時系列テキストデータの可視化技術の開発

Visualization of Time Series Text Data

保守員の故障日報やコールセンタログ，twitter やブログなど，大量の時系列テキストデータが活用できる状況となっている。従来は人手でデータの分類・整理を通してデータを活用してきたが，現在は人がすべてを把握できる量をはるかに超えたデータが活用できる状況である。これら大量データの活用による価値ある情報の抽出が注目されており，分析者のサポートのため機械的な集計や全体的な傾向の分析技術が必要となっている。本報告では大量テキストデータの全体像を俯瞰するため，時間ごとにデータを区切り，各時間のデータを前後の時間と同様の基準で話題を分類できる技術の開発を行った。サンプルデータで検証した結果，過去の話題の流れを追った時間ごとの話題の分類ができ，従来技術では難しかった新規話題の発見もできることが確認できた。

塚原 朋哉　Tsukahara Tomoya
高梨 勝敏　Takanashi Katsutoshi
宮内 秀彰　Miyauchi Hideaki
佐藤 俊也　Sato Shunya

１．はじめに

1.1 背景

テキスト情報が大量に生成され分析対象となっている現在，その全体像を俯瞰する技術として文書クラスタリングが活用されている。また，故障日報やコールセンタログ，雑誌や論文，twitter やブログの内容の分析など，時間とともに変化するテキスト情報を俯瞰する技術も重要である。

俯瞰する技術としては文書内容の関連度合いによってグループ分けする文書クラスタリング技術がある。また，分類すべき単位（カテゴリ）をあらかじめ定め，文書をカテゴリ別に分類するクラス分類技術がある。

時系列の文書を俯瞰するとき，カテゴリが定まっている場合にはクラス分類が適用できる。しかし，カテゴリが未知の場合や時間とともに話題が変化する場合には，常に分類規則のメンテナンスが必要となり運用のコストが高くなる。カテゴリが変化する場合には文書同士の関連度合いによって統計的に文書の仕分けを行う文書クラスタリングが有用である。

本報告では時間とともに話題の内容が変化する文書群の俯瞰に有効な文書クラスタリングを述べる。

1.2 従来技術

時間とともに話題が変化する文書群を俯瞰するためにクラスタリングを適用する場合，時間軸を距離に取り入れクラスタリングする手法がある 1)2)。時間間隔の近い文書同士には距離が短くなる制約を与えることで，時間方向に連続する話題のつながりであるクラスタの生成を促進する。この手法はニュースなど時間の変化に対して話題の変化も大きな文書群から継続した話題の抽出には向くが，ある時間でみた場合，小さなクラスタが多く抽出されやすく，論文集合の時間的推移や日報からの故障傾向の推移，コールセンタログからのクレームの分析など，期間単位の文書群全体の俯瞰には向かない。

期間単位の俯瞰のために，期間単位にクラスタリングを適用することが考えられる。しかし，期間単位にクラスタリングを適用すると，前後の時間でつながりのないクラスタが生成されやすく，データの時間変化を追うことが難しくなる。この問題を解決するために，制約付きクラスタリング 3)4)を用いた手法がある。対象期間の文書間の距離情報に過去のクラスタリングの結果を距離情報として加算する手法である。対象期間の文書をクラスタリングするとき，過去時点の文書とともにクラスタリングし，過去時点の文書同士の距離に過去時点で同じクラスタに属していた場合に距離を強める制約を設け，過去時点のクラスタと同様のクラスタを生成することを目的としている。制約付きクラスタリングは過去クラスタを反映できる一方，新規のクラスタの生成を抑制してしまう性質を持つ。

