人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第26回)
SIG-AM-26-09

単語の組み合わせによるテキスト集合のラベルの自動生成

Automatic Generation of Labels for Text Sets Based on Word Combinations

若園 紫乃[1*]　砂山 渡[1]
Shino WAKAZONO[1]　Wataru SUNAYAMA[1]

[1] 滋賀県立大学工学部
[1] School of Engineering, The University of Shiga Prefecture

Abstract: Although it is clear that text sets classified by unlabeled clustering are classified by some features, it is difficult to understand the content of the text sets by looking at the clusters alone. In this study, we propose a method to automatically generate labels based on the combination of words in the text set and the label starting words. We have confirmed that the labels generated by this method are effective in supporting the understanding of the contents of text sets.

1　はじめに

ラベル無しクラスタリングで分類されたテキスト集合は，何らかの特徴によって分類されていることは分かるが，クラスタだけを見てその内容まで理解することは難しい．このクラスタにラベルが付与されていれば内容の理解がしやすくなる．

現状では，テキストに限らず，データ集合へのラベル付けに関してはクラスタの内容を理解した人が手作業で行うことが多い．この方法では，内容を理解している場合でのみラベルが付与できるが，理解していない場合は，その内容の理解から始めなけらばならず手間がかかることが問題点として挙げられる．

近年では，ソーシャルジオデータのクラスタに別のソーシャルジオデータを用いてラベルを推定する研究 [1] や，位置情報とモーションセンシングデータを用いて二輪車モーションセンシングデータへのラベル付けを行う研究 [2] など，テキストデータに以外でもクラスタへの自動ラベル付けの研究が行われている．

テキスト集合へ自動的にラベルを付与する場合においても，いくつかの研究がある．その手法としては，テキスト中の特徴単語をそのままラベルとして付与するというものや，予めラベルが与えられたデータがある場合にそのラベル付きデータを学習させてラベルとして付与するというものがある．しかし，単語をそのままラベルとして付与する手法では，内容を適切に言い表すことができないという問題点がある．また，予め学習させるという手法では，正解ラベルを用意する必要があり，ラベルが無い集合の場合，はじめに手作業でラベルを付与しなければならないという問題点がある．

そこで本研究では，テキスト集合内に出現する単語の組み合わせによって，事前データの学習をすることなくテキスト集合に自動でラベルを生成する．この手法では，単語のみのラベル付けに比べて，複数単語の関係性に基づいてラベル付けを行うため，テキスト集合の内容をより適切に表すことができることが期待できる．また，事前データの学習が必要ないため，どのようなテキスト集合へも応用ができ，汎用性の高い手法になることが期待される．本研究においては，ラベル無しクラスタリングへのラベル付けに際して，まず，予めテキスト集合の内容が大まかにわかっているテキスト集合中の単語を入力として，テキスト集合を説明するラベルを自動生成する．

本論文では，2 章で関連研究について述べ，3 章で提案するラベルの自動生成システムについて述べる．4 章で提案システムの効果を検証した評価実験について述べ，5 章で本論文を締めくくる．

2　関連研究

2.1　データの解釈を支援する研究

モデルの構築とシミュレーションが理論に基づくデータの解釈を支援するかどうかを検討する研究がある [3]．この研究では，学習者が計算機モデルの作成，シミュレーションを行うことで，データの解釈の支援を行う．また，データの中から分析の手がかりとなる着目点をハイライトにより明示をする ことで，データ分析の支援

*連絡先：滋賀県立大学 工学部 電子システム工学科 若園 紫乃
〒 522-8533 滋賀県彦根市八坂町 2500
E-mail: ov23swakazono@ec.usp.ac.jp

