表の構造を利用した類義語抽出

関 恒仁† 嶋田和孝‡ 遠藤 勉‡
† 九州工業大学大学院情報工学研究科 ‡ 九州工業大学情報工学部

1 はじめに

Web 上において，テキストは情報を表す標準的な手段として用いられる．我々が Web 上のテキストから必要な情報を検索する場合の問題点として，テキスト中の表記の揺れや類義語などの存在が挙げられる．この問題の一般的な解決支援には，シソーラスが利用される．しかしながら，シソーラスの構築には人手を要するため，汎用性の高いシソーラスの構築には非常にコストがかかるといった問題がある．さらに，既存のシソーラスには，日々生成される新しい語や，専門性の高い語などは，殆ど記載されていないという問題もある．したがって，様々な分野や利用目的に対応したシソーラスの自動構築が必要となっている．

ここで，シソーラス自動構築を目的として類義語抽出を行う際の対象コーパスとして，文書ではなく表形式のデータを利用することを考える．表形式のデータ (以下，表データ) は，情報を分かりやすく表現するために様々な状況で利用されており，多くの言語情報を含んでいる．Web 上に存在する表データの具体例として，パソコンやデジタルカメラといった製品のスペック情報を記述した表データが存在する．パソコンのスペック情報を記述した表データの例を図 1 に示す．本研究では，主にこのような属性と属性値によって構造化された表データを対象に議論する．

属性		属性値
プロセッサ		インテル® Pentium®4 1.3GHz
2次キャッシュ		512KB
メモリ	メインメモリ	256MB(DDR SDRAM)
	ビデオメモリ	64MB
外部記憶	HDD	40GB(4200rpm／9.5mm)
	FDD	3.5型(1.44MB/720KB)
本体寸法		幅314×奥行き260×高さ34(mm)
質量		2.9 kg

図 1: パソコンのスペック情報を記述した表データ

表データにおける属性は，属性値に記述されたオブジェクトや数値表現を知るための重要な情報を明記している．例えば，図 1 のような表データにおいて我々は，プロセッサ，HDD などの属性に対応する属性値からパソコンのスペックの良し悪しを判別している．しかしながら，表データの属性についても，表現の多様性が存在する．例えば，“プロセッサ” と “CPU”，“HDD” と “ハードディスク” などは同じ意味を表す属性であり，他の属性に関しても同様に表現の多様性が存在する．このような属性表現の多様性を解消することは，その分野のテキスト情報や表データを機械的に利用する際に重要な問題となる．本研究では，このような表データ中で属性を表す語を対象語として，表現の多様性を解決するために類義語抽出を行う．

本研究は，属性と属性値によって構造化された表データを対象とする．この属性と属性値間の対応関係を利用し，類義語を自動抽出する手法を提案する．

2 索引語・対象語行列

表データ中の属性を表す語を類義語抽出処理の対象語として扱うために，まず，対象語をベクトルで表現する．

属性との明確な対応関係を持つ属性値は，属性表現を特徴付ける語から成り立つと考えられる．そこで，対象語に対応する属性値から，対象語を特徴付けると考えられる索引語を抽出する．具体的には，属性値の文字列に対して「茶筌」[1] を用いて形態素解析を行い，以下の品詞を抽出する．

- 名詞 (代名詞，数字を除く)
- 未知語
- 接頭詞
- 記号 (アルファベット)

数値表現については，特徴的なパターンを持ち得る離散的な数値と，特徴的なパターンを持たない連続値が混在し，その判別が困難であるためストップワードとした．

索引語を要素として対象語をベクトルで表現するとき，各対象語に共起する索引語の頻度情報に統計的な重みを加え，その数値をベクトルの要素とする．ここでの重み付けは，局所的重み L_{ij} を対象語 w_j に対する索引語 t_i の重み，大域的重み G_i を表データ全体における索引語 t_i の重みとすると，それぞれ以下のように表す．

$$L_{ij} = log(1 + f_{ij}) \quad (1)$$

$$G_i = \frac{F_i}{n_i} \quad (2)$$

ここで，f_{ij} は対象語 w_j に共起する索引語 t_i の頻度，F_i は全表データ集合における索引語 t_i の頻度，n_i は索引語 t_i と共起する対象語数を表す．これにより，対象語 w_j を表す対象語ベクトルにおける索引語 t_i を表した，索引語・対象語行列 X の要素 x_{ij} は以下のようになる．

$$x_{ij} = L_{ij} \times G_i \quad (3)$$

これらの処理で作成した索引語・対象語行列において，対象語を特徴付ける索引語間の意味的な関連付けを図るために，潜在的意味解析を利用する．

潜在的意味解析

潜在的意味解析は，高次元の空間にあるベクトルを低次元の空間へと射影することにより，ベクトルの次元を圧縮する技術である [3]．高次元空間では別々に扱われていた索引語が，低次元空間では相互に関連を持ったものとして扱われる可能性もあるため，索引語を意味や概念に基づいて関連付けることが可能となる．

