E-044

ソーシャルブックマークにおけるコメントの分類方式

Classification of Comment in Social Bookmark

高田 彰† 山田 剛一† 絹川 博之†

Akira Takada Koichi Yamada Hiroshi Kinukawa

１．はじめに

近年, 情報発信の場にはblogやSNSなど様々な場が用意されておりユーザの情報発信が容易になっているが，そのなかで，Webページに対するクリッピングを行う場としてソーシャルブックマーク（Social Bookmark ; SBM）がある．SBMではWebページへのリンクを保存・管理する他に, タグやコメントの情報を付加することができる．これら付加情報のうちコメントにはエントリのまとめや意見など様々な種類のものが存在しており，メタデータとして活用することができる．しかし，ユーザにとって付加情報が必ずしも有益なものではなく，無益なものも含まれている．もし，コメントの分類分けを自動で行うことができれば，ユーザの望むコメントのみを表示することができるようになり，余分な情報を閲覧する必要がなくなる．

このため, 本研究でははてなブックマーク[1] 上に存在するコメントを，本文に多く出現する語や文末に現れる語などの特徴から分類することにより，ユーザが特定のコメントのみ参照することができるシステムを作成する．

２．SBMにおけるコメントの特徴と用途

はてなブックマークのコメントを分類するに当たりコメントの特徴と用途を調査した結果，10種類に分類することができた．ここでは，現状の判定基準をふまえてそれぞれのコメントの特徴と用途を述べる．

2.1 判定基準が明確なもの

2.1.1 エントリ内容を比較することで判別できる分類

引用：エントリの内容を一部転載しているもので，引用された情報は本文の重要部分, ユーザの注目箇所である．この分類の特徴として，コメントの自立語（名詞，動詞）が同じ出現順序で本文に出現することが挙げられる．

2.1.2 文末に特徴がある分類

疑問：エントリの内容とユーザの見解が異なるためユーザが疑問に感じた内容について書いてあるもので，この情報からエントリの内容に説得力があるのか，あるいはエントリ中の理解が困難な部分がどこなのかわかる．なお，この分類は「～だろうか」「～？」といった特徴的な文末表現から判定できる．

†東京電機大学大学院情報メディア学専攻
Graduate School of Science and Technology for Future Life, Tokyo Denki University

希望：エントリの内容に関するユーザのコンテンツ配信者に対する希望，ユーザが今後行いたいことが書かれており，この情報からエントリを読んだユーザの希望がわかる．そして，この分類は「～たい」「～ほしい」といった特徴的な文末表現から判定できる．

推測：エントリに関してユーザが何らかの状況を推測しているもので，この情報からエントリに対する想像について知ることができる．なお，この分類は「～そう」「～だろうな」といった特徴的な文末表現から判定できる．

2.2 判定基準が明確になっていないもの

2.2.1 特定の語がある分類

評価：エントリの内容に対してユーザ独自の基準で評価したものであり, エントリの有効性を知ることができる．なお, この分類を示す語として「わかりやすい」「ひどい」などの良し悪しを示す語と「記事」「まとめ」などのエントリを示す語が挙げられる．

予想：エントリに関してユーザが今後の展開を推測しているもので, 推測で挙げられた特徴の他に「今後」「年後」などの未来を示す語がある場合が挙げられる．

2.2.2 エントリに出現する語が少ない分類

補足：エントリに関する情報が書かれているもので，この情報を読むことでよりエントリを理解することができる．この分類を示す情報として，他のエントリに対するリンクが貼られている場合が挙げられるが，それ以外の場合，関連語の情報が必要になると考えられる．

要約：エントリの内容をユーザ独自にまとめてあり，この情報を読むことで記事の内容をおおまかに理解することができる．この分類を示す構文情報として，体言止めのコメントが多いことが挙げられる．

2.3 その他の分類

意見・感想：エントリやコメントに関してユーザが感じたことが書かれてあるもので，ユーザ間でコミュニケーションを行う場合もっとも有効な分類である．今回，他の分類に属していない場合「意見・感想」と判定する．

単語：コメントが文章の形式になっていないもの．

３．コメントの自動分類システム

3.1 システムの概要

エントリに対してユーザが付加したコメントを取得し，その特徴ごとにコメントを分類するシステムを作成する

には，文構造を把握する必要がある．本研究では，日本語係り受け解析器 Cabocha[2] で出力された文構造を利用し，各分類を判定する．なお今回，判定基準が明確な分類のみを対象としたシステムの処理手順を示す．

