ISS-P-137

Web ニュース記事のカテゴリ分類の時代変化に関する検討

小原　由華† 　木原　健† 　小野　智司† 　大塚　作一†
†鹿児島大学大学院理工学研究科

1. はじめに

インターネットの発達により，Web ニュース記事の質や量と分類の方法が変化している[1][2]．そこで，本検討では，年代と共に記事の内容やカテゴリの分類基準がどのように変化しているのかを自動分類を用いて調査した．

2. 実験方法

本実験で対象とする文書データには，多様な記事が使用されている（すなわち，個々の単語の出現頻度が小さい）ため単純に単語ベクトルを使用すると分類が困難である．このため，まず，実験に使用する記事全体に潜在意味解析（LSI）を実行し，関連した概念の集合を生成し，次元圧縮を行う．つぎに，線形判別分析（LDA）を行い，分類を容易にする．最後に，サポートベクターマシン（SVM）を用いて教師有の分類を行う．

実験には goo ニュース[2]から，2009 年と 2014 年の各カテゴリ 300 記事，合計 10800 記事を使用した．また，各々半数を学習用，テスト用に使用した．

3. 実験結果と考察

各々学習とテストの年代の組み合わせ全て（4 種類）について実験を行った．図 1 と図 2 に学習 2009 年の分類結果を示す．表 1 にはその際のカテゴリ番号と対応したカテゴリ名を示す．まず，図 1 と図 2 より，「エンタメ」と「スポーツ」は，同じ年での分類，異なる年での分類でも，共に安定して高い分類結果となっていることが分かる．残りの組み合わせでも同様であった．このことから，「エンタメ」と「スポーツ」は年代が経っても記事の性質，分類の性質，共に変化していないと考えられる．また，この結果から，今回採用した上述の分類方法は信頼できると考えられる．つぎに，図 1 に示した同じ年の分類では対角線上に多く分類されている．つまり，分類結果と記事のカテゴリが一致しているものが多いということである．これに対し，図 2 の異なる年での分類では結果にばらつきが出た．これらのことから，年代によって記事の性質または分類の性質が変化していると考えられる．最後に，同じ年の分類でも，図 1 の A に示すように，特異的に他のカテゴリに誤分類されるものが見られた．この原因は，2009 年が日本がオリンピック招致を行った年であるため，「国際(～2010)/国際・科学(2011～)」の記事の多くが「スポーツ」に誤分類されたことが考えられる．

4. まとめ

異なる年代の記事で自動分類を行い，Web ニュース記事の性質と分類の性質の変化について調査した．その結果，(1) 年代によって「エンタメ」・「スポーツ」のように内容（記事の性質）に変化が小さいカテゴリと「政治」・「ビジネス」のように変化が大きいカテゴリがあること，(2) 分類基準は年代によって変化していること，が明らかとなった．

参考文献

[1] 松永ほか，“時代の変遷に伴う Web ニュース記事の情報変化の検討”，電子情報通信学会 2014 年総合大会情報・システムソサイエティ特別企画学生ポスターセッション，2014.

[2] goo ニュースサイト http://news.goo.ne.jp/

表 1.　カテゴリ番号とカテゴリ名の対応表

カテゴリ番号	カテゴリ名
1	国際(～2010)/国際・科学(2011～)
2	政治
3	ビジネス(～2010)/経済(2011～)
4	社会
5	地域
6	仕事術(2011～)
7	ライフ(～2010)/生活術(2011～)
8	トレンド(2011～)
9	エンタメ
10	スポーツ

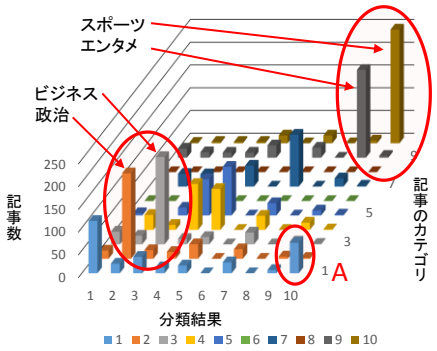

図 1.　学習 2009 年_テスト 2009 年の分類結果

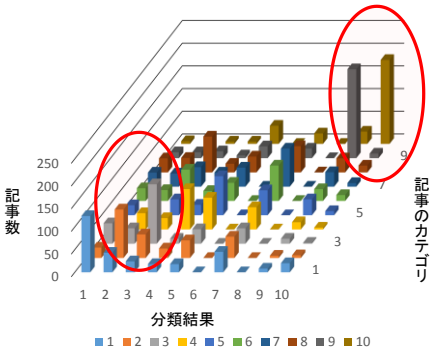

図 2.　学習 2009 年_テスト 2014 年の分類結果