Blog や Twitter に書かれた疑問を収集・提供する Web サイトの構築 *

荒井俊介 （学籍番号 200921723）
研究指導教員：辻慶太
副研究指導教員：芳鐘冬樹

1．はじめに

近年，Blog や Twitter のような Web 日記の利用者が増加してきている。この日記には著者の考えや意見など様々な事が書かれるが，本研究ではその中の日常的な疑問，特に「内容は分かるがタイトルが思い出せない」という疑問の書かれた日記記事に注目する。このような記事が誰かから回答を得る為には，その記事が回答を知っている人の目にとまる必要がある。だが，個々の Web 日記の記事を読む人の数は限られているため，疑問の多くは回答が得られないまま終わってしまう。

そこで本研究では Blog や Twitter に書かれた上記のような疑問を収集し，広く回答を呼び掛ける Web サイトを構築する。本サイトによって疑問の書かれた記事が回答を知る人の目にとまるようになり，回答が与えられるようになることを目指す。

本研究の意義として以下の 2 点が挙げられる。(1) 日記著者の疑問を解消できる，(2) 日記著者と回答者は同じ問題関心を持っている可能性が高く，回答を通じて新たな交流が生まれるかもしれない。

2．本サイトの有用性に関する予備的調査

本研究が構築を進めるサイトに関しては以下の3つの問題が憂慮される。即ち，(1) 疑問の書かれた記事の収集が収集困難なほど稀である，(2) 疑問の書かれた記事は回答のためのヒントが少なく，回答を行う事は困難である，(3) 疑問を書いた日記著者は，見知らぬ者からの回答を気味悪く感じ感謝することがない，という 3 点である。これらの問題を検証する為に以下の調査を行った。

(1)Google ブログ検索，Yahoo!ブログ検索で「本 タイトル 思い出せない」，「本 題名 思い出せない」をキーワードとして各検索結果の上位 200 記事の中に疑問の書かれた Blog 記事が何記事含まれているかを調べる，(2) 上で得られた疑問を Q&A サイトとデジタルレファレンスサービスに質問し回答可能性を調べる，(3) 上で得られた回答を疑問の書かれた記事にコメントとして書き込み日記著者の反応を確認する。

この結果，(1) 疑問の書かれた記事は全 800 記事中 16 記事発見出来た。(2) 上で得られた 16 記事の中の回答の出ていなかった 13 記事に関して質問を行い，Q&A サイトでは 6 記事，デジタルレファレンスサービスでは 8 記事で回答を得る事が出来た，(3) 上で得られた 8 記事の回答に関してコメント欄へ書き込みを行い，5 記事で日記著者から感謝のコメントを得る事が出来た。また感謝のコメントの無かった 3 記事は Blog の更新が止まっていたため，この回答を見ていない可能性が高い。

3．疑問の収集

予備的調査において疑問の書かれた記事は 800 記事中 16 記事にとどまった。Web 全体を考えれば相当数存在すると考えられるが，その収集には多くのコストがかかる事が予想される。そこで本研究では疑問の書かれた記事を以下の 2 ステップによって収集する。即ち(1)特徴的な表現による検索，(2)テキスト自動分類による抽出，の 2 つである。

3．1 特徴的な表現による検索

本研究において特徴的な表現とは，疑問の書かれた記事に多く現れるがそれ以外の記事にはあまり現れない単語列を指す。このような単語列をキーワードとしてサーチエンジンで検索を行えば検索結果の中に疑問の書かれた記事だけを多く含める事ができるであろう。

特徴的な表現を導出するため，学習用コーパスとして Twitter，Blog のそれぞれに関して，疑問の書かれた記事 100 個，それ以外の記事 100 個を用いた。このコーパスを形態素解析ソフト Mecab で分解し，コーパス内に現れる単語列の相対出現頻度

*“Developing a Website which Collects and Shows Questions Posted in Blogs and Twitter” by Syunsuke ARAI

