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

2004－ICS－136　(20)
2004／8／5

ラフ集合理論援用によるWebページのテキスト分類

板倉弘幸 †　田村雅樹 ‡　若木利子 †

† 芝浦工業大学大学院 電気工学専攻
‡ 北陸先端科学技術大学院大学 情報科学研究科

あらまし: 近年，WWW 上の Web ページは爆発的に増加しつつあり，それと共に, ディレクトリースタイルの検索エンジンを持つ Yahoo サービスのようなポータルサイトでは，膨大な Web ページを複数カテゴリーに自動分類するニーズが高まりつつある．本研究では, Web ページ分類に貢献する適切な単語素性 (feature) の撰択法に関して，ラフ集合理論の有効性を調べた．計算機実験による性能評価より，ラフ集合理論援用による属性選択法と分類器 (classifier) として線形核の Support Vector Machine を用いた組み合わせは，実用に耐えうる良い分類精度を保証しつつ，アドホックな閾値に依存しない高い次元圧縮 (属性選択) を可能にするなどの結果が得られた．

Rough Set-Aided Feature Selection for Automatic Web-Page Classification

Hiroyuki Itakura†　Masaki Tamura‡　Toshiko Wakaki†

†Shibaura Institute of Technology
‡Japan Advanced Institute of Science and Technology

abstruct: Recently Web-pages on World Wide Web are explosively increasing, and it is now required for portal sites such as Yahoo! service having a directory-style search engine to classify Web-pages into many categories automatically. This paper investigates how rough set theory can help select relevant features for Web-page classification. Our experimental results show that the combination of the *rough set-aided feature selection* method and the Support Vector Machine with linear kernel is quite useful for the practical purpose to classify Web-pages into many categories because the performance gives the acceptable accuracy achieving high dimensionality reduction without depending on arbitrary thresholds for the feature selection.

1　はじめに

近年，WWW 上の Web ページは爆発的に増加しつつあり，それと共に，WWW 上の検索エンジン機能をもつポータルサイト・サービスは，益々その需要が高まりつつある．例えば，ディレクトリー・スタイル検索エンジンを持つ Yahoo のようなポータルサイトでは，膨大な Web ページを階層構造を持つ複数カテゴリーに分類する作業が増大しつつある．しかしながら現在は，この Web ページのカテゴリー分類作業は人手で行なわれており，多大な開発時間や人件費を必要とする．それ故，Yahoo のようなポータルサイトでは，開発コストや開発時間の削減のために Web ページの自動分類のニーズが高まりつつある [9].

最近このようなニーズに対して，塚田らは，機械学習法に基づく Web ページの自動分類の方法 [9] を提案した．彼らのアプローチでは，先ず Yahoo! JAPAN の 5 個のトップカテゴリーから約 1,300 個の Web ページをダウンロードし，バスケット分析

により，各カテゴリー (分類クラス) の Web ページに或る閾値以上に頻繁に現れるアイテムの集合 (多頻度アイテム集合) を Web ページを特徴づける属性として生成している．そして，これらの分類クラスが既知の Web ページを訓練データとし，これらから生成された属性に関して決定木学習ツール C4.5 [6] を用いて分類規則を学習し，この結果，Yahoo! JAPAN のトップカテゴリーへの Web ページの分類に，実用に耐えうる分類精度が得られたことが報告されている．しかしながら塚田らの方法を適用する場合，多頻度アイテム集合という属性の生成において，最小指示度 (minimum support) と称される閾値を設定しなくてならない．Web ページ分類で実用に耐えうる，或は，できるだけ高い分類精度が得られるような最小指示度の適切な閾値は，そのような属性生成時に予め知ることができず，そのような最適な閾値は，種々の閾値に関して分類精度の評価をした結果から求めることができる．Web ページ分類という実用上の目的からは，要求される分類精度を与える適切な属性は，このようなアドホックな閾値に依存せずに導出されることが望ましい．

本研究ではアドホックな閾値を不用とする属性選択法として，ラフ集合理論 [4] を援用したアプローチを採用した．評価データとして，塚田らの研究で用いられた Yahoo の Web ページデータを用いた．そして Web ページを各ページの素性単語 (名詞) を属性とする文書ベクトルで表し，属性選択法としてはラフ集合援用による属性選択，TF-IDF 法，多頻度アイテム集合によるものを，分類器として決定木学習法の C4.5 と Support Vector Machine の TinySVM [12] を用いて，これらの組み合わせによる分類精度の計算機実検による評価を行った．この結果，ラフ集合理論援用による属性選択は元の属性数に対し 3%という高い次元削減が可能であり，かつラフ集合の属性選択と線形核 SVM の分類器との組み合わせの分類精度が最も良く，これは，塚田らの実験の最良の分類精度よりも高い，或は同等の精度を得ることができる，という良い結果を得ることができた [10] .

