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

2005－DBS－136（9）
2005－FI－79（9）
2005／5／19

e-マーケットプレース向け情報収集・抽出方式における人的コストの調査

楠村幸貴[†]　土方嘉徳[†]　西田正吾[†]

近年，Web 上には一般のユーザが記述した情報が大量に存在している．例えば e-マーケットプレースには，大量の売り手が多種多様な商品の情報を提供している．このためこのようなサイトから商品の情報を収集し統合してユーザに提示する技術が望まれている．精度の高い情報収集・抽出を行うためには商品についての知識を用いることが有効であるが，その入力には人的コストがかかる．実用的なシステムの開発にはシステムの構築に必要な知識を入力する人的コストとそれによって実現される精度・再現率が重要となる．我々は e-マーケットプレースを想定した複数の情報収集・抽出方式を実装し，精度・再現率と知識の入力にかかる人的コストの関係を調べる比較実験を行った．本論文ではその結果得られた e-マーケットプレースでの情報収集・抽出システムを構築する際の指針を明らかにする．

A Study of Human Workload in Information Collection and Information Extraction Methods for e-Market Place

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

With the recent development of the Internet, e-market places have been grown to big markets which have various items. There are so many items that a user cannnot select an item easily. E-market places need an information integration system which filters pages and extracts features of items from the pages. For constructing the system with high performance of filtering and extracting,the system developer needs to input a large size of knowledge about the items.This costs a lot of human workload. Therefore the balance of performance and human workload is important for a practical system design. We investigated some filtering methods and extracting methods from the viewpoint of not only the performance but also the human workload. This paper reports the results of the experiments and the analysis in designing information filtering/extracting system for e-market places.

1. はじめに

近年，ネットオークションなどの e-マーケットプレースが盛んであり，大量の一般のユーザが多種多様な商品の情報を提供している．これらのサイトでは，各ユーザが思い思いに商品の情報を記述する仕組みを取っているため，商品の情報についての記述は不均一であり，ユーザが商品を比較しづらいという問題がある．また，これらのサイトでは，ある程度商品の出品方法においてルールを設けているものの（30 文字程度で商品名をつける，商品にカテゴリを一つ指定する，など），ユーザが思い思いに出品情報を記述しカテゴリに登録するため，買い手が商品を検索する際の検索結果に検索要求と異なる商品（例えばパソコンのカテゴリに存在する「バッテリー」や「メモリ」，さらにスパムメールのような「宝くじを当てる方法」といった商品が挙げられる.）が混じることが多いという問題も存在している．従って，このようなサイトから商品の情報を収集・抽出し，情報を統合してユーザに提示する支援システムが望まれている．図 1 に情報収集・抽出システムの例を示す．情報収集・抽出システムは以下のように動作する．(1) ユーザが入力した検索要求を対象の e-マーケットプレースの Web サイトに送信し，Web ページを収集する．(2) ノイズページをフィルタリングする．(3) 正解ページから情報を属性と属性値の形で抽出する．(4) 抽出した情報を統合したインタフェースをユーザに提供する．

このような情報収集・抽出システムの構築を目指した場合，(2) と (3) の処理が問題となる．これらの処理を自動化するためには，少なからず商品についての知識をシステムに与える必要がある．例えばルールベー

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

スのアプローチでこれらの処理を行おうとすると，そのフィルタリング/抽出ルールをシステム構築者やユーザが用意しなければならない．また，機械学習ベースのアプローチを取る場合，フィルタリング/抽出ルールを学習するための訓練データを用意する必要がある．本論文ではこれらフィルタリングと情報抽出を行う際に必要となる対象についての知識（ルールや訓練データ）をドメイン知識として定義する．

精度と再現率の高いシステムを構築するためにはより多くのドメイン知識を用いることが有効である．しかし同時に，システム構築者やエンドユーザがその知識を用意する作業は大変なものになる．実用的なシステムの設計にはドメイン知識を入力する作業量（以下，人的コスト）とそれによって達成される精度・再現率のバランスが重要である．著者らが調査した限りでは，これまでこのような調査を行った研究は報告されていない．そこで本研究では，情報収集・抽出システムの構築にかかる人的コストとその精度・再現率の関係を明らかにすることを目的とする．我々はドメイン知識の量と質を変えた複数の情報収集・抽出方式を実装し，ネットオークションサイトから収集した実データに対して，各種方式の精度・再現率と人的コストの関係を調べる．これにより我々は情報収集・抽出システムを構築する際の設計指針を明らかにする．我々はこの知見が本研究の社会的な貢献になるものと期待している．

