Webディレクトリを利用した意味的関連語集合の作成

佐々木 稔, 三上 健太, 新納 浩幸
茨城大学工学部情報工学科

1　はじめに

新聞記事などの文書を内容ごとに分類する文書クラスタリングや，あるキーワードについて検索された例文集合に対して単語の意味ごとに分類する単語クラスタリングにおいて，分類する手がかりとして文書内に出現する単語や単語とある範囲内で共起する単語などが利用される．このような手がかりとなる単語に対して，頻度や文書の長さなどを利用した重み付け手法により数値化が行われ，出現単語の分布や重要性などの特徴を捉えることができる．

これらの特徴をベクトル空間モデルにおいて 2 つの文書または単語を比較する際，手がかりとなる単語は同じ文字列である必要がある．非常に近い意味を持っているとしても異なる文字列の単語であれば，それらは互いに独立した手がかりとして扱われる．また，グラフ構造を利用して単語間の類似性を数値化する場合においても，単語を表す節点間のつながりが文書データ内に明示的に表現されていないために，例えば「国道」と「県道」といったどちらも「道」を表す単語であっても，節点間のつながりを捉えることができない場合が存在する．また，共起単語の頻度差が大きい場合，クラスタリングを行うと高頻度語の影響が強く出てしまい，低頻度語の特徴はあまり活用されないこともある．このように単語を手がかりとしてクラスタリングを行うにはある程度の限界があり，上記の問題は短い文を扱う語義別用例分類や自動要約などにおいて顕著に表れる傾向にある．

これらの課題を解決する目的として，本稿では共起単語を使わずに単語や複合語の間にある概念的な関連性を捉えるために，World Wide Web 上のディレクトリ検索を利用した意味的関連語集合の作成方法を提案する．単語または複合語をディレクトリ検索し，その結果として得られるディレクトリ名を手がかりとしてベクトルを作成し，類似度計算を行う．ディレクトリ名を意味概念とみなし，語句の意味的な分布統計を考慮することで，語句がどのような背景で出現するのかを捉えることができる．このような分布を手がかりとしてクラスタリングを行った結果，類似した概念で利用されやすい単語のクラスタを求めることができる．また，従来の関連語自動抽出手法では単語に対する関連語を求めていたが，提案手法では複合語に対しても関連語を抽出することが可能である．

ディレクトリ検索を利用した語句の分布を求める研究はこれまでにいくつか提案されている．巨大な百科事典サイトとして知られる Wikipedia のカテゴリを利用して，見出し語間の類似度を計算する手法がある [3, 4]．見出し語に割り当てられたカテゴリ名の違いにより類似性を評価する方法であるが，見出し語のみ評価可能であるため，類似度の計算ができる単語の範囲は限られてしまう．また，その他の手法として，Google ディレクトリ (Dmoz) を利用して語句の類似性を評価するものも存在する [2]．Google ディレクトリを用いることで幅広い語句についてカテゴリ検索を行う事が可能で，固有表現については有効に類似度計算を行うことができる．しかし，一般名詞について利用可能かどうかはよく分かっていない．

本稿では，一般名詞を含めた幅広い語句に対してディレクトリ検索によりカテゴリ分布を求め，それを手がかりとしてクラスタリング手法により出現する概念が類似している単語集合の構築を行い，得られた分類が適切であるかどうかの評価を行う．

2　関連語集合の作成方法

本研究では，関連語集合を作成するにあたって，語句の持つ意味に注目している．語句がもつ意味として，その語句がどのような場面で用いられるか，どのような分野で使用されているのかといったようなジャンルを考慮することで，文脈を捉える手がか

りになると考える．語句のジャンルを抽出する方法として，World Wide Web 上で提供されているディレクトリ検索システムを利用する．まず，ある語句についてディレクトリ検索を行う．その検索結果として，その単語が出現する文書が存在するカテゴリ名を得ることができる．このカテゴリ名に含まれるラベルを抽出し，頻度を計算することで，カテゴリラベルを要素とするベクトルを作成する．このベクトルは，語句がどのようなジャンルに用いられているのかを表すベクトルである．集合内のすべての語句についてベクトルを作成し，それらを用いてクラスタリングを行う．そうすることで，出現するジャンルの類似した関連語集合を作成することができる．

