番組コメント解析クラスタリング結果の考察

A Numerical Consideration of Classification on Intelligence Circulation System for TV Program

有安　香子†　金次 保明†　浜田 浩行†
Kyoko ARIYASU†　Yasuaki KANATSUGU† and Hiroyuki HAMADA†

1. はじめに

近年、TV放送やVODなどの動画像を視聴しながら、インターネットを介してコメントを共有する「ソーシャルテレビ」という新しい視聴形態が広がりをみせている。我々は、放送番組に対するコメントを共有するだけでなく、コメント解析した結果も共有し合う「情報還流システム」の提案をおこなっている[1]。視聴者が番組を見ながら感想や意見を入力し、放送局側で字幕や出演者情報などの情報を補足しながらコメント解析をおこない、可視化グラフ[1]、漫画風ダイジェスト[2]、番組・ブログ推薦[3]など様々なサービスを生成し、視聴者と放送局、視聴者同士の繋がりを促進することを目的としたシステムである。対人魅力の類似性や熟知性などを考慮したサービス設計をおこない、同じ時間に同じ番組を見た視聴者同士がコミュニティを形成できるような環境を目指している。

本稿では、情報還流システムに必要な要素技術として、コメント内容の類似度によりユーザクラスタリングをおこなう手法を提案し、提案するクラスタリング手法の有効性を検証するための実験をおこなったので報告する。

2. コメント内容の類似性に基づくクラスタリング手法

情報還流システムでは、視聴者がテレビ番組を見ながら、ウェブサイトに集い感想等のコメントを書き込む。時間情報のついたデータをシステム入力とすることで、コメントと番組の時間方向の対応付けが可能となる。[1]で提案した解析手法を用い、コメント内容をコメント対象と感情表現の対として解析する。コメント対象とは、番組内の登場人物の誰に対するコメントかを表すものであり、感情表現とは、コメント内容を肯定・否定・驚き・悲しみなどの感情別に分類したものである。また、解析と同時に字幕データの時間情報を用いて、入力時間遅延を補正する。一連の解析処理により、視聴者の入力コメントは「ユーザＩＤ」「コメント入力時間」「コメント対象」「感情表現」の4要素からなるデータとしてシステムに蓄積される。

これらの解析結果を用いて、ユーザのコメント傾向をもとにしたクラスタリングをおこなう。各ユーザのコメント傾向の類似性を数値化し、この数値をもとにクラスタリングをおこなうこととする。

2-1. Simpson係数を用いた類似性の数値化

放送番組を見ながらコメントを共有する視聴スタイルでは、コメント数がユーザにより大幅に異なる、コミュニケーションを妨げることを目的とし、内容に相関性のない発言を繰り返すユーザ（荒し）が存在するなどの実態を念頭に置いて、クラスタリングなどのデータ処理をおこなう必要がある。また、情報還流システムでは、同じ番組を見て似たようなコメントを書き込んだ視聴者同士によるコミュニケーションを促進し、稀有な出会いを育てることを目的としているため、少数派コメントとして一致した場合、それらを重要視したクラスタリングすることが望ましい。これらの点を考慮し、構成要素数が大幅に異なる集合間の類似度の算出に適したSimpson係数法を改良し、コメント内容の類似性を数値化することとする。具体的には、各ユーザを、コメントを構成要素とした集合と捉え、ユーザ間の類似度を下記の方法により算出する。

n人のユーザのコメント群を$U_i(1\leq i\leq n)$と表すとする。U_iのコメント数をm_i個とすると、$U_i=\{W_{ij}\mid 1\leq i\leq n,\ 1\leq j\leq m_i\}=\{W_{i1}, W_{i2}\cdots W_{im_i}\}$で表される。ただし、各コメント$W_{ij}$は、$W_{ij}$=[コメント対象,感情表現]で表すこととする。

次に、各コメントW_{ij}が他のユーザのコメントとどの程度一致しているかを積算した値$|W_{ij}|$を求める。比較対象ユーザ$U_k=\{W_{kh}\mid 1\leq h\leq m_k\}$のコメント$W_{kh}$とコメント対象・感情表現ともに一致したときのみ1、それ以外は0となる$f(W_{ij},W_{kh})$を用いて、$|W_{ij}|=\sum_{k=1,k\neq i}^{n}\sum_{h=1}^{m_k}f(W_{ij},W_{kh})$を算出する。この$|W_{ij}|$を全コメント$(1\sim m_i)$について積算し、$U_i$のコメントが、他の全てのユーザとどの位一致したかを表す$|U_i|=\sum_{j=1}^{m_i}|W_{ij}|$とする。

