語をノードであらわす。Small World構造に基づくキーワード抽出法を用いて、貢献度が高い上位3単語をキーワードとして抽出し、そのうち3つの単語が、トピックスの単語と一致するかどうかを検証するという手順をとる。

4．実験結果

ロイター記事のトピックスの単語が、Small World構造に基づくキーワード抽出法を用いて抽出した3つの単語の中に含まれているかどうかを実

験した結果、ロイター記事 3718 件に対し、トピックスの単語と同じ単語を抽出することができた。対象となった記事は、338 件だった。

本研究では、同時に、その内容を手作業で吟味した。この結果、表２のようにトピックスと似た意味を持つ単語を抽出することができた。

Topics	キーワード	CB
earn	Said	1.01341
	Henderson	1.00804
	Company	1.00268
earn	Split	0.948276
	compani	0.931034
	said	0.913793
money-fx	bank	1.003559
	propos	1
	contract	0.992883
trade livestock	beef	1.305556
	japan	0.861111
	said	0.833333
crude ship	worker	0.96
	vote	0.96
	strike	0.94
trade veg-oil	oil	0.857143
	propos	0.857143
	foreign	0.809524
crude	said	1.074007
	british	1.030686
	oil	1.018051
trade iron-steel	export	2.025641
	seamless	1.282051
	pipe	1.25641
gnp	said	1.034884
	source	1
	finance	0.94186
	economic	0.930233

表２　抽出した単語

5. 考察

表２の veg-oil のように～oil などの単語は抽出した単語と一致しない。これは、短い記事の中では、～oil の～に相当する部分は、何回も繰り返さないだろうと考えていたため、記事からフレーズとして抽出しなかったことに起因する。

本実験では、トピックスの単語が表す意味より、単語同士を比較し、一致するかの一点を重視した。貢献度が高い単語とトピックスを、より高い確立で一致させるために、トピックスの単語と似た意味を持つ単語を語彙辞書として各トピックスに用意すれば、表２に示した例もトピックス別に分類が可能となるだろう。また、貢献度が高い上位３単語をキーワードとして抽出したが、記事の長さに応じて、出現回数、貢献度の閾値を変更していけば、精度の向上が期待でき、より良い結果を得ることができるだろう。

6. 結論

本研究では、ニュースコーパスを実験対象とし、Small World 構造に基づくキーワード抽出法でキーワードの抽出が可能か、また、抽出したキーワードを用いて記事分類が可能かを実験した。トピックスと同じ単語の抽出はおよそ１割程だが、表２に示したように、キーワードに対応する辞書を用意し検証を行えば分別できるといえるだろう。

今後の課題として、記事の長さの違いに対応させ一定の精度を保ったまま、単語一致から意味一致へとシフトさせ、検証していく必要がある。

[参考文献]

[1] 松尾　豊、大澤　幸生、石塚　満：“Small World 構造に基づく文章からのキーワード抽出” 情報処理学会誌、Vol.43, No.6, pp.1825-1833(2002)