3　実験

本研究で作成したプログラムを用いて，関連語集合を作成する実験を行った．

3.1　設定

本研究では，単語検索エンジンとして，Google ディレクトリ[1]を用いた．ディレクトリ検索で得られた結果を基に単語ベクトルを作成する．なお，単語ベクトルにおける各素性の重要度は，検索で得られたカテゴリ名の各ラベルの出現頻度とした．ただし，ディレクトリ名の上位にある “/Top/World/Japanese/” の部分は考慮しないこととする．また，地域情報である “/地域/アジア/日本/”，“/ニュース/新聞/地域/” の各カテゴリはサブカテゴリを含めて考慮しないこととする．次に，作成した単語ベクトルを基にクラスタリングを行う．本研究ではクラスタリングツールとして CLUTO を用いた [1]．CLUTO では，クラスタリング手法や類似度計算手法などがいくつか用意されており，オプション指定により選択して実行することが可能である．今回の実験では，クラスタリング手法として repeated bisections，類似度計算手法はコサインを用いて分類を行った．

3.2　データ

本研究では，Web 上のニュース記事，および分類語彙表 [5] よりデータを抽出し，それを実験データとした．具体的には，以下の 3 つのデータを用い，それぞれ実験を行った．

(1) ニュース記事から手作業で抽出した単語集合
(2) ニュース記事を MeCab で形態素解析し抽出した，単語および複合語からなる単語集合
(3) 分類語彙表で同じカテゴリに属する単語集合

3.3　評価方法

本実験の評価方法は，クラスタリングの正解率によるものとした．クラスタリングを行った際，同じカテゴリに属する単語集合のうち，どれだけの単語が正解とされるクラスタに割り振られているかを確率で表現する．正解率が高いほど，クラスタリングの精度が高いと言える．

4　実験結果と考察

前章で述べた 3 つのデータについて，それぞれ実験を行った．以下に結果と考察を述べる．

4.1　データ (1) の実験結果と考察

本実験では，実験に利用するニュース記事として，経済関連のニュースと IT 関連のニュースの 2 ジャンルのニュースを用いて実験を行った．このニュースデータは Web 上から無作為に抽出したものである．このニュースデータを手作業で単語集合に分割し，データを作成した．手作業であるため，単語だけではなく複合語も単語集合の中に含まれている．ニュースデータより抽出した単語は 95 件であり，カテゴリ数は 1613 個であった．なお，本実験ではクラスタリングをする際のクラスタ数は 5 とした．この実験の結果，正解率は 41.1%であった．同じ関連のニュース中の単語がほぼ全て 1 つのクラスタにまとまっているというわけではないため，経済関連，IT 関連と 2 つのくくりで捉えた場合の正解率はさほど高くないことがわかる．

次に，結果の内約を見ると，5 つのクラスタに分類された単語集合は，それぞれ意味を持っていることがわかった．5 つのクラスタはそれぞれ，会社関連，市場・株関連，IT 関連，経済関連，多義語を表す単語集合となった．正解率は 50%を下回っているが，抽出した単語が意味をもつ単語集合になってい