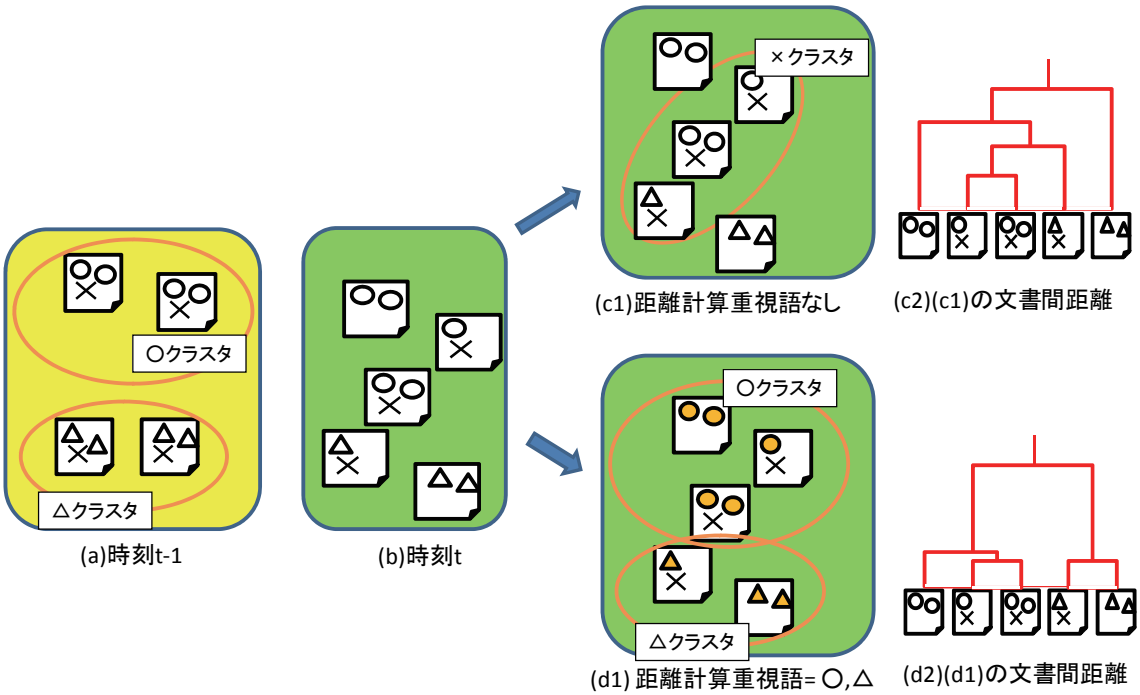

図 1　距離計算重視語によるクラスタリング

2．提案手法

本報告では，文書の俯瞰のために時間軸で区切った単位ごとに文書をクラスタリングする。その際，過去時点のクラスタの代表語を使って対象期間のクラスタリングをすることで，クラスタリングの観点が急激に変化しないことを目的とする。また，アラート情報を活用することで新規の話題の生成が促進できることを目的とする。

2.1 過去時点のクラスタ情報の反映

過去時点のクラスタ情報を反映させるために，クラスタリングで用いる文書間の距離を調整する。文書間の距離の調整には，文書間の距離の計算に用いる文書ベクトルの要素（特徴語の重み）の調整によって行う。クラスタとしてまとめたい特徴語（以下，距離計算重視語，と呼ぶ）の重みを大きくすることにより，距離計算重視語を中心としたクラスタを作成する。過去クラスタの内容を表す特徴語（以下，クラスタの代表語，と呼ぶ）を距離計算重視語とすることにより，対象期間のクラスタに過去クラスタの情報を反映させる。

クラスタの代表語の抽出には，クラスタを構成する文書群からの特徴語の抽出 5)を利用する。クラスタ内の全文書について，特徴語ごとにそれぞれの文書中の重要度の和を計算する。重要度の和が上位の特徴語をそのクラスタの代表語とする。

図 1 に概念図を示す。ファイルアイコンが検索対象の文書を表し，文書中の○や△が特徴語を表す。(a)が過去(時刻 t-1)時点でのクラスタリング結果とする。特徴語○と×を含む 2 文書が○クラスタとなり，特徴語△と×を含む 2 文書が△クラスタとなった例である。文書中の特徴語○と△がクラスタの代表語として抽出されている。(b)次の時刻 t で図に示す 5 文書が存在したとき，これら文書に対してクラスタリングを行うことを考える。距離計算重視語なしの場合，(c1)のように×クラスタができる可能性がある（文書間の距離を(c2)のデンドログラム（樹形図）で表す)。これに対し，(a)過去クラスタの代表を距離計算重視語として用いた場合，○を含む文書同士の距離が近くなり，また，△を含む文書同士の距離が近くなるため(d1)過去クラスタの観点を反映した○クラスタと△クラスタが生成されるクラスタリングが可能となる（文書間の距離を(d2)のデンドログラムで示す)。

