2D-2

BLOG のトラックバック構造における評価の可視化

石川　祥†　鈴木　佑介‡　関口　友樹‡　木村　昌臣‡
芝浦工業大学大学院†　芝浦工業大学‡

1.はじめに

近年普及しているブログを利用して個人が様々な話題に関する意見を Web 上で発信する機会が増えている．こうした意見の中には,商品やサービスに関する評判情報が多く含まれており，ブログを解析することによって製品の評価に関する情報を得られることが期待される．しかし，ブログの数は膨大であり,人手で評判情報を抽出するのは困難である．

そこで本研究では，ある話題に関するブログ記事をトラックバックをたどって自動で収集し，収集したブログ記事の本文から話題に対する評判情報を抽出するシステムを作成した．そして，トラックバック構造上に評判情報をマップすることにより話題に関する評価の分布を空間的に表現した．

2.システムの概要

2.1 トラックバックによる記事の収集方法

まず，収集したい話題のブログ記事（シード）を用意する．シードに対してトラックバックをしているブログ記事を収集する．さらに，その記事のデータを抽出し，記事の URL とそのトラックバック情報をブログ記事データベースに格納する．加えて，収集した記事にトラックバックをしている記事があれば同じ方法で収集し，以降この処理を繰り返す．ただし，すでに取得済みの記事にトラックバックをしている場合や，相互トラックバックがある場合にはループが出来て循環していまうため，その記事のトラックバックはたどることをやめる事とする．

2.2 記事のキーワードの抽出とグルーピング

記事の話題を取得するために記事のタイトルからキーワードを抽出する．まず，収集した記事からタイトルを取得し，茶筌[1]を使用して形態素解析を行い，名詞（代名詞と非自立の名詞は除く）を抽出する．さらに，全てのブログ記事のタイトルから名詞を抽出する．記事の内容をよく表し，全体としてよく現れる単語を取得するため，抽出された名詞の出現頻度をとり，その上位 10%を解析対象とする．こうして得られた単語は記事の内容を表すキーワードと考えられる．更に，関連した内容ごとにグループ化し，各グループを代表するキーワードでラベル付けを行った．これをもとに収集した記事のグループ化を行う．

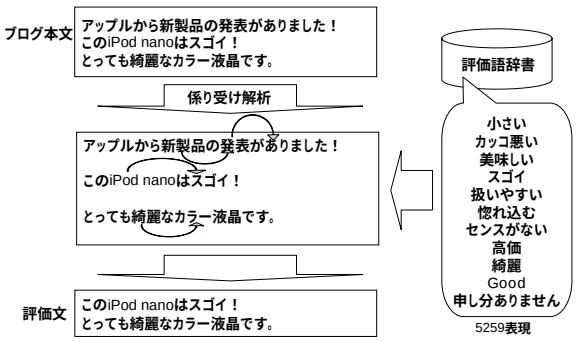

図 1.評価文の抽出方法

2.3 評価文抽出方法

商品やサービスに対する評価文として以下の二つの形式で記述されている文を抽出した．

1.<商品やサービスの名前の属性>「は/が/も」<評価語>
例）「<iPod nano>は<カッコいい！>」
2.<評価語> <商品やサービスの名前や属性>
例）「とても<オシャレな> <iPod>」

図 1 は評価文の抽出方法を示したものである．

まず，収集したブログ記事から本文を抽出し，南瓜[2]を用いて係り受け解析を行う．次に，係り受け情報の中から商品やサービスの名前や属性になり得る名詞または未知語に，助詞「は/が/も」が付属している文節の係り受け先の単語を評価語辞書と比較し，係り受け先の単語が辞書に含まれているならばその文を形式 1 の評価文として抽出する．また，商品やサービスの名前や属性になり得る名詞または未知語に対して係り受けをしている単語を評価語辞書と参照し，その単語が辞書に含まれているならば形式 2 の評価文として抽出する．なお，評価語辞書には小林らが作成した評価値表現辞書[3]を利用した．

2.4 トラックバック構造の可視化方法

可視化には我々が提案する水紋モデルを利用する．水紋モデルは次のステップで実現される．

Visualization of Distributions of Estimation in Trackback Structure of Weblog
†Sho Ishikawa　Graduate School of　Shibaura Institute of Technology
‡Yusuke Suzuki　Tomoki Sekiguchi　Masaomi Kimura
Shibaura Institute of Technology

1.始点となる根ノードを配置する.
2.根ノードとその各子ノードとの成す角度が均等になるように周囲に各子ノードを配置する.
3.配置した子ノードを self ノードとし self ノードとその親ノードを結ぶエッジを self ノード側に伸ばし，それを軸とした±45 度の範囲内に self ノードの各子ノードを均等に配置する.
4.全ノードが葉ノードになるまで 3.を繰り返す.

