ブログ中のリンクの評価極性判定

石野亜耶　　難波英嗣　　竹澤寿幸
広島市立大学大学院　情報科学研究科
{ishino, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

1. はじめに

一般の個人が、ネット上に手軽に情報を配信できる環境が整うとともに、個人が配信する情報の重要性が広く認知されるようになってきている。一方で、個人が配信する情報は玉石混交であるため、膨大な情報の中から、信頼性の高い情報をいかに効率的に見つけるかが、重要な課題となっている。本研究では、個人が意見を述べる機会の多いブログに焦点を当て、信頼性の高いブログを効率的に見つけるシステムの構築を目標としている。

現在、検索システムが行っている被リンク回数に基づいたページの順位付けのアルゴリズムでは、個人がどのような意見をもってリンクをしたかという情報が反映されていない。よって、内容とは関係ない誹謗中傷が集中しているブログなど、被リンク回数は多いが、信頼性が高いとはいえないサイトが、上位に表示されるという事態が起こっている。この問題を解決するため、本研究では、リンク元ブログの著者が、リンク先ブログに対して、どのような意見を持ってリンクをしたのか(評価極性)という点に注目する。図 1 は、ブログのリンク関係の一例である。ブログ A とブログ B は、共に被リンク件数は 3 件である。しかし、リンクの評価極性に注目すると、ブログの A の方がポジティブな評価を得ていることがわかるため、ブログ A の方が信頼性の高いブログだと判断できる。そこで本研究では、リンク先ブログに対する評価が書かれている個所(評価個所)に含まれる情報を用いて、リンクの評価極性を判定する手法を提案する。このように、リンクの評価極性を考慮することで、より信頼性の高いブログを発見することができると考えられる。

本論文の構成は以下の通りである。2 節では関連研究、3 節では提案手法、4 節では実験と結果について述べ、5 節で本稿をまとめる。

2. 関連研究

ブログ中に含まれるリンクの分類に関連する研究について述べる。リンクの分類の手法は、主に、Kale ら[Kale et al. 2007]の評価表現の比率に基づく手法(Kale 手法)と、Martineau ら[Martineau et al. 2008]の機械学習に基づく手法(Martineau 手法)に分けられる。

Kale らは、評価個所に含まれる肯定的あるいは否定的な単語(評価表現)の割合によって、リンクの評価極性を、ポジティブまたはネガティブに分類する手法を提案している。Kale らは、リンクの前後 X 文字を評価個所として使用しているが、リンク先ブログに対する評価は、リンクから離れた個所に記述

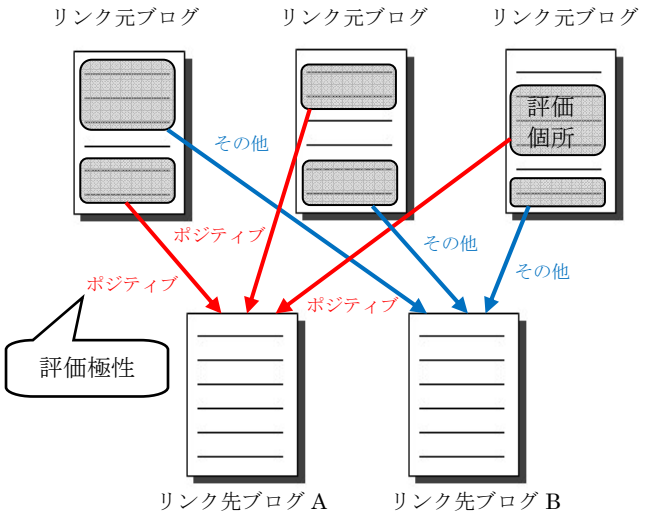

図 1　ブログのリンク関係

される場合もある。ブログ中から評価個所を適切に抽出することで、リンクの評価極性を正確に判定することが可能になると考えられる。よって本研究では、リンク先ブログに対する評価個所を、手掛かり語を用いて自動的に抽出する手法を提案する。

