E-036

Blog サイトのカテゴリ分類システム

Category classification system of blog site

備瀬竜馬† 笠原博和† 二本木智洋† 森本光昭† 高田政樹† 中川修†
Ryoma Bise Hirokazu Kasahara Tomohiro Nihongi Mitsuaki Morimoto Masaki Takada Osamu Nakagawa

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，Blog で流行のキーワードや Hub となる Blog サイト(様々なユーザが参照しており，影響力が高い Blog サイト)を提示するサービスが増えつつある．また，ある商品について，Blog で語られている情報を分析してマーケットリサーチに活用するサービスが増えつつある．

Blog サイトに着目した際に，特に記事に共通の主題はなく日々のできごとを書いてあるサイトもあれば，ある同じ主題に着目した記事ばかり書かれている Blog サイトがある．閲覧ユーザからすると自身と同じ興味を持ったブロガーのサイトを探したいというニーズがあるため，特定の主題に着目した記事が書かれている Blog サイトを提示することができれば価値があると考える．また，Blog 検索・分析サービスにおいて，Blog サイトのカテゴリ分類を行うことができれば，分野別の Hub となる Blog サイトの提示や，どの分野に興味を持っているブロガーがどんな意見を持っているか等のマーケットリサーチにも活用することができると考える．

そこで，本稿では，一定期間中の Blog サイト上の記事記述内容によって Blog サイトのカテゴリ分類を行うシステムを提案し，実際のデータに適用し，具体的な結果を示す．

2．提案システム

本システムは，まず，Blog 記事のカテゴリ分類を行い，特定の Blog サイトにおいて一定期間中に書かれた記事の特定カテゴリの割合が一定値を越えれば，そのサイトにそのカテゴリを紐付ける方法で実現した．方法の詳細を以下に示す．

2.1 Blog 記事の自動カテゴリ分類

文書自動分類の方法としては，ベイズ法，SVM 法，決定木による方法など多くのものが提案されている[1]．本システムでは，各カテゴリにベイズ法による 2 クラス自動分類器を用意し，対象文書に対して各カテゴリに属するかを各分類器で判定し，属すると判定された場合は，そのカテゴリのタグを対象記事に付与するという方法で実現した．

学習データとしては，Blog を書いたユーザが付与したタグ付き記事(フォークソノミー)を学習データとして用いた．フォークソノミーによるタグ付き記事はシステムで自動的に取得できるため，学習フェーズも自動化可能となる[2]．また，データとしては，全文ではなく RSS の情報(Blog 記事の一部で通常 200 文字程度)を用いた．

しかし，学習データとして，フォークソノミーによるタグ付き Blog 記事の RSS データを用いる場合，通常の文書分類と比べて以下のような問題がある．

・1 つの文書の長さが短い．
・一般的な文書と比べ，様々な単語が用いられる．
・カテゴリによっては，学習データが十分でない．

そのため，出現頻度が小さい単語を多数含んでしまう．出現頻度が小さい場合，特定のカテゴリに含まれる確率と含まれない確率の比の差が誤差である可能性が高くなり，適合度に影響を与えてしまうという問題がある．そこで，本稿では，一定の閾値以上となる出現頻度の単語のみを適合度の計算の対象とし，出現頻度が小さい単語は無視することで，少数しか出現しない単語による誤差の影響を軽減した．各カテゴリ分類の適合度を下記の式で定義し，適合度が閾値を超えるカテゴリのタグを対象文章に付与した．

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

ここで，P(ki|R)は特定のカテゴリに属する文書集合 R 中の文章が単語 ki を含む確率，¬R は R の補集合である．gi(dj)は，分類対象文章 dj が単語 ki を含みかつ P(ki|R)が一定値以上となる際には 1 を返し，それ以外には 0 を返す関数である[3]．

各カテゴリに分類された記事は，DB 上に登録することで，タグを付与する．どの分類器にも分類されなかった場合，その記事にはタグは付与されない．

2.2 Blog サイトのカテゴリ分類

図 1 に Blog サイトのカテゴリ分類システムの概要図を示す．まず，カテゴリ及び期間を指定し，その期間の記事を取得し，サイトごとに指定カテゴリに分類された記

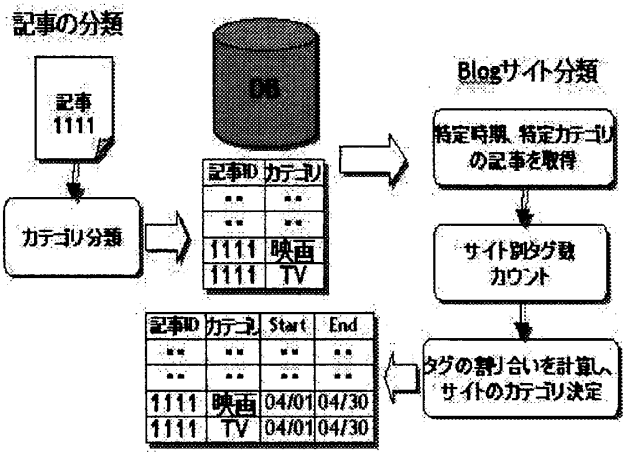

図 1．Blog サイト分類システム概要図

† 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

事数をカウントする．そして，一定タグ数以上のサイトのみを対象として，各サイトの指定期間内の記事数における指定カテゴリ記事の割合(主題度)を計算する．主題度が一定値以上となるサイトを指定カテゴリに分類する．

