最大被覆問題に基づくユーザレビュー分析手法

1G06H067-5　竹村隆
指導教員　後藤正幸

1　研究背景・目的

近年，インターネットの普及により，ユーザが自由に情報や意見を配信できる場が急増している．その 1 つにユーザが商品や店舗に対する評価（ユーザレビュー）を投稿可能な評価 Web サイトがある．このようなサイトが多数のユーザからの投稿により充実するにつれて，ユーザレビューが消費者の購買行動に大きな影響を与えるようになった．これに伴い，企業側は経営戦略としてユーザレビューの分析 [1] を行い，マーケティングに活用することが求められている．

一方で，蓄積されるユーザレビューの量は増加の一途をたどり，その全てを見通すには多大なる労力がかかる．そこで本研究では，ユーザの意見全体を把握できるようなユーザレビューの自動集約手法を提案する．具体的には，文書要約を最大被覆問題に帰着させる方法 [2] に注目し，最大被覆問題の観点から重要なユーザレビューの抽出を定式化し，その解法を与える．

2　文書要約

文書要約とは，与えられた単数あるいは複数の文書から，その内容を簡潔に表した要約文書を自動生成する技術である．一般に，良い要約を生成するためには，冗長性の除去と文法の整合性が大きな課題となる．

2.1　文書要約での最大被覆問題

文書要約の代表的な手法は重要文抽出である．重要文抽出では，出力となる要約文書において，少なくとも文レベルでの文法が保証される．重要文抽出の手法は様々な観点から研究されているが，出力された要約文書が全体の情報を担っているかを考慮していない．そこで高村ら [2] は，文書要約を最大被覆問題に帰着させ，組み合わせ最適化の視点から問題を大域的に解く研究を行っている．

このように帰着させる利点は二つある．まず，最大被覆問題では，文書に表現されている事象が要約文書により被覆されているか否かを直接モデル化できる．また，解くべき問題を正確に把握することにより，組み合わせ最適化の分野で開発された様々な知見や計算方法を利用できる．

2.2　文書要約のモデル [2]

モデル化の前処理として，文書 $\boldsymbol{d}$ は D 個の文で構成されているとし，文集合 $\boldsymbol{d} = \{s_1, s_2, \cdots, s_i, \cdots, s_D\}$ を作成し，文集合 $\boldsymbol{d}$ 内の i 番目の文を $s_i (i = 1, 2, \cdots, D)$ とする．また，文集合 $\boldsymbol{d}$ 内の全ての単語（語幹）V 個に対し，その集合を $\boldsymbol{T} = \{t_1, t_2, \cdots, t_v, \cdots, t_V\}$ とする．さらに各文 s_i の単語頻度ベクトルを $\boldsymbol{s}_i = \{\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, \cdots, \boldsymbol{e}_{iv}, \cdots, \boldsymbol{e}_{iV}\}$ と表す．

文書要約では，単語 t_v に対し，それを含む文が一つ以上選択された時に，「t_v は被覆された」と見なし，少数の文でより多くの単語を被覆することを目的とする．

[個数制約付き最大被覆問題]

まず，文書 $\boldsymbol{d}$ の中から選択する文の数は K 以下という制約を与える．x_i は文 s_i が選択された場合に $x_i = 1$ となり，それ以外の場合は $x_i = 0$ となる変数とする．a_{iv} は文 s_i 内の単語頻度ベクトル $\boldsymbol{e}_{iv} > 0$ の場合に $a_{iv} = 1$ となり，それ以外の場合は $a_{iv} = 0$ となる変数とする．これにより，単語 t_v が被覆されるための必要十分条件は，$\sum_i x_i a_{iv} \geq 1$ である．また，w_v は単語 t_v の重要度を表し，単語の出現頻度を用いて計算する．以下に文書要約のモデルを示す．

$$
\begin{aligned}
maximize \quad & \sum_{v=1}^{V} w_v z_v, \qquad (1)\\
s.t. \quad & \sum_{i=1}^{D} x_i \leq K;\ \forall j,\ \sum_{i=1}^{D} x_i a_{iv} \geq z_v,\\
& \forall i, x_i \in \{0, 1\};\ \forall v, z_v \in \{0, 1\}.
\end{aligned}
$$

ここで z_v は目的関数を簡略化するために導入した補助的な変数である．つまり，z_v は単語 t_v が被覆される場合に $z_v = 1$ となり，それ以外の場合は $z_v = 0$ となる．

