社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (1)
2003－ＦＩ－ 71 (1)
2003／5／22

ネットオークションを対象とした テキストマイニングエージェント「NTM-Agent」の評価

楠村幸貴[†] 土方嘉徳[†] 西田正吾[†]

近年，ネットオークションが盛んである．しかし，ネットオークションには大量の商品が存在しており，ユーザがその中から一つの商品を選択することは困難である．この問題に対し，本研究では商品の比較が容易になるようユーザに代わり商品の比較表を生成するエージェント（NTM-Agent:Net auction TextMing Agent）を構築した．NTM-Agent はネットオークションサイトで商品の検索し，ユーザの検索要求に適合しない商品をフィルタリングし，商品の特徴について説明してい文章からその特徴に関する情報を抽出して，商品の比較表を作成する．本稿では NTM-Agent の評価について述べる．

Evaluation of Text Mining Agent for Net Auction

Yukitaka Kusumura,[†] Yoshinori Hijikata[†] and Shogo Nishida [†]

Net auctions have been widely utilized with the recent development of the Internet. However, it is a problem that there are too many items for bidders to select the most suitable one. We aim at supporting the bidders on net auctions by automatically generating a table which contains the features of several items for comparison. We construct a system called NTM-Agent(Net auction Text Mining Agent). The system collects the web pages of items and extracts the items' features from the pages. After that, it generates the table which contains the extracted features. After the implementation of NTM-Agent, this research evaluated the system.

1. はじめに

近年，ネットオークションが盛んであり，非常に大量の商品が毎日出品されている．通常ネットオークションでは，ユーザはキーワードなどで商品を検索し，得られた複数の商品について出品者によって記述された商品の紹介文を読み，それらの商品を比較して入札する商品を決定する．しかし，商品の数が多くなってくるとこの作業は大変なものになる．

本研究ではこの問題に対し，商品の比較表を自動生成するエージェント（NTM-Agent:Net auction TextMing Agent）を提案した[1)2)]．NTM-Agent はユーザの検索要求に適合する商品の紹介ページを収集し，商品の特徴について説明している文章 (以下，商品紹介文) からその特徴に関する情報を抽出し，それらを用いて商品の比較表を生成する．

本研究では実用性の高い支援システムを目指すこととし，商品ごとに目的の情報を精度良く抽出できるようにするため，商品に関するドメイン知識☆1を用いる．具体的には，商品の特徴の属性を示すキーワード (以下，属性名) ☆2をドメイン知識として用い，対応する値 (以下，属性値) の抽出を行う．

このようにしてネットオークション上で商品の情報を収集し，紹介文から情報を抽出するには次の問題がある．

問題 1 出品商品の分類が不均一である
出品者が自由にタイトル☆3を付けて分類を行うため，目的の商品と異なる商品 (以下，ノイズ商品) ☆4が検索結果に混じる．

問題 2 紹介文の記述が不均一である
出品者によっては記述内容に属性名の省略がある (日本語の特性として主語が省略されがちであることがこの一因となっている)．また，出品者が

† 大阪大学大学院基礎工学研究科
Graduate School of Engineering Science, Osaka University

☆1 システムに事前に与えられる，特定の分野と問題などに関する知識
☆2 パソコンの場合,「CPU」や「メモリ」などのキーワード
☆3 ネットオークションでは出品者が商品に名前を付ける．その名前はその商品の紹介ページでタイトルとして表示される．
☆4 パソコンを検索した場合に検索結果に含まれるメモリ，キーボードなどの商品

自由に紹介文を記述するため，レイアウトに表，箇条書き，文章といった複数の記述パターンが混在する．

これらの問題に対して，次の解決方法を用いる．

解決策 1　タイトルと商品紹介文中のキーワードについての相関ルールでフィルタリングを行う．本研究ではマーケットバスケット分析を用いて相関ルールを生成する支援ツールを作成する．

解決策 2　属性名の抜けに対しては，属性名とその値の記述に関してその対応の簡単な記述例から学習を行う．学習後，属性名が書かれておらず，学習された属性値のみ書かれているテキストを発見すれば，その属性値を抽出する．不均一なレイアウトに対しては，表か箇条書きか文章かを判断して，その記述形態に最も適した方法で情報抽出を行う．

