手掛語に着目した問題文識別についてのナイーブベイズによる評価実験

酒井, 敏彦
九州大学

廣川, 佐千男
九州大学

出版情報：情報処理学会研究報告. 2013 (A-5-1), pp.1-5, 2013-03-15. 情報処理学会九州支部
バージョン：

手掛語に着目した問題文識別についてのナイーブベイズによる評価実験

酒井 敏彦[1,a)]　廣川 佐千男[1,b)]

概要：
研究者にとって関連論文調査は必須であり，特に，対象とする論文が何を問題ととらえているかという観点は重要である．したがって，論文概要から問題を記述している文（問題文）を自動的に抽出できれば，調査の精度と効率の向上が期待できる．SVM などの機械学習により判別機を構築できるが，判別性能を上げるには多量の学習例を人手で準備しなければならないという問題がある．筆者らは SVM を使ったこれまでの研究で，少数の学習例しかない場合でも，問題文を特徴付ける手掛り語集合に着目することで，全ての単語を用いるより判別性能を格段に向上できることを示している．本稿では，学習機としてナイーブベイズを用いた場合も，手掛り語集合が判別性能向上に有効であることを示す．さらにその性能向上率は，SVM の場合より高いことを示す．

Experiment Facused on Clue in Problem Sentences Identification Using Naivebayes

Abstract: It is important for researchers to investigate related work. In particular, it is important to consider the viewpoint that describes the problem of the paper they are aiming to find. Therefore, if we can automatically extract the problem sentences from paper abstract, our method is expected to improve the accuracy and the investigation and efficient. SVM is machine learning algorithm that can make classifications. But there are problems that we must first prepare many learning samples manually to increase classification efficiently. In past research we have proposed a method which can improve the classification efficiency by focusing on characterizing clue problem sentences. With only a samll number of learning samples we can obtain superior results. In this paper, we show that also by in addition to using naivebayes with clues is more effective than vectorizing by all words. Moreover, we show that the improved efficiency is higher than that when compared to SVM.

1. はじめに

近年，Web 上に公開される文書の量は爆発的に増え続けている．この問題に対応すべく，検索エンジンや情報検索の技術も発展している．研究を始めるときや研究成果をまとめるときにも検索エンジンを活用する関連研究の調査は必須となっている．しかし，指数関数的な論文の増加のため，キーワードで得られる検索結果の論文を 1 件ずつ読むことは困難である．研究者は通常自分の研究分野と同じ課題を扱っているかどうかの比較検討を行いながら関連研究の調査を行う．我々はキーワード以外の別の観点も検索の条件として設定を行うことが出来れば論文検索の精度と効率の向上につながると考える．本論文で最終的に目指していることは観点での検索を含むモーダル (様相) な検索が可能な検索手法または検索エンジンの開発である．モーダルな検索の例として，[14], [15] の研究がある．これらは複数の観点で検索を行い，検索結果を Facet として表示する．本稿は扱われている問題が何かという観点に着目する．

一般的なキーワードによる検索でなく，観点による検索，モーダルな検索を実現するには，対象とする文書中の各観点を表す部分の抽出が必要である．これらの情報抽出の研究では最適な手掛語を発見することは重要である．また，どの手掛語がクラスタリングにおいて重要な役割を果たしているのかを調べることも重要である．手掛語集合の効果は機械学習を通して評価することができる．従来の学習モ

[1] 九州大学
Kyushu University
[a)] 2IE11060Y@s.kyushu-u.ac.jp
[b)] hirokawa@cc.kyushu-u.ac.jp

デルとして決定木や最大エントロピー法などがあるが，これらの学習モデルは最適な属性選択を考慮して行わないと，過学習を起こしてしまうことがある．また，複数の単語の組み合わせで精度向上を図ることができたとしても，その組み合わせを探索するのは容易ではない．一般的に，SVM は属性選択において多くの属性を学習させても過学習しにくいとされている [3], [2]．しかし，現実には極少数しか学習データが入手できない事が多い．現実的には少ないデータで機械学習の効率化を図ることを考えなければならない．そこで，我々は学習データが少ない場合の手掛語集合の有効性について研究を行っている．これまでに我々は論文 [11], [12], [13] において学習データが少ない場合は属性の単語が多いとき F 値が頭打ちになり，すべての単語よりも限られた手掛語集合での結果のほうが良い性能を示し，手掛語集合の有効性を確認している．