以下，2 章で本論文で研究対象とする属性選択方法 (ラフ集合による属性選択法，TF-IDF 法など) の説明，3 章で 2 種類の分類器 (C4.5，SVM) の概要，4 章で計算機実験と評価結果，考察等の議論を述べ，5 章で論文のまとめを述べる．

2 属性選択方法

分類問題では，一般に “素性”(feature，“属性” と称することもある) の数が大きく，素性の中には “不適切”(irrelevant)，あるいは “冗長”(redundant) なものが存在する．M. Dash ら [3] の定義によると，“適切”(relevant) な素性とは，その素性を削除すると分類器の分類性能や分類精度が悪くなるような素性であり，不適切および冗長な素性は，適切ではない素性である．“不適切” な素性が存在すると，全素性を用いた分類器の分類性能を悪化させる．よって feature selector (以後，属性撰択方法と称する) のモチベーションは，(i) 選択された属性を用いて分類器 (classifier) を簡単化すること (ii) 分類器の分類精度を向上させる，または著しく低下させないこと (iii) 分類器が大量のデータを扱うことができるようにデータの次元を削減すること，となる．

これまで属性選択方法としての多くのアプローチが提案されており，或る閾値に依存して属性選択をする方法とそうでないものがある．本論文では，前者の方法として，TF-IDF 法と多頻度アイテム集合 [9] による属性選択法を，後者としてラフ集合援用による属性縮約 [4, 1, 7] の方法に関して比較研究を行う．

2.1 ラフ集合理論援用による属性縮約

ラフ集合理論援用による属性縮約 (*Rough Set-aided Dimensionality Reduction*，以後，RSDR と称する) は以下のように定義される．オブジェクトが行，属性が列の属性値表 (decision table) を T としたデータセットがある．U はデータセットの中のすべてのオブジェクト集合，A は $a : U \to V_a$ (但し, $a \in A$，V_a は a の値の集合) なる全ての属性の集合を表す．A は，条件属性の集合 C，および相互に排他的な $C \cup D = A$ である決定属性の集合 D に分割される．任意の $P \subseteq A$ について以下の同値関係 $I(P)$ が定義される．

$$I(P) = \{(x, y) \in U^2 \mid \forall a \in P \ a(x) = a(y)\}$$

これは $(x, y) \in I(P)$ において，オブジェクト x と y は属性集合 P によって識別することができないことを示す．“P の識別不能な同値関係”(*P-indiscernibility equivalence relation*) $I(P)$ の同値

類は $[x]_P$ で表される．所与の同値関係 $I(P)$ (但し，$P \subseteq C$) について，任意の $X \subseteq U$ に対する “下近似” $\underline{P}X$ は次式で定義される．

$$\underline{P}X = \{x \in \mathrm{U} \mid [x]_P \subseteq X\}$$

“D の *C-Positive Region*” は，次式で定義される．

$$\mathrm{POS}_C(D) = \bigcup_{X \in \mathrm{U}/D} \underline{C}X$$

これは C の属性値によって商集合 U/D の要素のクラスに確実に分類することができる U のオブジェクトの集合を意味している．

もし，$\mathrm{POS}_{(C-\{c\})}(D) = \mathrm{POS}_C(D)$ であれば，属性 $c \in C$ は属性値表 T において分類上必要ではない．そうでなければ属性 c は T において必要不可欠となる．属性集合 $R \subseteq C$ が，条件 $\mathrm{POS}_R(D) = \mathrm{POS}_C(D)$ を満たす $\subseteq$ に関する極小 (*minimal*) の属性集合であるならば，R は C の縮約 (*reduct*) と呼ばれる．しかし計算量やメモリの問題として，すべての縮約計算は NP-Hard となる [7]．この計算量の問題を克服するために，Web ページのテキスト分類のための属性選択として，我々は依存度 $\gamma_P(D)$ [1]