本稿では 2 章で関連研究について述べ，本研究との違いを明確にする．3 章では NTM-Agent の大まかな処理の流れと，NTM-Agent で使用するドメイン知識について述べる．4 章ではノイズ商品のフィルタリング方法について述べる．5 章では紹介文から商品の特徴を抽出する方法について述べる．そしてそれらを実装したシステムの構成について 6 章で述べ，7 章において実装したシステムの評価について述べる．最後に結論を 8 章で述べる．

2. 関 連 研 究

電子商取引の分野における支援サイトと研究について 2.1 節で述べる．そして，それらと本研究との違いについて 2.2 節で述べる．

2.1 従来のシステム

大量の商品に関する情報を整理して提供する支援サイトには価格.com[3)]，Libra[4)]，Bestlot.com[5)] が，研究例には Biddingbot[6)] と Shopbot[7)] がある．

価格.com と Libra は自サイト内のデータベースから商品を検索し，そのリストをユーザに提示する．出店者は定期的に商品とその特徴を検索サイトに提供しなければならない．

Bestlot.com と Biddingbot はユーザの検索要求を複数のオークションサイトに送信し，検索結果をまとめて表示してくれる．ただし，抽出する情報は商品名と価格のみである．

shopbot はオンラインショップの商品を Web 上から検索し商品の説明を表示する．オンラインショップの商品ページは一定の記述とレイアウトであることが多いため，shopbot はあらかじめ与えられた属性名の例を用いて商品ページ中にその属性名が記述される位置を学習し，抽出用のテンプレートを生成する．そしてそのテンプレートを用いて新たな商品に対して属性名の抽出を行う．

表 1　従来の Web 上のシステムとの違い

	Web ページの自動収集	情報の抽出	不均一なテキストの解析
価格.com Libra	×	×	×
Bestlot.com Biddingbot	○	△ (価格のみ)	×
Shopbot	○	○	×
NTM-Agent	○	○	○

2.2 本研究との比較

上述のシステムと本研究との違いをまとめると，表 1 のようになる．NTM-Agent は価格.com，Libra と異なり，自動的に Web 上から情報収集を行う．さらに，Bestlot.com，Biddingbot と異なり，価格のみでなく，製品の性能や状態に関する情報までを収集する．Shopbot は決まった記述とレイアウトを持つオンラインショップのページからしか情報抽出できないが，NTM-Agent は記述とレイアウトが不均一な商品の紹介文に対して文章の解析を行う．

3. NTM-Agent システムの概要

ユーザ側から見た NTM-Agent システムの大まかな処理の流れについて 3.1 節で述べる．また，NTM-Agent は処理を行う際にドメイン知識を用いている．このドメイン知識について 3.2 節で述べる．

3.1 処理の流れ

NTM-Agent の処理の流れを以下に示す (図 1 参照)．

(1) ユーザがシステムに検索キーワードまたはネットオークション内の検索結果のページ (以下，検索結果ページ) の URL を入力する．また，これと共に目的の商品のカテゴリ（「ノートパソコン」や「自動車」などのことで，3 のノイズの除去と 4 の情報抽出の際に用いる.）を入力する．

(2) キーワードまたは検索結果ページの URL をオークションのサイトに送信し，出品商品の検索結果を得る．

(3) 検索結果の中から，ノイズとなる商品を除く．

(4) 残った商品の紹介文から情報抽出を行い，比較表を作成する．

3.2 ドメイン知識

NTM-Agent は次のドメイン知識を用いて処理を行う．

探索用ドメイン知識　オークションサイトのリンク構

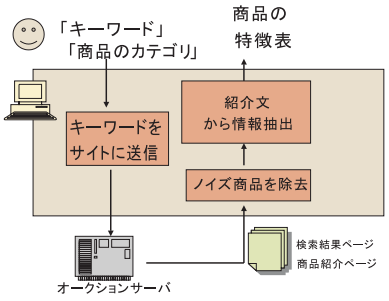

図 1 NTM-Agent の処理の流れ

造，オークションサイトの検索クエリーのテンプレート，検索結果ページと商品紹介ページのテンプレート

抽出用ドメイン知識 商品の特徴を示すキーワード (属性名)