本論文ではナイーブベイズを学習機とした場合の少数例による学習における属性選択の効果を評価した．特に論文概要の中で問題を記述する文 (本稿では問題文と呼ぶ) の抽出を考え，問題文抽出に有効な手掛語集合は何か，問題文判別に有効な特徴語は何かを考えた．特に，今回は一般的に単純な機械学習だと考えられているナイーブベイズによる機械学習で実験を行い，これまで得ている SVM についての評価との比較を行った．評価を行う手掛語集合としては (1) 問題文全体の特徴語，(2) 全単語でベクトル化し SVM で学習させたモデルにおける単語のスコア (これを SVM スコアと呼ぶ) の上位語，(3) 負のスコアで絶対値の上位語，(4) 正の上位と負の上位の両方を合わせた単語の 4 通りと全ての単語の 5 通りの単語集合でベクトル化を行い，分類性能を比較した．

2. 関連研究

SVM を使った分類で属性選択は重要なテーマとして多くの研究がある．SVM は属性数に依存せず，多くの属性からなるデータでも分類性能が悪くならないので，多くの分野で利用されている．例えば，[3] では，新聞記事の分類について，属性として使う単語の個数が多い程，分類性能がよいと述べられている．この結果から，文書分類に SVM を使う場合，属性選択について考える必要はほぼないといえる．しかし，[5] では，少数のサンプルしかないマイクロアレー遺伝子発現データの分類について，数千以上の属性に対し 40 個程度の属性選択でも同程度の分類性能が達成できると報告している．[6] では，全データをトレーニングデータとして学習したモデルにおける各属性のスコアを考え，1.0 に近い少数の属性だけに限定して再度モデルを作っても，ほぼ同程度の分類性能が達成できることを示している．このように，属性選択についての多くの研究は，対象を少ない属性により理解することが目的である．少数の属性でも全ての属性を使った場合と同程度の分類ができることを示しているが，少数の属性に限定することで，分類性能が向上するというものではない．また，これらの研究の多くでは，分類性能の評価におけるトレーニングデータの方が，テストデータよりサイズが大きい．例えば，[3], [7], [8] ではその比率は 1:1，[6] では 2:1，[9], [10] では 5:1，[5] では 9:1 である．

一方，ナイーブベイズを用いたテキスト分類の論文では一般的にベルヌーイモデルと多項式モデルが手法として用いられる。既存研究からはベルヌーイモデルよりも多項式モデルのほうが分類精度が高いと言われている。例えば、[17] や [18] ではニュース記事やスパムメールをデータとしてナイーブベイズの多項式モデルにおいて全ての単語を属性として用いるよりもネガティブな単語を含むベクトル化の方が精度が良いと述べられている。しかし、属性として使う単語の数が多い時は多項式モデルの方が分類性能が高いが、少ない時にはベルヌーイモデルの方が分類性能が高くなると報告している論文もある [16]。本研究では、ナイーブベイズの多項式モデルの方を用いている。

本研究では，少数例の学習では，少数の手掛語集合を使った方が，全ての単語を使ったベクトル化よりも，格段に判別性能が高くなることがあるという事例を示す．

3. 評価実験データ

今回データとして 2004 年から 2011 年に出版された電子情報通信学会[*1]の研究会論文 42,921 件を収集し，その中からランダムに 300 件を抽出した．300 件の論文概要を MeCab[*2]を用いて形態素解析した結果，全ての単語の数は 4,048 個であった．また，文の数は 1,344 文であった．SVM で学習させる際には，ラベル (-1 or +1) をつける必要がある．今回は論文概要の文において問題文である場合が+1 になる．逆に問題文でない場合は-1 になる．次に，正解データの作成を行うため，3 人の被験者に 300 件の論文概要を読んでもらい，各文が問題文であるか問題文でないかのチェックをしてもらった．判定基準として 2 人以上が問題文と判定したものを問題文とした．分類を行った論文構成要素は「背景」「問題」「関連研究」「目的」「手法」「結果」「その他」の 7 つの観点である．各文に対して，その文が複数の観点を含む場合もあるため，複数分類を行えるようにした．例えば，「A を実現するために，B の手法を提案する．」という文だと，A の部分が「目的」，B の部分が「手法」となり，この場合だと，文に対して 2 つの観点「目的」,「手法」を付与できる．この作業の結果，2 人以上が同じ観点に分類した文がその観点での正解文とする．結果として, 問題文は 149 文であった．

