概要

近年，ウェブ上に蓄積された膨大な電子文書から情報を得る機会が増えている．これをうけ，電子文書に含まれる情報の取捨選択を効率的にするための様々な手法が研究されている．吉谷ら [1] は，電子文書から人物のプロフィールに関する情報を固有表現抽出と情報統合の手法を利用して抽出した．平尾ら [2] は，複数文書を対象に，各文書に共通する単語やタイトルに出現する固有表現を素性として SVM による文のランキングを行い，重要文を抽出する複数文書要約を行った．これらの研究では対象となる文書で重要と思われる情報を抽出しているが，人によって必要とする情報は異なる．また，抽出されなかった情報の中にも有用な情報が埋もれている可能性がある．

そこで，抽出する情報を限定することなく，複数の文書に含まれる情報を種類別に表に整理する手法 [3] を過去に提案した．ここで複数の文書とは，同じ種類の文書を集めたものである．例えば異なる機種のスマートフォンの新製品記事に含まれる「メーカー」や「価格」などの情報を種類別に表に整理する．このように同種の複数の文書の情報を種類別に表に整理することは，情報の取捨選択に加え，文書間で情報を比較する際にも役立つ．この研究では，同種の複数の文書に含まれる情報を文単位で抽出し，情報を X-means 法 [4][5] というクラスタリング手法によって分類した結果を，行を文書，列をクラスタとする表に整理した．本稿ではこの手法を従来手法と呼ぶ．X-means 法とは $K = 2$ での K-means 法によるクラスタリングを繰り返しながら，BIC という指標を基に最適なクラスタ数を自動で推定するクラスタリング手法であり，人手でクラスタ数 (分類先の個数) を指定する必要がない．しかし，X-means 法によって推定されたクラスタ数は最適なクラスタ数に比べ小さい傾向にあり，この結果を整理した表は情報が 1 つの列にまとまりすぎており，表の精度が低いという問題があった．

そこで，本研究ではこの問題を改善するために，表の埋まり具合と情報の密集度のバランスを最適にする方法でクラスタ数を推定し，この結果を表に整理する手法を提案する．提案手法では，まず，情報を階層クラスタリングでクラスタリングする．次に，階層クラスタリングのクラスタ数が 1〜n までの結果をそれぞれ表に整理する．そして，クラスタ数 (列数) が 1〜n での各表について，表の埋まり具合と，整理された情報の密集度を求める．この二つの指標のバランスが最適になるときのクラスタ数を最適なクラスタ数と推定する．最後に推定された最適なクラスタ数での結果を表に整理する．本研究では以上の手法により，表の精度の向上を試みた．15 種類の複数文書を用いた実験の結果，従来手法では表の評価指標である F 値 の平均が 0.43 だったが，提案手法では 0.65

と向上し，提案手法の有効性が確認できた．

目 次

図 目 次

表 目 次

第1章　はじめに

近年，ウェブ上に蓄積された膨大な電子文書から情報を得る機会が増えている．これをうけ，電子文書に含まれる情報の取捨選択を効率的にするための様々な手法が研究されている．吉谷ら [1] は，電子文書から人物のプロフィールに関する情報を固有表現抽出と情報統合の手法を利用して抽出した．平尾ら [2] は，複数文書を対象に，各文書に共通する単語やタイトルに出現する固有表現を素性として SVM による文のランキングを行い，重要文を抽出する複数文書要約を行った．これらの研究では対象となる文書で重要と思われる情報を抽出しているが，人によって必要とする情報は異なる．また，抽出されなかった情報の中にも有用な情報が埋もれている可能性がある．

そこで，抽出する情報を限定することなく，文書に含まれる情報を種類別に表に整理する手法 [3] を過去に提案した．ここで複数の文書とは，同じ種類の文書を集めたものである．例えば異なる機種のスマートフォンの新製品記事に含まれる「メーカー」や「価格」などの情報を種類別に表に整理する．このように同種の複数の文書の情報を種類別に表に整理することは，情報の取捨選択に加え，文書間で情報を比較する際にも役立つ．この研究では，同種の複数の文書に含まれる情報を文単位で抽出し，情報を X-means 法 [4][5] というクラスタリング手法によって分類した結果を行を文書，列をクラスタとする表に整理した．本稿ではこの手法を従来手法と呼ぶ．X-means 法とは $K = 2$ での K-means 法によるクラスタリングを繰り返しながら，BIC という指標を基に最適なクラスタ数を自動で推定するクラスタリング手法であり，人手でクラスタ数 (分類先の個数) を指定する必要がない．しかし，X-means 法によって推定されたクラスタ数は最適なクラスタ数に比べ小さい傾向にあり，この結果を整理した表は情報が 1 つの列にまとまりすぎており，表の精度が低いという問題があった．

そこで，本研究ではこの問題を改善するために，表の埋まり具合と情報の密集度のバランスを最適にする方法でクラスタ数を推定し，この結果を表に整理する手法を提案する．提案手法では，まず，情報を階層クラスタリングでクラスタリングする．次に，階層クラスタリングのクラスタ数が 1〜n までの結果をそれぞれ表に整理する．そして，クラスタ数 (列数) が 1〜n での各表について，表の埋まり具合と，整理された情報の密集

度を求める．この二つの指標のバランスが最適になるときのクラスタ数を最適なクラスタ数と推定する．最後に推定された最適なクラスタ数での結果を表に整理する．本研究では以上の手法により，表の精度の向上を試みる．

第2章　従来手法

過去に行った研究 [3] では，同種の複数の文書に含まれる情報を文単位で抽出し，情報を X-means 法 [4][5] というクラスタリング手法によって分類した結果を表に整理した．

2.1　表に整理する手順

文の情報を表に整理する手順を以下に示す．また，手順の例を図 2.1 に示す．

手順 1 複数文書に含まれる文を句点区切りで抽出する．

手順 2 文のベクトルを計算する．

手順 3 文ベクトルを基に文を X-means 法でクラスタリングする．

手順 4 クラスタリングの結果を，行を文書，列をクラスタとする表に整理する．ここで，クラスタリングの際に，情報がどの文書に含まれていたかは考慮されないため，1 つのセルに複数の情報が含まれる場合がある．

手順 5 表の各列について，項目名を付与する．

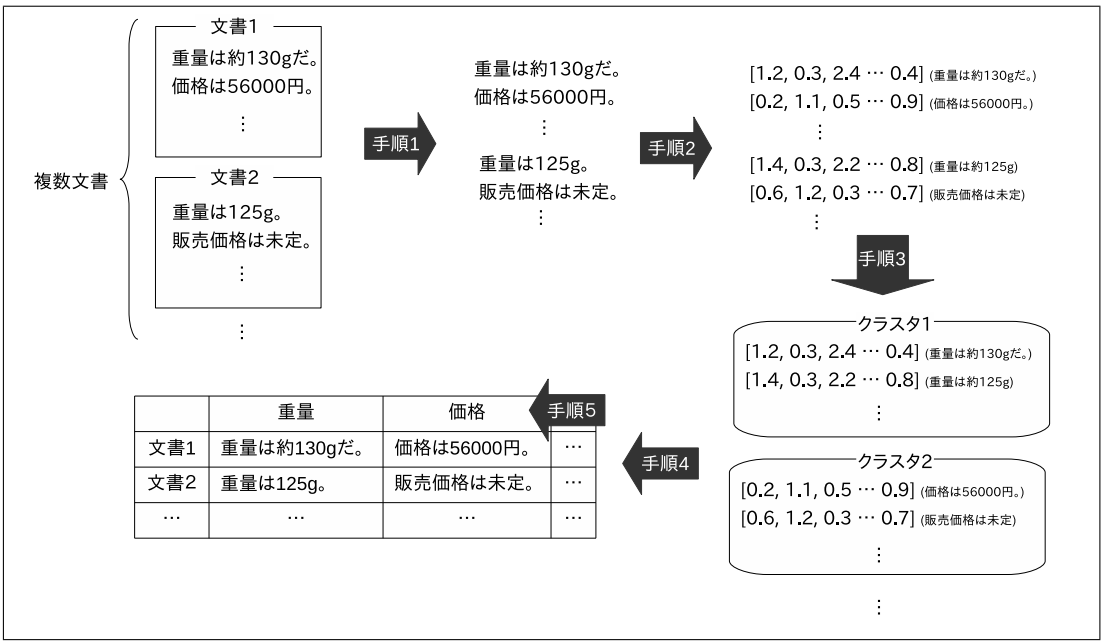

図 2.1: 表に整理する手順の例

2.2　手順 2 : 文のベクトルの計算方法

2.1 節の手順 2 における文のベクトルは以下の手順で求める．

1. 文を MeCab[1]を用いて形態素解析する．
2. 形態素解析結果のうち，品詞が名詞で，かつ，品詞分類 1 が代名詞，数，非自立，副詞可能でない単語を抽出する．
3. 抽出した単語のベクトルの総和を文のベクトルとする．ここで単語のベクトルは FastText[6] に Wikipedia の全記事を学習させて求める．

2.3　手順 3 : 文のクラスタリング

2.1 節の手順 3 では，X-means 法 [4][5] というクラスタリング手法を用いて文をクラスタリングする．X-means 法は，図 2.2 のように K=2 での K-means 法による分割を繰り返し，ベイズ情報量 BIC によって分割を停止するかを判定することで，クラスタ数を自動で決定するクラスタリング手法である．分割前のベイズ情報量 BIC，分割後のベイズ情報量 BIC' に対し，$BIC \leq BIC'$ ならば分割を停止する．p 変量正規分布を

$$f(\theta_i; x) = (2\pi)^{-p/2}|V_i|^{-1/2} \exp\left[\frac{1}{2}(x-\mu_i)^t V_i^{-1}(x-\mu_i)\right]$$

と仮定すると，BIC は以下のように定義される．ここで $\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$ は，p 変量正規分布の最尤推定値とし，μ_i は p 次の平均値ベクトル，V_i は $p \times p$ の分散共分散行列である．q はパラメータ空間の次元数で，V_i の共分散を無視すれば $q = 2p$ であり，無視しなければ $q = p(p+3)/2$ である．L は尤度関数で $L(= \Pi f()$ である．

$$BIC = -2\log(L\hat{\theta}_i; x_i \in C_i) + q \log n_i$$

また，分割後の BIC' は以下のように定義される．ここで $\hat{\theta}'_i = [\hat{\theta}^1_i, \hat{\theta}^2_i]$ は，分割後の 2 つの各クラスタにおける p 変量正規分布の最尤推定値とする．共分散を無視すると，各 p に対し平均と分散の 2 つのパラメータが存在するので，$q' = 2 \times 2p = 4p$ であり，無視しなければ $q' = 2q = p(p+3)$ である．

$$BIC' = -2\log(L\hat{\theta}'_i; x_i \in C_i) + q' \log n_i$$

[1] http://taku910.github.io/mecab/

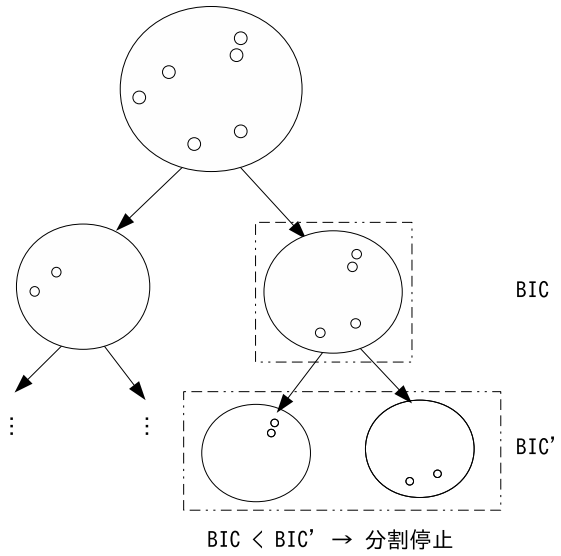

図 2.2: X-means 法のイメージ

2.4　手順 4：クラスタリング結果を表に整理

2.1 節の手順 3 のクラスタリング結果を，行を文書，列をクラスタとする表に整理する．ここで，クラスタリングの際に，情報がどの文書に含まれていたかは考慮されないため，1 つのセルに複数の情報が含まれる場合がある．また，表の可読性を高めるために，列を重要度の高い順にソートする．

クラスタリング結果を整理した表の列には，表 2.1 の列 1 のように関連する文だけで構成される重要度の高い列もあれば，列 2 のように関連しない文が混在した重要度の低い列もある．ここでは列の重要度を，密集率と文書カバー率に基づいて定義する．

表 2.1: 列の重要度の例

	列 1(クラスタ 1)	列 2(クラスタ 2)	…
文書 1	重量は約 130g だ。	価格は 56000 円。	…
文書 2	重量は 125g。	薄さは 8mm。	…
文書 3	重量はおよそ 100g と軽量。	画面サイズは 5 インチ。	…
文書 4	重量は 130g である。	価格は未定。	…
文書 5	重量は約 150g である。	1 月より発売予定。	…

まず密集率について，k 番目の列の密集率 d_k を式 2.1 のように定める．ここで，N_k は k 番目の列に含まれる文の総数であり，$S_{k,l}$ は k 番目の列に含まれる l 番目の文のベクトルであり，$S_{k,mean}$ は k 番目の列に含まれる文のベクトルの平均である．

$$d_k = \frac{1}{N_k} \sum_{l=1}^{N} \frac{S_{k,l} \cdot S_{k,mean}}{|S_{k,l}||S_{k,mean}|} \tag{2.1}$$

式 2.1 で求めた列の密集率 d_k を，式 2.2 を用いて，最小値が 0，最大値が 1 になるように正規化する．ここで，nd_k は k 番目の列の正規化された列の密集率であり，K は列の総数である．

$$nd_k = \frac{d_k - d_{min}}{d_{max} - d_{min}} \tag{2.2}$$

$$d_{min} = \min_{1 \le k \le K} d_k \tag{2.3}$$

$$d_{max} = \max_{1 \le k \le K} d_k \tag{2.4}$$

次に文書カバー率について，k 番目の列の文書カバー率 c_k を式 2.5 のように定める．p_k は k 番目の列において文を抽出できた文書の数であり，P は文書の総数である．

$$c_k = \frac{p_k}{P} \tag{2.5}$$

式 2.5 で求めた文書カバー率 c_k を，式 2.6 を用いて，最小値が 0，最大値が 1 になるように正規化する．ここで，nc_k は k 番目の列の正規化された文書カバー率であり，K は列の総数である

$$nc_k = \frac{c_k - c_{min}}{c_{max} - c_{min}} \tag{2.6}$$

$$c_{min} = \min_{1 \leq k \leq K} c_k \tag{2.7}$$

$$c_{max} = \max_{1 \leq k \leq K} c_k \tag{2.8}$$

k 番目の列の重要度 i_k を式 2.9 のように定義する．

$$i_k = nd_k \times \ nc_k \tag{2.9}$$

2.5 手順 5 : 列の項目名の付与

2.1 節の手順 5 では，各列に対し以下の手順で項目名を付与する．

1. 列に含まれる各文について，文に含まれる品詞が名詞の単語を抽出する．
2. 1 で抽出した各単語について，文書頻度を求める．
3. 文書頻度が最大の単語をクラスタの項目名として付与する．
4. 文書頻度が最大の単語が複数ある場合は，読点で区切って全て付与する．

