人工知能学会研究会資料
SIG-SW&ONT-A302

Boosting を用いた評判の信頼性評価方法

荒井幸代　村上陽平　杉本悠樹　田仲正弘
京都大学大学院情報学研究科社会情報学専攻
〒606-8501　京都市左京区吉田本町
{sachiyo,yohei,sugimoto,mtanaka}@kuis.kyoto-u.ac.jp

情報の信頼性評価の手段として評判に基づく方法を考える. 評判は, 統計的な量的側面と, セマンティクスに基づく質的側面の組み合わせによる評価であると考えられる. 統計的学習によって正しい評価を得るためには, 文書の収集対象を Web 全体に広げ, 能動的に評価情報を集めた上で, それらのセマンティクスに基づく分類が必要となる. 本研究では, 非定型の文書の分類問題に対して, 代表的な統計的分類手法の利用可能性をいくつかの実験から考察する. とくに, 理論的背景のあるブースティングによる評判の統合方法とその分類精度からその有効性を議論する.

1 はじめに

情報の氾濫の中, ユーザは要求を充足し, 信頼できる情報提供を必要としている. Web 利用において, 情報を信じる／信じないはユーザの情報リテラシーに委ねられている. リンク構造や文書構造から Web を質的に評価する方法は提案されている[Dhyani et.al]ものの, 形式的な質の良し悪しと, 内容の信頼度は必ずしも一致しない. 相次ぐインターネット上の誹謗中傷問題や詐欺事件はユーザが Web 情報の信頼性評価規範を持たないことに起因する.

信頼性評価には, (1)既存の知識を用いて論理的な整合性から判断する方法と (2)情報や情報源に関する評判に基づく方法の二つに大別できる. 前者は強力な方法ではあるが, オープンな Web 情報に対しては限界がある. 一方, 後者の評判（多勢）による信頼性判断は, 匿名性の高い Web 上では危険性を孕んでいる. また, 人気や権威のあるサイトに信頼が委ねられている現状はひとつ間違えば, 情報の切り取りや情報操作を受ける可能性もある.

我々はWeb情報に対する信頼のフィルタリング機能の実現を目標としている. あるトピックに関してWebに存在する情報を収集し, 内容の意味するところ（以後, セマンティクス）の相違に基づいて分類し, それらの分布を正確に提示することを考える. セマンティクスの分布を示すことによって, ユーザに情報の見方に関する物指しを与え, 意思決定を支援する. すなわち, ここでの信頼フィルタとは, 情報を切り取ることなく, ユーザに提示する機能である. 本研究では, その第一歩として, 従来文書分類に適用されてきた統計的学習法が, セマンティクスの相違による分類に適用できるのかどうかに関して実験を通じ, 考察する.

2 問題領域

2.1 社会的トピックの論調分類

ここでは, 社会的トピックに対する論調を分類, 説明する問題を考える. CD や本など商品情報に関しては, 評判サイトに集められたランキングの分布やレビュー情報を比較的容易に得ることができる. しかし, 評判サイトによる評価の問題点として, 統計量評価に十分なサンプル数が集められないことや, 「否定的評価は肯定的な評価に比べて集まりにくい」というポリアンナ効果[Resnick and Zeckhauser 2000]が指摘されている. 本研究では, サンプル収集対象を Web 全体に広げ, （つまり, あるサイトに集められた任意の提供情報に限定しない）, 能動的に評価情報を収集することによってサンプル数の不足を補い, Web 上の評価情報に基づいた信頼性評価を提供す

る.

また,文書からの情報抽出に関して, 従来から対象とされてきたドメインは商品の推薦情報や評判情報に関するものであった. ここでは, 商品が評価される属性集合をある程度絞り込むことができる. したがって, 文書の特徴量抽出が比較的容易であり, 統計的学習による分類精度が期待できる. 一方, 本研究では, ある社会的トピックに対する「Pro/Con」の分布を提示することが目的である. ここでは, 商品評価情報の分類とは異なり, (1)評価属性が多様であること, (2)評価属性値に関しても, 商品価格のように陽に与えられているとは限らない. したがって, 統計的学習に必要な文書の特徴の抽出が困難な領域である.