本論文の構成は次のようになっている．本論文ではまず 2 章において関連研究を述べ，3 章において実験のために作成した実験データについて述べる．そして 4 章と 5 章においてフィルタリング実験と情報抽出実験について述べる．さらに 6 章において，それらの結果から考察を行い，最後に 7 章においてまとめを行う．

2. 関 連 研 究

本研究は情報収集・抽出システムの要素技術であるノイズ商品のフィルタリングと商品の特徴の抽出について，それぞれの各種方式を比較するものである．このうちノイズ商品のフィルタリングは商品を正解商品とノイズ商品に分類するもので文書分類の分野と関連が深い．また商品の特徴の抽出は情報抽出の分野と関連が深い．これらの分野では公開された共通のデータセットを利用し，種々の手法を共通の課題で比較するということが行われてきた．まず文書分類の分野では新聞記事を分類するデータセットとして Reuters21578[1] やスパムメールフィルタリングのデータセットとして Ling-Spam anti-spam filtering corpus[2] がある．これまでこの分野では，この課題についていくつかの手法が試され，サポートベクターマシン[3)4)] やナイーブベイズ分類器[5)] で機械学習を行った手法が良い結果を達成することが示されている．次に，情報抽出の分野では MUC[6]，IREX[7]，NTCIR[8] という団体が新聞記事から地名や人名を抽出する課題を正解データと共に公開している．これらに対しては，サポートベクターマシンを用いた手法[9)10)] が良い結果を示すことが知られている．

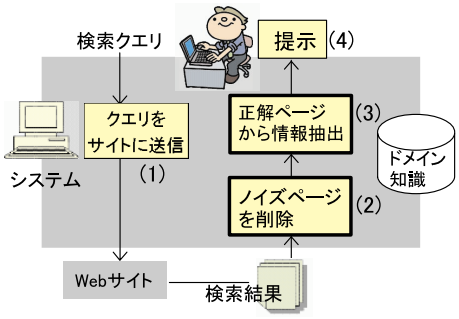

図 1 情報収集・抽出システム

これらの研究は学習に必要な訓練データの量と精度と再現率を比較するのみで，人的コストまでを意図して比較するものではない．本研究では学習ベースのアプローチだけでなく，人手でルールを記述する手法や人手でルールを修正する手法についても調べる．この点が従来の研究と本研究の違いである．

3. 実験データの作成

我々は代表的な e-マーケットプレースである Yahoo!Auctions[11] のカテゴリ-コンピュータから実験データを収集した．収集は商品情報をサイトから自動で集めるプログラムを作成し，約一ヶ月間一週間おきに行った．我々はその結果収集できた約 5 万件の商品のうち，出品者名が同じ商品を削除し，12080 件のみを使用することとした．これは同じ出品者が記述した紹介文はコピー&ペーストを使用して記述されていることが多く，機械学習を行う際にデータに偏りが生じないようにするためである．さらに我々は収集した商品に対してフィルタリングと情報抽出の正解データを付加した．フィルタリングの正解データは対象の商品（本実験用データではコンピュータ）であるかどうかを表す 2 値とし，情報抽出の正解データは商品の紹介文に含まれる抽出値それぞれに対し，属性の名前 (本実験用データでは CPU, メモリ，ハードディスク，ディスプレイ，OS，ドット抜けの数)，抽出する文字列，紹介文中での登場位置（文書先頭からの文字数）とした．これらの正解データの付加作業は時間がかか

るため専用のツールを開発し，実験者が 3 週間かけて入力した．その結果実験者が正解の判断がつかなかった 355 件を除き 11725 件の実験データを作成することができた．

4. フィルタリング実験

本実験では，前章で得られた実験データに対し，各商品がパソコン（以下，正解商品）であるかそれ以外の商品（以下，ノイズ商品）であるかを判別する問題を対象とする．判別の方法としては，機械学習を用いた方式（以下，学習ベース），人手で学習によって得られたルールを修正する方式（以下，修正学習ベース），教師信号を用いずに学習する方式（以下，クラスタベース），人手でルールを構築する方式 (以下，辞書ベース) の 4 つを用いる．本章ではまず各手法のアルゴリズムと想定しているドメイン知識の入力方法について 4.1 節で述べる．次に各種手法の精度・再現率と人的コストの関係を調べる実験について 4.2 節で述べる．