また、Kale らは、リンクの評価極性を判定する際に、評価表現を使用している。日本語の評価表現が登録されている辞書に、鍛治ら[鍛治他 2007]の評価表現辞書がある。評価表現辞書は、大規模な評価文コーパスにより自動構築されたものであり、形容詞、形容詞句と評価極性値のペアが約 10,000 組登録されている。評価極性値とは評価極性の強さを表す指標であり、この値が 0 より高いとポジティブな極性、低いとネガティブな極性を持つことになる。評価表現辞書は、形容詞・形容詞句のみが集められた辞書であるため、リンク先ブログに肯定的な評価を与える際によく使われる「お勧め」などの評価表現は登録されていない。このような問題を解決するため、本研究では、ブログ中のリンクの極性判定に特化した評価表現を登録した辞書であるリンク評価辞書の作成を行う。

Martineau らは、ブログ中のリンクについて、様々な観点から分類を試みている。実験では単語を素性とし、機械学習 SVM を用いて分類器を構築している。リンクの評価極性判定においても、機械学習は有効な手段だと考えられる。よって、本研究では、上記の Kale 手法、Martineau 手法それぞれに、リンク評価辞書を用いることで、リンクの評価極性判定を行う手法を提案する。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

3. ブログ中のリンクの評価極性判定

本節では、人手によるリンクの評価極性判定の手順、および提案手法の流れについて説明を行う。リンクの評価極性判定の手順については、3.1 節で説明を行う。本研究で提案する、ブログ中のリンクの評価極性判定の流れを以下に示す。これらの 2 つのステップについては 3.2、3.3 節でそれぞれ説明する。

(1) 評価個所の抽出
(2) 評価個所を用いたリンクの評価極性判定

3.1. リンクの評価極性判定の手順

一般的にリンクの評価極性は、ポジティブ、ネガティブ、肯定的とも否定的とも評価されていないものをその他として扱う。本研究で収集したブログ 840 件中、人手でリンクの評価極性がネガティブと判断されたブログは 5 件のみであった。よって本研究では、ネガティブをその他と合わせ、リンクの評価極性をポジティブまたはその他で判定する。

図 2 は、7 文目に他のブログへのリンクを含むブログの例である。ブログの記載内容から、リンク先ブログのタイトルが、「キュウママ日記」であることがわかる。リンク先ブログに対して、1 文目に、「魅力的な水彩画がいっぱい」という肯定的な評価表現が記述されているため、リンクの評価極性はポジティブと判定できる。本研究では、リンクの評価極性は、評価個所に肯定的な評価表現が含まれていればポジティブ、含まれていなければその他と判定する。

3.2. 評価個所の抽出

評価個所の抽出とは、リンク先ブログに対する評価が書かれている文を、リンク元ブログの文から抽出する処理である。本研究では、リンク先ブログに対する評価個所を、手掛かり語を用いて自動で抽出する手法を提案する。本節では、引用個所の自動抽出の基本的なルールについて説明を行う。

ブログを紹介する際には、リンク先ブログのタイトルを“「」”や“『』”などの記号で囲む場合が多くある。また、リンク先ブログを紹介する際に、“紹介”、“の HP”などの語が使われるため、これらの語を手掛かり語として使用する。以下に、引用個所の抽出ルールを示す。

1 「キュウママ日記」には、魅力的な水彩画がいっぱい！
2 季節の変化に合わせて私たちの目を楽しませてくれる野の花たち。
3 野の花には、人の手で丹精込めて育てられた花々にはない、独特な風情がありますよね。
・・・・・・・(略)・・・・・・・
4 芯の強さを秘めながらも、繊細で儚げに人知れず花を咲かせる野の花たち。
5 そんな野の花の特徴をよくとらえたキュウママさんの透明水彩画と楽しい記事は、こちらでご覧になれます。
6 「キュウママ日記」⇒
7 <ahref="http://blogs.yahoo.co.jp/kyumamanikoniko">
http://blogs.yahoo.co.jp/kyumamanikoniko

