MIIDAS: 情報の適合的選別による文書フィルタリング

5Q−9

石川 開　奥村 明俊　村木 一至
NEC C&C メディア研究所

1 はじめに

我々は、情報集配信サービスMIIDAS[1]により、ユーザーに必要な様々な情報をフィルタリング、分類・検索、活用・発信するためのプラットフォームを提供することを目指している。本稿では、このうち、我々が取り組んでいるテキスト情報に対するフィルタリングシステムの研究開発に関する報告を行なう。現在、尤度の推定にベイズの定理を用いた単語ベースモデル[2]によるフィルタリングシステムを実現し、新聞記事情報を対象としたフィルタリングサービスを運用している。単語ベースモデルにおいて、文書中に出現する全単語に対して単語空間を設定すると、大量の文書情報を扱う場合に単語数の増加による単語空間の次元数の爆発という困難が生じる。このため、単語空間の設定のための重要語の選別が不可欠となる。
単語空間中で、自立語以外の単語は単語自身が独立して意味を持たないため、単語ベースモデルのための単語空間として適切ではない。さらに単語ベースモデルにおける単語空間の設定には、以下の二つの問題がある。

(1) 一般的な語義を持つ単語や語義の曖昧性を持つ単語は、文脈に依存して語義や用法が変化する。単語ベースモデルでは、これらの単語の文脈上の語義や用法の違いを区別することができないため、語義によって異なる本来の統計分布とは異なる扱いとなる。これらの単語の単語空間を占める割合が高くなると、尤度の推定におけるこれらの単語からの寄与により、精度の低下が引き起こされる。

(2) 日本語の文書に関しては形態素解析の語切り誤りにより、カテゴリを特定する寄与が大きい固有名詞や複合語などが分割されてしまうという問題がある。
本稿では、上記の二つの問題に対して(1)文書のカテゴリを特定する寄与の大きい単語のみを選別する方法、および(2)固有名詞の追加による分割された固有名詞の一単語としての取扱い、の方法を用いることにより、上記の問題を解決したのでその方法と可能性について示す。

MIIDAS: Document filtering with word co-occurrence probability
ISHIKAWA Kai, OKUMURA Akitoshi, MURAKI Kazunori
C&C Media Research Laboratories, NEC Corporation

2 重要語の抽出

(1) 重要語は、関連記事から形態素解析により名詞とサ変を抽出し、その中で“このため”、“見込み”、“基本”のように、単語単独では十分に意味の特定されない単語を主観的に判断し、除去を行なった。(2) カテゴリを特定する寄与の大きい固有名詞に関して、形態素解析の語切り誤りにより分割される単語をまとまった一単語として扱うために、企業名や省庁名をリストで補った。さらに一部上場企業名に関しては、同じ企業の異なる略称による表記も扱うことができるように、一部上場企業名データベースを用いて単語を複数の表記へ増強した。

3 フィルタリングモデル

ユーザーモデルの学習は、トレーニングデータの記事中に重要語の出現する頻度情報から、各単語の関連記事、非関連記事に対する対数尤度をベイズの定理[2]を用いて推定する。出現頻度から出現確率を推定する際に最尤推定を用いると出現頻度の値の極端に小さい単語に対して適切な尤度が計算できないため、ラプラス推定量[3]を用いて確率の推定を行なう。このようにして求めた関連記事、非関連記事に対する対数尤度の比を取り、各重要語の関連記事への関連度を表現する重みとした。テストデータの記事の判定は、記事中に出現する重要語の重みの和によって対数対数尤度の比を計算し、閾値により判定を行なった。

4 評価

4.1 評価データ

新聞記事情報としては、一般誌(朝日、読売、毎日)96年10月6日～97年4月25日を用いた。図1に示すような新聞記事を人手で選別した記事ヘッドラインの集配信サービスをモデルとし、ヘッドラインから5W1H情報を用いた関連記事抽出技術[4]を用いて正解記事集合を生成した。

図１：記事ヘッドラインの一例

Ｎ社　手のひらサイズの光コネクタ加工装置を開発
Ｏ社　輝度2.5倍のプラズマディスプレイを開発

このようにして正解記事を含む新聞記事情報を二つに分け、96年10月6日～97年3月31日の記事をトレーニングデータ、97年4月1日～97年4月25日の記事をテストデータとして用いた(表1)。