4.1 各フィルタリング方式

ここでは実験で用いた 4 つの手法について述べる．辞書ベース以外の手法では商品を単語列で扱うが，これらは商品名と紹介文に対して形態素解析[12] を行い，低頻度語（5 回以下の単語）と高頻度語（上位 100 位までの単語）を削除することで作成した．以下ではそれぞれの手法のアルゴリズムと想定するドメイン知識の入力方法を示す．

4.1.1 辞書ベース

1. ドメイン知識の入力　入力者が正解商品又はノイズ商品（以下，クラス）と相関が高いと思うキーワードを辞書に入力する．辞書はクラスごとにテキストファイルとして用意する．

2. 分類　システムが各クラス用の辞書を用いて分類したい商品を調べ，辞書に含まれるキーワードが何回登場しているかを数える．その結果，辞書中のキーワードが多く登場したクラスに分類する．

4.1.2 クラスタベース

1. クラスタリング　システムが訓練データ中の商品に対し，正解データを用いずにクラスタリングを行う．クラスタリングアルゴリズムには k-means 法[13] を用いる．ベクトルの重み付けには tfidf 値[14] を用いる．

2. ドメイン知識の入力　入力者がそれぞれのクラスタを調べ，正解商品のクラスタかノイズ商品のクラスタかのクラス付けを行う．クラス付けは図 2 に示すような入力インタフェースを用いて行う．

3. 分類　システムが分類したい商品とそれぞれのクラスタの重心との距離を計算し，最も近いクラスタのクラスを割り当てる．

4.1.3 学習ベース

1. ドメイン知識の入力　入力者がそれぞれの商品へのクラス付けを行う．クラス付けは各商品に対して商品名，価格，正解商品か否かを意味するチェックボックスを表示するインタフェースを用い，クリック操作によって行う．

2. 学習　クラス付けされた商品を教師信号として用いて SVM で学習を行う．カーネル関数は精度が高くパラメータの少ない動径基底関数[15] を用いる．なお，商品のベクトルは tf・idf 値を用いて重み付けを行う．

3. 分類　商品を上記のベクトル空間に射影し，SVM によって予測を行う．

4.1.4 修正学習ベース

1. ドメイン知識の入力　学習ベースと同様，入力者が人手で商品へのクラス付けを行う．

2. 学習　NB で学習を行う．まず，単語 w_i のクラス c に対する条件付確率 $P(w_i|c)$ を 単語の種類数 K とクラス c に含まれる商品数 $N(c)$ と単語 w_i を持ちクラス c に含まれる商品の数 $N(w_i, c)$ を用いて次のように定義する．

$$P(w_i|c) = \frac{N(w_i, c) + \frac{1}{K}}{N(c) + 1}$$

さらにこの確率を用い，単語の正解商品らしさ $\alpha(w_i)$ を次の式で計算する．

$$\alpha(w_i) = \frac{P(w_i|\text{ 正解商品})}{P(w_i|\text{ 正解商品}) + P(w_i|\text{ ノイズ商品})}$$

3. ルールの修正　入力者が単語の確率分布の修正を行う．修正には単語の追加と削除が考えられるが，人間が単語に正解商品らしさ $\alpha(w_i)$ を与えることはどの程度の値を付けてよいのか分からず難しいため，ここでは単語の削除のみを行うこととする．図 3 にルールの修正を行うインタフェースを示す．各行には単語 w_i と $\alpha(w_i)$ が表示され，チェック

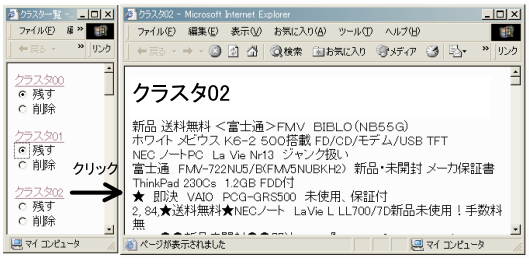

図 2　クラスタに対するクラス入力インタフェース

ボックスに削除するものを選択することでその単語は削除される．

4. 分類　分類したい商品 d に対し，商品名と紹介文に含まれる単語の条件付確率を用いて以下の正解商品らしさ $T_{true(d)}$ とノイズ商品らしさ $T_{false}(d)$ を計算し，大きい方に分類する．

$$T_{true}(d) = P(\text{正解商品}) \times \prod_{w_i \in d} P(w_i|\ \text{正解商品})$$

$$T_{false}(d) = P(\text{ノイズ商品}) \times \prod_{w_i \in d} P(w_i|\ \text{ノイズ商品})$$

