6T−06

# オークション支援システム MultiHammer における naive Bayes に基づく HTML 文書の分類について

山田 亮太　大囿 忠親　新谷 虎松
名古屋工業大学 知能情報システム学科
e-mail: {ryota,ozono,tora}@ics.nitech.ac.jp

## 1 はじめに

我々はオンラインオークションにおける効果的な取引の実現を支援することを目的として，情報収集エージェントを利用したオンラインオークション支援システム MultiHammer を開発してきた [1]．MultiHammer では情報収集エージェントが HTML 文書の処理を行う．HTML 文書の処理を行う際には，取得した HTML 文書が処理の対象として適切なものであるか否かを判別する必要がある．この判別を行うためには，処理の対象となる HTML 文書が適切な HTML 文書の集合に属するか否かを分類すれば良い．

また，情報収集エージェントは HTML 文書の記述規則を表す情報抽出パターンを用いて情報の構造化と抽出を行う．情報抽出パターンの作成には HTML 文書を記述規則ごとに分類する作業が必要となる．

本稿では，情報収集エージェントによる naive Bayes[2] に基づく HTML 文書の分類機構を提案する．

## 2 MultiHammer と HTML 文書の分類

一般に，オークションサイトでは異なる種類の情報を提示する HTML 文書や，記述規則の異なる HTML 文書が混在する．例えば，複数の財の概要を一覧として与える HTML 文書と，1 つの財の詳細な情報を与える HTML 文書とでは，提示する情報の種類や記述規則が異なっている．

情報収集エージェントはオークションサイトで提供される HTML 文書を取得して処理を行う．情報収集エージェントが HTML 文書を処理する場合，処理対象となる HTML 文書が提示する情報の種類ごとに処理の内容は異なる．HTML 文書の取得には予め与えられた URL から直接取得する場合と，直接取得した HTML 文書からハイパーリンクを辿って間接的に取得する場合がある．予め与えられた URL から HTML 文書を直接取得する場合，利用者により適切な種類の情報を提示する HTML 文書の URL が与えられていることが期待できる．一方，HTML 文書を間接的に取得する場合，ハイパーリンクの構造次第で異なる種類の情報を提示する HTML 文書が取得される場合がある．情報収集エージェントに URL を与える作業自体は困難な作業ではない．しかし，与える URL の数が増加するにしたがって，利用者の負担は増加する．利用者の負担軽減のため，情報収集エージェントが可能な限り自律的に HTML 文書により提示される情報の種類の判別を行うことが望ましい．

MultiHammer では情報収集エージェントはパターンマッチングを用いて情報の構造化と収集を行う．パターンマッチングには HTML 文書の記述規則を表すパターンを用いる．我々は共通の記述規則に基づく複数の HTML 文書を比較してパターンを抽出する手法を提案してきた [1]．パターンの抽出を行うためには，HTML 文書の集合を同じ記述規則に従う HTML 文書の集合に予め分類する必要がある．

Classifying HTML Documents using Naive Bayes Classifier for the Online Auction Support System MultiHammer
Ryota Yamada, Tadachika Ozono, Toramatsu Shintani
Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya, 466-8555

## 3 naive Bayes による HTML 文書の分類

### 3.1 naive Bayes

naive Bayes は Bayes 理論に基づく分類器である．naive Bayes は，個々のインスタンス $x$ を互いに独立な属性の連言として表現可能で，目標関数 $f(x)$ が有限の集合 $V$ から任意の値をとるといった状況で利用される．naive Bayes は，予め与えられる訓練データの集合から条件付き確率に関する学習を行い，互いに独立な属性の連言 $\langle a_1, a_2, ..., a_n \rangle$ として表現される新しいインスタンス $x$ について，$f(x)$ の値を予測する．naive Bayes により予測される値 $v_{NB}$ は次の式で得られる．

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

本研究で提案する分類機構の場合，インスタンス $x$ は “HTML 文書” に，目標関数 $f(x)$ のとる値は “HTML 文書が提示する情報の種類” または “HTML 文書の記述規則” に，それぞれ相当する．