図 2 リンクを含むブログ

(1) リンクが含まれている文、およびの前後 X 文を抽出する。(予備実験より X=2 とする。)
(2) 手順(1)で抽出された文に、手掛かり語が含まれていれば、手掛かり語の周辺文字列を、リンク先ブログを指し示す語(Keyword)として抽出する。例えば、手掛かり語“「」”、“のブログ”を含む“「キュウママ日記」”、“A さんのブログ”という語があれば、Keyword は“キュウママ日記”、“A さん”となる。
(3) Keyword が含まれている文を抽出する。

図 2 のブログを用いて、評価個所抽出ルールを説明する。ルール(1)により、リンクが含まれている文(7 文目)と前後 2 文(5、6 文目)を評価個所として抽出する。ルール(2)により、6 文目に“「」”という手掛かり語が含まれているため、“キュウママ日記”を Keyword とする。ルール 3 により、Keyword である“キュウママ日記”という単語が含まれている 1 文目を、引用個所として抽出する。よって、図 2 のブログから抽出される評価個所は、1、5、6、7 文目である。

3.3. 評価個所を用いたリンクの評価極性判定

本研究では、リンク先ブログを評価する際によく使われる評価表現を収集し、リンクの極性判定に特化したリンク評価辞書の作成を行う。作成手順を以下に示す。以下の手順により収集した 135 個の評価表現が、リンク評価辞書に登録されている。

(1) Web 5 億文データやブログデータから、「このブログは」という表現が含まれている文を収集する。
(2) ブログに対しポジティブな評価をあたえているのか、その他なのかを一文ずつ人手で判定する。評価が行われた文例を表 1 に示す。
(3) 手順(2)でポジティブと判定した文から、評価表現を人手で抽出し、リンク評価辞書に登録する。表 1 の例からは、二重下線で示した、「オススメ」、「情報満載」という評価表現が抽出される。

本研究では Kale 手法、Martineau 手法にリンク評価辞書を使用することで、リンクの評価極性判定を行う。Kale 手法では、評価個所に、リンク評価辞書に登録されている評価表現が含まれている場合には、リンクの評価極性はポジティブ、含まれていない場合にはその他と判断する。また、Martineau 手法では、機械学習に、リンク評価辞書に登録されている評価表現の有無を素性として与える。

表 1 手順(2)により判定された文例

評価極性	文例
ポジティブ	このブログはマジでオススメである。
	このブログは情報満載なのでリンクします。
その他	このブログは明確に「日記だ」といってますねえ。
	このブログは参加しにくい。

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

4. 実験

本章では、提案手法の有効性を示すためのいくつかの実験を行い、その結果を示す。本研究では、2 種類の実験を行った。実験 1 では、3.3 節の手順により作成したリンク評価辞書の有効性を確かめるための実験を行った。また、実験 2 では、3.2 節で提案した評価個所の抽出手法の有効性を確かめるための実験を行った。

4.1. データセットと評価尺度

ブログ中に含まれる 840 件のリンクに対して、人手による評価極性判定を行った結果を、実験に利用した。人手によるリンクの評価極性判定の結果を表 2 に示す。それぞれのリンクを含むブログには、評価個所を人手で抽出したデータ(ブログデータ 1)と、評価個所を抽出していないデータ(ブログデータ 2)がある。評価には、精度、再現率、F 値を使用した。

表 2　人手によるリンクの評価極性判定の結果

評価極性	ポジティブ	その他	合計
リンク件数	378	462	840

4.2. 実験 1: リンク評価辞書の評価

実験 1 では、3.3 節の提案手法により作成したリンク評価辞書の有効性を確かめるため、ブログデータ 1 を用いて実験を行った。

4.2.1. 実験手法

リンクの分類の手法は主に、Kale らの評価表現の比率に基づく手法(Kale 手法)と、Martineau らの機械学習に基づく手法(Martineau 手法)に分けられる。以下に実験 1 の手法について説明を行う。

