4P-2

動画共有サイトにおけるコメントを利用した動画自動タグ付け手法

山下 智紀† 矢吹 太朗 佐久田 博司
青山学院大学 理工学部 情報テクノロジー学科‡

1 序論

1.1 研究背景

動画共有サイトの普及に伴い，膨大な動画が Web 上に存在するようになった．そのため，大量の動画の中から，利用者が目的の動画を探すことが困難になっている．この問題の解決法として，動画共有サイトでは動画に付けられたタグ情報からの検索機能が設けられている．動画へのタグ付けは，利用者が手動で行っている場合が多い．そのため，動画に不適切なタグが付けられることもあり，検索を失敗させる原因となっている．また，増加し続ける動画に対して，継続的に手動でタグを付ける作業が煩雑だという問題もある．そのため，動画に適したタグを効率よく付ける方法が求められる．

1.2 研究目的

本研究では，動画に付けられたコメントを用いて，自動的にタグを推定する手法を提案する．コメントの活用方法として，コメント中の特徴語をタグとする方法が挙げられる．しかし，この方法にはコメント中に出現した単語しかタグの候補とされず，単調なタグが大量に生成されてしまう欠点がある．そこで本研究では，すでに他の動画に付けられているタグの中から候補となるタグを推定する．これは，既存の動画に付けられたコメントとタグの関連性を学習し，その結果をタグが未定の動画のコメントに適用することで実現される．これにより，人間によるタグ付け結果と似た，柔軟なタグ付けが行える．

1.3 関連分野

動画への自動タグ付け手法は機械学習に基づくテキスト分類とみなせる．機械学習には様々な手法があり，中でも実装の容易さと精度の高さから，ナイーブベイズ分類器（Naive Bayes classifier）がメールのスパム判定 [1] やブログのカテゴリ分類 [2] など様々な分野に応用されている．そのため，本研究でも学習手法としてナイーブベイズ分類器を用いる．

Automated Tagging System for Video Sharing Services.

† Tomonori YAMASHITA (t.yamashita2@gmail.com)
‡ Department of Integrated Information and Technology, College of Science and Engineering, Aoyama Gakuin University.

2 対象とするコンテンツ

本研究では，視聴者のコメントを利用したタグ推定を考えているため，動画に大量のコメントが付けられているニコニコ動画（http://www.nicovideo.jp/）を対象にする．ニコニコ動画は動画の投稿や共有，視聴が可能な動画共有サービスである．2012 年 1 月現在で，約 700 万の動画と約 30 億のコメントが投稿されている．

ニコニコ動画では，1 つの動画に対して，タグを最大 10 個まで登録することができる．タグの一部にカテゴリタグがあり，これによって動画は 30 個のカテゴリに分類されている．タグはニコニコ大百科と呼ばれるオンライン百科事典と連動しており，そこでそのタグに関する詳細を知ることができる．

3 提案手法

コメントから抽出した単語とタグを組とした訓練データをもとに，ナイーブベイズ分類器を用いて，訓練データからコメントとタグの関連性を学習する．学習がおわった分類器を使って，タグが未定の動画のコメントからタグを推定する．この手法により動画に適したタグを効率よく推定することができ，タグ検索における検索能力の向上が期待できる．本研究では，以下の手順でタグ推定を行う（図 1）．

1. 動画情報の取得
2. 訓練データの作成
3. ナイーブベイズ分類器によるタグ推定

各手順について以下で詳述する．

3.1 動画情報の取得

ニコニコ動画から動画に付けられたコメントとタグを取得する．コメントは動画 1 つにあたり最大 1000 件取得することができる．またタグを取得する際，タグ検索のノイズとなりうるものを除去するために，ニコニコ大百科に掲載されているタグのみを抽出する．これはニコニコ大百科に掲載されているタグが，タグ検索のノイズにならないものだと考えられるためである．

