2C-1

フォークソノミーを利用した 自動カテゴリ作成及び分類システムの提案

備瀬竜馬† 笠原博和† 二本木智洋† 森本光昭† 高田政樹† 中川修†

大日本印刷株式会社 情報コミュニケーション研究開発センター†

1．はじめに

Blog が広まるにつれて，指定キーワードに関連する Blog の検索サービスや Blog で流行のキーワードを提示するサービスが増えつつある．このようなサービスにおいて，Blog がカテゴリごとに分類されているとユーザが興味ある情報を閲覧する際の利便性が高まると考えられる．

そこで，様々な Web ページのカテゴリ分類手法が提案されている．例えば，人がカテゴリごとに設定したルールによって分類する手法や確率モデルや SVM 等の機械学習による手法等[1]が提案されている．しかし，これらの手法は，人がルールを定めたり，分類器作成のための学習データを用意する必要があるという問題がある．

一方，近年，個々のユーザが各自の視点で記事にタグを付与し，そのタグを利用して分類するフォークソノミーと呼ばれる方法が注目を集めている．しかし，インターネット上の一部の文書にしか付与されていないという問題点がある．また，タグをそのままカテゴリとして採用した場合，「日記」や「雑記」等の曖昧なタグや同様の概念のタグがあるという問題がある．

そこで，本稿では，フォークソノミーによってタグが付与された記事を利用し，曖昧なタグや同様の概念のタグを除外したカテゴリを自動的に生成するシステムを提案する．また，識別が難しいカテゴリ間に関しての分類精度向上を図ったシステムを提案し，その評価を行う．

2. 提案システム

2.1. カテゴリ選択フロー

図１にカテゴリ選択フローを示す．まず，インターネット上のタグ付けされた文書集合(分類器作成学習データ)を取得し，その中で一定数以上出現したタグをカテゴリ候補とする．そして，各カテゴリ候補についてそのカテゴリに属するか否かを判別するカテゴリ分類器を作成する．今回は，カテゴリ分類器として，カテゴリの文

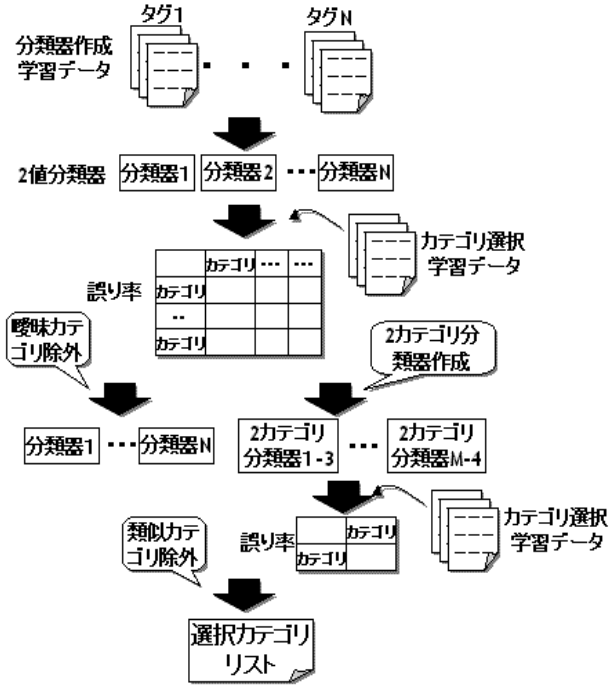

図 1．カテゴリ選択フロー

書に出現する名詞の偏りを利用する確率モデルを採用した．適合度 sim は以下のように定義され，sim が閾値を超えた場合，対象文書はそのカテゴリに属すると判断される．

$$sim = \sum_{gi(dj)=1} \log \frac{P(k_i \mid R) + \alpha}{P(k_i \mid \neg R) + \alpha} + \sum_{gi(dj)=1} \log \frac{P(\neg k_i \mid R) + \alpha}{P(\neg k_i \mid \neg R) + \alpha}$$

ここで，P(ki|R)は特定のカテゴリに属する文書集合 R 中の文書が単語 ki を含む確率，P(¬ki|R)は R 中の文書が単語 ki を含まない確率であり，¬R は R の補集合である．gi(dj)は，分類対象文書 dj が単語 ki を含む際には 1 を返し，含まない場合には 0 を返す関数，αは分母が 0 にならないように付加する微小値とする．

次に，インターネット上から各カテゴリのタグが付与されたカテゴリ選択学習データを一定数取得し，作成した各カテゴリ分類器がそのデータのカテゴリを誤って分類した文書数の割合(誤り率)を各タグに対して求める．そして，その誤り率が一定値(25%)を超えるタグ数が全タグ数の 30%を超えた場合，そのカテゴリは曖昧なカテゴリであるとみなし，カテゴリ候補から除外する．

Proposal of automatic category generation and classification system using Folksonomy.