$$\gamma_P(D) = \frac{\| \mathrm{POS}_P(D) \|}{\| \mathrm{U} \|}$$

というヒューリスティックスを使用した以下の QUICKREDUCT アルゴリズム [4] を用いて縮約を計算する．この $\gamma_P(D)$ は完全なヒューリスティックスではないので，このアルゴリズムで極小の縮約を必ずしも生成するとは限らない．データセットの次元を非常に縮小することにおいて有用だが，極小の縮約に近い一つの縮約のみを見つける．QUICKREDUCT アルゴリズムは，C が n 次元の場合，最悪計算量は $O(n!)$ だが，平均計算量は，実験的にほぼ $O(n)$ であることが報告されている [1]．

QuickReduct(C, D, R)

Input: C はすべての条件属性集合
　　D は決定属性集合
output: 縮約された属性集合 $R(R \subseteq C)$

1. $R \leftarrow \phi$
2. **do**
3. 　$T \leftarrow R$
4. 　$\forall x \in (C - R)$
5. 　　**if** $\gamma_{R \cup \{x\}}(D) > \gamma_T(D)$
6. 　　$T \leftarrow R \cup \{x\}$
7. 　$R \leftarrow T$
8. **until** $\gamma_R(D) = \gamma_C(D)$
9. **return** R

[1] 任意の A に対して，$\| A \|$ は A の要素数を表す．

2.2 TF-IDF による属性選択

TF-IDF 値は，情報検索における索引語としての単語の重みを表す値で，(i) 一つの文書内に多く出現した単語はその文書の内容とより関係がある．(ii) より多くの文書に出現した単語は重要ではない．この 2 点を考慮し TF-IDF の値は決定され，本研究では下記の式を用いる．

$$tfidf(t,d) = \frac{tf(t,d)}{\Sigma_{s \in d} tf(s,d)} \left(\log \frac{N}{df(t)} + 1 \right)$$

t は単語，d は文書，N は全文書数を表す．$tf(t,d)$ は文書 d の中での単語 t の出現頻度で，$df(t)$ は単語 t が出現した文書数を表す．

TF-IDF 値を用いた属性選択方法は次のように定義される．ベクトル空間モデルでは，文書 d は，それに出現する単語 t を一つずつ座標軸に対応付けたベクトルで表現され，t の軸の成分の値は，それの TF-IDF 値がある閾値を超えていた場合に 0 でない値 (本研究では 1) を，それ以外は 0 の値を持つ．

2.3 バスケット分析に基づく属性生成

塚田らの研究 [9] では，データマイニングの分野で有名なバスケット分析に基づき，Web ページから属性値表のデータを作成するために多頻度アイテム集合を属性として生成する．バスケット分析はアイテムの集合を属性とみなし，設定した閾値 (支持度:support) よりも高いアイテム集合を導く．アイテム集合の支持度とは全トランザクション数に対し，アイテム集合 I を包含するトランザクション数の割合と定義されている．最小支持度 (minimum support) より大きな支持度を持つアイテム集合は多頻度アイテム集合と呼ばれる，塚田らは，この Web ページから抽出された多頻度アイテム集合を Web

ページのそれぞれのカテゴリーを反映する属性としている．

このように属性を生成した後，属性値表 (行列)T は以下のように定義される．c をあるクラス，T_c はクラス c の属性値表，$Page_i^c$ は，クラス c の i 番目の Web ページを意味し，そのページに含まれる名詞の集合 (アイテム集合) で表現される．$Attribute_j$ はクラス c より生成された j 番目の多頻度アイテム集合 $Itemset_j^c$ とする．$T_c[i,j]$ は $Itemset_j^c \subset Page_i^c$ であれば 1，そうでなければ 0 となる．最後に，T はすべてのクラス c の属性値表 T_c をあわせたものとなる．

3 分類器

2 章で述べた属性選択方法の有効性を評価するために，2 種類の分類器を使用する．分類器には決定木学習法 C4.5 と Support Vector Machines を用いた．