[*1] http://www.ieice.org/jpn/index.html
[*2] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

4. 問題文を特徴付ける手掛語集合の提案

仮説として全ての単語で学習させるよりも手掛語集合のような問題文判別に適した単語のみで学習させ，分類を行ったほうが良いと考えた．そこで今回 GETA による特徴語と SVM による単語スコアの 2 つの手掛語集合を考えた．

4.1 GETA における特徴語

GETA[*3]は大規模な索引付きデータを対象として，類似度を高速に計算できる汎用連想計算エンジンである．この GETA を用いることで入力した単語を含む文書集合を求め，更に，その文書集合の特徴語を求めることができる．GETA を利用するため，GETA のインデックスを作成した．例として，GETA における問題文の特徴語上位 10 個を表 1 に示す．表 1 の上位を見ると人間が問題文を判別する際に用いるであろう単語が出てきている．

表 1 問題文における GETA の特徴語

ランキング	単語	単語スコア
1	しかし	14.15
2	い	8.12
3	ない	7.45
4	困難	7.42
5	低下	6.60
6	問題	6.59
7	なる	5.06
8	が	4.95
9	しまう	4.86
10	ため	4.80

4.2 SVM スコアの上位語と下位語

今回 SVM のツールとして SVM-light[*4]を用いた．SVM-light とは，C 言語で記述されたサポートベクターマシンを利用するためのツールである．また，カーネルに線形カーネルを用いた．SVM-light では学習データは+,-の符号の後に単語 id とその文における単語の正規化した重みの列として一行で表す．分類のための各単語の重要度は SVM で得られたモデルによる単語の重みを用いた．SVM による問題文判定の評価指標として Precision, Recall, F-measure を用いた．今回は閾値 α を決め，SVM による文の推定値が α 以上の文を推定問題文として判定することによって評価指標とした．

全ての単語でベクトル化した際に学習データとテストデータを同じにした場合の単語のスコアが上位と下位の単語を手掛語集合とした．具体的には表 2, 3 の単語である正の上位 10 個，負の上位 10 個の単語などがある．

*3 http://geta.ex.nii.ac.jp/geta.html
*4 http://svmlight.joachims.org/

表 2 SVM の単語スコア上位 10 個

ランキング	単語	単語スコア
1	しかし	1.2524
2	まだ	0.4477
3	難しい	0.4295
4	欠点	0.3180
5	低下	0.2760
6	招く	0.2386
7	後者	0.2359
8	頻繁	0.2168
9	ISI	0.1804
10	安全	0.1797

表 3 SVM の単語スコア下位 10 個

ランキング	単語	単語スコア
4039	有効	-2.6113
4040	結果	-2.6460
4041	用い	-2.6858
4042	示す	-2.7369
4043	手法	-2.7637
4044	シュミレーション	-2.8957
4045	本	-2.9148
4046	本稿	-3.1270
4047	提案	-3.2176
4048	実験	-3.2314

5. ナイーブベイズにおける手掛語集合の有効性評価

ナイーブベイズ手法とはベイズの定理を適用することで実現できる単純な確率的分類器であり，スパムメールの判別等に用いられている．事前に教師あり学習の設定により各単語が属する項目の事前確率を求めることで，テストを行うデータの分類を行うことができる．本章ではナイーブベイズを用いて機械学習を行い論文概要の問題文識別を行う．