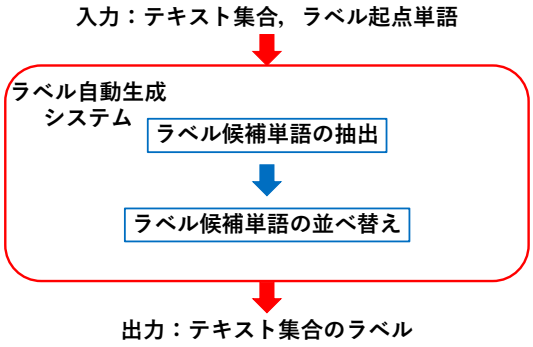

図 1: ラベル自動生成システムの構成図

を行う研究がある [4]．この研究では，平均等の基準値からのズレが大きいデータを着目点として明示しデータ分析の支援を行う．本研究ではデータ集合にラベルを付与することでデータ分析の支援を行う．

2.2 テキスト集合へのラベル付けを行う研究

番組シーン集合にラベルを付与する研究がある [5]．この研究では，クラスタの特徴語と差分語を用いていくつかの単語ラベルを付与するが，本研究では単語を組み合わせて 1 つのラベルとして付与する．また，バグレポートの文章に意味的ラベルを付与する研究がある [6]．この研究では，予め意味的ラベルを付与した文章を教師データとして機械学習によってラベルを付与するが，本研究では教師データを必要とせずにラベル付けを行う．

3 ラベル自動生成システム

3.1 システムの概要

本研究で提案するラベル自動生成システムの構成図を図 1 に示す．本システムでは，テキスト集合の単語を入力とし，その入力単語が含まれるテキストよりラベルの候補単語を抽出する．そして，抽出した単語の並び替えによりラベルを生成して出力する．なお，本システムでは，ラベルを付与するテキスト集合の他に，関連するいくつかのテキスト集合が用意されていることを前提とする．例えば．ラベルを付与するテキスト集合が「あんぱんの作り方」の場合，関連するテキスト集合として，「鯛焼きの作り方」，「水羊羹の作り方」などの和菓子の作り方に関するテキスト集合を用意する．

表 1: TFIDF 順位

TFIDF 順位	単語名	TFIDF 値
1	生地	337.0
2	発酵	69.0
3	丸める	63.0
4	包む	48.0
5	ない	37.0
6	等分	33.0
7	ポイント	26.0
8	塗る	23.0
9	パン	22.1
10	材料	22.0

3.2 ラベル候補単語の抽出

まず，入力単語としてラベルを付与したいテキスト集合内の単語を入力する．この時入力する単語は，テキスト集合内の高頻度語などの特徴的な単語を用いる．入力された単語を含むテキストをラベルを付与したいテキスト集合，他の関連するテキスト集合からそれぞれ抽出する．

抽出されたテキストより，各単語の TF 値を計算する．本システムでは，TF 値を各単語のテキスト頻度とする．テキスト頻度を用いる理由は，一つのテキストに偏ってその単語が大量に出現した場合，結果に影響が出ることを防ぐためである．DF 値は各単語がラベルを付与したいテキスト集合以外の関連テキスト集合に出現する場合は 2，出現しない場合は 1 を取る．

以上で求めた TF 値，DF 値を用いて TFIDF 値を計算する．単語 w の TF 値を $TF(w)$，DF 値を $DF(w)$，テキスト集合数を n とすると，TFIDF 値は以下の式 (1) で求められる．なお，式 (1) 中で $+1$ をの処理を行う理由は，DF 値が 2 の時に TFIDF 値が 0 になることを防ぐためである．

$$TFIDF(w) = TF(w)(\log_{10} \frac{n}{DF(w)} + 1) \quad (1)$$

ここで算出した TFIDF 値の上位 3 単語を抽出し，これをラベル候補単語とする．表 1 に「あんぱんの作り方」に関するテキスト集合の「生地」という単語を入力単語とした時の TFIDF 順位の例を示す．この「あんぱんの作り方」に関するテキスト集合は，クックパッドから収集した 100 テキストで構成されたものである．表 1 の場合，上位 3 単語が「生地」「発酵」「丸める」となるので，この 3 単語をラベル候補単語とする．