[1] http://chasen.naist.jp/hiki/ChaSen/

3　クラスタリング

これまでに述べた処理によって得られた対象語ベクトルについて，意味的に類似する対象語同士をいくつかのクラスタに自動分類する．基本的なクラスタリング手法としては，球面 k 平均アルゴリズム [4] を利用する．

球面 k 平均アルゴリズムは，ユークリッド空間内における概念ベクトルと対象語ベクトル間の内積を類似度として，多次元空間の単位球を分割することによりクラスタリングを行うものである．m 次元ベクトル空間における単位球面上に n 個の対象語ベクトル $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n$ が与えられたと仮定する．この時，$\pi_1, \pi_2, ..., \pi_k$ は対象語ベクトルを k 個のクラスタに分割した際の各クラスタとする．ここで，各クラスタは意味的に類似する対象語同士で構成された類義語グループとも考えられる．このとき，あるクラスタ π_j $(1 \leq j \leq k)$ の概念ベクトル $\mathbf{c}_j$ は，π_j の重心ベクトルのノルムを正規化したものである．

本アルゴリズムでは，クラスタ π_j の結合密度を以下の式で評価する．

$$\sum_{\mathbf{x} \in \pi_j} \mathbf{x}^T \mathbf{c}_j \tag{4}$$

また，クラスタリングによって与えられるクラスタを評価する指標として，全クラスタの結合密度の総和を目的関数とし，以下に示す目的関数が局所的に最大となるまでクラスタリングを繰り返す．

$$Q = \sum_{j=1}^{k} \sum_{\mathbf{x} \in \pi_j} \mathbf{x}^T \mathbf{c}_j \tag{5}$$

3.1　クラスタリングの拡張

球面 k 平均アルゴリズムは，分割最適化クラスタリング手法において代表的な手法である k-means 法のバリエージョンの 1 つである．これら分割最適化手法は，高速なクラスタリング手法として知られているが，初期分割の基準となる初期種子点を定めずに初期分割をランダムに行うと，同じクラスタ数においても結果に揺れが生じてしまうという問題がある．また，クラスタ数を予め与えなければならないといった問題もある．最適なクラスタ数を定めることができれば，これらの手法を用いることで良好な結果を得ることができるが，対象となるデータを変更する度に最適なクラスタ数を人手で定めるのは未知データへの適用が困難であるといえる．したがって，初期種子点を求めることや，最適と考えられるクラスタ数を推定することが効率的な類義語抽出を行う上で重要となる．本研究では，表データ中の属性を表す語について表現の多様性を解決することを目的としているため，表データの性質を基に，初期種子点問題やクラスタ数問題の解決を試みる．

3.1.1　初期種子点

対象となる表データは属性数が一定ではないことが多く，予め最適なクラスタ数を定めることは困難であるが，それぞれの表データにおける属性数に着目すると，クラスタ数は属性数が最小の表データから得られた対象語の数以上になると考えられる．つまり，最小属性数の表データから得られた対象語数は，考えられる最小のクラスタ数といえる．そこで，対象となる表データ中で属性数が最小となる表データから得られた対象語ベクトルを初期種子点としてクラスタリングを開始する．

3.1.2　クラスタ分割処理

最小属性数の表データから得られた対象語ベクトルを初期種子点とすることで，考えられる最小のクラスタ数でクラスタリングを行うことになる．ここで，考えられる類義語グループ数に対して，与えるクラスタ数が小さい場合，1 つのクラスタに複数の類義語グループが存在したり，類義語グループ中に不要な対象語が分類されると考えられる．このとき, このようなクラスタに属する対象語ベクトルのばらつきは大きくなるため，クラスタの分割処理を行う必要がある．このクラスタ分割処理について詳しく述べる．ここで，各クラスタに属する対象語ベクトルのばらつきを評価する指標として，各クラスタにおける偏差平方和を利用する．本研究では，これをクラスタ内平方和と呼ぶ．あるクラスタ π_j の重心ベクトルを $\mathbf{m}_j$ とするとき，クラスタ内平方和 $ss(j)$ を以下の式で定義する．