3.1 決定木学習法: C4.5

訓練データが属性値表として与えらると，決定木学習はクラスが未知の新しいデータを分類することができる決定木を生成する．新しい事例は，決定木の各ノードで評価され，その結果進むべき枝を決定する．C4.5 [6] は，各ノードで評価する “テスト” を利得比率と呼ばれる統計基準を使用し，ルートから利得比率の高いテストをノードとして生成して決定木を構築する．葉ノードは，テストの代わりにクラス・ラベルを含む．新しい事例が葉ノードに達すると，決定木学習法はそこに格納されたラベルを使用してそれを分類する．

3.2 Support Vector Machines

まず始めに，線形核の SVM を紹介する．d 次元特徴空間の中で訓練データは，各々のデータをベクトルとした $\{\mathbf{x_1},\ldots,\mathbf{x_m}\}$ で表すことができる．そして，$\{y_1,\ldots,y_m\} \in \{-1,1\}$ はそれぞれ $\mathbf{x_i}(\mathbf{1} \leq \mathbf{i} \leq \mathbf{m})$ に対応したクラス変数を表し，$y_i = 1$ は正例を表し，-1 は負例を表す．その後，これらの訓練ベクトル集合から学習されたクラスの境界線を使用して，未知のクラスのベクトルを分類する．線形核 SVM ではこの境界線は次の式，

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b.$$

で表されるように超平面となる．$\mathbf{w}$ は d 次元の重みベクトルで，b はバイアスを表す．

一般に，SVM によって訓練されたクラス予測は $prediction(\mathbf{x}) = sign(f(\mathbf{x}))$ で表される．$sign(z)$ は $z \geq 0$ であれば 1，そうでなければ -1 を返す関数である．また，$f(\mathbf{x})$ をサポートベクトル $\mathbf{s_i}$ の集合 SVs を用いて表現すると次の式が導かれる．

$$f(\mathbf{x}) \quad = \sum_{\mathbf{s_i} \in SVs} \alpha_i y_i \mathbf{K}(\mathbf{s_i}, \mathbf{x}) + b.$$

この $\mathbf{K}(\mathbf{z},\mathbf{x})$ はカーネル (核) と呼ばれるものである．線形核の場合，$\mathbf{K}(\mathbf{z},\mathbf{x}) = \mathbf{z}^T\mathbf{x}$ となり，分類予測は $sign(\mathbf{w}^T\mathbf{x}+b)$ と表すことができる．このとき重みベクトル $\mathbf{w}$ は次のように表すことができる．

$$\mathbf{w} = \sum_{\mathbf{s_i} \in SVs} \alpha_i y_i \mathbf{s_i}.$$

本研究では線形核のほかにも次のような多項式核 $(p = 2, 3)$ も使用した．

Polynomial Kernel $\quad \mathbf{K}(\mathbf{z},\mathbf{x}) = (\mathbf{z}^T\mathbf{x}+1)^p.$

この SVM のツールとして奈良先端科学技術大学院大学で開発された TinySVM [12] を用いた．

4 計算機実験と評価

4.1 使用データ

塚田らは Yahoo! JAPAN の 14 のトップカテゴリーより 5 つのカテゴリーを選択し，その Web ページに対して分類実験を行っている．そのカテゴリーとは芸術と人文 (Ah)，ビジネスと経済 (Be)，教育 (Ed)，政治 (Go)，健康と医学 (He) の 5 つである．(図 1 参照)

彼らはカテゴリーごと約 250 個の Web ページ (計 1270 個) を無作為にダウンロードし，その Web ページに対して HTML タグの削除，形態素解析などの前処理を行う．形態素解析には日本語形態素解析ツールである茶筌 [11] を用い，各ページは次で示すようにカテゴリー c ごとに名詞の集合の $Page_i^c$ $(1 \leq i \leq n_c)$ で表される．

$$class_c \Leftrightarrow \{Page_1^c,\ldots,Page_i^c,\ldots,Page_{n_c}^c\},$$
$$Page_i^c = \{word_{i,1}^c,\ldots,word_{i,j}^c,\ldots\},$$

この $Page_i^c$ は $class_c$ に属する i 番目の Web ページから抽出された名詞の集合である．また，$word_{i,j}^c$ は $Page_i^c$ の j 番目の名詞であることを示す．

