

## 口コミを用いた企業の特徴の抽出方法の提案 An extraction method of the features of the company by using reviews

菱田隆彰<sup>†</sup>, 炭竈桂輔<sup>‡</sup>, 遠藤正隆<sup>‡</sup>, 中嶋裕一<sup>‡</sup>, 三浦哲郎<sup>‡</sup>  
Takaaki Hishida<sup>†</sup>, Keisuke Sumigama<sup>‡</sup>, Masataka Endo<sup>‡</sup>, Yuichi Nakashima<sup>‡</sup>, Tetsuro Miura<sup>‡</sup>

**Abstract** Review is effective to obtain more specific information about what we interested. However, from a large amount of the reviews that have been posted on the review site, it is difficult to obtain the information that we want. In this paper, we propose a method of extracting the features of the company from reviews of the career site.

### 1. 緒言

自身が勤めたことのない企業の優れた点や問題点を知ることが難しい。就職・転職を望む人々にとってできる限り詳細な企業の実態を知ることが人生を決める上で重要な作業と言えるだろう。インターネットが普及し、ユーザが発するさまざまな情報が公開されるようになった。企業に対する口コミもその一つであり、その口コミは企業の実態を知る有用なツールとして活用されている。しかし、他人が書いた膨大な量の口コミを整理して読むことは容易ではなく、ユーザが短時間で企業の特徴を把握することは困難な作業である。企業の口コミから特徴となる情報を自動的に抜き出す仕組みを構築することができれば、ユーザは口コミを全て読まなくても、企業のあらましを知ることができるだろう。本研究では、転職サイトなどの企業に対しての口コミから、企業の特徴に相当する情報を抽出する手法を提案する。

### 2. 口コミの分析

#### 2.1 関連研究

ブログやSNSなどを用いてユーザがWebサイト上で自身の意見を掲載する機会が増えている。そして、その意見は口コミとして他のユーザに対してその行動に影響を与えるようになった。また、特定の分野の口コミを集めた口

コミサイト数多く登場し、Eコマースサイトでは販売している商品に対して口コミを掲載ことが当たり前になっている。Webサイトに掲載された口コミの内容は、ユーザからの直接的な意見として有用であるが、多くのユーザの意見を整理し、全体の意見として読み取るのは難しい。そのため、口コミの内容の分析や可視化の方法については多くの検討がなされている。

長野ら[1]は、ブログ記事に含まれている口コミ自然言語処理(形態素解析, 構文解析)で解析し、主語(品名), 属性(評価視点), 属性値(評価表現)の抽出を行った。そして、ある分野で使われる語彙及び、語彙間の関係を体系化したオントロジーの辞書を用いて照合し、属性、属性値の組がオントロジー上に存在するかどうかを調べることによって、その対象商品に対する評価の判定を行い、口コミが“良い”もしくは“悪い”のかを分類して提示する手法を提案している。

立石ら[2]は、Webの意見をレーダーチャートの形式で要約するために、3つの機能(Web文書から意見に該当する箇所を抽出する機能, 抽出した意見を着眼点の軸で分類する機能, 抽出した意見を肯定か否定の軸へ分類する機能)を実装した。

清水[3]は、ゲームソフトのレビューサイトに投稿される口コミをテキストマイニング(クラスター分析, 形態素解析, テキスト分類, スコアリング)を用いて評判の分析を行った。本口コミ文に形態素解析を施し、手作業で分類, 辞書作成をした後に評判要素を抽出, 分析している。

goo 評判分析サービス[4]は評判情報インデクシングを利用して、weblogの記事を対象に「分析する」、「比較をする」、「関連語を探す」という機能をサービスしている。

<sup>†</sup> 愛知工業大学 情報科学部 情報科学科(豊田市)  
<sup>‡</sup> 株式会社リオ (名古屋市)

評判情報インデクシング技術では, 評判要素を「評価対象」, 「評価ポイント」, 「評価表現」として, 辞書を用いて分類, 分析している。

## 2.2 口コミによる企業の分析

企業の就業環境に関する口コミサイト, いわゆる転職サイトの口コミは, 整った文章で書かれており, 比較的長文で, 複数の話題に触れている場合が多く, 投稿者の経験を踏まえた明確な評価が述べられているという特徴を持つ。ユーザは口コミ内の大量の文から, それぞれ必要とする話題に着目し, その企業の評価を抽出する作業が必要になるが, その作業は簡単ではない。