今回，4 つの手掛語集合について学習させる単語数を変化させることで F 値がどのように影響を受けるかという比較実験を行った．手掛語集合の単語数を N とする．この N を変化させることで F 値にどのような変化が起きるかを調べた．それぞれ 4 つの手掛語は GETA による特徴語の上位において「各文書における単語のスコア上位」，SVM スコア「ポジティブな単語とネガティブな単語」，「ポジティブな単語」，「ネガティブな単語」の 4 通りである．これらと「全ての単語」によるベクトル化で比較する．それぞれ簡略化のため「All」「GETA」「Pn+Nn」「P2n」「N2n」と呼ぶことにする．例えば，N=1 の場合，手掛語は「GETA」は GETA の特徴語上位 2 個，「Pn+Nn」はポジティブな単語 1 個とネガティブな単語 1 個，「P2n」はポジティブな単語 2 個，「N2n」はネガティブな単語 2 個となる．一般の

N の場合，各文についてのベクトルを作る時に 2N 個の単語に限定する．1 文に最大 70 個しか単語はないので，実験で N=60 のとき，つまり，120 個に限定して作ったベクトルは全ての単語を使ったものになっている．これらのことに加え，学習比率を 1 割の場合と 9 割の場合に分けることで手掛語集合の性能の比較を行った．

図 1 は 9 割学習データでナイーブベイズによる学習を行い，残りの 1 割で評価を行った F 値の結果を示す．この図から全ての単語を用いた場合，F 値は 0.098591 を示しており，手掛語集合を使った方が良い F 値となっていることがわかる．しかし，4 つの手掛語集合において明確な差はみられない．図 2 は学習データ 1 割による F 値の結果を示す．この図から全ての単語を用いた場合，F 値は 0.009976 を示しており，手掛語集合を使った方が良い F 値となっていることがわかる．学習データが 1 割の場合の手掛語集合の選び方で F 値に差がある．P2n は N を大きくするにつれて F 値が単調増加している．また，GETA による手掛語集合も N を大きくするにつれて F 値が単調減少している．しかし，N が小さいときには F 値は低い．N2n と Pn+Nn による手掛語集合は N を大きくするにつれて F 値が単調減少している．しかし，N が小さいときには高い F 値となっている．このことから問題文を識別するときには問題の観点を明確に表すポジティブな単語だけでなくネガティブな単語が必要であることがわかる．

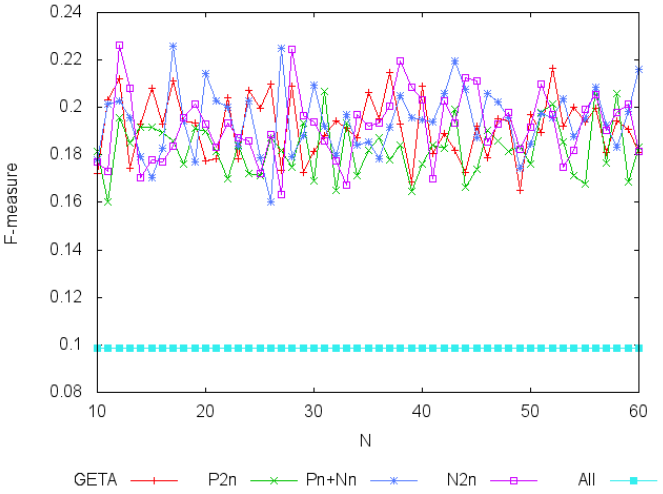

図 1 ナイーブベイズによる学習データ 9 割による F 値の結果

6. 手掛語集合による識別性向上率

ナイーブベイズと SVM において特に性能が良かった P2n,Pn+Nn に関して F 値の比較を行った．すべての単語を用いてベクトル化する方法をベースラインとする．図 3 と図 4 は以前に著者らが本論文と同条件下で SVM において実験を行った結果である [11], [12]．このときにも学習データが 9 割のときには手掛語集合が全ての単語でのベクトル化と同程度の性能を示したが，学習データが 1 割のときには手掛語集合が全ての単語でのベクトル化よりも良い性能を示した．図 5 と図 6 はそれぞれ SVM とナイーブベイズによる手掛語集合が全ての単語によるベクトル化よりもどの程度性能が向上したかを示した結果を表す．縦軸がそれぞれの項目における F 値をベースラインでの F 値で除算した値を表している．この比率が高いほど，ベースラインよりも良い性能を示していることを表す．横軸は単語の数 N である．