本研究では, 自然言語処理を最小限にとどめ, 人手で文書の特徴抽出を行い, これを統計的機械学習を適用して, Pro/Con の二分を試みる. 文書の特徴抽出は, セマンティクスの抽出に他ならず, 特徴抽出手順が形式化できれば, 文書の機械可読性実現をめざした Semantic Web 研究においても貢献が期待できる. 本稿では,「少年法改正」に関する意見を Web 情報から取り出し, 賛否に分類するための「特徴抽出」→「機械学習による分類」→「精度評価」→「特徴抽出」を試行錯誤的な繰り返し実験の経過を報告し, 半構造データ分類に必要な特徴抽出法と適用可能な機械学習法に関する知見を得ることにつなげる.

2.2 関連研究

2.2.1 セマンティックウェブコミュニティにおける信頼性研究

セマンティックウェブ研究においても, 情報の信頼性の検証はトラスト(Trust)の研究の一環として重要な課題である. その情報が信頼できるかどうかは, 情報の内容に依存するため, セマンティクスによる支援は有用であると考えられる. 現在開発されている情報の信頼性の検証技術として USC(University of Southern California)で研究されている TRELLIS と呼ばれる情報分析ツールが挙げられる[Gil 02]. TRELLI はユーザの情報分析や意思決定過程で, 各情報源に対する信頼性の評価や, 派生する情報の構造化, 分析結果のセマンティクアノテーションを行うツールである.

まず分析過程や意思決定過程で用いた情報を, 自然言語やあらかじめ用意された形式言語によって構造化し, *unit* を生成する. 次に *unit* に挿入された情報源属性に付与された信頼度に基づいて, 情報源の総合的な評価を引き出す.

また, ユーザは情報分析中に関連するトピックについて, 他の人たちが考慮し分析した情報や情報源を検索することが可能であり, その分析結果を自分の分析に取り込むこともできる. つまり, 取り込まれた情報や情報源は利用者の目的に合わせて再利用される. 最終的に, 利用者が TRERISS を用いて行った分析結果は, XML や RDF, そして DAML+OIL の 3 種類のセマンティックマークアップ言語で提供される. これらの分析結果はその後, Web ドキュメントとして公開することで他のツールにも利用可能になる.

TRELLIS では, あるトピックの分析結果をその分析過程とともに開示することが, 他者に対する情報の信頼性を提供している. しかし, 導かれた分析結果は, 分析者の主観に基づいて取捨選択された情報から成り立っており, この情報の切り取りが他者の信頼性評価にどのように影響するかは議論されていない. 本研究では, 情報の分析過程の開示と, その情報に対する多様な見方を提供し, どの情報が信頼できるのかをユーザ自らが判断できるような支援機能の実現をめざしている.

2.2.2 推薦システム (Recommendation System)

RS は, (1)協調フィルタリング(Collaborative Filtering; CF)に基づく方法 [Resnick 97][Schafer 99]と, (2)コンテンツに基づいて, 情報に含まれるアイテムと利用者のニーズの内容比較による方法の二つがあることは既に述べた. 前者は様々な領域[Goldberg 92]で, その効果を示しているが, 以下の 4 つの問題が指摘されている.

- Cold Start: 初期段階では, 利用者に類似した嗜好の持ち主を見つけるのが困難.
- Sparsity:推薦すべき項目(item)数が多いとき, 各項目に関するデータが集めにくい
- First Rater:rate 値が無い場合には利用できない.
- Popularity Bias: 特殊な嗜好を持ったユーザに対する推薦は困難.

