検索エンジンを利用した単語クラスタリング

大城亜里沙
茨城大学工学部
システム工学科

新納浩幸
茨城大学工学部
システム工学科

佐々木稔
茨城大学工学部
情報工学科

1 はじめに

本論文では Web をコーパスとして利用するための可能性や問題点を探るために，検索エンジンを利用した単語クラスタリングを行う．

コーパスを利用した統計的な自然言語処理はある程度の成功をおさめたが，いくつかの課題も残されている．その 1 つはコーパスのスパース性である．新聞記事 10 年分を調べても，日常的な単語「おにぎり」は数回しか出現しないという報告もある．一方，近年，Web が急速に普及し，Web 上の情報は莫大な量である．そこで Web を巨大なコーパスとみなして，従来行われていた統計的な自然言語処理を Web を利用して行うことが期待できる．Web は情報が莫大であるために，コーパスのスパース性は回避できると考えられる．

本論文では単語クラスタリングを題材にして，Web をコーパスとして利用するための可能性や問題点を探る．

Web のページは新聞記事のように単純なテキストではないので，単にページを収集しただけではコーパスとして利用することは難しい [3]．そこで本論文では検索エンジンを利用する．検索エンジンを利用した場合，ある単語群をクエリとして用いれば，その単語群を含むページ数を得ることができる．単語が他の単語とどの程度共起するかを表す共起分布を考えると，類似の単語どうしの共起分布は類似しているという性質がある．ここで共起の定義として単語がどの程度まで離れていてもよいかは様々である．通常は文や周辺数単語内ということが多い．ここでは同じページ内に出現していれば共起していると考える．すると，単語 A と単語 B が共起している頻度は検索エンジンに “A B” というクエリを渡すことで得ることができる．基本的に任意の 2 単語の共起の頻度を得ることができれば，単語を特徴ベクトルとして表現できるので，単語クラスタリングは実現できる．

本実験では 5 つのクラスを想定し，それぞれのクラスから 5 単語取り出し，合計 25 単語についてクラスタリングを行った．比較のために新聞記事 1 年間分のコーパスを利用したクラスタリングも行った．新聞記事を利用した場合，クラスタリングの結果は良好であったが，コーパスのスパース性の問題が確認できた．Web を利用したクラスタリングは，新聞記事を利用したものほどうまくはいかなかった．相関比が最大になるように基底の単語を設定し直すことで改良を試みたが，この場合も理想的なクラスタリング結果とは若干異っていた．うまくクラスタリングできない原因を考察する．

2 単語クラスタリング

2.1 特徴ベクトルの作成

テキストデータから注目する単語に対応する表層的な特徴を抽出し，個々の特徴の重みを表す数値を要素とした単語の特徴ベクトルを作成する．

クラスタリングの対象である n 語を $w_1, \cdots, w_n$ で表す． 単語を表現する特徴数を m とし，単語 w_i の特徴ベクトル w_i を下記のように表す．

$$w_i = (v_{i1}, \cdots, v_{im})$$

$v_{ij}(j = 1, \cdots, m)$ は，j 番目の特徴と単語 w_i の関係の強さを表す数値で，特徴の重みと呼ぶ．

ここでは検索エンジンを利用して特徴の重みを設定する．単語の特徴としては，ある単語 e_j と共起するかどうかを考え，単語 w_i の e_j に関する重み v_{ij} を測る．単語 e_j の設定方法は様々な手法があるが，ここでは空間上で基底をなすような単語群を直接選ぶことにした．論文 [1] で抽出された 922 クラスの単語クラスタの各クラスタから代表的な単語を選んだ．結果 922 単語選ばれ，単語 w_i は 922 次元の特徴ベクトルで表現されることになる．google を利用して，w_i と e_j の共起頻度をカウントし，正規化することで，特徴ベクトルを作成した．

またクラスタリング対象の単語群を以下のように定めた．まず，互いに関連性の少ない 5 つの概念のクラスを設定し，そのそれぞれに対して関連性のある 5 つの単語を選出した計 25 単語を設定した．つまり理想的には，クラスタリングにより 5 つの単語からなる 5

つのクラスタリングができるはずである．設定した概念のクラスとその単語は以下の通りである．

動物 プードル， チワワ， 犬， 猿， ゴリラ
食べ物 カレー， ラーメン， スパゲッティー， 焼きそば， ハンバーグ
感情や人生観 幸福， 満足， 愛情， 結婚， 運命
繁栄しているもの 情報， 知識， 手段， 交通， 設備
地名 今帰仁， 沖縄， ブリスベン， オーストラリア，オーストリア

2.2 クラスタリング手法

クラスタリング手法は大きく階層的手法と分割的手法に分類される [2].

