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

2007－NL－180 (21)
2007／7／25

文字列を特徴量とし反復度を用いたテキスト分類

平田 勝大† 岡部 正幸‡ 梅村 恭司§

† § 豊橋技術科学大学 情報工学系 〒441-8580 愛知県豊橋市雲雀ヶ丘 1-1
‡ 豊橋技術科学大学 情報メディア基盤センター 〒441-8580 愛知県豊橋市雲雀ヶ丘 1-1
E-mail: † hira@ss.ics.tut.ac.jp, ‡ okabe@imc.tut.ac.jp, § umemura@tutics.tut.ac.jp

あらまし 標準的なテキスト分類では、文書の特徴として単語を使用するが、文字列を特徴として使用する研究もたくさんある。文字列を特徴としたテキスト分類では、文書の部分文字列数が膨大であることから、どの文字列を特徴として使用するのが良いかという問題がある。相互情報量に基づく条件付確率によって文字列を抽出すると効果があるという報告より、本研究では、文字列を反復度という統計量を用いることで抽出し、テキスト分類の性能の向上を目指す。

キーワード テキスト分類，特徴量抽出，Suffix Tree，Support Vector Machine

Extracting String Features with Adaptation for Text Classification

Katsuhiro Hirata† Masayuki Okabe‡ and Kyoji Umemura§

† § Information & Computer Sciences, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580 Japan
‡ Information and Media Center, Toyohashi University of Technology, 1-1 Hibarigaoka, Tenpaku-cho, Toyohashi-shi, Aichi, 441-8580 Japan
E-mail: † hira@ss.ics.tut.ac.jp, ‡ okabe@imc.tut.ac.jp, § umemura@tutics.tut.ac.jp

Abstract In traditional methods for text classification, words are used as a set of features for a document. However there are many string-based approaches. In the string-based approaches, the number of all substrings of documents would be extremely large and we don't know which substring is important for text classification. Previous research reports that using conditional probabilities based on mutual information for extracting features is effective. We reports extracting features with adaptation and that the method is more useful for text classification.

Keyword Text Classification, Feature Extraction, Suffix Tree, Support Vector Machine

1. はじめに

文書を内容ごとに分類することは、大量の文書を扱う情報社会では重要である。

標準的な機械学習によるテキスト分類は、文書の特徴量に文書中の単語を使用する[8]。しかしながら、テキスト分類において、文書の特徴量として文字列を使用する研究もたくさんある[3]。文字列を特徴量に使用することには次のような利点がある。

- 語形変化のような単語内での特徴と連語のような複数の単語による特徴の両方を特徴量として使用することができる。特に、ジャンル分類や著者分類のようなトピックのないテキスト分類において有用である。
- 日本語のような単語境界がわからない言語において、単語に分ける必要がない。単語を特徴量として使用する手法では、単語分割における誤りがテキスト分類の分類結果を悪化させてしまう。
- 句読点のような、英字や漢字以外の文字も特徴量に利用することができる。多くのスパムメールは"Viagra"を"V1a.gr.a"と置き換えたりするため、特にスパムフィルタリングにおけるテキスト分類において有用である。
- Webページとメール文書のような異なる種類の文書でも同様の手法で扱うことができる。これは、計算機内のさまざまな種類の文書を分類するデスクトップ検索ツールにおいて有用といわれている。

文書中の部分文字列を特徴として機械学習でテキスト分類を行うと効果があるという報告があるが、どの文字列を文書の特徴として使用するのが良いのかわからないという問題がある。

本研究では、反復度という統計量を用いてテキスト文書の特徴となる文字列を抽出する。この文字列を用いて、Support Vector Machine(SVM)[9]でテキスト分類を行う。さらに、英語のテキスト分類において、条件付確率を用いて抽出した文字列を特徴とした時の分類結果と比較し、反復度を用いることで分類結果が改善したことを報告する。