探索用ドメイン知識はオークションサイトごとに作成され，NTM-Agent がオークションのサイト内を探索して必要な商品ページを取得し，その中から商品紹介文部分，タイトル部分，価格部分，入札期限部分を抽出するために用いられる．

抽出用ドメイン知識は商品のカテゴリごとに作成され，NTM-Agent がそれらのキーワードを検索し，属性値がどこに記述されているかを探すための手がかりとして用られる．抽出用ドメイン知識には抽出したい属性の属性名の類義語を記述しておき，抽出の際記述されたキーワードを検索して，対応する属性値が記述されている文や行を取得する．

4. ノイズ商品のフィルタリング

ノイズ商品のフィルタリング方法について 4.1 節で述べる．その際用いるキーワードを用いたルールの生成について 4.2 節で述べる．

4.1 フィルタリング方法

本研究では正解商品と相関の強いキーワード (A) とノイズ商品と相関の強いキーワード (B) をそれぞれ事前に登録しておき，商品のタイトルと商品紹介文に対して次のような 2 種類のルールを用いる．

正解商品用のフィルタリングルール 「A が含まれているならば，正解商品」

ノイズ商品用のフィルタリングルール 「B が含まれているならば，ノイズ商品」

これらのルールにより，商品の集合に対して次の 2 種類のフィルタリング方法を切り替えながら用いる．

正解商品用のフィルタリング 正解商品用のフィルタリングルールのみを用いて，正解商品を取り出しそれ以外の商品をすべて削除する．この方法は判定できない商品をすべて削除するので，ノイズ商品を多く削除できるが正解商品を誤って削除してしまう可能性が高い．

ノイズ商品用のフィルタリング ノイズ商品用のフィルタリングルールのみを用いて，ノイズ商品と判定された商品のみを削除する．この方法は判定できない商品を残すため，削除できるノイズ商品は少ないが誤って正解商品を削除してしまう可能性は低い．

この 2 種類のフィルタリング方法の切り替えは検索結果の商品の数によって行う．つまり，検索結果の商品の数が閾値（100 件）以上存在する場合正解商品用のフィルタリングを行い，商品の数が閾値未満の場合はノイズ商品用のフィルタリングを行う．これは次の考察によるものである．商品の数が小さい場合はノイズ商品が含まれていてもユーザにとって問題にはならず，ノイズ商品が残っていてもより多くの商品を提示することが好ましい．反対に商品の数が大きい場合はノイズ商品の数も大きくなり似通った商品も複数含まれることから，すべての正解商品を取り出すことよりもノイズ商品の数を減らすことを優先させるべきである．

4.2 相関ルールの生成

商品のタイトルと商品紹介文のキーワードチェックで用いるルールは事前に目的の商品のカテゴリごとに用意しておかなければならない．しかし，商品のカテゴリごとにこれらのルールを発見することは困難である．本研究ではシステムに相関ルールの生成を支援するツールを付加することでこの問題に対処する．本ツールは，商品の特徴表を提供するサービスプロバイダや，ディフォルトのルールをカスタマイズするユーザ向けのものである．本節 (4.2 節) ではこれらの人々をユーザと呼ぶことにする．本ツールは特定の種類の商品に依存しない，汎用的なツールとして作成する．相関ルールの抽出には最も基本的なアルゴリズムであるマーケットバスケット分析[9] を用いる．相関ルールの生成支援の流れを次に述べる．ユーザがルールを追加したい商品のカテゴリを含む検索結果が返されたときに，システムはその検索結果を元に教師信号入力インターフェース (図 2 参照) を表示する．ユーザはインターフェース上の商品のタイトルや Web 上の紹介ページを参考に，それぞれの商品が正解商品とノイズ商品どちらであるかを選択する．システムはユーザの入力を教師信号とし，マーケットバスケット分析によりタイトル中のキーワードと商品紹介文中のキーワードそれぞれについて信号と共起しているものを学習し，

図 2 教師信号入力インターフェース

- 表
  - <TABLE>タグ中
    - <TR> 「~属性名~ 」<TD>「~ ○○○ ~」
    - <TR> 「~属性名~ 」<TD> 「~ ○○○ ~」
- 箇条書き
  - 「~属性名~○○○~」
 or HTML中の改行
  - 「~属性名~ ○○○ ~」
 or HTML中の改行
- 文章
  - その他

