2P-6

Blog からの評判抽出に関する研究

霜田　雄一†　　成田　祐一‡
日本大学大学院工学研究科†　　日本大学工学部‡

1．はじめに

現在、インターネット上では Blog と呼ばれる日記型のサービスが一般ユーザに爆発的に普及してきている。平成 17 年 5 月に出された総務省の資料[1]によると、今後も右肩上がりで Blog 利用者数が増加すると予想されている。この Blog に書かれている情報を活用しようと試みが増加しており、Blog を対象とした検索エンジンとして、BlogWatcher[2]等の Blog 検索エンジンが登場してきている。これらは、興味のある Blog を検索できる。また、ホットキーワードの表示や評判情報検索[3]を行えるものもある。Blog の特徴は、一般ユーザでも簡単に更新することができる。即時性が高く書き込みの内容が社会的なイベントと連動する傾向がある。

一方、世の中の動向を知る上で我々の指標となるものに世論調査がある。しかし、世論調査は、人手による作業なので集計するまでに時間がかかってしまうという問題がある。

以上の背景を踏まえ、本研究では、Blog を大衆の社会的な出来事に対する意見が即座に反映される情報源として捉え、Blog 全体から評判情報検索することにより、Blog の意見を判別し自動的に世論調査を行うシステムを提案する。

2．提案手法

2.1．評価表現辞書

評判情報検索を行う上で、評価表現を抽出する辞書はシステムの精度を決める大きな鍵となる。既存の研究では、一般的な辞書を作成するあまりに特定の分野の評判情報検索を行うことが難しくなっている。そこで、Blog から評判情報検索を行うユーザは検索を行うキーワードに対してある程度の基礎知識があると仮定し、評価表現辞書をユーザ自身に入力してもらう。世論調査という目的の為「肯定」「否定」のように、対比させることが出来る辞書を複数作成させる。

A study of the reputation extraction from weblogs
†Yuichi SHIMOTA
‡Yuichi NARITA
†Graduate School of Engineering, Nihon University
‡Callege of Engineering, Nihon University

2.2．フィルタリング

広告や新聞記事を引用しただけの Blog 等は客観的な情報であり、検索対象から外れる。また、Blog は日記として書かれているものが多く、ネガティブな Blog に対してフィルタリングが必要となる。本手法では以下のフィルタリングをかける。

2.2.1．Blog 本文の切出し

RSS を解析して収集した Blog を対象に、評判を抽出するとノイズが多く入ってしまう。その原因は、殆どの Blog は、図 1 のような構造をしており、本文以外にも広告や過去の Blog のタイトル等が書き込まれている為である。そこで、正規表現により Blog 本文を切出す。

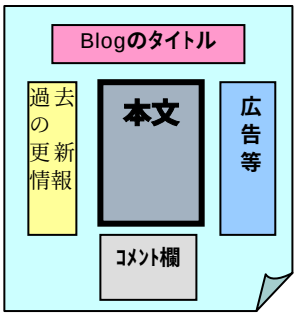

図 1．Blog の構造

2.2.2．ノイズとなる Blog のフィルタリング

Blog を評価する上で、Blog に書かれている情報にどれだけ信頼があるか判定する必要がある。しかし、信頼度を数字で表すのは困難であり明らかなノイズを排除する方式を採る。ある社会的な出来事に対して意見を述べている Blog には、ある範囲内の文章量で書かれいる傾向があり、Blog 本文の長さをフィルタリングの要素とする。また、同一筆者による更新頻度が極端に多い Blog は、何らかの広告である可能性がかなり高く、排除する。

2.3．世論調査における Blog の判定方法

世論調査の質問内容は、作成したインターフェースからユーザによって入力されたキーワードと作成された複数の評価表現辞書により作成する。

キーワードの TF 値が 1 以上の Blog に対して質問についてどのような答えを持っているか判定する。ユーザにより作成された辞書毎に評価値を算出する。評価値を Score、1 つのある評価表現辞書の評価表現の出現数を RF、パラメータとしてα、β、γを与えたとき、ある Blog の作成さ