2. 関連研究

2.1. 生成アプローチ

文字列によるテキスト分類での生成アプローチは、それぞれの文書がマルコフ連鎖モデル[1]によって生成されると仮定する。各クラス c に対応するマルコフ連鎖モデル M_c を学習文書から学習する。テスト文書 d が与えられたとき、マルコフ連鎖モデル M_c から文書 d がクラス c に分類される確率を計算し、この確率が最大になるクラスに分類する。

マルコフ連鎖モデルには、固定長と可変長があり、固定長 n の場合は、n グラムモデル[5]とよく呼ばれる。N グラムモデルは、情報検索などにおいて広く利用されている。自然言語処理での n グラムは単語単位での連鎖を扱うことが多いが、本研究では文字単位の n グラムに注目する。Peng ら(2004)[4]は、さまざまなテキスト分類において文字単位の n グラムモデルを利用している。

2.2. 識別アプローチ

生成アプローチとは異なり、識別アプローチは生成モデルを仮定せず、分類のための関数を直接生成する。特に、単語に基づくテキスト分類では、識別アプローチのひとつである SVM が生成アプローチよりもよい性能を示している[10,11]。

文字列に基づく識別アプローチでは、文書に対する部分文字列が膨大であるという問題があったが、Zhang ら(2006)[3]がこの問題を解決している。この方法を次に示す。

3. 文字列を用いたテキスト分類

文書中のすべて部分文字列の数は、文書の文字数 n の時 $n(n-1)$個と非常に膨大であり、文書の特徴として用いるには多すぎる。Zhang ら(2006)[3]は、Suffix Tree を用いて文字列を統計的に等しい集合でまとめ、さらに、条件付確率等を用いた条件によって特徴量に用いる文字列の数を大幅に減らす。これらの特徴量を使用し、SVM によってテキスト分類を行う。

3.1. Suffix Tree

文字列を扱うデータ構造の一つに suffix tree[2]がある。このデータ構造は、文字列 S 中の接尾辞を木構造に構築したデータ構造である。接尾辞とは、文字列 S の最後尾の文字を e とすると、文字列 S 中の文字 x から e までの文字列からなる文字列 S の部分文字列である。文字列 dbababc$のすべての接尾辞を図 1 に示す。

d
db
dba
dbab
dbaba
dbabab
dbababc
dbababc$

図 1.文字列 dbababc$のすべての接尾辞

suffix tree は、根から葉への有向グラフであり、葉でない節点は 2 つ以上の子を持つ。枝は、空でない部分文字列が割り当てられ、根から葉までの経路が一つの接尾辞を表す。図 2 に文字列 dbababc$に対する suffix tree を示す。

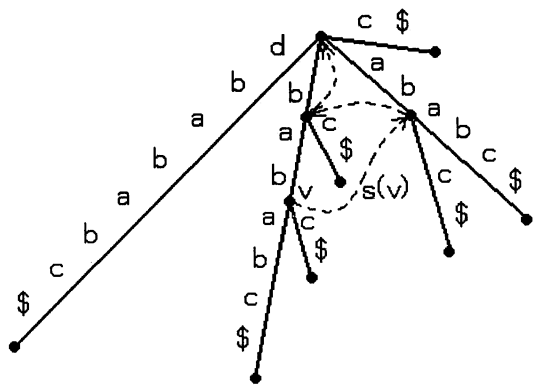

図 2. 文字列 dbababc$の Suffix Tree

suffix tree の葉は、文字列 S の長さが n の時、n 個である。文字列 S の部分文字列 P が S 中で出現する頻度は、根からの経路に割り当てられた文字列が P で終わる節点 v を根とする部分木の葉の数と等しい。

x を 1 文字、α を 0 文字以上の文字列として、根と葉でない節点 v への根からの経路が文字列 $x\alpha$ である

とき、根から経路が α である節点 u へのリンクを suffix link という。図 1 に節点 v からの suffix link $s(v)$を示す。suffix link によって 1 文字短い接尾辞をたどることができ、suffix tree の作成や最大共通部分文字列に利用される。

