シソーラスを利用した文書クラスタリングにおける次元圧縮アルゴリズムの性能評価

Evaluation of Dimensionality Reductions in Document Clustering Using Term-Vectors with Thesaurus

酒井将太†　　新美彩彦†
†公立はこだて未来大学システム情報科学部

1. はじめに

テキストマイニングにおいては、多次元ベクトルモデルによる解析が盛んに行われている。多次元ベクトルモデルを利用した方法では文書中に含まれる単語の出現頻度を利用すると、データスパースネスな多次元ベクトルとなり、性能や計算時間に問題が出ることは既に広く知られている。また、近年では様々なシソーラスが構築されており、シソーラスを利用し単語の持つ意味を用いた研究も盛んである[1]。そこで本研究ではシソーラスを利用し単語の上位語、その単語の属する領域を検索した上で、それを元に特徴ベクトルを構築することを試みる。これにより、多次元ベクトルの性能を改善する事ができると期待される。さらに、シソーラスを用いた特徴ベクトルに対して代表的な次元圧縮・選択・変換アルゴリズムを適用し、シソーラスの利用と次元圧縮・選択・変換アルゴリズムの関係の評価を行った。

2. 次元圧縮・選択・変換アルゴリズムの性能評価

本章では、文書クラスタリング問題に対し、提案するシソーラスを使った多次元ベクトルの構築と次元圧縮・選択・変換アルゴリズムの組み合わせについて述べる。まず、利用するシソーラスについて述べた後、シソーラスを用いた特徴ベクトルの構築法を提案する。構築した特徴ベクトルに対し、次元圧縮・選択・変換アルゴリズムを適用する。その後文書クラスタリングを行う。

2.1 シソーラス

本研究ではシソーラスとして日本語WordNet[2][3]を用いた。日本語WordNetは日本語の概念辞書であり、約57,000の概念と93,000の語を収録している。個々の概念はsynsetという単位にまとめられており、それらが意味的に結びついている。このシソーラスを用い、文書中に出現する単語のsynsetを特徴ベクトルの要素とすれば文書中に含まれる意味、概念を元にした特徴ベクトルとなり、単語の出現頻度を元にした特徴ベクトルよりも文書の内容をより良く特徴づける特徴ベクトルが構築できると考える。

2.2 特徴ベクトル

文書から以下の2つの特徴ベクトルの構築を行った。

1) 文書-単語ベクトル:文書中に出現する単語を要素としたベクトル。文書idを行にとり、出現する単語を列に取るように特徴ベクトルを構築した。

2) 文書-synsetベクトル:文書中に出現する単語をシソーラスで検索し、得られたsynsetを要素としたベクトル。

文書-単語ベクトルにおいては、文書を形態素解析し、得られた形態素列を要素とし、重み付けにはtfidfを用いた。

文書-synsetベクトルにおいては、まず文書を形態素解析し形態素列を得る。その後得られた全単語列に対し日本語WordNetで「上位語」、「領域」を検索する。多くの場合1つの単語に対して複数の上位語・領域が存在する。

そしてその単語のsynset,上位語のsynset、領域のsynsetをベクトルの要素とし、前者同様重み付けにtfidfを用いた。また,態素解析器にはSen[4]を用いた。

2.3　次元圧縮アルゴリズム

構築した2つのベクトル(文書-単語ベクトル、文書-synsetベクトル)に対して以下の3つのアルゴリズムを適用した。

1)　潜在的意味インデキシング[5]:文書単語ベクトルおよび文書-synsetベクトルVに対し、特異値分解によって

$V = U\Sigma D$と分解した。ここで

U，Dは左あるいは右特異ベクトル、

Σは特異値を対角成分に持つ対角行列である。特異値分解で得られたベクトルを利用し、以下のように変換した。

$V_k = U_k \Sigma_k D_k$

ここでU_K, D_Kは最初のk個の左あるいは右特異ベクトルであり、Σ_kは大きいほうからk個の特異値を対角成分に持つ対角行列である。

この変換を行うことで近似行列を作成した。またkの値は30とした。

2)　主成分分析:2つのベクトルに対し、不偏分散共分散行列から主成分を求め、累積寄与率が0.8以上となる主成分までの主成分スコアを採用した。

3)　属性選択法:カイ2乗値を用いた属性選択により、カイ2乗値が高いほうから順にm個の属性を選択した。mは(特徴ベクトルの次元数)*(0.3)とした

2.4 文書クラスタリング

2.1で述べた2つのベクトルに対し、2.2で述べた4つの次元圧縮/選択アルゴリズムを適用し、ウォード法による階層的クラスタリングを適用し文書クラスタリングを行った。