2.6 従来手法の問題点

情報を表に整理する場合，表 2.2 の列 1 のように情報が 1 列にまとまりすぎたり，表 2.3 の列 2，列 3 のように情報が細かく分類されすぎることのない，表 2.4 のようなバランスの良い表が望ましいと考えられる．一方で，従来手法では X-means 法によって推定されたクラスタ数 (表の列数) が最適なクラスタ数に比べに小さい傾向にあった．この結果，表 2.2 の列 1 でメモリーの情報とストレージの情報が混在しているような，情報が 1 つの列にまとまりすぎた表が得られることが多く，このことが表の精度が低い原因であった．

表の精度を向上させるためには，表 2.2 や表 2.3 のいずれにも偏ることなく，これらのバランスを最適にするような結果が得られるようにクラスタ数 (列数) を推定する必要がある．

表 2.2: 情報が 1 列にまとまりすぎた例

列 1	列 2
メモリーは 4GB 内蔵ストレージは 64GB	発売日は 10 月上旬
メモリーが 3GB ストレージが 32GB	発売日は 9 月
メモリーが 3GB 内蔵ストレージが 32GB	発売日は 1 月
メモリーが 3GB 内蔵ストレージが 32GB	8 月より発売

表 2.3: 情報が細かく分類されすぎた例

列 1	列 2	列 3	列 4
メモリーは 4GB	内蔵ストレージは 64GB		発売日は 10 月上旬
メモリーは 3GB		ストレージが 32GB	発売日は 9 月
メモリーは 3GB	内蔵ストレージが 32GB		発売日は 1 月
メモリーは 3GB	内蔵ストレージが 32GB		8 月より発売

表 2.4: 最適な表の例

列 1	列 2	列 3
メモリーは 4GB	内蔵ストレージは 64GB	発売日は 10 月上旬
メモリーは 3GB	ストレージが 32GB	発売日は 9 月
メモリーは 3GB	内蔵ストレージが 32GB	発売日は 1 月
メモリーは 3GB	内蔵ストレージが 32GB	8 月より発売

第3章　提案手法

従来手法の問題点を解消するため，本研究では表の埋まり具合と情報の密集度のバランスを最適にする方法でクラスタ数を推定し，この結果を表に整理する手法を提案する．なお，従来手法で用いていた X-means 法では，初期値依存性によって得られる表が実行ごとに異なっていたため，提案手法では初期値依存性のない階層クラスタリングの手法を用いる．

3.1　提案手法の手順

提案手法の手順を以下に示す．また，手順の例を図 3.1 に示す．

手順 1 複数文書に含まれる文を句点区切りで抽出する．

手順 2 文のベクトルを計算する．

手順 3 文ベクトルを基に文を Ward 法による階層クラスタリングでクラスタリングする．

手順 4 階層クラスタリングによって得られた各クラスタ数でのクラスタリング結果を基に表に整理する．ここで，クラスタリングの際に，情報がどの文書に含まれていたかは考慮されないため，1 つのセルに複数の情報が含まれる場合がある．クラスタ数 k での表の埋まり具合を式 (3.1) から，情報の密集度を式 (3.2) からそれぞれ求める．$|c_{k,i}|$ はクラスタ数 k での表の i 番目の列に含まれる文の総数，$d_{k,i,j}$ はクラスタ数 k での表の i 番目の列の j 番目の文のベクトル，C_k はクラスタ数 k での表の列の総数，$cosine(x, y)$ は x，y のコサイン類似度を求める関数を表す．

$$cover_k = \frac{\text{クラスタ数 } k \text{ での表の埋まっているセルの数}}{\text{クラスタ数 } k \text{ での表のセルの総数}} \tag{3.1}$$

$$density_k = \min_{j \neq h}(cosine(d_{k,i,j}, d_{k,i,h})) \quad i = 1, \cdots, C_k \quad j, h = 1, \cdots, |c_k, i| \tag{3.2}$$

ここで，全てのクラスタ数での $cover_k$ の集合を $COVER$，$max(COVER)$ を集合 $COVER$ の最大値，$min(COVER)$ は集合 $COVER$ の最小値とする．各クラスタでの $cover_k$ を式 (3.3) で 0〜1 の範囲に正規化する．

$$norm(cover_k) = \frac{cover_k - min(COVER)}{max(COVER) - min(COVER)} \tag{3.3}$$

同様に，全てのクラスタ数での $density_k$ の集合を $DENSITY$，$max(DENSITY)$ を集合 $DENSITY$ の最大値，$min(DENSITY)$ は集合 $DENSITY$ の最小値とする．各クラスタでの $density_k$ を式 (3.4) で 0〜1 の範囲に正規化する．

$$norm(density_k) = \frac{density_k - min(DENSITY)}{max(DENSITY) - min(DENSITY)} \tag{3.4}$$

クラスタ数 k での表の $Score_k$ を式 (3.5) より求める．$Score_k$ が最大となるときのクラスタ数 k を最適なクラスタ数として採用する．

$$Score_k = norm(cover_k) \times norm(density_k) \tag{3.5}$$

手順 5 手順 4 で採用されたクラスタ数でのクラスタリングの結果を，行を文書，列をクラスタとする表に整理する．

手順 6 表の各列について，項目名を付与する．

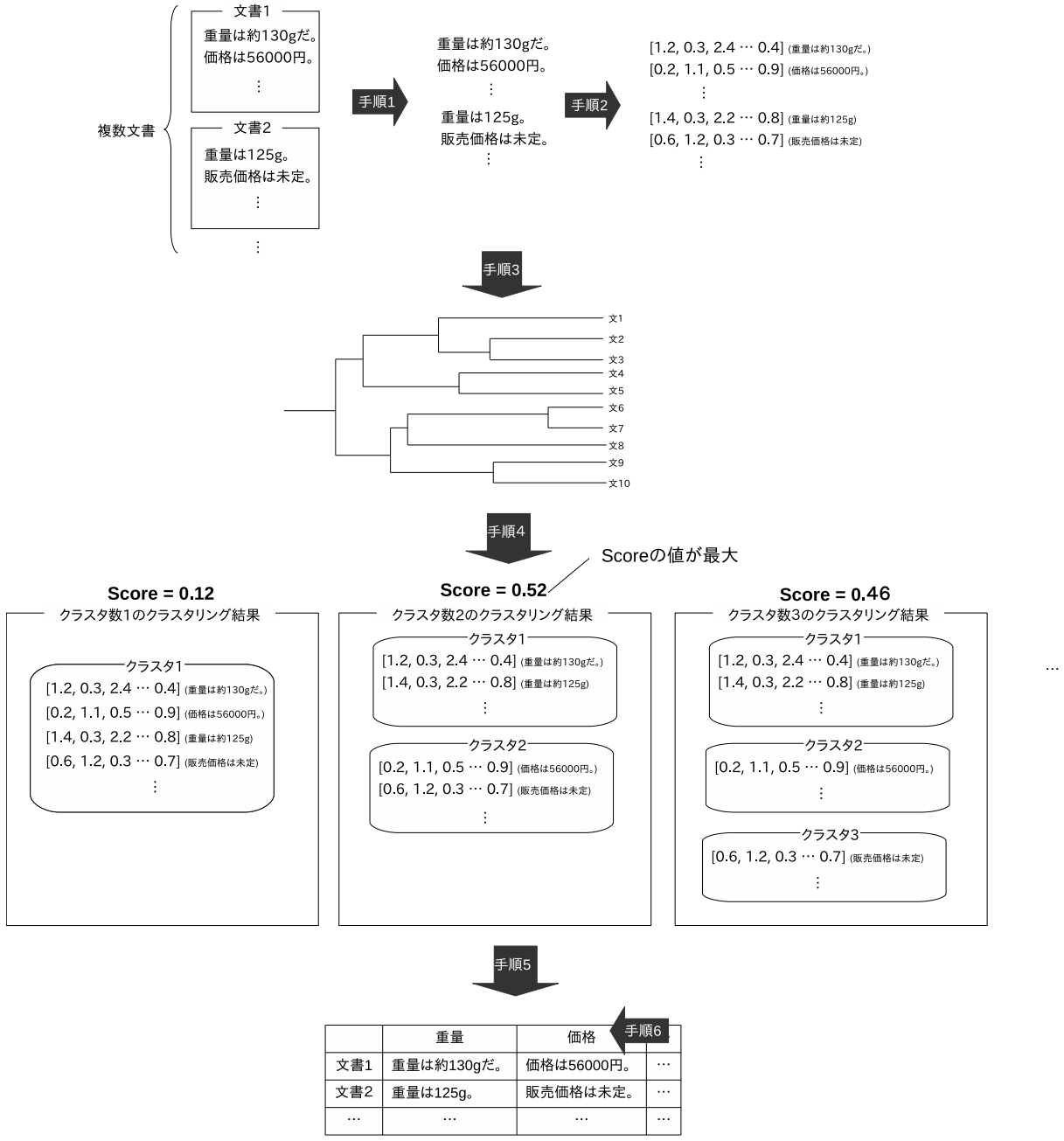

図 3.1: 提案手法の手順の例

3.2 階層クラスタリング

階層クラスタリングは，距離の最も近いクラスタ同士の統合を繰り返すクラスタリング手法である．階層クラスタリングはクラスタ間の距離の定義の違いによっていくつかの手法が存在するが，今回は Ward 法を用いた．Ward 法ではクラスタ C_1 とクラスタ C_2 の距離 $D(C_1, C_2)$ を以下のように定義する．

$$D(C_1, C_2) = E(C_1 \cup C_2) - E(C_1) - E(C_2)$$

$$\mathrm{E}(C_i) = \sum_{\mathbf{x} \in C_i} (d(\mathbf{x}, \mathbf{c}_i))^2$$

$$c_i = \sum_{\mathbf{x} \in C_i} \mathbf{x} / |C_i|$$

第4章　実験

従来手法を用いて生成された表と，提案手法を用いて生成された表を評価し比較する．なお，従来手法で用いている X-means 法は，初期値依存性により実行ごとに異なる表が得られる．今回は表を 1,000 回生成し，全ての表の評価結果のうち，その値が最大値，平均値，最小値となる結果を比較対象として用いた．

4.1　実験環境

4.1.1　FastText の学習環境

文のベクトルの計算で用いる単語のベクトルは，FastText[6] の学習によって求める．FastText は隠れ層と出力層からなる 2 層のニューラルネットワークで，隠れ層が単語の分散表現に相当する．FastText の学習データとして，Wikipedia の全 1,061,375 記事を用いた．なお学習データは図 4.1 のように，日本語は全角，アルファベットと数字は半角に統一し，MeCab で形態素単位に分かち書きしている．FastText の学習で用いたパラメータは，学習モデルを skip-gram，ベクトルの次元数を 300 とした．他のパラメータ値はデフォルト値を用いた．

表 4.1: 学習データの詳細

記事数	行数
1,061,375	22,794,659

<doc id=”5” url=”https://ja.wikipedia.org/wiki?curid=5” title=”アンパサンド” >

アンパサンド

アンパサンド (, &) とは「…と…」を意味する記号である。英語の に相当するラテン語の の合字で、 (et cetera = and so forth) を と記述することがあるのはそのため。Trebuchet MS フォントでは、と表示され ”et” の合字であることが容易にわかる。

その使用は 1 世紀に遡ることができ (1)、5 世紀中葉 (2,3) から現代 (4-6) に至るまでの変遷がわかる。
Z に続くラテン文字アルファベットの 27 字目とされた時期もある。

アンパサンドと同じ役割を果たす文字に「の et」と呼ばれる、数字の「7」に似た記号があった (, U+204A)。この記号は現在もゲール文字で使われている。

記号名の「アンパサンド」は、ラテン語まじりの英語「& はそれ自身 ”and” を表す」(& per se and) のくずれた形である。英語以外の言語での名称は多様である。

日常的な手書きの場合、欧米でアンパサンドは「 ε 」に縦線を引く単純化されたものが使われることがある。

また同様に、「t」または「+（プラス）」に輪を重ねたような、無声歯茎側面摩擦音を示す発音記号「」のようなものが使われることもある。

プログラミング言語では、C など多数の言語で AND 演算子として用いられる。以下は C の例。
PHP では、変数宣言記号（$）の直前に記述することで、参照渡しを行うことができる。
</doc>

図 4.1: 学習データの例

4.1.2　実験データ

実験で用いる同種の複数文書として，以下の 15 種類の複数文書を用いる．各複数文書の詳細を表 4.2 に示す.

