6T-2

単語共起を用いたベイジアンフィルタによる中国語文章フィルタリングについて

トウ トクエツ † 吉村 卓也 ‡ 藤井 雄太郎 † 伊藤 孝行 †‡

† 名古屋工業大学大学院産業戦略工学専攻 ‡ 名古屋工業大学情報工学科

1 はじめに

現在，中国では既に世界一のネットユーザーを擁する国となった．ブログや掲示板，SNS など，ユーザーが自由に発信するサイトは急速に普及しており，多くの Web サイトでは，有害な情報に対しての対策をとっていない．そこで本稿では，中国語の有害な文書である負例と有害では無い文書である正例から, 文書の形態素解析を行い，形態素から共起関係を抽出して辞書を作成する．作成した形態素辞書と共起辞書を用いて単語間の共起情報を用いたベイジアン分類手法を採用した中国語の有害な書き込みを自動的判別する情報フィルタリングを実装した．中国語での二単語共起分類手法の精度の高さは比較実験により検証した．

2 関連研究

Paul Graham 氏のスパムメールフィルタリングはベイズ確率を本質的に使う手法であり [1]，文章から特徴単語として確率の高い上位 15 単語を抽出し，メールに含まれる特徴的な単語の結合確率を計算することでフィルタリングを行う．

Gary Robinson 氏では Paul Graham 氏のモデルを改善するため，特徴単語ではなくすべての確率を用いて確率を求めると述べている [2].

二単語間の共起情報を利用した分類手法 [3] は文章中の単語を単位としてではなく，共起 (単語の組み合わせ) の出現確率を抽出し，共起データベースを構築し，Paul Graham 氏あるいは Gray Robinson 氏と同様な手法で文章を判別する．

Chinese Harmful Sentence Filtering based on Co-occurrence words for Bayesian filter

Deyue Deng †, Takuya Yoshimura ‡, Yutaro Fujii †and Takayuki Ito †‡

†Techno-Business School , Nagoya Institute of Technology 466-8555, Nagoya, Japan

‡Dept. of Computer Science, Nagoya Institute of Technology 466-8555, Nagoya, Japan

{deng , yoshimura.takuya , fujii }@itolab.nitech.ac.jp , ito.takayuki@nitech.ac.jp

3 実装手法

3.1 中国語の特徴

中国語は語形変化（活用）が生じず，語順が意味を解釈する際の重要な決め手となる孤立語である．孤立語とは，言語を形態論的な特徴から分類したときの類型の一つで，理想的には 1 語が 1 形態素に対応する，単一の形態素からなる単語を語順や独立性の高い助詞・前置詞などの文法機能によって結びつける言語である．中国語には時制を表す文法カテゴリーが存在しない．また，格による語形変化がないのが孤立語の特徴である．孤立的な特徴をもつ言語としては他に英語などがある。

もともとベイズ分類器は強い独立性仮定に基づいた単純な確率的分類器である．つまり，各特徴単語の独立性が高いほど，より理論的値に近いことである．その点から見ると，中国語にはベイズ分類を応用するのは適切と考えられる．

3.2 形態素解析とデータベースの構築

本稿では，有害文章判別を目的として, 多くの文章を用いでデータベースを構築した. 今回, データベース構築の元となる正例は中国のポータルサイト「新浪網」[1], 「網易」[2],「搜狐」[3],「百度」[4]等のニュースの文章を 10000 件を用いて，負例は成人向け小説の文章を 10000 件を用いている．形態素解析は imdict-Chinese-Analyzer[5] を利用している．

本稿で作成したデータベースについて述べる．本稿では文章を形態素解析を行って，分割した各形態素をすべて一意の数値（ID）へとハッシュし用いている．ハッシュを用いている理由は，共起データベースの構築および保存の際にデータベースのサイズを抑えるためである．

[1] http://www.sina.com/
[2] http://www.163.com/
[3] http://www.sohu.com/
[4] http://www.baidu.com/
[5] http://code.google.com/p/imdict-chinese-analyzer/

