2021年度（令和3年度）　修士論文

クラスタリング法を用いた単語レベルでの重要情報抽出の改良

令和 4 年 2 月

鳥取大学大学院　持続性社会創生科学研究科
工学専攻　情報エレクトロニクスコース

自然言語処理研究室

M20J4046H　FU JIAJUN

概要

近年, インターネットの発展と伴い, ネット上の情報が急速的に増えている. このような膨大な情報から重要な情報を抽出して, 整理する研究が重要になっている. このような重要な情報は情報検索する際に重要な単語をキーワードとして使うこともできる. 赤野らの研究 [1] では Wikipedia 全データを用いて, 単語レベルで重要な情報を文書から抽出して, 表に整理する. この研究では Wikipedia 全データを単語レベルで分割して, これらの単語を k-means 法でクラスタリングし, クラスタリングの結果と処理したい文書を比較し, 重要な単語レベルの情報の種類を人手で選択して, 表に整理する. しかし, この研究では人手でクラスター数を決めていて, 最適なクラスター数になっておらず, 情報の重要度も計算していなく, 人手で決める必要があるという問題がある. 岡崎らの研究 [2] では文レベルで重要な情報を文章から抽出し, 表に整理する. 岡崎らの研究 [2] ではデータの密集度とカバー率を使って, クラスタリング際に最適なクラスター数とクラスターの重要度を計算して, 重要な情報を表に整理する. 本研究では岡崎らの研究成果 [2] を使って, 赤野らの研究の問題点を解決することで, 単語レベルで重要な情報を文書から抽出し, 表に整理する. 本研究では提案手法と 15 種類の複数文書を用いた実験の結果, 赤野らの研究 [1] では f 値の平均値は 0.21 であり, 提案手法では 0.60 以上であり, 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, インターネットの発展と伴い, ネット上の情報が急速的に増えている. このような膨大な情報から重要な情報を抽出して, 整理する研究が重要になっている. このような重要な情報は情報検索する際に重要な単語をキーワードとして使うこともできる. 赤野らの研究 [1] では Wikipedia 全データを用いて, 単語レベルで重要な情報を文書から抽出して, 表に整理する. この研究では Wikipedia 全データを単語レベルで分割して, これらの単語を k-means 法でクラスタリングし, クラスタリングの結果と処理したい文書を比較し, 重要な単語レベルの情報の種類を人手で選択して, 表に整理する. しかし, この研究では人手でクラスター数を決めていて, 最適なクラスター数になっておらず, 情報の重要度も計算していなく, 人手で決める必要があるという問題がある. 岡崎らの研究 [2] では文レベルで重要な情報を文章から抽出し, 表に整理する. 岡崎らの研究 [2] ではデータの密集度とカバー率を使って, クラスタリング際に最適なクラスター数とクラスターの重要度を計算して, 重要な文レベルの情報を表に整理する. 文レベルで重要な情報を表に整理する場合, 文が長くて, 重要な情報が見づらい状況がある. 本研究では岡崎らの研究成果 [2] を使って, 赤野らの研究の問題点を解決することで, 単語レベルで重要な情報を文書から抽出し, 表に整理する. 本研究での主張点は以下の 3 点である.

新規性

赤野らの研究 [1] は人手でクラスタ数 1000 で設定して, 人手で重要な列を選択する. 本研究では自動でクラスタ数を決定して, 重要度の順に列を自動的に並べ替える.

有用性

赤野らの研究 [1] では人手で重要な列を選択する必要がある. 本研究では手作業がなくても, 自動的に重要な列を選択することができる.

性能

本研究で整理した表の性能を F 値で評価すると, 平均値は 0.64 である.F 値で赤野らの研究 [1] を評価すると, その平均値は 0.21 である.

第2章　従来手法

2.1　従来手法1(赤野)

赤野ら [1] の情報抽出の研究では同じ種類の複数の文書を用いて, クラスタリング法を利用して, 単語レベルの重要な情報を文書から抽出して, 抽出した情報を表に整理する.

2.1.1　従来手法1の手順 (赤野)

過去の研究の文の情報を整理する方法の手順を以下に示す.

手順1　処理したい文書を収集する.

手順2　事前準備として, Wikipedia 全データを用いて, MeCab でデータを単語レベルで分割する. Word2vec を用いて単語をベクトルに変換する. 人手でクラスター数を 2,000 で設定して, データを k-means 法でクラスタリングする. クラスタリングすることで, よく似ている単語が同じクラスターに分類するクラスター数が 2,000 のクラスタリング表を作成する. 具体例を表 2.1 で示す.

手順3　MeCab を用いて, 文書データを単語レベルで分割して, 名詞単語以外の単語を削除する.

手順4　手順2 でできた表と手順3 で処理した文書を利用して, 重要な情報を結果の表に整理する. 具体的なやり方として, クラスタリング結果に基づく単語のクラスターを表の列とし, 文章を表の行とし, 単語レベルで分割した文書に出現するクラスターの単語を該当する行と列の箇所に埋める. 具体例を表 2.2 で示す. 例えば, 表 2.2 の欄 1 の (マグニチュード, 地震) の意味は文書番号 1 そしてクラスター 1718 に含まれているの意味である.

手順5　そして, 人手で重要な列を結果の表から選ぶ.

2.1.2 MeCab

MeCab は京都大学情報学研究科と日本電信電話株式会社コミュニケーション科学基礎研究所共同研究ユニットプロジェクトを通じて開発されたオープンソース形態素解析エンジンである. 特徴として, 辞書, コーパスに依存しない, 条件確率確率場 CRF に基づく高い解析性能と考えられる. 過去の研究では MeCab で単語を分割する. そして, 名詞の識別も MeCab でできる. 入力の例と出力の例を以下で示す.

入力の例:

自然言語処理は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術である.

出力の例

自然 名詞, 形容動詞語幹,*,*,*,*, 自然, シゼン, シゼン
言語 名詞, 一般,*,*,*,*, 言語, ゲンゴ, ゲンゴ
処理 名詞, サ変接続,*,*,*,*, 処理, ショリ, ショリ
は 助詞, 係助詞,*,*,*,*, は, ハ, ワ
、 記号, 読点,*,*,*,*,、 ,、 ,、
人間 名詞, 一般,*,*,*,*, 人間, ニンゲン, ニンゲン
が 助詞, 格助詞, 一般,*,*,*, が, ガ, ガ
日常 名詞, 一般,*,*,*,*, 日常, ニチジョウ, ニチジョー
的 名詞, 接尾, 形容動詞語幹,*,*,*, 的, テキ, テキ
に 助詞, 副詞化,*,*,*,*, に, ニ, ニ
使っ 動詞, 自立,*,*, 五段・ワ行促音便, 連用タ接続, 使う, ツカッ, ツカッ
て 助詞, 接続助詞,*,*,*,*, て, テ, テ
いる 動詞, 非自立,*,*, 一段, 基本形, いる, イル, イル
自然 名詞, 形容動詞語幹,*,*,*,*, 自然, シゼン, シゼン
言語 名詞, 一般,*,*,*,*, 言語, ゲンゴ, ゲンゴ
を 助詞, 格助詞, 一般,*,*,*, を, ヲ, ヲ
コンピュータ 名詞, 一般,*,*,*,*, コンピュータ, コンピュータ, コンピュータ
に 助詞, 格助詞, 一般,*,*,*, に, ニ, ニ
処理 名詞, サ変接続,*,*,*,*, 処理, ショリ, ショリ
さ 動詞, 自立,*,*, サ変・スル, 未然レル接続, する, サ, サ
せる 動詞, 接尾,*,*, 一段, 基本形, せる, セル, セル
一連 名詞, 一般,*,*,*,*, 一連, イチレン, イチレン
の 助詞, 連体化,*,*,*,*, の, ノ, ノ
技術 名詞, 一般,*,*,*,*, 技術, ギジュツ, ギジュツ
で 助動詞,*,*,*, 特殊・ダ, 連用形, だ, デ, デ
ある 助動詞,*,*,*, 五段・ラ行アル, 基本形, ある, アル, アル
EOS

2.1.3 k-means

クラスタリング法は教師なし学習の方法の一つである. クラスタリングの目的はラベルがないデータのパターンを探す. クラスタリング法には色んなアルゴリズムがある, 赤野の研究で用いられる方法はK-means法である. k-means法のアルゴリズムを紹介する. まずは人手でクラスター数を決める. クラスターの数をKで示す. データを $x_1, x_2, x_3 ... x_n$ で

示す, クラスターは $c_1, c_2, c_3...c_k$ で示す. クラスターの中心ベクトルを $\mu_1, \mu_2, \mu_3.....\mu_k$ で示す. 手順を以下で示す.

手順 1　ランダムで K 個のベクトルを選ぶ. この K 個のベクトルを k 個のクラスターの中心ベクトル μ として扱う.

手順 2　あるデータ x と K 個のクラスターの中心ベクトルとの距離 (ユークリッド距離) を計算する. データ x を距離が最も近いクラスターに分類する.

手順 3　クラスターの中心ベクトルを更新する. 更新方法はクラスターにあるすべての単語ベクトルの elementwise-mean[1]をこのクラスターの中心ベクトル μ として扱う.

手順 4　手順 2 と手順 3 をすべでのデータを 1 個つず用いて繰り返す.

k-means 法の結果の例を図 2.1.3 で示す.

2.1.4　過去研究の問題点 (赤野)

- 問題 1:クラスタリングの性能はクラスター数によって変わる. 人手でクラスター数を決めると, クラスタリングの性能が低いという問題がある.
- 問題 2:情報の重要度を計算する必要がある. 過去の研究では列にある単語の延数で列を並べ替えて, 人手で重要な列を選択する必要がある.

2.2　従来手法 2(岡崎)

本研究で最適なクラスター数と列の重要度を計算する際に岡崎らの研究成果が必要である. 岡崎らの研究 [2] では文レベルで重要な情報を文書から抽出する. 岡崎らの研究 [2] では表の埋まり具合と情報の密集度のバランスで最適なクラスター数を推定して, クラスターの重要度も表の埋まり具合と情報の密集度で計算する. 最適なクラスター数とクラスターの重要度の情報を用いて, クラスタリングの結果を表に整理する. 手順を以下で示す.