の差によって特徴的な表現を導出する。

この方法の効率を確かめる為に，導出された特徴的な表現を用いてサーチエンジン（Blog に関しては Google ブログ検索と Yahoo!ブログ検索，Twitter に関しては Twitter 検索を用いる。）で検索を行い，検索結果の上位 100 記事の中に疑問の書かれた記事が何記事含まれているかを調べた。

3．2 テキスト自動分類による抽出

疑問の書かれた記事とそれ以外の記事を機械的に分類する事が出来れば 2 つの記事が混ざった記事集合から疑問の書かれた記事だけを収集する事が出来る。このような問題を解く手法としてテキスト自動分類がある。本研究では決定木，ブースティング，Naive Bayes，SVM の 4 つのテキスト自動分類手法を用いて分類を行い，性能を比較した。この分類にはデータマイニングツール Weka を用いた。

学習用コーパスには，Blog に関しては疑問の書かれた記事とそれ以外の記事をそれぞれ 150 個ずつ，Twitter に関してはそれぞれ 800 個ずつ用いた。このコーパスを形態素解析ソフト Mecab で分解し，(1) 1 単語を素性とした場合，(2) 2 単語列を素性とした場合，(3) 3 単語列を素性とした場合，の 3 つの場合の分類性能を調べた。

本研究では疑問の書かれた記事とそれ以外の記事の分類における F 値を導出し，その加重平均を評価尺度として用いた。F 値には 10 回交差検定を行った上でその平均値を用いた。

4．実験結果と Web サイトの構築

4．1 特徴的な表現による検索

導出した特徴的な表現の中で最も精度が良かったのは，Twitter に関しては「タイトルが思い出せない」で，100 記事中 13 記事，Blog に関しても「タイトルが思い出せない」で，Google ブログ検索では 16 記事，Yahoo!ブログ検索では 19 記事であった。

2 節の予備的調査において，「本 タイトル 思い出せない」を検索ワードとして用いた場合，Google ブログ検索，Yahoo!ブログ検索のいずれを用いても，検索結果の上位 100 記事中に 2 記事しか疑問の書かれた記事を発見する事が出来なかった事を考えると，「タイトルが思い出せない」という表現を用いる事で疑問の書かれた記事の収集効率を向上させる事が出来たと言える。

4．2 テキスト自動分類による抽出

Twitter に関して最も性能の良かった分類手法は素性として 1 単語を用い，決定木によって分類を行った場合であり，その性能は，精度 94.8%，再現率 94.3%，F 値 0.941 であった。また，Blog に関して最も性能の良かった分類手法は素性として 1 単語を用い，ブースティングによって分類を行った場合であり，その性能は，精度 94.3%，再現率 94.3%，F 値 0.943 であった。

これらの分類手法を用いる事で，高い精度で疑問の書かれた記事だけを抽出する事が可能になった。

4．3 構築した Web サイトの有用性

最後に，収集した疑問の書かれた記事を Web サイトに提示し，回答を呼び掛ける事でどの程度回答が行われ，感謝が返されるかを調査した。

Twitter に関して 31 記事，Blog に関して 30 記事を提示し，Twitter 上で多くのフォロワーを持つ知人に宣伝を依頼する形で呼び掛けを行ったところ，Twitter では 5 記事，Blog では 6 記事に回答が行われた。また Twitter に関しては 5 件全て，Blog に関しては 2 件に対して感謝のコメントが返された。このとき Blog において感謝のコメントが返されなかった記事は，書かれた時期が古い記事であった。

5． おわりに

本研究に関して以下の 2 つの結果が得られた。(1) 提案手法によって疑問の書かれた記事の収集コストを大幅に軽減する事が可能である，(2) 本 Web サイトによって疑問の書かれた記事へ回答者を誘導する事が可能である。また運用を工夫する事で回答率は高まる事が予測できる。

文献

[1]永田昌明，平博順（2001）．“テキスト分類：「学習理論の見本市」．”情報処理，42(1)，pp．32-37

[2]安形輝，池内淳，石田栄美，宮田洋輔，上田修一（2009）．“学術論文に特化した検索エンジンの開発―機械学習による英語論文の自動判定―．”2009 年日本図書館情報学会研究大会発表要綱