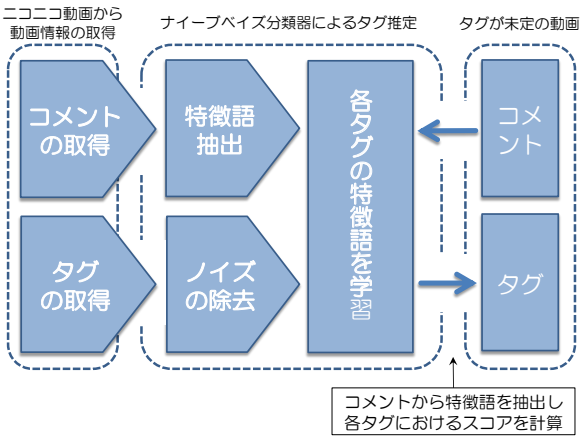

図 1　ナイーブベイズ分類器を用いたタグ推定の概要

3.2　訓練データの作成

コメントは Bug-of-words（単語の集合）として扱い，単語の順番を考慮せずに出現回数だけを学習に用いる．形態素解析を用いて，コメントから名詞，形容詞，副詞を抽出し，単語とタグを組とした訓練データを作成する．形態素解析には MeCab 0.98[3] を用いる．MeCab で使われている辞書は，一般的な単語のみからなっているため，崩した日本語やネット上で使われる特殊な言葉を期待した通りに分割できない．そのため本研究では，はてなキーワードと日本語版 Wikipedia の項目名を辞書に追加する．

3.3　ナイーブベイズ分類器によるタグ推定

ナイーブベイズ分類器には多項モデルと多変数ベルヌーイモデルがある．多項モデルはテキスト分類において，多変数ベルヌーイモデルに比べ性能が高いと報告されている [4]．そのため本研究では，多項モデルを用いる．多項モデルでは，各単語の出現回数がモデルに組み込まれる．単語とタグを組合せとした訓練データを学習し，各単語の出現回数を数えて，各タグが生成される確率を計算する．

3.3.1　TF-IDF

どのタグにも出現する単語は，各タグを特徴付けるという意味ではあまり役にたたない．TF-IDF は，各単語に対する重み付け手法であり，特定のタグにのみ出現する単語の重要度を上げる役割を果たす．そのため，TF-IDF 値を単語の重みとして，単語の出現回数に掛け合わせる．

3.3.2　加算スムージング

タグを推定する際，訓練データのボキャブラリに含まれない単語を 1 つでも含んでいると，そのタグが生成される確率が 0 になってしまう．そのため，出現回数に一律の値 δ を加える加算スムージングを用いる．

4　評価実験

4.1　実験方法

700 万件の全動画を取得するのは困難であるため，本研究では，動物カテゴリタグが付けられた動画に絞り，10,000 件分の情報を取得した．取得できたコメントの総数は 400 万，タグの総数は 5,090 であった．コメントから抽出した単語とタグを組とした訓練データをもとに，コメントとタグの関連性を学習した．訓練データの総単語数は 22,360 となった．

4.2　評価方法

推定結果の評価には，F 値を用いた．F 値は，推定の正確さ（精度）と推定の網羅性（再現率）を組み合わせた尺度である．評価実験では，訓練データとは別に用意した，動物カテゴリタグの付けられた動画 100 件を対象とし，テストデータとして与えた．正解データは元々動画に付けられているタグから検索ノイズとなりうるタグを除去したものとした．

4.3　実験結果

加算スムージングに用いた値 δ，TF-IDF を用いた場合においての，推定結果の F 値を比較する．評価実験を行った結果を表 1 に示す．

表 1　タグ推定における F 値

δ	多項モデル	多項モデル (TF-IDF)
1	0.332	0.287
0.1	0.334	0.476
0.01	0.330	0.481

5　考察

表 1 より，TF-IDF を単語の重みとして用いることで，推定精度の向上を得られることが分かった．加算スムージングの値によっても，推定精度が変化しており，今後は最適な値を検討していく必要がある．誤って推定されたタグを調べると，多くが正解のタグに類似しているタグであった．本実験では，動画に付けられているタグを正解データとして扱ったが，付けられているタグ以外にも，その動画に適したタグは多く存在するため，評価方法についても検討する必要がある．

6　結論

本研究では，ナイーブベイズ分類器を用いた，コメントによるタグ推定の手法を提案した．推定精度には多少の不満があるものの，本手法により，適切なタグを効率よく付けることが可能であり，従来の手動によるタグ付けの問題点を解決することができた．

参考文献

[1] 佐々木稔, 新納浩幸. 文書分類を用いたスパムメール判定手法. 情報処理学会研究報告, Vol. 2004, No. 93, pp. 75–82, 2004.
[2] 平野耕一, 古林紀哉, 高橋淳一. 日本語圏ブログの自動分類. 情報処理学会研究報告, Vol. 2005, No. 117, pp. 21–26, 2005.
[3] 工藤拓. Mecab （和布蕪）. http://mecab.sourceforge.net/.
[4] Kamal Nigam Andrew McCallum. A comparison of event models for naive bayes text classification. *AAAI/ICML-98 Workshop on Learning for Text Categorization*, pp. 41–48, 1998.