E-007

Wikipedia のカテゴリを利用した Web 検索結果のフィルタリングの検討
A study on web search filtering using Wikipedia category information

榊原 敬吾† 山本 けい子‡ 田村 哲嗣‡ 速水 悟‡
Keigo Sakakibara Keiko Yamamoto Satoshi Tamura Satoru Hayamizu

1. はじめに

近年，インターネットの普及に伴い，Web 上に存在する情報は急激に増加している．この膨大な情報の中から必要な情報を見つけ出すために，ユーザは Google や Yahoo!等の検索サイトを利用する．このようなキーワードを利用した検索では，同字異義語を含むページも検索結果に含まれていたり，適切な検索キーワードを考えるためには，ユーザの技術や知識がある程度必要といった問題がある．そのため，Web 検索の結果には，ユーザにとって不要な情報が含まれる可能性がある．不要な情報が検索結果に含まれている場合，必要な情報が埋もれたり，見落とされたりすることがある．この問題に対して，検索結果をクラスタリングして検索の支援をする研究[1][2]が行われている．また，Web ページのジャンルを推定する研究[3]や，Wikipedia を利用した研究[4]も行われている．

そこで，本研究では Wikipedia[5]のカテゴリを利用して Web 検索結果からユーザの指定したカテゴリに属するページのみを抽出する Web 検索支援システムを構築した．本論文では，Wikipedia を学習データに用いて，Web 検索結果に対して SVM(Support Vector Machine)により指定したカテゴリの文書とそれ以外の文書に分類する実験と評価を行い，Wikipedia を学習データとして用いることが有効であることを示す．

2. システム・提案手法

2.1 システム概要

作成したシステムの概要を図 1 に示す．

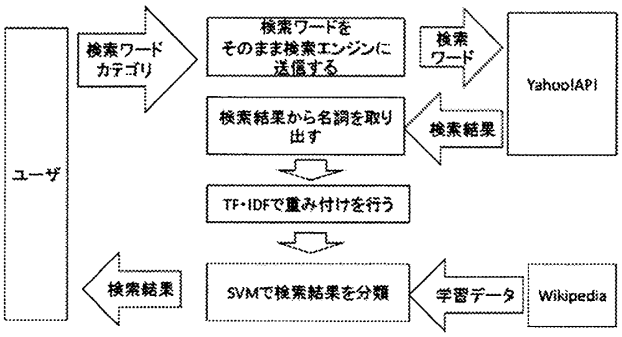

図 1 システム概要

Wikipedia 日本語版には，2008 年 6 月の段階で 50 万本を超える記事が存在している．これらの記事は，常に人手で更新されているため，最新の話題にも対応可能であり，さらに，一定の規則に基づいて記述されているためカテゴリ等の情報が比較的容易に取得できるといった特徴がある．

Wikipedia のカテゴリは，ある一つのカテゴリに注目すると，図 2 のような木構造になる．そのため，あるカテゴリに関連する文書を取得するためには，この木構造を何度か辿る必要がある．しかし，深くまで辿り過ぎると最初のカテゴリと全く関係のないカテゴリの文書まで取得してしまう．今回の実験では，指定したカテゴリに関する文書を集める際には，3 階層までの文書を取得した．

実験に用いた Wikipedia の文書は 2007 年 9 月のものを用いた．

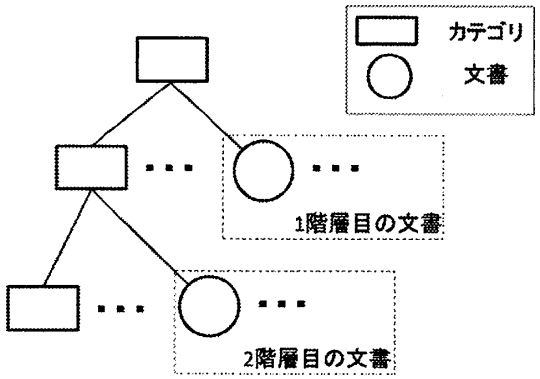

図 2 Wikipedia のカテゴリの階層構造の例

2.2 文書ベクトル

文書ベクトル $\mathbf{a}_j$ には，(1),(2)式で求められる TF*IDF 値を用いた．要素 a_{ij} は文書 j における単語 i の TF*IDF 値を正規化した値で，n は総単語数である．

文書ベクトルに用いた単語は，MeCab[*1] によって形態素解析を行って得られた名詞すべてである．ただし，この中から記号や数字だけの単語は除いた．

$$\mathbf{a}_j = \begin{pmatrix} a_{1j} & a_{2j} & \cdots & a_{nj} \end{pmatrix} \quad (1)$$

$$a_{ij} = \frac{TF_{ij} * IDF_i}{\left|\mathbf{a}_j\right|} \quad (2)$$

3. 実験

3.1 学習データ

学習データに用いる文書は 2.1 で述べた方法で取得した．指定したカテゴリに属さない文書に関しては，残った Wikipedia のすべての文書から同数の文書をランダムで選

† 岐阜大学大学院工学研究科
Graduate School of Engineering, Gifu University
‡ 岐阜大学工学部
Faculty of Engineering, Gifu University