本稿では, 転職サイトにおける特定の企業の口コミからその企業の特徴を分析する手法を提案する。具体的には, 口コミ内に含まれる投稿者の企業に対する評価を述べた文を抽出し, その内容の良し悪しを分類する手法と, 口コミ内の用語の傾向を分析することで企業の傾向とする手法の2つの手法を提案する。

## 3. 企業の評価の抽出

### 3.1 分析手順

本節では, 口コミ内に含まれる特徴語や評価語に注目し, 企業の特徴を端的に抽出する手法を考える。ユーザが最も欲しい情報の一つは, その企業についてユーザの興味のある特定の状況が“良い”か“悪い”である。企業に対する口コミには“給料”や“残業”などユーザの興味の対象となる企業の特徴を示す語(特徴語)や, その特徴に対して“多い”や“少ない”などはっきりとした評価を表す語(評価語)が含まれている場合が少なくない。しかし, 口コミ内の文章には様々な説明が含まれており, 特徴を捉えるには冗長な内容となる。従って, 特徴語と特徴語に対する適切な評価語の組み合わせが文中に含まれる場合,

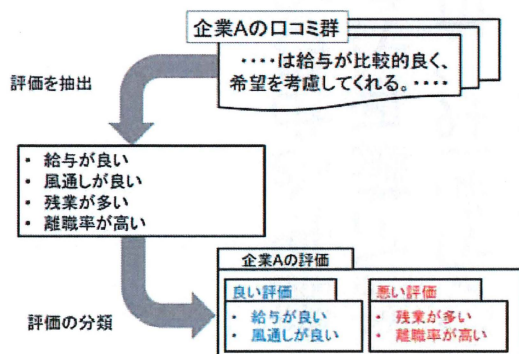


図 3.1 企業の評価抽出の概要

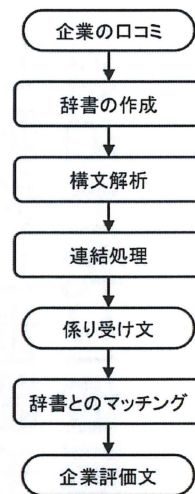


図 3.2 企業の評価抽出の流れ

その組み合わせのみを抽出することができれば, 企業の評価を簡単に入手できることになる。また, その評価内容の良し悪しを分類して集計することができれば, その企業が全体としてどのような印象を持たれているかを知ることができる。全体の概要を図 3.1 に示す。

本稿では, 企業の口コミに対して構文解析を行い, あらかじめ用意した特徴語と評価語が含まれるかどうかを調査する。必要な情報が含まれる場合, その内容を端的な文章として構成して出力する。全体の流れを図 3.2 に示す。

### 3.2 事前準備: 企業の特徴に関する辞書の作成

企業の口コミには, “給料”, “残業” など企業の特徴を示す特徴語が多く含まれている。また, その特徴語を説明する形容詞, “多い”や“良い”など, が含まれる場合, その組み合わせは企業の評価を表す文章になると考えられる。それぞれの特徴に対する評価となるには, 適切な単語の組み合わせで文章が構成されてなくてはならない。従って事前準備として, 企業の特徴を表すのに適切な関連する用語の辞書を作成する。

作成する辞書には, 企業の特徴となる単語(特徴語)とその単語を形容することで評価を表す単語(評価語)のペアを登録した。例えば, 辞書に登録した語句のペアとしては, (“残業”, “多い”), (“離職率”, “高い”) などがある。加えて, 各ペアの評価の良し悪しを区別するため, “良い”, “悪い” のラベルを付加する。

### 3.3 企業評価文の作成

分析にはまず, 口コミの構文解析を行い, 連結処理を行う。構文解析には日本語係り受け解析器 CaboCha[5]を利用

する。ロコミを句点で文に分解し、さらにその文を読点で分解して分析を行う。CaboChaの出力には文節番号と係り先番号が含まれるため、これらを利用して係り受け関係にある文節同士を繋げた文（以下、係り受け文と呼ぶ）を作成する。

作成した係り受け文と事前準備で作成した辞書との照合を行い、企業の特徴を表す内容かどうかを判別する。係り受け文のうち、辞書に含まれている特徴語と評価語のペアが共に含まれているものを抽出し、企業評価文の候補とする。候補の中には、係り受けの関係や辞書の組み合わせによって、評価文とならないものや、重複して抽出してしまう文があるため、それらを取り除いた結果を企業評価文とする。選出した企業評価文は評価の良し悪しを区別し結果として出力する。