なお，上位 3 単語に動詞が 2 個以上出現した場合は，動詞を上位 1 つのみと限定し次に順位が高い動詞以外の単語を使用する．

表 2: 出現順序頻度

出現順序	出現回数
生地→発酵	83
生地→丸める	92
発酵→生地	78
発酵→丸める	56
丸める→生地	88
丸める→発酵	73

3.3 ラベル候補単語の並べ替え

3.2 節で抽出したラベル候補単語のテキスト集合中での出現順序を調べる．はじめに，ラベル候補単語の 2 単語ずつの出現順序を調べる．表 2 に TFIDF 順位が表 1 の場合のラベル候補単語の出現順序頻度の例を示す．

出現順序頻度の合計値を計算し，合計値が最も大きい順に並び替える．表 2 の場合，「生地→発酵」，「丸める→生地」の頻度合計が 171 で最大となるので，「丸める→生地→発酵」の順に並べ替えを行う．

なお，ラベル候補単語に動詞が含まれる場合は動詞を最後に配置し，その上で出現順序頻度の合計値が大きい順に並べ替える．このようにする理由は，一般的な文の構成は主語，動詞の順であり，ラベルもこれと同じ形式に揃えることでより内容の理解につながると考えたためである．

以下の表 3 に入力単語とラベルの出力例を示す．表 3 に示すようにラベルは，ラベル候補単語どうしを「＋」でつないで出力する．

表 3: 入力単語とラベルの出力例

テキスト集合	入力単語	ラベル
あんぱんの作り方	包む	発酵＋生地＋包む
鯛焼きの作り方	牛乳	牛乳＋生地＋加える
マカロンの作り方	シート	クッキング＋シート＋絞る
鶴の恩返しのあらすじ	織る	美しい＋部屋＋織る
桃太郎のあらすじ	退治	団子＋退治＋行く

4 生成ラベルの評価実験

本研究で提案したシステムによって生成されたラベルがテキスト集合内容の理解を支援できるかを評価するため，評価実験を行った．その方法と実験結果を示す．

表 4: テキストデータの詳細

データ名	テーマ	内容
和菓子の作り方	「あんぱん」「鯛焼き」「大福・饅頭」「どら焼き」「水羊羹」の 5 種類のお菓子の作り方について	「あんぱん」「鯛焼き」「大福・饅頭」「どら焼き」「水羊羹」のそれぞれの材料，手順などを説明したテキストをクックパッドから 1 種類あたり 100 テキストずつ用意した．
洋菓子の作り方	「マカロン」「ティラミス」「チーズケーキ」「アップルパイ」「スイートポテト」の 5 種類のお菓子の作り方について	「マカロン」「ティラミス」「チーズケーキ」「アップルパイ」「スイートポテト」のそれぞれの材料，手順などを説明したテキストをクックパッドから 1 種類あたり 100 テキストずつ用意した．
童話のあらすじ	日本の童話「かぐや姫」「鶴の恩返し」「さるかに合戦」「桃太郎」「浦島太郎」のあらすじについて	日本の童話「かぐや姫」「鶴の恩返し」「さるかに合戦」「桃太郎」「浦島太郎」の，それぞれの概要やあらすじなどについて書かれたテキストをネット上から 1 種類あたり 50 テキストずつ用意した．

表 5: 各テキスト集合からの単語

あんぱん	鯛焼き	大福・饅頭	マカロン	ティラミス
包む	牛乳	白玉粉	シート	出来上がる
発酵	スプーン	冷める	乾燥	スポンジ
焼く	焼く	ない	クリーム	レシピ
ない	加熱	片栗粉	潰す	ポイント
パン	薄い	上がる	焼く	チーズ
オーブン	生地	生地	冷ます	ケーキ
レシピ	小麦粉	レシピ	挟む	冷蔵庫
ポイント	ポイント	加える	来る	
		ポイント	冷める	
			冷蔵庫	

アップルパイ	かぐや姫	鶴の恩返し	桃太郎	浦島太郎
シート	羽衣	恩返し	征伐	乗る
煮る	消える	心理	退治	竜宮
パイ	記憶	作家	金銀	乙姫
レモン	地球	愚か	借りる	玉手箱
ポイント	求婚	快い	宝物	太郎
薄い	宝物	贈る	下女	思い出す
冷凍	無理	売れる	団子	考える
冷蔵庫	人間	助ける		良い
	住む	買う		
	知る			
	逃げる			