2.2 新規話題のクラスタ生成の促進

過去クラスタの代表語を距離計算重視語とした場合，新規の話題が過去クラスタの構成に取り込まれ新しいクラスタを作ることが難しくなる。そのため，新規内容（アラート情報）を表す特徴語も距離計算重視語として用いることで新規話題のクラスタの生成を促進する。

新規内容の判定には，過去のある一定期間には現れず，

対象期間にはじめて現れた特徴語（以下，新規語，と呼ぶ）で判定する。新規語のうち重要度が上位の新規語を対象期間の距離計算重視語として用いる。

2.3 従来手法との比較

制約付きクラスタリング 3)4)では，過去時点の文書も含めて対象期間のクラスタの計算を行う。このとき，過去時点の文書同士が同じクラスタに属していた場合，対象期間の距離計算においてそれら文書同士の距離を近くする制約により，過去時点のクラスタ情報を反映させている。提案手法では，過去クラスタの代表語を距離計算重視語として抽出し，距離計算重視語を含む文書同士の距離を近くすることにより過去時点のクラスタ情報を反映させている点が異なる。

また，従来手法では過去時点のクラスタ情報を反映することにより新規話題が抑制される。提案手法でも過去クラスタの代表語を距離計算重視語としたことにより新規話題が抑制される。そこで，新規話題を表す新規語を抽出し，新規語も距離計算重視語として用いることで新しいクラスタの生成を促進することを試みる。

３．評価実験

3.1 実験目的

(1) 距離計算重視語によるクラスタ生成の変化を確認する。距離計算重視語なし（特徴語の重みに統計量を利用した文書間距離）の文書クラスタと，距離計算重視語ありの文書クラスタを比較する。

(2) 過去クラスタの代表語を距離計算重視語として用いることで対象期間のクラスタに過去時点のクラスタの構成が反映されているかを確認する。比較対象として過去クラスタの代表語を用いずに，対象期間ごとにクラスタリングした結果と比較する。

(3) 過去クラスタの代表語に加えて新規語を距離計算重視語とすることで，新規話題のクラスタが生成できるかを確認する。比較対象として過去クラスタの代表語を距離計算重視語としたクラスタリング結果と比較する。

3.2 評価データ

評価実験ではサンプルデータに対する評価を行った。サンプルデータは通販のコールセンタログを仮定して作成した製品テスト用のデータである。150 レコードで，期間として 4 月～9 月までを想定しており，問い合わせ元の属性情報 13 種類（「年代」や「商品名」など）と，自然文 2 種類（「問い合わせ内容」と「回答」）からなるデータである。このうち，「問い合わせ内容」の自然文を対象に文書クラスタリングを行う。各データには商品を対象とした問い合わせの場合には，商品名の属性に対象の商品名が付与されている。商品名の属性値は 21 種類＋商品名なし 1 種類の合計 22 種類である。

実験目的の(1)のカテゴリ情報として商品名を距離計算重視語として用いる。なお，商品名が問い合わせ内容に記述されている場合と記述されていない場合があり，また，商品に対する問い合わせでない場合（住所変更についての問い合わせなど）は商品名自体の値はない。

3.3 実験方法

評価データの 1 件を 1 文書として XML を作成した。XML では各属性・自然文を個別のタグに格納した。自然文は形態素解析を行い，品詞情報に基づき，名詞・動詞・形容詞・副詞を抽出し，名詞同士の連結，接頭語・接尾語の名詞への連結による複合語の作成，名詞と形容詞・動詞との係り受け解析を行い，連結した単位の文字列を特徴語とした。

特徴語の重要度の計算には 1 つの文書内で使われる回数が多いほど，また，特定の文書にのみ出現するほどそれらの文書を特徴づける（重要と考える）指標である tf-idf に類する統計値を使った。文書間の距離には問い合わせ内容の特徴語を要素とする文書ベクトルに対して，ベクトル要素の各特徴語で検索を行い，特徴語の重要度と文書スコアの積和を用いた。特徴語が距離計算重視語の場合は特徴語の重要度を 100 倍（部分一致する場合は 75 倍）に変更した。クラスタリングは最短距離法を使い，距離の近いクラスタ同士の統合により文書クラスタを作成した。距離の近いクラスタ同士の統合は，大きなクラスタ同士の統合をしない（統合対象のクラスタ内の文書数が 3 文書以下の場合統合する）としてクラスタを生成した。