4.2 評価実験

次に我々は 4 種類のフィルタリング手法について，ドメイン知識の量を変化させながら精度 P_f と再現率 R_f を求めた．これらは次の式によって計算される．

- 精度 $P_f = \mid B \mid / \mid A \mid$
- 再現率 $R_f = \mid B \mid / \mid C \mid$

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

実験では実験データ 11725 件のうち 90%の 10553 件を訓練データ，残り 10%の 1172 件を評価データとした．また，実験で用いるフィルタリング用のルールは実験者が作成することとし，それぞれ以下のような手順で実験を行った．

辞書ベース： 実験者が訓練データの一部を参考にしながら思いつかなくなるまでルールの入力を行い，全部で 212 個のルールを用意した．実験ではこのうち N 個のルールを選び出してフィルタリングを行い，精度と再現率を調べる．N は 10 から 212 まで変化させ，それぞれの場合で精度と再現率を調べた．このときキーワードを選び出す順番は入力された順番と同じになるようにした．

クラスタベース: 訓練データに対し付加された正解データを用いずにクラスタリングを行い，実験者が各クラスタにクラス付けを行った．クラスタの数は 2 個の場合から 300 個の場合の間で変化させ，それぞれの場合に対して精度と再現率を調べた．

図 3　ルールの修正インタフェース

学習ベース： 訓練データの数を商品 100 件から 11725 件まで変化させ，それぞれの場合で学習を行い，それぞれの場合において精度と再現率を調べた．

修正学習ベース： 学習ベースと同様，訓練データの数を商品 100 件から 11725 件まで変化させてそれぞれの場合で学習を行った．その後の人手でのルールの修正は実験者が行った．訓練データの数は商品 100 件から 11725 件までの間で変化させている．その結果生成された 1500 個から 4000 個のルール（訓練データ数によって異なる）のうちそれぞれ 100 個程度の単語を削除した．

なお，修正学習ベースでは，修正を行わなかった場合（以下,「NB（修正無し）」）についても，分析を行うこれらの手法は用いているドメイン知識の質が異なるため，これらの結果を一つのグラフに並べることができない．そこで本研究ではそれぞれの手法が必要としているドメイン知識の入力時間を被験者実験によって調べた．この被験者実験は日常的にコンピュータを扱っている 20 代〜30 代の男女 10 名に対して行われ，ルールの記述（辞書ベース），ルールの修正（修正学習ベース），クラスタに対するクラス付け（クラスタベース），商品に対するクラス付け（学習ベースと修正学習ベース）という 4 つの作業にどれだけの時間がかかるかを調べた．その結果，各作業にかかる時間の推定を 95%の信頼区間で行った．結果を図 4 に示す．この結果より各種手法の人的コストに対する精度・再現率の値をグラフに並べた．この結果のうち，精度と人的コストのグラフを図 5 に，再現率と人的コストのグラフを図 6 に示す．なお，作業時間が 500 分以上になると，値にほとんど変化が無いため割愛している．

まずクラスタベースの精度は低く，作業時間が増えても増加が見られない．これは生成されたクラスタが正解商品とノイズ商品をうまく分けられていないためである．次に NB を用いた手法（修正学習ベースの手法と NB《修正無し》の手法）を見ると，再現率が高い代わりに精度が低い．つまり誤って正解商品を削除することが少ない分多くのノイズ商品を削除できない．これは商品のクラスに含まれる単語の分布の違いが原因になっている．本問題はコンピュータの検索結果からコンピュータ以外の商品を削除する問題であり，正解商品に比べノイズ商品の多様性が高い．このためノイズ商品にはコンピュータと全く関係がなく頻度の低いものが多い（「温泉旅行」や「英会話のテープ」等）．