また、U_iとU_kのコメント対象・感情表現ともに一致したときのみ上記と同じよう積算し、U_iとU_kのコメントの一致度を表す、$|U_i\cap U_k|=\sum_{j=1}^{m_i}\sum_{h=1}^{m_k}f(W_{ij},W_{kh})$を算出する。

以上の算出結果から、U_iとU_kのコメントの類似度、

$$Sim(U_i,U_k)=\frac{|U_i\cap U_k|}{\min(|U_i|,|U_k|)} \quad (1)$$

を求める。なお、$\min(|U_i|,|U_k|)=0$である場合には、(1)式の値を0とする。

他のユーザと相関性のない、荒しユーザは分母の$\min(|U_i|,|U_k|)=0$となるので、(1)＝0のユーザを排除することで、荒しユーザをクラスタリングから除外することが可能となる。また、同様に分母の$\min(|U_i|,|U_k|)$により、稀有なコメントが一致したユーザ同士の(1)の値が高くなるので、(1)の値の高いユーザ同士を同じクラスタに配することで、偶然の出会いを基にコミュニティを形成するきっかけとして利用することができる。この様に2ユーザ間の類似性をあらわす(1)を全てのユーザについて計算した結果をユーザマトリックスに保持する。

2-2. 類似性に基づくクラスタリング手法

各ユーザ間のコメント内容の類似性を計算した係数で構成されるユーザマトリクスを用いて、以下の手順でクラスタリングをおこなう。

全ユーザをt個のクラスタ$(C_1\cdots C_t)$に分ける場合、まず全ユーザ中相関の最も低いペアを選び、各々クラスタ1・クラスタ2に配す。次に、クラスタ1・2のユーザに対して相関性の高いユーザを各クラスタに追加する。配した全てのユーザをマトリクスから除去する。この作業の中で処理されなかったユーザに関しては、新たに最も相関の低いペアを選びクラスタの種としたクラスタを追加生成し、上記作業を繰り返す。この作業を、クラスタ数が上限値に達するか、ユーザがマトリクスに存在しなくなるまで繰り返す。この様にして、クラスタ内のユーザ同士の相関性が高く、また、クラスタ間の相関性が低くなるよう、クラスタにユーザを配置していく。Fig. 1に処理の流れを示す。

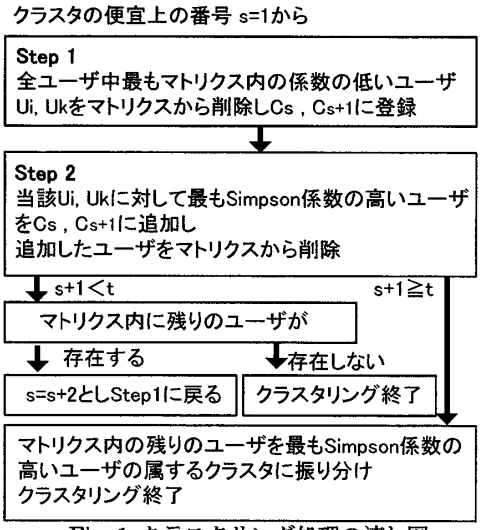

Fig. 1 クラスタリング処理の流れ図

3. クラスタリング結果検証実験

3.1. 実験手法

大河ドラマ「天地人」第 3 話を対象番組とし、延べ 50 人の被験者が情報還流システムをもちいて、番組を視聴しながら 2000 コメントの入力をおこなった。便宜上、これらのデータを不作為データと呼ぶこととする。

3.2. 不作為データのクラスタリング結果

不作為データのクラスタリング結果より、出演シーンの多い主役級の役柄の登場人物 A・B（以下 A・B と表記）、出演シーンが少なく内容もサブストーリ的に独立した役柄の登場人物 C・D（以下 C・D と表記）について、以下の 3 つのコメント傾向が確認された。

①最も構成ユーザ数の多いクラスタに配されたユーザは A・B に対し肯定的なコメントが多い

②2 番目に構成ユーザ数の多いクラスタに配されたユーザは A に肯定的・B に否定的なコメントが多い

③構成人数の少ない少数派クラスタに C・D に対するコメントをするユーザが存在した

3.3. 検証用データの作成

2. で提案するクラスタリング手法の有効性を検証するため、3.2 のコメント傾向を基に、検証用データを作為的に生成した。作成した検証用のデータの内容を Table 1 に記す。

