Yahoo!知恵袋に投稿されたテキストに対する著者判別

† 西村 涼　† 渡辺 靖彦　†† 村田 真樹　† 岡田 至弘
† 龍谷大学大学院　理工学研究科　情報メディア学専攻
r_nishimura@afc.ryukoku.ac.jp, {watanabe,okada}@rins.ryukoku.ac.jp
†† 独立行政法人 情報通信研究機構
murata@nict.go.jp

1　はじめに

インターネット上には文章によるコミュニケーションが可能なサービスは多い。例えば、Yahoo!知恵袋[1]というサービスでは、質問を書いておくと、このサービスを利用している他のユーザがその質問に対する答えを書いてくれる。このようなサービスは会員登録制であることが多く、ID によるユーザの管理が適切なサービスの運営に重要な役割を果たしている。しかし、一人のユーザが複数の ID を使うことで、適切なサービスの運営が困難になることがある。例えば、悪質なユーザであると ID で判別しサービスの利用を停止しても、異なる ID を利用して同様の行為を繰り返すユーザもいる。管理側はこのようなユーザをきちんと管理しなければ、適切な利用を行っているユーザにサービスを見放されるおそれがある。しかし、このようなユーザを管理するにはアクセスログなどを詳しく解析しなければならず、コストがかかる作業となる。複数の ID を利用している悪質なユーザをメールのスパムフィルタリングのようにフィルタリングする技術があれば管理コストの削減が期待できる。

インターネット上の文書を機械学習を用いて分類することを目的としている研究は多い [1][2]。池田ら [1] は blog テキストを対象として著者の性別の推定を行っているが、分類の目的の点で本研究と異なる。また、坪井ら [2] は E-mail を対象に機械学習法を利用して著者判別を行っているが、対象としているテキストの点で本研究と異なる。

本研究では、Yahoo!知恵袋に投稿された文章からその文章を書いたユーザを機械学習法を用いて特定する方法を提案する。本研究で提案する手法でユーザを特定できれば、複数 ID を使っているユーザを監視することが可能となる。

2　実験データと素性

2.1　実験データ

国立情報学研究所から提供されている Yahoo!知恵袋のデータ[2] を用い、「パソコン、周辺機器」「病気、症状、ヘルスケア」「政治、社会問題」の 3 つのカテゴリから実験データを作成した。提供されているデータは、2004 年 4 月から 2005 年 10 月までに Yahoo!知恵袋に投稿された質問約 311 万個、回答約 1347 万個から構成されている。この中から「パソコン、周辺機器」「病気、症状、ヘルスケア」「政治、社会問題」の 3 つのカテゴリに投稿されたすべての質問、それぞれ 171848 個、84364 個、78777 個を取り出し、実験データとして利用した。また、3 つのカテゴリに投稿されたすべての回答も、それぞれ 474687 個、289578 個、403306 個を取り出し、実験データとして利用した。

表 1: 利用する素性の種類

素性ラベル	説明
s1	本文の形態素解析の結果（例：「おいしい」「ご飯」「を」「食べ」「た」）
s2	それぞれの文の形態素解析の結果とその文番号
s3	本文から取り出した文字 3-gram
s4	それぞれの文から取り出した文字 3-gram とその文番号
s5	それぞれの文の文頭の 1〜10 文字（例：「し」「しか」「しかし」）
s6	それぞれの文の文末の 1〜10 文字（例：「だ」「だっ」「だった」）
s7	PrefixSpan を利用した頻出系列パターン（2.3 節）

2.2　素性

表 1 にユーザを推定するのに利用する素性を示す。表 1 の本文とは、解析対象の文章全体のことである。また、それぞれの文とは、解析対象の文章を構成するそれぞれの文のことである。s1 と s2 は形態素解析した結果を素性としたものである。このとき、s2 はある文の形態素が別の文に現われた場合に別の素性として扱うため、文の番号を付与している。s1 はそのような区別は行わない。また、形態素解析には、Mecab[3]を用いた。

s3 と s4 で文字 n-gram として 3-gram を採用しているのは、小高ら [3] の先行研究で文字 n-gram のうち日本語には 3-gram が良いとされているからである。ただし、s4 も s2 と同じように別の文で現われる文字 3-gram は別の素性として扱うため、文の番号を付与している。また、著者ごとの特徴的な表現は文頭や文末に現われることが多いと考え、s5 と s6 の素性を与えた。

s7 は Pre xSpan を利用した素性であり、これについては次節で詳しく説明する。

[1] http://chiebukuro.yahoo.co.jp/
[2] http://research.nii.ac.jp/tdc/chiebukuro.html
[3] http://mecab.sourceforge.net/