4.1 実験準備

本実験では，「和菓子の作り方」「洋菓子の作り方」「童話のあらすじ」のテキストを用意した．それぞれの詳細については表 4 に示す．この中から 10 のテキスト集合を本実験に使用した．表 5 は実験に用いたテキスト集合中の各テキスト集合からの単語である．これをラベル無しのテキスト集合とし，これらの単語をベース単語と定義する．

このテキスト集合に 3 章で述べたシステムを用いてラベルを付与した．本実験では，入力単語を変えた 3 パターンのラベルを生成した．表 6 に各テキスト集合に付与したラベルを示す．入力単語は表 5 に示されている単語のテキスト中における出現頻度が高い単語上位 3 つとした．

表 6: 各テキスト集合への付与ラベル

あんぱん	鯛焼き	大福・饅頭	マカロン	ティラミス
発酵＋生地＋包む	牛乳＋生地＋加える	バット＋片栗粉＋まぶす	クッキング＋シート＋絞る	材料＋スポンジ＋クリーム
発酵＋生地＋オーブン	生地＋焼く＋美味しい	等分＋生地＋包む	感想＋オーブン＋焼く	ケーキ＋材料＋スポンジ
発酵＋オーブン＋焼く	薄い＋生地＋塗る	白玉粉＋砂糖＋加える	クリーム＋完成＋挟む	ココア＋冷蔵庫＋時間
アップルパイ	**かぐや姫**	**鶴の恩返し**	**桃太郎**	**浦島太郎**
材料＋シート＋冷凍	貴公子＋求婚＋公達	美しい＋部屋＋織る	黄表紙＋成功＋征伐	背中＋女性＋乗る
シナモン＋砂糖＋煮る	結婚＋無理＋難題	約束＋人間＋助ける	団子＋退治＋行く	乙姫＋玉手箱＋開ける
材料＋砂糖＋レモン	童歌＋知る＋地球	夫婦＋買う＋値段	団子＋宝物＋夫婦	仙女＋乙姫＋考える

4.2 実験手順

本実験では，テキスト集合を表す単語を見てテキスト集合の内容を推定し，説明してもらった．ここで，それぞれのテキスト集合に対し，テキスト集合のベース単語のみの場合とラベルを追加した場合の 2 種類を用意した．なお，各テキスト集合に対してベース単語は約 8 個，ラベルは 3 個である．これを 11 人の理系学生を被験者として行った．

本実験では，被験者 A から E をグループ A，被験者 F から K をグループ B の 2 グループに分け，テキスト集合単語のラベルの有無を変えて実験を行った．表 7 に各被験者グループが説明したテキスト集合と使用したシステムを示す．「和菓子の作り方」「洋菓子の作り方」のテキスト集合ではその作り方を，「童話のあらすじ」のテキスト集合ではそのあらすじを，表示されている単語から推定して説明してもらった．なお，説明文を書く順番は表 7 の上から順であり，グループ A はラベル無しのテキスト集合，グループ B はラベル有りのテキスト集合を先に行った．

表 7: 各被験者グループが説明したテキスト集合と使用したシステム

テキスト集合	グループ A	グループ B
あんぱん	ラベル無し	ラベル有り
鯛焼き		
マカロン		
かぐや姫		
桃太郎		
大福・饅頭	ラベル有り	ラベル無し
ティラミス		
アップルパイ		
鶴の恩返し		
浦島太郎		