・ 強盗事件に関する新聞記事 20 件

2016 年度の毎日新聞から見出しに「強盗：」を含む記事をランダムに 20 件抽出したデータ

・ 外為・株式に関する新聞記事 20 件

2016 年度の毎日新聞から見出しに「外為・株式：」を含む記事をランダムに 20 件抽出したデータ

・ 地震に関する新聞記事 20 件

2016 年度の毎日新聞から見出しに「地震」と「震度」を含む記事をランダムに 20 件抽出したデータ

・ 交通事故に関する新聞記事 20 件

2016 年度の毎日新聞から見出しに「交通事故：」を含む記事をランダムに 20 件抽出したデータ

・ リコールに関する新聞記事 20 件

2016 年度の毎日新聞から見出しに「リコール：」を含む記事をランダムに 20 件抽出したデータ

・ スマートフォンに関する新製品記事 20 件

2018 年 1 月 15 日時点での「価格.com」のスマートフォンカテゴリーにおける最新の新製品ニュース記事 20 件を抽出したデータ

・ スマートフォンに関する新製品記事 20 件

2018 年 1 月 15 日時点での「価格.com」の薄型テレビ液晶テレビカテゴリーにおける最新の新製品ニュース記事 20 件を抽出したデータ

・ デジタルカメラに関する新製品記事 20 件

2018 年 1 月 15 日時点での「価格.com」のデジタルカメラカテゴリーにおける最新の新製品ニュース記事 20 件を抽出したデータ

・ロボット掃除機に関する新製品記事 20 件

2018 年 1 月 15 日時点での「価格.com」の掃除機カテゴリーにおけるロボット掃除機に関する最新の新製品ニュース記事 20 件を抽出したデータ

・エアコンに関する新製品記事 20 件

2018 年 1 月 15 日時点での「価格.com」のエアコン・クラーカテゴリーにおける最新の新製品ニュース記事 20 件を抽出したデータ

・城に関する Wikipedia の記事 20 件

2017 年 6 月 1 日時点での Wikiedia のカテゴリー「日本の 100 名城」に含まれる全ページのうち，ランダムに抽出した 20 記事の要約部を抽出したデータ

・恐竜に関する Wikipedia の記事 20 件

2017 年 6 月 1 日時点での Wikiedia のカテゴリー「ジュラ紀の恐竜」に含まれる全ページのうち，ランダムに抽出した 20 記事の要約部を抽出したデータ

・力士に関する Wikipedia の記事 20 件

2017 年 6 月 1 日時点での Wikiedia のカテゴリー「高校相撲部出身の大相撲力士」に含まれる全ページのうち，ランダムに抽出した 20 記事の要約部を抽出したデータ

・山に関する Wikipedia の記事 20 件

2017 年 6 月 1 日時点での Wikiedia のカテゴリー「日本百名山」に含まれる全ページのうち，ランダムに抽出した 20 記事の要約部を抽出したデータ

・野球チームに関する Wikipedia の記事 20 件

2017 年 6 月 1 日時点での Wikiedia のカテゴリー「アメリカ合衆国の野球チーム」に含まれる全ページのうち，ランダムに抽出した 20 記事の要約部を抽出したデータ

表 4.2: 実験で用いる各複数文書の詳細

	文書数	総文数	1 文あたりの平均文字数
新聞記事 (強盗)	20	128	39.3
新聞記事 (外為・株式)	20	124	49.2
新聞記事 (地震)	20	91	37.2
新聞記事 (交通事故)	20	143	41.7
新聞記事 (リコール)	20	89	56.7
新製品記事 (スマホ)	20	313	46.3
新製品記事 (テレビ)	20	273	49.0
新製品記事 (カメラ)	20	340	52.0
新製品記事 (ロボット掃除機)	20	235	47.7
新製品記事 (エアコン)	20	255	62.3
Wikipedia(城)	20	94	31.2
Wikipedia(恐竜)	20	77	49.9
Wikipedia(力士)	20	103	28.7
Wikipedia(山)	20	76	31.5
Wikipedia(野球チーム)	20	68	46.9

4.2 評価方法

生成された表の精度を以下の手順で評価する．列の F 値 の計算例を図 4.2 に示す．

1. あらかじめ作成した正解の表の各列に注目する．
2. 注目している列に含まれるデータを最も多く含む実験の表の列を抽出する．
3. 式 4.1，式 4.2，式 4.3 から適合率，再現率，F 値を求める．
4. 2，3 をすべての正解の列に対して行い，各列の F 値の平均を求め，これを実験の表の評価結果とする．

$$\text{適合率} = \frac{\text{正解の表の列と実験の表の列に共通して含まれる文の数}}{\text{実験の表の列に含まれる文の数}} \tag{4.1}$$

$$\text{再現率} = \frac{\text{正解の表の列と実験の表の列に共通して含まれる文の数}}{\text{正解の表の列に含まれる文の数}} \tag{4.2}$$

$$F\text{値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} \tag{4.3}$$

・正解の表の5列目に注目しているとする.

・正解の表の5列目に含まれる情報を最も多く含む実験の表の列は4列目とする.

	列5(正解の表)
文書1	
文書2	・得意技は右四つ、寄り
文書3	・得意技は押し
文書4	
文書5	・得意手は突き・押し
文書6	・得意技は左四つ、寄り
文書7	・得意手は右四つ、突っ張り、叩き込み
文書8	・得意手は突っ張り、押し
文書9	・得意手は押し、左四つ、寄り
文書10	・得意手は突っ張り、右四つ、上手投げ
文書11	
文書12	・得意手は右四つ、寄り、上手捻り、首投げ
文書13	
文書14	・得意は押し
文書15	・得意手は突き、押し、叩き、引き
文書16	・得意手は左四つ、上手出し投げ、右四つ、寄り
文書17	
文書18	
文書19	
文書20	

	列4(実験の表)
文書1	
文書2	・得意技は右四つ、寄り
文書3	
文書4	
文書5	
文書6	・得意技は左四つ、寄り
文書7	・得意手は右四つ、突っ張り、叩き込み
文書8	
文書9	・得意手は押し、左四つ、寄り
文書10	・得意手は突っ張り、右四つ、上手投げ
文書11	
文書12	・得意手は右四つ、寄り、上手捻り、首投げ
文書13	
文書14	
文書15	
文書16	・得意手は左四つ、上手出し投げ、右四つ、寄り
文書17	
文書18	
文書19	
文書20	

$$\text{適合率} = \frac{\text{正解の表の列と実験の表の列に共通して含まれる文の数}}{\text{実験の表の列に含まれる文の数}} = \frac{7}{7}$$

$$\text{再現率} = \frac{\text{正解の表の列と実験の表の列に共通して含まれる文の数}}{\text{正解の表の列に含まれる文の数}} = \frac{7}{12}$$

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}} = \frac{2 \times 7/7 \times 7/12}{7/7 + 7/12} \fallingdotseq 0.74$$

図 4.2: 列の F 値 の計算例

4.3 実験結果

提案手法を用いて得られた表のうち，評価結果の F 値 が最も良かった表の一部を表 4.3，表 4.4 に示す．最も悪かった表の一部を表 4.5，表 4.6，表 4.7 に示す．表の「・」に続く文が 1 つの情報である．また，クラスタリングの際に，情報がどの文書に含まれていたかは考慮されないため，1 つのセルに複数の情報が含まれる場合がある．

表 4.3: 評価結果が最も良かった表：Wikipedia(力士)

	1 列目 (身長)	2 列目 (最高)	3 列目 (出身)
文書 1	・身長 183 ｃｍ、体重 191 ｋｇ	・最高位は東十両 12 枚目	・政風基嗣は、長崎県長崎市出身で尾車部屋所属の現役大相撲力士
文書 2	・身長 175 ｃｍ、体重 130 ｋｇ、血液型はＯ型	・最高位は西幕下 2 枚目	・栃の山博士は、東京都立川市出身で、千賀ノ浦部屋の元大相撲力士で、現世話人
文書 3	・現役時代の体格は身長 179 ｃｍ、体重 149 ｋｇ、血液型はＡＢ型	・最高位は東幕下 4 枚目	・栃乃里隆光は、石川県石川郡野々市町出身で春日野部屋に所属していた元大相撲力士
文書 4	・身長は 180 ｃｍ、体重は 160 ｋｇ	・最高位は西前頭 6 枚目	・誉富士歓之は、青森県西津軽郡鰺ヶ沢町出身で伊勢ヶ濱部屋所属の現役大相撲力士
文書 5	・身長 192 ｃｍ、体重 120 ｋｇ ・大相撲力士時代は身長 192 ｃｍ、体重 120 ｋｇ		・田上明は、日本の実業家、元プロレスラー、元大相撲力士
文書 6	・本名は、三好正人、身長 178 ｃｍ、体重 196 ｋｇ、血液型Ａ型	・最高位は西幕下 2 枚目	・朝陽丸勝人は、大阪府枚方市出身で高砂部屋所属の元大相撲力士
文書 7	・現役時代の体格は 182 ｃｍ、92 ｋｇ	・最高位は西前頭 11 枚目	・吉井山朋一郎は、福岡県田川郡糸田町出身で、出羽海部屋に所属した大相撲力士
文書 8	・本名は小塚一、身長 186 ｃｍ、体重 147 ｋｇ	・最高位は西前頭 2 枚目、血液型Ｏ型	・朝乃翔嚆矢は、神奈川県小田原市出身で若松部屋所属の元大相撲力士
文書 9	・身長 185 ｃｍ、体重 148 ｋｇ	・最高位は西十両 2 枚目	・大岳宗正は滋賀県草津市出身の元大相撲力士
文書 10	・全盛期の体格は 187 ｃｍ、144 ｋｇ	・最高位は東小結	・大翔鳳昌巳は、北海道札幌市豊平区平岸出身で立浪部屋所属の元大相撲力士
文書 11	・現役時代の体格は身長 180 ｃｍ、体重 129 ｋｇ、血液型はＯ型	・最高位は西幕下 16 枚目	・玉大輝剛志は、石川県鳳珠郡能登町出身で片男波部屋に所属していた元大相撲力士
文書 12	・現役時代の体格は 179 ｃｍ、116 ｋｇ	・最高位は西関脇	・開隆山勘之亟は、秋田県南秋田郡昭和町出身で、1960 年代に活躍した大相撲力士
文書 13	・身長 182 ｃｍ、体重 146 ｋｇ	・最高位は東十両 12 枚目	・大翔鵬清洋は、モンゴル・ウランバートル市出身で、追手風部屋所属の大相撲力士
文書 14	・身長 182 ｃｍ、体重 183 ｋｇ、血液型はＡ型、星座は蟹座	・最高位は西前頭 7 枚目	・木村山守は、和歌山県御坊市出身で春日野部屋所属だった元大相撲力士
文書 15	・現役時代の体格は 183 ｃｍ、150 ｋｇ	・最高位は西前頭筆頭	・薩洲洋康貴は、鹿児島県指宿市出身で、1980 年代に活躍した大相撲力士
文書 16	・現役時代の体格は 178 ｃｍ、115 ｋｇ	・最高位は東関脇	・栃東知頼は、福島県相馬郡日立木村出身の元大相撲力士
文書 17	・現役時代の体格は 175 ｃｍ、117 ｋｇ	・最高位は東小結	・智ノ花伸哉は、熊本県八代市出身で立浪部屋に所属した大相撲力士
文書 18	・身長 182 ｃｍ、体重 126 ｋｇ、血液型はＯ型	・最高位は東三段目 51 枚目	・加賀ノ花麻衣は、石川県加賀市出身で千賀ノ浦部屋に所属していた元大相撲力士
文書 19		・最高位は西幕下 2 枚目	・琴藤沢和穂は、高知県高知市出身で佐渡ヶ嶽部屋に所属した元大相撲力士
文書 20			・逆鉾昭廣は鹿児島県姶良市出身で井筒部屋所属の元大相撲力士 ・なお、実際の出身地は東京都墨田区である

表 4.4: 評価結果が最も良かった表：Wikipedia(力士)(続き)

	4 列目 (四つ)	5 列目 (得意)
文書 1	・得意技は右四つ、寄り	
文書 2		・得意技は押し
文書 3		
文書 4		・得意手は突き・押し
文書 5		
文書 6	・得意技は左四つ、寄り	
文書 7	・得意手は右四つ、突っ張り、叩き込み	
文書 8		・得意手は突っ張り、押し
文書 9	・得意手は押し、左四つ、寄り	
文書 10	・得意手は突っ張り、右四つ、上手投げ	
文書 11		
文書 12	・得意手は右四つ、寄り、上手捻り、首投げ	
文書 13		
文書 14		・得意は押し
文書 15		・得意手は突き、押し、叩き、引き
文書 16	・得意手は左四つ、上手出し投げ、右四つ、寄り	
文書 17		
文書 18		
文書 19		
文書 20		

表 4.5: 評価結果が最も悪かった表：新聞記事 (交通事故)