図 3 表，箇条書き，文章の定義

ルールとして出力する．出力されたルールは GUI 上に表示されるため，ユーザは最後にそれらのルールを確認し編集することが可能である．

5. 紹介文からの情報抽出

本研究では記述形式が異なる紹介文に対し，記述形式を判別してそれぞれに適した抽出を行う．また属性名の記述が無い紹介文については，属性名の記述のある紹介文からの抽出の際に属性値のキーワードについて学習を行い，学習したキーワードを用いて抽出を行う．記述形式ごとの抽出ついては 5.1 節で詳しく述べる．属性名の記述抜けのための学習については 5.2 節で詳しく述べる．

5.1 記述形式ごとの抽出

表，箇条書き，文章を図 3 のように定義して，それぞれに適した抽出を行う．

この定義を基に表は<TR>タグと<TD>タグごと，箇条書きは
タグと HTML 中の改行ごと，文章は「,」「.」「/」などの区切り記号ごとにテキストを区切る．区切られた文の中から抽出用ドメイン知識を用いて属性値が含まれる文を特定し，形態素解析を行い，数値や固有名詞を☆を優先して名詞を抽出する．ただし，文章に対して名詞が存在しない場合,「ありません。」や「きれいです。」などの述語の記述を抽出するために文末の用言を優先して抽出する．

☆ 形態素解析システムの辞書から判別する

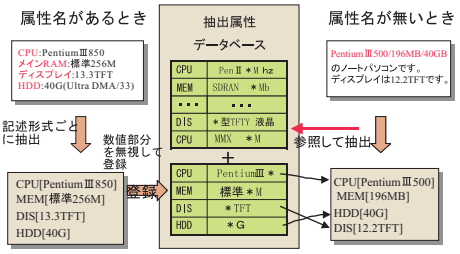

図 4 属性値の学習

5.2 属性名の記述抜けのための学習

属性名の記述が抜けていると，抽出する属性値が記述されている部分を属性名のキーワードを用いて検索できない．そこで，属性名と属性値の組み合わせについて学習しておき，抽出の際に学習した属性値を検索してそれを抽出する．

図 4 に属性値の学習とそれを用いた抽出方法を示す．属性名が記述されている場合，5.1 節で述べた方法で属性値を抽出する．この際，抽出した属性値を属性名と組み合わせて，システム内のデータベース（抽出属性データベース）に保存しておく．このとき，数字部分については記号「＊」に変換しておき，どのような数値でも対応できるようにしておく．商品紹介文中に属性名が記述されていない場合，抽出属性データベースを参照し，そこに保存してあるキーワードと商品紹介文中のキーワードのマッチングを行い，抽出する属性値を特定する．

6. 実装したシステム

本研究は現実のネットオークションでユーザを支援するシステムを目指した．そこで設計にあたり，ユーザのマシンの負担を少なくするべきと考え，システムを NTM-Client と NTM-Server に分け，サーバクライアント方式で実装した．図 5 にシステムの構成を示し，以下に処理の流れを示す．

- ： ユーザが NTM-Agent の Web ページ中のアプレットに欲しい商品に関する検索キーワード（または検索結果のページの URL）と商品のカテゴリを入力する．NTM-Client は受け取ったキーワードまたは URL を NTM-Server に HTTP で送信する．
- ： NTM-Server は NTM-Client からのリクエストを受け付け，探索モジュールへ送る．
- ： 探索モジュールはオークションサイトのリンク構造，HTML の構造を記述したサイト探索用ドメイン知識を参照してオークションサイトにアクセスし，検索結果のページを得る．

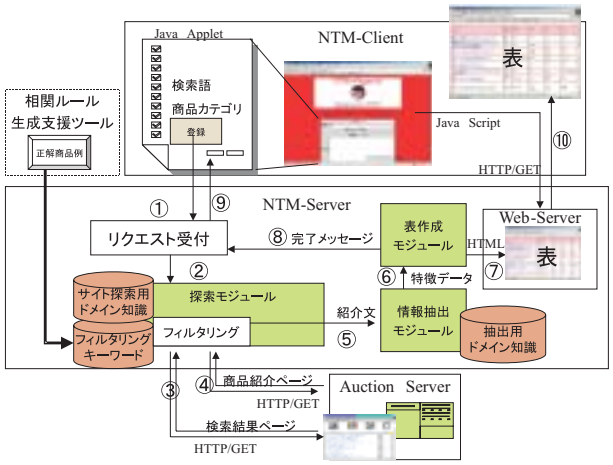

図 5 システムの構成

: フィルタリング用のキーワードを用いて検索結果のページ中の各商品のタイトルについてチェックを行い，ノイズ商品を省く．その後，残った商品について，それぞれの商品紹介ページをサイトから収集する．