3.2 システムの処理

システム処理の流れ（図１参照）として，
（1）RSS フィードからエントリ，コメント情報を取得．
（2）エントリからコメント，アフェリエイトの情報を除去し，本文のみを抽出，この際，本文をCabochaで解析し，本文に出現している表現を把握する．
（3）エントリから抽出した本文とコメントをそれぞれCabochaで解析して構文情報を取得する．
（4）分析された構文情報，文末表現を用いてコメントを分類．

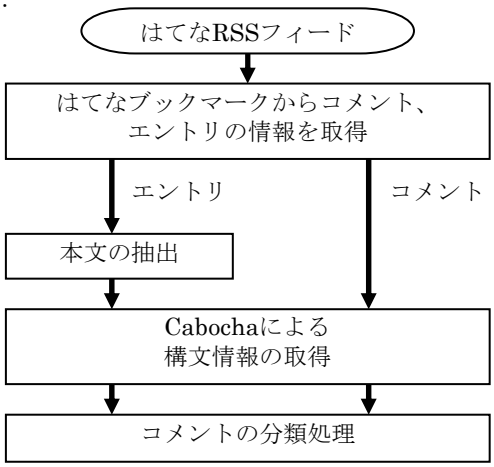

図１． システムの処理手順

3.3 コメントの分類処理

コメントとエントリの構文情報，文末表現を用いることによってコメントの分類を決定するが，この際，誤って他の分類と判断されないようにするため，判断基準が明白なものから順に分類する．このため，「引用」「疑問」「希望」「推測」「意見・感想」の順で分類分けを行う．

４． 実験

4.1 実験方法

はてなブックマークにではエントリを追加する際，自動的に「社会」「政治・経済」「生活・人生」「スポーツ・芸能・音楽」「科学・学問」「コンピュータ・IT」「ゲーム・アニメ」「おもしろ」いずれかのカテゴリに分類される．今回，それぞれのカテゴリから2つのエントリを抽出し，16 エントリに存在する 1207 コメントについて分類の内訳を調査した．なお，条件としてブックマーク数が100を超えているエントリの中から無作為に選択し，そのエントリに存在するコメントに対して，本システムを用いて解析し，その結果と調査結果を比較することで精度，再現率を求める．

$$\text{精度} = \frac{\text{システムによって検出された正解コメント数}}{\text{システムによって検出された数}}$$

$$\text{再現率} = \frac{\text{システムによって検出された正解コメント数}}{\text{そのエントリに存在する正解コメント数}}$$

4.2 実験結果

本システムを用いて解析した結果，表２の結果になった．

表２． コメント分類の精度，再現率

	正解データ	解析結果	正解数	精度	再現率
引用	171	159	136	0.8553	0.7953
疑問	63	99	61	0.6162	0.9683
希望	38	22	18	0.8182	0.4737
推測	38	48	25	0.5208	0.6579
意見・感想	721	720	607	0.8431	0.8419
単語	61	68	54	0.7941	0.8852

4.3 評価

今回，「疑問」の精度を下げた要因として「皮肉」のコメントが「疑問」に分類されたことが挙げられる．これは「皮肉」のコメントでは，否定したい物事に対し他のユーザから同調の意見を得るため，疑問形で他のユーザにコメントが記述されているためである．

また，「希望」の再現率を下げた要因として，「～してほしい」「～したい」などの文末表現に加えて「～なんだが」「～です」などの語の付属が挙げられる．

５．まとめ

SBMに存在するコメントの分類わけを行うにあたり，本論文では文末表現と名詞，動詞の出現頻度，出現順序を用いた．この結果，「引用」「疑問」「推測」「希望」の分類を判定することができた．しかし，現状では「要約」「補足」「評価」「予想」の分類は判定基準が定まっていないことから，今後「評価」「予想」「要約」で用いられることが多い表現や構文，語を把握する必要性がある．また判定基準が明確な分類に対して，現状の判定基準を改良することにより精度・再現率を上げる．

参考文献

[1]はてなブックマーク：http://b.hatena.ne.jp/
[2]Cabocha ：http://chasen.org/~taku/software/cabocha/
[3] Jaehui Park at all. : Web content summarization using social bookmarks - a new approach for social summarization, Proceeding of the 10th ACM workshop on Web information and data management, pp. 103-110, New York, NY, USA, ACM. (2008)