2.3 PrefixSpan

Pre xSpan とは、系列パターンを高速に抽出できる手法であり、多くの文書分類の研究で利用されている。例えば、松本ら [4] は評価文書を肯定的な立場、否定的な立場に分類するタスクで用い、坪井ら [2] は、メールの集合を送信者ごとに分類するタスクに適用している。系列パターンは、文中に出現する連続または非連続な単語列のパターンで、連続な単語列のパターンしか抽出できない n-gram では得られないような言い回しなどの表現を得ることができる。

本研究では、系列パターンを取り出す対象として形態素を選んだ。出現パターン素性として、実験のために取り出した文章の中で、出現回数 5 以上、アイテム数 3 以上、最大ギャップ数 1、最大ギャップ長 1 の系列パターンを取り出した。系列パターンの抽出には、Pre xSpan-rel[4] を用いた。

3 質問者特定の実験と有効な素性の分析

本研究では、Yahoo!知恵袋に投稿された質問と回答を分けて、それぞれの文章を書いたユーザを特定することを目的とする。ユーザの特定には、最大エントロピー法 (MEM) とサポートベクトルマシン (SVM) の 2 つの機械学習の方法を用いる。MEM として maxent[5]を用い、SVM として TinySVM[6] を用いた。SVM のカーネルとして 1 次の多項式カーネルを選び（予備実験において 2 次の多項式よりもおおむね性能が良いことを確認している)、ソフトマージンパラメータ C を 1 として実験した。実験は 10 分割クロスバリデーションで行った。

最初に、Yahoo!知恵袋に投稿された質問文を対象として、その質問文を書いたユーザを特定する実験を行う。まず、実験を行うためにそれぞれのカテゴリから質問を投稿している回数が多いユーザを 10 人取り出す。次に、この 10 人の中で質問の投稿数が多い 5 人を取り出す。この 5 人をユーザを特定する実験に利用し、特定したいユーザとそれ以外の 9 人とを区別する実験を 5 人分行うこととする。実験に利用するカテゴリは、「パソコン、周辺機器」「病気、症状、ヘルスケア」「政治、社会問題」の 3 種類であり、それぞれのカテゴリにおける質問数が多いユーザの質問数を表 2 に示す。ただし、この表でのユーザ A、B、C、D、E は、それぞれのカテゴリでの質問の投稿数が多い 5 人のユーザを便宜的に表しているだけであり、カテゴリ間で同一のユーザというわけではない。

これらのデータに素性を与え、ユーザを特定する実験を行った。実験データの構成は、特定したいユーザが書いた質問文（正例）の数にそれ以外のユーザが書いた質問文（負例）の数が合うようにした。例えば、ユーザ A を特定する実験では、ユーザ A の質問文を正解事例とし、それ以外の 9 人の質問文からランダムで不正解の事例を 1 つ選んだ。実験結果を表 3 に示す。表 3 から、それぞれのカテゴリにおけるユーザの判別精度の平均値として 88%程度が得られたことがわかる。

次に、実運用を想定した実験を行う。分類器には MEM を利用する。特定したいユーザの質問文が複数ある場合において、分類器への入力にそれらすべての質問文を利用する。複数の質問文をひとつずつ分類器に与え、それぞれで得られた確率値の積をもとに分類先を決定する。このような方法を利用して、表 3 の実験で得られた分類器に質問文を 2 個、3 個、4 個、5 個ずつと変化させて与える実験を行った。実験は表 3 での実験と同じ条件で、分類器に入力する質問の数のみを変化させた。それぞれのカテゴリにおける 5 人のユーザの判別精度を平均し、グラフ化した結果を図 1 に示す。この結果から入力に与える質問文を 5 個にすると 99%程度の精度でユーザを特定できることがわかった。

また、機械学習として MEM を用いて有効な素性の分析を行った。MEM でもとまる α 値を正規化した値 (ここでは正規化 α 値と呼ぶ。二分類においてこの値の和が 1 になるように正規化している。) をもとめた [5]。正規化 α 値は、その値が大きいほどその素性が重要であり、小さいほど重要でないことを示す。素性は分析しやすいように s1、s3、s5、s6、s7 を選んだ。s2 と s4 を省いたのは、s1 と s2、s3 と s4 は似ているため分析しにくいと考えたからである。それ以外の実験条件は変えていない。