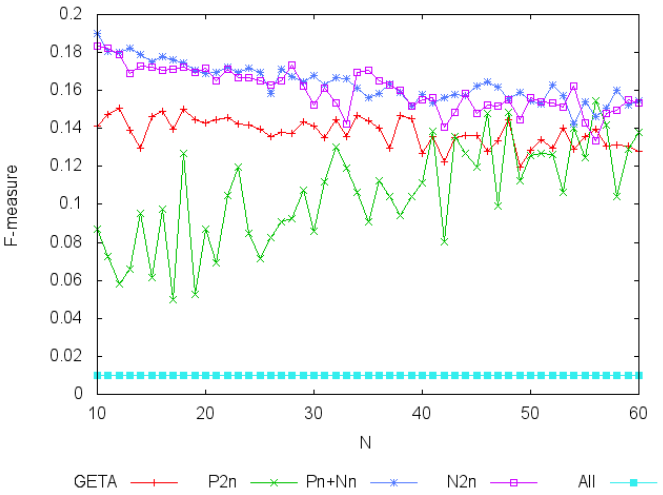

図 2 ナイーブベイズによる学習データ 1 割による F 値の結果

図 5 から学習比率を 1 割にした場合のほうが 9 割にした場合よりも値が高いことが確認できる．学習比率 1 割のとき，Pn+Nn も P2n の場合も値は 1.5 前後を示しており，ベースラインよりも 1.5 倍性能が良いことがいえる．逆に，学習比率 9 割のときは Pn+Nn も P2n も値 1 に満たないためベースラインのほうが性能が良いことがわかる．

図 6 からこちらも学習比率を 1 割にした場合のほうが 9

割にした場合よりも値が高いことが確認できる．加えて，値は一番良いときで 19 を示しており，図 5 のときよりも性能が格段に良いことがわかる．学習比率 1 割のときの Pn+Nn の F 値が一番高く，単調減少している．このことから，N が小さい時，つまり手掛語集合が限定する方が効いているということになる．ネガティブな単語を増やしているので，徐々に値が下がったと推測できる．学習比率 1 割のときの P2n の F 値が二番目に高く，こちらは単調増加している．このことから，ポジティブな単語を増やせば増やすほど，問題文を推定しやすいことを示した．学習比率 9 割のときは値は 2 程度にとどまり，学習比率 1 割のときのほうが格段に良い．

したがって，学習比率が小さい場合には，SVM に限らず，ナイーブベイズでも手掛り語が有効であることがわかった．すべてのキーワードを使うベースラインのとき SVM では F 値が約 1.5 倍になる．分別性能が SVM より低いナイーブベイズではすべてのキーワードを使ったとき F 値が約 19 倍となり，ナイーブベイズでは手掛り語の効果が大きいことが分かった．

7. おわりに

本研究では論文概要から問題の観点を抽出する際に手掛語集合がどの程度有効であるかということを SVM とナイーブベイズの機械学習における結果を比較することで行った．結果として，学習データの比率が大きい場合は SVM やナイーブベイズでの手掛語集合を用いたときの F 値は高々 2 倍程度にしかならない．しかし，学習データの比率が小さい場合には，SVM に限らず，ナイーブベイズでも手掛語集合が有効であることがわかった．SVM における全ての単語でのベクトル化と手掛語集合との比率は高々約 1.5 倍となる．さらに分類性能が SVM より低いといわれているナイーブベイズにおいては全ての単語でのベクトル化と手掛語集合との比率は高々約 19 倍となり，ナイーブベイズでは手掛語集合の効果が大きいことがわかった．したがって，学習データの比率が小さい場合には手掛語集合が有効であることが SVM やナイーブベイズ共通に確認することができた．