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.

[1]ベクトルにある要素ごとに平均を求める.

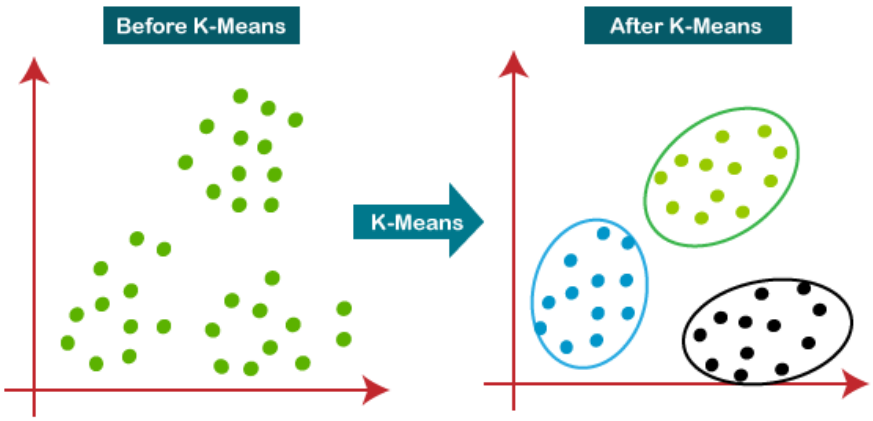

図 2.1: k-means 概略図

表 2.1: クラスター数を 2000 で設定したのクラスタリング表

クラスター 1	クラスター 2	クラスター 3	...
圧曲	concagua	阿三	
圧濃	Adamaoua	逢沢	
加圧水	Adamawa	為一	
火入れ	Alin	維四	
寒剤	Alphonsus	右一	
汽力	Andes	宇能	...
空気泡	Ararat	宇八	
軽水	Asha	卯一	
軽水炉	Aspiring	浦山	
原子力	Athos	営来	
...	...	...	

表 2.2: 実験で生成されたテーブル (従来手法 (データ:地震))

文書番号	1718	1311	1036	428	...
文書 1	マグニチュード, 地震	震源, 気象庁	推定	規模	...
文書 2	地震		推定	規模	...
文書 3	地震		推定	規模	...
文書 4	マグニチュード, 地震	震源, 気象	推定	規模	...
文書 5	マグニチュード, 地震	震源, 気象庁	推定	規模	...
文書 6	マグニチュード		推定	規模	...
文書 7	マグニチュード, 地震	震源, 気象庁	推定	規模	...
文書 8					...
文書 9	マグニチュード	震源, 気象庁	推定		...
文書 10	地震		推定	規模	...
文書 11	地震		推定	規模	...
文書 12	マグニチュード, 地震	震源, 気象庁	推定	規模	...
文書 13	地震		推定	規模	...
文書 14	地震		推定	規模	...
文書 15	マグニチュード, 地震	気象庁, 震源	推定	規模	...
文書 16	マグニチュード, 地震	気象庁, 震源	推定	規模	...
文書 17	マグニチュード, 地震	気象庁, 震源	推定	規模	...
文書 18	マグニチュード, 地震	気象庁, 震源, 震度, 観測	推定	規模	...
文書 19	マグニチュード, 地震	気象庁, 震源, 震度, 観測	推定	規模	...
文書 20	マグニチュード, 地震		推定	規模	...

手順3　人手でクラスター数を1から１０００まで設定して, 文ベクトルを基に文を階層クラスタリングで複数回クラスタリングする.

手順4　手順3の結果に基づいて, 行を文書, クラスターを列とする表の埋まり具合と情報の密集度のバランスを用いて, これらの複数のクラスタリングの表の中から最適なクラスター数の表を選択する.

手順5　手順4で得られた表の列 (クラスター) の重要度を計算して, 表の列を重要度で並べ替える. 地震のデータでできた結果の表を表2.3と表2.4に示す.

2.2.1　階層クラスタリング

階層クラスタリング法 [3] は最初一つのデータを一つのクラスターとして扱う. 距離が最も近い二つのクラスターを接合する. この手順を設定したクラスター数に到達するまで繰り返す.

2.2.2　最適なクラスター数を計算する方法 (岡崎)

岡崎らの研究 [2] で使った最適なクラスター数を計算する方法を紹介する前に, 二つの重要な概念を紹介する必要がある. 一つ目は情報のカバー率 ($cover_k$) である. これは表の空欄ではないセルの割合である. 数式は式2.1に示す. 表Kはクラスター数kで作った表である. $cover_k$ はクラスター数kで作った表のカバー率である.

$$cover_k = \frac{\text{表 } K \text{ の空ではないセルの数}}{\text{表 } K \text{ のセルの総数}} \tag{2.1}$$

二つ目は情報の密集度 ($density_k$) である. 情報の密集度 ($density_k$) とは表の各列にあるデータ間の最小類似度である. 数式2.2に示すと. W_{kij} はクラスター数kでできた表のi番目の列のj番目のデータの意味である. $|C_k|$ はクラスター数kでできた表の列の数である. $|C_{ki}|$ はクラスター数kでできた表の列のIのベクトルの数である.

$$density_k = min\left(cos\left(W_{kij}, W_{kih}\right)\right) \tag{2.2}$$

$$i = 1, 2, ..., |C_k| \quad j, h = 1, ..., |C_{ki}|$$

表 2.3: 文レベルでクラスタリング結果 (岡崎らの研究 (地震)(列 1))

文書番号	マグニチュード
文書 1	気象庁によると、震源の深さは約 50 キロ、地震の規模を示すマグニチュードは 5・0 と推定される
文書 2	地震の規模はマグニチュード 4．4 と推定される
文書 3	地震の規模はマグニチュード 3・4 と推定される, 地震の規模はＭ 5・8 と推定される
文書 4	気象庁によると、震源の深さは約 10 キロで、地震の規模を示すマグニチュードは 6・1 と推定される
文書 5	気象庁によると、震源は同県熊本地方、震源の深さは約 10 キロ、地震の規模を示すマグニチュードは 4・7 と推定される
文書 6	マグニチュードは 4．6 と推定される
文書 7	気象庁によると、震源の深さは約 13 キロ、地震の規模を示すマグニチュードは 5・2 と推定される
文書 8	
文書 9	気象庁によると、震源の深さは約 10 キロ、マグニチュードは 5・5 と推定される
文書 10	地震の規模はマグニチュード 3．7 と推定される
文書 11	地震の規模はマグニチュード 5．5 と推定される
文書 12	気象庁によると、震源は日向灘、震源の深さは約 60 キロ、地震の規模を示すマグニチュードは 4・4 と推定される
文書 13	地震の規模はマグニチュード 4・1 と推定される
文書 14	地震の規模はマグニチュード 3・1 と推定される
文書 15	気象庁によると、震源の深さは約 10 キロ、マグニチュードは 4．7 と推定される
文書 16	気象庁によると、震源の深さは約 9 キロ、地震の規模を示すマグニチュードは 5・8 と推定される
文書 17	気象庁によると震源の深さは約 30 キロ、地震の規模を示すマグニチュードは 4・6 と推定される
文書 18	気象庁によると、震源の深さは約 20 キロ、地震の規模を示すマグニチュードは 4・2 と推定される
文書 19	気象庁によると、震源の深さは約 50 キロ、地震の規模を示すマグニチュードは 5・3 と推定される
文書 20	地震の規模を示すマグニチュードは 5・5 と推定される

表 2.4: 文レベルでクラスタリング結果 (岡崎らの研究 (地震)(列 2))

文書番号	発生情報
文書 1	20 日午前 7 時 25 分ごろ、茨城県南部を震源とする地震があり、さいたま市や水戸市などで最大震度 4 を観測した
文書 2	26 日午前 4 時 52 分ごろ、茨城、栃木、埼玉、千葉の各県で震度 3 の地震があった
文書 3	27 日午前 3 時 28 分ごろ、長野県で震度 3 の地震があった, また、同 7 時 57 分ごろ、宮城、福島、茨城、栃木の各県で震度 3 の地震があった
文書 4	1 日午前 11 時 39 分ごろ、三重県南東沖を震源とする地震があり、和歌山県古座川町で震度 4 を観測した
文書 5	1 日午前 6 時 33 分ごろ、熊本市西区・南区や熊本県宇城市、上天草市で震度 4 を観測する地震があった
文書 6	25 日午後 1 時 51 分ごろ長野県の小谷村と小川村で震度 4 の地震があった
文書 7	31 日午後 7 時 46 分ごろ、熊本県熊本地方を震源とする地震があり、熊本市西区と熊本県宇城市で震度 5 弱を観測した
文書 8	熊本県・大分県を中心に続いている一連の地震で、10 日にこれまでの主な活動領域よりやや南西に離れたところで地震が発生した
文書 9	19 日午後 5 時 52 分ごろ、熊本県熊本地方を震源とする地震があり、同県八代市で震度 5 強、同県氷川町と同県芦北町で震度 5 弱を観測した
文書 10	26 日午前 9 時 49 分ごろ、高知県で震度 3 の地震があった
文書 11	26 日午後 2 時 13 分ごろ、北海道函館市で震度 4 の地震があった
文書 12	22 日午前 3 時 33 分ごろ、大分県佐伯市で震度 4 を観測する地震があった
文書 13	22 日午前 7 時 59 分ごろ、千葉県銚子市で震度 3 の地震があった
文書 14	25 日午後 8 時 48 分、兵庫県養父市で震度 3 の地震があった
文書 15	22 日午後 2 時 34 分ごろ、茨城県北部を震源とする地震があり、同県常陸太田市で震度 4 を観測した
文書 16	18 日午後 8 時 41 分ごろ、熊本県阿蘇地方を震源とする地震があり、熊本県阿蘇市などで震度 5 強を観測した
文書 17	5 日午前 7 時 41 分ごろ、神奈川県東部を震源とする地震があり、川崎市川崎区や東京都町田市で震度 4、神奈川県や東京都の広い範囲で震度 3 を観測した
文書 18	18 日午前 8 時 50 分ごろ、茨城県沖を震源とする地震があり、水戸市などで震度 3 を観測した
文書 19	27 日午後 11 時 47 分ごろ、茨城県北部を震源とする地震があり、同県日立市や常陸太田市などで震度 5 弱を記録した
文書 20	16 日午後 9 時 23 分ごろ、茨城県南部を震源とする地震があり、同県小美玉市で震度 5 弱を記録したほか、東北から中部地方の広い範囲で揺れを観測した

この二つの数値の正規化結果の掛け算の結果をこの表の Score として扱う. 式 2.3 に数式を示す. 式 2.4 に x は全部のデータの意味である.

$$Score_k = norm(density_k) * norm(cover_k) \tag{2.3}$$

$$norm(x_n) = \frac{x_n - min(x)}{max(x) - min(x)} \tag{2.4}$$

2.2.3 列の重要度の計算方法 (岡崎)

列の重要度を計算するときは列の情報のカバー率と情報の密集度を用いて計算する.