4.3 結果と考察

各テキスト集合において．ラベルの有無によって内容説明文に変化があるかを検証するために，テキスト

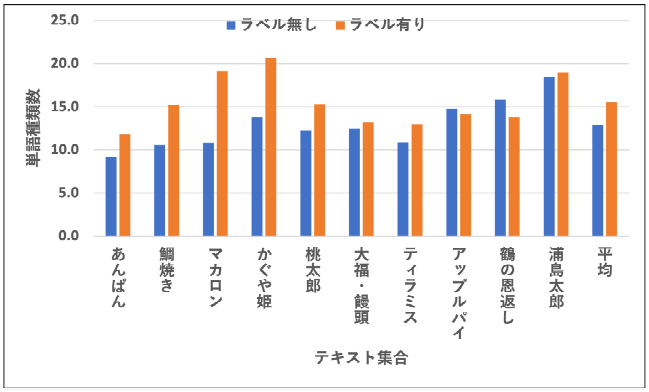

図 2: テキスト集合ごとの内容説明文の単語種類数

集合ごとの内容説明文の単語種類数を図 2，文字数を図 3 に示す．

図 2 より，単語種類数はラベル無しよりラベル有りの方が多くなった．図 3 より，文字数はラベル無しよりラベル有りの方が多くなった．これは，ラベルの付与によって説明に使用できる単語の種類が増加し，それに伴い内容をより長文で書けるようになったためと考えられる．

続いて，説明に使われた単語にどれだけラベル単語が使用されたかを検証するために，テキスト集合ごとの内容説明文の単語種類数の内訳を表 8 に示す．

表 8 より，ラベル有りの場合にベース単語の数が少なくなり，ベースとラベルの共通単語とラベル単語の数が多くなった．このことから，ラベル有りの場合の説明文には，ラベルで付与された単語が多く使われていることが分かる．ベースとラベルの共通単語が多く使われたことから，ラベルの付与によってラベル単語が説明に有効に使われたと考えられる．また，ベースとラベルの共通単語が多くなったことから，ベース単語の中でも着目すべき単語を示唆する効果があったと考えられる．

ラベルの付与が内容理解に有効であったかを検証するために，「マカロンの作り方」に関する説明文の一例を表 9 に，「かぐや姫のあらすじ」に関する説明文の一例を表 10 に示す．表 9 より，説明文の生地の作り

表 8: テキスト集合ごとの単語種類数の内訳

	ラベル無し				ラベル有り			
テキスト集合	総数	ベース単語	ベースとラベルの共通単語	ラベル単語	総数	ベース単語	ベースとラベルの共通単語	ラベル単語
あんぱん	9.2	0.2	3.4	0.4	11.8	0.2	4.2	1.0
鯛焼き	10.6	1.2	3.6	0.0	15.2	1.3	3.8	1.3
マカロン	10.8	2.0	3.2	0.0	19.2	1.8	4.8	3.2
かぐや姫	13.8	4.0	2.0	0.2	20.7	3.5	2.5	1.5
桃太郎	12.2	1.8	3.0	0.0	15.3	1.3	3.0	0.3
大福・饅頭	12.5	1.3	3.2	0.0	13.2	0.4	4.0	3.6
ティラミス	10.8	1.8	2.2	0.0	13.0	0.4	2.8	2.4
アップルパイ	14.8	2.2	2.7	0.0	14.2	1.6	3.2	2.0
鶴の恩返し	15.8	4.2	1.7	0.2	13.8	2.6	1.6	2.8
浦島太郎	18.5	3.3	3.0	0.0	19.0	3.0	3.4	2.8
平均	12.9	2.2	2.8	0.1	15.5	1.6	3.3	2.1

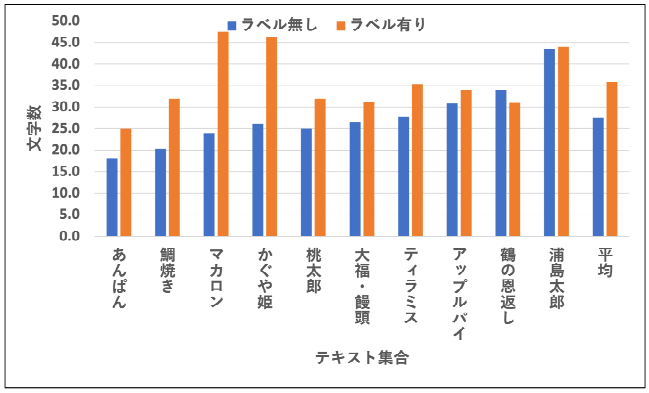

図 3: テキスト集合ごとの内容説明文の文字数

方に関する内容に注目すると，ラベル無しの場合では，乾燥させるという内容しか書かれていないことが分かる．一方ラベル有りの場合では，生地は絞り出すことや，オーブンで焼くことが書かれており，より工程が詳しく書かれていることが分かる．