$$ss(j) = \sum_{\mathbf{x} \in \pi_j} \|\mathbf{x} - \mathbf{m}_j\|^2 \tag{6}$$

$ss(j)$ は，クラスタ π_j に属する対象語ベクトルのばらつきが大きい程に大きな値をとる．ここで，全クラスタ中でクラスタ内平方和が最大となるクラスタ π_i を仮定し，クラスタ π_i 空間の分割処理について述べる．全クラスタ中でクラスタ内平方和が最大であったクラスタ π_i は，クラスタ中に複数の類義語グループが存在していると考えられる．そこで，クラスタ π_i に属する対象語ベクトル中で，最遠距離にある 2 つの対象語ベクトルを初期種子点としてクラスタリングを行うことによってクラスタ π_i 空間を 2 分割する．π_i を 2 分割した結果を利用して再びクラスタリングを行うことにより，全体のクラスタ数は 1 つ増加したことになる．この処理を繰り返すことで，逐次的にクラスタ数を増加させていく．具体的なアルゴリズムを以下に示す．

i. 最小属性数の表データから得られた対象語ベクトル k_f 個を初期種子点とし，$k = k_f$ でクラスタリングを行なう．
ii. クラスタ内平方和が最大となるクラスタ π_i を求め，π_i に属する対象語ベクトルの中で最遠距離にある 2 つの対象語ベクトル $\mathbf{x}_{i1}$，$\mathbf{x}_{i2}$ を求める．
iii. $\mathbf{x}_{i1}$ と $\mathbf{x}_{i2}$ を初期種子点として，クラスタ π_i に属する対象語ベクトルに対してクラスタリングを行なう．
iv. ii のクラスタリング結果における概念ベクトル集合からクラスタ π_i の概念ベクトル $\mathbf{c}_i$ を削除し，iii のクラスタリング結果の 2 つの概念ベクトル $\mathbf{c}_{i1}$ と $\mathbf{c}_{i2}$ を追加する．
v. iv で得た概念ベクトル集合を初期種子点とし，$k = k+1$ で再度クラスタリングを行い，ii へ戻る．(以降，k が十分に大きくなるまで，ii〜v を繰り返し行う．)

3.1.3　クラスタ数の推定

近年，類義語抽出を目的とした単語クラスタリング手法の提案が多く行なわれており，そのようなクラスタリング手法において分割最適化手法が利用されることも多い [1] [2]．しかしながら，これらのクラスタリング手法において重要な問題となるクラスタ数の最適性についてはあまり議論されていない．クラスタ数は，対象語群に対して意味的な類似性を持つクラスタがいくつ生成されるべきかを示す重要な値となる．特に，本研究は日々増え続ける Web 上

の情報を対象としているため，未知データに対してクラスタリングを行なうことを想定し，最適なクラスタ数を動的に推定することが重要である．

本研究では，逐次的にクラスタ数を増加させながらクラスタリングを繰り返し行い，各クラスタ数におけるクラスタリング結果を統計的に評価することにより，最適と考えられるクラスタ数の推定を行う．まず，クラスタ数 k におけるクラスタリング結果を評価する指標として，総平方和 $W(k)$ を以下のように定義する．

$$W(k) = \sum_{j=1}^{k} ss(j) \tag{7}$$

$W(k)$ は，式 (6) で定義したクラスタ内平方和を k 個のクラスタについて総和したものである．各クラスタのばらつきが小さい程に，$W(k)$ も小さくなるため，クラスタ数 k におけるクラスタリング結果全体を評価する指標となる．しかし，1 つの対象語ベクトルが単独でクラスタを構成する場合には，そのクラスタのクラスタ内平方和が 0 となり，総平方和も減少する傾向がある．このような傾向はクラスタ数が大きくなる程に起こりやすい．そこで，クラスタ数を $k-1$ から k へ，さらには k から $k+1$ へと変化させた場合のクラスタリング結果における総平方和の変化率を計ることで最適と思われるクラスタ数の推定を行う．

クラスタ数の推定には，Krzanowski と Lai の提案した統計量 (以下，KL 統計量と呼ぶ) を用いる [5]．クラスタ数 k における KL 統計量 $KL(k)$ は，対象となるベクトル空間の次元を m とすると，式 (8) で定義される．

