O-028

文書に付与された特徴語に基づく 特定の話題に関する評価文書群抽出手法に関する検討

A Study on Extraction Method of Evaluation Documents Related to Specific Topic Using Words Tagged on Document

加藤 亮 † 吉川 大弘 † 古橋 武 † 奥山 賢治 ††
Ryo Kato Tomohiro Yoshikawa Takeshi Furuhashi Kenji Okuyama

1. はじめに

近年，ブログや Web 掲示板，SNS などを通じて個人が手軽に情報の発信や収集を行うことが可能となっている．これらの情報において，例えばブログ上には，企業や消費者にとって有益な情報となるような，特定の対象製品等に関する評価文書が数多く含まれている．しかし一方で，その製品に対する広告記事や，日常の様子を書いた文書など，企業や消費者にとって特に解析の必要のない文書も多数存在している．そのため単純にキーワードに基づいて検索を行った場合，解析者は上記の有益な文書とその他の文書とが混在したものを取得することになり，その中から有益な文書のみを効率よく見つけ出すことは容易ではない．そのため，解析者の求める文書と，その他の文書とを自動で分類する手法の開発が求められている．文書分類を統計的に行う方法の一つとして，あらかじめラベル付けされたデータを用いて学習を行い，未知のデータを分類するというものがある．この分類手法では，bag-of-words を仮定し，単語の頻度情報に基づいて分類を行うのが一般的である．しかしこの方法では，文書の主題と関係なく，高頻度の単語に影響されやすいという問題がある．

本稿では，特定の話題に関する評価文書群の抽出を目的として，文書の主題を捉えた文書分類を行うシステムを提案する．評価文書の分類方法として，知識整理を行うために文書に付与されている，整理対象を代表するような短い単語（タグ）に着目する．提案システムでは，ユーザが文書に付与したタグ（ユーザタグ），あるいは Tag-LDA により自動で付与したタグ（モデルタグ）を素性とし，ナイーブベイズ分類器を用いて文書分類を行う．本稿では，提案システムを実際のブログにおける分類問題に適用し，その有用性を検証する．

2. 提案システム

文書分類の素性選択において，文書内の単語ではなく，タグに着目した分類システムはこれまであまり報告されていない．本稿では，分類のための素性となるタグの付与器として Tag-LDA[1]，分類器としてナイーブベイズ分類器 [2] を用いた文書の分類システムを提案する．付与されたタグにより，文書の主題を捉えた分類を行うことで，分類精度の向上を目指す．

2.1 Tag-LDA

知識整理を行う方法として，整理対象を代表するようなタグを付与する方法が用いられる．近年では，このタグを文書に自動で付与し，知識整理を行う手法が報告されている [3][4]．それらの中でも，トピックモデルを用いた研究が近年注目され，また成果を挙げている [1]．トピックモデルとは，単語とタグの背景にトピックを仮定し，トピックの持つ単語とタグに対する確率分布に基づいて単語とタグが生成されるとした確率モデルである．文献 [1] では，ブログデータの一部を学習に用いたタグ付与実験を行い，従来のタグ付与手法よりも高い適合率と再現率でタグ付与が行えることを示している．タグ付与を行う過程を以下に示す．

タグの付与された学習データをモデルが学習し，トピックにおける単語とタグの確率分布を求める．これに基づいて，タグの付与されていない未学習の文書データに対するトピック分布を推定し，文書 d が与えられたときのタグ t の出現確率 $p(t|d)$ を得る．$p(t|d)$ の値の高いものから上位 n 個のタグを選択し，タグとして文書 d に付与する．本稿では，$n = 5$ とした．

2.2 ナイーブベイズ分類器

ナイーブベイズ分類器 [2] は，文書の分類方法として最も一般的な手法の一つである．分類問題においては，他にも SVM(Support Vector Machine) を用いた手法なども提案されている [5]．SVM は，分類精度が高いことが知られているが，本稿では，文書分類において，選択する素性の比較を第一の目的としているため，解析の行いやすい単純なモデルであるナイーブベイズ分類器を用いる．

ナイーブベイズ分類器では，文書 D が与えられたとき，カテゴリ C が得られる事後確率を算出する．この事後確率を最大化するようなカテゴリを選択することで文書分類を行う．以下の式 (1) 中では，カテゴリ C での語彙 v の出現回数を n_v^C，カテゴリ C での単語出現回数を n^C，カテゴリ C である文書数を n_C^D，全文書数を n^D，単語の種類数を V，スムージング項を α と表記している．スムージング項は，$n_v^C = 0$ であった場合に，適用した文書の事後確率が 0 となるのを防ぐために導入している．本稿では，$\alpha = 0.1$ とした．

$$\begin{aligned} p(C|D) &= \frac{p(D|C)\cdot p(C)}{p(D)} \\ &\propto p(D|C)\cdot p(C) \\ &\propto \Pi_i^N p(w_i = v|C)\cdot p(C) \\ &\propto \Pi_i^N \frac{n_v^C + \alpha}{n^C + V\cdot\alpha}\cdot\frac{n_C^D}{n^D} \end{aligned} \quad (1)$$