距離計算重視語に用いる過去クラスタの代表語はクラスタ内文書の重要度の高い特徴語のうち，上位 5 語を用いた。

距離計算重視語に用いる新規語は，過去期間すべてと比較して，対象期間で初めて現れた名詞のうち，多くの文書に含まれる順に上位 5 語を用いた。

(a)距離計算重視語適用なし

No	各クラスタの代表語	クラスタ内文書の商品名属性値
1	[変色,返品,黒い,白桃,どう] (98)	白桃(16), チーズケーキ(4), トマト(6), ワカメ(2), スイカ(3), その他(67)
2	[品切れ,購入する,いつも,商品,商品=購入する] (25)	チーズケーキ(6), 大福(1), その他(18)
3	[おいしい,とても,もっと,大変,つる] (13)	大福(7), チーズケーキ(1), スイカ(1), その他(4)
4	[振込む,代金,間に合う,振り込む,期限] (3)	商品名なし(3)
5	[カタログ未着] (3)	商品名なし(3)
6	[まったく,OP,OP=言う,折り返し,言う] (2)	商品名なし(2)
7	[電話番号,住所] (2)	商品名なし(2)
8	[希望,住所] (2)	商品名なし(2)
9	[注文キャンセル] (2)	商品名なし(2)

(b)距離計算重視語=商品名

No	各クラスタの代表語	クラスタ内文書の商品名属性値
1	[変色,黒い,返品,白桃,送る] (79)	白桃(16), チーズケーキ(1), トマト(1), スイカ(2), その他(59)
2	[配送,買う,入る,つぶれる,トマト] (13)	トマト(5), スイカ(1), その他(7)
3	[指定する,ワカメ,ワカメ=指定する, 三陸産ワカメ,三陸産ワカメ=指定する] (2)	ワカメ(2)
4	[品切れ,購入する,いつも,チーズケーキ,商品] (23)	チーズケーキ(9), 大福(1), その他(13)
5	[クレジット,支払う,頼む,電話,送料] (6)	その他(6)
6	[おいしい,とても,もっと,大変,つる] (13)	大福(7), チーズケーキ(1), スイカ(1), その他(4)
7	[振込む,代金,間に合う,振り込む,期限] (3)	商品名なし(3)
8	[カタログ未着] (3)	商品名なし(3)
9	[まったく,OP,OP=言う,折り返し,言う] (2)	商品名なし(2)
10	[電話番号,住所] (2)	商品名なし(2)
11	[希望,住所] (2)	商品名なし(2)
12	[注文キャンセル] (2)	商品名なし(2)

表 1　距離計算重視語適用有無の比較。左から(a)距離計算重視語の適用なし，(b)距離計算重視語＝商品名を適用。カッコ内の数字はクラスタ・属性に紐づく文書数。赤字は本文で参照している商品名を表す。

3.4 従来手法との比較

実験目的に対して評価データを用いて実験を行った。

3.4.1 距離計算重視語

距離計算重視語の影響をみるために，評価データの全期間に対して，通常の統計値を用いてクラスタリングした結果と，評価データの商品名を距離計算重視語として用いてクラスタリングした結果を表 1 に示す。表ではクラスタごとに，クラスタの代表語（上位 5 語），クラスタ内の商品名とその商品名の文書数を示している。クラスタ内の商品名は代表的なものを表に記述しており，その他の商品・商品名なしは「その他」としてまとめている。

表 1(a)No.1 のクラスタでは，距離計算重視語適用前には大きなクラスタだったが，距離計算重視語の適用により，「トマト」や「ワカメ」を代表とする新しいクラスタが生成されることが確認できた（(b)No.2, No.3)。また，「チーズケーキ」が「チーズケーキ」グループ（(b)No.4）に集まることも確認できた。一方「スイカ」や(b)No.1, No.6 の「チーズケーキ」など商品ごとのカテゴリにならない文書も存在したが，これらは文書中に商品名が含まれていないためであった。