	1 列目 (交差点)
文書 1	・29 日午後 1 時 5 分ごろ、愛知県北名古屋市鍛治ケ一色西 2 の県道と市道の交差点で、乗用車と軽乗用車が出合い頭に衝突した
文書 2	・8 日午前 8 時ごろ、埼玉県上里町嘉美の町道で保育園の園児を送迎するバスが軽乗用車と衝突し、横転した
文書 3	・28 日午前 8 時ごろ、横浜市港南区大久保 1 の市道で車 3 台が絡む事故があり、はずみで軽トラックが横転し、集団登校中の小学生 9 人を巻き込んだ ・県警によると、死亡したのは近くに住む市立桜岡小 1 年、田代優さん
文書 4	・27 日午前 7 時 45 分ごろ、兵庫県加古川市西神吉町中西の交差点で、軽乗用車と衝突したタクシーが弾みで登校中の小学生の列に突っ込んだ
文書 5	・県警高速隊は、下り線を逆走していた同県さぬき市の男性の軽乗用車を発見 ・県警高速隊によると、さぬき市の高松道上り線で同日午後 9 時半ごろ、乗用車の女性＝徳島県松茂町＝が逆走している車を発見
文書 6	・9 日午後 3 時 40 分ごろ、広島県庄原市東城町の中国自動車道下り線で、バレーボール全日本男子の次期監督に内定している中垣内祐一さん＝大阪市平野区＝運転の乗用車が、工事規制中の警備員の男性をはねた
文書 7	・1 日午前 0 時 50 分ごろ、栃木市都賀町家中の北関東自動車道下り線・栃木都賀ジャンクション―都賀インターチェンジ間で、走行車線を走っていた乗用車が愛知県稲沢市の男性運転のトラックに追突した
文書 8	・2 日午前 2 時 10 分ごろ、北海道室蘭市東町 5 の国道 36 号交差点で、乗用車が道路脇の信号機の支柱に衝突して大破していると 110 番があった ・道警室蘭署によると、死亡したのは登別市新生町 1 の会社員、長尾卓弥さんと東京都江東区東陽 5 の会社員、平石隆祥さん、室蘭市小橋内町 2、自営業、山下知弥さん ・3 人は室蘭市出身で、小中学校時代の同級生 ・平石さんは市内の実家に帰省中だった
文書 9	・26 日午前 5 時 45 分ごろ、大阪府寝屋川市池田北町の国道 1 号交差点で、横断歩道を自転車で通行していた男性が左折中の大型トラックにひかれて死亡した ・南容疑者は京都府京田辺市内から寝屋川市内の営業所へ、外壁用のブロックを運んでいた
文書 10	・20 日午後 7 時ごろ、東京都大田区蒲田本町 1 の環状 8 号線で、観光バスが中央分離帯にある信号機の柱に衝突した ・バスを運行したのは埼玉県久喜市の「夢湖観光バス」 ・20 日朝にＪＲ蒲田駅を出発し、河口湖などを巡った後、蒲田駅へ戻る途中だった
文書 11	・26 日午前 6 時 40 分ごろ、大阪市旭区中宮 1 の市道交差点で、横断歩道を歩いていた 80 代くらいの女性が車にはねられた
文書 12	・4 日午後 9 時半ごろ、大阪市住吉区万代東 3 の府道で、あべの橋発遠里小野橋行きの大阪市営バスが道路脇の電柱などに接触した
文書 13	・8 日午後 9 時 55 分ごろ、香川県観音寺市柞田町の国道 11 号で、大型トレーラーが、地元の祭りで引いていた太鼓台に後ろから突っ込んだ ・現場はＪＲ観音寺駅から南東に約 2・4 キロ
文書 14	・2 日午前 2 時 5 分ごろ、愛知県岡崎市駒立町の新東名高速道路上り線で、故障のため路側帯に停車していた観光バスに大型トラックが追突した ・死亡した運転手は、大阪府太子町葉室、大谷秀雄さんと、大阪市城東区中浜 1 の染谷文彦さん ・バスは運転手 2 人と乗客 27 人が乗り、1 日夜に大阪市内を出発し、千葉県浦安市の東京ディズニーシーに向かっていた ・現場は岡崎サービスエリアから東京方面へ約 3 キロ
文書 15	・12 日午後 5 時ごろ、兵庫県宝塚市小浜 2 の国道 176 号で、いずれも 18 歳の男女 4 人が乗った乗用車が中央分離帯のガードレールに衝突、出火した
文書 16	・16 日午後 3 時半ごろ、奈良県川上村大迫の国道 169 号大迫トンネルで、ワゴン車と乗用車が正面衝突し、火災が起きた
文書 17	・8 日午前 2 時 45 分ごろ、兵庫県西宮市浜脇町の阪神高速神戸線下りで、中型トラックが大型トレーラーに追突し、トラックを運転していた同県南あわじ市の会社員、殿本亘幸さんが死亡した
文書 18	・8 日午前 7 時 55 分ごろ、静岡県磐田市中泉の県道交差点で、登校中に横断歩道を渡っていた市立磐田中部小学校 2 年の大石萌衣さん＝同所＝と、同級生の男子児童の 2 人がライトバンにはねられた
文書 19	・10 日午前 8 時 45 分ごろ、大阪府島本町山崎の名神高速上り線左ルートの天王山トンネル内で、路線バスや大型トラックなど計 5 台が絡む多重衝突事故があった
文書 20	・26 日午前 9 時半ごろ、大津市蛍谷の名神高速道路下り線で、高速バスが前のトラックに追突 ・バスは、ＪＲ福井駅発大阪・梅田行きで、京福バスが運行

表 4.6: 評価結果が最も悪かった表：新聞記事 (交通事故)(続き)

	2 列目 (片側)	3 列目 (自動車運転処罰法)
文書 1		・県警西枇杷島署は自動車運転処罰法違反の疑いで、乗用車の同市徳重東出、パート、大口久美子容疑者を現行犯逮捕した ・同法違反の過失致死傷容疑に切り替えて調べる
文書 2		
文書 3		・同署は軽トラックの運転手に自動車運転処罰法違反の疑いもあるとみて、詳しく事情を聴く
文書 4		
文書 5		・同県警によると、男性は免許はあるが、普段は運転していないという
文書 6		
文書 7		
文書 8	・現場は片側 2 車線のカーブ	
文書 9		・府警寝屋川署は、トラックを運転した京都市伏見区淀池上町、会社員、南隆樹容疑者を自動車運転処罰法違反の疑いで現行犯逮捕した ・「人通りが少なく油断した」と容疑を認めている
文書 10	・現場は片側 2 車線のほぼ直線の道路	・同署はバスの運転手、菅原正容疑者＝東京都足立区＝を自動車運転処罰法違反容疑で現行犯逮捕した
文書 11		
文書 12	・大阪府警住吉署などによると、現場は片側 2 車線の直線道路	
文書 13	・同署によると、現場は見通しの良い片側 1 車線の直線道路	・香川県警観音寺署は、トレーラーを運転していた愛媛県大洲市、大川貴之容疑者を自動車運転処罰法違反容疑で現行犯逮捕した ・容疑を認めているという
文書 14	・片側 2 車線で見通しは良いという	・県警高速隊は、トラックを運転していた福岡市博多区西春町 1 の会社員、斎藤信夫容疑者を自動車運転処罰法違反容疑で逮捕した
文書 15	・現場は片側 2 車線の直線道路	
文書 16		
文書 17		
文書 18		・県警磐田署は、ライトバンを運転していた浜松市南区金折町の会社員、河合秀幸容疑者を自動車運転処罰法違反で現行犯逮捕した
文書 19		
文書 20	・高速隊によると、現場は片側 2 車線の直線	

表 4.7: 評価結果が最も悪かった表：新聞記事 (交通事故)(続き)

	4 列目 (軽傷)	5 列目 (交差点)
文書 1		
文書 2		・現場は、見通しのよい十字路で、信号機はなかった
文書 3	・軽トラックを運転していた同市磯子区の男性と軽乗用車を運転していた女性、同乗していた 30 代男性も軽傷	
文書 4	・タクシーに乗っていた同県姫路市内の男子中学生 2 人と軽乗用車を運転していた女性も軽傷を負った	・現場は信号機のある県道と市道の交差点 ・同署によると、東進していた軽乗用車と北進中のタクシーが出合い頭に衝突、タクシーが交差点の北西角付近にいた男児らをはねたという ・同署が当時の信号の状況などを調べている
文書 5	・女性が両足を打撲、トラックの男性＝香川県東かがわ市＝にけがはなかった	
文書 6		
文書 7	・栃木県警高速隊によると、トラックの男性は軽傷	
文書 8		
文書 9		
文書 10	・警視庁蒲田署によると、乗客の 20〜70 代の男女 24 人がけがをしたが、いずれも軽傷とみられる ・同署などによると、バスには乗客 28 人のほか、運転手と添乗員 1 人ずつが乗っていた	・赤信号で停車していたが、青に変わったので発車したら衝突した」と供述しているという
文書 11		・車は交差点を北から東に左折した際に女性をはね、そのまま逃走したという
文書 12	・男性運転手が運転中に意識がもうろうとなり、病院に救急搬送された ・運転手は「意識がもうろうとした」と話しているといい、同署が事故原因を調べている ・市交通局によると、バスに搭載されているドライブレコーダーに、運転手の映像が記録されていた ・直後に運転手は前のめりになり、ハンドルにもたれかかった ・運転手は、病院へ搬送される途中で意識を取り戻した ・運転手には持病がなかった	・現場近くの交差点で信号待ちのため停車し、信号が青に変わると発車
文書 13		
文書 14	・故障に対処するため車外にいたバスの男性運転手 2 人が、バスと側壁に挟まれ、出血性ショックのため間もなく死亡した ・バス内にいた乗客の大阪市の女子大学生と、大型トラックの男性運転手の計 2 人も切り傷など軽傷を負った	
文書 15	・兵庫県警宝塚署によると、後部座席には 2 人乗っており、女子高校生が意識不明の重体、会社員の男性が鎖骨を折って重傷	
文書 16	・乗用車に乗っていた 3 人のうち、西東京市の無職、藤井純代さんが頸椎損傷で死亡し、運転していた夫が右足骨折、親族の女性＝奈良県桜井市＝が頭に打撲のけがをした	
文書 17		
文書 18		
文書 19	・この事故で大型トラックの 30 代の男性運転手が左足に軽傷を負った	
文書 20	・計 4 台が絡む玉突き事故になり、バスの運転手や乗客ら計 8 人が病院に搬送された ・滋賀県警高速隊などによると、バスの男性運転手が骨折している可能性があるが、7 人は軽傷という	

4.4 評価結果

従来手法と提案手法によって生成された表の評価結果を表 4.8 に示す．15 種類の複数文書を用いた実験では，提案手法の評価結果の平均が 0.65 であり，従来手法の最大値，平均値を上回る結果となった．

表 4.8: 各表の評価結果

	従来手法			提案手法
	最大値	平均値	最小値	
新聞記事 (強盗)	0.74	0.58	0.30	0.66
新聞記事 (外為・株式)	0.54	0.37	0.13	0.63
新聞記事 (地震)	0.68	0.49	0.19	0.72
新聞記事 (交通事故)	0.55	0.39	0.23	0.51
新聞記事 (リコール)	0.59	0.41	0.18	0.65
新製品記事 (スマホ)	0.70	0.58	0.28	0.78
新製品記事 (テレビ)	0.51	0.35	0.13	0.64
新製品記事 (カメラ)	0.46	0.31	0.06	0.58
新製品記事 (ロボット掃除機)	0.47	0.35	0.11	0.60
新製品記事 (エアコン)	0.52	0.40	0.16	0.51
Wikipedia(城)	0.82	0.52	0.19	0.76
Wikipedia(恐竜)	0.60	0.42	0.23	0.78
Wikipedia(力士)	0.73	0.52	0.22	0.84
Wikipedia(山)	0.55	0.35	0.21	0.65
Wikipedia(野球チーム)	0.56	0.40	0.20	0.51
平均	0.60	0.43	0.19	0.65

また，評価結果の差が有意かを調べるために，全ての評価結果の組み合わせで，対応のある両側 t 検定を行った．有意水準は 0.05 とした．有意差検定の結果を表 4.9 に示す．表 4.9 より，従来手法と提案手法の評価結果の間に有意差があるという結果が得られた．

表 4.9: 有意差検定の結果

	最大値	平均値	最小値	提案手法
最大値				
平均値	**0.000**			
最小値	**0.000**	**0.000**		
提案手法	**0.024**	**0.000**	**0.000**	

第5章　考察

5.1　従来手法との比較

15 種類の複数文書を用いた実験では，提案手法の評価結果の平均が 0.65 と従来手法の最大値，平均値を上回る結果となった．また，有意水準を 0.05 として対応のある両側 t 検定を行った結果，従来手法と提案手法の評価結果の間に有意差があるという結果が得られた．提案手法では表の埋まり具合と情報の密集度のバランスを最適にすることで，表の整理により適したクラスタ数 (表の列数) の推定を行うことができ，この結果，表の精度を向上させることができたと思われる．

実際に，従来手法と提案手法でそれぞれ推定されたクラスタ数 (表の列数) と，正解の表の列数を比較した結果，表 5.1 のようになった．また，従来手法と提案手法でそれぞれ推定されたクラスタ数 (表の列数) と，正解の表の列数の差の絶対値を比較した結果は，表 5.2 のようになった．結果から，従来手法では, 正解の表の列数との差が平均で 6.7 あったが，提案手法ではこの差が 3.8 に縮まった．よって，従来手法において，X-means 法により推定されたクラスタ数 (表の列数) が小さい傾向にあった問題は提案手法では改善され，より正解の表の列数に近づいたことが分かった．

また，今回，表の精度を表の各列の F 値 の平均として求めたが，このときの各列の適合率の平均と再現率の平均を求めたところ，表 5.3 のような結果が得られた．表 5.3 より，従来手法の平均値での結果では，再現率が平均で 0.79 であるのに対し，適合率が平均で 0.39 と低い．これは，従来手法で得られる表の列の特徴が，情報の取りこぼしが少ない一方で関連しない情報を含みやすいことを示している．一方で，提案手法では，再現率が平均で 0.68 であるのに対し，適合率が平均で 0.75 であり，二つの指標の隔たりが小さいため，バランスの良い表が得られていることが分かる．

表 5.1: 生成された表の列数

	従来手法			提案手法	正解の表
	最大値	平均値	最小値		
新聞記事 (強盗)	9	9	5	10	8
新聞記事 (外為・株式)	8	4	2	14	13
新聞記事 (地震)	9	6	2	11	10
新聞記事 (交通事故)	8	5	2	11	10
新聞記事 (リコール)	8	5	2	14	11
新製品記事 (スマホ)	34	26	7	40	24
新製品記事 (テレビ)	25	12	5	24	25
新製品記事 (カメラ)	18	11	2	24	33
新製品記事 (ロボット掃除機)	20	11	3	23	27
新製品記事 (エアコン)	15	12	3	13	16
Wikipedia(城)	11	6	2	16	9
Wikipedia(恐竜)	4	3	2	12	8
Wikipedia(力士)	13	6	5	13	10
Wikipedia(山)	5	3	2	9	9
Wikipedia(野球チーム)	6	8	4	7	9

表 5.2: 最適な表の列数との差

	従来手法-正解の表			提案手法-正解の表
	最大値	平均値	最小値	
新聞記事 (強盗)	1	1	3	2
新聞記事 (外為・株式)	5	9	11	1
新聞記事 (地震)	1	4	8	1
新聞記事 (交通事故)	2	5	8	1
新聞記事 (リコール)	3	6	9	3
新製品記事 (スマホ)	10	2	17	16
新製品記事 (テレビ)	0	13	20	1
新製品記事 (カメラ)	15	22	31	9
新製品記事 (カメラ)	7	16	24	4
新製品記事 (掃除機)	1	4	13	3
新製品記事 (エアコン)	2	3	7	7
Wikipedia(城)	4	5	6	4
Wikipedia(恐竜)	3	4	5	3
Wikipedia(力士)	4	6	7	0
Wikipedia(山)	3	1	5	2
平均	4.1	6.7	11.6	3.8

表 5.3: 表の再現率 (平均) と適合率 (平均)

	従来手法 (平均値)			提案手法		
	適合率の平均	再現率の平均	F 値 の平均 (表の精度)	適合率の平均	再現率の平均	F 値 の平均 (表の精度)
新聞記事 (強盗)	0.57	0.71	0.58	0.67	0.72	0.66
新聞記事 (外為・株式)	0.28	0.89	0.37	0.62	0.72	0.63
新聞記事 (地震)	0.45	0.81	0.49	0.76	0.77	0.72
新聞記事 (交通事故)	0.31	0.76	0.39	0.53	0.71	0.51
新聞記事 (リコール)	0.38	0.74	0.41	0.74	0.64	0.65
新製品記事 (スマホ)	0.60	0.69	0.58	0.87	0.75	0.78
新製品記事 (テレビ)	0.33	0.73	0.35	0.67	0.72	0.64
新製品記事 (カメラ)	0.24	0.71	0.31	0.55	0.77	0.58
新製品記事 (ロボット掃除機)	0.32	0.84	0.35	0.60	0.75	0.60
新製品記事 (エアコン)	0.39	0.68	0.40	0.48	0.71	0.51
Wikipedia(城)	0.50	0.89	0.52	0.80	0.78	0.76
Wikipedia(恐竜)	0.31	0.80	0.42	0.93	0.70	0.78
Wikipedia(力士)	0.50	0.89	0.52	0.90	0.87	0.84
Wikipedia(山)	0.26	0.89	0.35	0.62	0.82	0.65
Wikipedia(野球チーム)	0.44	0.77	0.40	0.52	0.80	0.51
平均	0.39	0.79	0.43	0.68	0.75	0.65