[1] http://www.google.co.jp/dirhp?hl=ja

表 1: データ (2) のデータ数および正解率

	単語数	カテゴリ数	正解率
ニュース 1	659	3594	62.2%
ニュース 2	878	4354	49.8%

るため，本実験の結果は有用なものと言えるのではないかと考える．

4.2 データ (2) の実験結果と考察

本実験では，ニュース記事として，経済，社会，IT，スポーツ，エンタテインメントと 5 つのジャンルのニュースを用いて実験を行った．このニュースデータは前項と同じように，Web 上から無作為に抽出したものである．ニュースデータは，各ジャンルから 2 ニュースずつ，計 10 個のニュースからなる．今回は，上述の構成のニュースデータを 2 つ用意し，MeCab による形態素解析を行い，名詞を抽出した後，その名詞と，連続名詞からなる複合語で構成される単語集合を作成，クラスタリングを行った．前項の実験と同じく，本実験でもクラスタリングの際のクラスタ数は 5 とした．なお，MeCab による形態素解析，複合語作成，クラスタリングデータ作成のプロセスは全てプログラムにより自動で行った．実験に使用したニュースデータのデータ数を，表 1 に示す．

2 つのニュースデータの正解率の平均値は 56.0%であった．この数値も高い正解率とは言い難い．しかし，結果の内約を見てみると，5 つのクラスタの中には，同じジャンルの単語が集まっているクラスタを確認することができた．今回の結果では，スポーツ，IT，経済の 3 つのジャンルの単語は，比較的同じクラスタに集まっていた．このように同じジャンルの単語が集まっているにも関わらず，正解率があまり高くない理由としては，単語集合中の単語および複合語の中に，あまり意味をもたない単語が含まれているからだと考える．数値が連続してできた複合語や，連続名詞であるが，複合語にしても意味がないもの，例えば「90 円 50 銭」という文章で言えば「円 50」という複合語，こういったものを単語集合から除けば，もっと正解率は高くなると考える．

表 2: データ (3) のデータ数および正解率

	単語数	カテゴリ数	正解率
1	86	992	86.0%
2	201	1946	58.2%
3	71	978	67.6%
4	52	898	73.1%
5	77	928	57.1%
6	73	1091	68.5%
7	64	644	84.4%
8	47	646	80.9%
9	29	557	82.8%
10	54	875	77.8%

4.3 データ (3) の実験結果と考察

本実験では，分類語彙表より抽出した単語集合を用いて実験を行った．この実験においては抽出した単語集合の抽出条件として，カテゴリ内に複合語が含まれていることとした．この条件の基で単語集合を抽出することにより，単語のみに限らず，複合語についても，関連語集合を作成することが可能であることを示すことができると考えた．分類語彙表から 5 つのカテゴリを選択し，そこに属する語句をすべて抽出し，1 組の正解データとした．この正解データを 10 組用意し，それぞれに対して実験を行った．なお，前項の実験と同じく，本実験でもクラスタリングの際のクラスタ数は 5 とした．実験の結果を表 2 に示す．

表 2 に示した 10 組のデータの正解率の平均値は 73.6%であった．本研究での実験において，一番良い正解率を得ることができた．また，10 個の実験データ中 6 個については，1 つのカテゴリに 1 つのクラスタが対応するという理想的な結果が得られた．残りの 4 個の実験データについては，1 つのクラスタに 2 つのカテゴリが対応するといった状況がみられた．これは，データ中に似たようなカテゴリが存在していたため，同じクラスタに 2 つのカテゴリが対応してしまっているのだと考える．例えば，同じ経済という大きなくくりの中の，保険と貯金という 2 つのカテゴリを同じ実験データ中に含めたため，この 2 つのカテゴリが同じクラスタに割り振られてしまったという状態である．しかし，保険と貯金というカテゴリではなく,「経済」というカテゴリとして考えると，1 つのクラスタに 1 つのカテゴリが対応していると言うことができる．よって，本実験結果は，クラスタリングによって，複合語を含めた関連語集合の作成に成功していると言えると考える．

ここで，クラスタリングを行うことである程度適

表 3: 実験に利用した単語集合

カテゴリ名	単語集合
箱など	郵便箱, 郵便受け, 郵便私書箱, 私書箱, ポスト, 状差し, 投票箱
武器	爆弾, 時限爆弾, 原子爆弾, 原爆, 水爆, 中性子爆弾, 核兵器
道具・置物・像など	花火, 線香花火, ねずみ花火, 仕掛け花火, 打ち上げ花火, 爆竹
標章・標識・旗など	勲章, 文化勲章, 喪章, 略章
機械・装置	機関, 蒸気機関, 内燃機関, 外燃機関, ディーゼル機関

