他カテゴリとの差異を考慮したユーザレビュー要約手法

1X09C084-5　中村壮悟
指導教員　後藤正幸

1　研究背景と目的

近年，評価 Web サイトには，商品やサービスに関するユーザレビューが大量に投稿されており，企業のマーケティング活動において，これらのユーザレビュー分析の重要性は増している [1]．しかし，レビューの数は膨大であり人手での把握が困難であるため，高村らの文書要約手法 [2] を拡張したユーザレビュー要約手法 [3] が研究されている．この手法は，特定の企業や商品やサービスなど単一のカテゴリ (以下，対象カテゴリ) を対象に，ユーザレビューに含まれるユーザ意見をより被覆するような代表的なレビューを抽出することでユーザ全体の意見をまとめた要約を自動生成するものである．

ユーザレビューの要約は，ある特定の企業や商品，サービスといった対象カテゴリに対するユーザの評価をまとめるために適用されることが多い．その際，他の企業や商品，サービス（以下，他カテゴリ）と異なる特徴的なユーザ意見を発見することのメリットは大きい．一方，評価 Web サイトでは，他カテゴリに対するユーザレビューも容易に取得できるため，これらの情報を活用することにより，分析対象カテゴリに特徴的に現れるユーザ評価を抽出することが可能である．

そこで，本研究では，被覆度に加え，他カテゴリとの差異を考慮するため，新たに特徴度という概念を導入したユーザレビューの要約手法を提案する．さらに，複数の宿泊施設に対するユーザレビューに提案手法を適用し，評価実験によりその有効性を示す．

2　高村らの文書要約手法 [4]

高村ら [2] は，文書全体の内容を最も被覆する文の組み合わせを最適な要約と定義し，施設配置問題と呼ばれる整数計画問題の解として要約文を与える手法を提案している．

いま，要約対象文書は文によって構成されており，i 番目の文を $\boldsymbol{s}_i$ とする．このとき，要約文の組み合わせを，以下の施設配置問題の解として定式化する．

【定式化】

$$\begin{aligned} \max \quad & \sum_{i,j} e_{ij} z_{ij}, \\ s.t. \quad & z_{ij} \leq x_i, \forall\ i,j; \textstyle\sum_i a_i x_i \leq K; \\ & \textstyle\sum_i z_{ij} = 1, \forall\ j; z_{ii} = x_i, \forall\ i; \\ & x_i \in \{0,1\}, \forall\ i; z_{ij} \in \{0,1\}, \forall\ i,j. \end{aligned} \tag{1}$$

(a_i:文 $\boldsymbol{s}_i$ の単語数，K：出力される単語数)

ここで，x_i は，文 $\boldsymbol{s}_i$ が要約文として選択された場合に $x_i = 1$，それ以外の場合に 0 を取る変数，z_{ij} は，文 $\boldsymbol{s}_i$ が文 $\boldsymbol{s}_j$ を被覆する場合に z_{ij}=1，それ以外の場合に 0 をとる変数である．また，文 $\boldsymbol{s}_i$ の文 $\boldsymbol{s}_j$ に対する被覆度 e_{ij} を，

$$e_{ij} = \frac{|S(\boldsymbol{s}_i) \cap S(\boldsymbol{s}_j)|}{|S(\boldsymbol{s}_j)|}, \tag{2}$$

と定義する．ここで $S(\boldsymbol{s})$ は，文 $\boldsymbol{s}$ に含まれる単語集合である．被覆度 e_{ij} を含む目的関数 $\sum_{i,j} e_{ij} z_{ij}$ を最大化する文の組合せとして，x_i が 1 となった文 $\boldsymbol{s}_i$ が要約文として出力される．

3　提案手法

3.1　概要

本研究では，他カテゴリのユーザレビューとの差異に着目し，分析対象カテゴリで特徴的に現れるユーザ意見を抽出するために，新たに特徴度という基準を提案する．具体的には，他カテゴリでは出現していないような文を特徴的な文とし，特徴度で重みを与える．この特徴度と併せて対象カテゴリの被覆度を考慮することで，対象カテゴリのユーザレビューの内容を被覆し，かつ他カテゴリには特徴的な文を要約文として抽出することを目的としている．

また，自動要約では，一般的に出力する文書の長さに制約条件を与える．高村らの手法は，単語数の上限で制約を与えた．提案手法では，制約を文数とし被覆度を定義する．以上のことを考慮して最適化問題で定式化を行い，貪欲アルゴリズムを利用した疑似多項式方式によって解法を与える．