5.2 階層クラスタリングの最適な結果との比較

5.2.1 列数の推定精度の改善についての考察

提案手法によって推定されたクラスタ数と，階層クラスタリングにおける最適なクラスタ数の近さを調査した．具体的には，階層クラスタリングの各クラスタ数でのクラスタリング結果を整理した表を全て評価し，最良の評価結果となるときのクラスタ数を調べ，これと比較した．結果を表 5.4 に示す．

表 5.4: 階層クラスタリングでの最適なクラスタ数と表の評価結果

	最適な結果		提案手法の結果	
	クラスタ数 (最適)	評価結果	クラスタ数	評価結果
新聞記事 (強盗)	7	0.69	10	0.66
新聞記事 (外為・株式)	22	0.68	14	0.63
新聞記事 (地震)	11	0.72	11	0.72
新聞記事 (交通事故)	29	0.63	11	0.51
新聞記事 (リコール)	33	0.67	14	0.65
新製品記事 (スマホ)	25	0.81	40	0.78
新製品記事 (テレビ)	37	0.67	24	0.64
新製品記事 (カメラ)	43	0.66	24	0.58
新製品記事 (ロボット掃除機)	61	0.65	23	0.60
新製品記事 (エアコン)	21	0.63	13	0.51
Wikipedia(城)	10	0.81	16	0.76
Wikipedia(恐竜)	10	0.80	12	0.78
Wikipedia(力士)	12	0.88	13	0.84
Wikipedia(山)	15	0.73	9	0.65
Wikipedia(野球チーム)	17	0.74	7	0.51

結果を見ると，新製品記事 (ロボット掃除機) や新製品記事 (カメラ)，新聞記事 (リコール) での推定されたクラスタ数が最適なクラスタ数に比べ非常に小さい値となっている．これらの複数文書は表 5.5 のように，いずれも正解の表に占める空欄の割合が多い．そのため，提案手法において表の埋まり具合を過度に考慮したことが悪く働き，このように推定されるクラスタ数が大幅に少なくなったと考えられる．この問題を解消するには，表の埋まり具合と表の密集度の重みを文書に応じて調整する必要がある．具体的には複数文書の文書間の類似度を求めるなどして表の空欄の割合を推定し，これを基に表の埋

まり具合と密集度の重みを調整することで，階層クラスタリングにおける最適なクラスタ数により近づくと考えられる．

表 5.5: 各複数文書の正解の表に占める空欄の割合

複数文書	空欄の割合
新聞記事 (強盗)	0.26
新聞記事 (外為・株式)	0.59
新聞記事 (地震)	0.57
新聞記事 (交通事故)	0.56
新聞記事 (リコール)	0.65
新製品記事 (スマホ)	0.41
新製品記事 (テレビ)	0.54
新製品記事 (カメラ)	0.59
新製品記事 (ロボット掃除機)	0.63
新製品記事 (エアコン)	0.53
Wikipedia(城)	0.57
Wikipedia(恐竜)	0.53
Wikipedia(力士)	0.52
Wikipedia(山)	0.56
Wikipedia(野球チーム)	0.58

5.2.2 表の精度の向上に向けた課題

一方で，表 5.4 から，最適なクラスタ数が選ばれたとしても評価結果が 0.70 に届いていない場合があり，クラスタ数の推定方法を改善するだけでは大幅な表の精度の向上は見込めない．今後，表の精度を向上させていくには，階層クラスタリングによる分類の精度を上げる必要がある．分類の精度を上げるために考えられることは二つある．

一つ目は文のベクトルの精度を高めることである．現在，文のベクトルを計算する際，文中の名詞の単語ベクトルを同じ重みで足し合わせているが，これを重要な単語ほど重みを大きくするなどしてベクトルの精度を高めることが考えられる．

二つ目は階層クラスタリングによって得られた樹形図を異なる距離でカットすることである．今回の実験では階層クラスタリングでのクラスタ数ごとのクラスタリング結果を得るために樹形図を水平にカットする方法を用いている．階層クラスタリングでは，類似した (距離の近い) 文から順にクラスタにまとめられるため，字面の似た文は樹形図の

下の階層で，すでに同じクラスタに統合されていると考えられる．一方で，同じ種類でも字面の似ていない文は樹形図の比較的上の階層で統合されると考えられる．このように，同種の文同士の距離が文の種類ごとに大きく異なる場合は，図 5.1 のように樹形図を距離に基づき水平にカットする方法では対応できない．この問題を解決するには，図 5.2 のように，樹形図を異なる距離に基づいてカットする必要がある．

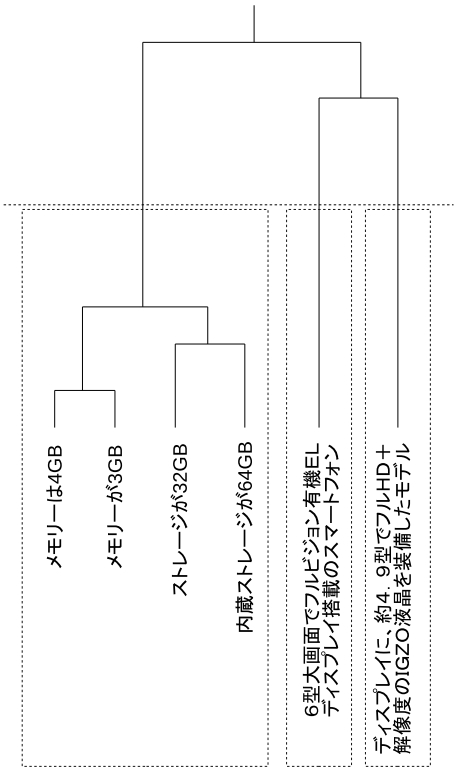

図 5.1: 樹形図を水平にカットする例

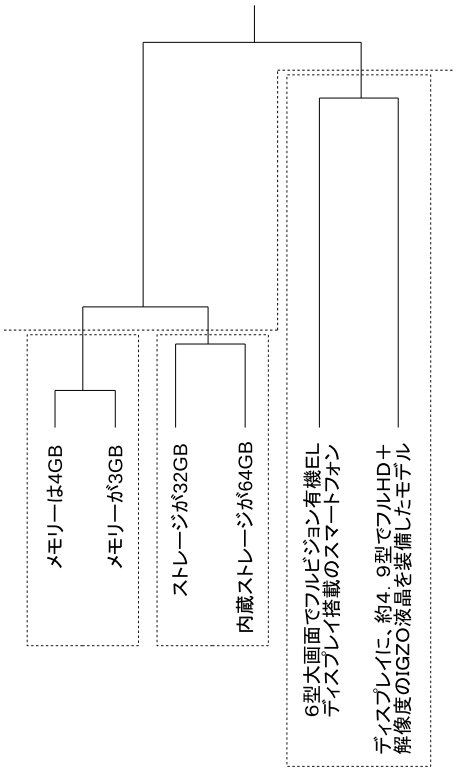

図 5.2: 樹形図を異なる距離でカットする例

5.3 追加実験 (他のクラスタ数の推定方法との比較)

今回，階層クラスタリングの結果から表の埋まり具合と密集度に基づき最適なクラスタ数を推定する手法を提案したが，最適なクラスタ数を推定する指標はすでにいくつか提案されている．そこで，次の 3 つのクラスタ数の推定手法を提案手法の手順 4 に適用して同様の実験を行った．

5.3.1 手法 1 : シルエット分析

シルエット分析 [7] は，クラスタ内のデータの凝集性とクラスタ間の離散性に基づく指標である．シルエット分析では，クラスタ数ごとに，全てのデータのシルエット係数を計算し，この平均値が最大となるときのクラスタ数を選択する．データ i のシルエット係数 $s(i)$ は式 5.1 で表される．ここで，$a(i)$ はデータ i とデータ i の属するクラスタに含まれる各データとの距離の平均，$b(i)$ はデータ i の属さないクラスタのうち，データ i と最も距離の近いクラスタに含まれる各データとデータ i との距離の平均を表す．ここで，データ i と最も距離の近いクラスタは，データ i とクラスタに含まれる各データとの距離の平均が最小となる場合のクラスタである．また，$max\{a(i), b(i)\}$ は $a(i)$ と $b(i)$ のうち，大きい方の値を表す．

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} \tag{5.1}$$

5.3.2 手法 2 : Upper Tail 法

Upper Tail 法は Mojena[8] によって提案された，階層クラスタリングの結果に対し統計的な停止規則に基づいて最適なクラスタ数を決定する手法である．Upper Tail 法では，まず，クラスタ数 j が 2〜$N-1$ となる場合の，それぞれのクラスタリング結果に対し，基準値 α_j を求める．ここで，α_j はクラスタ数 j でのクラスタリング結果における各クラスタの重心点間の距離のうち，最小の距離を表す．次に j の値を 2 からはじめて，以下の条件

$$\alpha_j \leq \bar{\alpha} + ks_\alpha$$

を満たさなくなるまで j の値を 1 ずつ増やしていく．停止したときの j が最適なクラスタ数として選ばれる．ここで，$\bar{\alpha}$ と s_α はそれぞれ，全ての基準値 α_j の平均と不偏分散の平方根を表す．k の値については，Mojena[8] では，データ数が 60〜120 の場合，2〜4 の値を用いている．また，志津ら [9] は 1 群のデータ数が 30〜50 前後の場合は $k = 3$ が

よいと報告している．これらを参考に，今回は 1 群のデータ数が文書数以下 (20 以下) になると仮定して，$k = 1$ で実験を行う．

5.3.3 手法 3 : BIC に基づく方法

2.3 節の X-means 法と同様の方法を階層クラスタリングの結果に対して適用し，最適なクラスタ数を推定する．具体的には，階層クラスタリングの結果得られた樹形図の根から順に二分割するごとに，分割前の BIC と分割後の BIC' をそれぞれ 2.3 節と同様に求め，$BIC \leq BIC'$ ならば分割を停止する．

5.3.4 実験結果

シルエット分析と Upper Tail 法を用いた場合の評価結果を表 5.6 に示す．また，評価結果の差が有意かを調べるために，対応のある両側 t 検定を行った．有意水準は 0.05 とした．有意差検定の結果を表 5.7 に示す．

表 5.6: 各表の評価結果

	従来手法 (X-means 法)			階層クラスタリング (Ward 法)			
	最大値	平均値	最小値	シルエット分析	Upper Tail 法	BIC	提案手法
新聞記事 (強盗)	**0.74**	0.58	0.30	0.64	0.61	0.69	0.66
新聞記事 (外為・株式)	0.54	0.37	0.13	**0.64**	0.55	0.36	0.63
新聞記事 (地震)	0.68	0.49	0.19	0.71	**0.72**	0.32	**0.72**
新聞記事 (交通事故)	0.55	0.39	0.23	**0.57**	0.51	0.45	0.51
新聞記事 (リコール)	0.59	0.41	0.18	0.61	0.64	0.51	**0.65**
新製品記事 (スマホ)	0.70	0.58	0.28	0.76	**0.78**	0.66	**0.78**
新製品記事 (テレビ)	0.51	0.35	0.13	0.49	0.44	0.47	**0.64**
新製品記事 (カメラ)	0.46	0.31	0.06	**0.60**	0.21	0.26	0.58
新製品記事 (ロボット掃除機)	0.47	0.35	0.11	0.49	0.37	0.32	**0.60**
新製品記事 (エアコン)	**0.52**	0.40	0.16	0.18	0.45	0.39	0.51
Wikipedia(城)	**0.82**	0.52	0.19	0.80	0.80	0.29	0.76
Wikipedia(恐竜)	0.60	0.42	0.23	0.65	0.72	0.40	**0.78**
Wikipedia(力士)	0.73	0.52	0.22	0.79	**0.84**	0.73	**0.84**
Wikipedia(山)	0.55	0.35	0.21	0.60	0.63	0.30	**0.65**
Wikipedia(野球チーム)	**0.56**	0.40	0.20	0.43	0.48	0.36	0.51
平均	0.60	0.43	0.19	0.60	0.58	0.43	**0.65**

表 5.7: 有意差検定の結果

	最大値	平均値	最小値	シルエット分析	UpperTail 法	BIC	提案手法
最大値							
平均値	**0.000**						
最小値	**0.000**	**0.000**					
シルエット分析	0.895	**0.000**	**0.000**				
UpperTail 法	0.506	**0.000**	**0.000**	0.699			
BIC	**0.000**	0.876	**0.000**	**0.004**	**0.005**		
提案手法	**0.024**	**0.000**	**0.000**	**0.037**	**0.026**	**0.000**	

5.3.5　シルエット分析を用いた階層クラスタリングの考察

シルエット分析を用いた階層クラスタリングの評価結果は，平均が 0.60 と従来手法を上回ったものの，提案手法よりは低い値となった．特に新製品記事 (エアコン) での評価結果が 0.18 と非常に低かった．新製品記事 (エアコン) は 1 文あたりの平均文字数が 62.3 文字と 15 種類の文書の中で最も多い．本研究では文ベクトルを含まれる単語ベクトルの総和として算出していることから，文字数の多い文では，文をよく表す重要な単語が他の多くの単語に埋もれてしまい，精度の低い文のベクトルが算出されてしまう．シルエット分析ではクラスタ内のデータの凝集性とクラスタ間の離散性が考慮されるがいずれも文のベクトルを基に計算されるため，このような精度の低い文のベクトルに大きく影響され低い評価結果となったと考えられる．一方で，1 文あたりの平均文字数が少ない文書では良い評価結果が得られている．よって，シルエット分析を用いた方法は簡潔な文のみを含む文書に対しては有効であると思われる．