■Kale 手法

Kale らの手法を用いて実験を行った。

Kale_LinkDic（提案手法）：本研究で作成したリンク評価辞書を用いて実験を行った。

Kale_SentDic[y]：評価表現辞書を用いて実験を行った。評価表現辞書には評価表現と評価極性値が対になって登録されている。この評価極性値に制限 y を与えることで、評価表現の利用範囲を指定する。例えば y=5 のとき、評価極性値が−5 以下または 5 以上の評価表現を使用する。評価個所に出現する評価表現の評価極性値の合計が 0 より大きければポジティブ、それ以外であればその他と判定した。

■Martineau 手法

Martineau らの手法を用いて実験を行った。機械学習には TinySVM を用いた。2 次の多項式カーネルを使用し、4 分割交差検定を行った。

Martineau_LinkDic（提案手法）：本研究で作成したリンク評価辞書に登録されている評価表現の有無を素性として与えた。

Martineau_SentDic（提案手法）：評価表現辞書に登録されている評価表現の有無を素性として与えた。

Martineau_Base：評価表現辞書、リンク評価辞書の有効性を確かめるため、辞書を素性として与えずに実験を行った。素性には、単語のみを使用した。

4.2.2. 実験結果と考察

実験 1 のリンク評価辞書の評価実験の結果を表 3 に示す。本研究では、不正確な評価が増えるより、正確な評価を得る方が情報源の評価として望ましいと考える。よって表 3 には、Kale_SentDic[y]の実験においては、最も精度の高い Kale_SentDic[11]と、最も F 値の高い Kale_SentDic[3]を記載した。

Kale 手法に、リンク評価辞書を用いた実験(Kale_LinkDic)が、精度・再現率ともに最も高い数値を記録した。また Martineau 手法においても、リンク評価辞書を素性として与えた実験(Martineau_LinkDic)が、精度・再現率ともに最も高い数値を記録した。これらの結果から、リンク評価辞書の有効性を示すことができたといえる。また、Kale_LinkDic と Martineau_LinkDic を比較すると、Kale_LinkDic の方が良い結果を得ることができた。ブログ著者がリンク先ブログを紹介する際には、「お勧め」や「ステキ」のような特定の評価表現が使われる場合が多い。そのため Martineau 手法よりも、評価表現の有無のみで評価極性を判定する Kale 手法の方が、良い結果が得られたと考えられる。

実験結果の最も良い Kale_LinkDic の、判定誤りの主な原因としては、表層的な情報のみを使用した

表 3　実験 1: リンク評価辞書の評価実験の結果

	実験手法	精度(%)	再現率(%)	F 値(%)
Kale 手法	**Kale_LinkDic（提案手法）**	85.2	90.3	87.8
	Kale_SentDic[11]	72.0	9.5	40.8
	Kale_SentDic[3]	55.9	67.7	61.8
Martineau 手法	**Martineau_LinkDic（提案手法）**	81.7	74.8	78.3
	Martineau_SentDic（提案手法）	78.8	74.1	76.5
	Martineau_Base	78.0	71.6	74.8

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved

手法の限界が挙げられる。Kale_LinkDic による失敗例を図 3 に示す。図 3 で示した失敗例のブログには、リンク先ブログに対する評価が書かれていないため、人手によるリンクの評価極性はその他と判断される。しかし失敗例のブログには、リンク評価辞書に登録されている「ステキ」という評価表現が含まれているため、Kale_LinkDic によるリンクの評価極性はポジティブと判定された。

先日いつも仲良くしていただいている
テンファミリー＋オチビの「もりりんさん」から、　ステキなプレゼントが届きました～～

図 3　Kale_LinkDic による失敗例

リンク評価辞書は、リンク先ブログに対する評価によく使用される評価表現を登録した辞書である。リンク評価辞書を使用することで、ブログに対する評価を表す評価表現をもれなく使うことができるため再現率を向上させる効果と、ブログ以外の評価によく使われる評価表現を排除することで精度を向上させる効果があったと考えられる。しかし失敗例に含まれる「ステキ」という評価表現は、リンク先ブログに対する評価にもよく使われるが、物や人物など幅広い対象にもよく使われる評価表現である。よってこの問題を解決するためには、何に対して評価表現が使われているのかを解析するなど、より深い意味処理が必要不可欠である。