† 名古屋大学工学研究科
†† 東邦ガス株式会社

3. 実験

3.1 使用データ

ガスコンロに対するユーザの評価文書の抽出実験を行った．実験には So-net ブログ * を用いた．「ガスコンロ」,「ガス」,「ビルトインコンロ」などの単語を検索キーワードとして，タグや文書内単語を含む文書情報を 548 件抽出した．ガスコンロの評価に関する文書データを正例，その他を負例として人手で分類を行い，24 件を正例，524 件を負例とした．抽出データ全体に対して正例が少ない不均衡データであるため，本稿では負例の数を減らして正例の数と同じとすることでアンダーサンプリングを行い実験を行った．

3.2 方法

本実験では，以下に示す 4 種類の素性を用いて比較実験を行った．ただし，素性のデータ内での最小出現回数を 2 回とし，出現回数が 1 回の素性は除外した．

素性 1　文書内単語における名詞による素性

素性 2　ユーザタグによる素性

素性 3　モデルタグ ($n = 5$) による素性

素性 4　ユーザタグとモデルタグの組合せによる素性

タグ数に制限がないため，付与されたユーザタグが 1 つしかない文書データが存在する．また，最小出現回数を満たすタグが 0 個となることで，素性が 0 個となってしまう文書データも存在する．そのため素性 4 では，ユーザタグの数が 5 に満たないものに対してはモデルタグを付与し，タグの総数が 5 となるようにした．それぞれの統計情報を表 1 に示す．ただし表 1 は，最小出現回数を満たしたタグの数である．また実験は，leave-one-out cross-validation により，上記の 4 種類の素性の精度を比較した．

表 1: それぞれの素性の統計情報

素性	総出現回数	種類数
素性 1	2827	716
素性 2	126	33
素性 3	216	15
素性 4	253	53

3.3 結果

それぞれの素性に対する分類正解数を表 2 に示す．表の（）内は正解率である．従来手法である素性 1 は，すべての素性の中で正例正解数が最も高く，負例正解数が最も低い値を示した．そのため，文書内の単語（名詞）は，正例を分類するのに有効な素性であるといえる．一方で，素性 2〜4 は素性 1 と比べて，正例正解数は低いものの，負例正解数は高い．そのため，タグは，負例を除外するのに有効な素性であると考えられる．また素性 4 では，素性 1 に対して合計の正解数が 2 つ勝っていることがわかる．

*https://blog.so-net.ne.jp/_tag/

表 2: それぞれの素性に対する正解数

素性	正解数	正例正解数	負例正解数
素性 1	33 (0.69)	23 (0.95)	10 (0.42)
素性 2	29 (0.60)	13 (0.54)	16 (0.67)
素性 3	32 (0.67)	14 (0.58)	18 (0.75)
素性 4	35 (0.73)	18 (0.75)	17 (0.71)

ナイーブベイズ分類器では式 (1) に基づいて分類が行われるが，式 (1) の値 $p(C|D)$ は，分類における確信度と捉えることができる．そこでこの確信度を基に，素性 1〜4 において確信度の高い分類結果を採用するハイブリッド分類方法について検討する．このハイブリッド分類の仕方は $(2^4 - 5)$ 通りの組合せが考えられる．表 3 に，その中で最も正解数が高かった組み合わせである，{素性 1，素性 4} の精度を示す．

表 3: 従来素性とハイブリッドの正解数における比較

素性	正解数	正例正解数	負例正解数
素性 1	33 (0.69)	23 (0.95)	10 (0.42)
ハイブリッド	39 (0.81)	22 (0.92)	17 (0.71)

ハイブリッド分類では，従来手法である素性 1 の正例正解数を上回ることはできなかったものの，素性 1 や素性 4 を上回る合計正解数を得ることができた．

4. おわりに

本稿では，ナイーブベイズ分類器を用いた文書分類において，素性にタグを利用する手法について検討した．ユーザタグとモデルタグを組合せることで，従来の文書中の単語を素性とするものと比較して正解数を上昇させることができることを示した．また，分類時の確信度を用いたハイブリッド分類方法により，さらに精度を向上させることができることを示した．本稿での実験ではデータ数が少なかったため，今後はより大規模なデータにおいて検討を行っていく予定である．

参考文献

[1] X. Si, et al:Tag-LDA for scalable real-time tag recommendation, Joural of Computational Information Systems 6, pp. 23-31, 2009

[2] P. Domingos, et al:On the optimality of the simple Bayesian classifier under zero-one loss, Machine Learning 29, pp.103-130, 1997.

[3] C. H. Brooks, et al:Improved annotation of blogosphere via autotagging and hierarchical clustering, WWW, pp. 625-631, 2006.

[4] T. Ohkura, et al:Browsing System for Weblog Articles based on Automated Folksonomy, WWW, 2006

[5] T. Joachims:Text categorization with support vector machines, ECML, 1998.