表 4: 実験結果におけるカテゴリとクラスタの対応

クラスタ番号	単語数
0	武器
1	標章・標識・旗など
2	道具・置物・像など
3	機械・装置
4	箱など

切な関連語集合が作成できたデータの一例を示す. ここでは,「生産物および用具」という大きなジャンル分けの中の 5 つの小カテゴリを実験対象とした. 使用したデータのカテゴリおよびそれに属する単語集合を表 3 に示す. このデータに対して，CLUTO を用いて実験を行ったところ，各カテゴリに 1 つずつクラスタが割り当てられた. クラスタとカテゴリの対応を表 4 に示す. 表 3 で示した単語集合中に，5 つの単語は正解とされるカテゴリに属さず，別のカテゴリに割り振られた. そのため，正解率は 82.8%となった.

4.4 問題点

本研究で行った 3 つの実験より，問題点も発見することができた. まず第一に，データ (2) において，MeCab での形態素解析により抽出した単語および複合語に含まれている，あまり意味をもたない単語についてである. 本研究で作成したプログラムでは，名詞および連続名詞を抽出し，ディレクトリ検索を行っている. よって，データ (2) の実験結果で示したような，数値だけの単語や，数値と単語の連続名詞のうち意味をもたないものも検索対象となり，クラスタリングも行われてしまう. それらの単語がクラスタリング結果より求める正解率に影響を与えていると考えられる. この問題を解決するためには，MeCab により抽出した単語，複合語が意味をもつかどうかを判定し，意味のあるものだけを選択する必要がある. まず，簡単な解決策として，単語が数値の場合は検索対象としないということが挙げられる. これだけでも，正解率に与える影響が少なくなるのではないかと考える.

次に，データ (3) において，似たようなカテゴリの単語集合を同じ実験データに含むとカテゴリが一緒になるという問題が挙げられる. この問題は，前述のように，対象となるカテゴリを 1 つ上のカテゴリの部分集合とみなすことで，大きなくくりではあるが，関連語集合が作成されているとみなすことができる. しかし，細かなカテゴリにおいても，1 つのクラスタに 1 つのカテゴリが割り振られるという結果を得るためには，プログラムの改良を必要とする.

5 おわりに

本稿では，語句集合に対してディレクトリ検索によりカテゴリ分布を求め，それを手がかりとして出現する概念が類似している単語集合の構築を行った. 実験の結果，得られたクラスタの分類が適切であるかどうかの評価を行ったところ，分類語彙表から抽出した語句集合では正解率の平均が 73.6%であった. これより，同じ概念を持つ語句について，Web ディレクトリのカテゴリが意味的な特徴を持ち，同じクラスタに分類されやすい性質を持っていることが分かった. また，どのような文脈においても出現しやすい語句についてはひとつのクラスタにまとまりやすい性質を持つことも分かった. 今後は，前節で示した問題点を改良して，語句に対するより詳細な出現傾向を捉えることをめざし，さらに高精度な意味的関連語集合が作成できるようにする予定である.

参考文献

[1] George Karypis. CLUTO - a clustering toolkit. Technical Report #02-017, University of Minnesota, Department of Computer Science, 2003.

[2] Jiahui Liu and Larry Birnbaum. Measuring semantic similarity between named entities by searching the web directory. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 461–465, 2007.

[3] Simone Paolo Ponzetto and Michael Strube. An api for measuring the relatedness of words in wikipedia. In *ACL '07: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 49–52, 2007.

[4] Michael Strube and Simone Paolo Ponzetto. Wikirelate! computing semantic relatedness using wikipedia. *AAAI'06: proceedings of the 21st national conference on Artificial intelligence*, pages 1419–1424, 2006.

[5] 国立国語研究所. **分類語彙表 増補改訂版**. 大日本図書, 2004.