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \tag{8}$$

$$DIFF(k) = (k-1)^{2/m} W(k-1) - k^{2/m} W(k) \tag{9}$$

$KL(k)$ は，クラスタ数 k を変化させたクラスタリング結果における総平方和の変化率を，ベクトル空間の次元に適応させて評価する指標となっている．

KL 統計量は本来，$KL(k)$ が最大となる k を最適なクラスタ数とみなすが，本研究では，クラスタ数 k がある閾値以上大きくなってから KL 統計量の計測を行い，その計測中で $KL(k)$ が最大となる k を最適なクラスタ数とする．これは，3.1.2 節で述べた，クラスタ数 $k = k_f$ でのクラスタリングでは，多くのクラスタにおいてクラスタ内平方和は大きくなっているため，分割処理における総平方和の変化が大きくなり，KL 統計量が高い値をとる傾向にある．したがって，k_f に近いクラスタ数での KL 統計量は最適なクラスタ数を判断する指標としての信頼性が高いとはいえない．言い換えると，クラスタ数が十分大きくなってからのクラスタリングにおいて，KL 統計量が高い値をとるクラスタ数が，最適なクラスタ数として信頼性が高いといえる．

4 実験・考察

提案手法に基づく類義語抽出処理の有効性を示すために，実際に Web 上から収集した表データに対して類義語抽出の実験を行った．使用した表データは，パソコン (PC)，デジタルカメラ (DC)，プリンタ (PR) の 3 種類の製品カテゴリについて，製品メーカの Web サイトから収集した各製品のスペック情報を記述した表データである．各製品カテゴリにおいて，実験に使用した表データ数，対象語の異なり数などの詳細を表 1 に示す．表 1 中の属性・属性値対とは，表データ中の属性とそれに対応する属性値の組のことで，属性・属性値対の抽出は類義語抽出処理の事前処理として行っている．各製品カテゴリにおいて，入力となる表データから属性・属性値対を抽出し，索引語・対象語行列の作成を行った．それぞれの索引語・対象語行列において，潜在的意味解析を利用して近似行列を作成し，対象語ベクトルのクラスタリングを行った．クラスタリング処理では，閾値以上のクラスタ数において KL 統計量が最大となるクラスタ数を最適クラスタ数とした．ここで，KL 統計量の計測を開始するクラスタ数の閾値は，各製品カテゴリの表データ中で，最大属性数の表データから得られた対象語数とした．

表 1: 実験に用いた表データの詳細

カテゴリ	表データ数	属性・属性値対	対象語数
PC	20	646 組	239 語
DC	19	583 組	189 語
PR	18	393 組	130 語

PC カテゴリにおける実験で，KL 統計量の計測結果をグラフで表したものを図 2 に示す．各製品カテゴリでの KL 統計量の計測結果において，KL 統計量が高い値を示したクラスタ数を順にまとめたものを表 2 に示す．表 2 において，* は実験結果から人手で判断した最適なクラスタ数である．PC および DC カテゴリの場合，KL 統計量が最大となるクラスタ数と人手で判断した最適なクラスタ数が一致した．PR カテゴリについては，KL 統計量が最大となるクラスタ ($k = 31$) では，十分に分割できていないクラスタがいくつか存在した．PR カテゴリにおいて，人手によって最適と判断されたクラスタ数は，$k = 41$ であった．これは，KL 統計量が 2 番目に大きくなるクラスタ数と一致した．

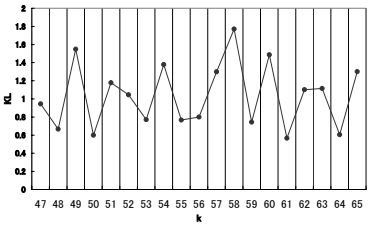

図 2: PC の表データにおける KL 統計量

表 2: KL 統計量の計測結果

カテゴリ	1 位	2 位	3 位
PC	**58***	49	60
DC	**59***	56	53
PR	31	**41***	38

*** … 最適クラスタ数**

このようにして得られた最適クラスタ数におけるクラスタリング結果を，類義語抽出の結果とした．各製品カテゴリの表データから類義語抽出処理によって抽出された類義語を，それぞれ図 3，図 4，図 5 に示す．本研究における類義語抽出の実験では，精度を数値的に評価することが困難である．これは，正解データを作成する場合に類義語の判別に主観が大きく作用するためである．そこで，今回の実験結果については，抽出された類義語を示すだけに留まった．

