5C-6

ニュースキュレーションサービスのための ネットコメント要約手法の提案

池田和史†、服部元†、滝嶋康弘†

†KDDI 研究所 〒356-8502　埼玉県ふじみ野市大原 2-1-15

1. まえがき

近年、インターネット上にあふれる多数の情報を集約し、有益な情報のみを利用者に提示するキュレーションサービスが注目されている。コンテンツに対するネット上のコメントを人手で集約するサービスも提供されており、他の閲覧者がどのようにコンテンツを捉えたかが重要視される傾向にある。本稿では、このような利用者のニーズに応じたキュレーションサービスを実現するためのコメント要約手法を提案する。提案手法では、コメントの主観的な表現に着目することで、コンテンツがどのように捉えられたかを端的に表す要約語と、代表的なコメントを提示する。主観評価実験により提案手法とベースライン手法を比較し、有効性を確認した。

2. 関連研究

テキスト情報を要約する研究に関しては、自然言語処理の分野を中心に多くの研究がなされている。一般的な文章に対しては、文章構造に基づいて要約を生成する手法が提案されている[1]。近年では、ソーシャルメディア上の情報を要約する手法も提案されており、複数の投稿をもとに、発生した 1 つの事象について説明する文章を自動生成する手法が提案されている[2]。

本稿では、投稿者の主観的な意見を含むコメントを抽出し、その要約を生成することで、コンテンツがどのように受け止められたかを容易に把握可能とし、コンテンツを魅力的に提示する手法を提案する。

3. 提案手法

提案手法の全体像を図 1 に示す。提案手法は、従来手法などを用いて行う関連データ収集と、提案手法の特徴である主観表現抽出、要約表現抽出、代表コメント提示からなる。本稿では、コンテンツとしてニュースを、コメントとしてツイートを対象として扱うが、ニュース以外のテレビ番組や音楽などの様々なコンテンツ、およびコンテンツに関連する掲示板や SNS 上のコメントに対しても、データ収集の方法を変更することで、提案手法を適用することは可能である。図 1 中で各方式に付与された番号は、方式を説明する本稿の章番号と対応する。

3.1. 関連データ収集

ニュースとニュースに関連するネット上のコメントを収集する。各ニュースサイトが提供する RSS サービスなどを用いて、定期的にニュースを収集する。ニュース記事の URL および、ニュース記事のタイトルを含むツイートを search API を用いて検索することで、ニュース記事を引用するツイートを収集する。

Summarization Techniques of Online Opinions for News Curation Services

† Kazushi Ikeda, Gen Hattori, Yasuhiro Takishima, KDDI R&D Laboratories, Inc.

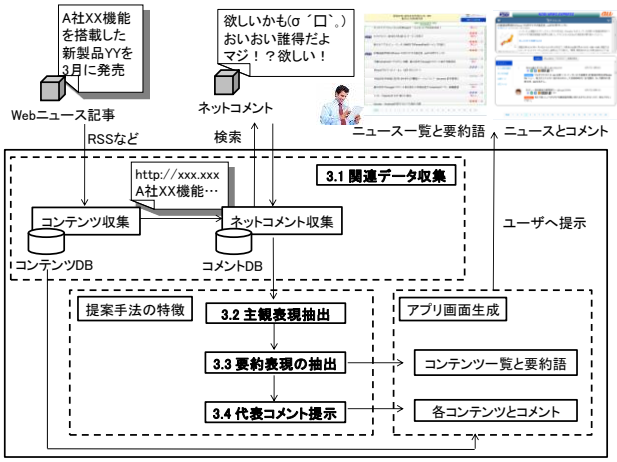

図 1　従来手法と提案手法の動作概要

表 1　主観表現パターンと具体例

パターン 1: 顔文字を含むもの	
えっ(;゜⊿゜)誰得？	
欲しいかも(σ´口｀。)	
パターン 2: 文字が連続しているもの	
うおおおおお！欲しいいい！	→ 5 文字
おいおいおい誰得だよ	→ 6 文字
パターン 3: 記号による主観表現	
マジ！？でも欲しい！！	→ 4 文字
えっ？意味が分からない！！	→ 3 文字
パターン 4: 文体による感情表現	
うわぁ、誰得な仕様はやめろよ	→ 命令形
欲しいけど、安かったらなぁ	→ 仮定形

3.2. 主観表現抽出

各ツイートが主観表現を含む度合いを言語的な指標を用いて算出する。投稿者の主観が現れる指標として、文献[3]を参考に次の 4 指標を利用する。それぞれの指標の算出方法を下記に示し、その具体例を表 1 に示す。

- パターン 1: コメント中に含まれる顔文字の個数を e とする。頻繁に用いられる顔文字表現 20,000 件を登録した辞書を用いて、顔文字を検出する。
- パターン 2: コメント中で連続して出現する文字列の文字数を c とする。コメントを形態素解析し、連続して出現する形態素の文字数を指標として用いる。
- パターン 3: コメント中に出現する顔文字以外の記号の数を s とする。感情表現を表す記号として代表的な「！」および「？」の出現回数を指標として用いる。
- パターン 4: 文体に関する指標を w とする。コメントの主観性はコメント末尾の形容詞、形容動詞、動詞の活用形に現れやすく、活用形が命令形、仮定形、未然形