### 3.2 HTML 文書の属性集合化

naive Bayes を利用する際，インスタンスをどのような属性の集合として表現するかが重要となる．

一般に，Web サイトにおいて情報が HTML 文書として提示されるのは，人間による情報の理解を助けるためである．通常，人間は HTML 文書を閲覧する際，Web ブラウザを利用する．Web ブラウザを通して HTML 文書を閲覧する場合，人間の目に映るのは，整形された “HTML タグ以外の文字列” である[1]．

HTML 文書に含まれる “HTML タグ以外の文字列” を観察すると，閲覧者に伝えるべき “内容” と，“内容” の理解を助けるための “案内” の 2 つに大別できる．例えば，“1,000 円” という “内容” に “最高入札価格” や “落札価格” などの “案内” が付加され，閲覧者の理解を助ける．提示する情報の種類が異なる HTML 文書では，異なった “内容” の集合を提示する．したがって，登場する “案内” の集合も異なるものとなる．また，同じ種類の情報を提示する HTML 文書では，閲覧者の混乱を防ぐため，共通の “案内” と記述規則を用いて “内容” が提示されることが多い．

以上の考察より，本研究では，“案内” に相当する文字列の集合を選出し，インスタンスである HTML 文書中に各 “案内” が存在するか否かを属性とする．

[1] 画像などの “埋め込まれる” オブジェクトについては考えない．

“案内” に相当する文字列は，次に示す (1) から (4) の手順で訓練データ集合から自動的に選出する．

(1) 訓練データ集合 $T = \{t_1, ..., t_n\}$ に含まれる全ての HTML 文書から “HTML タグ以外の文字列” を選出し，それぞれ出現頻度を計算．(2) 選出した “HTML タグ以外の文字列” の集合 $E_0$ から出現頻度の低いものを除外した集合 $E_1$ を作成．(3)$E_1$ から $n$ 回以上登場したものを除外した集合 $E_2$ を作成．(4)$E_2$ から予め用意した不要語リストに含まれるものを除外した集合を “案内” に相当する文字列の集合として選出．

(2) は，“内容” を除外するための処理である．提示される “内容” は HTML 文書ごとに異なるため，出現頻度は低くなると考えられる．(3) では，全ての HTML 文書に共通して含まれる “案内” を HTML 文書の分類に役立たないものと見做し，それらを除外している．

### 3.3 HTML 文書の分類プロセス

まず，情報収集エージェントは次の (i) から (v) のプロセスを経て HTML 文書の分類のための学習を行う．

(i) 利用者から “HTML 文書 $d$ の URL$u_d$” と “HTML 文書 $d$ が提示する情報の種類または HTML 文書 $d$ の記述規則 $c_d$” の対 $\langle u_d, c_d \rangle$ の集合 $T = \{\langle u_1, c_1 \rangle, ..., \langle u_n, c_n \rangle\}$ を訓練データ集合として受け取る．(ii) 各 $u_i$ で指定される HTML 文書 $i$ を取得する $(1 \leq i \leq n)$．(iii)HTML 文書の集合から “案内” に相当する文字列の集合 $G = \{g_1, ..., g_m\}$ を選出する (3.2 参照)．(iv) 各 $i$ について $G$ の各要素 $g_1, ..., g_m$ を含むか否かを検査し，属性の集合 $A_i = \{a_{i1}, ..., a_{im}\}$ を得る $(1 \leq i \leq n)$．(v) 各 $i$ についての $c_i$ と $A_i$ の対 $\langle c_i, A_i \rangle$ の集合 $S$ を用いて naive Bayes で条件付き確率の学習を行う $(1 \leq i \leq n)$．

学習を完了したエージェントは，新しく HTML 文書 $x$ を取得するごとに次の (vi) から (vii) のプロセスを経て HTML 文書の分類を行う．

(vi)$x$ について $G$ の各要素 $g_1, ..., g_m$ を含むか否かを検査し，属性の集合 $A_x = \{a_{x1}, ..., a_{xm}\}$ を得る．(vii)$A_x$ を naive Bayes に入力し，分類結果 $v_{NB}$ を得る．

