NTM-Agent: テキストマイニングによるネットオークションの商品比較支援

楠村　幸貴†　　土方　嘉徳†　　西田　正吾†

† 大阪大学大学院基礎工学研究科　〒 560–8531　豊中市待兼山町 1–3

E-mail: †kusumura@nishilab.sys.es.osaka-u.ac.jp, ††hijikata@sys.es.osaka-u.ac.jp, †††nishida@sys.es.osaka-u.ac.jp

あらまし　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特徴表を生成する支援を行う．そのため我々は，ユーザの検索要求に適合する商品の Web ページを収集し，商品の特徴について説明している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特徴表を作成するエージェント（NTM-Agent:Net auction TextMing Agent）を構築した．

キーワード　テキストマイニング，情報抽出，ネットオークション，電子商取引

NTM-Agent: Support of Item Comparison in Net Auction by Using Text Mining Techniques

Yukitaka KUSUMURA†, Yoshinori HIJIKATA†, and Shogo NISHIDA†

† Graduate School of Engineering Science, Osaka University　1–3 Machikaneyama, Toyonaka, Osaka 560-8531, JAPAN

E-mail: †kusumura@nishilab.sys.es.osaka-u.ac.jp, ††hijikata@sys.es.osaka-u.ac.jp, †††nishida@sys.es.osaka-u.ac.jp

Abstract　Net auctions have been widely utilized with the recent development of the Internet. However, it has a problem that there are too many items for bidders to select the most suitable one. We aim at supporting bidders on net auctions by automatically generating a table containing the features of some items for comparison. We constructed a system called NTM-Agent(Net auction Text Mining Agent). The system collects the information of items, filters the items which is different from the item of the user's target, and extracts their features from their text information by text mining methods. After that it generates the table containing extracted features.

Key words　Text Mining，information extraction，net auctions，e-commerce

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(注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) (注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(注3)を付けて分類を行うため，目的の

(注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

(注2)：パソコンの場合,「CPU」や「メモリ」などのキーワード

(注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．

商品と異なる商品 (以下，ノイズ商品) (注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. 章では問題 1 に対処する方法としてノイズ商品のフィルタリングについて述べる．5. 章では問題 2 に対処する方法として紹介文からの情報抽出について述べる．6. 章では実装したシステムの構成と実行例について述べる．最後にまとめを 7. 章で行う．

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．

(注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

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	○	○	○

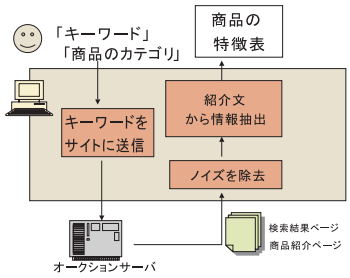

図 1　NTM-Agent の処理の流れ

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は 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

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．

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予め相関ルールの生成ツール(注5)によって生成される．このツールはユーザに教師信号を入力させ，マーケットバスケット分析により商品と相関の高いキーワードを出力するものである．

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

商品紹介文はタグによって表，箇条書き，文章に判別される．判別された後，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(注6)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．

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．

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

ユーザが Web ブラウザ上で手軽にシステムを使用できるよう，我々は NTM-Agent を Java サーブレットとして実装した．図 2 にシステムの構成を示し，以下に処理の流れを示す．

1: ユーザは NTM-Agent の Web ページ中のフォーム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．また価格についての条件を入力することも可能である．NTM-Server はユーザからのリクエストを受け付け，探索モジュールへ送る．

2: 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．このとき，予め学習したフィルタリング用のルールを用いてノイズ商品を省く．さらに，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．

3: 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，さらにフィルタリング用のルールを用いてノイズ商品を除く．