である場合、主観表現を含む*(w=1)*とし、それ以外の場合、主観表現を含まない*(w=0)*とした。

これらの指標から、コメントの主観度合い P を算出する。一般的な算出式は(1)のように記述できる。ここで、α、β、γ、δ は各指標の重みを表すパラメータであり、人手によって主観性の有無をラベリングされた教師データを用いて最適化する。

$$P = \alpha e + \beta c + \gamma s + \delta w \qquad (1)$$

3.3. 要約表現の抽出

主観度合いが閾値以上であるコメント中で特徴的に出現する語を、当該ニュースがどのように捉えられたかを端的に表す要約語として抽出する。特徴的に出現する語の抽出には、TFIDF を用いる。本稿では、TF を当該ニュースに関連する主観的なコメントにおける各単語の出現頻度、DF を全てのニュースに関連するコメント中の各単語の出現頻度とする。また、主観的な要約語を優先的に提示するため、算出された TFIDF 値に対し、語の品詞に基づいて優先度を設定することも有効と考えられる。本稿では、主観の表れやすい形容詞、形容動詞、動詞が優先されるよう、それぞれ重みを設定した。これにより、表 1 に例示したツイートからは、「欲しい」や「誰得」といった要約語が抽出される。

3.4. 代表コメント提示

3.3 章で抽出した要約語を含むコメントおよび、3.2 章で算出した主観度合いが高いコメントを、当該ニュース記事がどのように受け止められたかを表す代表的なコメントとして提示する。ユーザは要約語を閲覧した後に、ニュース記事およびコメントを閲覧することを想定しているため、要約語を含むコメントを最優先に提示し、次に主観度合いの降順にコメントを提示する。

4. 性能評価実験

4.1. 実験の手順と環境

提案手法による要約語および代表的なツイートの提示の有効性を確認するため、5 人の被験者を対象に主観評価実験を実施した。比較対象とするベースライン手法、評価実験に用いたデータ、評価方法について説明する。

ベースライン手法として、要約語の生成については、(1)当該ニュース記事に関連する全ツイートに対して TFIDF を用いて要約語を抽出する手法、(2)当該ニュース記事に対して TFIDF を用いて要約語を抽出する手法、を比較対象とする。代表ツイートの提示については、(a)ニュース記事に関連するツイートを投稿時刻の昇順に時系列で提示する手法、(b)リツイートが多い順に提示する手法、を比較対象とする。

実験に利用したデータとして、Yahoo! ニュースに掲載されたニュース記事 20 件と、当該ニュースに関連するツイート約 3,400 件を用いた。

評価方法は、各手法で抽出した要約語を各ニュースのタイトルと合わせて提示し、その中でニュースに最も興味を持つような要約語を 1 つ選択させた。各手法で優先度の高い 10 件のツイートをニュース記事と合わせて提示し、その中で最も魅力的と感じる手法を 1 つ選択させた。ニュース記事 20 件に対して被験者 5 人が最良の手法を 1 つ選択するため、最大で 100 点の評価となる。

表 2: 被験者実験による要約語抽出性能比較

手法	獲得票数
提案	**89**
ツイート全体の重要語	7
記事の重要語	4

表 3: 各手法による要約語抽出例

ニュース要旨	提案	ツイート	記事
不審な Android アプリに要注意	**怖い**	不審	Android
スマホ版「ドラクエ」初日で 100 万ダウンロード	**なつかしい**	ダウンロード	ドラクエ
2014 年 W 杯、日本の対戦グループ発表	**厳しい**	コートジボワール	W 杯

表 4: 被験者実験による代表ツイート提示性能比較

手法	獲得票数
提案	**92**
時系列	3
リツイート	5

4.2. 実験結果

要約語抽出手法について、被験者実験の結果を表 2 に示す。提案方式が 100 点中 89 点を獲得し、最良の方式であったと言える。ニュース記事に対する要約語の具体例を表 3 に提示する。提案手法では、ニュース記事がどのように受け止められたかを端的に把握することができる。これに対し、ツイートにおける重要語やニュース記事における重要語を抽出するベースライン手法では、ニュースタイトルに含まれる語が要約語として提示される場合が多く、付加情報が得られにくいことが分かった。

次に、代表ツイート提示について、被験者実験の結果を表 4 に示す。提案手法が 100 点中 92 点を獲得し、最良の方式であったと言える。提案手法では、ニュース記事の閲覧者がどのように感じたかといった主観的な意見が多く含まれるのに対し、時系列やリツイート数に基づいて提示するベースライン手法では、引用のみのツイートが多く、有益な情報が得られにくいことが分かった。

5. まとめ

本稿では、コンテンツに対する主観的なコメントを抽出し、ニュースがどのように捉えられたかを端的に表す要約語と、代表的なコメントを提示する手法を提案した。被験者実験により、提案手法はベースライン手法と比べて、より魅力的な要約語および代表ツイートを提示可能であることが確認された。今後の課題として、主観表現の抽出に用いた指標の精査や、年代や性別、IT リテラシの異なる多様なユーザの受容性を考慮した指標の重みづけなどが必要と考えられる。また、多様なコンテンツ、コメントへの適用可能性の評価なども今後の課題である。

参考文献

[1] R. Barzilay and M. Elhadad, “Using Lexical Chains for Text Summarization”, Proc. of the ACL Workshop on ISTS, 17, 1, pp. 10-17, 1997.

[2] B. Sharifi, et. al, “Summarizing Microblogs Automatically”, NAACL HLT, pp. 685-688, 2010.

[3] B. Marina, et. al, “A :) Is Worth a Thousand Words: How People Attach Sentiment to Emoticons and Words in Tweets”, Proc. of SocialCom, pp.345-350, 2013.