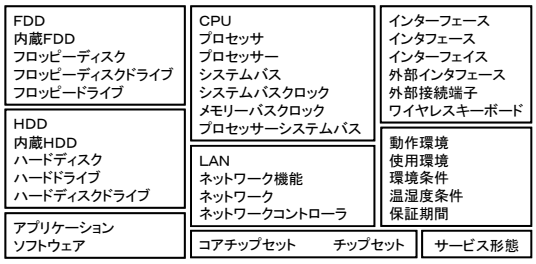

図 3: PC の表データから得られた類義語

図 4: DC の表データから得られた類義語

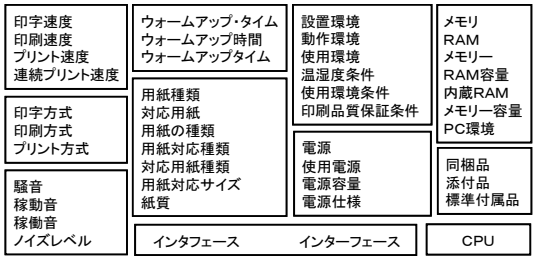

図 5: PR の表データから得られた類義語

以下，実験結果に対する考察を行う．クラスタ数推定についての考察としては，PC および DC カテゴリでは KL 統計量が最大値を示したクラスタ数において，それぞれ最適クラスタ数が得られた．PR カテゴリにおいては，KL 統計量が最大となったクラスタ数が最適クラスタ数とはいえなかったが，次に高い値を示したクラスタ数おいて，最適クラスタ数が得られた．これにより，KL 統計量が高い値を示すクラスタ数において，最適クラスタ数の推定が上手く行えているといえる．しかし，初期種子点として利用した対象語に最適クラスタ数が依存する可能性も考えられる．そこで，初期種子点の変更がクラスタ数推定にどれほど影響があるかを確認するために，追加実験として初期種子点とする表データをいくつか変更してクラスタリングを行った．追加実験における各製品カテゴリの最適クラスタ数の平均値を表 3 に示す．追加実験の結果においても，PC，DC カテゴリの最適クラスタ数は表 2 に示した最適クラスタ数と比較的近い値で収束した．PR カテゴリにおいては，比較的小さいクラスタ数で収束する傾向にあったが，KL 統計量が次に高い値を示したクラスタ数まで考慮に入れると，その中で最適クラスタ数が得られていた．また，最適クラスタ数におけるクラスタリング結果についても，それほど大きな精度の差は見られなかった．

表 3: 追加実験における最適クラスタ数の平均値

製品カテゴリ	平均クラスタ数
パソコン	56.1
デジカメ	55.8
プリンタ	36.2

図 3，図 4，図 5 に示した類義語抽出結果についての考察としては，各製品カテゴリの特徴的な語において，意味的に類似する対象語同士を類義語として抽出することができており，比較的良好な結果を得たといえる．また，各製品カテゴリの表データにおいて，特殊な属性表現で意味的に類似する語が他に存在しないような対象語も，単独でクラスタを構成するようにクラスタリングが行われており，クラスタ分割処理も有効に機能しているといえる．しかし，処理対象とする表データの規模や種類を変更した実験を行い，本手法の有効性を確認するための更なる考察が必要であると考えられる．

5 おわりに

本稿では，Web 上に存在する表データの中で，特に製品のスペック情報を記述した表データを対象に，属性表現の多様性を解決するために類義語を自動抽出する手法の提案を行った．

3 つの製品カテゴリに対して，Web 上から収集した表データを基に類義語抽出実験を行ったところ，良好な結果が得られていることが確認できた．今後は，対象とする表データの規模や種類を変更した実験や，実際に抽出した類義語を辞書として利用することにより，本研究で提案した類義語抽出手法の有効性を確認する必要がある．

参考文献

[1] 佐々木稔, 新納浩幸,“ 単語クラスタリングの語義判別問題への応用”. 情報処理学会自然言語処研究会, 154-21, pp. 145-152, 2003.

[2] 大城亜里沙, 新納浩幸, 佐々木稔, “ 検索エンジンを利用した単語クラスタリング”. 言語処理学会第 10 回年次大会, pp. 17-20, 2004.

[3] 北研二, 津田和彦, 獅々堀正幹,“ 情報検索アルゴリズム”. 共立出版, 2002.

[4] I. S. Dhillon and D. S. Modha,“ Concept decompositions for large sparse text data using clustering”. Technical report, IBM Almaden Research Center, 1999.

[5] Krzanowski, W. J. and Lai, Y. T,“ A criterion for determining the number of clusters in a data set”. Biometrics,44, pp. 23-34, 1985.