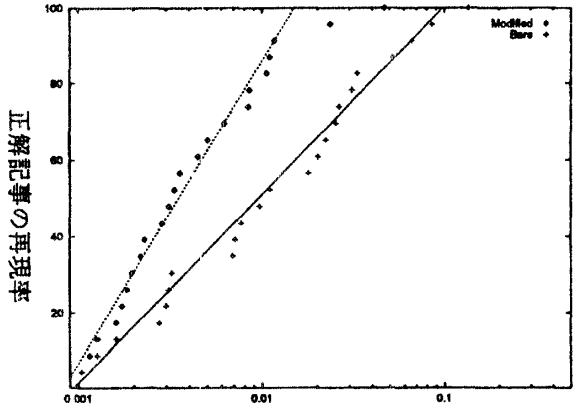
表１：記事データと正解記事

	総記事数	正解記事数
トレーニングデータ	54755	193
テストデータ	8678	23

4.2　結果

2節で述べた重要語抽出法により、356単語からなる単語空間(Modified)を生成した。比較評価のために名詞とサ変の抽出のみにより、3907単語からなる単語空間(Bare)を生成した。これらの2つの単語空間(Modified,Bare)を用いてトレーニングデータからモデルの学習を行ない、テストデータでの評価を行なった。図2では、フィルタリングにより抽出した記事数の総記事数に対する割合を横軸(対数表示)に取り、正解記事の再現率を縦軸に取っている。例えば、正解記事の再現率80%に対し、単語空間(Bare)では適合率5.3%、単語空間(Modified)では28%となっている。このようにプロファイルの縮小化と精度の向上が同時に達成されるという結果を得た。

図２：改良した単語空間による精度の向上



4.3　考察

一般に、単語ベースモデルでは、モデルの学習において単語の出現確率を推定するために統計的に十分な情報量が得られないというデータスパースネス問題がある。この問題は、2節で述べた重要語の選別と、出現確率の推定にラプラス推定量[3]を用いた確率の推定によって一部改善されている。しかし、2節の重要語選別では単語の出現頻度の統計的信頼性を考慮していないため、尤度推定の精度を下げるような単語の除去が完全には行なわれていないという問題がある。この問題を解決するために、次のような相互情報量を用いた単語の選別を行なう。記事の判定$C=\{p,n\}$と各単語$w \in W$に関する相互情報量

$$I(W,C)=\sum_{w\in W}\sum_{c\in C}P(w,c)\log\frac{P(w,c)}{P(w)P(c)} \qquad (1)$$

を計算し、次の閾値θを超えるものを選別する[5]。

$$\theta=\beta\cdot\frac{(k_W-1)\cdot(k_C-1)\cdot\log_2 N}{2\cdot N} \qquad (2)$$

ここで、βは重み変数$(0\leq\beta\leq 1)$、k_Wは単語空間Wの単語数、$k_C=2$、Nはデータ数。この方法により、記事の判定に対する寄与が高く、かつ統計的に信頼性の高い単語を抽出できることが予想される。今後、この手法を用いた改良を進めていく予定である。

5　おわりに

本稿では、尤度の推定にベイズの定理を用いた単語ベースモデル[2]によるフィルタリングシステムにおいて、文書のカテゴリを特定する寄与の大きい単語のみを選別する方法、および固有名詞の追加を組み合わせた重要語抽出法を用いる方法について提案および評価を行なった。また、記事の判定に対する単語の統計的信頼性を考慮した重要語抽出の改善方法を提案した。今後は、本研究で有効性の示された単語選別技術を、単語ベースモデルを用いたフィルタリングシステムにおける単語空間の自動獲得技術の実現に応用することを検討している。

参考文献

[1] 奥村明俊, 池田崇博, 村木一至:"MIIDAS: 情報の選別的共有のためのオントロジ構築とその増進的学習", 情報処理学会第55回全国大会, 5Q-08, 1997.

[2] Louise Guthrie, Elbert Walker, and Joe Guthrie. Document classification by machine: Theory and practice. Proc. COLING'94,pages 1059-1063, 1994.

[3] Williams A.Gale and Kenth W.Church. Poor estimates of context are worse than none. Proc.DARPA Spee. Nat. Lang. Worksh.,pages 283-287, 1990.

[4] 池田崇博, 奥村明俊, 村木一至:"MIIDAS: 情報の選別とEasyReadingのためのエピソード", 情報処理学会第55回全国大会,5Q-10, 1997.

[5] 李航, 竹内純一: "証拠の強さと信頼度を考慮した日本語同形異音語の読み分け", 情報処理学会研究報告 97-NL-119, pp.53-59, 1997.