しかし，このような頻度の低い商品の単語は学習がしづらい．このためノイズ商品を判別するためには正解商品と相関の高い単語が無いことが決め手となることが多い．NB を用いた手法は商品に含まれる単語のみを手がかりとして分類するため，このような負の相関を利用することができない．これに対し，SVM はバランスよく精度と再現率が同等に増加していくことがわかる．SVM は商品をキーワードの頻度ベクトルとして扱うため，商品に含まれない単語についても相関を調べることができるためである．

また，辞書ベースの方法は作業時間が増えるに連れて再現率が下がっている．これは，ルールの数が増えてくると，その中には必ずしもそうとは言い切れないルールが混じってしまう可能性が高くなるためである．例えば,「用」という単語は「Thinkpad 用メモリ」といったノイズ商品を削除するために有効であるが，中には「自宅用パソコン」といった商品も存在する．人手で記述したルールでは，このような細かい例外に対応できず誤って正しい商品を削除してしまう特徴がある．

5. 情報抽出実験

本実験では，パソコンの紹介文から CPU, メモリ，ハードディスク，ディスプレイ，OS，ドット抜けの数という特徴（以下，属性）についてそれぞれの値（以下，属性値）を抽出する問題を対象とする．抽出手法としては，抽出例から機械学習を行う方式（学習ベース），抽出対象の文字列の辞書を用いる方式（辞書ベース），ページの構造と手がかり語についての知識を用いる方式（ヒューリスティックベース）の 3 つを用いる．本章ではまず各手法のアルゴリズムと本実験で想定しているドメイン知識の入力作業の詳細について 5.1 節で述べる．次に人的コストと精度，再現率の関係を調べるために行った実験について 5.2 節で述べる．

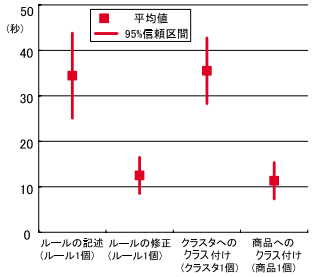

図 4　フィルタリングに必要な作業時間

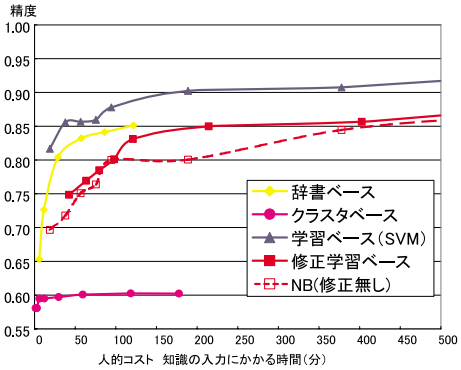

図 5　フィルタリングの精度

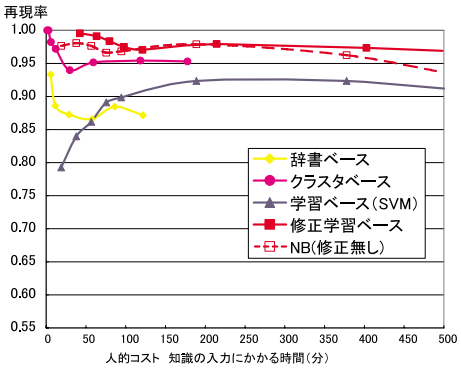

図 6　フィルタリングの再現率

5.1 抽 出 方 法

ここでは実験で用いた 3 つの手法について述べる．なお，フィルタリング実験と同様，辞書ベース以外の手法では紹介文に対して形態素解析を行って単語を取り出し，低頻度語と高頻度語を省いたものだけを用いて実験を行った．以下ではそれぞれの手法を述べる．

5.1.1 辞書ベース

以下に辞書ベースの手順を示す．

1. ドメイン知識の入力　属性値の辞書を用意する．辞書はそれぞれの属性ごとに，値の例となる文字列を記述したテキストファイルである．辞書では数字も含め文字列すべてを記述する必要がある．人的コストを考えると正規表現を用いることも考えられるが，抽出対象の文字列に曖昧性をもたせると，その分抽出エラーが増える．この点を考慮しながら正規表現を入力することは困難である．このため実験では単純化するために正規表現を用いないこととした．

2. 抽出　辞書を用いてキーワードマッチングを行い，最長一致したものを抽出する．

5.1.2 学習ベース

学習ベースは SVM を用いた山田らの手法[9] である．山田らの実験では，抽出位置の表現方法として IOB1, IOB2，IOE1，IOE2 の 4 種類，解析方向として右向