3　提案手法

本研究では，ユーザレビューからの重要意見抽出を最大被覆問題として定式化し，具体的に解く方法を示す．

3.1　意見抽出での最大被覆問題

従来の重要文抽出では，1 つのトピックについて書かれた文書から，文書の内容を代表する重要文を抽出するものである．しかし，ユーザレビューの集合は多数のユーザによって独立に書かれた文書の集まりであり，全ての文書が同じトピックについて書かれているとは限らない．さらに，ユーザレビューに含まれる意見の重要性は，同じ意見を持つユーザの人数によって測られるべきである．そのため，従来の重要文抽出手法ではユーザの傾向を把握することが困難である．そこで直接，ユーザ全体の代表意見を，同様意見を持つユーザ数に基づいて抽出することでユーザレビュー内の重要意見を把握する方法を考える．また，意見抽出では文の文法及び，整合性が保障される利点がある．

提案手法では，少数の意見でより多くのユーザの意見内容を被覆することを目的とする．具体的には被覆対象を文とし，類似度が高い文は被覆されたとする．したがって，重要度は単語の出現頻度ではなく類似した文の総数で測る．

3.2　意見抽出のモデル化

ある商品について N 人のユーザが評価する．まず，全ユーザが持つユーザレビュー全体を $\boldsymbol{R} = \{r_1, r_2, \cdots, r_i, \cdots, r_N\}$ で表し，意見集合と呼ぶ．ここで意見集合 $\boldsymbol{R}$ 内の i 番目の意見を $r_i (i = 1, 2, \cdots, N)$，$r_i$ 内の文を $s_{ij} (j = 1, 2, \cdots, M_i)$ と表す．次に出力する意見の数は K 以下という制約を与える．x_i は r_i が選択された場合に $x_i = 1$ となり，それ以外の場合は $x_i = 0$ である変数とする．$a_{ij.kl}$ は文 s_{ij} と文 s_{kl} の類似度を表し，w_{ij} は文 s_{ij} の重要度を表す．本研究では，文 s_{ij} の重要度は他の文との類似度の合計とする．また，文間の類似度は一般的に用いられているコサイン尺度により計算す

る．意見集合 $\boldsymbol{R}$ 内の全ての単語（語幹）V 個に対し，各文 s_{ij} 毎に単語頻度ベクトル $\boldsymbol{s}_{ij} = \{\boldsymbol{e}'_{ij1}, \boldsymbol{e}'_{ij2}, \cdots, \boldsymbol{e}'_{ijv}, \cdots, \boldsymbol{e}'_{ijV}\}$ を作成する．以下に意見抽出のモデルを示す．

$$maximize \quad \sum_{i=1}^{N}\sum_{j=1}^{M_i} w_{ij} z_{ij}, \tag{2}$$

$$s.t. \quad \sum_{i=1}^{N} x_i \leq K; \;\; w_{ij} = \sum_{k=1}^{N}\sum_{l=1}^{M_k} a_{ij.kl},$$

$$a_{ij.kl} = \frac{\boldsymbol{s}_{ij} \cdot \boldsymbol{s}_{kl}}{|\boldsymbol{s}_{ij}|\,|\boldsymbol{s}_{kl}|},$$

$$\forall i, j, \; \sum_{k=1}^{N}\sum_{l=1}^{M_k} x_i a_{ij.kl} \geq z_{ij},$$

$$\forall i, x_i \in \{0, 1\}; \; \forall 4, z_{ij} \in \{0, 1\}.$$

2.2 節と同様に，z_{ij} は目的関数を簡略化するために導入した補助的な変数である．また，K は圧縮率 γ が決定された際，全体の意見数 N に対し $K = \lceil \gamma \cdot N \rceil$ と定まる．

3.3 最適化問題の解法

3.2 節の最適化のため，ここでは貪欲アルゴリズムを用いる．全意見集合 $\boldsymbol{R}$ に対し，出力意見の集合 $\boldsymbol{B}$ を表す．W_i は意見 r_i 内の文 s_{ij} が被覆する文の重要度の和を表し，W'_i は意見 r_i 内の文 s_{ij} 内で被覆する文の中で，$\boldsymbol{B}$ 内の文で被覆されていない文の重要度の和を表す．このアルゴリズムは W'_i と意見 r_i 内の文数 M_i の比 W'_i/M_i が最大となる r_i を順に選択していく．