: 集めた商品紹介ページからテンプレート (サイト探索用のドメイン知識に記述されている) を用いて商品紹介文を取り出し，フィルタリング用のキーワードを用いてチェックを行い，ノイズ商品を除く．その後，商品紹介文を抽出モジュールに送る．

: 商品に対する抽出用ドメイン知識を参照し，商品の特徴値を抽出する．

: 抽出した商品の特徴値を元に商品の特徴表を HTML ファイルとして作成する．

: システムの処理が終了したことをリクエスト受付モジュールに伝える．

: 処理の終了をユーザマシン上のアプレットに送信する．

: ユーザが NTM-Agent の Web ページ上のボタンをクリックすると出力の Web ページが表示される．

7. 評　　価

NTM-Agent を数量的に評価するため，NTM-Agent が用いた手法について評価を行った．ノイズ商品のフィルタリングについて 7.1 節で述べ，商品紹介文からの情報抽出について 7.2 節で述べる．

7.1 ノイズ商品のフィルタリングについての評価

支援ツールによって生成された相関ルールがどの程度有効であるかどうかを評価するため，フィルタリングの精度と再現率を調べた．フィルタリングの精度と再現率は次の等式によって計算される．

- フィルタリング精度 $P_f = | B | / | A |$
- フィルタリング再現率 $R_f = | B | / | C |$

ただし数式中の A，B，C は次の集合を意味する．

A： フィルタリングによって取得できた商品の集合
B： A のうちの正解商品の集合
C： 検索結果中のすべての正解商品の集合

これらの指標の結果を比較するため，ランダムにフィルタリングを行った場合の精度と再現率を計算した．この精度はすべての商品のうちの正解商品の割合であり，再現率は NTM-Agent と同じ数だけ商品を取得したと仮定したときの，すべての正解商品のうちの獲得できた正解商品の割合である．

我々のフィルタリング手法は検索結果に含まれる商品の数によって用いる相関ルールを変更するものである．ノイズ商品のフィルタリングルールに対しては 30 件の商品で実験を行い，正解商品用のフィルタリングルールに対しては 100 件の商品で実験を行った．実験の手順は次の通りである．

(1) 支援ルールを用い相関ルールを生成した．ルールの詳細は次の通りである．
　カテゴリ: パソコン，自動車，ベビー服
　教師信号: それぞれ 100 件☆
　ルールの数: パソコン 20 個, 自動車 7 個，ベビー服 8 個
(2) 表 2 に示す検索語で商品の検索を行う．
(3) ノイズ商品のフィルタリングを行い，フィルタリングの精度と再現率を計算する．

図 6-(a) に商品の数は少ない場合のフィルタリングの精度と再現率を示す．図 6-(b) に商品の数が多い場合のフィルタリングの精度と再現率を示す．ベビー服についてのみ着目すると，精度と再現率共にランダムにフィルタリングを行った場合とほとんど変化が無い．この原因は一般的にベビー服のノイズ商品の数が小さいためである (表 3 参照)．ノイズ商品の少ないカテゴリの商品に対しては，支援ツールが相関ルールを学習する際，教師信号中にも同様の偏りがあるため正しいルールを学習できない．そのため，これらの商品には本フィルタリング手法は有効でない．

次にパソコンと自動車について考察する．まずノイズ商品用のフィルタリングについての精度と再現率に

☆ パソコン，自動車，ベビー服に対してそれぞれ「thinkpad」「SOTEC」「Pen4」,「トヨタ」「ボルボ」「セダン」,「mikihouse」「おむつカバー」「ベビー服」を検索語として収集した商品から学習を行った．