列の情報のカバー率 ($cover_i$) は列の空ではないセルの割合である. 数式を式 2.5 に示す.

$$cover_i = \frac{\text{列 } i \text{ の空ではないセルの数}}{\text{列 } i \text{ のセルの総数}} \tag{2.5}$$

列の情報の密集度 ($density_i$) は列 i にあるデータの間の最小類似度である. 数式で示す. $|C_i|$ は列 i にあるデータの数である.

$$density_i = min\left(cos\left(W_{ij}, W_{ih}\right)\right) \tag{2.6}$$

$$j, h = 1, ..., |C_i|$$

列 i の重要度 ($Important_i$) はこの二つの値の正規化値の掛け算の結果である.

$$Important_i = norm(density_i) * norm(cover_i) \tag{2.7}$$

第3章　提案手法

3.1　提案手法の概要

赤野らの研究 [1] では以下の問題がある. クラスタリングで重要な単語レベルの情報を文書から抽出する際に, 人手でクラスター数を決めて, 最適なクラスター数になっていない. 重要な列も人手で選択する必要がある. 最後にできた表の精度が低い. 岡崎らの研究 [2] では最適なクラスター数と列の重要度を計算して, 文レベルで重要な情報を表に整理するが, 文が長くなると, 重要な情報が見づらいのような場合がある. 本研究では最適なクラスター数と列の重要度の計算方法を使って, 重要な単語レベルの情報を文書から抽出して, 表に整理する.

3.2　提案手法の手順

従来手法の問題点を解決するため, 本研究ではクラスタリングを 2 回 (1 回目は文レベルでクラスタリング, 2 回目は単語レベルでクラスタリング) して, 岡崎らの研究 [2] で用いられる最適なクラスター数を計算する方法と Silhouette 法或いは UpperTail 法で最適なクラスタ数を推定し, 最適なクラスター数で作った表を最適な表とし, 最適な表の各列の重要度を計算して, 表にある全ての列を重要度で並べ替える. 提案手法の手順を以下に示す.

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　2.2 節の岡崎らの方法を用いて, 文ベクトルをクラスタリングする. 結果を表に整理する. 一部の結果の表を表 2.3 と表 2.4 に示す. 手順 1～ 順 3 を図 3.1 に示す.

手順 4　手順 3 でできた列を重要度で並べ替えた表の 1 番目から 6 番目の列を選択して, 列ごとに処理する. 処理方法として, これらの文を MeCab と termExtract を用いて, 単

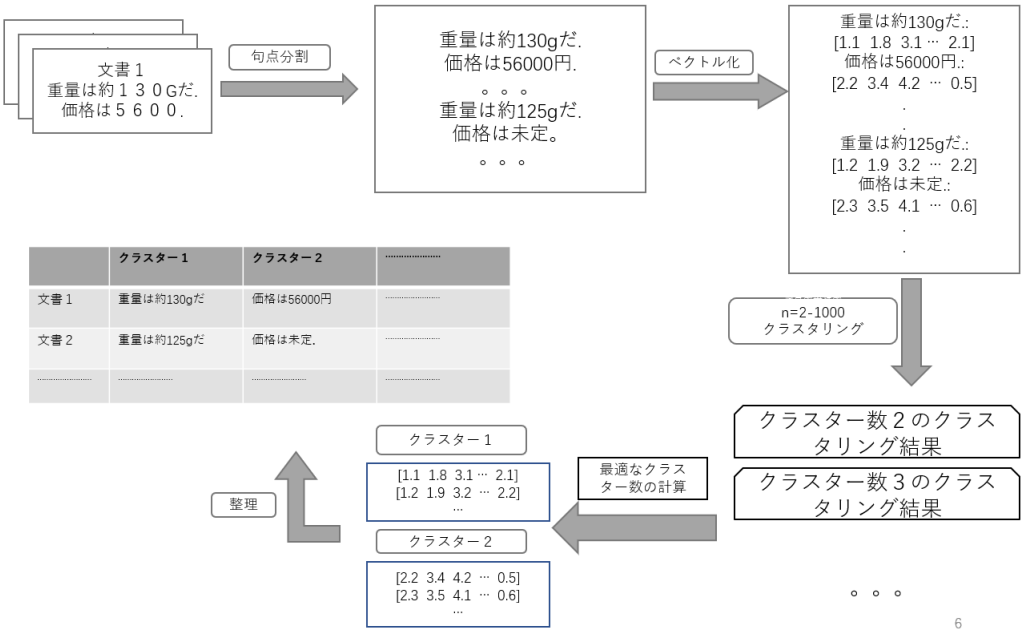

図 3.1: 手順１～手順３ (1 回目クラスタリング) の図

語レベルで分割する. 名詞単語以外の単語を削除する. 残った単語を FastText を用いて, ベクトル化する. これらの単語ベクトルを階層クラスタリング法を用いてクラスター数を 1~1,000 まで設定して複数回クラスタリングする. クラスタリング結果に基づき, 岡崎らの方法, Silhouette 法或いは UpperTail 法で最適なクラスター数を計算して, 最適なクラスター数で作ったクラスタリング結果を選ぶ. 手順 4 を図 3.2 に示す.

手順５　手順 4 で採用された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. そして, 従来手法 2.2.3 で紹介した列の重要度の計算方法を用いて, 列の重要度を計算して, 表の列を重要度で並べ替える. 地震データを用いて, 結果の一部を表 3.1 と表 3.2 に示す.

3.3　単語分割の手法

MeCab だけを使う場合, 文書を分割し過ぎることがよくある. 例えば, 専門用語の自然言語処理が含まれる文を MeCab で分割すると,「自然　言語　処理」で分割する. このような問題を解決するため, 連続している名詞単語を一つの名詞として扱い, かつ termextract を用いて, 専門用語を識別する. 専門用語の多くは単語を組み合わせて、複雑な概念を表すこ

表 3.1: 1 回目のクラスタリング結果の列 1 のデータをを用いて作った出力テーブル (岡崎らの方法で最適なクラスター数を計算した)

文書番号	列 1	列 2	列 3	...
文書 1	マグニチュード	50 キロ	さ, 5 ・ 0	...
文書 2	マグニチュード 4・4		4・4	...
文書 3	マグニチュード 3 ・4		3・4	...
文書 4	マグニチュード		5・8	...
文書 5	マグニチュード	10 キロ	さ,6・1	...
文書 6	マグニチュード	10 キロ	さ,4・7	...
文書 7	マグニチュード		4・6	...
文書 8		13 キロ	さ,5・2	...
文書 9	マグニチュード			...
文書 10	マグニチュード 3 ・ 7		3・7	...
文書 11	マグニチュード 5.5		3・7	...
文書 12	マグニチュード	60 キロ	さ,4・4	...
文書 13	マグニチュード		4・1	...
文書 14	マグニチュード		3・1	...
文書 15	マグニチュード	10 キロ	さ,4・7	...
文書 16	マグニチュード	9 キロ	さ,5・8	...
文書 17	マグニチュード	30 キロ	さ,5・5	...
文書 18	マグニチュード	20 キロ	さ,4・6	...
文書 19	マグニチュード	50 キロ	4・2	...
文書 20	マグニチュード	10 キロ	5・3	...

表 3.2: 1 回目のクラスタリング結果の列 2 のデータをを用いて作った出力テーブル (岡崎らの方法で最適なクラスター数を計算した)

文書番号	列 1	列 2	列 3	...
文書 1	20 日午前 7 時 25 分ごろ	震源, 地震	最大震度 4	...
文書 2	26 日午前 4 時 52 分ごろ	地震	震度 3	...
文書 3	27 日午前 3 時 28 分ごろ,7 時 57 分ごろ	地震	震度 3	...
文書 4	1 日午前 11 時 39 分ごろ	震源, 地震	震度 4	...
文書 5	1 日午前 6 時 33 分ごろ	地震	震度 4	...
文書 6	25 日午後 1 時 51 分ごろ長野県	震源, 地震	震度 4	...
文書 7	31 日午後 7 時 46 分ごろ	地震, 地震	震度 5 弱	...
文書 8	10 日	震源, 地震		...
文書 9	19 日午後 5 時 52 分ごろ	地震	震度 5 強, 震度 5 弱	...
文書 10	26 日午前 9 時 49 分ごろ	地震	震度 3	...
文書 11	26 日午後 2 時 13 分ごろ	震源, 地震	震度 4	...
文書 12	2 日午前 3 時 33 分ごろ	震源, 地震	震度 4	...
文書 13	22 日午前 7 時 59 分ごろ	震源, 地震	震度 3	...
文書 14	5 日午後 8 時 48 分	地震	震度 3	...
文書 15	22 日午後 2 時 34 分ごろ	震源, 地震	震度 4	...
文書 16	8 日午後 8 時 41 分ごろ	震源, 地震	震度 5	...
文書 17	5 日午前 7 時 41 分ごろ	震源, 地震	震度 4	...
文書 18	18 日午前 8 時 50 分ごろ	震源, 地震	震度 3	...
文書 19	27 日午後 11 時 47 分ごろ	震源, 地震	震度 5 弱	...
文書 20	16 日午後 9 時 23 分ごろ	震源, 地震	震度 5	...

