2007 年度卒業研究概要

重要文抽出に基づくユーザーコメントの分析手法に関する研究

後藤 正幸 研究室

0432166 原田 繁幸

1. 研究背景

近年、急速なインターネットの普及により、人々は自らのブログや評判サイトにおけるユーザレビュー等で気軽に情報を発信できるようになった。これらはマーケティング分野でも顧客の生の意見として大きな価値があり、この膨大なデータから如何にして有用な情報を引き出すかが企業経営の重要な課題となっている。しかしながら、インターネット上に投稿される消費者コメントは日々増加する一方であり、それら膨大な量の情報を人手で分類整理することは容易ではない。そのため、どのような意見が多く、どれを自社の参考にすれば良いのか、判断が非常に難しいという現状がある。この要求に対応しようとするものがテキストマイニングの技術であり、現在、非常に注目を集めている。

2. 研究目的

テキストマイニングの分析手法の一つとして重要文抽出がある。これは文章中の重要な部分を抜き出す手法である。本研究では、渡辺による先行研究[1]を踏まえ、意見集約という目的に適した形の重要文抽出を行い、大量のユーザーコメントデータの分析を行う方法について研究を行う。[1]では、ある程度回答傾向が定まったアンケートデータを分析対象としていた。しかしブログ、評判サイト等ではユーザーが自由に文章を書くため、意見の表現方法が非常に多岐に渡る。ユーザーコメントに使用される単語数も格段に増えるため従来の手法ではうまく結果が導き出せないという課題があった。そこで本研究では、文章データから抽出する単語を【名詞】【形容詞】【動詞】に限定し、さらに出現頻度によって絞込みを行う方法により、ユーザーコメントデータに適した重要文抽出手法を提案することを目的とする。

3.従来手法

―類似度、類似度閾値γを用いた重要文抽出アルゴリズム―

本研究で扱う渡辺の手法[1]文書間の類似度と類似度閾値γを用いる重要文抽出の基本的な考え方は以下の通りとなる。

[定義１]（単語ベクトル） 文書 x における各単語の出現回数を要素とするベクトルを単語ベクトル $\boldsymbol{v}_x$ とする。

[定義２]（類似度） 文書 x, y の類似度 $sim(x, y)$ を単語ベクトル $\boldsymbol{v}_x, \boldsymbol{v}_y$ の余弦で定義する。

$$sim(x, y) = \frac{\boldsymbol{v}_x \cdot \boldsymbol{v}_y{}^T}{\sqrt{(\boldsymbol{v}_x \cdot \boldsymbol{v}_x{}^T)(\boldsymbol{v}_y \cdot \boldsymbol{v}_y{}^T)}} \qquad \cdots (1)$$

多くの文書と類似度の大きい文書は、文書集合全体の多くの内容を含んでいると考えられる。よって、類似度の平均により文書の重要度を与える。

[定義３]（重要度） 文書 x の重要度 $imp(x)$ を次式で定義する。ただし、m は文書の総数である。

$$imp(x) = \frac{1}{m-1} \sum_{y \neq x} sim(x, y) \qquad \cdots (2)$$

【類似度を用いた重要文抽出アルゴリズム】

Step1 分析を行う文書集合の形態素解析を行って単語を抽出し、各文書を単語ベクトルで表す。
Step2 文書集合中の各文書間の類似度 $sim(x, y)$ を式(1)より計算する。
Step3 未抽出文について、文書の重要度 $imp(x)$ を式(2)より計算する。
Step4 未抽出文のうち重要度の最も大きい１文書を抽出する。
Step5 Step4 で抽出した文書との類似度がγ以上の文書を除外し、それを類似文書とする。
類似文書数を数えて抽出文書の付随情報として付加する。
Step6 残り全ての文書が除外された場合は終了。
Step7 Step4 に戻り、残った文書の中から再び重要度の高いものを抽出する。

4. 提案手法

本研究では従来研究の手法に加え、品詞と単語出現頻度による単語の絞込みを行う。具体的には分析対象のユーザーコメントを形態素解析し、単語に細分化、その中から単語出現頻度 2 回以上の【名詞】、【動詞】、【形容詞】を抽出し、重要文抽出を行以下に手順を示す。

【提案手法】

Step1　分析を行う文書集合の形態素解析を行って単語を抽出する。
Step2　文書集合から抽出した単語の【名詞】【形容詞】【動詞】を抽出する。
Step3　抽出した【名詞】【形容詞】【動詞】の中で単語出現頻度が一定回数以上の単語を抽出する。
Step4　抽出した単語を用いて文書を単語ベクトルで表す。
Step5　その後は従来手法と同じ計算を行う。