すなわち, 各情報, あるいは評価項目に対して, 十分な Rating データが存在しなければ, 精度の良い推薦を生成することができないという問題を抱えている. また, たとえ十分なデータが存在しても, 悪意のある Rating による Fraud や Deception の影響を受けやすい. 一方, コンテンツに基づいた方法は, 必要とする情報の特徴を抽出し, 既存の情報と照合することによって推薦を生成する枠組みであるため, 上記の 4

つの問題は生じない. しかし, コンテンツの「信頼」をどのような形で裏付けるのかが問題となる. CF が十分なデータ量を前提に量的「信頼」が期待できる枠組みであるのに対して, コンテンツの「信頼」は推薦する根拠を含んだ質的な信頼評価が期待できる.
データの不足を補う方法として, CF とコンテンツに基づく方法を統合することによって解消する枠組みも提案されている[Melville 01]. また, 近年, 利用者のプライバシーを守ることや, peers の間で情報を交換することを目的とした分散環境で CF を実現する枠組みが提案されている.

2.2.3 評判システム (Reputation System)

評判システムは, eBay などオンライン取引における“信用と協調”の実現を目指したメカニズムとして注目されている. オンライン取引においては, これまでの (オフラインの取引における) サービスの品質を保証するためのメカニズムでは不十分である.モバイルやユビキタスコンピューティングの普及により, 疎に結合されたコンピュータネットワーク社会 (ソフトエージェント社会や p2p ネットワーク) における“法と秩序”を生み出すメカニズムとしての期待が大きい.

eBay.com をはじめ, slashdot.org (”news for nerds”), affereo.org などが既に評判システムとして影響力を持っており, ますます信用 (Trust) に関する Formal な議論が必要である. Trust 尺度に関する議論は,[Mui 02]に詳しい.

以上のように推薦をはじめとする評判メカニズムの研究は, Web 上の情報を集積し, 信用 (Trust) できる情報を提供するために形式化されたシステムである. 問題点として指摘されているように信頼の評価方法や信頼できる情報にたどり着く方法については未だ手探りの状態である.

3 実験方法

「少年法改正」をキーワードとして Goo から検索された集合を, {賛成, 中立, 反対}に分類する. 各文書を茶筌(ChaSen)を用いて形態素解析し, 名詞と形容詞を分類に用いた

分類に必要な特徴ベクトルは TF/IDF による方法, 人手によって賛成, 反対, 中立を代表する語を抽出する方法の二つを用いている. また, 分類手法は, ボトムアップクラスタリング, EM, および, Adaboost を適用した. 以下, 各方法について説明する.

3.1 ボトムアップクラスタリング法による分類

まず, TFIDFによって文書からN個の特徴ベクトルを作成する. 文書d_jに属する語t_iのTFIDFは次のように求められる.

$$TFIDF(t_i,d_j)=TF(t_i,d_j)\times IDF(t_i)$$
$$TF(t_i,d_j)=\frac{d_j\text{に出現する}t_i\text{の個数}}{d_j\text{の全語の数}} \quad (1)$$
$$IDF(t_i)=\log\left(\frac{\text{文書の総数}}{t_i\text{が出現する文書数}}\right)+1$$

文書がd_1からd_N, 語がt_1からt_Mまである場合, 文書d_jの特徴ベクトルv_iを次のように定める.

$$v_i=(TFIDF(t_1,d_i),TFIDF(t_2,d_i),...,TFIDF(t_M,d_i))^T \quad (i=1\ldots N)$$

ここで語とは, いずれかのドキュメント中に 1 回以上出現する名詞または形容詞を示す. ただし, 活用のある品詞についてはその基本形とした.

TFIDF により作成した N 個の特徴ベクトルを, ボトムアップクラスタリング法でクラスタリングした. アルゴリズムを図1に示す.