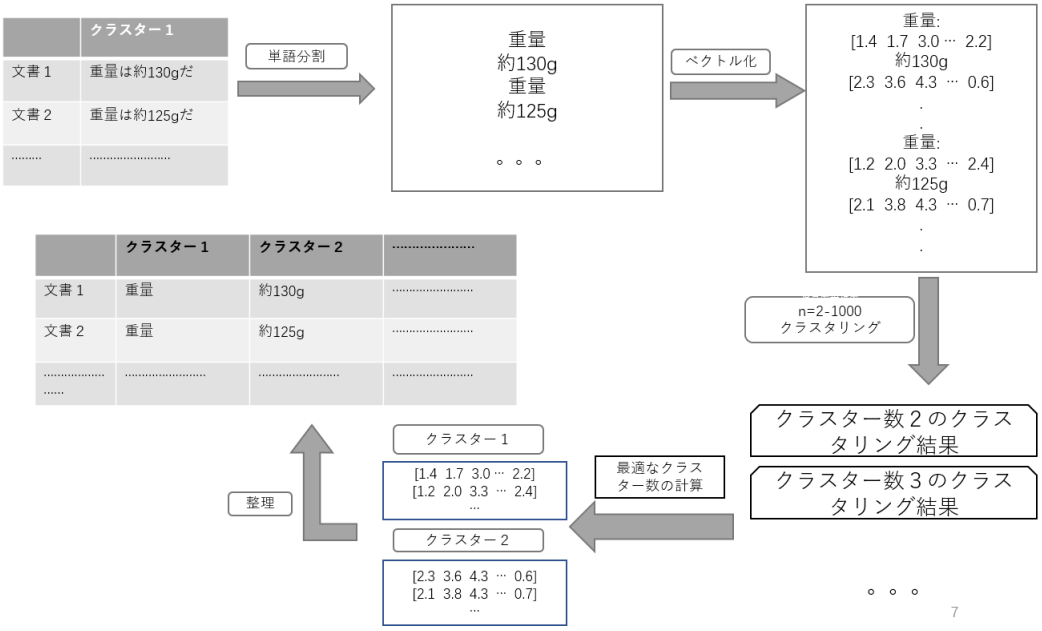

図 3.2: 手順４～手順５ (2 回目クラスタリング) の図

とが多くなる。特に「茶筅」の場合は単語を品詞単位で細かく分割するため、そのまま使うには難しい. その問題点を解決したソフトは termextract である.MeCab と termextract を用いて, 処理する例を以下で示す.

入力の例

自然言語処理は、人間が日常的に使っている自然言語をコンピュータに処理させる一連の技術である

出力の例

自然言語処理 人間 日常的 自然言語 コンピュータ 処理 一連 技術

3.4 Silhouette 法

事前に複数のクラスター数で複数回データをクラスタリングしたとする. Silhouette 法 [4] はデータの凝集性 (cohesion)a(x) と分離性 (sepeartion)b(x) を用いて, 最適なクラスター数を決める. まず, データ x の凝集性 a(x) の計算方法を紹介する. 数式 (a(x)) を式 2.1 に示す. $|C_X|$ はクラスター X にある単語の総数である. $d_{x,y}$ はデータ x とデータ y のユー

クリッド距離である.

$$a(x) = \frac{1}{|C_X| - 1} \sum_{y \in C_X, x \neq y} d_{x,y} \tag{3.1}$$

データ x の分離性 (sepeartion)b(x) の計算方法を紹介する. これはデータ x と他のクラスターにあるデータの最小平均距離である.

$$b(x) = min_{X \neq Y} \frac{1}{|C_Y|} \sum_{y \in C_Y} d_{x,y} \tag{3.2}$$

このデータ X の凝集性 $a(x)$ と分離性 $b(x)$ を用いて, データの silhouette 係数 s(x) を計算する.

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \tag{3.3}$$

すべでのデータに対して, この silhouette 係数の総和を計算して, この総和が最も大きクラスター数を最適なクラスタとして扱う. 地震データの 1 回目クラスタリング結果の列 2 を用いて, 提案手法を用いて, Silhouette 法で最適なクラスター数を計算した結果を表 3.3 に示す.

3.5 *UpperTail* 法

UpperTail 法 [5] は Mojena（1977）によって提案され, 統計的な停止規則を用いた階層的クラスタ分析におけるクラスタ数決定のための重要な方法である. これらの規則は, 基準値 α となる分布が N − 1 個あるのを利用する. ここではクラスタ間の距離のみが基準値として用いられ, 基準値の値は α_{n1}, すなわちクラスタが 2 つの場合から α_{n-1} 個まで取りうる. 停止規則は, クラスタ数（j）が 2 個から始めて増加させていき, $\alpha_j \leq \hat{\alpha} + k_{s_\alpha}$ の条件を満たすまで繰り返される. ここで,$\hat{\alpha}$ と s_α は, それぞれの基準値 α の分布の平均と不偏分散である. k は α の分布の平均と不偏分散に基づき, 上部棄却域を決める定数である. なお, Mojena（1977）においては, k の値は 60 から 120 までのデータ数に応じて 2 から 4 の値をとっている. 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. 地震データの 1 回目クラスタリング結果の列 2 を用いて, 提案手法を用いて, UpperTail 法で最適なクラスター数を計算した結果を表 3.4 に示す.

表 3.3: 1 回目のクラスタリング結果の列 1 のデータをを用いて作ったテーブル (Silhouette 法で最適なクラスター数を計算した)

文書番号	列 1	列 2	列 3	...
文書 1	20 日午前 7 時 25 分ごろ	震源, 地震	最大震度 4	...
文書 2	26 日午前 4 時 52 分ごろ	地震	震度 3	...
文書 3	27 日午前 3 時 28 分ごろ,7 時 57 分ごろ	地震	震度 3	...
文書 4	1 日午前 11 時 39 分ごろ	震源, 地震	震度 4	...
文書 5	1 日午前 6 時 33 分ごろ	地震	震度 4	...
文書 6	25 日午後 1 時 51 分ごろ長野県	震源, 地震	震度 4	...
文書 7	31 日午後 7 時 46 分ごろ	地震, 地震	震度 5 弱	...
文書 8	10 日	震源, 地震		...
文書 9	19 日午後 5 時 52 分ごろ	地震	震度 5 強, 震度 5 弱	...
文書 10	26 日午前 9 時 49 分ごろ	地震	震度 3	...
文書 11	26 日午後 2 時 13 分ごろ	震源, 地震	震度 4	...
文書 12	2 日午前 3 時 33 分ごろ	震源, 地震	震度 4	...
文書 13	22 日午前 7 時 59 分ごろ	震源, 地震	震度 3	...
文書 14	5 日午後 8 時 48 分	地震	震度 3	...
文書 15	22 日午後 2 時 34 分ごろ	震源, 地震	震度 4	...
文書 16	8 日午後 8 時 41 分ごろ	震源, 地震	震度 5	...
文書 17	5 日午前 7 時 41 分ごろ	震源, 地震	震度 4	...
文書 18	18 日午前 8 時 50 分ごろ	震源, 地震	震度 3	...
文書 19	27 日午後 11 時 47 分ごろ	震源, 地震	震度 5 弱	...
文書 20	16 日午後 9 時 23 分ごろ	震源, 地震	震度 5	...

表 3.4: 1 回目のクラスタリング結果の列 1 のデータをを用いて作った出力テーブル (UpperTail 法で最適なクラスター数を計算した)

文書番号	列 1	列 2	列 3	...
文書 1	20 日午前 7 時 25 分ごろ	震源, 地震	最大震度 4	...
文書 2	26 日午前 4 時 52 分ごろ	地震	震度 3	...
文書 3	27 日午前 3 時 28 分ごろ,7 時 57 分ごろ	地震	震度 3	...
文書 4	1 日午前 11 時 39 分ごろ	震源, 地震	震度 4	...
文書 5	1 日午前 6 時 33 分ごろ	地震	震度 4	...
文書 6	25 日午後 1 時 51 分ごろ長野県	震源, 地震	震度 4	...
文書 7	31 日午後 7 時 46 分ごろ	地震, 地震	震度 5 弱	...
文書 8	10 日	震源, 地震		...
文書 9	19 日午後 5 時 52 分ごろ	地震	震度 5 強, 震度 5 弱	...
文書 10	26 日午前 9 時 49 分ごろ	地震	震度 3	...
文書 11	26 日午後 2 時 13 分ごろ	震源, 地震	震度 4	...
文書 12	2 日午前 3 時 33 分ごろ	震源, 地震	震度 4	...
文書 13	22 日午前 7 時 59 分ごろ	震源, 地震	震度 3	...
文書 14	5 日午後 8 時 48 分	地震	震度 3	...
文書 15	22 日午後 2 時 34 分ごろ	震源, 地震	震度 4	...
文書 16	8 日午後 8 時 41 分ごろ	震源, 地震	震度 5	...
文書 17	5 日午前 7 時 41 分ごろ	震源, 地震	震度 4	...
文書 18	18 日午前 8 時 50 分ごろ	震源, 地震	震度 3	...
文書 19	27 日午後 11 時 47 分ごろ	震源, 地震	震度 5 弱	...
文書 20	16 日午後 9 時 23 分ごろ	震源, 地震	震度 5	...

第4章　実験環境

4.1　単語をベクトル化するツール

本研究では単語をベクトル化する必要がある. 本研究で用いる単語をベクトル化するツールは Fasttext である. Fasttext[6] は隠れ層が一つのニューラルネットワークである. 学習データは Wikipedia の全 1,061,375 記事である. 学習モデルは skip-gram で, ベクトルの次元数は 300 とした.

4.2　実験データ

本研究はクラスタリングを 2 回して, 単語レベルの重要な情報を文書から抽出する. 1 回目のクラスタリングは岡崎らの研究に基づいて, 文書を文レベルでクラスタリングする. 本実験は直接人手で 1 回目クラスタリング結果の正解テーブルを作って, これらの正解テーブルを 1 回目クラスタリングの結果として扱う. 人手で作った正解テーブルを用いて, 2 回目のクラスタリング実験だけを行う, 単語レベルの重要な情報を文書から抽出する. 地震の文書の例を図 4.2 に示す, 岡崎の正解テーブルを表 4.1 と表 4.2 に示す.

文書データの詳細を以下で示す.

- 1. 入力データ:強盗事件に関する新聞記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新聞記事の詳しい: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.