5. ユーザレビュー分析への適用と評価

5.1 実験概要

本研究では旅行のクチコミサイト 4travel.jp（http://4travel.jp/）のホテルのユーザーコメントを分析対象として扱う。4travel.jp の自由記述のコメントは専門的な用語が使われておらず、一般的な用語が多く使われているので分析対象としては適していると言える。今回の実験で用いたデータは札幌ホテルのユーザーコメント 133 件分である。ユーザーコメント全体の傾向として「景色が良い」「アメニティが充実している」等のポジティブな意見が多く占め ている。提案手法の [Step2]において抽出する【名詞】【動詞】【形容詞】は 2 回以上の頻度を持った単語を抽出する。

5.2 分析結果（一部抜粋）

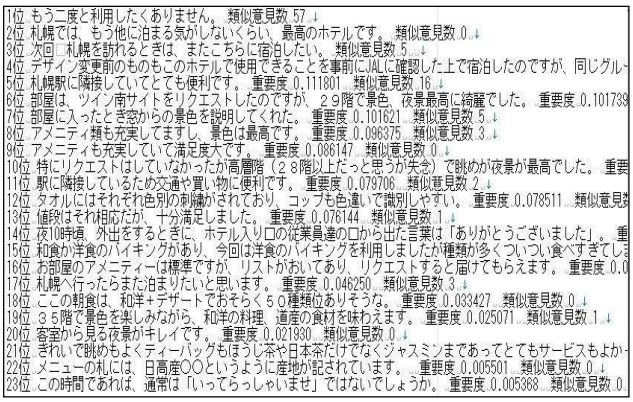
1位 もう二度と利用したくありません。 類似意見数 57
2位 札幌では、もう他に泊まる気がしないくらい、最高のホテルです。 類似意見数 0
3位 次回、札幌を訪れるときは、またこちらに宿泊したい。 類似意見数 5
4位 デザイン変更前のものもこのホテルで使用できることを事前にJALに確認した上で宿泊したのですが、同じグル
5位 札幌駅に隣接していてとても便利です。 重要度 0.111801 類似意見数 16
6位 部屋は、ツイン南サイトをリクエストしたのですが、２９階で景色、夜景最高に綺麗でした。 重要度 0.101739
7位 部屋に入ったとき窓からの景色を説明してくれた。 重要度 0.101621 類似意見数 5
8位 アメニティ類も充実してますし、景色は最高です。 重要度 0.096375 類似意見数 3
9位 アメニティも充実していて満足度大です。 重要度 0.086147 類似意見数 0
10位 特にリクエストはしていなかったが高層階（２８階以上だっと思うが失念）で眺めが夜景が最高でした。 重要
11位 駅に隣接しているため交通や買い物に便利です。 重要度 0.079706 類似意見数 2
12位 タオルにはそれぞれ色別の刺繍がされており、コップも色違いで識別しやすい。 重要度 0.078511 類似意見数
13位 値段はそれ相応だが、十分満足しました。 重要度 0.076144 類似意見数 1
14位 夜10時頃、外出をするときに、ホテル入り口の従業員達の口から出た言葉は「ありがとうございました」。 重
15位 和食か洋食のバイキングがあり、今回は洋食のバイキングを利用しましたが種類が多くついつい食べすぎてし
16位 お部屋のアメニティーは標準ですが、リストがおいてあり、リクエストすると届けてもらえます。 重要度 0.0
17位 札幌へ行ったらまた泊まりたいと思います。 重要度 0.046250 類似意見数 3
18位 ここの朝食は、和洋＋デザートでおそらく５０種類位ありそうな。 重要度 0.033427 類似意見数 0
19位 ３５階で景色を楽しみながら、和洋の料理、道産の食材を味わえます。 重要度 0.025071 類似意見数 1
20位 客室から見る夜景がキレイです。 重要度 0.021830 類似意見数 0
21位 きれいで眺めもよくティーバッグもほうじ茶や日本茶だけでなくジャスミンまであってとてもサービスもよか
22位 メニューの札には、日高産○○というように産地が記されています。 重要度 0.005501 類似意見数 0
23位 この時間であれば、通常は「いってらっしゃいませ」ではないでしょうか。 重要度 0.005368 類似意見数 0