*1 http://mecab.sourceforge.net/

んで取り出した．本実験ではカテゴリとして「医療・映画・スポーツ・政治・生物」の 5 つを用いた．

学習データの文書数を表 1 に示す．

表 1　学習データの文書数

カテゴリ	カテゴリに属する文書	属さない文書	合計
医療	5,634	5,634	11,268
映画	7,349	7,349	14,698
スポーツ	14,750	14,750	29,500
政治	12,333	12,333	24,666
生物	8,898	8,898	17,796

文書ベクトルの次元数は約 12 万～23 万次元であり，これらを LIBSVM[*2] により学習を行った．SVM には線形カーネルを用いた．

3.2 テストデータ

テストデータは Yahoo!API[*3] を用いて収集した Web ページのタイトルと概要文を用いた．それらの中から，3.1 に示した 5 つのカテゴリに属するものをそれぞれ人手で抽出した．このとき，検索に使ったキーワードは，医療カテゴリを例とした場合，「インフルエンザ」や「高血圧」等である．1 つのカテゴリにつき 500 件集め，指定外のカテゴリの文書として，指定したカテゴリ以外の残り 4 カテゴリの 2,000 文書の中からランダムで選んだ 500 件を用いた．

学習データの文書数を表 2 に示す．

表 2　テストデータの文書数

カテゴリ	カテゴリに属する文書	属さない文書	合計
医療	500	500	1,000
映画	500	500	1,000
スポーツ	500	500	1,000
政治	500	500	1,000
生物	500	500	1,000

3.3 実験結果

3.1 で示した学習データを用いて LIBSVM により学習を行い，3.2 で示したテストデータに対して，指定したカテゴリに属すか属さないかを分類した結果の適合率・再現率・F 値を表 3 に示す．

表 3　実験結果

カテゴリ	適合率	再現率	F 値
医療	93.10	91.80	92.45
映画	97.15	95.40	96.27
スポーツ	97.20	90.40	93.68
政治	92.86	78.00	84.78
生物	79.66	83.80	81.68
平均	92.06	87.88	89.77

*2　http://www.csie.ntu.edu.tw/~cjlin/libsvm/

*3　Yahoo!デベロッパーネットワーク http://developer.yahoo.co.jp/

4. 考察

実験の結果から，平均で適合率が 92%，再現率が 88%，F 値が 90%とほぼ正しく分類できていることがわかる．これにより，Wikipedia を学習に用いた本手法は有効であることが確認できた．

政治カテゴリの再現率が低い原因は，テストデータの政治の分野が学習データの分野よりも広く，学習データを 3 階層まで取得しただけでは足りなかったことが考えられる．また，ブログ等の様にくだけた文章で政治の話題を扱っているページは分類誤りをしやすかった．

生物カテゴリに関しては，適合率，再現率ともに低かった．この原因としては，テストデータの選択方法が考えられる．生物カテゴリのテストデータとして集めた文書は，ペットに関する文書が多かった．これらの文書は，人が見ても生物カテゴリに入れるべきか判断に迷うものが多い．実際，今回の実験では，ペットに関する文書の分類誤りが多く，結果として再現率の低下につながったと考えられる．

また，医療カテゴリに属する文書を生物カテゴリと誤判定してしまう傾向があった．これは Wikipedia の生物カテゴリに属するとして取得した文書に，医療カテゴリにも属する文書が含まれていたためと考えられる．つまり，Wikipedia のカテゴリにおいて，生物カテゴリと医療カテゴリは比較的近い存在だったため分類誤りをしやすかったと考えられる．

5. 今後の課題

今回の実験の対象としたカテゴリは，学習データが多く取れる大きなカテゴリを扱った．今後はもっと学習データが少ない，小さなカテゴリを対象にした実験も行う必要がある．

本手法は，学習データの文書数が多い場合，学習に時間がかかる．システムとして利用する場合，学習データの多い大きなカテゴリはあらかじめ学習しておくか，学習方法を見直して計算時間を減らす工夫が必要である．

学習データの選択方法は，今回は 3 階層までの文書を用いたが，これは医療カテゴリを基準に決めたため，他のカテゴリを扱う場合に最適とは言えない．そのため，カテゴリに応じて適切な階層の文書を取得できるように，取得方法を見直す必要がある．

生物カテゴリの実験の様に，指定したカテゴリに属するか判断が難しい文書も存在するため，そのような文書に対しては，指定したカテゴリに属する確率を求める方法を考える等の対策も必要である．

参考文献

[1]國貞暁　他，「要約情報の類似度を用いた WEB 検索支援システム」，人工知能学会第 21 回全国大会，2007.

[2]柘植覚　他，「サポートベクターマシンによる適合性フィードバックを用いた情報検索」，情報処理学会論文誌，Vol.44, No.1, pp59-67, 2003.

[3]伊東敏章　他，「Web 検索結果のページ選択を支援するジャンル分類システム」，言語処理学会第 14 回年次大会，2008.

[4]隅田飛鳥　他，「WWW 文書集合から自動抽出した意味的関係を用いた大規模な検索用ディレクトリの試作」，言語処理学会第 13 回年次大会，2007.

[5]ウィキペディア（Wikipedia），http://ja.wikipedia.org/wiki/メインページ