4.3. 実験 2: 評価個所抽出手法の評価

実験 2 では、3.2 節で提案した評価個所の抽出手法の有効性を確かめるため、ブログデータ 2 を用いて実験を行う。

4.3.1. 実験手法

Kale_LinkDic_Auto として、評価個所抽出ルールを用いて、ブログデータ 2 から評価個所を自動で抽出し、評価極性を判定する実験を行った。評価極性の判定は、4.2 節で述べた、人手で抽出した評価個所を使用した場合に、リンクの評価極性判定の実験結果が最も良い Kale_LinkDic を用いた。

4.3.2. 実験結果と考察

実験 2 の Kale_LinkDic_Auto では精度 86.5%、再現率 87.6%とともに高い数値を記録した。評価個所抽出ルールを使用して抽出した評価個所を用いた実験(Kale_LinkDic_Auto)と、人手により抽出した評価個所を用いた実験(Kale_LinkDic)の実験結果の比較を行うため、実験結果を表 4 にまとめる。Kale_LinkDic_Auto は、Kale_LinkDic よりも再現率が 2.7 ポイント低下したが、ほぼ同程度の実験結果が得られたといえる。よって本研究で作成した評価個所の抽出手法は有効であるといえる。

本研究で提案した評価個所抽出ルールは、手掛かり語を利用し、リンク先ブログを指し示す語(Keyword)を抽出することで評価個所を抽出した。しかし、ブログ著者により Keyword の記述方法が大きく異なるため、Keyword が正確に抽出できない場合がある。また、評価個所全てに Keyword が出現しているわけではない。よって現段階の評価個所抽出ルールでは、評価個所の抽出にもれが出る可能性がある。この問題を解決するためには、リンク元ブログの記述内容だけではなく、リンク先ブログの記述内容に注目する必要がある。例えば、リンク先ブログとリンク元ブログに出現する単語を比較し内容の一致度を測ることで、より正確に評価個所の抽出ができるようになると考えられる。

表 4　Kale_LinkDic_Auto と Kale_LinkDic の比較

	精度 (%)	再現率 (%)	F 値 (%)
Kale_LinkDic_Auto (提案手法)	86.5	87.6	87.1
Kale_LinkDic (提案手法)	85.2	90.3	87.8

5. おわりに

本研究では、ブログ中のリンクの評価極性判定を行うための手法を提案した。この提案手法の流れは(1)評価個所の抽出、(2)評価個所を用いたリンクの評価極性判定という 2 つのステップに分かれている。

まず、(2)のリンクの評価極性判定について述べる。(2)では、リンクの評価極性判定に特化したリンク評価辞書の作成を行った。リンク評価辞書を Kale 手法に使用した実験(Kale_LinkDic)では精度 85.2 %、再現率 90.3%という結果を得ることができ、リンク評価辞書の有効性を示すことができた。

次に、(1)の評価個所の抽出について述べる。(1)では、手掛かり語を用いた評価個所の抽出ルール作成し、評価個所を自動で抽出する手法を提案した。評価個所抽出ルールにより抽出された評価個所のデータを用いて行った実験(Kale_LinkDic_Auto)では、人手により抽出された評価個所を用いた実験(Kale_LinkDic)の結果とほぼ同程度の結果を得た。よって、本研究で提案した評価個所抽出ルールの有効性を示すことができたといえる。

参考文献

[Kale et al. 2007] Anubhav Kale, Amit Karandikar, Pranam Kolari, Akshay Java, Tim Finin and Anupam Joshi. “Modeling Trust and Influence in the Blogosphere Using Link Polarity”, Proceedings of International Conference on Weblogs and Social Media, 2007.

[Martineau et al. 2008] Justin Martineau and Matthew Hurst. “Blog Link Classification”, Proceedings of International Conference on Weblogs and Social Media, 2008.

[鍛治他 2007] 鍛治伸裕, 喜連川優. “自動構築した評価文コーパスからの評価表現辞書の構築”, 日本データベース学会 Letters, Vol.6, No.1, pp.41-44, 2007.

Copyright(C) 2012 The Association for Natural Language Processing.
All Rights Reserved