3.4.2 過去クラスタの代表語

評価データの 4 月～9 月のうち，月を単位として文書のクラスタリングを行った。このとき，クラスタリング対象期間の前の期間（5 月でクラスタリングする場合，4 月）のクラスタの代表語（クラスタの重要語上位 5 語）を距離計算重視語として用いた。

表 2 にデータの最初の期間である 4 月(a)と次の期間である 5 月(b)～(d)の結果を示す。(b)は距離計算重視語を用いていない通常のクラスタリング結果である。(c)は 4 月のクラスタの代表語を距離計算重視語として用いた結果である。

(c)No.1 のクラスタは(a)No.1 と同じく「商品が品切れ」や「届かない」ことに関するクレームのクラスタであり，No.2 の「システム面」のクレームのクラスタと異なるクラスタとして生成されている。(b)No.1 ではこの 2 つが混ざっている。過去クラスタの代表語を距離計算重視語として用いる効果が現れている。

3.4.3 新規語

表 2(d)が過去クラスタの代表語に加えて新規語を距離計算重視語として用いたクラスタリング結果である。(d)No.1 は「商品が届かない」クレームを表し，No.2 は「品切れ」と「カタログ」に関するクラスタを表している。新規語(商品・お届け日)により内容を絞った新しいクラスタが生成されている。

3.5 考察

3.4.1 節の商品名を距離計算重視語として用いた場合，商品名を反映した新しいクラスタができた。平均的な特徴語の重みで大きなクラスタとなっていた文書群から，クラスタ形成の指針となる特徴語が与えられた効果である。評価データの商品名を分類として用いる場合や，論

(a)分析対象期間＝4月

No	各クラスタの代表語
1	[品切れ,購入,たつ,言う,チーズケーキ] (5)
2	[住所] (4)
3	[引っ越す,新住所] (2)
4	(クラスタ外文書) (3)

(b)分析対象期間＝5月

No	各クラスタの代表語
1	[届く,指定する,商品=届く,商品,お届け日] (17)
2	[住所,電話番号] (3)
3	[キャンセルする,注文=キャンセルする,注文,いくら,醤油付け] (3)
4	(クラスタ外文書) (1)

(c)分析対象期間＝5月、代表語あり

No	各クラスタの代表語	4月の代表語
1	[届く,指定する,商品=届く,商品,お届け日] (11)	品切れ(3), チーズケーキ(2)
2	[間に合う,振り込む,振込む,代金,忘れる] (5)	–
3	[住所,電話番号,引越し,新住所,送る] (4)	住所(3), 新住所(1)
4	[キャンセルする,注文=キャンセルする,注文,いくら,醤油付け] (3)	–
5	(クラスタ外文書) (1)	–

(d)分析対象期間＝5月、代表語・新規語あり

No	各クラスタの代表語	4月の代表語	新規語
1	[届く,指定する,商品=届く,商品,お届け日] (7)	品切れ(1)	商品(4), お届け日(1)
2	[食べる,販売数,いつも,販売,品切れ] (5)	品切れ(2), チーズケーキ(2), 新住所(1)	カタログ(2)
3	[間に合う,振り込む,振込む,代金,忘れる] (5)	–	パスワード(2), お客様番号(3)
4	[住所,電話番号] (3)	住所(3)	–
5	[キャンセルする,注文=キャンセルする,注文,いくら,醤油付け] (3)	–	–
6	(クラスタ外文書) (1)	–	–

表 2　距離計算重視語適用による比較。(a), (b)適用なし，(c)過去クラスタの代表語を適用，(d)過去クラスタの代表語＋新規語を適用。カッコ内の数字はクラスタ・特徴語に紐づく文書数。