(1) 階層的手法

階層的手法は，さらに分岐型 (divisive) と凝集型 (agglomerative) に分けられるが，ここでは後者のみを扱う．この手法は，1 個の対象だけを含む n 個のクラスタがある初期状態から，クラスタ間の距離 (非類似度) 関数に基づき，最も距離の近い二つのクラスタを逐次的に併合する．そして，この併合を，すべての対象が一つのクラスタに併合されるまで繰り返すことで階層構造を獲得する．

以下は代表的な Ward 法の距離関数 $\boldsymbol{D}(\boldsymbol{C}_1, \boldsymbol{C}_2)$ を表す．

$$\boldsymbol{D}(\boldsymbol{C}_1, \boldsymbol{C}_2) = \boldsymbol{E}(\boldsymbol{C}_1 \bigcup \boldsymbol{C}_2) - \boldsymbol{E}(\boldsymbol{C}_1) - \boldsymbol{E}(\boldsymbol{C}_2)$$

ただし，$\boldsymbol{E}(\boldsymbol{C}_i) = \sum_{\boldsymbol{x} \in \boldsymbol{C}_i} (\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{c_i}))^2$

Ward 法は，各対象から，その対象を含むクラスタのセントロイドまでの距離の二乗の総和を最小化する．

(2) 分割的手法

分割的手法は，分割の良さの評価関数を定め，その評価関数を最適にする分割を探索する．可能な分割の総数は単語数に対して指数的なので，実際は準最適解を求める．代表的な k-means 法では，セントロイドをクラスタの代表点とし

$$\sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i} (\boldsymbol{D}(\boldsymbol{x}, \boldsymbol{c_i}))^2$$

の評価関数を最大化する．最適解の探索のアルゴリズムを図 1 に示す．対象のクラスタへの割当てと代表点の再計算を交互に繰り返して行っている．この手法は山登り法で，局所最適解しか求められないため，ランダムに初期値を変更して，評価関数を最大にする結果を選択する．

1. k 個の代表点 $c_1, \cdots, c_k$ をランダムに選択
2. $\forall x \in \boldsymbol{X}$ を $min_i \boldsymbol{D}(\boldsymbol{x}, \boldsymbol{c}_i)$ なる代表点に割当て
3. $\boldsymbol{if}$ 代表点への割当が変化しない $\boldsymbol{then}$ 終了 $\boldsymbol{else}$ 各クラスタのセントロイドを代表点にしてステップ 2. へ

図 1: k-means 法アルゴリズム

3 実験

3.1 コーパスを利用したクラスタリング

比較実験のためにコーパスを利用したクラスタリングを行なう．

まずコーパスを利用して特徴ベクトルを作成する．単語の特徴としては，単語 e_j と 1 文中で共起するかどうかを考え，単語 w_i の e_j に関する重み v_{ij} を，コーパス中で w_i と e_j が共起した文の数で測る．単語 e_j の設定方法は，検索エンジンを用いたものと同じ単語を利用した．

コーパスとしては，95 年度版毎日新聞 1 年間分の記事を用いた．次に w_i と e_j の共起頻度をカウントし，正規化することで，特徴ベクトルを作成した．

ward 法と k-means 法でクラスタリングを行った結果を図 2，3 に示す．

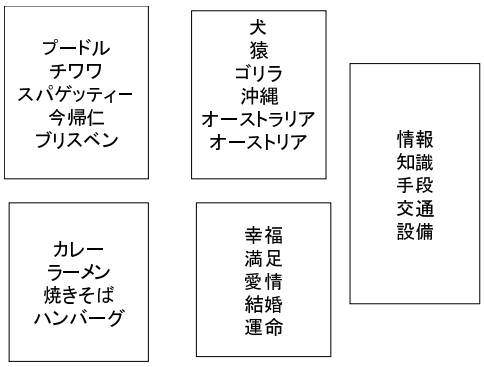

図 2: ward 法 (コーパス)