(注5)：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．

(注6)：形態素解析システムの辞書から判別する

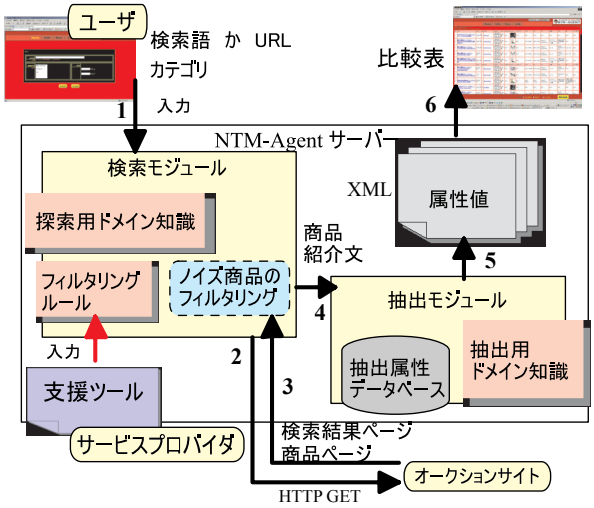

図 2　システムの構成

4:　残った商品について，商品紹介文が抽出モジュールに送られる．商品に対する抽出用ドメイン知識を参照し，商品の特徴値の抽出を行う．

5:　抽出した商品の特徴値を XML 形式で保存する．

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

6.1　システム実行例

NTM-Agent の入力ウインドウを図 3 に示す．NTM-Agent の使用方法は次の通りである．

（1）　目的の商品に関する検索語もしくは，オークションサイトの検索結果ページへの URL を 1 に入力する．

（2）　目的の商品のカテゴリを 2 のコンボボックスから選択する．

（3）　3 の検索ボタンをクリックする．

（4）　「検索が終了しました」というメッセージを確認したら、4 のボタンをクリックして出力ページを表示する．

NTM-Agent の出力ページを図 4 に示す．特徴表の一行目の各列には商品の属性名が表示され，二行目以降にはそれぞれの商品のタイトルと各属性が表示される．商品のタイトルはネットーオークションの商品紹介ページへのリンクとなっている．また，商品紹介文中に属性名の記述が無く，抽出できなかった属性値には「n/a」という表示がされる．

7.　ま　と　め

近年 Web 上には，掲示板やネットオークションのようにインフォーマルなテキストが存在している．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．本システムでは，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．これらはネットオークションというドメインだけでなく，インフォーマルなテキストを対象とした Web アプリケーションの開発に有効であり，今後このような技術の需要は高まっていくと考えられる．

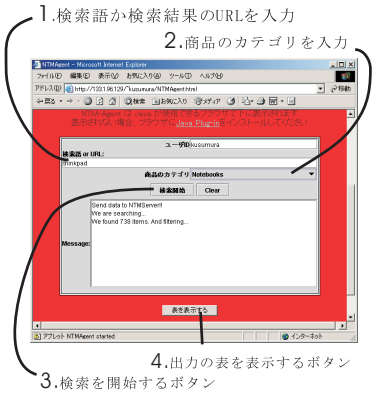

図 3　入力インターフェース

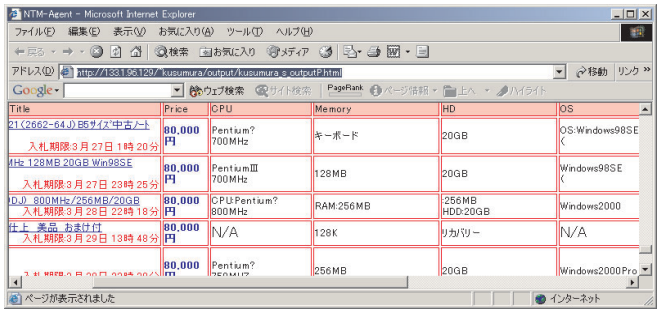

図 4　出力インターフェース

文　　献

[1] Y.Kusumura, Y.Hijikata, S.Nishida: ”NTM-Agent:Text Mining for Net Auction ”, The 2003 International Symposium on Applications and the Internet(SAINT2003), pp.356-359,2003.

[2] Y.Kusumura, Y.Hijikata, S.Nishida: ”Text Mining Agent for Net Auction”, ACM Symposium on Applied Computing(SAC 2004), 2004(Selected).

[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, 服部宏充, 新谷虎松: “エージェント間の協調的入札機構に基づく複数オークション入札支援システム BiddingBot”, 人工知能学会論文誌, 人工知能学会, Vol.17, No.3, pp. 247-258, 2002.

[7] R.B.Doorenbos, O.Etzioni, D.S.Weld: ”A Scalable Comparison-Shopping Agent for the World Wide Web”, in Proceedings of the First International Conference on Autonomous Agents, pp. 39-48, 1997