きと左向きの 2 種類，計 8 種類の手法を実験している．本研究では予めこの 8 種類について予備実験を行い，その結果が最も良かった IOE2 の右向き解析を用いることとした．この方法によって学習ベースの手法は次のような手順で抽出を行う．

1. ドメイン知識の入力　人手で抽出対象に対してクラス付けを行う．クラス付け作業は図**??**で示したインタフェースで行う．

2. 学習　クラス付けされた紹介文データを元に SVM で学習を行う．カーネル関数動径基底関数を用いる．

3. 分類　商品を上記のベクトル空間に射影し，SVM によって抽出を行う．

5.1.3　ヒューリスティックベース

商品の属性として記述される情報は属性名（例：「CPU」）と属性値（例：「Pen3 800MHz」）のセットによって記述されている．我々は，代表的な e-マーケットプレースにおいて商品の紹介文の調査を行い，紹介文の記述形式は表，箇条書き，文章の 3 つに分類できることがわかった．このことから，本手法では商品のドメインに依存しない属性名と属性値の位置関係を記述形式ごとに予め登録しておき，ドメイン知識として属性名のみを用いて抽出を行う．具体的な処理を以下に示す．

1. ドメイン知識の入力　属性名の辞書を用意する．辞書は属性ごとに属性名の類義語をテキストファイルに記述したものである．

2. 抽出　紹介文を予め定義したタグ構造によって表，箇条書き，文章に判別する．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属性名の辞書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。」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　評 価 実 験

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

- 精度 $P_e = \mid B \mid / \mid A \mid$
- 再現率 $R_e = \mid B \mid / \mid C \mid$

ただし数式中の A ，B ，C は次の集合を意味する．

A ： 抽出したすべての属性値の集合．

B ： 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．

3 章で作成したデータセットのうち，属性値が存在した商品 5467 件を用いて実験を行った．我々はこれらのデータを 8 つに分け，うち 7 つに当たる 4686 件を訓練データとし，残り 1 つに当たる 781 件を評価データとした．実験課題としては 6 つの属性（CPU，メモリ，HD，OS，ディスプレイ，傷の有無）の情報を抽出することとした．それぞれの手法における抽出用ルールは次のように構築した．

辞書ベース： 訓練用データ 4686 件に含まれる属性値をそのまま辞書に登録して用いる．実験では辞書に登録するキーワードの数を 100 個から 6832 個と変化させてそれぞれの場合で精度と再現率を計算した．このとき，辞書に登録する順序が実験結果に影響を与えるため，最善の順番で登録していく場合（辞書のサイズをN個とするとき，訓練用データ中の抽出値のキーワードの頻度を調べ，高いものの順に上位N個を登録する）とランダムに登録していく場合（辞書のサイズをN個とするとき，訓練用データ中の抽出値からランダムにN個を選択して登録する）の 2 種類で実験を行い，その平均値を計算した．

学習ベース： 訓練データの数を商品 10 件から 4686 件まで変化させ，それぞれの場合で学習を行って抽出を行い，精度と再現率を計算した．

ヒューリスティックベース： 実験者が訓練データの一部を参考にしながら 96 個の属性名を登録し，精度と再現率を計算した．

フィルタリング実験と同様，本実験でもこれらの各種方式に必要なドメイン知識の入力時間を被験者実験によって調べた．被験者実験はフィルタリング実験と同じ 10 名に対して行われ，属性値の記述（辞書ベース），属性名の記述（ヒューリスティックベース），抽出対象のクラス付け（学習ベース）という 3 つの作業にどれだけの時間がかかるかを調べた．その結果を図 7 に示す．この結果より各種手法の人的コストに対する精度・再現率の値をグラフに並べた．この結果のうち，精度と人的コストのグラフを図 8 に，再現率と人的コストのグラフを図 9 に示す．なお，ヒューリスティックベースについては見やすくなるよう x 軸と平行に横線を引いた．