ここで，期間を指定して分類を行ったのは，同じ主題に着目して Blog 記事を書き続けているブロガーも多いと考えられるが，長い期間やより細かいカテゴリでみると興味が移り変わっていくサイトもあると考えるためである．このように，一定期間の Blog サイトを分類することで，過去の特定の時点でのカテゴリ別 Blog ランキングの作成やその期間におけるブロガーの興味の推移を可視化することに応用できると考える．また，Blog サイトのランキングへの応用を考えた場合，通常，全サイトに対して分類を行うことが考えられるが，Blog サイトは膨大な数あり，一定期間ごとに全ての Blog サイトに対してカテゴリ付けを行うのは負荷が高いため，対象カテゴリに属する記事が一定数以上含む Blog サイトのみを分類対象としている．

3. 実験

本実験では，2007 年 4 月 1 日～31 日までの 1 ヶ月間の間で，BLOG360[4]で収集した約 260 万件の記事，約 30 万サイトを実験対象とした．学習データとしては，2007 年 3 月の記事に人手で付与されているタグの中から日記，雑記等の曖昧なタグを除いた上位 10 個をカテゴリとして選定した．上記データに対して，提案システムを用いて，記事分類及び Blog サイト分類を行った．自動分類結果のサンプルについてカテゴリの妥当性を判断した結果としては，8 割程度の記事が妥当なカテゴリに分類されていた．

表 1 に記事及び Blog サイトのカテゴリ分類結果を示す．結果を見ると，選定したカテゴリのどれにも分類されなかった記事数，Blog サイト数が圧倒的に多いのがわかる．記事数に比べ Blog サイト数はさらに少なくなることがわかる．食とサッカーに関する記事は同程度の数書かれているが，Blog サイトではサッカーは食に比べて倍近く多い．このことから Blog サイトに一貫して書かれやすいカテゴリと様々な話題を書く人に書かれる方が多いカテゴリがあることが伺える．また，複数タグ付与されているサイトをカウントした結果，(音楽，映画)，(本，映画)，(アニメ，TV)等のカテゴリの組み合わせが多かった．

	記事数	サイト数
音楽	17689	188
アニメ	15678	199
ゲーム	3353	26
本	17040	333
映画	17716	258
TV	9776	36
食	24950	273
ニュース	9266	48
野球	14732	170
サッカー	26219	509
未分類	2415950	297428

表 1．記事及び Blog サイトのカテゴリ分類結果

	平均被リンク数	被リンク割合
音楽	1.527	0.282
アニメ	7.166	0.648
ゲーム	2.500	0.577
本	2.015	0.364
映画	1.938	0.357
TV	4.056	0.472
食	2.168	0.366
ニュース	11.646	0.521
野球	4.560	0.533
サッカー	1.576	0.352
ランダム	0.957	0.250

表 2．被リンク数の比較

このような結果から，ブロガーが興味あるカテゴリの組み合わせの分析等にも応用できるかと考える．さらに，これらの結果を評判分析等に利用することで，どんなことに興味を持っているユーザがどんな意見を言っているか等の分析に応用できると考える．

また，閲覧ユーザは日常の出来事を日記のように綴っている Blog サイトより，特定分野に詳しい人の Blog サイトを探したい場合があると考える．例えば，映画に関するクチコミを探しているユーザは，映画に詳しい人の記事を見たいというニーズがあると考えられる．そのため，特定の主題に関して多く記事を書いている Blog サイトは，一般な日常の出来事を書いている Blog サイトより参照されているのではないかと考えられる．そこで，ランダムに取得した Blog サイトと各カテゴリに分類された Blog サイトの 1 サイトあたりの平均被リンク数と 1 つでも他サイトからの被リンクがあるサイトの割合を求め，比較した．結果を表 2 に示す．この結果，全てのカテゴリがランダムに取得した Blog サイトよりも被リンクが多いことがわかる．また，ニュース，アニメ，ゲームのカテゴリは比較的多く参照されていることがわかる．

4. まとめ

本稿では，Blog サイトをカテゴリに自動分類するシステムを提案した．そして，実際の Blog サイトを分類し，カテゴリごとのサイト数の傾向や被リンク数の傾向を調査した．今後は，分類した Blog サイトの情報を活用してカテゴリ別の Blog ランキングやブロガーの属性付き評判分析等に応用していきたいと考える．

参考文献

[1]加沢 他．テキスト分類 - 学習理論の見本市，情報処理学会誌 Vol.42 P.32～37 2001 年 1 月

[2]備瀬 他．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，情報処理学会第 69 回全国大会

[3]Ricard Baeza-Yates 他．Modern Information Retrieval.

[4]ブログ解析によるクチコミ追跡サイト BLOG360(http://blog360.jp/)