表 4 に「パソコン、周辺機器」カテゴリにおける結果を示す。ただし、全角の空白は「△」記号に置き換えてある。この結果からユーザごとの特徴的な素性について議論する。ユーザ A では、s3 の「すが、」から文節をこの表現で区切りやすいことがわかる。また、「ますか」という表現で質問をしやすいこともわかる。ユーザ B では、s6 の「んですか？」からこの表現を文末に使うことが特徴的なことがわかる。ユーザ C は「けど、」という区切りを使いやすい。ユーザ D は、ユーザ B と異なり「のですか？」という表現を文末に使いやすことがわかる。ユーザ E は他のユーザと異なり「いいのでしょうか？」という長く特徴的な表現がある。

4 回答者特定の実験と有効な素性の分析

Yahoo!知恵袋に投稿された回答文を対象として、質問者の特定と同様の方法で実験を行った。実験を行うためにそれぞれのカテゴリから回答を投稿している回数が多いユーザを 10 人取り出した。ユーザの特定は、10 人のうち投稿数が多い 5 人をそれ以外の 9 人から区別することによって行う。カテゴリごとのユーザ数は表 5 に示す。これらのデータに素性を与えて回答者を特定する実験を

[4]http://prefixspan-rel.sourceforge.jp/
[5]http://www2.nict.go.jp/x/x161/members/mutiyama/software.html#maxent
[6]http://chasen.org/ taku/software/TinySVM/

表 2: それぞれのカテゴリにおける質問の投稿回数が多い 10 人のユーザとその質問数

カテゴリ名	A	B	C	D	E	その他 (5 人)	合計
パソコン、周辺機器	407	1178	452	340	734	1342	4453
病気、症状、ヘルスケア	200	236	543	362	237	783	2361
政治、社会問題	587	667	531	808	1021	2287	5901

表 3: 1 つの質問文からの質問者特定の実験結果

特定したいユーザ	パソコン、周辺機器		病気、症状、ヘルスケア		政治、社会問題	
	MEM	SVM	MEM	SVM	MEM	SVM
A	**90.04%**	89.43%	**86.25%**	85.75%	92.50%	**92.67%**
B	94.56%	**95.16%**	88.98%	**90.67%**	**86.50%**	**86.50%**
C	84.45%	**84.73%**	**94.01%**	93.92%	87.94%	**88.22%**
D	**86.02%**	84.41%	84.80%	**84.94%**	88.49%	**88.86%**
E	86.10%	**86.71%**	**82.91%**	**82.91%**	**83.83%**	83.79%
平均	**88.23%**	88.08%	87.21%	**87.63%**	87.85%	**88.01%**

表 4: 5 人の質問者を特定する場合に有効であった上位 10 個の素性（ただし、空白記号は「△」で表している）

A		B		C		D		E	
素性	有効度	素性	有効度	素性	有効度	素性	有効度	素性	有効度
s1:、	0.871	s1:事	0.689	s1:ｘｐ	0.696	s1:△	0.722	s1:って	0.732
s1:△	0.845	s3:しえて	0.674	s1:、	0.673	s3:Ｗｉｎ	0.660	s1:の	0.726
s1:って	0.743	**s3:おしえ**	0.671	**s3:るのは**	0.639	**s6:のですか？**	0.644	s3:という	0.684
s3:って、	0.725	s1:おしえ	0.671	s1:こと	0.629	s3:どうし	0.644	**s3:いいの**	0.670
s5:W	0.629	**s6:んですか？**	0.669	s5:ロ	0.625	**s1:どうして**	0.643	s1:という	0.661
s3:か？△	0.629	s3:タイピ	0.667	s1:一番	0.619	s3:うして	0.637	s1:物	0.657
s3:ますか	0.624	s3:イピン	0.667	**s3:けど、**	0.618	s1:可能	0.630	**s6:いいのでしょうか？**	0.648
s3:すが、	0.619	s3:ピング	0.664	**s1:まで**	0.616	s1:や	0.625	s3:ネット	0.645
s3:ＵＳＢ	0.614	s1:て	0.650	s3:ちらが	0.615	s1:し	0.617	s6:いのでしょうか？	0.641
s1:ＵＳＢ	0.614	s3:方を教	0.640	s3:ｗｉｎ	0.614	s1:「	0.617	s1:ある	0.635