れたある評価表現辞書の評価値を以下の式で算出し、意見を判別する。

$$Score = \alpha\, TF + \beta\, RF + \sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{\gamma}{keyword_i \text{と} repute_j \text{の距離}}\right)$$

$$(keyword_i \text{と} repute_j \text{の距離} > 0) \quad (1)$$

今回は式(1)のαを 20、βを 10、γを 1000 として評価値を算出した。作成された辞書の中で、評価値が一番高い評価表現辞書の名前を質問の答えとして判定する。また、RF が 0 の場合はキーワードに対して意見を持っていない Blog とみなし、判定を行わないものとする。

3．システム概要

本システムは、提案した手法を用いて自動世論調査をする。本システムの構成図を図 2 に示し、流れを説明する。

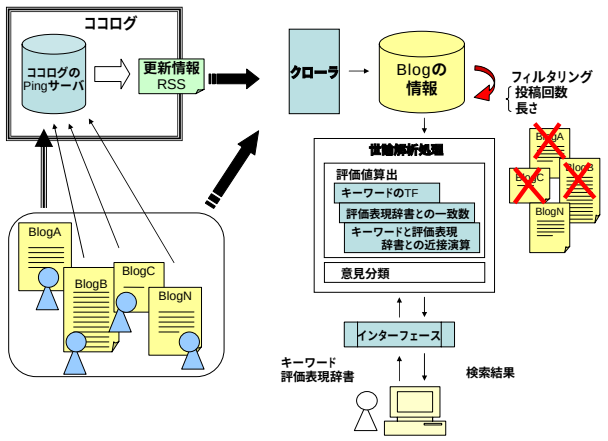

図 2.システムの概要図

① RSS ファイルを定期的に解析し ping サーバに投稿された全 Blog の本文と RSS の内容をデータベースに蓄積する。
② ノイズとなる Blog をフィルタリングする。
③ ユーザにキーワードと複数の評価表現辞書の名前と評価表現を入力してもらう。
④ ユーザの入力と蓄積したデータを基に式(1)により Blog を判定し、結果を表示する。

4．評価実験と今後の課題

4.1．評価実験と結果

ココログ[4]の RSS ファイルを用いて、2005 年 12 月中に収集した Blog を対象として評価実験を行った。評価実験に用いたキーワードは「自民党」である。またユーザの入力として作成した評価表現辞書の例を表 1、表 2 に示す。そして、これらを用いて本システムで調査した結果を表 3 に示す。尚、本手法を用いたシステムにおける意見をもつ Blog を抽出する精度は先行研究[5]で評価してあり、その精度は 65%である。

表 3 は、12 月に投稿された Blog の中で、「自民党」について「賛成」していると思われる Blog 数と「反対」していると思われる Blog 数を週単位と月単位で示したものである。

表 1. 「賛成」の評価表現の入力例

賛成	支持する

表 2. 「反対」の評価表現の入力例

反対	支持しない	支持できない

表 3. 評価実験の結果

期間	賛成	反対
1 ヵ月間	11	39
検索日-1 週間前	4	10
1 週間前-2 週間前	1	10
2 週間前-3 週間前	3	7
3 週間前-4 週間前	1	5

4.2 今後の課題

評価実験の結果、表 3 のように自動世論調査をすることが出来た。しかしながら、本システムは、評判情報検索の精度により評価が変わってしまう。今後は評判情報検索の精度の向上を図って行く必要がある。また、ユーザが簡単に評価表現辞書を作成できる手法を考えていく必要がある。

5．まとめ

本論文では、Blog からの評判抽出の研究として、Blog 全体から評判情報検索することにより、自動的に世論調査を行うシステムの提案をした。本システムを用いることにより、Blog から自動的に世論調査を行うことが出来た。今後は現在分かっている問題点を分析していき、システムを改善していく。

参考文献

[1] ブログ･SNS の現状分析及び将来予測(総務省) http://www.soumu.go.jp/s-news/2005/pdf/050517_3_1.pdf
[2]BlogWatcher http://blogwatcher.pi.titech.ac.jp/
[3] 立石健二他,インターネットからの評判情報検索,人工知能学会誌,Vol19,No.3,pp317-323,(2004)
[4]ココログ http://www.cocolog-nifty.com/
[5]霜田雄一,成田祐一,Blog からの評判抽出システムの構築に関する研究,平成 17 年度第 4 回情報処理学会東北支部研究会(4),(2006)