図1. 従来の重要文抽出結果(従来手法)


1位 部屋から札幌の夜景が一望出来きて爽快でした。 重要度 0.149979 類似意見数 34 全体
2位 南向きの角部屋は、札幌の夜景とJRの電車の両方を見ることができます。 重要度 0.14997
3位 また宿泊した部屋も３０階南サイトだったので夜景がとても綺麗です。 重要度 0.145948
4位 ホテルと札幌駅が繋がっていますので隣接する大きなショッピング街まですぐです。 重要
5位 札幌では、もう他に泊まる気がしないくらい、最高のホテルです 重要度 0.137936 類似意
6位 部屋は、ツイン南サイトをリクエストしたのですが、２９階で景色、夜景最高に綺麗でした
7位 お部屋は若干、狭い気がするけどホテル全体の高級感客室から見える景色は最高（すべての
8位 中部国際空港から新千歳空港までのエアーとホテル朝食（３店から選べます）付きで４４９
9位 部屋は、はっきりいって広くないですが、でも眺望は抜群！朝は、うっすらと雪の積もった
10位 夜10時頃、外出をするときに、ホテル入り口の従業員達の口から出た言葉は「ありがとう
11位 客室も清潔で、従業員の応対も、さすが日航ホテル。 重要度 0.110341 類似意見数 0
12位 部屋に入ったとき窓からの景色を説明してくれた。 重要度 0.105449 類似意見数 0 全
13位 他のホテルにありがちなトイレとバスの間をカーテンで仕切るということはありません 与
14位 スパも宿泊者半額で利用出来、温泉だそうです。 重要度 0.097893 類似意見数 0 全体
15位 部屋は最高級クラスでもちろんきれいです。 重要度 0.090128 類似意見数 0 全体順位
16位 スパも利用しました、宿泊者は割引料金で利用出来てゆったりと時間が過ごせて心も身体
17位 アメニティも豊富で、お願いすれば化粧水セット（資生堂）・ボディスポンジ・子供用ス
18位 札幌へ行ったらまた泊まりたいと思います。 重要度 0.083417 類似意見数 1 全体順位
19位 部屋はきれいで清掃も行き届いており、フェラガモのアメニティはお持ち帰りした。 重
20位 札幌が初めてな私は、まず到着後すぐにコンシェルジュに相談しB的確にアドバイスをして
21位 特にリクエストはしていなかったが高層階（２８階以上だっと思うが失念）で眺めが夜景
22位 朝食はバイキングを利用。 重要度 0.080080 類似意見数 3 全体順位 55
23位 夜景も格別です 難を言えばススキノの繁華街まで距離があることです 駅前だからしょう

図2.【名詞】【形容詞】【動詞】での
重要文抽出を行った抽出結果(提案手法)

6.結果の考察

従来手法で分析を試みると(図1)、第1位の筆頭意見として「もう二度と利用したくありません」という文章が抽出されてしまう。これは不要語を含んだまま重要文抽出を行ってしまった結果であり。ユーザーコメント全体の傾向とは異なる文章が抽出されてしまった。一方、提案手法の結果(図2)においては、単語の出現頻度が多い【名詞】、【形容詞】、【動詞】の 3 種類を抽出するため、「部屋から札幌の夜景が一望できて爽快だった」等のユーザーコメント全体の流れと同じようなコメントを抽出する事ができた。また提案手法では単語の出現頻度の多かった「夜景」や「景色」「朝食」と言った単語を含んだ文章が上位を占めている。この事から提案手法による重要文抽出を行うとユーザーコメント全体の流れを代表するユーザーコメントを抽出する事ができる。

7.結論と今後の課題

結論、出現頻度が多い【名詞】【形容詞】【動詞】を抜き出し、重要文抽出することで分析対象のユーザーコメント全体の傾向を把握することができた。しかし、この手法では単語が必ず 1 度は出現しないと分析の対象にすることができないため、文字数が極端に少ないユーザーコメントに対しては正確な分析ができない可能性がある。今後の課題としては文字数が少ないユーザーコメントではどのように処理するのか、また様々な種類のユーザーコメントデータでどの様な結果が求められるのかより深く調べていくことが求められる。

参考文献

[1]渡辺智幸:"情報検索技術を用いたアンケートデータの分析手法に関する研究 ", 2005 年度 武蔵工業大学 環境情報学部卒業論文, (2006)
[2]吉村賢治: 自然言語処理の基礎, 株式会社サイエンス社,(2006)