1 for $i=1$ to N do
  $c_i \leftarrow \{v_i\}$
end for
2 $C \leftarrow \{c_1,c_2,...,c_N\}$
3 $j \leftarrow N$
4 if $\#C=k$ then exit
5 $x \leftarrow \underset{(c_i,c_j)\in C\times C;i<j}{\arg\min}\ d(c_i,c_j)$
6 $j \leftarrow j+1$
7 $x=(c_a,c_b)$ と書くと, $c_j \leftarrow c_a \cup c_b$
8 $C \leftarrow (C\setminus\{c_a,c_b\})\cup\{c_j\}$
goto 4

図1 ボトムアップクラスタリング法

クラス間の距離がもっとも短いものを融合していき, 最終的に k 個のクラスタを得たところで終了する. 今回の実験では, k=3 とした.

上記アルゴリズムにおいて, クラス間の距離の定義を変え実験を行った. 採用した定義は, クラスの平均ベクトル間のユークリッド距離と, クラスの平均ベクトル間のコサイン類似度がベースの距離である.

μ_i,μ_j をそれぞれ c_i,c_j の平均ベクトルとすると, 平均ベクトルのユークリッド距離は式(2)で表される.

$$d(c_i,c_j)=\|\mu_i-\mu_j\|$$
$$\|a\|=\sqrt{\sum_{k=1}^{M}a_k^2} \quad (2)$$

コサイン類似度ベースの距離は式(3)を用いる.

$$d(c_i, c_j) = 1 - \cos\theta$$

$$\cos\theta = \frac{\mu_i \cdot \mu_j}{\| \mu_i \| \times \| \mu_j \|} \qquad (3)$$

3.2 EM アルゴリズムによる分類

形態素解析結果から名詞と形容詞のみを取り出し出現頻度の高いものから 500 語を選び, 式(1)で計算される TF 値 $d_{i,j}$ を要素とする N×500 行列 D を作る. 総文書数は N とする. 次にデータ行列 D を特異値分解する. $m \times n$ 行列の特異値分解は $D=U\Sigma V^T$ で定義される. U は $m \times r$ 直行行列, V は $n \times r$ 直行行列である. ここで $r = rank(D)$ である. Σ は $r \times r$ 対角行列で, Σ の対角要素が特異値となる. $\Sigma = diag(\sigma_1, \sigma_2, \cdots \sigma_r)$ とした場合, 特異値は $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ を満たす.

ここで U , V を列ベクトルの集合で表現して,

$$U = [u_1 u_2 \ldots u_r]$$

$$V = [v_1 v_2 \ldots v_r]$$

とすると, 文書 i を表現する k 次元ベクトル $(k < r)$ d_i は, $d_i = (\sigma_1 v_{j1}, \sigma_2 v_{j2} \ldots \sigma_k v_{jk})^T$ となる.

以上の方法を用いてデータ行列 D を 50 に縮小した行列を D' とし, 行列 D' に対して EM アルゴリズムを適用し, 分類を試みた.

EM アルゴリズムは適当に選んだパラメータの初期値から始めて, 尤度を最大化するようにパラメータ更新のステップを繰り返す.

各クラスの平均が μ_c, 分散が Σ_c, 各クラスの事前分布が $p(c) = \omega_c$ であるような k 次元正規混合分布に対しては, 以下のような更新式となる.

$$\omega_c^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} p(c \mid x_i; \theta^{(t)})$$

$$\mu_c^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(c \mid x_i; \theta^{(t)})}{\sum_{c=1}^{K} p(c \mid x_i; \theta^{(t)})} x_i$$

$$\Sigma_c^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \frac{p(c \mid x_i; \theta^{(t)})}{\sum_{c=1}^{K} p(c \mid x_i; \theta^{(t)})} (x_i - \mu_c)^T (x_i - \mu_c)$$

ここで $D' = [x_1 x_2 \ldots x_r]^T$ とし, EM アルゴリズムを適用する. 各ベクトルは文書に対応している. クラスの数は3であるとして上記のパラメータ更新の反復を行い, その結果最も帰属度の高いクラスに属するものとした.