このモデルを利用するとトラックバックにより得られる記事の拡がり方を可視化することができる．更に，2.2 節で求めた記事のグループ化に基づいてノードの色を決定し，ブログ記事の主題の分布を視覚的に表現する．また， 2.3 節で求めた評価文に基づいてノードの色を決定し可視化をすることによってブログ記事の評価の分布を視覚的に表現する.

3.実験

2 章で説明した提案システムを用いて実験を行った．本実験では映画「ダ・ヴィンチ・コード」の話題の記事をシードとし，50000 件のブログ記事を収集し，トラックバック構造における話題の分布を可視化した.

さらに，収集したブログ記事からそれぞれの話題に対する評価文を抽出し，トラックバック構造と評判情報を特徴付けした可視化を行い評価の分布を調べた.

4.結果・考察

図 2 はトラックバック構造における話題の分布を可視化した結果である．可視化の結果から，中心のシードから近い所では，「ダ・ヴィンチ・コード」の記事が多く現れているが，トラックバックをたどるごとに，その他の映画の話題の記事が現れていることが分かる．このようにトッラクッバクによる話題の拡がり方を可視化によって表現した.

この中から「ダ・ヴィンチ・コード」と「博士の愛した数式」についての評価の分布を調べた．その結果（図 3）によると「ダ・ヴィンチ・コード」では「面白い映画」，「役者が良い」といった肯定的評価を表すノードと，「つまらない映画」や「展開が早い，難しい」といった否定的評価を表すノードが固まることなくバラバラに出現していることが見て取れる．一方，「博士の愛した数式」では，「良い映画」，「役者が良い」といった肯定的評価が，「つまらない映画」などの否定的評価よりも多く出現していて，同じ評価の記事同士が固まって現れている様子が見て取れる．これにより対象に応じてブログ記事の発信者の評価のバラつき度合いを直感的に理解することができる.

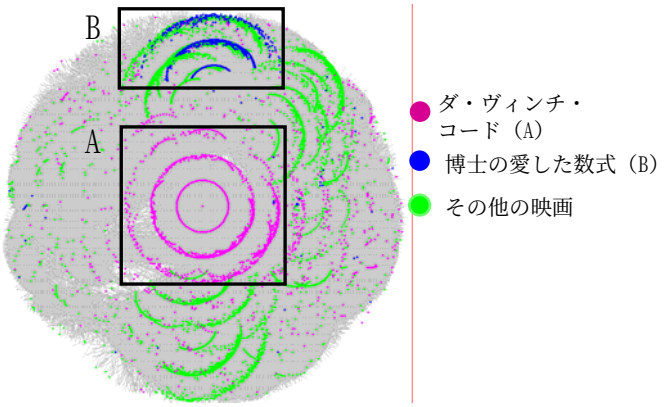

図 2.トラックバック構造における話題分布の可視化

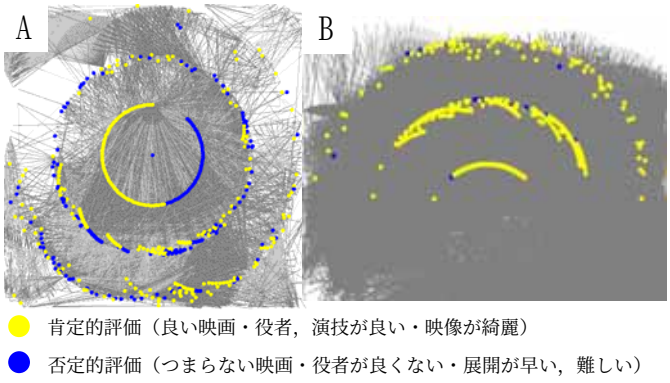

図 3.トラックバック構造における評価分布の可視化

また,評価文抽出の評価のため，無作為にブログ記事を 100 件選択し，人手による評価文の抽出と本システムとを比較した．人手で抽出された 452 件の評価文のうち本システムでは 341 件抽出でき，抽出精度は 75.4%であった.

5.まとめと今後の課題

本稿ではトラックバックを利用してブログ記事を収集し，トラックバック構造におけるそれぞれの話題についての評価の分布を得ることが出来ることを示した．今後は評価文の抽出精度の向上や評価対象をもとに評価の分布を検索し表示させるシステムの実現を目指す.

参考文献

[1] 形態素解析システム茶筌.
http://chasen.naist.jp/hiki/ChaSen/
[2] CaboCha/南瓜.
http://chasen.org/~taku/software/cabocha/
[3] 小林のぞみ，乾健太郎，松本裕治，立石健二，福島俊一．意見抽出のための評価表現の収集．自然言語処理，Vol.12，No.2，pp.203-222，2005.07
[4] 鈴木泰裕，高村大也，奥村 学:Weblog を対象とした評価表現抽出，人工知能学会，セマンティックウェブとオントロジー研究会，SIG-SWO-A401-02，2004.