本研究では，塚田らによって集められた 1270 個の Web ページを使用した．しかし塚田らの研究では 1270 個の Web ページから，さらに 1000 個のページを選択している．塚田らとはわずかに異なるがほぼ同じデータを用い，分類評価を行っている．

また，1270 個の Web ページから抽出された異なるアイテムの総数 (つまり名詞の個数) は，11385 個であった．このように前処理済みデータセット $\cup Page_i^c$ は Web ページを分類するための訓練データとなり，2 章で述べた種々の属性選択方法が適用される．

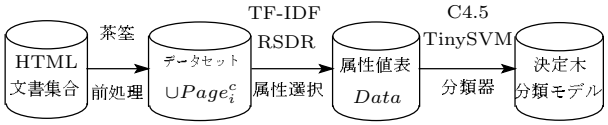

図 1: データと処理の流れ

4.2 性能評価の尺度

次に示す 4 つの値は，分類精度を評価する尺度として使用される [8].

- TP: 正例を正例であると判断した文書の数
- TN: 負例を負例であると判断した文書の数
- FP: 正例を負例であると判断した文書の数
- FN: 負例を正例であると判断した文書の数

これらの値を用い，分類評価として正答率 (*Accuracy*)，誤差率 (*Error rate*)，適合率 (*Precision*)，再現率 (*Precision*)，F_1 値が次式で定義される．

$$
\begin{aligned}
\text{正答率} &= \frac{TP+TN}{TP+TN+FP+FN} \\
\text{適合率} &= \frac{TP}{TP+FP} \\
\text{再現率} &= \frac{TP}{TP+FN} \\
F_1\ \text{値} &= \frac{2}{\frac{1}{\text{適合率}}+\frac{1}{\text{再現率}}}
\end{aligned}
$$

適合率と再現率はトレードオフの関係にあるが，どちらも考慮する必要がある．そこで，本研究では，適合率と再現率の調和平均である F_1 値を分類性能の最も重要な評価基準として採用した．

4.3 実験を行うための準備

4.1 節で述べた 5 つのカテゴリーからなる Yahoo データを用い，次の問題意識を持ち実験を行った．

(i) TF-IDF による属性選択方法や多頻度アイテム集合を属性とした方法と比較して，RSDR の方法はどの程度，Web 文書のテキスト分類に有効であるといえるか?

(ii) SVM のカーネルはどの核が Web 文書分類において性能が最も良いか.

(iii) QUICKREDUCT アルゴリズムの性能はどうか.

4.3.1 属性選択方法の適用

図 1 に示すように，前処理済みデータセット $\bigcup Page_i^c$ に 2 章で述べた各属性選択方法を適用し，選択された属性のみを用いた縮約データセット $Data$ を縮約された属性値表として生成する．

この実験では TF-IDF 値による属性選択と RSDR を使った 4 つの組み合わせにより，属性選択の評価を行う．

4.3.2 分類精度の評価

分類精度は，縮約データセット $Data$ を使用し，分類器 C4.5 または TinySVM を用いて以下のように評価される．

最初に，塚田らの実験で行われたように，各カテゴリー c に対して縮約データセット $Data$ より，そのカテゴリーの正例，負例の 2 クラスからなるデータセット $Data_c$ を構築する．つまり，正例は c に属する Web ページのオブジェクト，負例は c に属さないオブジェクトとなる．このような $Data_c$ は，2 クラス分類で分類評価するために使用される．その主な理由の 1 つに，分類器 TinySVM はデータを正例と負例の 2 クラス分類しかできないが，分類対象の Web ページ (オブジェクト) は 5 クラス (カテゴリー) に分類されており，分類対象の Web ページには複数のカテゴリーに属するものも存在する．2 ク

ラス分類を行うと Web ページは 1 つのカテゴリーに属さず，複数のカテゴリーに属する結果となる場合もあるが，そのようなページがあるのは上記のように当然であるので，この方法を使用した．