3.3 Adaboost を用いた C4.5 による分類

まず, C4.5 を用いて複数の決定木を作成するために, 賛成, 反対を特徴付けていると考えられる語(もしくは言い回し)を属性とし, その語の有無により各文書の特徴ベクトルを生成する. 今回は, トピックを「少年法改正」に絞り, 賛成･反対を特徴付けていると考えられる語として, 以下の 59 語を属性として用いた.

{欠陥, 欠陥がある, 厳しさ, 厳しい, 不十分, 残虐, このままで, 必要, 改正は必要, 改正が必要, 改正に賛成, 改正に賛成です, 改正に取り組む, 改正されなければ, 取り組む, 取り組むべき, 人権, 人権派, 被害者, 被害者無視, 民事訴訟, 訴訟, 賛成, 賛成派, 賛成です, 賛成派です, ようやく改正, 減るのでしょうか, 弊害, 強行, 厳罰, 厳罰より, 厳罰化, 改悪, 反対, 反対の意, 反対です. , 反対の立場です, 改正反対, 法案反対, 「改正」反対, 断固, 声明, 決議, 廃案, 日教組, 日弁連, 署名, 検察官, 不安定, 拘束, 不利益, 証拠, 動機, 暴行, 援助, 環境, 健全, 成長}.

上記の語の有無により各文書の特徴ベクトルを生成し, ランダムに 50 個の特徴ベクトルを選択する. 選択された特徴ベクトルを決定木の学習用の訓練データ X とし, 残りの特徴ベクトルを誤り率(汎化誤差)の測定用のテストデータ V とする.

訓練集合: $\{(x_1, y_1), \ldots, (x_{50}, y_{50})\}$.

テスト集合: $\{(v_1, y_1), \ldots, (v_{50}, y_{50})\}$.

ただし教師信号を $x_i \in X$, $v_i \in V$, $y_i \in Y$={“賛成”, “中立”, “反対”}とする.

ここで t 回目の学習における訓練例題 (x_i, y_i) の重みを $D_t(i)$ としたときの C4.5 への AdaBoost の適用法の具体的な手順を以下に示す.

1. C4.5 により訓練データ X を用いて, 決定木 $h_t: X \rightarrow Y$ $(1 \leqq t \leqq 5)$ を作成.
2. $D_1(i)=1/50$ で初期化.
3. t=1, …, 5 に対して
   - 分布 D_t において以下の誤り率を最小にする決定木 $h_t: X \rightarrow Y$ を選択.
     $$\varepsilon_t = \Pr_{D_t}\{h_t(x_i) \neq y_i\} = \sum_{i: h_t(x_i) \neq y_i} D_t(i)$$
   - 誤り率を用いて信頼度 $\alpha_t \in R$ を計算.
     $$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$
   - 以下の式で分布 Dt を更新.
     $$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

ただし, Z_tは$\sum_i^{50} D_{t+1}(i)$規格化因子

$$\exp(-\alpha_t y_i h_t(x_i)) = \begin{cases} \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} & \text{誤判別された例題} \\ \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} & \text{正解した例題} \end{cases}$$

4. 全ての決定木の出力を信頼度で重みづけし, 多数決をとる. $H(v) = \arg\max_{y \in Y} \sum_{t=1}^{5} \alpha_t I(h_t(v) = y)$

4 実験結果および考察

4.1 ボトムアップクラスタリング

ボトムアップクラスタリングで得られた 3 個のクラスタに対し, それぞれデータ数 98 個, 1 個, 1 個になった. 実験は 2 種類の距離の定義を用いたが, いずれの結果も同じとなった. TFIDF によるベクトル生成とボトムアップクラスタリングでは, 賛否を識別するには全く不十分であるといえる.

4.2 EM アルゴリズム

EM アルゴリズムによって分類されるクラスをクラス 1・クラス 2・クラス 3 としたとき, 賛成・中立・反対のラベル付けを行ったデータセットとの比較の結果は以下のようになる. EM アルゴリズムによる分類のクラスと, データセットの賛成・中立・反対のクラスとの対応は明らかではないため, すべての組み合わせを表1に示す.