#### 4. 企業の傾向の分析

##### 4.1 分析手順

ロコミから企業の傾向を分析するためには、特定の企業に対する全てのロコミをそれぞれがどのような話題であるかを分類し、話題毎のロコミの総数を比較することが必要である。今回、ロコミの自動分類には、SVM (support vector machine)を利用した機械学習による分類を用いた。

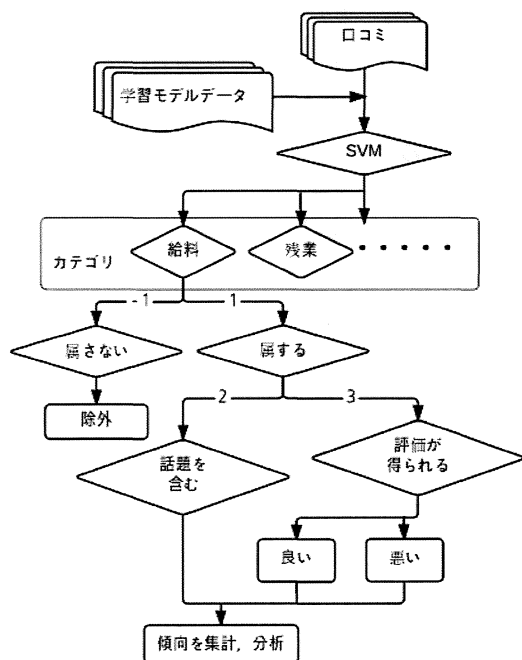


図 4.1 企業の傾向分析手法の概要

分析する内容は、カテゴリという形で予め話題を選定しておき、あるロコミがどのカテゴリに属するのか SVM を使って判定する。この際の事前準備としてカテゴリの選出、SVM への学習用モデルデータの作成、モデルデータ作成のための辞書作成が必要となる。

分析全体の流れを図 4.1 に示す。収集した企業のロコミ全てに対し、SVM を用いて 2 段階の処理により分類する。まず、そのロコミの話題が用意したカテゴリに属するか属さないかを分類する。次に、属すると分類されたロコミは形態素解析を行い辞書との比較により、そのカテゴリの話題を含むロコミにすぎないのか、企業に対するなんらかの評価が得られるロコミであるのかを分類する。次に、評価が得られると分類されたロコミは、企業に対する評価文を抽出した上で、“良い評価”と“悪い評価”に分類する。また、話題を含むロコミであった場合、直接的な評価はなくても企業の特長となる情報がいくらか得られる可能性が高いと考える。ここで、話題を含むロコミを“その他の評価”とし、それぞれのカテゴリで得られた“その他の評価”、“良い評価”、“悪い評価”の比率を比較することで、企業の傾向を知ることができる。

##### 4.2 事前準備：SVM による学習

ロコミから企業の傾向を得るためには、あるロコミがあるカテゴリに属するかを分類する必要がある。ロコミを分類するために、SVM を用いる。SVM とはパターン認識器の一つであり、与えられた未知のデータを、学習モデルを元に分類することができる。

今回は企業の特徴として 5 つのカテゴリ（残業、スキル、女性、給料、風通し）を用意し、それぞれのカテゴリで出現する単語を抽出した辞書を元に学習モデルを作成する。辞書の作成には Vorkers[6]を含む企業のロコミサイトから収集したロコミを利用する。それぞれのモデルデータには、3 つのラベル（属さない、話題を含む、評価が得られる）を付加する。学習に必要な入力データは、カテゴリごとに用意したサンプルのロコミに対して形態素解析を行って得られた単語の出現頻度を用いた。

##### 4.3 特徴の分類

4.2 節で作成したそれぞれの学習モデルを元に、収集した企業のロコミ全てに対してカテゴリの分類とロコミの抽出を行う。各ロコミは最終的に“カテゴリに属する”、“カテゴリに属さない”、“話題を含む”、“評価を得られる”の 4 つに分類される。この内の“評価が得られる”と分類されたロコミはさらに、3 章で述べた企業評価文抽出方法を用いて、“良い評価”と“悪い評価”に分類する。分類されたロコミの数をそれぞれ集計し図 4.2 のようにグラフ化することで、そのカテゴリの企業の傾向を示すことができる。