次に，データセット $Data_c$ を用いて分類器 C4.5 や TinySVM で n 分割交差検定による分類評価を行う．この方法は，$Data_c$ の中のオブジェクト集合を，ほぼ等しい要素数の n 個の部分集合に分割する．n 個の各部分集合について，それぞれをテストデータとし，そのテストデータ以外の $n-1$ 個の集合の和集合を訓練データとする．その n 個のサンプルに対して n 回同様な分類評価を行い，それらの評価値の平均を用いてこの研究のアプローチの評価値とする．塚田らの研究の実験では 4 分割交差検定を行っているので本研究でも同様の検定を行った．図 2 と図 3 は 5 つのカテゴリーごとの $Data_c$ を 4 分割し分類器 C4.5 と TinySVM で F_1 値を計算し，それを平均したものを示す．各図の中の 4 本のグラフは属性選択方法の 4 つの組み合わせに相当する．

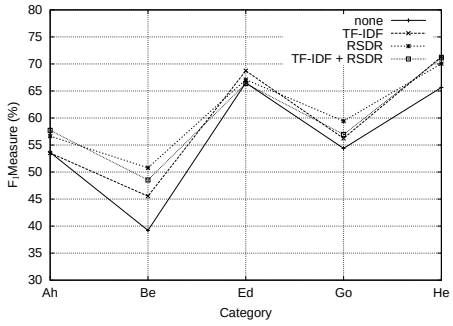

図 2: C4.5 によるクラス別 F_1 値の比較

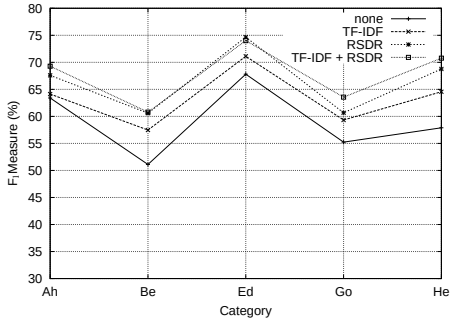

図 3: SVM によるクラス別 F_1 値の比較

表 1: 分類性能

分類器	属性選択方法			精度 (%)			
	R	T	属性数	正答率	適合率	再現率	F_1値
C4.5	×	×	11385	86.36	77.85	43.91	55.88
	×	0.4	1113	87.01	76.81	49.57	60.11
	○	×	336	87.23	78.16	49.94	60.80
	○	0.25	371	87.18	78.81	48.80	60.14
SVM (線形核)	×	×	11385	85.26	67.17	52.91	59.09
	×	0.4	1113	87.10	72.63	58.47	64.76
	○	×	336	88.19	77.62	58.16	66.46
	○	0.25	371	88.65	70.92	59.05	67.69
SVM (2次多項式)	×	×	11385	81.78	55.62	48.04	51.08
	×	0.6	638	84.74	64.94	53.03	58.14
	○	×	336	85.10	58.52	56.50	60.41

表 2: 塚田らのアプローチの分類性能

分類器	多頻度アイテム集合の属性集合		精度 (%)
	Minsup	属性数	F_1 値
C4.5	10%	823	66.3
	20%	78	57.8
	30%	19	53.0

4.4 実験結果と議論

表 1 は分類器 C4.5，線形核 SVM，2 次多項式核 SVM による 5 カテゴリー分類精度の平均を表し，4.3.1 節で示した属性選択法の 4 つあるいは 3 つの組み合わせで評価を行った．この表で R と T はそれぞれの QUICKREDUCT を用いた RSDR と TF-IDF 値による属性選択 (以後 TF-IDF と称する) を表す．また，その方法が適用されたかしないかを○と×で表現した．TF-IDF 法の場合は，閾値を可変にした実験を行い，F_1 値が最も高い時の閾値を○の代わりにその値を表記した．

一方，塚田らのアプローチによって得られた分類精度は，我々のアプローチの結果を評価するために表 2 に示す．”Minsup”は多頻度アイテム集合を生成するための閾値である最小支持度 (minimum support) を示す．また，F_1 値は [9, p. 310] にある分類精度結果をもとに計算したものである．彼らの実験では，4.1 節で言及したように，1000 個の Web ページを訓練データとし，分類器に C4.5 が使用されている．

上記の表の分類精度より，以下の結果が得られる:

1. 属性選択の有効性

   C4.5 と SVM のいずれの分類器でも属性選択を適用したほうが高い F_1 値が得られるので，

RSDR，TF-IDF あるいは両方の属性選択方法の適用は，分類精度向上に貢献するといえる．その上，RSDR は各分類器において TF-IDF の閾値を変えた実験を行って導き出された最も良い F_1 値よりも高い値が得られたことから，TF-IDF 法より有効であると言える．また，塚田らの属性選択方法と比較すると RSDR の方が閾値に独立なので実用的である上，我々の RSDR+線形核 SVM の方が高い F_1 値が得られたので RSDR の方が有効と考える．

2. 選択された属性数

   RSDR は，分類精度を下げることなく，また閾値を必要とせずに，元の属性数 11385 個に対し，約 3%にあたる 336 個に次元を削減することができた．ほかの属性選択方法と比べると圧倒的な圧縮率と言える．

3. 分類精度

   我々のアプローチでは，RSDR と線形核 SVM を用いた F_1 値が 66.46%，“RSDR と TF-IDF の両者の属性選択”+線形核 SVM を用いたものが 67.69%となった．この数値は塚田らの実験で最も良かった F_1 値の 66.3%と比較すると，それ以上，あるいは同等の精度がえられていることがわかる．これらの分類精度は Web 文書分類において，実用に耐えうる acceptable なものである．

4. 分類器の性能

   線形核 SVM，2 次多項式核 SVM，C4.5 の分類器を比較すると，それぞれの対応する属性選択において F_1 値は常に線形核 SVM が高いことより，線形核 SVM が Web 文書分類において最も適していると言える．属性選択をしない場合，C4.5 は 2 次多項式核 SVM よりも F_1 値が高いが，RSDR による属性選択した場合は，2 次多項式核 SVM が F_1 値 60.41%，C4.5 が 60.80 となり，ほぼ同じ精度が得られている．

図 4 は TF-IDF 閾値によって精度がどのように変化するかを示す．TF-IDF 閾値は 0.0 から 1.0 まで (属性数は 11385 から 258) 変化させ，それぞれ C4.5，線形核/2 次多項式核/3 次多項式核の SVM で分類を行い評価をした．この図より，TF-IDF 閾値 0.0 から 0.8 の間では線形核 SVM が最も良く，2 番目に C4.5，3 番目に 2 次多項式核 SVM が良く，3 次多項式核 SVM は最も悪い分類精度が得られていることが分かる．各分類器ごとの最も F_1 値が高い最適な TF-IDF 閾値は表 1 に示されている．

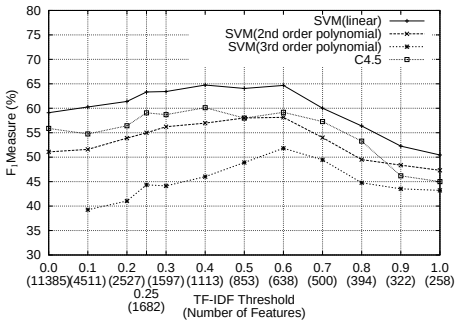

図 4: TF-IDF による属性選択の性能

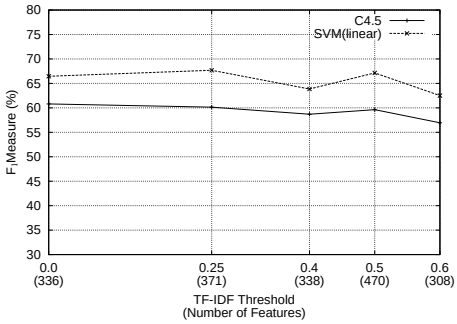

図 5: TF-IDF と RSDR の組み合わせによる性能

さらに，TF-IDF 法による属性選択を行った後に RSDR による属性選択を行った場合の結果を図 5 と表 3 に示す．図 5 は TF-IDF 閾値を 0.0 から 0.6 に変化させた結果を表し，表 3 は TF-IDF での属性選択の後で RSDR を行った場合と行わない場合の属性数を表す．これらの結果より，RSDR が TF-IDF のような閾値を持たずに分類に適切な属性を選択できるということを再確認できる．

図 6 は，RSDR 法で縮約属性集合を計算する QUICKREDUCT プログラムの実行時間を示す．11385 個の属性より 336 個の縮約属性を算出する計算に，Linux 上 (CPU2.4GHz Pentium IV) で 428 秒かかる．通常 RSDR の計算は一度のみ行われるものなので，実用上，問題の無い性能であると言える．