ヒューリスティックベースは少量の人的コストである程度の精度と再現率を実現していることがわかる．

☆ 形態素解析システムの辞書から判別する

辞書ベースの方法は，精度がドメイン知識の増加によって変化していない．これは新たなキーワードはマッチングのパターンが増えるだけであり，再現率にしか寄与しないためである．辞書ベースは初めからかなり精度が高い．本実験の対象は新聞記事などの多様な情報を含む文章に比べ特定の特定の商品の紹介文という小さいドメインに限定しているため，紹介文に登場する値のパターンに関しては限定されている．このため少ない量のドメイン知識である程度の再現率が得られる．しかし，それ以降の向上を目指すためには，登場する可能性が低いような属性値（特異な記述「ペンティの８００」や誤記述「ペンティアム 500MHB」など）を記述しなければならず，再現率の向上は困難である．このため，最終的にはある程度の揺らぎを許容するルールを作成する学習ベースが有効であったと考えられる．

6. 考　　察

我々はフィルタリングと情報抽出機能において，必要な人的コストと精度・再現率の関係を調べた．しかし，これらの方式は人的コストの量と達成される精度と再現率の関係において特徴が異なっている．実際のアプリケーションを設計する場合は，アプリケーションによってこれらの方式を使い分ける必要がある．方式を使い分ける方針について議論するため，各手法の精度と再現率について，それぞれ大きいか小さいかを，人的コストの大きさによってまとめた．表１にフィルタリング手法の結果を表２に情報抽出手法の結果を示す．各セルがそれぞれの場合での精度と再現率の高さを表している．なおデータが無い部分については「-」とした．情報収集・抽出システムを設計する際には，実現できる人的コストによって表の列を決定し，目指すアプリケーションが精度と再現率のどちらを優先すべきかによって表の行を決定することで方式の選択が可能となる．

例えば，単純に抽出した値を表としてユーザに提示するアプリケーションを考えた場合，フィルタリングでは商品の量が問題となる．つまり商品数がノイズ商品よりも多い場合，結果が見づらくなるため誤って正しい商品を削除することよりもより多くのノイズページを削除する，精度を重視した手法（学習ベース）が望まれる．これに対し，ページの数が少ない場合はノイズページが混じってもユーザにとってはより多くの情報が得られる方が好ましいため，再現率を重視した手法（人的コストが小ならば辞書ベース，人的コストが大ならば学習ベース）を選択すべきである．また，抽出に関しては，ユーザが属性値を確認できるため誤った属性値が含まれていてもそれほど問題とならないことが多い．このため，誤った属性値を含みつつもより多くの情報を収集するという再現率が優先する手法（人的コストが小ならばヒューリスティックベース，それ以外ならば学習ベース）が好ましい．また，エージェントのように抽出した結果を用いて意思決定を行うようなアプリケーションを考えた場合，抽出した値をユーザが確認せずに処理するため，誤った値が混入しないように，フィルタリング・情報抽出共に精度を重視した手法（フィルタリング：学習ベース，抽出：辞書ベース）を用いるべきである．このように，情報