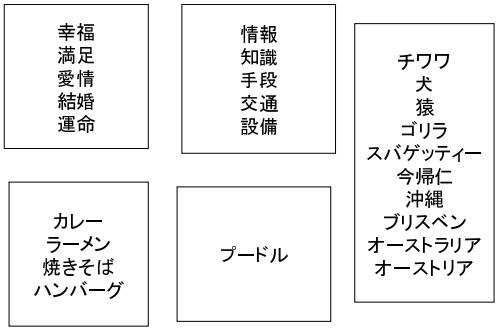

図 3: k-means 法 (コーパス)

コーパスを利用したクラスタリング結果をみると，どちらの手法も実験結果「犬」,「猿」,「ゴリラ」は同クラスになり,「沖縄」,「オーストラリア」,「オーストリア」も同クラスになっている．また，どちらの手法も「プードル」,「チワワ」,「スパゲッティー」,「今帰仁」,「ブリスベン」のクラスが誤っている．これら以外の単語については適切にクラスタリングされている．誤りの原因を探るため各単語の頻度を調べた．以下に各単語とその頻度を（ ）に示す．

動物 プードル (4)，チワワ (4)， 犬 (2422)， 猿 (124)，ゴリラ (123)

食べ物 カレー (185)，ラーメン (382)，スパゲッティー (0)， 焼きそば (44)， ハンバーグ (39)

感情や人生観 幸福 (635)， 満足 (1641)， 愛情 (416)，結婚 (2970)， 運命 (631)

繁栄しているもの 情報 (47095)， 知識 (2327)， 手段 (3864)， 交通 (10652)， 設備 (5449)

地名 今帰仁 (0)，沖縄 (11242)，ブリスベン (0)，オーストラリア (2904)， オーストリア (642)

「スパゲッティー」,「今帰仁」,「ブリスベン」は頻度 0 で,「プードル」,「チワワ」は頻度 4 であり，頻度が低いと誤分類されるが，頻度が高い単語は適切にクラスタリングされていることがわかる．よって，コーパスを利用したクラスタリングを行うと，コーパスのスパース性の影響があることが確認できる．

3.2 Web を利用したクラスタリング結果

google を利用して，単語 w_i と単語 e_j の共起頻度をカウントし，正規化することで，特徴ベクトルを作成した．

ward 法と k-means 法でクラスタリングを行った結果を図 4，5 に示す．

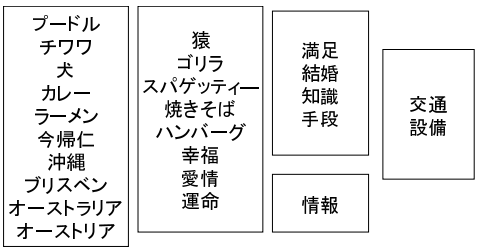

図 4: ward 法 (Web)

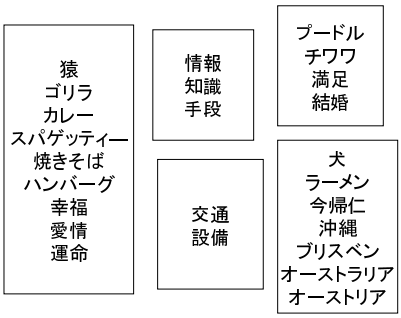

図 5: k-means 法 (Web)

Web を利用したクラスタリングでは，コーパスを用いたものほどうまくクラスタリングされなかった．

3.3 最適な次元の選出

Web を利用したクラスタリングは，コーパスを利用したものほどうまくはいかなかった．各クラス間の相関比の平均が最大になるように次元を削減して，改良を試みた．相関比が最大になるように基底の単語を設けると，384 次元が最適な次元となった．384 次元で ward 法と k-means 法でクラスタリングを行った結果を図 6，7 に示す．

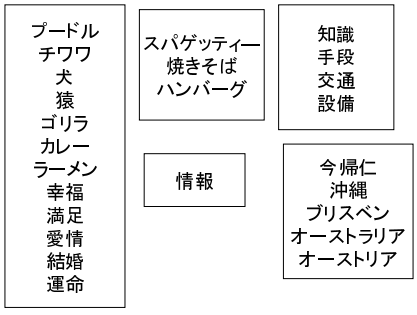

図 6: ward 法 (改良 Web)