## 4 実験

オークションサイトでは，複数の財の情報を一覧するための HTML 文書 (一覧ページ) と，1 つの財の情報を詳細に記した HTML 文書 (詳細ページ) が提供される．情報エージェントはそれぞれの HTML 文書に対して異なる処理を行うため，両者を分類する必要がある．本稿で提案した HTML 文書の分類機構の性能を評価するために，eBay Japan[2]，Auction.excite[3]，そして，Yahoo! Auction[4] から一覧ページと詳細ページを収集し，分類の実験を行った．

オークションサイトでは，一覧ページから一覧ページに含まれる財の詳細ページへとリンクが張られている．本稿では，1 つの一覧ページと，その一覧ページから直接リンクされている詳細ページ全てを一つのまとまりとして，“1 セット” とする．

訓練データとしては，それぞれのオークションサイトから無作為に一覧ページを 10 件選択し，各 10 セット，合計で 30 セット 907 件の HTML 文書を用意した．訓練データの内訳は表 1 に示す通りである．

[2]http://www.eBayJapan.co.jp/
[3]http://auctions.excite.co.jp/
[4]http://auctions.yahoo.co.jp/

表 1: 訓練データの内訳

| | 一覧ページ | 詳細ページ |
|---|---|---|
| eBay Japan | 10 件 | 458 件 |
| Auction.excite | 10 件 | 296 件 |
| Yahoo! Auctions | 10 件 | 123 件 |
| 合計 | 30 件 | 877 件 |

表 1 に示した訓練データにより学習を行った後，それぞれのオークションサイトから訓練データとして用いていない一覧ページを無作為に 10 件選択し，各 10 セット，合計で 30 セット 971 件の HTML 文書の分類実験を行った．実験の結果を表 2 に示す．

表 2: 一覧ページと詳細ページの分類実験の結果

| | 入力数 | 正解数 | 正解率 |
|---|---|---|---|
| eBay Japan 一覧 | 10 | 10 | 1.00 |
| Auction.excite 一覧 | 10 | 10 | 1.00 |
| Yahoo! Auctions 一覧 | 10 | 10 | 1.00 |
| eBay Japan 詳細 | 487 | 487 | 1.00 |
| Auction.excite 詳細 | 300 | 300 | 1.00 |
| Yahoo! Auctions 詳細 | 154 | 154 | 1.00 |
| 全体 | 971 | 971 | 1.00 |

表 2 に示した通り，提案した HTML 文書の分類機構により用意した 971 件の HTML 文書を全て正しく分類することができた．これは，オークションサイトごとに “案内” として用いる語が大きく異なることと，各サイトにおいて一覧ページに含まれず詳細ページにのみ含まれる “案内” が多数存在することによる．

情報収集エージェントのための HTML 文書記述規則パターンの抽出を行う場合，分類対象間で，より多くの “案内” が共通することが想定される．共通する “案内” が多数存在する状況での性能については，今後，さらなる検証を行う必要がある．

## 5 まとめ

本稿では，オークション支援システム MultiHammer における情報収集エージェントのための HTML 文書の分類機構を提案した．提案した機構では，HTML 文書から情報の内容を説明する “案内” となる要素を選出し，“案内” の有無に基づいて HTML 文書を属性の集合に変換した上で，naive Bayes による分類を行った．

提案した機構について実験を行ったところ，良好な結果を得ることができた．今後，異なる分類対象での実験を行い，より詳細な評価を行う予定である．

## 参考文献

[1] R. Yamada, H. Hattori, T. Ito, T. Ozono and T. Shintani, “MultiHammer: A Virtual Auction System based on Information Agents,” In the Proc. of the Pacific Asian Conference on Intelligent Systems 2001 (PAIS2001), pp.73–77, 2001.

[2] Tom M. Mitchell, “Bayesian Learning,” Machine Learning, pp.154–200, McGraw-Hill, 1997.