まず、3.2 に記したコメント傾向が正しいか否かを確認するため、①の傾向を模した A/B、③の傾向を模した C を、コメント対象を単数に絞り作成した。次にコメント対象が複数の場合を確認するため、①の傾向を模した AB、②の傾向を模した $A\bar{B}$、③の傾向を模した CD を作成した。これらの作為データが、意図したクラスタに配されることを前提とした時、稀有なコメントをしたユーザ同士が同じクラスタに配されることを確認するため、①・③どちらの特徴にもあてはまるが、少数派のコメント対象に言及する AC のデータを作成した。また、番組内容と関係のない文章をコメントする荒しユーザを模した作為データも生成した。

3.4. 検証用データを含めたクラスタリング結果

検証用データを含めたクラスタリング結果を Table2 に記す。A/B/C/ AB/ $A\bar{B}$/CD の結果より、コメント内容の類似性を基に、多数派意見を代表するクラスタから少数派クラスタまで、コメント内容と全体のコメント傾向に応じたクラスタリングがなされることを確認した。

Table 1 ユーザ別検証用データの特徴

userID	コメント対象	傾向	内要
A	単数	①	A 肯定 10 件
B	単数	①	B 肯定 10 件
C	単数	③	C 肯定 10 件
AB	複数	①	A 肯定 6 件, B 肯定 5 件
$A\bar{B}$	複数	②	A 肯定 6 件, B 否定 5 件
CD	複数	③	C 肯定 6 件, D 肯定 5 件
AC	複数	①or③	A 肯定 5 件, C 肯定 5 件
荒し	-	-	関係のない文章を 3 件

Table 2 クラスタリング結果

クラスタの特徴	配された検証ユーザ
①構成ユーザ数最多のクラスタ	A　B　AB
②2 番目にユーザの多いクラスタ	$A\bar{B}$
③C・Dに言及する少数派クラスタ	C　CD　AC

Table 3 登場人物数と生成されたクラスタ数

対象番組名	登場人物数	生成されたクラスタ数
天地人第 3 話	20 人	9
新撰組第 9 話	16 人	12

その上で、AB/CD/AC の結果を比較することにより、少数派の発言（誰もが言及する A や B でなく C に関しての言及）をしたことが重要視されたクラスタリングがなされたことを確認した。また、荒しユーザがクラスタリングから除外されたことも確認した。

3.5. 別番組での検証

同様の実験を大河ドラマ「新撰組!」第 9 話（延べ 62 人・952 コメント）に関してもおこなった。3.4 と同様にコメント傾向に応じたものとなった。Table 3 に実験において、コメント対象となった登場人物数と生成されたクラスタ数を記す。実験対象番組はドラマという性質上、コメント対象となる登場人物が多く、コメント傾向が発散しやすい。2 つの番組を比較すると、新撰組！は話数が進んでいるため、ストーリに対してコメント内容が分散しており、コメント対象となる人物数が少ないにも関わらず、生成されたクラスタ数が多いことが確認された。これらの結果より、Fig.1 の処理により、コメント内容の分散に応じクラスタ数を決定し、クラスタリングがなされたと思われる。

4. まとめ

視聴者が番組を見ながら入力したコメントから様々なサービスを生成し、視聴者と放送局、視聴者同士の繋がりを促進することを目的とした情報還流システムにおいて、コメント内容の類似度によりユーザクラスタリングをおこなう手法を提案した。今後はコメント対象が少ない種類の番組など、対象番組のジャンルを広げた検証や、コメント傾向が極端に偏った番組などを検証し、結果に応じてクラスタリング処理の改良をおこなう予定である。

参考文献

[1] 有安 他,"情報還流プロトタイプシステム試作",電子情報通信学会技術研究報告, vol. 108, no. 378, CQ2008-60, pp. 5-9, 2009/01/15

[2] 有安 他,"コメント解析結果を反映した漫画風番組ダイジェスト"，第 8 回情報科学技術フォーラム， 第 3 分冊, pp.687-688, 2009/09/02

[3] 有安 他, "ソーシャルテレビに関する一提案", 電子情報通信学会技術研究報告,A8-4, HCG Symposium 2009, 2009/12/12