5.3.6　Upper Tail 法を用いた階層クラスタリングの考察

Upper Tail 法を用いた階層クラスタリングの評価結果は，0.58 と従来手法を上回ったものの，提案手法，シルエット分析を用いた方法よりは低い値となった．Upper Tail 法は設定した k の値に大きく影響される．今回は k を 1 として実験を行ったが，この値の適切さを調べるために，k の値を 0〜4 の範囲で 0.1 刻みで変化させて実験を行った．15 の複数文書での結果をそれぞれ図 5.3，図 5.4，図 5.5，図 5.6，図 5.7，図 5.8，図 5.9，図 5.10，図 5.11，図 5.12，図 5.13，図 5.14，図 5.15，図 5.16，図 5.17 に示す．結果から，今回設定した k の値は新聞記事や Wikipedia から抽出した複数文書に対しては概ね適切であったと思われる．一方で，新製品記事に関する複数文書については，適切な k の値は「スマートフォンに関する新製品記事」を除く 4 種類の複数文書で 0.5 前後であり，今回設定した k の値は適切ではなかった．これらの複数文書の特徴として含まれる文の平均文字数が多いことが挙げられる．本研究では文ベクトルを含まれる単語ベクトルの総和として算出していることから，文字数の多い文では，文をよく表す重要な単語が他の多くの単語に埋もれてしまい，精度の低い文のベクトルが算出されてしまう．そのため，これらの複数文書で適切な k の値が設定できなかった原因は文のベクトルの精度が低いことが影響していると考えられる．よって，Upper Tail 法での最適な k の値を設定する際は，想定される 1 群のデータ数と同時に文の平均文字数から推定される文ベクトルの精度も考慮する必要があると考えられる．