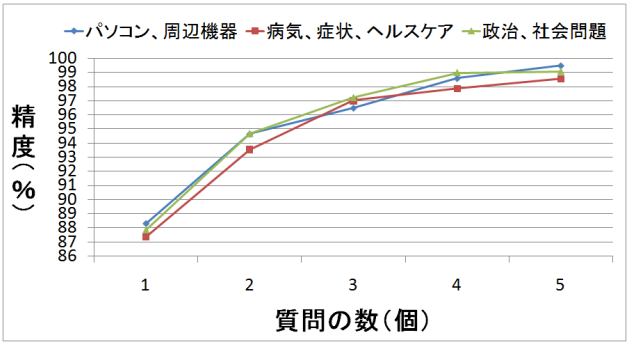

図 1: 入力に与える質問の数による判別精度の変化

行った。実験結果を表 6 に示す。実験結果から平均的に 94%程度でユーザを特定できることがわかった。回答者の特定は質問者の特定よりも精度良く行えた。これは、学習データの量が関係していると考えられる。また、カテゴリ間で精度にあまり差がないこともわかった。これは質問者を特定する時と同じ傾向であった。

MEM を用いて回答文が複数ある場合について実験を行った。表 6 の実験で得られた分類器に回答文を 2 個、3 個、4 個、5 個ずつ与えた。実験は、表 6 の実験と同じ条件で、分類器に入力する回答文の数のみを変化させた。それぞれのカテゴリにおける 5 人のユーザの判別精度を平均し、グラフ化した結果を図 2 に示す。この結果から、質問者の特定の実験と同様に回答者の特定においても判別したいユーザの文章が複数あることによって 99%程度の精度が得られることがわかった。

機械学習として MEM を用いて有効な素性の分析も行った。実験で用いる素性は質問者の特定の実験と同じである。「パソコン、周辺機器」カテゴリにおける 5 人の回答者を特定する時に役立った素性について議論する。表 7 に結果を示す。ただし、全角の空白は「△」記号に置き換えてある。全体的な傾向として、質問文を書くときにはかしこまって書いているのに対して、回答文を書くときにはかしこまらずに自由に書いていることがわかった。これは、質問するユーザと回答するユーザの立場の違いが関係していると考えられる。質問を書くときには、回答を得られやすくするため、文章表現を丁寧に書くのが普通だからである。

ユーザごとの書きぶりも特徴的である。ユーザ A は数多くの「・」記号を使って文章を書く傾向が見られる。ユーザ B は、文末記号である「。」を「。。。。。。。。。。」のように書くのが特徴的である。また、「、△△」のように全角の空白を「、」の後ろに書くのが最もユーザ B を特定するのに重要な書きぶりだという結果も得られている。ユーザ C は 5 人のうちで最も多くの記号を使って回答文を書いている。例えば、「s5:＞」から文頭に「＞」を付けて書いていることが特徴的という結果が得られた。ユーザ D は「れば、」「ならば」「らば、」の表現を使いやすい。ユーザ E もユーザ B と同じく文末記号をいくつかつなげて文末に書く傾向がある。ただし、ユーザ B よりも文末記号をつなげる数が少ないようである。

表 5: それぞれのカテゴリにおける回答の投稿回数が多い 10 人のユーザとその回答数

カテゴリ名	A	B	C	D	E	その他 (5 人)	合計
パソコン、周辺機器	2789	3845	13656	3986	2687	12557	39519
病気、症状、ヘルスケア	1399	1541	1468	1460	4542	5461	15871
政治、社会問題	3827	3293	4270	3057	3947	12796	31190

表 6: 1 つの回答文からの回答者の特定の実験結果

特定したいユーザ	パソコン、周辺機器		病気、症状、ヘルスケア		政治、社会問題	
	MEM	SVM	MEM	SVM	MEM	SVM
A	90.46%	**90.85%**	93.60%	**93.74%**	**88.62%**	88.43%
B	**96.83%**	96.76%	95.39%	**95.91%**	**94.23%**	94.10%
C	**96.17%**	96.14%	**97.71%**	97.54%	**94.77%**	94.43%
D	94.76%	**94.90%**	**96.09%**	96.06%	93.52%	**93.60%**
E	87.53%	**87.56%**	**95.40%**	95.09%	**94.73%**	94.65%
平均	93.15%	**93.24%**	95.63%	**95.66%**	**93.17%**	93.04%