- 2. 入力データ:地震に関する新聞記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新聞記事の詳しい: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

地震の文章の例

<doc title=”熊本地震：水俣で震度３”>　熊本県・大分県を中心に続いている一連の地震で、１０日にこれまでの主な活動領域よりやや南西に離れたところで地震が発生した。
　気象庁によると、１０日午前４時４１分から５時７分ごろにかけて、熊本県天草・芦北地方を震源とする地震が断続的に計３回あり、水俣市で震度３、芦北町や天草市などで震度２を観測した。この地域を震源とした震度１以上の地震が観測されたのは先月１４日以降初めて。気象庁は熊本地震の活動の一つとしている。【円谷美晶】
……
　◇熊本県を中心とした地震の回数（回）
　カッコ内は１０日に起きた地震回数（１１日午前０時現在）
震度
７　　２　（０）
６強　　２　（０）
６弱　　３　（０）
５強　　４　（０）
５弱　　７　（０）
４　　８６　（０）
１～３　１２７０（１５）
計　　１３７４（１５）
</doc>

図 4.1: 処理結果の例

表 4.1: 地震での正解テーブル

文書番号	マグニチュード
文書 1	気象庁によると、震源の深さは約 50 キロ、地震の規模を示すマグニチュードは 5・0 と推定される
文書 2	地震の規模はマグニチュード 4．4 と推定される
文書 3	地震の規模はマグニチュード 3・4 と推定される
文書 4	27 日午前 3 時 28 分ごろ、長野県で震度 3 の地震があった
文書 5	気象庁によると、震源地は新潟県上越地方で、震源の深さは約 10 キロ
文書 6	
文書 7	気象庁によると、震源の深さは約 10 キロで、地震の規模を示すマグニチュードは 6・1 と推定される
文書 8	
文書 9	気象庁によると、震源は同県熊本地方、震源の深さは約 10 キロ、地震の規模を示すマグニチュードは 4・7 と推定される
文書 10	
…	…

表 4.2: 地震での正解テーブル

文書番号	発生日時	…
文書 1	20 日午前 7 時 25 分ごろ、茨城県南部を震源とする地震があり、さいたま市や水戸市などで最大震度 4 を観測した	…
文書 2	26 日午前 4 時 52 分ごろ、茨城、栃木、埼玉、千葉の各県で震度 3 の地震があった	…
文書 3		…
文書 4		…
文書 5		…
文書 6		…
文書 7	1 日午前 11 時 39 分ごろ、三重県南東沖を震源とする地震があり、和歌山県古座川町で震度 4 を観測した	…
文書 8		…
文書 9	1 日午前 6 時 33 分ごろ、熊本市西区・南区や熊本県宇城市、上天草市で震度 4 を観測する地震があった	…
文書 10		…
…	…	…

表 4.3: 文書データの詳しいの表

記事種類	記事数	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

- 3. 入力データ:交通事故に関する新聞記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新聞記事の詳しい: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

- 4. 入力データ:リコールに関する新聞記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新聞記事の詳しい: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

- 5. 入力データ:スマートフォンに関する新製品記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新製品記事の詳しい: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

- 6. 入力データ:スマートフォンに関する新製品記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新製品記事の詳しい: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

- 7. 入力データ:デジタルカメラに関する新製品記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新製品記事の詳しい: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

- 8. 入力データ:ロボット掃除機に関する新製品記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新製品記事の詳しい: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

- 9. 入力データ:エアコンに関する新製品記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  新製品記事の詳しい: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

- 10. 入力データ:城に関する Wikipedia の記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  Wikipedia の記事の詳しい: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

- 11. 入力データ:恐竜に関する Wikipedia の記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  Wikipedia の記事の詳しい: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

- 12. 入力データ:力士に関する Wikipedia の記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  Wikipedia の記事の詳しい: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

- 13. 入力データ:山に関する Wikipedia の記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  Wikipedia の記事の詳しい: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

- 14. 入力データ:野球チームに関する Wikipedia の記事 20 件に基づき, 人手で作った文レベルの正解テーブル (1 回目クラスタリングの結果として)
  Wikipedia の記事の詳しい: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

第5章　評価

5.1　balance F-Score

本実験で用いられる評価方法は balance F-score である. この方法は precision の値と recall の値の調和平均を計算することで手法の性能を示す方法である. precesion の意味は抽出されたデータの中に正解データの割合である. recall の意味は抽出された正解データと正解データの総数の割り算の結果である. precesion の計算方法は数式 5.1 で示す. recall の計算方法は数式 5.2 で示す.

$$precesion = \frac{\text{正解データと実験で抽出されるデータが一致しているデータの数}}{\text{実験で抽出されるデータの数}} \tag{5.1}$$

$$recall = \frac{\text{正解データと実験で抽出されるデータが一致しているデータの数}}{\text{正解データの総数}} \tag{5.2}$$

そして, この二つの値の調和平均 F1 を計算する. 数式 5.3 で示す.

$$F1 = \frac{2 * precesion * recall}{precesion + recall} \tag{5.3}$$

この f1 値が高ければ高い程, 性能が高いことを意味する.

本実験の場合, precesion と recall の分子の部分は実験で生成されたテーブルと正解テーブルの対応する欄に一致しているデータの総数である. precesion の分母は実験で生成されたテーブルのデータの総数, recall の分母は正解テーブルのデータの総数である. 具体例は表 5.1 で示す. この具体例の実験データの部分は毎日新聞の地震記事から抽出した場所に関するデータである. 正解データは人手で毎日新聞の地震記事がら場所に関する単語を抽出する. 表 1 の F1 値を計算すると, precesion の結果は 0.7（14 ／ 20）である. recall の結果は 0.46（14/30）である. f1 の結果は 0.55 である.

5.2　正解テーブルの作り方

本実験は 2 回クラスタリングをすることで（ 1 回目は文レベルのクラスタリング, 2 回目は単語クラスタリング）, 重要な情報を文書から抽出する. 一つの正解テーブルは 1 回

表 5.1: 地震記事評価例

文書番号	実験で抽出されるデータ	正解表のデータ	一致しているテータの数
文書 1	茨城県南部 水戸市	茨城県南部 さいたま市 水戸市	2
文書 2	県	茨城 栃木 埼玉 千葉	0
文書 3	長野県 県	長野県 宮城 福島 茨城 栃木	1
文書 4	三重県南東沖 和歌山県古座木川町	三重県南東沖 和歌山県古座川町	2
文書 5	熊本県熊本地方 熊本市西区 熊本県宇城市 上天草市	熊本市西区 南区 熊本県宇城市 上天草市	3
文書 6		長野県の小谷村 小川村	0
文書 7	熊本県熊本地方 熊本市西区 熊本県宇城市	熊本県熊本地方 熊本市西区 熊本県宇城市	3
文書 8	熊本県 大分県 南西	熊本県, 大分県	2
文書 9	熊本県熊本地方 県八代市	熊本県熊本地方 同県八代市 同県氷川町 同県芦北町	1
文書 10	高知県	高知県	1
総数	20	30	14

目のクラスタリングの結果の表の一列（この列は人手で選択する）に基づいて, 人手で作成する. 人手でこの列について重要な情報の種類を考えて, 人が重要と思うデータの種類と関連しているデータを列から抽出して, 正解テーブルを作る. 具体例を示す. 正解テーブルを作る時の根拠, 1 回目のクラスタリングの結果の表を表 5.2 と表 5.3 に示す. この 1 回目のクラスタリングの表に基づいて作った正解テーブルを表 5.4 と表 5.5 に示す. 具体例の実験データは毎日新聞から抽出した交通事故に関するデータである.

この表 5.2 と表 5.3 は毎日新聞の地震記事を利用して, 1 回目のクラスタリング（文レベル）の結果の一部である. 表 5.2 を使って作った表は表 5.4 に示す. 表 5.3 を使って作った表は表 5.5 に示す.

表 5.4 は表 5.2 に基づいて作った正解テーブルである. 表 5.2 は文書から事故の発生時間や場所と車のタイプなどの情報が含まれる文を抽出した. 正解テーブルを作る時, これらの重要な情報の種類（事故の発生時間や場所と車のタイプ）を考えて, 表 5.4 を作った. 表 5.5 は表 5.3 に基づいで作った正解テーブルである. 表 5.5 は文書から容疑者の名前や罪の名前などを含む文を文書から抽出した. 正解テーブルを作る時, これらの重要な情報の種類（容疑者の名前や罪の名前）を考えて, 表 5.5 を作った.

5.3 評価手順

評価手順を以下に示す.

手順 1　作成した正解テーブル[1]の各列に注目する.

手順 2　注目している列と出力デーブル[2]の各列の F1 値を計算する.

手順 3　手順 2 で計算した結果の中で最も高い数値をこの注目している列の F1 値として扱う.

手順 4　手順 2 と手順 3 の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, これを出力テーブルの評価結果とする.

[1]人手で実験データに基づく作った正解のテーブルを正解テーブルと呼ぶ.
[2]2 回クラスタリングすることで機械で抽出された重要な情報が含むテーブルを出力テーブルと呼ぶ.