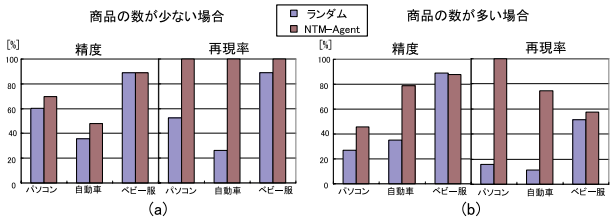

図 6 フィルタリングの精度 P_f と再現率 R_f

商品のカテゴリ	ノイズ商品の割合（200 件中）
パソコン	65 % (130 件)
自動車	53 % (106 件)
ベビー服	8 % (16 件)

表 3 ノイズ商品の割合

着目する．これは検索結果の商品が少ない場合に用いる方法であり，精度よりも再現率を優先した手法である．このとき，パソコンと自動車どちらの精度もランダムにフィルタリング行った場合に比べそれほど上がってはいないが，再現率についてはどちらも大きく向上していることがわかり，本フィルタリング手法の有効性が言える．さらに，正解商品用のフィルタリングについての精度と再現率に着目する．これは検索結果の商品が多い場合に用いる方法で，再現率よりも精度を優先した手法である．このとき，再現率についてはランダムに比べ大きく上回っているものの，100%ではないため，正解商品をいくつか削除してしまっていることがわかる．しかし，優先した精度についてはノイズ商品用のフィルタリングに比べ大きく向上しているため，本フィルタリング手法の有効性が言える．これらのことから，２つのフィルタリング手法の有効性とそれを切り替えて用いることの効果が確認された．

7.2 商品紹介文からの情報抽出についての評価

本節では三つの評価実験を行った．一つ目の実験は本研究で提案した情報抽出の手法がどの程度有効であるかを調べるために行った．これについて 7.2.1 節で述べる．二つ目の実験は属性データベースが商品紹介文からの情報抽出においてどの程度有効であるかを調べるために行った．これについて 7.2.2 節で述べる．これらの実験はすべての属性に対する抽出精度の平均値を調べることによって行った．しかしながら，抽出の困難さは属性によって異なる．このため，三つ目の実験ではそれぞれの属性ごとに精度を調べた．これにより抽出手法について考察を行う．これについて 7.2.3 節で述べる．

7.2.1 抽出の精度と再現率

抽出手法を評価するため，抽出の精度と再現率を求めた．抽出の精度と再現率は次の等式によって計算される．

	精度 [%]	再現率 [%]
パソコン	60.3	61.9
自動車	71.3	52.2
ベビー服	56.6	54.2
平均	62.7	56.1

表 5 情報抽出の精度 P_e と再現率 R_e

(1) 抽出の精度 $P_e = | B | / | A |$

(2) 抽出の再現率 $R_e = | B | / | C |$

ただし数式中の A , B , C は次の集合を意味する．

A : NTM-Agent が抽出したすべての属性値の集合.

B : NTM-Agent が抽出した正しい属性値の集合.

C : 実験で用いたすべての商品紹介文に含まれる属性値の集合.

表に 4 実験条件を示す．実験結果を表 5 に示す．抽出の精度と再現率は共に約 60%となっていることがわかる．インフォーマルなテキストを対象としていることを考えると，これらの値は高いと考えられる．

7.2.2 属性データベースの評価

属性データベースの効果を調べるため，抽出の精度と再現率を求めるときと同じ条件で，属性データベースを用いた場合と属性データベースを用いなかった場合の精度と再現率を計算し，その比である改善率を計算した．改善率は次の数式により計算される．

(1) 改善率 $I = | B | / | A |$

ただし A と B は次の集合である．

A 属性データベースを用いなかった場合の抽出の精度 (または再現率)

B 属性データベースを用いた場合の抽出の精度 (または再現率)

この実験は 7.2.1 節で述べた条件と同じ条件で行った．表 6 に実験結果の改善率を示す．この再現率は平均で 1.1 倍になっているため，これにより新たに 1 割の属性値を抽出することができる．このことから，本手法の有効性が検証された．また，個別に改善率を見ると，パソコンの再現率が特別大きいことがわかる．この原因はパソコンの属性データベースへの登録が多いためと考察される．属性データベースへの属性値の登録はより抽出が簡単で精度の高い表と箇条書きから行われる．パソコンの商品紹介文には表や箇条書きの記述が多いため，パソコンの商品紹介文からはより多くの属性値を学習することができる．これらのことから，この手法は表や箇条書きが多い商品カテゴリに有効である．