今後の課題としては，他のデータでも同様の現象が確認できるかを行う必要がある．

参考文献

[1] Cortes, C., Vapnik, V., Suppert Vector Networks, Machine Lerning, Vol.20, pp.273-297, 1995
[2] 工藤拓, 松本裕治, Support Vector Machine による日本語係り受け解析, 情報処理学会研究報告 自然言語処理研究会報告 2000(65), pp. 79-86, 2000
[3] Taira, H., Haruno, M., Feature Selection in SVM Text Classification, Proc. AAI99, pp. 480-486, 1999
[4] Sadamitsu, K., Saito, K., Imamura, K., Kikui, G., Entity Set Expansion using Topic information, Proc. ACL

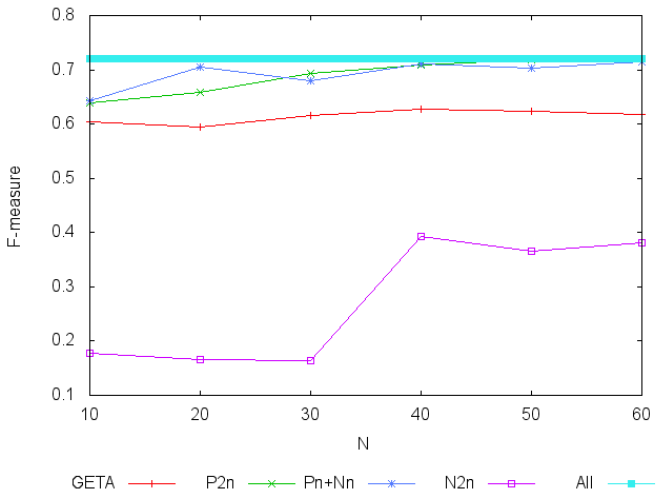

図 3 SVM による学習データ 9 割による F 値の結果

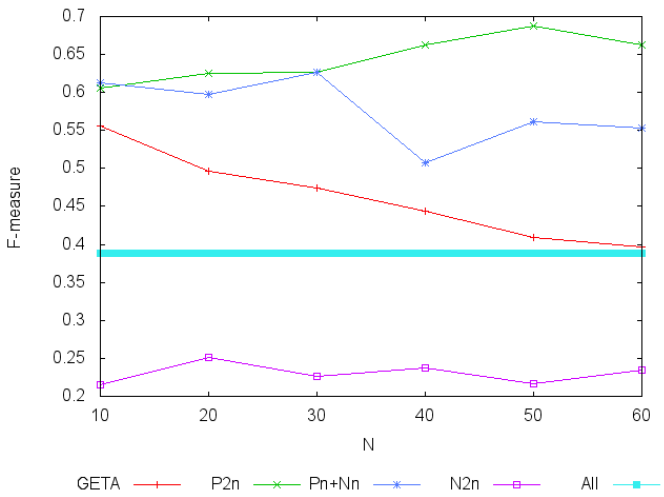

図 4 SVM による学習データ 1 割による F 値の結果

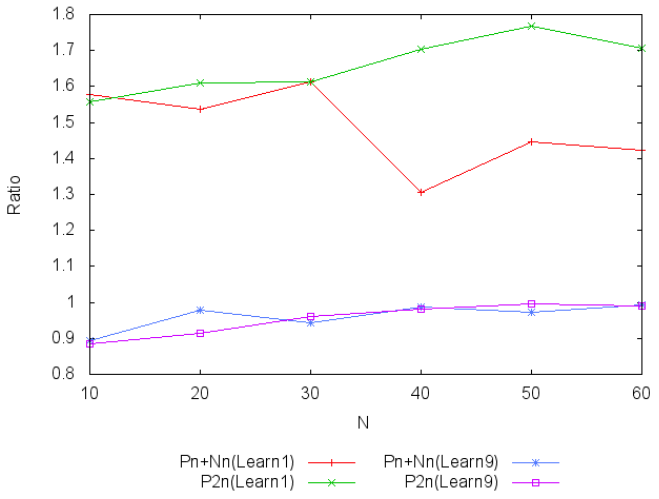

図 5 SVM におけるベースラインとの比率

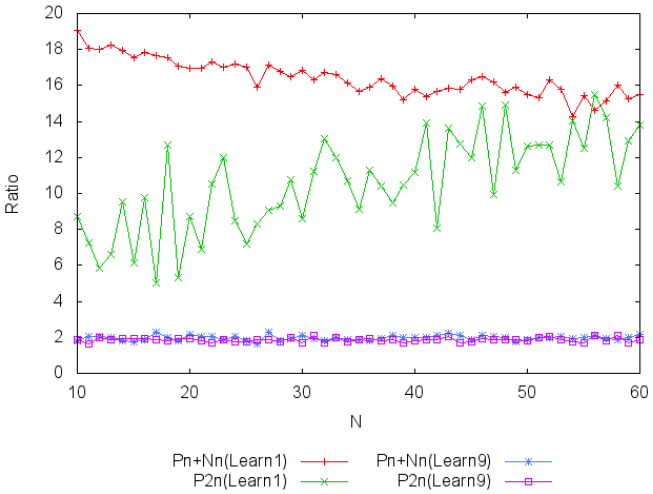

図 6 ナイーブベイズにおけるベースラインとの比率

2011, pp. 726-731, 2011

[5] Alonso-Gonzalez,C.J., Moro, Q.I., Prieto, O.J., Simon, M. A., Selectiong Few Genes for Microarray Gene Expression Classification, Springer LNCS 5988, pp.111-120, 2010

[6] Hermes, L., Buhmann, J.M., Feature Selection for Support Vector Machines, Proc. Pattern Recognition, Vol. 2, pp. 712-715, 2000

[7] Qi, B., Zhao, C., Youn, E., Nansen, C., Use of weighting algorithms to improve traditional support vector machine based classifications of reflectance data Optics Express, Vol. 19, No. 27, pp. 26816-26826, 2011

[8] Nguyen, M.H., De la Torre, F., Optimal Feature Selection for Support Vector Machines, Journal of Pattern Recognition archive, Vol. 43, No. 3, pp. 584-591, 2010