3.3 有害度の計算

本稿で提案する中国語の文章フィルタリングは以下のように有害度を計算する．

まずは形態素 w_i の負例での出現確率 $p(w_i)$ を式 (1) に示す．ただし，n_{bad} は負例文書の総数，n_{good} は正例文書の総数，b_i は負例での w_i が 出現した回数,g_i は正例での w_i が出現 した回数である．

$$p(w_i) = \frac{(b_i/n_{bad})}{a(g_i/n_{good}) + (b_i/n_{bad})} \tag{1}$$

Paul Graham 氏による文章 D の有害度 P(D) を求める式を式 (2) に示す．

$$P(D) = \frac{\prod_{i=1}^{15} p(w_i)}{\prod_{i=1}^{15} p(w_i) + \prod_{i=1}^{15}\{1 - p(w_i)\}} \tag{2}$$

Gray Robinson 氏の方法で文章 D の有害度 I を求める式を式 (3)，式 (4)，式 (5)，および式 (6) に示す．

$$f(w_i) = \frac{s \times x + n \times p(w_i)}{s + n} \tag{3}$$

$$H(D) = 1 - \{ \prod_{w_i \in D} f(w_i)\}^{\frac{1}{n}} \tag{4}$$

$$S(D) = 1 - \{ \prod_{w_i \in D} (1 - f(w_i))\}^{\frac{1}{n}} \tag{5}$$

$$I = \frac{S - H}{S + H} \tag{6}$$

共起情報を用いる場合，式 (1) の w_i の負例での出現確率を共起 $p(w_1, w_2)$ の負例での出現確率に入れ替えることになる．

4 評価実験と考察

表 1 に各手法に対する再現率，適合率，正確度および F 値を示す．

評価実験では，テストデータの有害度を求め，判定を行う．

本稿の実験では収集した学習データの量が少ないため交差検定の手法を用いた．テストデータに安全な文章および有害な文章各 1 万件を平均 10 組に分け，順番で一つの組をテストデータにして，他の組は学習データにする．同じ手続きを 10 回繰り返し，各回で算出された精度の平均を各方式フィルタの推定精度と認める．

Gray Robinson 氏の手法では Paul Graham 氏の手法と比較して適合率は 6.06%, 正確度は 4.25%,F 値 (F-measure) という総合的指標は 3.67%向上したことが分かった．共起情報を用いる場合，Gray Robinson 氏の手法は Paul Graham 氏の手法より適合率は 0.71%, 正確度は 0.92%，F 値は 0.91%向上した．

表 1: 実験結果

手法	再現率	適合率	正確度	F 値
Graham 方式 (閾値 0.9)	99.37%	83.10%	89.58%	0.9051
Robinson 方式 (閾値 0.5)	99.79%	89.16%	93.83%	0.9418
Graham 共起方式 (閾値 0.9)	97.66%	92.49%	94.87%	0.9500
Robinson 共起方式 (閾値 0.6)	98.79%	93.20%	95.79%	0.9591

２単語間の共起情報を用いることで，Gray Robinson 方式と Paul Graham 方式二つの手法とも F 値が高い値となっている．

本中国語フィルタリング全体の評価として，四つの判別手法を適用した結果，判別率（適合率）は最大 93.20% という値が出た．既存の日本語フィルタリング [3] の最大判別率 84%と比べ，精度が上げた．本フィルタリングは有効であると言える．

5 まとめと今後の課題

本稿では，中国語の有害文書を自動的判定するため四つの手法を実装した．評価実験では，ベイジアンフィルタと共起フィルタ両方ともよい結果を得た．単語間の共起情報を利用することにより，精度が向上する結果も示した．しかし，各サンプル組を実験した結果は均一ではない．原因はベイジアン手法の計算を単位としての単語及び共起関係の中に，特徴を持っていないものがたくさん存在するためである．そのため，単語と共起関係の適切な重み付けは今後の課題である．

参考文献

[1] Paul Graham ,“A PLAN FOR SPAM”, http://www.paulgraham.com/spam.html

[2] Gary Robinson,“A Statistical Approach to the Spam Problem”, http://www.linuxjournal.com/article/6467?page=0,0

[3] 安藤哲志，藤井雄太郎，伊藤孝行,“複数単語間の共起情報を用いた有害文章判定手法の提案”，2010 年度人工知能学会全国大会（第 24 回）