表 5.2: 交通事故に関する 1 回目のクラスタリング結果 (列 1)

文書番号	列 1
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
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 5.3: 交通事故に関する 1 回目のクラスタリングの結果 (列 2)

文書番号	列 2
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	
文書 6	
文書 7	
文書 8	
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	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	
文書 13	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	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	
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	

表 5.4: 1 回目のクラスタリングの列 1 に基づく正解テーブル

文書番号	時間	場所	車
文書 1	29 日午後 1 時 5 分ごろ	愛知県北名古屋市	乗用車, 軽乗用車
文書 2	8 日午前 8 時ごろ	埼玉県上里町嘉美	バス, 軽乗用車
文書 3	28 日午前 8 時ごろ	横浜市港南区大久保 1	車
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	軽乗用車
文書 5	9 日夜	香川県内	軽乗用車
文書 6	9 日午後 3 時 40 分ごろ	広島県庄原市東城町	乗用車
文書 7	1 日午前 0 時 50 分ごろ	栃木市都賀町家中	乗用車
文書 8	2 日午前 2 時 10 分ごろ	北海道室蘭市東町 5	乗用車
文書 9	26 日午前 5 時 45 分ごろ	大阪府寝屋川市池田北町	自転車
文書 10	20 日午後 7 時ごろ	東京都大田区蒲田本町 1	観光バス
文書 11	26 日午前 6 時 40 分ごろ	大阪市旭区中宮 1	車
文書 12	4 日午後 9 時半ごろ	大阪市住吉区万代東 3	大阪市営バス
文書 13	8 日午後 9 時 55 分ごろ	香川県観音寺市柞田町	大型トレーラー
文書 14	2 日午前 2 時 5 分ごろ	愛知県岡崎市駒立町	観光バス, 大型トラック
文書 15	12 日午後 5 時ごろ	兵庫県宝塚市小浜 2	乗用車
文書 16	16 日午後 3 時半ごろ	奈良県川上村大迫	ワゴン車, 乗用車
文書 17	8 日午前 2 時 45 分ごろ	兵庫県西宮市浜脇町	中型トラック, 大型トレーラー
文書 18	8 日午前 7 時 55 分ごろ	静岡県磐田市中泉	ライトバン
文書 19	10 日午前 8 時 45 分ごろ	大阪府島本町山崎	路線バス, 大型トラック
文書 20	26 日午前 9 時半ごろ	大津市蛍谷	高速バス, トラック

交通事故の実験データで 1 回クラスタリングの 1 番目の列（表 5.2）を利用して, 提案手法を使って実験で生成されたテーブルを表 5.6 に示す. 表 5.6 と表 5.4 を例として, F1 の計算方法を以下で示す.

正解テーブル（表 5.4）の列 1 と出力テーブル（表 5.6）の列 1 , 列 2 , 列 3 , 列 4, 列 5, 列 6 の 6 個の F1 値を計算して, この中に数値が最も高い数値を正解テーブルの列 1 の f 1 値として扱います. 同じ手順で正解テーブル（表 5.4）の列 2 と出力テーブル（表 5.6）の列 1 , 列 2 , 列 3 , 列 4, 列 5, 列 6 を使って, 正解テーブル（表 5.4）の列 2 の F1 値を計算する. 同じ手順で正解テーブルの全部の列の F1 値を計算することができる. これらの値の平均値を正解テーブル（表 3 ）の F1 値として扱う.

1 回目のクラスタリング結果の列の F1 値を計算することができる. これらの F1 値のなかに最も高い三つの F1 値の平均値をこの実験データの F1 値として扱う. すべでのデータの F1 結果を表 5.7 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

表 5.5: 1 回目のクラスタリングの列 2 に基づく正解テーブル

文書番号	名前	罪
文書 1	大口久美子容疑者	過失致死傷容疑
文書 2		
文書 3		自動車運転処罰法
文書 4		
文書 5		
文書 6		
文書 7		
文書 8		
文書 9	南隆樹容疑者	自動車運転処罰法
文書 10	菅原正容疑者	自動車運転処罰法
文書 11		
文書 12		
文書 13	大川貴之容疑者	自動車運転処罰法
文書 14	斎藤信夫容疑者	自動車運転処罰法
文書 15		
文書 16		
文書 17		
文書 18	河合秀幸容疑者	自動車運転処罰法
文書 19		
文書 20		

表 5.6: 1 回目のクラスタリングの列 1 に基づく出力テーブル

文書番号	列 1	列 2	列 3	...
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	乗用車, 乗用車	...
文書 2	8 日午前 8 時ごろ	埼玉県上里町嘉美	乗用車	...
文書 3	28 日午前 8 時ごろ	横浜市港南区大久保 1	トラック	...
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	乗用車	...
文書 5	9 日夜	香川県内	乗用車 乗用車 乗用車 トラック	...
文書 6	9 日午後 3 時 40 分ごろ	広島県庄原市東城町 大阪市平野区	乗用車	...
文書 7	1 日午前 0 時 50 分ごろ	栃木市都賀町家中 栃木都賀ジャン... 愛知県稲沢市	乗用車, トラック	...
文書 8	2 日午前 2 時 10 分ごろ	北海道室蘭市東町 5	乗用車	...
文書 9	26 日午前 5 時 45 分ごろ	大阪府寝屋川市池田北町	大型トラック	...
文書 10	20 日午後 7 時ごろ	東京都大田区蒲田本町 1		...
文書 11	26 日午前 6 時 40 分ごろ	大阪市旭区中宮 1	車	...
文書 12	4 日午後 9 時半ごろ	大阪市住吉区万代東 3		...
文書 13	8 日午後 9 時 55 分ごろ	香川県観音寺市柞田町	大型トレーラー	...
文書 14	2 日午前 2 時 5 分ごろ	愛知県岡崎市駒立町	大型トラック	...
文書 15	12 日午後 5 時ごろ	兵庫県宝塚市小浜 2	乗用車	...
文書 16	16 日午後 3 時半ごろ	奈良県川上村大迫	ワゴン車, 乗用車	...
文書 17	8 日午前 2 時 45 分ごろ	兵庫県西宮市浜脇町 県南 市	中型トラック 大型トレーラー トラック	...
文書 18	8 日午前 7 時 55 分ごろ	静岡県磐田市中泉 市立磐田中部小学校 2 年	ライトバン	...
文書 19	10 日午前 8 時 45 分ごろ	大阪府島本町山崎	大型トラック	...
文書 20	26 日午前 9 時半ごろ	大津市蛍谷	トラック	...

は 0.05 とした. 有意差検定の結果を表 5.8 に示す.

表 5.7: 性能評価

データセット	提案手法 (クラスター数計算方法: 岡崎ら)	提案手法 (クラスター数計算方法: silhouette)	提案手法 (クラスター数計算方法: UpperTail)	従来手法 赤野ら
強盗	0.67	0.63	0.61	0.21
力士	0.64	0.61	0.51	0.25
山	0.74	0.72	0.67	0.39
スマホ	0.67	0.67	0.64	0.12
城	0.70	0.68	0.80	0.33
地震	0.82	0.75	0.81	0.14
野球チーム	0.32	0.37	0.34	0.11
ロボット掃除機	0.63	0.62	0.58	0.10
恐竜	0.64	0.62	0.61	0.18
エアコン	0.63	0.70	0.73	0.06
カメラ	0.63	0.54	0.52	0.29
テレビ	0.62	0.62	0.61	0.11
交通事故	0.90	0.73	0.69	0.24
外為	0.81	0.71	0.52	0.26
リコール	0.71	0.68	0.63	0.32
平均値	0.68	0.64	0.62	0.21

表 5.8: 有意差検定の結果

データセット	提案手法 (クラスター数計算方法: 岡崎ら	提案手法 (クラスター数計算方法: silhouette)	提案手法 (クラスター数計算方法: UpperTail)	従来手法 赤野ら
岡崎ら		0.003	0.017	0.000
silhouette			0.996	0.000
UpperTail				0.000
赤野ら				

第6章　考察

6.1　従来手法との比較

従来手法では Wikipedia 全データを用いて, 人手でクラスタスー数 2000 を設定して k−means 法を利用してクラスタリングする. このクラスタリングの結果と処理したい文書を比較して, 重要な情報が含まれている表を作る. この方法の欠点は直接文書のデータでクラスタリングするのではなくて, Wikipedia 全データでクラスタリングする. クラスタリング数も事前に設定し, 最適なクラスター数になっていない可能性がある. そして, 列の重要性が列にある単語の延数で判断する. この方法で列の重要性を判断すると, 重要でない列を重要と判断する可能性が高い. 最終的に作成された表に空欄が多い, 重要でない列が多く現れる, 一つの列に一つの単語だけが複数回現れるといった欠点がある. 本研究では Wikipedia 全データを用いてクラスタリングするのではなくて, 直接文書データを用いてクラスタリングする. クラスタリングする時も最適なクラスタリング数でクラスタリングする. 列の重要性もカバー率と密集度で計算する.

本研究では F 値で結果を評価する. F 値は適合率と再現率の調和平均である. 結果を見ると, 従来手法の赤野らの手法では精度が全体的に低い. 最低値が 0.10, 最高値が 0.39, 平均は 0.19 である. Silhouette 法で最適なクラスター数を計算して, クラスタリングの結果では最低値 0.37, 最高値は 0.75, 平均は 0.63 である, UpperTail 法の最低値, 最高値, 平均値は 0.34,0.80,0.63 である. 岡崎の情報のカバー率と密集度で最適なクラスター数を計算する方法では最低値, 最高値, 平均値が 0.32,0.90,0.66 である. 従来手法の方法と提案手法の最適なクラスター数と列の重要度を用いて重要な情報を抽出する方法と比べて, 全体的に性能が低い.

6.2　最適なクラスター数を計算する三つの方法の間の比較

同じデータ (カメラ, エアコン) を用いて, 三つ最適なクラスター数を計算する方法を用いて作った表を以下に示す. エアコンのデータを三つの方法を用いて処理した結果は表

6.1, 表 6.2, 表 6.3 に示す. カメラのデータを三つの方法を用いて処理した結果は表 6.4, 表 6.5, 表 6.6, 表 6.7, 表 6.8, 表 6.9, 表 6.10 に示す. これらの結果を見ると, この三つの方法の結果がよく似ている. 三つの方法で計算する最適なクラスター数がほぼ同じのが原因である. 三つの方法の計算結果がほぼ同じ原因は真の最適なクラスター数に近いためかも知れない. 重要な情報が含まれる列も抽出されていた. 二つ目の原因は 2 回目のクラスタリングの入力データは文の特徴に関係し, 文に含まれる単語の数が少ないのが原因の一つかも知れない.

表 6.1: 出力テーブル 1 (データ:エアコン (列:メーカー発表列))(最適なクラスター数計算方法:岡崎ら)

文書番号	列 1	列 2	列 3	...
文書 1	発表	ｎｏｃｒｉａＸＳシリーズ	エアコン	...
文書 2	発表		ルームエアコン	...
文書 3	発表	シリーズ 6 機種 シリーズ 5 機種 シリーズ 6 機種 シリーズ 6 機種	家庭用ルームアコン	...
文書 4	発表		窓用ルームエアコン	...
文書 5	発表	シリーズ	エアコン	...
文書 6	発表	Ｘシリーズ	ルームエアコン	...
文書 7	発表	シリーズ 17 モデル	家庭用ルームエアコン	...
文書 8	発表	シリーズ	エアコン	...
文書 9	発表	シリーズ	ルームエアコン	...
文書 10	発表	シリーズ	ルームエアコン	...
文書 11	発表	Ｘシリーズ	ルームエアコン	...
文書 12	発表	Ｘシリーズ		...
文書 13	発表	シリーズ	ルームエアコン	...
文書 14	発表	Ｒシリーズ, Ｐシリーズ	ルームエアコン	...
文書 15	発表	ＵＸシリーズ	住宅用マルチエアコン	...
文書 16	発表	Ｚシリーズ	ルームエアコン	...
文書 17	発表	Ｘシリーズ		...
文書 18	発表	シリーズ	ルームエアコン	...
文書 19	発表		人感センサー機能ルームエアコン ルームエアコン	...
文書 20	発表	ｎｏｃｒｉａＸＳシリーズ		...