文のカテゴリ情報など分類すべき観点がある程度明確な場合，距離計算重視語の効果があることがわかった。

表 1(b)の新しいクラスタができた中で, No.5 クラスタ「クレジット・電話」ができたのが確認できた。クラスタ内の文書に共通する商品は存在しないため，距離計算重視語の直接の効果ではないが，商品名による新規クラスタができた際に，元クラスタとの関連度が下がり別クラスタになった。距離計算重視語によるクラスタの変化に伴い，副次的にクラスタ分割されたケースであり，距離計算重視語による絶対的なクラスタリングではなく，文書内容によって柔軟にクラスタを形成していることがわかる。

新規語の場合，対象の限定（商品・お届け）と，新しい観点（カタログ）により，クラスタを分割し意味を限定することができている。過去クラスタの代表語「品切れ」でひとまとまりにされていた内容を分割することに成功している。ただし，品切れでまとまっていたクラスタが分割されているとみることもでき，過去クラスタのまとまりを重視する場合には負の作用として働いている。

4．まとめ

過去の文書クラスタの代表語を距離計算重視語としてクラスタリングに利用した。これにより過去時点のクラスタ構成の観点を対象期間のクラスタリングに取り入れることができることを確認した。

本手法ではクラスタリング対象期間の直前期間のクラスタ代表語のみを距離計算重視語として利用した。それ以前の時点の情報は累積的に過去のクラスタ構成に影響を与えているため，間接的に過去情報を取り込んでいることになる。一方，クラスタリング対象期間からさかのぼって複数期間のクラスタの代表語を明示的に用いる手法も考えられる。今後検証が必要である。

過去クラスタの代表語・新規語の選択によって生成されるクラスタが変動する。適切な特徴語を選択すれば，新カテゴリをクラスタとして抽出できるが，誤った新規語の選択ではユーザの望むクラスタリングの結果とはならない。また，過去クラスタの保持と，新規クラスタの生成は相反する目的ともなるため，特徴語の抽出と，どの程度利用するかの重み付け方法の考察が必要である。

距離計算重視語によりクラスタの中心となる概念の指定ができたが，過去時点で同じクラスタ・異なるクラスタに存在していた情報を反映していない。文書間の距離を調整する制約については今後の検討としたい。

本報告では最終的に生成されるクラスタと，クラスタの代表語で評価を行ったが，特徴語として複合語を用いている効果や最短距離法のクラスタリングなど，距離計算重視語のみの効果を評価しにくくなっている。今後，定量的な評価方法の考案と実データへの適用をすすめていく。

５．おわりに

本報告ではサンプルデータに対する実験を通して時系列データが同様の観点でクラスタリングされることを確認した。現在は，実データに対する適用をすすめて評価を行っている。twitter のデータからの商品とその評判情報の抽出，プロジェクト計画書やメールからのプロジェクトごと・お客様ごとの傾向情報の抽出とその活用の実験をすすめている。さらに大規模のデータに対応するため，分散クラスタ構成での適用を検討中である。

参考文献

1) 菊池匡晃，他，階層型クラスタリングを用いた時系列テキスト集合からの話題推移抽出，DEWS2008 B3-3
2) 浜田玲子，他，時系列性を持つテキストのクラスタリング，NLC2012-2，PRMU2012-22，pp.7-12，2012-6。
3) 榊剛史，他，制約付きクラスタリングを用いた論文分類，人工知能学会全国大会論文集，JSAI2006，2006.
4) 水野珠季，他，制約付きクラスタリングによるデータの時系列変化の把握，人工知能学会全国大会論文集，JSAI2009，2009.
5) 塚原朋哉，情報の視覚的検索方法，FIT（情報科学技術フォーラム）2003, E-043, pp.179-180, 2003.

塚原 朋哉　　1997 年入社
ナレッジソリューショングループ
CoreExplorer，テキスト検索・分析システムの研究・開発
tomo@hitachi-to.co.jp

高梨 勝敏　　1995 年入社
地域復興貢献室
東北地域の復興貢献活動, 知識管理システムの研究開発
takana@hitachi-to.co.jp

宮内 秀彰　　2008 年入社
ナレッジソリューショングループ
CoreExplorer，テキストマイニングツールの開発・適用
hideaki.miyauchi.01@hitachi-to.co.jp

佐藤 俊也　　1993 年入社
ナレッジソリューショングループ
CoreExplorer，テキスト・数値情報の分析ソリューションの適用
shu_sato@hitachi-to.co.jp