表 3: TF-IDF と RSDR の組み合わせによる選択属性数

属性選択		TF-IDF 閾値				
R	T	0	0.25	0.40	0.50	0.60
×	○	11385	1682	1113	835	638
○	○	336	371	388	470	308

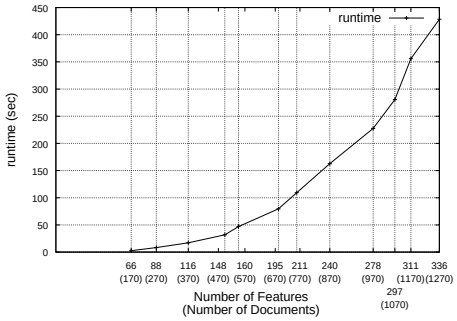

図 6: 属性数と RSDR 実行時間

5 議論

本研究では，Web ページの自動テキスト分類を行うためにラフ集合援用による属性縮約の方法 (以下 RSDR 法と称する) の有効性を評価した．その結果，RSDR 法と線形核 SVM の組み合わせによる分類が分類精度を著しく改善することがわかった．また，そのときの属性数は縮約前の属性数の約 3%となり，高い次元削減を誇る．その上，多くの属性選択方法で必要となるアドホックな閾値は不要である．RSDR は閾値に独立に 1 度の処理で分類に適切な属性集合を得ることができるので，Web ページのテキスト分類において非常に実用的に有用である．RSDR の処理は通常何度も計算をする必要がないので，QUICKREDUCT の最悪の場合の時間量 $O(n!)$ は，それほど深刻な問題ではないと考える．実際，この実験によって RSDR を計算した結果，実行時間は 482 秒だった．

SVM では，線形核 SVM が多項式核 SVM より Web ページ分類においてはより良い分類性能が得られた．その結果，Web ページ分類においては，多項式核 SVM よりも線形核 SVM の方が適していると結論される．

最後に，本研究では Yahoo トップレベルのカテゴリーに関する Web ページの自動分類を対象としているが，将来的にはディレクトリースタイル検索エンジンの階層構造に沿った Web ページの階層的分類を行うための方法を提案することが課題である．

謝辞

大阪大学の元田教授および鷲尾助教授には，本研究で用いた Yahoo データの提供と貴重な議論を賜わりました．前橋工科大学の N. Zhong 教授には，本研究に関する御支援を賜わりました．ここに深く感謝の意を表します．

参考文献

[1] A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorization. *Applied Artificial Intelligence*, 15:843–873, 2001.

[2] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods.* Cambridge University Press, 2000.

[3] M. Dash and H. Liu. Consistency-based search in feature selection. *Artificial Intelligence*, 151:155–176, 2003.

[4] J. Katzberg and W. Ziarko. Variable precision extension of rough sets. In W. Ziarko (ed.) *Fundamenta Informaticae, Special Issue on Rough Sets*,Vol. 27, No. 2-3:155–168, 1996.

[5] Z. Pawlak. *Rough sets: Theoretical Aspect of Reasoning About Data.* Kluwer Academic Publishers, 1991.

[6] J. R. Quinlan. *C4.5: Programs for Machine Learning.* San Mateo, CA: Morgan Kaufmann, 1991.

[7] Q. Shen and A. Chouchoulas. Rough set-based dimensionality reduction for supervised and unsupervised learning. *International Journal of APPLIED MATHEMATICS AND COMPUTER SCIENCE*, Vol.11, No.3:583–601, 2001.

[8] K. Shima. Identifying Discriminative Features form High-Dimensional Data using Support Vector Machines. *Ph. D. Thesis*, University of Tokyo, Tokyo, Japan, 2003.

[9] T. Tsukada, M. Washio and H. Matoda. Automatic web-page classification by using machine learning methods. *Proceedings of the First Asia-Pacific Conference on Web Intelligence (WI2001)*, LNAI 2198:303–313, 2001.

[10] T. Wakaki, H. Itakura and M. Tamura. Rough Set-Aided Feature Selection for Automatic Web-Page Classification. to appear in *Proceedings of 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI2004).*

[11] chasen. `http://chasen.aist-nara.ac.jp/`.

[12] TinySVM. `http://cl.aist-nara.ac.jp/~taku-ku/software/tinysvm/`.