表1 EM によるクラスとそのラベルの精度

対応するラベル			適合率(上)・再現率(下)		
クラス 1	クラス 2	クラス 3	クラス1	クラス 2	クラス 3
賛成	中立	反対	0.21	0.37	0.10
			0.57	0.20	0.04
賛成	反対	中立	0.21	0.33	0.70
			0.57	0.32	0.14
中立	賛成	反対	0.51	0.30	0.10
			0.65	0.35	0.04
中立	反対	賛成	0.51	0.33	0.20
			0.65	0.32	0.09
反対	賛成	中立	0.29	0.30	0.70
			0.64	0.35	0.14
反対	中立	賛成	0.29	0.37	0.20
			0.64	0.20	0.09

一見クラス 3 が中立に分類されるとした場合比較的高い性能が得られているように見えるが, 実際にはそうではない. なぜなら EM アルゴリズムによる分類ではクラス3に分類されるのはデータセットの 1 割程度であり, 含まれるデータ数はかなり少ない. 一方データセットに含まれるデータのなかでは中立に分類されているデータは 6 割程度と多いので, この数字から有効な分類が行われたとは言い難い.

考えられる問題として, 特徴ベクトルに用いる語の選択の困難さがある. 100 のサンプルを形態素解析した結果得られた形容詞と名詞はおよそ 2 万 5 千種類であったが, このうち 5000 語程度の語が 10 回以上出現し, 9000語程度の語が5回以上出現している. このため出現頻度で上位から選ばれた 500 語はページの分類の点で適切ではない可能性がある(データセットにおいて出現頻度が上位 500 の語は 180 回以上出現している).

しかし用いる語を増やすと特異値分解に要する計算量が著しく増える上に, あまり多い次元のデータには EM アルゴリズムを適用できないため最終的に利用する次元数は落とさざるを得ない. 頻度で上位から5000語を選んで同様な分類を行った場合にも, 適合率に目立った改善は見られなかった.

4.3 Adaboostを用いた C4.5

C4.5 により訓練集合から作成された, 5つの決定木は以下の通りである.

決定木 1

```
訴訟 = 1: neu (4.0)
訴訟 = 0:
|   厳しい = 1: con (2.0)
|   厳しい = 0:
|   |   反対 = 0:
|   |   |   検察官 = 0:
|   |   |   |   必要 = 0: neu (10.0/4.0)
|   |   |   |   必要 = 1: pro (8.0/4.0)
|   |   |   検察官 = 1:
|   |   |   |   必要 = 0: con (2.0)
|   |   |   |   必要 = 1: pro (2.0)
|   |   反対 = 1:
|   |   |   健全 = 1: neu (2.0)
|   |   |   健全 = 0:
|   |   |   |   援助 = 0: con (7.0/2.0)
|   |   |   |   援助 = 1: neu (2.0)
```

決定木 2

```
訴訟 = 1: neu (6.0/1.2)
訴訟 = 0:
|   反対 = 0:
|   |   検察官 = 0:
```

```
|   |   |   必要 = 0: neu (14.0/6.8)
|   |   |   必要 = 1: pro (10.0/6.5)
|   |   検察官 = 1:
|   |   |   必要 = 0: con (2.0/1.0)
|   |   |   必要 = 1: pro (3.0/2.1)
|   反対 = 1:
|   |   証拠 = 1: con (2.0/1.0)
|   |   証拠 = 0:
|   |   |   健全 = 1: neu (2.0/1.0)
|   |   |   健全 = 0:
|   |   |   |   環境 = 0: con (7.0/3.4)
|   |   |   |   環境 = 1: neu (4.0/2.2)
```