†Media Technology Research Center. Dainippon Printing Co.,Ltd.

さらに，誤り率が高いカテゴリの組み合わせに関しては，新たに 2 カテゴリ分類器を作成する．例えば，「野球」分類器によるサッカータグ文書の誤り率が高い場合，不適合文書としてサッカータグ文書を与えて学習し，2 カテゴリ分類器を作成する．不適合文書としてサッカーの文書を与えるので野球かサッカーかの分類精度は高くなると考えられる．

また，類似のタグの組み合わせに関しては，上記のような 2 カテゴリ分類器でも誤り率は高いと考えられる．そこで，さらに評価を行い，誤り率が改善されてない場合(30%以上の場合)は，類似カテゴリであると判断して，タグの出現頻度が少ない方を除外する．

2.2. カテゴリ分類フロー

図 2 にカテゴリ分類フローを示す．分類対象文書に関してカテゴリ分類器によって属するのに適していると判断されたカテゴリを対象文書のカテゴリ候補とする．そして，候補のうちの組み合わせで 2 カテゴリ分類器が存在する場合，その 2 カテゴリ分類器で分類し，不適合と判断された方のカテゴリをカテゴリ候補から除外する．そして，残りのうち分類器のスコアが高い方をカテゴリとして登録する．また，残りの候補がない場合は未分類とする．このように，2 カテゴリ分類器を利用することによって分類精度が向上すると考えられる．

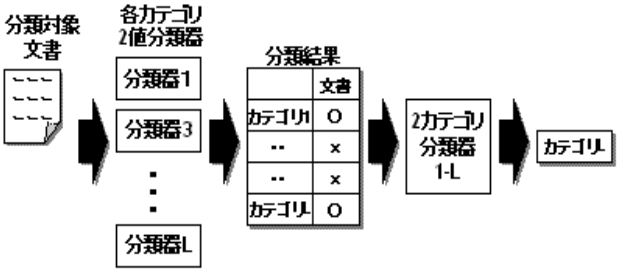

図 2. カテゴリ分類フロー

3. 実験

データとして，2006 年の 9/1～11/30 までの間で BLOG360[2]で収集した約 180 万記事を対象とした．また，学習データ及び評価データとするタグ付きの記事には，Blog の著者が記したタグ付きの記事とした．400 以上出現しているタグをカテゴリ候補として，実験を行った．以下が本システムによるカテゴリ選択結果である．

・**システムが選択したタグ**

音楽，ゲーム，本，アニメ，映画，仕事，
テレビ，漫画，サッカー，ニュース，
ハロプロ，食，学校，野球

・**システムが除外したタグ**

日記，雑記，日常，その他，生活，読書，
メモ，雑談，戯言，徒然，TV，雑感，雑，
ひとりごと，独り言，自転車，つれづれ，
game，ネタ，つぶやき

この結果を見ると，人がカテゴリとして判断しないであろうタグはシステムでも除外されていることがわかる．また，TV，game というタグも除外されており，（TV，テレビ）（game，ゲーム）という同様の概念であるタグの判別も適切であることがわかる．また，候補となるタグ数を増やした場合も，9 割程度は人間が判断するカテゴリと合致した．

本システムのカテゴリ分類精度を評価するため，人手で文書を分類し，システム評価用データとした．用意した文書数は 2,000 文書である．また，2 カテゴリ分類器を利用しない場合とした場合の比較を行い，2 カテゴリ分類器による分類精度向上の効果を検証した．比較は，再現率の平均が同程度の値となるようパラメータを調整し，適合率を比べた．

再現率，適合率の平均値を表 1 に示す．結果を見ると，再現率が 6 割程度のとき，2 カテゴリ分類器を利用したシステムの方が利用していない場合より適合率が向上しているのがわかる．

2 カテゴリ分類	なし	あり
再現率	0.59507	0.593171
適合率	0.771091	0.842194

表 1. 再現率及び適合率

4. まとめ

フォークソノミーによって付与されたタグ付き記事を学習データとして利用し，カテゴリとして不適切なタグ及び(TV，テレビ)等の同様の概念であるタグを除外するシステムを提案し，実験した結果，カテゴリとして不適切なタグを除外できることを検証した．また，分類精度が高くないカテゴリの組み合わせに関しては，その組み合わせを判別するための 2 カテゴリ分類器を作成することで，適合率が改善することを実験的に検証した．

参考文献

[1]高村，松本.SVM を用いた文書分類と構成的機能学習法.情報処理学会論文誌 データベース，Vol.44，SIG3(TOD 17)pp.1-9,2003
[2]ブログ解析によるクチコミ追跡サイト
BLOG360(http://blog360.jp/)