3. 実験

本章では提案した手法の有効性を検証するために行った実験について述べる。

3.1 実験設定

2章で述べた手順を元に実験を行った。対象としたデータセットは楽天データ公開 [6]で公開されたインターネットショッピングサイトの商品データ(データセットA)とインターネットショッピングAmazon[7]においてユーザーが投稿した商品に関するレビュー(データセットB)である。

データセットAはあらかじめ階層構造に分類されている。あらかじめ分類されているカテゴリを正解とし、実験によってクラスタリングされた結果と比較することによって検証した。全商品データから100商品のデータをランダムサンプリングし、それを1セットとし、30セット用意した。

データセットBは1セットに約200件のレビューが存在し10セット用意した。

3.2 両ベクトルの違い

文書-単語ベクトルと文書-synsetベクトルについて、次元数と密度について計測を行った。次元数とは各ベクトルの列数、密度については(ベクトルの非ゼロ要素)/(ベクトルの全要素数)として計測した。計測された結果を表1、表2に示す。

表1 データセットAの
特徴ベクトルの次元数と密度

ベクトル	平均次元数	平均密度
文書-単語	924.8	0.042
文書-synset	1400.2	0.073

表2 データセットBの
特徴ベクトルの次元数と密度

ベクトル	平均次元数	平均密度
文書-単語	1049.3	0.035
文書-synset	1302.1	0.065

表1、表2より、データセットA、B共に平均次元数は文書-単語ベクトルのほうが低い。しかし平均密度は文書-synsetベクトルの方が高くなった。文書-synsetベクトルは文書中に出現する1つの単語に対して日本語WordNetによって検索された「上位語」・「領域」が複数存在するために次元数が増えたものと考えられる。また、文書-synsetベクトルの平均密度が増えたことについては、2.1で述べたようにより良く文章の意味を特徴づける特徴ベクトルが構築されたと考えられる。

しかし2つのベクトルの密度はともに数%であり、スパースネス問題を劇的に解決できていない。

3.3 性能評価

2.4で示した階層型クラスタリングの結果について述べる。

データセットAについては、1セットに約20個のジャンルが含まれているので、約20のクラスにクラスタリングした。

データセットBについては10のクラスにクラスタリングした。図1,図2に各クラスタに含まれる文書数についてのグラフを示す。図1、図2は横軸がクラスタ番号、縦軸がそのクラスタに含まれる文書数を表している。図1より、データセットAを使用した場合はどの方法を用いてクラスタリングを行ってもクラスタに含まれる文書数の分布に差は見られなかった。

図2より、データセット2を使用した場合については、文書-概念IDベクトルに主成分分析、属性選択のアルゴリズムを適用した場合はクラスタに含まれる文書数の分布にばらつきがでた。