表 6.2: 出力テーブル 2(データ:エアコン (列:メーカー発表列))(最適なクラスター数計算方法:UpperTail)

文書番号	列 1	列 2	列 3	...
文書 1	発表	2018 年 1 月 6 日	ｎｏｃｒｉａＸＳシリーズ	...
文書 2	発表	11 月 1 日		...
文書 3	発表	3 月下旬		...
文書 4	発表	11 月 18 日		...
文書 5	発表		シリーズ	...
文書 6	発表	3 月上旬	シリーズ 17 モデル	...
文書 7	発表	10 月下旬	シリーズ	...
文書 8	発表	2 月上旬,3 月中旬,4 月上旬	シリーズ	...
文書 9	発表	10 月下旬	シリーズ	...
文書 10	発表	10 月 25 日		...
文書 11	発表	3 月上旬		...
文書 12	発表	2 月上旬,4 月上旬	シリーズ	...
文書 13	発表	10 月 1 日		...
文書 14	発表	10 月下旬		...
文書 15	発表	2018 年 1 月 25 日		...
文書 16	発表	11 月 1 日,12 月中旬,12 月下旬	シリーズ	...
文書 17	発表	5 月 2 日		...
文書 18	発表	7 月 30 日	シリーズ	...
文書 19	発表			...
文書 20	発表	ｎｏｃｒｉａＸＳシリーズ		...

表 6.3: 出力テーブル 3 (データ:エアコン (列:メーカー発表列))(最適なクラスター数計算方法:silhouette)

文書番号	列 1	列 2	列 3	...
文書 1	発表	2018 年 1 月 6 日	ｎｏｃｒｉａＸＳシリーズ	...
文書 2	発表	11 月 1 日		...
文書 3	発表	3 月下旬		...
文書 4	発表	11 月 18 日		...
文書 5	発表		シリーズ	...
文書 6	発表	3 月上旬	シリーズ 17 モデル	...
文書 7	発表	10 月下旬	シリーズ	...
文書 8	発表	2 月上旬,3 月中旬,4 月上旬	シリーズ	...
文書 9	発表	10 月下旬	シリーズ	...
文書 10	発表	10 月 25 日		...
文書 11	発表	3 月上旬		...
文書 12	発表	2 月上旬,4 月上旬	シリーズ	...
文書 13	発表	10 月 1 日		...
文書 14	発表	10 月下旬		...
文書 15	発表	2018 年 1 月 25 日		...
文書 16	発表	11 月 1 日,12 月中旬,12 月下旬	シリーズ	...
文書 17	発表	5 月 2 日		...
文書 18	発表	7 月 30 日	シリーズ	...
文書 19	発表			...
文書 20	発表	ｎｏｃｒｉａＸＳシリーズ		...

表 6.4: 出力テーブル 4(データ:カメラ (列:メーカー発表列))(最適なクラスター数計算方法:岡崎ら)

文書番号	列 1	列 2
文書 1		ＯＭＮＩｓｈｏｔＯＣＡＭ
文書 2	発表	
文書 3	発表	ＰｏｗｅｒＳｈｏｔＧ, ＰｏｗｅｒＳｈｏｔＧ１ＸＭａｒｋＩＩＩ
文書 4	発表	ｚＥＹＥ,1
文書 5	発表	
文書 6	発表	コンスーマーエレクトロニクスショー, これ
文書 7	発表	コンスーマーエレクトロニクスショー
文書 8	発表	ＥＸ
文書 9	発表	ニコンイメージングジャパン, ＣＯＯＬＰＩＸＷ 300
文書 10	発表	タフネスコンパクトデジタルカメラ, ＯＬＹＭＰＵＳＴｏｕｇｈＴＧ,5
文書 11	発表	ＲＩＣＯＨＷＧ,50
文書 12	発表	ＬＵＭＩＸＤＣ, ＬＵＭＩＸＤＭＣ
文書 13	発表	ＰｏｗｅｒＳｈｏｔＳＸ 730 ＨＳ
文書 14	発表	ＲＩＣＯＨＴＨＥＴＡＳＣＴｙｐｅＨＡＴＳＵＮＥＭＩＫＵ
文書 15	発表	ＲＩＣＯＨＴＨＥＴＡ, ＲＩＣＯＨＴＨＥＴＡＶ, こと
文書 16	発表	ＲＩＣＯＨＴＨＥＴＡ, ＲＩＣＯＨＴＨＥＴＡＶ
文書 17	発表	ＲＩＣＯＨＴＨＥＴＡ, ＮＡＢＳＨＯＷ 2017
文書 18		400, サンワダイレクト
文書 19		タフネスカメラ, ＲＩＣＯＨＷＧ,50
文書 20		Ｒ 210 ＮＺＷＡＸＪＰ

表 6.5: 出力テーブル 4(データ:カメラ (列:メーカー発表列))(最適なクラスター数計算方法:岡崎ら)

文書番号	列 3	...
文書 1	方位 360 度撮影,4 K対応ＶＲカメラ	...
文書 2		...
文書 3	プレミアムコンパクトカメラ	...
文書 4	撮影, タフカメラ	...
文書 5	デジタルカメラ	...
文書 6		...
文書 7	小型カメラ	...
文書 8	デジタルカメラ	...
文書 9	コンパクトデジタルカメラ	...
文書 10		...
文書 11	コンパクトデジタルカメラ	...
文書 12	光学 30 倍ズームレンズ, コンパクトデジタルカメラ	...
文書 13	コンパクトデジタルカメラ	...
文書 14	360 度カメラ	...
文書 15	360 度カメラ	...
文書 16	360 度カメラ	...
文書 17	60 度, 撮影, 高画質,360 度,4 K動画撮影	...
文書 18	天球 360 度カメラ	...
文書 19		...
文書 20	360 度カメラ	...

表 6.6: 出力テーブル 5(データ:カメラ (列:メーカー発表列))(最適クラスター数計算方法:UpperTail)

文書番号	列 1
文書 1	
文書 2	発表
文書 3	発表,12 月 8 日
文書 4	発表
文書 5	発表
文書 6	ドイツ, ベルリン, 開催, ＩＦＡ 2017, 発表, 正式発表
文書 7	ドイツ, ベルリン, 現地時間 9 月 1 日, 開催, ＩＦＡ 2017, 発表
文書 8	発表
文書 9	発表
文書 10	6 月下旬発売, 発売日, 決定
文書 11	発表
文書 12	発表
文書 13	発表
文書 14	発表
文書 15	10 周年, 記念, 期間限定, 発表
文書 16	発表
文書 17	発表, 主催
文書 18	
文書 19	
文書 20	

表 6.7: 出力テーブル 5(データ:カメラ (列:メーカー発表列))(最適クラスター数計算方法:UpperTail)

文書番号	列 2
文書 1	ＯＭＮＩｓｈｏｔＯＣＡＭ
文書 2	
文書 3	ＰｏｗｅｒＳｈｏｔＧ, ＰｏｗｅｒＳｈｏｔＧ１ＸＭａｒｋＩＩＩ
文書 4	ｚＥＹＥ,1
文書 5	
文書 6	コンスーマーエレクトロニクスショー
文書 7	コンスーマーエレクトロニクスショー
文書 8	ＥＸ
文書 9	ニコンイメージングジャパン, ＣＯＯＬＰＩＸＷ 300
文書 10	タフネスコンパクトデジタルカメラ, ＯＬＹＭＰＵＳＴｏｕｇｈＴＧ,5
文書 11	ＲＩＣＯＨＷＧ,50
文書 12	ＬＵＭＩＸＤＣ, ＬＵＭＩＸＤＭＣ
文書 13	ＰｏｗｅｒＳｈｏｔＳＸ 730 ＨＳ
文書 14	ＲＩＣＯＨＴＨＥＴＡＳＣＴｙｐｅＨＡＴＳＵＮＥＭＩＫＵ
文書 15	ＲＩＣＯＨＴＨＥＴＡ, ＲＩＣＯＨＴＨＥＴＡＶ, こと
文書 16	ＲＩＣＯＨＴＨＥＴＡ, ＲＩＣＯＨＴＨＥＴＡＶ
文書 17	ＲＩＣＯＨＴＨＥＴＡ, ＮＡＢＳＨＯＷ 2017
文書 18	400, サンワダイレクト
文書 19	タフネスカメラ, ＲＩＣＯＨＷＧ,50
文書 20	Ｒ 210 ＮＺＷＡＸＪＰ

表 6.8: 出力テーブル 5(データ:カメラ (列:メーカー発表列))(最適クラスター数計算方法: 方法 UpperTail)

文書番号	列 3	...
文書 1	方位 360 度撮影,4 K対応ＶＲカメラ	...
文書 2		...
文書 3	プレミアムコンパクトカメラ	...
文書 4	撮影, タフカメラ	...
文書 5	デジタルカメラ	...
文書 6		...
文書 7	小型カメラ	...
文書 8	デジタルカメラ	...
文書 9	コンパクトデジタルカメラ	...
文書 10		...
文書 11	コンパクトデジタルカメラ	...
文書 12	光学 30 倍ズームレンズ, コンパクトデジタルカメラ	...
文書 13	コンパクトデジタルカメラ	...
文書 14	360 度カメラ	...
文書 15	360 度カメラ	...
文書 16	360 度カメラ	...
文書 17	60 度, 撮影, 高画質,360 度,4 K動画撮影	...
文書 18	天球 360 度カメラ	...
文書 19		...
文書 20	360 度カメラ	...