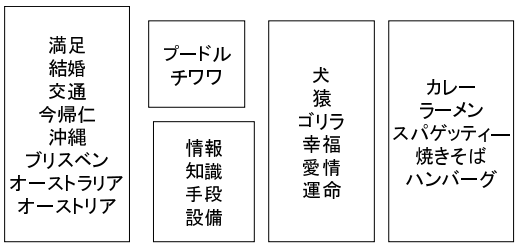

図 7: k-means 法 (改良 Web)

相関比の平均が最大になるように基底の単語を設定し直すことで改良を試みたが，理想的なクラスタリング結果とは若干異っていた．

4 考察

Web を利用したクラスタリングがあまりうまくいかなかった原因は 2 つ考えられる．

第 1 の原因は，基底の単語の選出方法である．ここでの基底の単語選出は，コーパス中の単語をクラスタリングすることから行なわれている [1]．このために，利用した基底はコーパスに対しては有効であったが，Web に対しては有効でなかった可能性がある．最適な次元でクラスタリングを行っても理想的な結果を得ることができなかったのも同じ理由からだと考えられる．

第 2 の原因は，共起の定義である．Web を利用したクラスタリングでは共起の定義を同じページ内に出現することとしている．しかしこれは共起の定義としては荒い可能性がある．

ちなみに，以下のような実験を行なってみた．

最適な次元となる 384 次元のデータから，同じクラスで距離が最も離れている単語の対を 1 つ選んだ．ここでは「情報」と「交通」であった．次にこの 2 単語で最も距離が離れた次元を選んだ．この次元に対応する単語は『浴室』であった．“情報 浴室” と “交通 浴室” をクエリにして google で検索し，提示された上位 20 ページの内容を調べてみた (実験 A)．括弧内の数値は上位 20 ページ中のページ数である．

「情報」と『浴室』の検索結果
住関連用品 (7)，リフォーム (6)，宿・旅館施設 (3)，住宅設備と建材 (2)，生活情報 (2)
「交通」と『浴室』の検索結果
宿・旅館施設 (7)，賃貸物件 (4)，住宅設備 (4)，旅情報 (1)，(中国語のため内容不明 (4))

次に，同じクラスで距離が最も近い単語の対を 1 つ選んだ．ここでは「プードル」と「チワワ」であった．次にこの 2 単語で最も距離が近い次元を選んだ．この次元に対応する単語は『暫定』であった．“プードル 暫定” と “チワワ 暫定” をクエリにして google で検索し，提示された上位 20 個のページの内容を調べてみた (実験 B)．

「プードル」と「暫定」の検索結果
エッセイ・日記等 (15)，動物の画像 (1)，クイズ (1)，動物の病状 (1)，書籍 (2)
「チワワ」と「暫定」の検索結果
エッセイ・日記等 (15)，ペットグッズ (3)，メキシコ年表 (1)，ボクシング (1)

実験 A ではページの内容が異なり「情報」と「交通」の類似性が反映されていない．実験 B ではページの内容が似通っており「プードル」と「チワワ」の類似性が反映されている．つまりここでの共起の定義はある部分では有効に機能しているが，別の部分ではまったく機能していない．

今後は基底の単語の選び方や有効な共起の定義及びその定義に基づく共起頻度を Web から得る方法を考えたい．

5 おわりに

ここでは Web をコーパスとして利用するための研究の一貫として，検索エンジンを利用したクラスタリングを行なった．

新聞記事を利用したクラスタリングは良好であったが，コーパスのスパース性の問題があることが確認できた．Web をコーパス代わりにするために，検索エンジンを利用したクラスタリングを行ったが，コーパスを利用したものほど うまくはいかなかった．

原因は基底の単語の選出方法とここでの共起の定義が荒いためであると考えられる．今後は基底の単語の選び方や有効な共起の定義及びその定義に基づく共起頻度を Web から得る方法を考えたい．

参考文献

[1] M. Sasaki and H. Shinnou: “Automatic thesaurus construction using word clustering”, PACLING-03, pp.55-62, 2003.

[2] 神嶌敏弘：“データマイニング分野のクラスタリング手法 (1) －クラスタリングを使ってみよう!－”, 人工知能学会誌, Vol.18, No.1, pp.59–65, 2003.

[3] 関口洋一, 山本和英：“Web コーパスの提案”, 情報処理学会自然言語処理研究会,NL-157-17, pp.123–130, 2003.