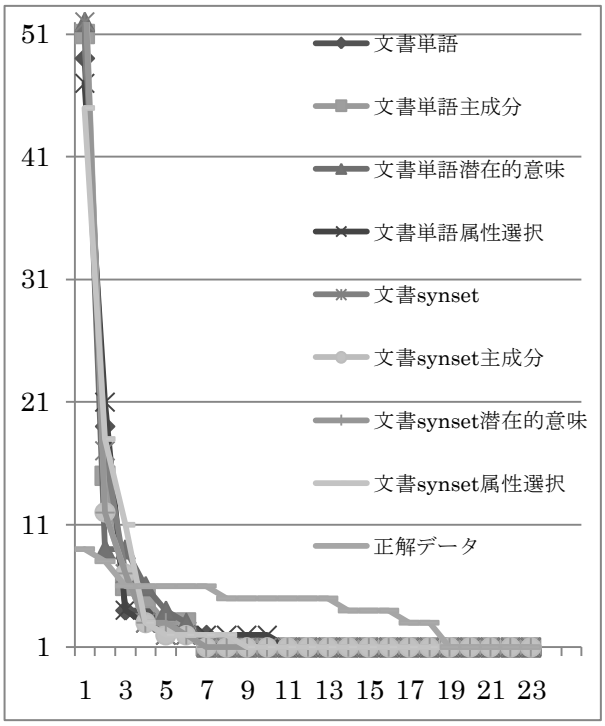

図1 データセットAのクラスタと
含まれる文書数

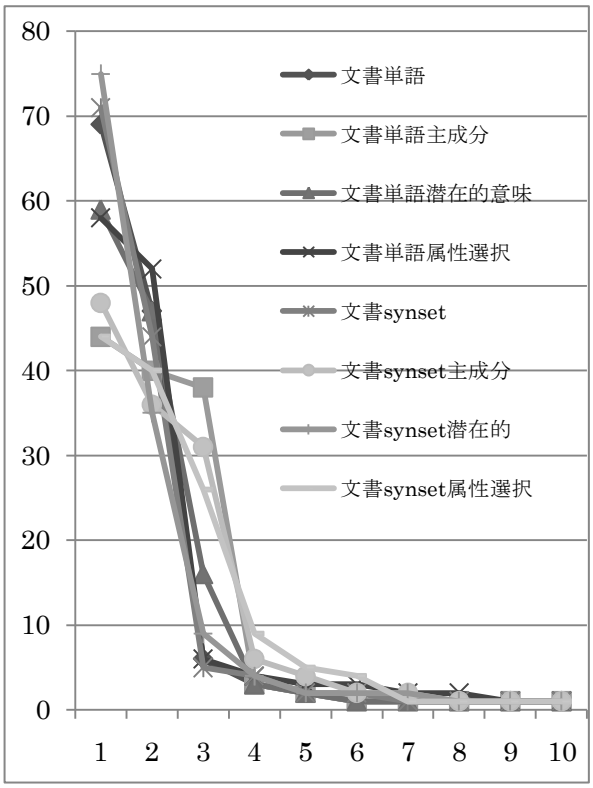

図2 データセットBのクラスタと
含まれる文書数

4. 考察

3.3より、データセットBを用いた場合はクラスタリング結果に違いが出た。データセットAは商品に関するデータ、Bは商品に関する「良い」「悪い」といった感性表現を含むデータであり、データセットの性質の違いによるものと考えられる。また、データセットAは商品データであり、商品サイズなどの表記に「150」「200」といった単語列が多く見られたが、これらの単語列の上位語を日本語WordNetで検索すると全て「数字」という上位語が得られるため本研究においてはそういった表現がノイズとなってしまいうまくクラスタリングできなかったと考えられる。

データセットBについては、文書-synsetベクトルに対し主成分分析と属性選択のアルゴリズムを適用しクラスタリングをした場合に分布にばらつきが出た。データセットBは投稿者の意見を含むレビューでありどの方法が一番よくクラスタリングできているかは判断が難しいが、この方法を取ることにより何らかの違いが出た。

5. 終わりに

本研究ではシソーラスを利用し、単語の上位語、その単語の属する領域を検索した上で、それを元に特徴ベクトルを構築する手法を試みた。これにより多次元ベクトルの性能を改善することができると期待される。さらにシソーラスを用いた特徴ベクトルに対して代表的な次元圧縮・選択・変換アルゴリズムを適用し、シソーラスの利用と次元圧縮・選択・変換アルゴリズムの関係の評価を行った。 2つのデータセットを用いた実験では平均次元数は文書-単語ベクトルの方が低いが、平均密度は文書-synsetベクトルの方が高くなった。文書-synsetベクトルは文書中に出現する1つの単語に対して日本語WordNetで検索された「上位語」・「領域」が複数存在するため次元数が増えたと考えられる。

データセットのクラスタリングについては、現在、結果をまとめているところである。

参考文献

[1] 村松祐希 , 山本和英 (2010)語彙知識を用いた日本語テキスト含意認識評価セット構築と認識実験 言語処理学会第16回年次大会　発表論文集

[2]日本語WordNet http://nlpwww.nict.go.jp/wn-ja/ (最終アクセス日 2011年1月22日)

[3]Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi and Kyoko Kanzaki (2009) Enhancing the Japanese WordNet in The 7th Workshop on Asian Language Resources, in conjunction with ACL-IJCNLP 2009, Singapore. pp8

[4]形態素解析システム Sen http://www.mlab.im.dendai.ac.jp/~yamada/ir/MorphologicalAnalyzer/Sen.html (最終アクセス日 2011年1月22日)

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990) Indexing by Latent Semantic Analysis, Journal of the Society for Information Science, 41(6), 391-407, 1990.

[6]楽天データ公開 http://rit.rakuten.co.jp/rdr/ (最終アクセス日 2011年1月22日)

[7]Amazon.co.jp http://www.amazon.co.jp/(最終アクセス日 2011年1月22日)