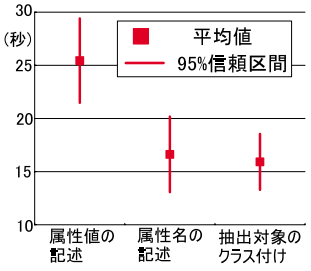

図 7　抽出手法に必要な作業時間

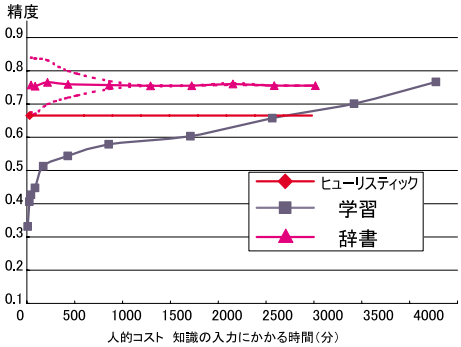

図 8　情報抽出の精度

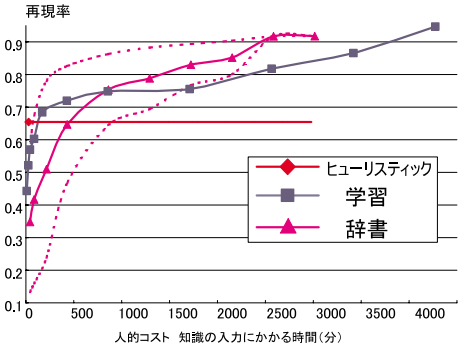

図 9　情報抽出の再現率

収集・抽出システムを設計する際にはこの表を用いて方式を選択することで，人的コストと精度・再現率のバランスの取れたシステム構築が可能となる．

またこれらの精度・再現率をフィルタリングと情報抽出の間で比較すると，情報抽出にかかる人的コストはフィルタリングにかかる人的コストに比べて圧倒的に大きい．これは文章の内容までを解析する情報抽出問題の難しさを示している．また，フィルタリングでは学習ベースの方式が常に有効であるのに対し，情報抽出では辞書ベースの方式にも有効性が見られる．辞書ベースの知識の入力について考えると，フィルタリングで用いる辞書の構築には人間が有効なキーワードを考える必要があるが，情報抽出では抽出する対象の例を入力すれば良いという違いがある．このため情報抽出の辞書は人間にとって作成しやすい．また学習ベースの方式を用いる場合，フィルタリングでは文書に含まれる単語でベクトルを作成するが，情報抽出では属性値の付近の語だけでベクトルを作成する．このため情報抽出の学習ではデータのスパース性の影響が大きく，学習効率が悪い．これらのことが情報抽出において単純な方法である辞書ベースの手法に有効性が見られた理由だと考えられる．情報抽出についての技術はまだまだ発展段階であり，さらなる技術の向上が望まれていると言える．

表 1 フィルタリング方式の選択方針

人的コスト 作業時間	小 50 分以内	大 50 分以上
辞書ベース	精度△再現率△	精度△再現率△
学習ベース	精度○再現率×	精度○再現率△
修正学習ベース	- -	精度△再現率○

表 2 抽出方式の選択方針

人的コスト 作業時間	小 700 分以内	中 3000 分以内	大 3000 分以上
辞書ベース	精度○ 再現率×	精度○ 再現率○	- -
学習ベース	精度× 再現率×	精度× 再現率○	精度○ 再現率○
ヒューリスティックベース	精度△ 再現率△	- -	- -

7. ま と め

我々は不均一な Web サイトに対しての情報統合支援を目指し，情報収集・抽出システムを開発する際に必要となるフィルタリングと情報抽出について，種々の方式を精度再現率と人的コストの関係から明らかにした．これにより，新たな情報サービスを提供したいサービスプロバイダがより少ない人的コストで性能の良いシステムを設計できるものを期待している．また，本研究で得られた知見が新たな情報収集・抽出の手法を研究する際の手がかりとなれば幸いである．

参 考 文 献

1) Reuters21578 Test collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/.
2) Ling-Spam anti-spam filtering corpus. http://idl.ils.unc.edu/ efrom/data/lingspam/.
3) T. Joachims. Transductive inference for text classification using support vector machines. *Proceedings of ICML-99*, pp. 200–209, 1999.
4) H.Taira and M.Haruno. Feature selection in svm text categorization. *Proceedings of AAAI-99*, pp. 480–489, 1999.
5) P.Pantel and D.Lin. Spamcop: A spam classification and organization program. In *Learning for Text Categorization: Papers from the 1998 Workshop*, Madison, Wisconsin, 1998. AAAI Technical Report WS-98-05.
6) N.A.Chinchor. Overview of muc-7/met-2. *Proceedings of MUC-7*, 1999.
7) IREX:Information Retrieval and Extraction Exercise. http://cs.nyu.edu/cs/projects/proteus*ash*irex/.
8) NII-NACSIS Test Collection for IR Systems. http://research.nii.ac.jp/ntcir/.
9) 山田寛康, 工藤拓, 松本裕治. Support vector machine を用いた日本語固有表現抽出. 情報処理学会論文誌, Vol. 43, No. 1, pp. 43–53, 2002.
10) K.Takeuchi and N.Collier. Use of support vector machines in extended named entity recognition. *Proceedings of CoNLL-2002*, pp. 119–125, 2002.
11) Yahoo!Auctions. http://auctions.yahoo.co.jp/.
12) 茶筌. 日本語形態素解析システム. http://chasen.naist.jp/hiki/ChaSen/.
13) J.MacQueen. Some methods for classification and analysis of multivariate observations. *5th Berkeley symposium on mathematics, statistics and Probability*, pp. 281–296, 1967.
14) W.R.Herch. *Infomation Retrieval:A Health Care Perspective*. New York, Springer, 1996.
15) S. S. Keerthi and C.J.Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural Computation*, pp. 1667–1689, 2003.