7.2.3 属性ごとの抽出の比較

抽出の精度と再現率を求める実験で用いた商品につ

商品数	パソコン	自動車	ベビー服
30 件	「Lavie NX」	「カローラ　GT」	「ベビー服　男」
100 件	「Gateway」	「カローラ」	「オーバーオール」

表 2　実験で用いた検索語

カテゴリ	パソコン	自動車	ベビー服
検索語	「Vaio」	「Stepwagon」	「baby clothing」
商品紹介文中の属性値	118 個	138 個	142 個
抽出する属性	8 個	7 個	7 個
属性名のキーワード	40 個	34 個	26 個
属性データベースへの学習	1000 件の商品から	1000 件の商品から	1000 件の商品から

表 4　実験の条件

	精度	再現率
パソコン	1.07	1.26
自動車	1.02	1.03
ベビー服	1.04	1.05
平均	1.04	1.11

表 6　属性データベースによる改善率 I

いて属性ごとに精度と再現率を調べた．その結果を表 7 に示す．抽出モジュールでは数値と固有名詞を優先して抽出を行うため，CPU，走行距離，サイズなどの数値と固有名詞を属性値とする属性の精度と再現率が高い．それに対し，精度と再現率が小さい属性についてその原因を考察した結果，次の理由があげられる．

属性名の記述が少ない　属性名が記述されてない場合，NTM-Agent は事前に学習した属性データベースから抽出を行うが，属性名が記述されることが少ない属性に対しては，属性データベースに学習することができず抽出が困難になる．色のような属性は「白のカローラです.」「赤いおむつカバーです.」といった記述がほとんどであり,「白の」や「赤い」に対する「色」という属性名が書いていないことが多い．そのため，抽出用ドメイン知識を用いて属性名を検索することができず属性データベースに学習できなかった．

文章中にのみ登場する　文章は他の形式よりも抽出が困難であり，文章中に記述されることの多い属性は精度が低い．傷，ドット抜け，汚れといった属性は表や箇条書きには記述されず,「傷やドット抜けは全く無い美品です.」「多少のシミはご容赦ください.」といった複雑な文も含まれていた．そのため，文末の用言を抽出するという単純な方法では抽出できなかった．

属性名自体が属性値である　NTM-Agent は抽出時に属性名の付近の名詞を抽出するという一定の抽出方法を取っている．このため，属性名自体が属性値であるような記述がある場合，それに対応して属性名を抽出することができない．つまり，CDD，新品/中古，性別といった属性は「その他：CDD　FDD　USB」「新品タグ付き」「男の子用おむつカバー」のような属性名自体が属性値となっている記述が多い．これらに対し「CDD」「新品」「男の子」という抽出用ドメイン知識を用いて抽出を行った場合，NTM-Agent は最も近い数値，固有名詞，名詞を抽出するため,「FDD」「タグ」「おむつ」という属性値を抽出してしまう．

さらに，この結果から抽出の精度と再現率を上げる方法を考察した．以下にその方法を述べる．

抽出用ドメイン知識に属性値の例を追加する　属性名の記述が少ない属性に対しては属性名を全く使わずに属性データベースに属性値を登録して抽出を行う方法が考えられ，色のような限られた属性値しか取らない属性には有効である．

構文解析を行い，係り受け情報を利用する　文章中にのみ存在し，複雑な文が多い属性（傷，ドット抜け，汚れ）に対して，より高度なテキスト解析を行う方法が考えられる．ただし，ネットオークション上のテキストは口語的であり，正しく構文解析できるとは言いがたい．そのためそれらを正しく機能させるためには経験的な工夫を必要とするものと考えられる．

抽出用ドメイン知識に属性値の型を付加する　属性名自体が属性値である属性に対して属性名を取り出すよう，ドメイン知識に属性の型を定義する方法が考えられる．つまり，抽出用ドメイン知識に属性の型 (タイプ A とタイプ B) を定義し，タイプ A の属性に対しては属性名付近の数値，固有名詞を抽出し，タイプ B の属性に対しては属性名をそのまま抽出を行うこととする．これにより，属性名が属性値となっているような属性も抽出することができる．