決定木 3

```
訴訟 = 1: neu (6.0/1.2)
訴訟 = 0:
|   反対 = 0:
|   |   検察官 = 0:
|   |   |   動機 = 1: pro (2.0/1.8)
|   |   |   動機 = 0:
|   |   |   |   必要 = 1: pro (9.0/5.5)
|   |   |   |   必要 = 0:
|   |   |   |   |   厳罰 = 1: neu (3.0/1.1)
|   |   |   |   |   厳罰 = 0:
|   |   |   |   |   |   被害者 = 0: neu (6.0/2.3)
|   |   |   |   |   |   被害者 = 1: pro (4.0/2.2)
|   |   検察官 = 1:
|   |   |   必要 = 0: con (2.0/1.0)
|   |   |   必要 = 1: pro (3.0/2.1)
|   反対 = 1:
|   |   証拠 = 1: con (2.0/1.0)
|   |   証拠 = 0:
|   |   |   健全 = 1: neu (2.0/1.0)
|   |   |   健全 = 0:
|   |   |   |   環境 = 0: con (7.0/3.4)
|   |   |   |   環境 = 1: neu (4.0/2.2)
```

決定木 4

```
訴訟 = 1: neu (6.0/1.2)
訴訟 = 0:
|   証拠 = 1: con (4.0/2.2)
|   証拠 = 0:
|   |   反対 = 0:
|   |   |   検察官 = 0:
|   |   |   |   必要 = 0: neu (13.0/5.7)
|   |   |   |   必要 = 1: pro (10.0/6.5)
|   |   |   検察官 = 1:
|   |   |   |   必要 = 0: con (2.0/1.0)
|   |   |   |   必要 = 1: pro (2.0/1.0)
|   |   反対 = 1:
|   |   |   動機 = 1: neu (2.0/1.0)
|   |   |   動機 = 0:
|   |   |   |   検察官 = 1: con (2.0/1.0)
|   |   |   |   検察官 = 0:
|   |   |   |   |   賛成 = 0: con (7.0/4.4)
|   |   |   |   |   賛成 = 1: neu (2.0/1.0)
```

決定木 5

```
訴訟 = 1: neu (6.0/1.2)
訴訟 = 0:
|   反対 = 0:
|   |   検察官 = 0:
|   |   |   必要 = 0: neu (14.0/6.8)
|   |   |   必要 = 1: pro (10.0/6.5)
|   |   検察官 = 1:
|   |   |   必要 = 0: con (2.0/1.0)
|   |   |   必要 = 1: pro (3.0/2.1)
|   反対 = 1:
|   |   証拠 = 1: con (2.0/1.0)
|   |   証拠 = 0:
|   |   |   動機 = 1: neu (2.0/1.0)
|   |   |   動機 = 0:
|   |   |   |   検察官 = 1: con (2.0/1.0)
|   |   |   |   検察官 = 0:ppp
|   |   |   |   |   賛成 = 0: con (7.0/4.4)
|   |   |   |   |   賛成 = 1: neu (2.0/1.0)
```

表2に，決定木を単独で用いた場合とAdaboostを適用した場合の性能比較結果を示した．

表2　各決定木誤差とAdaboostによる誤差

	訓練誤差	汎化誤差
決定木 1	0.28	0.54
決定木 2	0.28	0.52
決定木 3	0.22	0.58
決定木 4	0.26	0.50
決定木 5	0.28	0.48
H	0.22	0.58

表2から，Adaboostを用いると訓練誤差は小さくなるが，汎化誤差が大きくなるという結果となり，過学習が起こっている可能性がある．しかし，表3に示すように，実際にブースティングの回数より，どれだけ性能が変化しているのかを確認すると，1 回目以降性能に変化が無いことが分かる．

表3　ブースティング回数による誤差変化

	訓練誤差	汎化誤差
1 回目	0.22	0.58
2 回目	0.22	0.58
3 回目	0.22	0.58
4 回目	0.22	0.58
5 回目	0.22	0.58

これは，表4に示すように2回目以降選択された決定木の信頼度が極めて小さく，ブースティングの最終結果に全く影響を与えなかったためと考えられる．

表4 決定木の信頼度

選択順序	決定木	信頼度
1番目	決定木3	0.633
2番目	決定木4	0.213
3番目	決定木1	-0.0099
4番目	決定木5	-0.0668
5番目	決定木2	-0.1021