[**貪欲アルゴリズム**]

$\boldsymbol{B} \leftarrow \phi$
while $|\boldsymbol{B}| < K$ **do**
  $\beta \leftarrow \arg\max_{i, r_i \in \boldsymbol{R} \setminus \boldsymbol{B}} W'_i/M_i$
  insert r_β into $\boldsymbol{B}$
end while
output $\boldsymbol{B}$

4 実験及び考察

4.1 実験設定

本研究では，文の被覆に注目して出力したユーザ意見（提案手法）と，従来の重要文抽出の観点から単語の被覆に注目して出力したユーザ意見（従来手法）を比較する．これにより提案手法の有意性を検討し，評価を行なう．また，形態素解析には MeCab[3] を用いた．

[**対象データ**]
宿泊施設のオンラインサイト「じゃらん.net」[4] の 2009 年 9 月 4 日から 12 月 4 日までのユーザ意見

[**対象宿泊施設**]
東京ディズニーシー・ホテルミラコスタ（47 件）
東京ディズニーランドホテル（84 件）
ディズニーアンバサダーホテル（275 件）

[**圧縮率**]
各ホテルのユーザ意見全体をそれぞれ 5%〜30%（5% 毎）に圧縮して出力する．

[**評価方法**]
被験者 45 人（各ホテル 15 人）により，各ホテルの内容を把握する項目（値段，立地等）を作成し，出力意見に対し各項目の被覆率と 1 意見当たりの平均項目数（平均項目数）により評価を行う．

$$\text{被覆率} \quad = \frac{\text{出力意見に含まれる項目数}}{\text{全項目数}} \tag{3}$$

$$\text{平均項目数} = \frac{\text{出力意見に含まれる項目数}}{\text{出力意見数}} \tag{4}$$

4.2 実験結果及び考察

提案手法と従来手法の結果を以下の表に示す．

表 1. ディズニーアンバサダーホテルにおける実験結果

	従来手法		提案手法	
圧縮率	平均項目数	被覆率	平均項目数	被覆率
0.05	4.50	0.600	2.14	0.467
0.10	4.82	0.733	2.18	0.533
0.15	4.60	0.767	2.33	0.700
0.20	4.58	0.767	2.47	0.767
0.25	4.39	0.867	2.71	0.867
0.30	4.30	0.900	2.81	0.967

表 1 は，アンバサダーホテルについての実験結果である．被験者により事前に作成された項目数は 30 項目である．

表 1 より，圧縮率 0.20 の場合，提案手法，従来手法ともに被覆率 0.767 であるが，平均項目数は提案手法が少なくなっている．これは提案手法が短いユーザ意見で全体を被覆できていることを示す．特に，圧縮率 0.30 の場合，提案手法は平均項目数が低く，かつ被覆率が高い．従って，被覆率に対する提案手法の有意性を示すことができた．

同様に，ホテルミラコスタに対しては，圧縮率 0.15 以上で提案手法が従来手法を上回る結果を得た．東京ディズニーランドホテルでの圧縮率 0.30 の場合，従来手法では平均項目数 7.92，被覆率 1.000 に対し，提案手法は平均項目数 5.65，被覆率 0.964 であり，東京ディズニーランドホテルに対しては提案手法の有効性を明確に示すことができなかった．

しかし，平均項目数から従来手法では単語を効率よく被覆するために長文の意見が優先的に選択されていることがわかる．項目数が大きい程，含まれる単語数も増えて長文になる．そのため被覆率が高い値をとる．ただし受け手側は長文の意見を 1 つ提示されるより，短文で簡潔な意見を幾つか提示されることを求める傾向があるため，1 意見当たりで語られる項目数が平均的に少ない提案手法の優位性が示せた．

5 結論及び今後の課題

本研究では，文書要約における手法を消費者のユーザレビューに適用し，最大被覆問題に基づくユーザレビューの自動集約手法を提案した．その結果，圧縮率が高い場合では提案手法の優位性を示すことができた．

今後の課題として，被覆率の向上及び平均項目数の減少と，より大量のユーザ意見に対して実験を行い，有効性を検証していくことが挙げられる．

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] “MeCab,” http://mecab.sourceforge.net/
[4] “じゃらん.net,” http://www.jalan.net/