パソコン	CPU	MEM	HDD	OS	モニタ	CDD	傷	ドット抜け
精度 [%]	93.4	55.6	68.4	86.7	45.5	33.3	30.8	33.3
再現率 [%]	83.3	60	72.2	81.3	76.9	28.6	28.6	14.3

自動車	走行距離	年式	車検	色	排気量	AT/MT	傷
精度 [%]	91.3	85	76	15.4	66.7	83.3	20
再現率 [%]	75	65.4	65.5	9.1	16.7	52.6	12.5

ベビー服	サイズ	素材	枚数	色	新品/中古	性別	しみ
精度 [%]	84.2	66.7	50	26.1	56	40	25
再現率 [%]	74.4	58.3	42.9	54.5	45.2	50.0	20.0

表 7 属性ごとの抽出の精度と再現率

8. 結　　論

本論文では，現実のネットオークション上で商品紹介文から必要な情報を抽出し，それらを分かりやすく表形式で表示するエージェントを提案した．本システムにおいてノイズ商品の問題に対して，タイトルと商品紹介文中のキーワードを用いて相関ルールによりフィルタリングを行う手法とその際に用いる相関ルール作成をマーケットバスケット分析により支援するツールを用いた．また，複数の記述形式の問題に対して，記述形式を判別してそれらに最も適した方式で抽出を行った．さらに属性名の記述抜けの問題に対して，属性データベースに抽出した属性値の記述を学習させ，属性名の記述が無い属性値からの抽出方法を提案した．その結果，次のことが示唆された．

- 相関ルールを用いたフィルタリング方法はノイズ商品が少ないカテゴリの商品には有効ではなかったが，ノイズ商品がある程度含まれているカテゴリの商品については，本研究で用いたフィルタリングの手法は有効である．
- 情報抽出の精度と再現率は共に 6 割ほどであり，インフォーマルなテキストを対象とした手法としては，本研究で用いた情報抽出の手法は有効である．
- 属性データベースを用いることでさらに 1 割の属性値を抽出することができ，属性データベースへの学習は情報抽出に有効である．

また，本システムには次のような課題があると考えられる．商品紹介文から商品の特徴を抽出する際の手がかりとなるキーワードを，商品のカテゴリごとにサービスプロバイダもしくはユーザが手作業で入力しなければならない．抽出する特徴を表現するキーワードを思いつくこと自体はそれほど困難ではないが，商品ごとに行わなければならない問題と，同義語を列挙しなければならない問題がある．この問題に対しては，シソーラスを使い手がかりとなるキーワードを補完することや，属性名と属性値の対等が明確な記述形式 (表や箇条書きなど) から新たな属性を学習することが，解決方法として考えられる．今後はこれらのサービスプロバイダの支援を目指す研究を行っていく予定である．

参 考 文 献

1) 楠村幸貴, 土方嘉徳, 西田正吾「NTM-Agent:ネットオークションを対象としたテキストマイニングエージェント」，データベースとＷｅｂ情報システムに関するシンポジウム (DBWeb2002)，pp375-pp382，2002 年 12 月．
2) Yukitaka Kusumura, Yoshinori Hijikata, Shogo Nishida: ”NTM-Agent:Text Mining for Net Auction ”, The 2003 International Symposium on Applications and the Internet(SAINT2003), pp.356-359,2003.
3) 価格.com ，http://www.kakaku.com/
4) Libra， http://www.libra.ne.jp/
5) bestlot.com， http://www.bestlot.com/
6) 伊藤孝行，福田直樹，新谷虎松:”マルチエージェント入札支援システム BiddingBot におけるエージェント間の協調的入札機構について”，第 8 回マルチエージェントと協調計算ワークショップ (MACC-99)，1999
7) Doorenbos，R. B.，Etzioni，O.，and Weld，D. S. : ”A Scalable Comparison-Shopping Agent for the World Wide Web”，in Proceedings of the First International Conference on Autonomous Agents，pp. 39-48，1997
8) 日本語形態素解析システム「JUMAN」http://www-nagao.kuee.kyoto-u.ac.jp/nl-resource/index-j.html
9) R. Agrawal, T. Imielinski, A. Swami: ”Mining Associations between Sets of Items in Massive Databases”，in Proceedings of ACM SIGMOD Conference on Management of Data，pp207-216，1993