3.2　提案手法の定式化

本研究においては，高村らの手法と同様に要約の最小単位を文とする．そこで，分析対象のカテゴリ c_α に属する要約対象となるユーザレビューを文ごとで切り分け，i 番目のレビュー文を $\boldsymbol{s}_{\alpha i}$ とする．また，その他のカテゴリ c_l に属する k 番目の文も同様に $\boldsymbol{s}_{lk}$ とする．レビュー文 $\boldsymbol{s}_{\alpha i}$ と他カテゴリの文 $\boldsymbol{s}_{lk}$ との差異性を表す特徴度 q_i を，単語の被覆度に着目して以下のように定義する．

$$q_i = 1 - \max_{k,l} \frac{|S(\boldsymbol{s}_{\alpha i}) \cap S(\boldsymbol{s}_{lk})|}{|S(\boldsymbol{s}_{\alpha i})|} \tag{3}$$

この特徴度 q_i は，$\boldsymbol{s}_{\alpha i}$ と $\boldsymbol{s}_{lk}$ の出現単語が完全一致した場合 0 となり，全く一致しない場合は 1 となる．次に，要約文として出力する文書の長さに文数で制約を与えるため，式 (2) に対して，分母に $\boldsymbol{s}_{\alpha i}$ の単語数を考慮し，c_α に属するレビュー文 $\boldsymbol{s}_{\alpha i}$ と $\boldsymbol{s}_{\alpha j}$ の被覆度 g_{ij} を，

$$g_{ij} = \frac{2 \times |S(\boldsymbol{s}_{\alpha i}) \cap S(\boldsymbol{s}_{\alpha j})|}{|S(\boldsymbol{s}_{\alpha i})| + |S(\boldsymbol{s}_{\alpha j})|}, \tag{4}$$

と定義する．本研究では，被覆度 g_{ij} と特徴度 q_i の和を w_{ij} とし，目的関数を $\sum_{i,j} w_{ij} z_{ij}$ とする．また，要約文として出力する文の数を K' 文と制約する．以上を用いて，整数計画問題で定式化を行う．

【定式化】

$$\begin{aligned} \max \quad & \sum_{i,j} w_{ij} z_{ij}, \\ s.t. \quad & z_{ij} \leq x_i, \forall\ i,j; \textstyle\sum_i x_i \leq K'; \\ & \textstyle\sum_i z_{ij} = 1, \forall\ j; z_{ii} = x_i, \forall\ i; \\ & x_i \in \{0,1\}, \forall\ i; z_{ij} \in \{0,1\}, \forall\ i,j. \end{aligned} \tag{5}$$

4　評価実験

本実験では，宿泊施設に関するユーザレビューを対象として，提案手法の有効性を検証する．立地条件が類似した 5 つの宿泊施設をそれぞれカテゴリと考え，各施設について他施設との特徴を考慮したレビュー文抽出を行う．

実験 1 では，高村らの手法，提案手法を適用し，被覆度と特徴度の 2 つ観点から定量的に評価を行う．

実験 2 では，高村らの手法，提案手法に対し，学生アンケートによる評価を行う．アンケートにおいては，定式化で用いた被覆度と特徴度が人間の主観にも合致するか否かに加え，マーケティングでの有用性という 3 つの観点から定性的に評価を行う．

4.1　実験データ

評価実験に用いるデータは以下の通りである．

＜対象データ＞

評価 Web サイト「楽天トラベル」[3] のユーザレビュー

＜対象宿泊施設とレビュー文数＞

c_1：	ヨコハマグランドインターコンチネンタル	2,072 文
c_2：	横浜ロイヤルパークホテル	13,805 文
c_3：	パンパシフィック横浜ベイ	4,151 文
c_4：	横浜ベイシュラトン	13,178 文
c_5：	リッチモンドホテル横浜馬車道	14,485 文

4.2　実験条件および評価指標

本実験では，1 つの宿泊施設を分析対象のカテゴリ，分析対象以外の 4 つの宿泊施設を他カテゴリとし，5 つの宿泊施設それぞれを要約対象とした場合について実験を行った．

[実験 1]

全ての対象データを実験データとして用いる．

＜制約条件＞

提案手法においては，出力する文の数を $K' = 50$ とする．また，提案手法と比較するため，高村らの手法では，出力単語数 K として，レビュー文の 1 文あたりの単語平均値に 50 をかけた値を設定した．