[9] Shen, K.-Q., Ong, C.-J., Li, X.-P., Wilder-Smith, E.P.V., Feature selection via sensitivity analysis ov SVM probabilistic outputs, Machine Learning, Vol. 70, pp. 1-20, 2008

[10] Grinblat, G.L., Lzetta, J., Granitto, P.M., SVM Based Feature Selection: Whey Are We Using the Dual?, Springer LNCS 6433, pp. 413-422, 2010

[11] Toshihiko Sakai, Sachio Hirokawa, Feature Words that Classify Problem Sentence in Scientific Article,The 14th International Conference on Information Integration and Web-based Applications & Services (iiWAS2012), pp. 360-367, 2012

[12] 酒井敏彦, 廣川佐千男, 手掛語による論文概要中の問題文の特徴付け, 第 2 回テキストマイニング・シンポジウム研究会, 信学技報 (IEICE Technical Report) Vol. 112, No. 196, pp. 73-78, 2012

[13] 廣川佐千男, 酒井敏彦, 少数例による学習における属性選択の効果について, 電気学会情報システム研究会データからの知識発見とその応用, 電気学会情報システム研究会資料, Vol. IS-12, No. 39-44.46-55, pp. 45-49, 2012

[14] Hearst, M. Clustering versus Faceted Categories for Information Exploration, in Communications of the ACM 49(4), pp. 59-61, 2006

[15] 廣川佐千男, 関隆宏, 安元裕司, 山田泰寛, 教員データに対する多面的検索システム, 電子情報通信学会技術研究報告. DE, データ工学 105(173), No. 67-72, 2005

[16] Andrew McCallum and Kamal Nigam, A comparison of Event Models for Naive Bayes Text Classification, Proc. AAAI-98 Workshop on Learning for Text Categorization, pp. 41-48, 1998

[17] Karl-Michael Schneider, On word frequency information and negative evidence in naive bayes text classification, Proc. EsTAL 2004, LNAI 3230, pp. 474-485, 2004

[18] Karl-Michael Schneider, A comparison of event models for Naive Bayes anti-spam e-mail filtering, Proc. EACL '03 Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, Vol1. pp. 307-314, 2003