表 6.9: 出力テーブル 6 (データ:カメラ (列:メーカー発表列))(最適なクラスター数計算方法:方法:silhouette)

文書番号	列 1
文書 1	
文書 2	発表
文書 3	発表,12 月 8 日
文書 4	発表
文書 5	発表
文書 6	ドイツ, ベルリン, 開催, ＩＦＡ 2017, 発表, 正式発表
文書 7	ドイツ, ベルリン, 現地時間 9 月 1 日, 開催, ＩＦＡ 2017, 発表
文書 8	発表
文書 9	発表
文書 10	6 月下旬発売, 発売日, 決定
文書 11	発表
文書 12	発表
文書 13	発表
文書 14	発表
文書 15	10 周年, 記念, 期間限定, 発表
文書 16	発表
文書 17	発表, 主催
文書 18	
文書 19	
文書 20	

表 6.10: 出力テーブル 6 (データ:カメラ (列:メーカー発表列))(最適なクラスター数計算方法:方法:silhouette)

文書番号	列 2	列 3	...
文書 1			...
文書 2			...
文書 3			...
文書 4			...
文書 5			...
文書 6			...
文書 7		小型カメラ	...
文書 8		デジタルカメラ	...
文書 9			...
文書 10	リコー	イメージング	...
文書 11	リコー		...
文書 12			...
文書 13	リコー		...
文書 14	リコー	リコーイメージング	...
文書 15	リコー	リコーイメージング	...
文書 16		リコーイメージング	...
文書 17	リコー	60 度, 撮影, 高画質,360 度,4 K動画撮影	...
文書 18		リコーイメージング	...
文書 19			...
文書 20			...

6.3 正解テーブルにはない重要な列

本研究では人手で正解テーブルを作って, 人手で作った正解テーブルと実験で生成されたテーブルを比較して,F 値で提案手法の性能を評価する. 正解テーブルにはないが, 実験で生成されたテーブルを見ると, 重要と思う列があるかも知れない. 例えばエアコンのデータで正解テーブルを作る時,「時間」,「シリーズ」,「数」 3 種類のデータが重要と思うので, この 3 種類のデータに基づいて, 正解テーブルを作った. 実験で生成されたテーブルを見ると,「エアコンの種類」も重要なデータかも知れない.

正解テーブルにはないが, 実験で生成されたテーブルを見ると重要と思う列の数を表 6.11 に整理する. 列が Silhouette, 行がエアコンの欄にある 1 の意味は aircon のデータで作った正解テーブルにはないが, 実験で生成されたテーブルを見ると, 新たに重要と思う列の数が 1 であることを意味する.

表 6.11: 正解テーブルにはない重要な列の数の表

データ種類	岡崎ら	UpperTail	silhouette
エアコン	0	2	1
強盗	0	0	0
カメラ	0	0	0
ロボット掃除機	0	0	0
恐竜	1	1	1
地震	1	1	1
野球チーム	0	0	0
山	0	0	0
城	0	0	0
スマホ	0	0	0
力士	0	0	0
交通事故	1	0	0
テレビ	0	0	0
リコール	2	1	2
外為	1	1	1
平均	0.4	0.4	0.4

6.4　F 値が低いの原因

表 5.7 を見ると, 野球チームの F 値の評価結果が低い. その原因は同じ種類のデータが同いクラスターに分類されていない. 野球チームデータの実験で生成されたテーブルの一部を表 6.12 と表 6.13 に示す. 実験で生成されたテーブルを見ると, 列 1 と列 3 が同じクラスターに分類するのはずである. データの詳細を見ると, 列 1 は野球チームの傘下に関する情報である. 列 3 も野球チームの傘下に関する情報であるが,「A 級チーム」などの情報もある. 単語を間違って分割するのがデータが上手くクラスタリングされていないのは原因と考えられる. 列 2 と列 4 のデータも同じクラスターに分類するのはずであるが, 実験で生成されたテーブルには同じクラスターに分類されていない. 実験で決めたクラスター数が最適なクラスター数より多いのが原因と考えられる.

表 6.12: 野球チームの出力テーブル (1)

データ種類	列 1	列 2	列 3	列 4
文書 1	コロラド・ロッキーズ傘下			
文書 2	タンパベイ・レイズ傘下			
文書 3	ＭＬＢフィラデル フィア・フィリーズ傘下			
文書 4	カンザスシティ・ロイヤルズ傘下			ミッドウェストリーグ
文書 5		メジャーリーグ	セントルイス・カージナルス傘下Ａ級チーム	
文書 6	コロラド・ロッキーズ傘下			サウス・アトランティックリーグ
文書 7		メジャーリーグ	テキサス・レンジャーズ傘下Ａ級チーム	
文書 8	ジャイアンツ傘下			
文書 9	レッズ傘下			
文書 10	セントルイス・カージナルス傘下			
文書 11	ダイヤモンドバックス傘下			サウス・アトランティック・リーグ
文書 12		メジャーリーグ	ピッツバーグ・パイレーツ傘下Ａ級チーム	
文書 13	メッツ傘下			
文書 14	ミルウォーキー・ブルワーズ傘下			ミッドウェストリーグ東部地区
文書 15	トロント・ブルージェイズ傘下		Ａ級チーム	パシフィック・コーストリーグ

表 6.13: 野球チームの出力テーブル (2)

データ種類	列 1	列 2	列 3	列 4
文書 16		メジャーリーグ	カブス傘下ＡＡＡ級チーム	サザンリーグ
文書 17		メジャーリーグ	シカゴ･ホワイトソックス傘下ＡＡ級チーム	パ シ フィック・コーストリーグ
文書 18		メジャーリーグ	オークランド･アスレチックス傘下ＡＡＡ級チーム	パ シ フィック・コーストリーグ
文書 19		メジャーリーグ	ドジャース傘下ＡＡＡ級チーム	
文書 20			傘下チーム	

第7章　追加実験

本研究では人手で正解テーブルを作って, F 値で提案手法の有効性を確認する. そして, 正解テーブルにはない提案手法で発見された重要な列もあるかもしれない. 追加実験で,15 種類のデータを用いて, 事前に正解テーブルを作るではなくて, 実験で生成されたテーブルを見てから, 正解テーブルを作って, F 値で提案手法の性能を評価する. その結果は表 7.1 に示す.

表 7.1 を見ると,「地震」「恐竜」「エアコン」「交通事故」「リコール」「外為」の F 値の結果では元と比べて, 高くなった.「テレビ」「野球チーム」「城」「山」「力士」「カメラ」「強盗」「カメラ」「ロボット掃除機」の結果は元と同じである. 原因は結果の表を見て, 正解テーブルにない重要な列がひとつもないである.

表 7.1: 追加実験での性能評価

データセット	提案手法 (クラスター数計算方法: 岡崎ら)	提案手法 (クラスター数計算方法: silhouette)	提案手法 (クラスター数計算方法: UpperTail)
強盗	0.67	0.63	0.61
地震	0.89	0.82	0.84
ロボット掃除機	0.63	0.62	0.58
恐竜	0.65	0.71	0.64
エアコン	0.77	0.78	0.73
カメラ	0.63	0.54	0.52
力士	0.64	0.61	0.51
山	0.74	0.72	0.67
スマホ	0.67	0.67	0.64
城	0.70	0.68	0.80
野球チーム	0.32	0.37	0.34
テレビ	0.62	0.62	0.61
交通事故	0.92	0.74	0.71
外為	0.84	0.78	0.53
リコール	0.71	0.68	0.63
平均値	0.69	0.66	0.63

第8章　おわりに

過去の単語レベルで重要な情報を抽出する赤野ら研究 [1] では Wikipedia 全データを用いて, 人手でクラスター数を 2,000 で設定して, k-means 法でクラスタリングする. このクラスタリングの結果と処理したい文書データを比較して, 重要な情報を表に整理する. 過去の研究では最終的に作成された表に空欄が多く, F 値で評価すると, 精度が低いという問題がある.

本研究では 2 回目のクラスタリングをすることで, 最適なクラスタ-数と情報の重要度を計算して, 過去の研究と比べて, 重要な情報を含む, 精度が高い表を作る. 本研究では最適なクラスター数を計算する方法として, 岡崎らの研究成果, Silhouette 法, UpperTail 法を用いる. この三つの計算方法を用いて作った結果はよく似ている, 2 回目の文レベルでクラスタリングする際に文の中に単語の数が少ないのが原因と考えられる. そして, この三つの結果の精度の平均は 0.70 に届かないが, 過去の研究と比べて, 精度が高くなり, 提案手法の精度の最低値も 0.30 以上で, 部分的に重要な情報が抽出されたと見える. 提案手法の最高値では 0.90 であり, 重要な情報がほとんど抽出された. 提案手法の有効性を確認することができる.

第9章　謝辞

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.

参考文献

[1] Hokuto Akano, Masaki Murata, and Qing Ma. "Detection of inadequate descriptions in Wikipedia using information extraction based on word clustering". *IFSA-SCIS 2017*, pp. 1–6, 2017.

[2] 岡崎健介, 村田真樹, 馬青. "複数文書からの文レベルの情報の書き漏らしの検出". 言語処理学会第 25 回年次大会, pp. 359–362, 2019.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *"The Elements of Statistical Learning"*. Springer, 2009.

[4] Rousseeuw P. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.". *Journal of Computational and Applied Mathematics, 20(1)*, pp. 53–65, 1987.

[5] Mojena R. "Hierarchical grouping methods and stopping rules: an evaluation*". *The Computer Journal*, Vol. 20, No. 4, pp. 3059–363, 1977.

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. "Enriching word vectors with subword information". *In Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146, 2017.

[7] Renato Cordeiro de Amorim and Christian Hennig. "Recovering the number of clusters in data sets with noise features using feature rescaling factors". *Information Sciences 324*, pp. 126–145, 2015.

[8] Ayaka S. and Matsuda S. "Comparison of automatic cluster number determination methods in cluster analysis.". *Academia Information Sciences and Engineering*, pp. 17–34, 2011.

[9] Chang C.-H. Kayed M. Girgis M.R. and Shaalan K.F. "A survey of web information extraction systems.". *IEEE Transactions on Knowledge and Data Engineering, 18 (10)*, pp. 1411–1428, 2006.