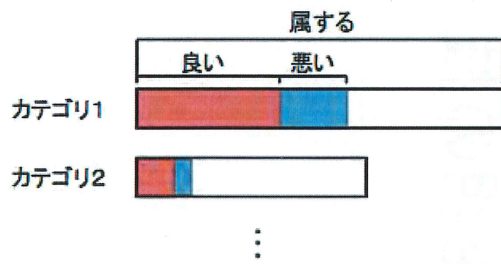


図 4.2 カテゴリ毎の可視化

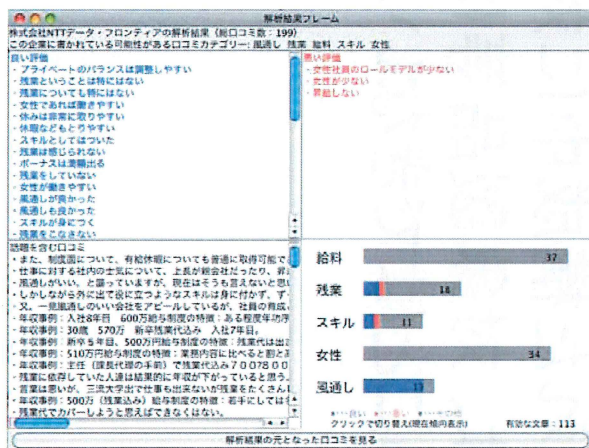


図 5.1 試作システムによる企業分析結果の例

## 5. 企業分析システムの試作

前述した2つの分析手法を実装した企業分析システムの試作を行った。企業の分析結果として3つの情報を表示することができる。その分析結果の一例を図5.1に示す。

①には、その企業の企業評価文が表示される。3章に示した提案手法により得ることのできた企業評価文を良い情報(青色)と悪い情報(赤い)で色分けし提示する。企業の長所や短所を、短時間で把握することが可能となる。

②には、話題を含む口コミが表示される。4章に示した提案手法によって話題を含むと判断された口コミ内の文章がそのまま抜き出して表示される。この文章を確認することで、対象となる話題について何らかの情報を得ることができ、企業の雰囲気を掴むのに役立つ。

③には、口コミ全体の集計結果がグラフ表示される。

②と同様に4章の分析結果を元に集計したカテゴリ毎の数値を集計し、積み上げ型の棒グラフとして表示する。青は良い評価、赤は悪い評価、灰色はその他の評価を表している。このグラフによって、どのカテゴリの口コミの割合が多いかを把握することができる。

評価文を見たとき、評価の元となった口コミを知りたくなる場合がある。そのような場合に、画面下部のボタンを押すことで、評価文の元となった口コミを見ることができる。元の口コミを見ることで、より詳細な情報を得られる可能性がある。また、元の口コミも①と同様に良い情報と悪い情報で分けて同じ順序で表示し、評価文との対応が容易に把握できる。

## 6. 結言

口コミサイトに投稿された内容が、ユーザの行動に大きな影響を与えていることはよく知られている。就職や転職を希望する者も企業に対する口コミサイトの情報を有益な情報として利用されている。しかし、多くの口コミに含まれる様々な意見を整理して全体としての印象を理解するのは手間のかかる作業となる。

本稿では、転職サイトの企業に対する口コミに対して、SVMを用いた文章の分類と特徴語と評価語の辞書を用いた特徴的な文章の抽出を行うことで、企業の特徴に相当する端的な情報を抽出する方法を示した。また、提案手法を実装し、登録されている口コミを分析、可視化する企業分析システムの試作を行った。今後はコーパスに利用する語の見直しや、辞書の充実を図り、分析精度を向上させる方法の検討が必要である。

## 参考文献

- [1] 長野伸一, 溝口由美子, 稲葉真純: インターネットから評判情報を抽出する口コミ分析技術ユビ de コミミハサンダー, 東芝レビュー, Vol.63, No.10, pp.35-38, 2008.
- [2] 立石健二, 福島俊一, 小林のぞみ, 上出将行, 高橋哲朗, 乾孝司, 藤田篤, 乾健太郎, 松本裕: Web 文書集合からの意見情報抽出と着眼点に基づく要約生成, 情報処理学会 情報学基礎研究会報告 2004(93), 1-8, 2004.
- [3] 清水航: レビューの語の重みを考慮したテキストマイニングによるゲームソフトの評判分析, 法政大学大学院工学研究科システム工学専攻, 修士論文, 2011.
- [4] goo 評判分析サービス, <http://blog.search.goo.ne.jp/wpa/>
- [5] CaboCha/南瓜, Yet Another Japanese Dependency Structure Analyzer, <http://code.google.com/p/cabocha/>
- [6] Vorkers, <http://www.vorkers.com/>