＜評価指標＞

出力された要約文の単語種類数を基準とした被覆度および特徴度によって評価する．

$$\text{被覆度} = \frac{\text{要約文の単語種類数}}{\text{ユーザレビューの単語種類数}} \tag{6}$$

$$\text{特徴度} = 1 - \frac{\text{他カテゴリに含まれる要約文の単語種類数}}{\text{他カテゴリのユーザレビューの単語種類数}} \tag{7}$$

[実験 2]

対象データから，各宿泊施設ごとに 100 文のデータをランダムで抽出した．このランダムサンプリングを 30 回行い，それぞれを実験データとして用いる．

＜制約条件＞

$K' = 20$ とし，高村らの手法では実験 1 と同様に出力する文の数を算出した．

＜アンケート項目＞

設問 1 出力された要約を元のユーザレビューと比べて，どちらの手法が代表的な意見を要約として抽出しているか．

設問 2 出力された要約についてどちらの手法が他の宿泊施設のユーザレビューでは指摘がないような，特徴的な意見が要約として抽出しているか．

設問 3 出力された要約文について提案手法と高村らの手法を比較し，宿泊施設の経営に有益な情報となっているか．

＜評価指標＞

高村らの手法と提案手法の要約を比較し，各設問とも 5 段階で評価してもらう．高村らの手法がよい場合を −2, イーブンを 0，提案手法がよい場合を +2 とし，各設問ごとに回答の平均値により評価を行う．

4.3　実験結果と考察

実験 1 及び実験 2 の結果を表 1 に示す．

表 1. 実験結果

	実験 1				実験 2		
	高村らの手法		提案手法		設問 1	設問 2	設問 3
	被覆度	特徴度	被覆度	特徴度			
c_1	0.106	0.034	0.093	0.028	-0.258	1.419	1.677
c_2	0.039	0.052	0.053	0.109	0.323	1.419	1.548
c_3	0.088	0.031	0.073	0.047	0.226	1.194	1.452
c_4	0.042	0.041	0.045	0.100	0.258	1.355	1.581
c_5	0.040	0.020	0.037	0.018	0.387	1.484	1.163
平均	0.063	0.035	0.060	0.060	0.187	1.374	1.574

表 1 より，分析対象のカテゴリにおける被覆度を維持しつつ，他カテゴリとの特徴度を考慮できていることがわかる．また，提案手法の要約結果には 2 つの特徴が見られた．1 つ目は,「部屋」や「サービス」という宿泊施設のユーザレビューに頻出する単語に付随して,「ベイブリッジ」や「直結」など，特定のカテゴリに特徴的な景観や立地条件のサービスなど各企業がアピールする単語を含む文が抽出されていたことである．2 つ目は，提案手法では，高村らの手法よりも批判的な意見が高村らの手法より多く含まれたことである．一般的に，ユーザレビューには肯定的意見が否定的意見より多く，被覆度を考慮した高村らの手法では，肯定的な意見が多く抽出された．一方，提案手法では，特徴度を考慮したため，他カテゴリでは指摘されていないような批判的なレビュー文が抽出されたと考えられる．

カテゴリ c_1 に対しては，実験 1 では特徴度が低く，実験 2 では設問 1 が負の値を示した．この宿泊施設は，立地条件やサービスよりも外観が象徴的であるが，他カテゴリと比較して特徴的なレビュー文が少なかったためと考えられる．

5　まとめと今後の課題

本研究では，企業のマーケティングを目的として，他カテゴリとの差異を考慮したユーザレビュー要約手法を提案した．評価実験から，高村らの手法の被覆度を保持しつつ，カテゴリの特徴を考慮した提案手法の有効性を示した．

今後の課題として，被覆度や特徴度の観点から，出力する文数を自動推定する方法などが挙げられる．

参考文献

[1] 田邊亘, 後藤正幸,“宿泊施設の戦略構築を支援するユーザレビュー分析に関する一考察,”武蔵工業大学環境情報学部情報メディアセンタージャーナル,Vol.9,pp.91-101,2008.

[2] 高村大也, 奥村学,“文書要約の最大充足化問題によるモデル化,” 情報処理学会研究報告,Vol.46, pp.23-30,2008.

[3] 竹村隆, 雲居玄道, 後藤正幸,“最大被覆問題に基づくユーザレビュー集約手法に関する一考察,”経営情報学会, 秋季全国研究発表大会,D4-1,2010.

[4] 高村大也, 奥村学,“施設配置問題による文書要約のモデル化,” 人工知能学会論文誌,Vol.25(1),pp.174-182,2010.

[5] “楽天トラベル,” http://travel.rakuten.co.jp/.