また，表 10 より，説明文の宝物に関する内容に注目すると，ラベル無しの場合では宝物が与えられているのに対し，ラベル有りの場合では，無理難題で与えられなかったことが分かる．実際のかぐや姫の内容では，宝物を手に入れることは無理難題であるため，ラベル無しの説明文は不適当であると言える．一方，ラベル有りの場合では，無理難題であることが書かれているため，内容を適切に表せていると言える．

以上のことから，ラベルを付与した場合，ベース単語だけでは説明できなかった内容についても説明することができ，内容理解に有効であったと考えられる．

表 9: 「マカロンの作り方」に関する説明文

	説明文
ラベル無し	生地をシートの上で乾燥させて冷ましてクリームを挟む
ラベル有り	クッキングシートに生地を絞り，オーブンで焼く．冷めたらクリームを挟み，冷蔵庫で冷やす．

表 10: 「かぐや姫のあらすじ」に関する説明文

	説明文
ラベル無し	地球に住む人がかぐや姫の存在をしり，求婚を求め宝物を与えている．
ラベル有り	貴公子が求婚するが宝物が無理難題で結婚できないのが「かぐや姫」のあらすじ

5 おわりに

テキスト集合におけるラベルを単語の組み合わせによって自動生成するシステムを実装した．

また，実装したラベル自動生成システムによってテキスト集合にラベルをつけることでテキスト内容の理解を支援できるかを検証するために評価実験を行った．被験者にラベルが無いテキスト集合からの単語と，ラベルが有るテキスト集合からの単語を見て内容について説明を書いてもらった．この実験より，テキスト集合へのラベルの付与が内容理解の支援に有効であるということが分かった．

これを受けて今後の課題として，ラベルの無いテキスト集合に対しても内容理解の支援に有効であるラベル生成システムを目標としていきたい．

参考文献

[1] 荒川豊，福田晃：ソーシャルジオデータのクラスタリング結果に対する自動的な意味付けに関する一検討，第 75 回全国大会講演論文集，Vol.2013，No.1，pp.7–8（2013）

[2] 神村吏，木谷友哉：位置情報を用いた二輪車モーションセンシングデータへの正解データ自動ラベリング手法の一提案，情報処理学会研究報告. マルチメディア通信と分散処理研究会報告，Vol.2013-DPS-157，No.6，pp.1–6（2013）

[3] 齋藤ひとみ，三輪和久，神崎奈奈，寺井仁，小島一晃，中池竜一，森田純哉：理論に基づく実験結果の解釈の支援 認知科学の授業実践におけるモデル構築の効果に関する検討，人工知能学会論文誌，Vol.30，No.3，pp.547–558（2015）

[4] 中川拓郎，砂山渡，畑中裕司，小郷原一智：着目点の明示によるデータ分析支援，第 18 回人工知能学会インタラクティブ情報アクセスと可視化マイニング研究会，pp.34–39（2018）

[5] 三浦菊佳，松井淳，山田一郎，後藤淳，宮﨑太郎，宮崎勝，住吉英樹：番組のシーン集合へのラベリングの検討，情報処理学会全国大会講演論文集，Vol.78th，No.2，pp.2.23–2.24（2016）

[6] 野寄祐樹, 鷲崎弘宜, 深澤良彰, 鹿糠秀行, 大島敬志, 土屋良介：バグレポートの検索性向上のための機械学習による文章単位の自動ラベリング，第 80 回全国大会講演論文集，Vol.2018，No.1，pp.207–208（2018）