全ての決定木の判別結果にあまり差異がないため，信頼度が低い．以上の結果を特徴ベクトルの是非と，アルゴリズムの観点から考察する．

特徴ベクトル

各決定木で用いられる特徴語のうち，異なる語の数は13であった．つまり，決定木で利用されている語が集中しているため，決定木間での判別結果の差が出ないと考えることができる．

5個の決定木で現れた特徴語 {訴訟, 反対, 検察官, 必要}
3個の決定木で現れた特徴語 {健全, 証拠, 同期}
2個の決定木で現れた特徴語 {環境, 賛成}
1個の決定木で現れた特徴語 {厳しい, 援助, 厳罰, 被害者}

アルゴリズム

決定木生成が静的であるため，更新された例題の重みのもとで予め生成された決定木から誤り率を最小にするものを選択するだけになっている．本来であれば，更新された例題の重みのもとで，誤り率を最小とする決定木を学習する必要がある．

5 今後の課題

本稿では，社会的トピックを扱った文書のセマンティクス抽出における統計的機械学習の有用性を確かめるための実験結果を紹介した．テキストマイニングで最もよく利用されているアルゴリズムから3つを選んだが，いずれも，単純なTF/IDFから抽出した特徴からは，単純な分類でさえ出来ない状態である．これらの方法を有用にするためには，特徴ベクトル抽出が鍵である．共起関係や係り受けなど，文書構造を解析した意味の抽出などが現段階では一般的であるが，これに関しても本稿で取り上げた社会的トピックに対しては精度があまり期待できない状況である．アルゴリズムでは，ブースティングによるアンサンブル学習がこのような分野において適していると思われる．ブースティングの応用に際してマルチクラスに拡張したAdaboostも提案されている[Schapire 00]が，クラス数の問題とは別に，ブースティングによる学習効果と汎化に関する議論が必要である．C4.5はノイズに頑健であるが，ブースティングによる精度の改善がノイズへの頑健性に与える影響についても調べる必要がある．

参考文献

[Dhyani et.al] Dhyani, D., Ng, W.K., and Bhowmic,S.S.: A Survey of Web Metrics, ACM Computing Surveys, Vol.34, No.4, December 2002, pp.469-503 (2002).

[Gil 02] Gil, Y. and Ratnakar, V. : Trusting Information Sources One Citizen at a Time, Proceedings of the First International Semantic Web Conference, pp. 162-176 (2002).

[Goldberg 92] D. Goldberg, D. Nichols, B. M. Oki, D. Terry “Using collaborative filtering to weave an information tapestry.” *Communication of the ACM, 35*(12), pp. 61-70 (1992).

[Melville 01] P. Melville, R.J. Mooney and R.Nagarajan, “Content-Boosted Collaborative Filtering,” *Proc. Of the SIGIR-2001 Workshop on Recommender Systems* (2002).

[Mui, 02] L. Mui, M. Mohtashemi, A. Halberstadt , “A Computational Model of Trust and Reputation,” *Proc. 34th Hawaii International Conference on System Sciences* (2002).

[Resnick 97] P.Resnick and H.R.Varian, “Recommender systems. “ *Communications of the ACM, 40(3):* pp.56-58 (1997).

[Resnick and Zeckhauser 00] P. Resnick, R. Zeckhauser, “Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay’s Reputatoin System,” *Working Paper for the NBER Workshop on Empirical Studies of Electronic Commerce* (2002).

[Schafer 99] J.B. Schafer, J.Konstan, and J. Riedl, “Recommender systems in e-commerce.”, *In Proceeding of the ACM Conference on Electronic Commerce, Pittsburgh PA.* (1999).

[Schapire 00] Schapire, R.E. and Singer, Y., BoosTexter : A Boosting-based System for Text Categorization", *Machine Learning,* Vol. 39, Number 2/3, pp.135-168 (2000).