表 7: 5 人の回答者を特定する場合に有効であった上位 10 個の素性（ただし、空白記号は「△」で表している）

A		B		C		D		E	
素性	有効度	素性	有効度	素性	有効度	素性	有効度	素性	有効度
s1:△	0.823	**s3:、△△**	0.914	**s5:＞**	0.960	s1:△	0.822	s6:△	0.933
s3:・・・	0.805	s5:△△	0.884	s3:とは、	0.828	**s3:れば、**	0.749	s1:下記	0.728
s1:・・・・・・・・・・・・・・・・・・・・・・・・	0.770	**s3:。△△**	0.812	**s1:』**	0.771	s3:ので、	0.729	**s6:。。。。。**	0.685
s1:　”	0.769	s1:、	0.751	**s1:『**	0.771	s1:その	0.706	s3:△ ‥	0.315
s6:か。	0.724	s3:？△△	0.747	**s6:か？**	0.764	**s3:ならば**	0.706	s1:,	0.681
s3:タブで	0.718	s1:△	0.746	s3:ｔ；&	0.759	s1:こと	0.704	s6:？	0.673
s1:選択	0.686	s1:ご	0.730	s1:＞	0.748	s3:**らば、**	0.701	**s3:かな？**	0.670
s3:・△・	0.328	s1:貴方	0.716	s3:△：△	0.252	s3:はない	0.689	s1:当方	0.667
s3:が、△	0.656	**s6:。。。。。。。。。。**	0.701	s6:か。	0.738	s1:もの	0.685	s6:る。	0.661
s1:アプリ	0.654	s3:す。△	0.685	s1:、	0.736	s1:私	0.685	s1:検索	0.657

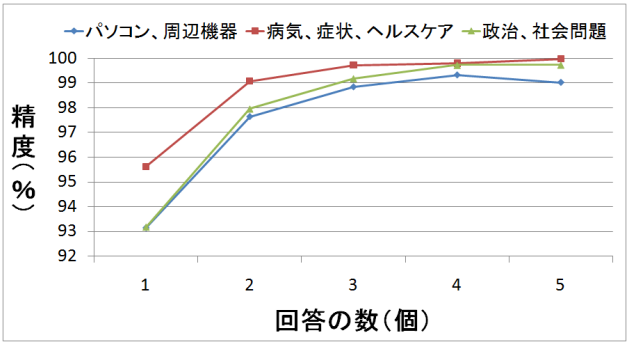

図 2: 入力に与える回答の数による判別精度の変化

5　おわりに

本研究では、Yahoo!知恵袋に投稿された質問・回答文を対象として、その文章を書いたユーザを判別する研究を行った。実験の結果、1 つの質問・回答文を分類器に与える実験では、それぞれ 88%と 94%程度の精度でそれらを書いたユーザを判別することができた。一方、複数の質問・回答文を分類器に与える実験では、両方とも 99%という高い精度で判別できることがわかった。

MEM に与えた素性を分析することで、ユーザごとに特徴的な書きぶりがあり、それが判別に役立っていることがわかった。分析結果から 3-gram の素性と文頭・文末の文字列素性は特徴的な書きぶりをとらえるのに有効であることを確認した。

謝辞

本研究を実施するにあたり、ヤフー株式会社が国立情報学研究所にて研究用に公開した Yahoo! 知恵袋のデータを利用させていただきました。ここであらためて感謝とお礼を申し上げます。

本研究の一部は、日本学術振興会科学研究費補助金基盤 (C)「心豊かなコミュニケーションを促進する質問作成支援システムの作成」(課題番号 20500106) の助成を受けて行われたものです。

参考文献

[1] 池田, 南野, 奥村: “blog 著者の性別推定”, 言語処理学会 第 12 回年次大会, (2006).

[2] 坪井, 松本: “Authorship Identification for Heterogeneous Documents”, IPSJ-NL-148, (2002).

[3] 小高, 村田, 高, 諏訪, 白井, 高橋, 黒岩, 小倉: “n-gram を用いた学生レポート評価手法の提案”, 電子情報通信学会論文誌, Vol. J86-D-I No.9, 2003.

[4] 松本, 高村, 奥村: “単語の系列及び依存木を用いた評価文書の自動分類”, FIT2004, pp.212-214, (2004).

[5] Murata, M., Nishimura, R., Doi, K., Kanamaru, T. and Torisawa, K.: “Analysis of the Degree of Inportance of Information Using Newspapers and Questionnaires”, IEEE NLPKE-08, pp.137-144, (2008).