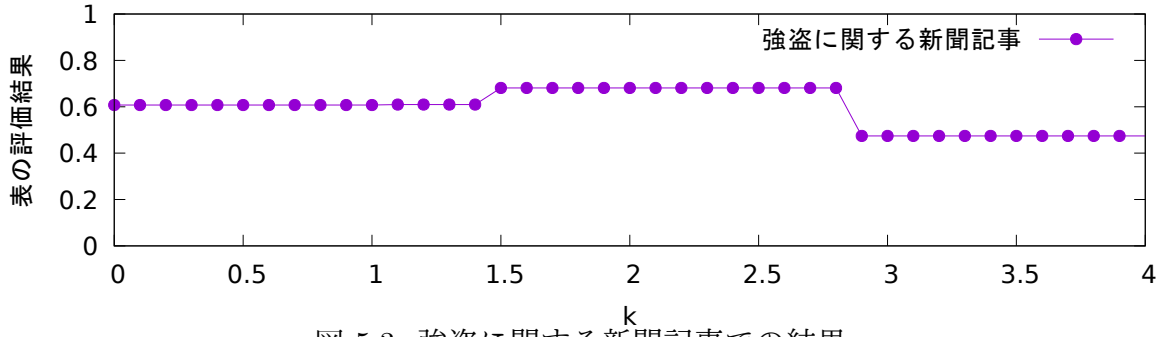

図 5.3: 強盗に関する新聞記事での結果

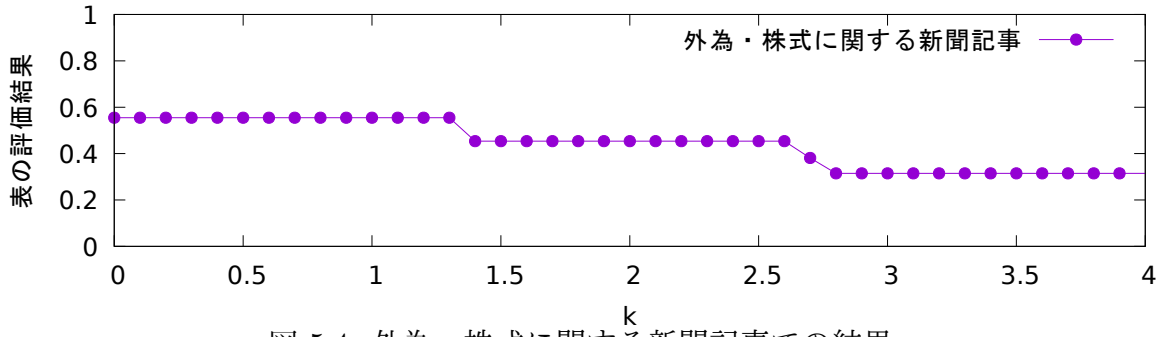

図 5.4: 外為・株式に関する新聞記事での結果

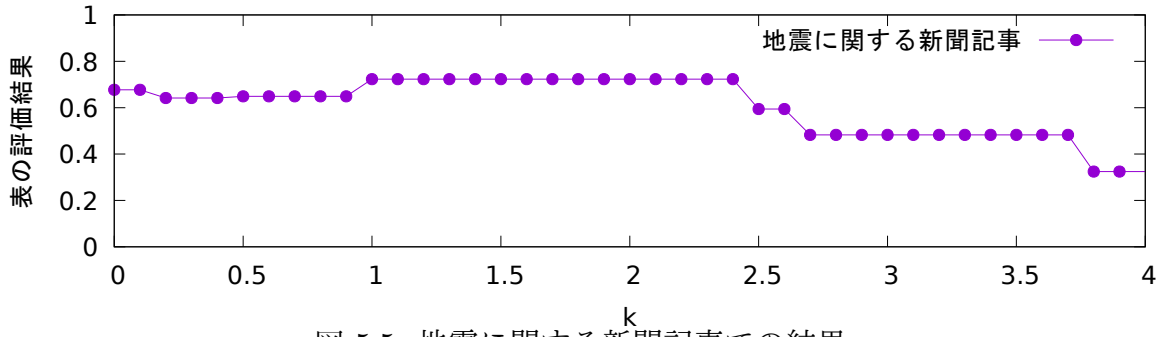

図 5.5: 地震に関する新聞記事での結果

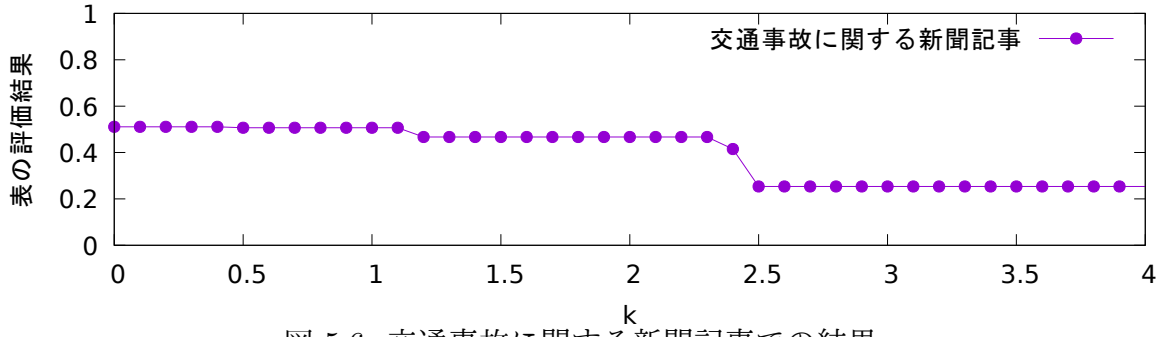

図 5.6: 交通事故に関する新聞記事での結果

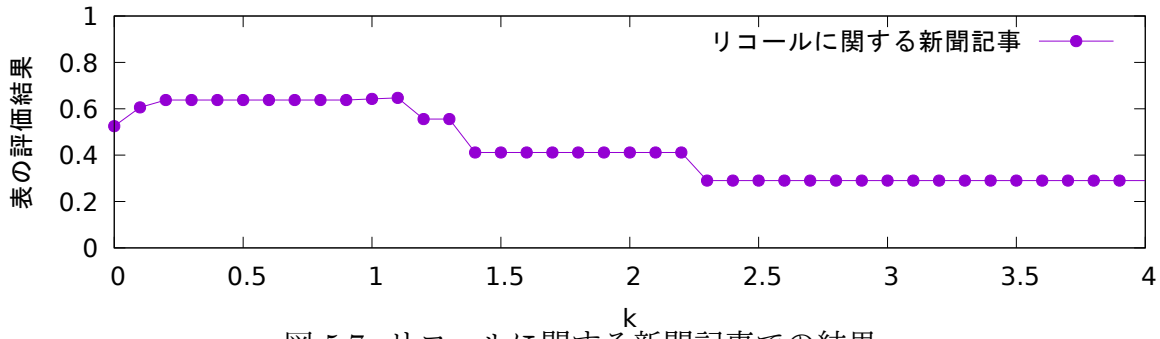

図 5.7: リコールに関する新聞記事での結果

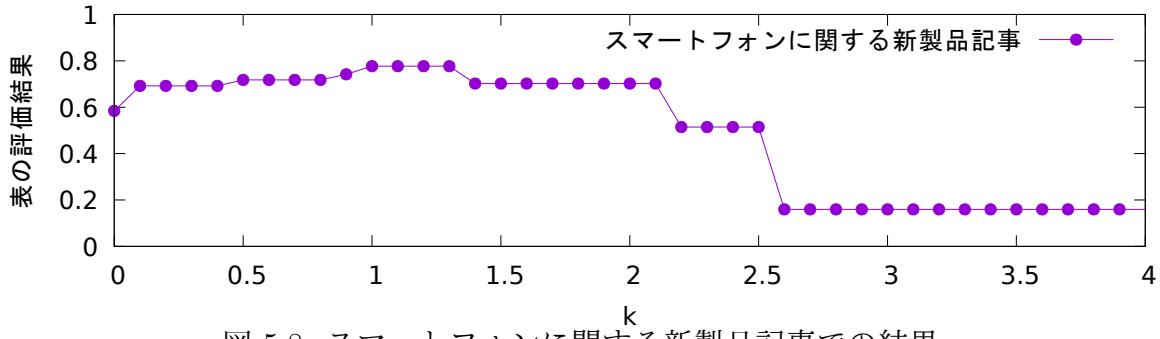

図 5.8: スマートフォンに関する新製品記事での結果

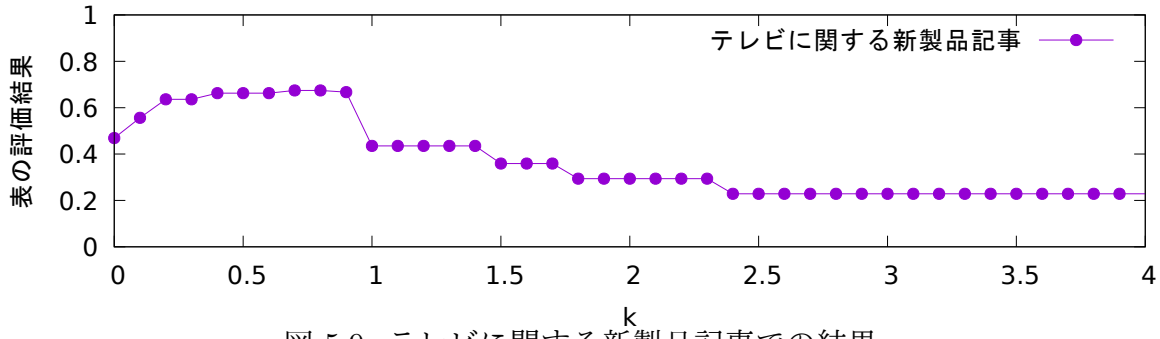

図 5.9: テレビに関する新製品記事での結果

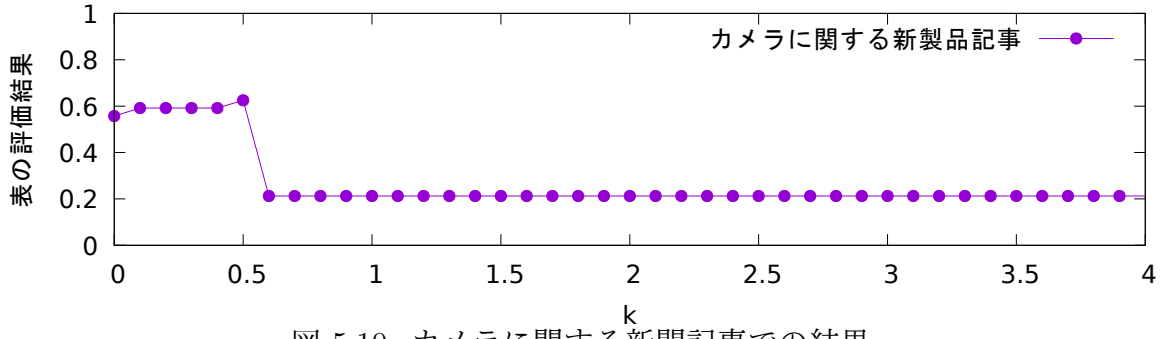

図 5.10: カメラに関する新聞記事での結果

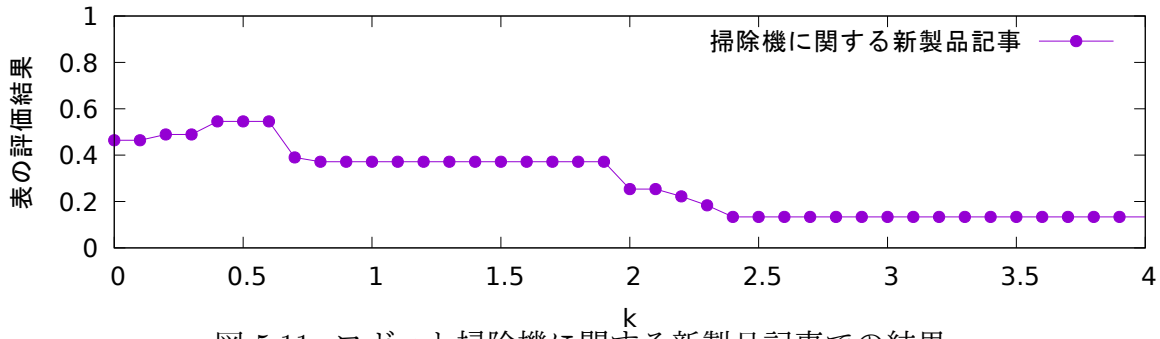

図 5.11: ロボット掃除機に関する新製品記事での結果

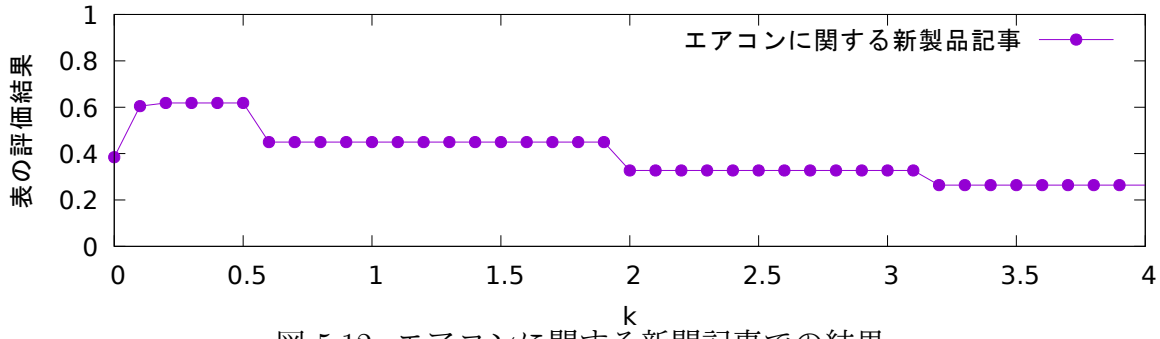

図 5.12: エアコンに関する新聞記事での結果

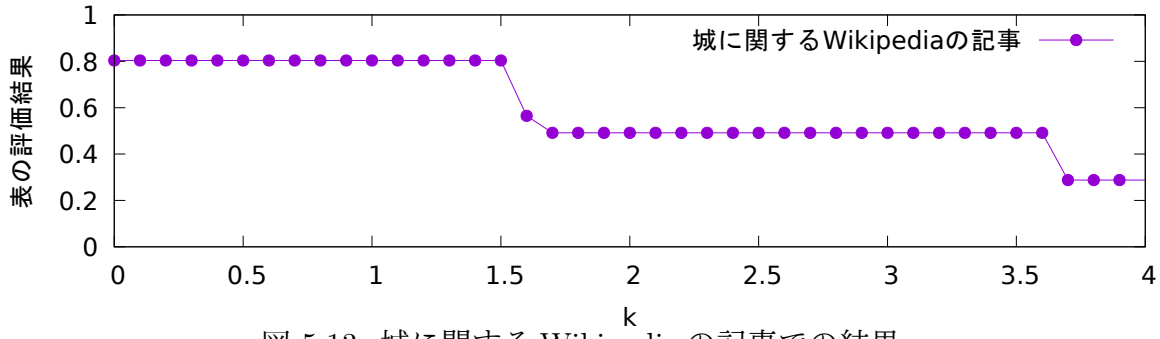

図 5.13: 城に関する Wikipedia の記事での結果

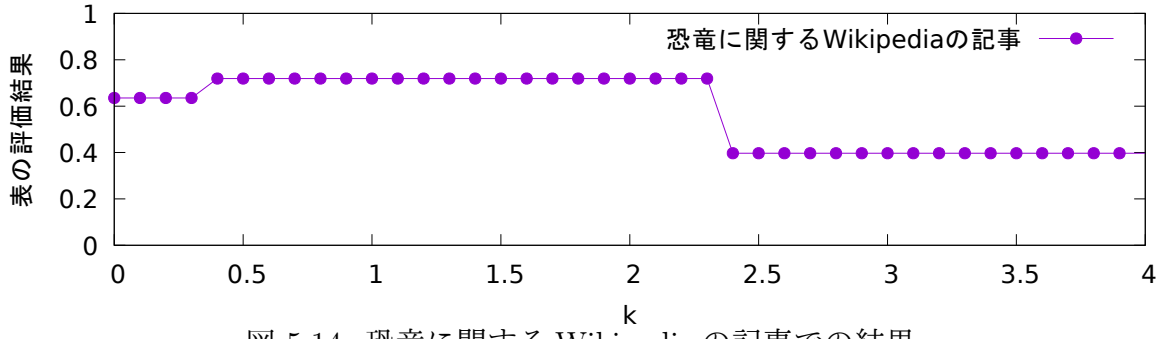

図 5.14: 恐竜に関する Wikipedia の記事での結果

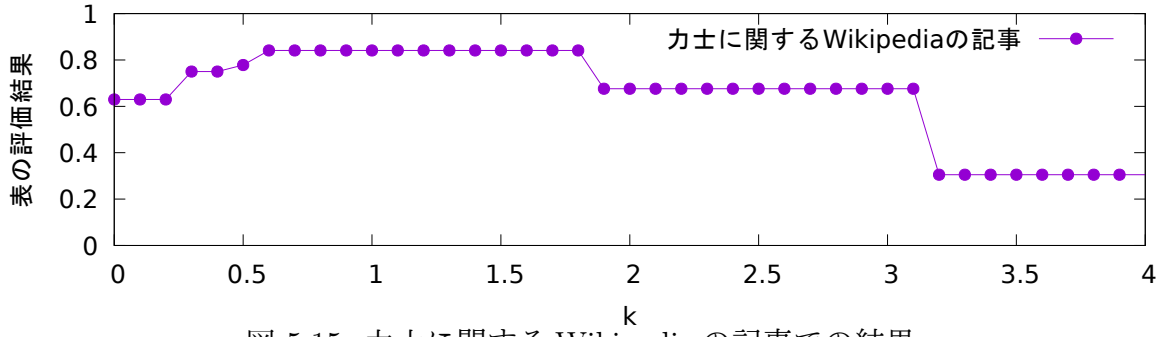

図 5.15: 力士に関する Wikipedia の記事での結果

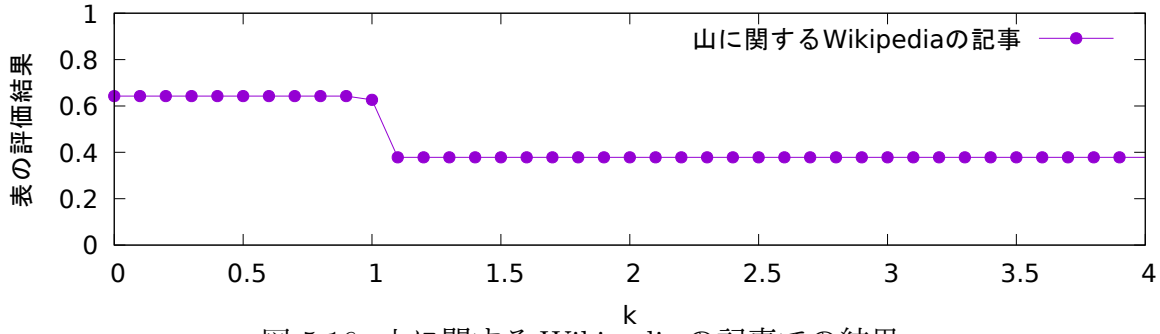

図 5.16: 山に関する Wikipedia の記事での結果

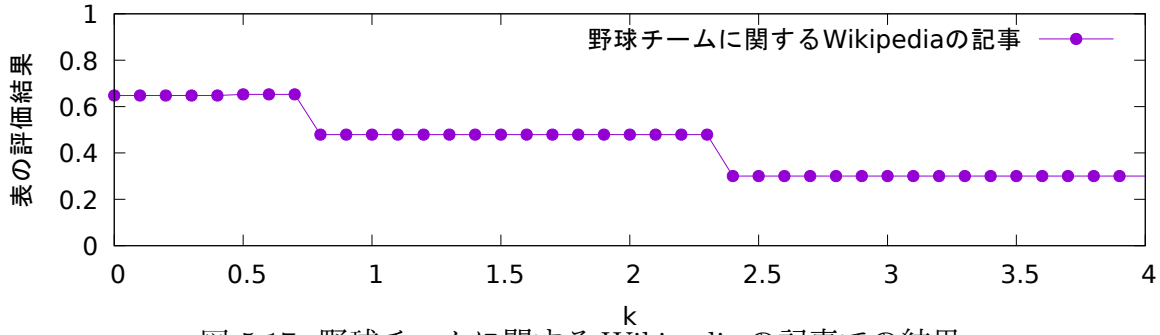

図 5.17: 野球チームに関する Wikipedia の記事での結果

5.3.7　BIC に基づく方法を用いた階層クラスタリングの考察

BIC に基づく方法を用いた結果は 0.43 と，従来手法の X-means 法を用いた場合の平均と同等の結果になった．この結果から，X-means 法 (K-means 法) と階層クラスタリングの分類方法の違いはではなく，今回提案したクラスタ数の推定方法によって表の精度を高めることができたと思われる．

第6章　おわりに

過去に，X-means 法を用いて，複数文書に含まれる文の情報を，表に整理する手法を提案した．X-means 法は BIC に基づいて最適なクラスタ数を推定するクラスタリング手法である．しかし，X-means 法によって推定されたクラスタ数 (表の列数) は最適なクラスタ数に比べ小さい傾向にあり，この結果を整理した表は情報が 1 つの列にまとまりすぎており，表の精度が低いという問題があった．

そこで，本研究ではこの問題を改善するために，文の情報を分類した階層クラスタリングの結果に対し，表の埋まり具合と情報の密集度のバランスを最適にする方法でクラスタ数を推定し，この結果を表に整理する手法を提案した．

提案手法では，まず，情報を階層クラスタリングでクラスタリングする．次に，階層クラスタリングのクラスタ数が 1〜n までの結果について，これを整理した表の，表の埋まり具合と，整理された情報の密集度を求める．この二つの指標のバランスが最適になるときのクラスタ数を最適なクラスタ数と推定する．最後に推定されたクラスタ数の結果を表に整理する．本研究では以上の手法により，表の精度の向上を試みた．. 15 種類の複数文書を用いた実験の結果，従来手法において X-means 法により推定されたクラスタ数が小さい傾向にあった問題は提案手法では改善され，より最適なクラスタ数に近づいたことが確認できた．これにより，従来手法では表の評価結果の平均が 0.43 だったが，提案手法では 0.65 と向上し，提案手法の有効性が確認できた．

一方で，最適なクラスタ数が推定できたとしても表の評価結果が 0.70 に届かない場合があるなど，クラスタ数の推定方法を改善するだけでは大幅な表の精度の向上は見込めないことも明らかになった．今後，表の精度を向上させていくには，階層クラスタリングによる分類の精度を上げる必要があると考えられる．よって，階層クラスタリングの分類の精度を上げる方法の検討が今後の課題である．

謝辞

最後に，2 年間研究を進めるに当たって，本研究のご指導を頂きました鳥取大学工学部知能情報工学科，自然言語処理研究室の村田真樹教授，村上仁一准教授そして自然言語処理研究室の皆様に深く感謝するとともに心から御礼申し上げます．また，参考にさせていただいた論文の著者の方々に対して深く感謝申し上げます．

付録　正解の表

表 6.1: Wikipedia(力士) の正解の表の一部

	最高位	身体的特徴	出身
文書 1	・最高位は西幕下 2 枚目		・琴藤沢和穂は、高知県高知市出身で佐渡ヶ嶽部屋に所属した元大相撲力士
文書 2	・最高位は東十両 12 枚目	・身長 183 ｃｍ、体重 191 ｋｇ	・政風基嗣は、長崎県長崎市出身で尾車部屋所属の現役大相撲力士
文書 3	・最高位は西幕下 2 枚目	・身長 175 ｃｍ、体重 130 ｋｇ、血液型はＯ型	・栃の山博士は、東京都立川市出身で、千賀ノ浦部屋の元大相撲力士で、現世話人
文書 4	・最高位は東幕下 4 枚目	・現役時代の体格は身長 179 ｃｍ、体重 149 ｋｇ、血液型はＡＢ型	・栃乃里隆光は、石川県石川郡野々市町出身で春日野部屋に所属していた元大相撲力士
文書 5	・最高位は西前頭 6 枚目	・血液型はＢ型 ・身長は 180 ｃｍ、体重は 160 ｋｇ	・誉富士歓之は、青森県西津軽郡鰺ヶ沢町出身で伊勢ヶ濱部屋所属の現役大相撲力士
文書 6	・最高位は西幕下 2 枚目	・本名は、三好正人、身長 178 ｃｍ、体重 196 ｋｇ、血液型Ａ型	・朝陽丸勝人は、大阪府枚方市出身で高砂部屋所属の元大相撲力士
文書 7	・最高位は西前頭 11 枚目	・現役時代の体格は 182 ｃｍ、92 ｋｇ	・吉井山朋一郎は、福岡県田川郡糸田町出身で、出羽海部屋に所属した大相撲力士
文書 8	・最高位は西前頭 2 枚目、血液型Ｏ型	・本名は小塚一、身長 186 ｃｍ、体重 147 ｋｇ ・最高位は西前頭 2 枚目、血液型Ｏ型	・朝乃翔嚆矢は、神奈川県小田原市出身で若松部屋所属の元大相撲力士
文書 9	・最高位は西十両 2 枚目	・身長 185 ｃｍ、体重 148 ｋｇ	・大岳宗正は滋賀県草津市出身の元大相撲力士
文書 10	・最高位は東小結	・全盛期の体格は 187 ｃｍ、144 ｋｇ	・大翔鳳昌巳は、北海道札幌市豊平区平岸出身で立浪部屋所属の元大相撲力士
文書 11	・最高位は西幕下 16 枚目	・現役時代の体格は身長 180 ｃｍ、体重 129 ｋｇ、血液型はＯ型	・玉大輝剛志は、石川県鳳珠郡能登町出身で片男波部屋に所属していた元大相撲力士
文書 12	・最高位は西関脇	・現役時代の体格は 179 ｃｍ、116 ｋｇ	・開隆山勘之亟は、秋田県南秋田郡昭和町出身で、1960 年代に活躍した大相撲力士 ・伊勢ヶ濱部屋に所属していた
文書 13	・最高位は東十両 12 枚目	・身長 182 ｃｍ、体重 146 ｋｇ	・大翔鵬清洋は、モンゴル・ウランバートル市出身で、追手風部屋所属の大相撲力士
文書 14	・最高位は西前頭 7 枚目	・身長 182 ｃｍ、体重 183 ｋｇ、血液型はＡ型、星座は蟹座	・木村山守は、和歌山県御坊市出身で春日野部屋所属だった元大相撲力士
文書 15	・最高位は西前頭筆頭	・現役時代の体格は 183 ｃｍ、150 ｋｇ	・薩洲洋康貴は、鹿児島県指宿市出身で、1980 年代に活躍した大相撲力士 ・井筒部屋に所属していた
文書 16	・最高位は東関脇	・現役時代の体格は 178 ｃｍ、115 ｋｇ	・栃東知頼は、福島県相馬郡日立木村出身の元大相撲力士 ・春日野部屋所属
文書 17	・最高位は東小結	・現役時代の体格は 175 ｃｍ、117 ｋｇ ・血液型はＡ型	・智ノ花伸哉は、熊本県八代市出身で立浪部屋に所属した大相撲力士
文書 18		・身長 192 ｃｍ、体重 120 ｋｇ ・大相撲力士時代は身長 192 ｃｍ、体重 120 ｋｇ	・田上明は、日本の実業家、元プロレスラー、元大相撲力士
文書 19			・逆鉾昭廣は鹿児島県姶良市出身で井筒部屋所属の元大相撲力士 ・なお、実際の出身地は東京都墨田区である
文書 20	・最高位は東三段目 51 枚目	・身長 182 ｃｍ、体重 126 ｋｇ、血液型はＯ型	・加賀ノ花麻衣は、石川県加賀市出身で千賀ノ浦部屋に所属していた元大相撲力士

表 6.2: Wikipedia(力士) の正解の表の一部 (続き)

	得意技	本名
文書 1		・本名は藤沢和穂
文書 2	・得意技は右四つ、寄り	・本名は北園基嗣
文書 3	・得意技は押し	・本名は山田博士
文書 4		・本名は矢鋪光太郎
文書 5	・得意手は突き・押し	・本名は三浦歓之
文書 6	・得意技は左四つ、寄り	
文書 7	・得意手は右四つ、突っ張り、叩き込み	・本名は吉井朋一郎
文書 8	・得意手は突っ張り、押し	
文書 9	・得意手は押し、左四つ、寄り	・本名は横江英樹
文書 10	・得意手は突っ張り、右四つ、上手投げ	・本名は村田昌巳
文書 11		・本名は久山毅
文書 12	・得意手は右四つ、寄り、上手捻り、首投げ	
文書 13		・本名はチミデレゲゼン・シジルバヤル
文書 14	・得意は押し	
文書 15	・得意手は突き、押し、叩き、引き	・本名は吉崎克幸
文書 16	・得意手は左四つ、上手出し投げ、右四つ、寄り	・本名は志賀駿男
文書 17		・本名は成松伸哉
文書 18		
文書 19		・本名は、福薗好昭
文書 20		・本名は川下源二

表 6.3: 新聞記事 (交通事故) の正解の表の一部

	概要
文書 1	・29 日午後 1 時 5 分ごろ、愛知県北名古屋市鍜治ケ一色西 2 の県道と市道の交差点で、乗用車と軽乗用車が出合い頭に衝突した
文書 2	・8 日午前 8 時ごろ、埼玉県上里町嘉美の町道で保育園の園児を送迎するバスが軽乗用車と衝突し、横転した
文書 3	・28 日午前 8 時ごろ、横浜市港南区大久保 1 の市道で車 3 台が絡む事故があり、はずみで軽トラックが横転し、集団登校中の小学生 9 人を巻き込んだ
文書 4	・27 日午前 7 時 45 分ごろ、兵庫県加古川市西神吉町中西の交差点で、軽乗用車と衝突したタクシーが弾みで登校中の小学生の列に突っ込んだ
文書 5	・9 日夜、香川県内を走る高松自動車道の上下線を軽乗用車が約 2 時間逆走し、別の乗用車に接触したほか、避けようと停車した乗用車にトラックが衝突する事故を引き起こした
文書 6	・9 日午後 3 時 40 分ごろ、広島県庄原市東城町の中国自動車道下り線で、バレーボール全日本男子の次期監督に内定している中垣内祐一さん＝大阪市平野区＝運転の乗用車が、工事規制中の警備員の男性をはねた
文書 7	・1 日午前 0 時 50 分ごろ、栃木市都賀町家中の北関東自動車道下り線・栃木都賀ジャンクション―都賀インターチェンジ間で、走行車線を走っていた乗用車が愛知県稲沢市の男性運転のトラックに追突した
文書 8	・2 日午前 2 時 10 分ごろ、北海道室蘭市東町 5 の国道 36 号交差点で、乗用車が道路脇の信号機の支柱に衝突して大破していると 110 番があった
文書 9	・26 日午前 5 時 45 分ごろ、大阪府寝屋川市池田北町の国道 1 号交差点で、横断歩道を自転車で通行していた男性が左折中の大型トラックにひかれて死亡した
文書 10	・20 日午後 7 時ごろ、東京都大田区蒲田本町 1 の環状 8 号線で、観光バスが中央分離帯にある信号機の柱に衝突した
文書 11	・26 日午前 6 時 40 分ごろ、大阪市旭区中宮 1 の市道交差点で、横断歩道を歩いていた 80 代くらいの女性が車にはねられた
文書 12	・4 日午後 9 時半ごろ、大阪市住吉区万代東 3 の府道で、あべの橋発遠里小野橋行きの大阪市営バスが道路脇の電柱などに接触した
文書 13	・8 日午後 9 時 55 分ごろ、香川県観音寺市柞田町の国道 11 号で、大型トレーラーが、地元の祭りで引いていた太鼓台に後ろから突っ込んだ
文書 14	・2 日午前 2 時 5 分ごろ、愛知県岡崎市駒立町の新東名高速道路上り線で、故障のため路側帯に停車していた観光バスに大型トラックが追突した
文書 15	・12 日午後 5 時ごろ、兵庫県宝塚市小浜 2 の国道 176 号で、いずれも 18 歳の男女 4 人が乗った乗用車が中央分離帯のガードレールに衝突、出火した
文書 16	・16 日午後 3 時半ごろ、奈良県川上村大迫の国道 169 号大迫トンネルで、ワゴン車と乗用車が正面衝突し、火災が起きた
文書 17	・8 日午前 2 時 45 分ごろ、兵庫県西宮市浜脇町の阪神高速神戸線下りで、中型トラックが大型トレーラーに追突し、トラックを運転していた同県南あわじ市の会社員、殿本亘幸さんが死亡した
文書 18	・8 日午前 7 時 55 分ごろ、静岡県磐田市中泉の県道交差点で、登校中に横断歩道を渡っていた市立磐田中部小学校 2 年の大石萌衣さん＝同所＝と、同級生の男子児童の 2 人がライトバンにはねられた
文書 19	・10 日午前 8 時 45 分ごろ、大阪府島本町山崎の名神高速上り線左ルートの天王山トンネル内で、路線バスや大型トラックなど計 5 台が絡む多重衝突事故があった
文書 20	・26 日午前 9 時半ごろ、大津市蛍谷の名神高速道路下り線で、高速バスが前のトラックに追突

表 6.4: 新聞記事 (交通事故) の正解の表の一部 (続き)

	現場	容疑
文書 1		・県警西枇杷島署は自動車運転処罰法違反の疑いで、乗用車の同市徳重東出、パート、大口久美子容疑者を現行犯逮捕した ・同法違反の過失致死傷容疑に切り替えて調べる
文書 2	・現場は、見通しのよい十字路で、信号機はなかった	
文書 3		・同署は軽トラックの運転手に自動車運転処罰法違反の疑いもあるとみて、詳しく事情を聴く
文書 4	・現場は信号機のある県道と市道の交差点	
文書 5		
文書 6		
文書 7		
文書 8	・現場は片側 2 車線のカーブ	
文書 9		・府警寝屋川署は、トラックを運転した京都市伏見区淀池上町、会社員、南隆樹容疑者を自動車運転処罰法違反の疑いで現行犯逮捕した
文書 10	・現場は片側 2 車線のほぼ直線の道路	・同署はバスの運転手、菅原正容疑者＝東京都足立区＝を自動車運転処罰法違反容疑で現行犯逮捕した ・現場から車が走り去るのが目撃されており、大阪府警旭署はひき逃げ事件として捜査を始めた
文書 11		・現場から車が走り去るのが目撃されており、大阪府警旭署はひき逃げ事件として捜査を始めた
文書 12	・大阪府警住吉署などによると、現場は片側 2 車線の直線道路	
文書 13	・同署によると、現場は見通しの良い片側 1 車線の直線道路 ・現場はＪＲ観音寺駅から南東に約 2・4 キロ	・香川県警観音寺署は、トレーラーを運転していた愛媛県大洲市、大川貴之容疑者を自動車運転処罰法違反容疑で現行犯逮捕した
文書 14	・片側 2 車線で見通しは良いという ・現場は岡崎サービスエリアから東京方面へ約 3 キロ	・県警高速隊は、トラックを運転していた福岡市博多区西春町 1 の会社員、斎藤信夫容疑者を自動車運転処罰法違反容疑で逮捕した
文書 15	・現場は片側 2 車線の直線道路	
文書 16		
文書 17		
文書 18		・県警磐田署は、ライトバンを運転していた浜松市南区金折町の会社員、河合秀幸容疑者を自動車運転処罰法違反で現行犯逮捕した
文書 19		
文書 20	・高速隊によると、現場は片側 2 車線の直線	

表 6.5: 新聞記事 (交通事故) の正解の表の一部 (続き)

	負傷者
文書 1	・軽乗用車に乗っていた、いずれも近くに住む水野照子さんと弟の英雄さんが搬送先の病院で死亡し、英雄さんの妻で運転していた和子さんが意識不明の重体
文書 2	・バスに乗っていた園児や保育士 12 人が病院に搬送されたが、全員軽傷という
文書 3	・軽トラックを運転していた同市磯子区の男性と軽乗用車を運転していた女性、同乗していた 30 代男性も軽傷 ・軽傷とみられる児童 8 人は、男児 4 人、女児 4 人で 1〜5 年生 ・11 人は軽傷とみられる ・神奈川県警港南署によると、うち男児 1 人が軽トラックの下敷きになり、搬送先の病院で死亡が確認された ・ 横浜市消防局によると、他の児童 8 人を含む 11 人も病院に搬送された ・県警によると、死亡したのは近くに住む市立桜岡小 1 年、田代優さん
文書 4	・タクシーに乗っていた同県姫路市内の男子中学生 2 人と軽乗用車を運転していた女性も軽傷を負った ・いずれも打撲などで軽傷とみられる ・車に乗っていた 3 人もけがをした ・県警加古川署などによると、市立西神吉小学校の 1〜4 年生の男児 7 人が病院に搬送された
文書 5	・女性が両足を打撲、トラックの男性＝香川県東かがわ市＝にけがはなかった ・男性にけがはなかった
文書 6	・中垣内さんは車が横転し、軽いけが ・男性は頭を強く打ち、重傷とみられる
文書 7	・栃木県警高速隊によると、トラックの男性は軽傷 ・運転席から男性の遺体が発見された
文書 8	・乗っていた男性 3 人は病院に運ばれたが、死亡が確認された ・道警室蘭署によると、死亡したのは登別市新生町 1 の会社員、長尾卓弥さんと東京都江東区東陽 5 の会社員、平石隆祥さん、室蘭市小橋内町 2、自営業、山下知弥さん 3 人は室蘭市出身で、小中学校時代の同級生
文書 9	・寝屋川署によると、死亡したのは成人とみられ、黒っぽい服を着ていた ・同署が身元の確認を急いでいる
文書 10	・警視庁蒲田署によると、乗客の 20〜70 代の男女 24 人がけがをしたが、いずれも軽傷とみられる ・同署などによると、バスには乗客 28 人のほか、運転手と添乗員 1 人ずつが乗っていた
文書 11	・女性は頭を強く打ち、病院に運ばれたが意識不明の重体
文書 12	・乗客 10 人にけがはなかった
文書 13	・近くの教員、冨田規之さんが死亡し、太鼓台を引いていた 20 人以上が搬送され、うち数人が重傷という
文書 14	・故障に対処するため車外にいたバスの男性運転手 2 人が、バスと側壁に挟まれ、出血性ショックのため間もなく死亡した ・バス内にいた乗客の大阪市の女子大学生と、大型トラックの男性運転手の計 2 人も切り傷など軽傷を負った ・死亡した運転手は、大阪府太子町葉室、大谷秀雄さんと、大阪市城東区中浜 1 の染谷文彦さん
文書 15	・兵庫県警宝塚署によると、後部座席には 2 人乗っており、女子高校生が意識不明の重体、会社員の男性が鎖骨を折って重傷 ・助手席にいた男子高校生は軽傷とみられる ・車を運転していた同県伊丹市のとび職の男性が、全身を強く打って搬送先の病院で死亡した
文書 16	・乗用車に乗っていた 3 人のうち、西東京市の無職、藤井純代さんが頸椎損傷で死亡し、運転していた夫が右足骨折、親族の女性＝奈良県桜井市＝が頭に打撲のけがをした ・県警吉野署によると、2 台とも全焼 ワゴン車に乗っていた性別不明の 2 人が遺体で見つかった
文書 17	
文書 18	・男子児童も顔に軽傷 ・大石さんは搬送先の病院で死亡が確認された
文書 19	・この事故で大型トラックの 30 代の男性運転手が左足に軽傷を負った
文書 20	・計 4 台が絡む玉突き事故になり、バスの運転手や乗客ら計 8 人が病院に搬送された ・滋賀県警高速隊などによると、バスの男性運転手が骨折している可能性があるが、7 人は軽傷という

参考文献

[1] 吉谷仁志, 黄瀬浩一, 松本啓之亮. サポートベクトルマシンを用いた新聞記事からのプロフィール情報抽出. 電気学会論文誌C（電子・情報・システム部門誌）, Vol. 124, No. 11, pp. 2260–2266, 2004.

[2] Tsutomu Hirao, Hideto Kazawa, Hideki Isozaki, Eisaku Maeda, and Yuji Matsumoto. Machine learning approach to multi-document summarization. *Journal of Natural Language Processing*, Vol. 10, No. 1, pp. 81–108, 2003.

[3] 岡崎健介, 村田真樹, 馬青. 複数文書からの重要情報の抽出と表の作成. 言語処理学会第 24 回年次大会, pp. 240–243.

[4] Dan Pelleg and Andrew W. Moore. X-means : Extending k-means with efficient estimation of the number of clusters. *Proc. of the 17th International Conference on Machine Learning, 2000*, pp. 727–734, 2000.

[5] 石岡恒憲. クラスター数を自動推定する k-means アルゴリズムの拡張について. 応用統計学, Vol. 29, No. 3, pp. 141–149, 2000.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[7] Peter Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, No. 1, pp. 53–65, 1987.

[8] R. Mojena. Hierarchical grouping methods and stopping rules: an evaluation*. *The Computer Journal*, Vol. 20, No. 4, pp. 359–363, 1977.

[9] 志津綾香, 松田眞一. クラスター分析におけるクラスター数自動決定法の比較. *Academia Information sciences and engineering*, Vol. 11, pp. 17–34, 2011.