3.2. 部分文字列集合

テキスト分類に用いるほとんどの機械学習は、特徴量として次のような統計量を用いる[12]。ただし、$|D|$はコーパス D の文書数を表す。

- 文字列 t が文書 d に出現する頻度である $tf(t,d)$
- コーパス D 中で文字列 t が出現する文書数 $df(t,D)$
- $tfidf(t,d) = tf(t,d) \cdot \log(|D|/df(t,D))$

したがって、これらの統計値が等しい文字列は機械学習の特徴量としては、区別する必要はない。そこで、これらの統計的に等しい文字列をひとつの特徴としてまとめてしまうことを考える。コーパス D 中の部分文字列 t の出現頻度は、Suffix Tree 中では根からの経路に割り当てられた文字列が t で終わる節点 v の葉の数と等しい。以上より、統計的に等しい部分文字列集合を $SG(v)$と定義し、$freq(SV(v))$をコーパス D における部分文字列集合の出現頻度と定義する。

3.3. 条件付確率による特徴量抽出

次の 5 つの条件によってさらに特徴量に用いる $SG(v)$を減らす。

(1)$freq(SG(v))$が最小頻度 l 未満の $SG(v)$は除く
(2)$freq(SG(v))$が最大頻度 m 以上の $SG(v)$は除く
(3)節点 v の子が最小の子の数 b 未満の $SG(v)$は除く
(4)節点 v の親が u であるとき、$freq(SG(v))/freq(SG(u))$が最大親子条件付確率 p 以上である $SG(u)$は除く
(5)節点 v の suffix link が u を指しているとき、$freq(SG(v))/freq(SG(u))$が最大 suffix-link 条件付確率 q 以上である $SG(u)$は除く

条件(3)は $SG(v)$中の文字の種類に相当し、文字列の文脈依存を反映している。含まれる文字の種類が多いほうがより文脈から独立な文字列であり、文字の種類が少ない文字列より文章の特徴として有用である。条件(4),(5)は $SG(v)$と $SG(u)$との相互情報量[6]に比例する確率であり、冗長性の高い部分文字列集合を除く。2 つの文字列の相互情報量が高ければ、2 つのうちどちらか 1 つを特徴として使用すれば十分である。以上が Zhang ら(2006)の手法である。

4. 反復度による特徴量抽出

本研究では条件(3),(4),(5)の代わりに反復度[7]を用いた条件によって特徴量に用いる文字列を減らすことを提案する。反復度とは、文書中で繰り返す文字列は文書の特徴を表す重要な文字列であるという仮定に基づく統計量であり、コーパス D 中に文字列 t が 2 回以上出現する文書数を $df_2(t,D)$ とすると、反復度 $adapt(t,D)$は次のように定義される。

$$adapt(t,D)= df_2(t,D)/df(t,D)$$

反復度は、表 1 に示すように語の境界において大きく減少する統計量であり、キーワードの自動抽出などに使用される統計量である[13]。

表 1.語の境界における反復度の変化

文字列	Df	df_2	反復度
メ	52424	22324	0.426
メデ	4632	2200	0.475
メディ	4580	2178	0.476
メディア	4434	2132	0.481
メディアを	560	88	0.157
メディアを用	83	12	0.145
メディアを用い	83	12	0.145
メディアを用いた	64	6	0.094

反復度による抽出の条件には次のような条件を用いる。ただし、$adaptation(SV(v))$はコーパス D における部分文字列集合の反復度とする。

- $adaptation(SV(v))$が最小反復度 a 未満の $SV(v)$は除く

本研究では、反復度を文字列の重要度として使用する。条件付確率では、統計的に冗長な文字列を除いていたが、反復度を用いる場合は、ひとつの文書の中でほとんど一度しか出現しない文字列を除くことになる。

5. Support Vector Machine

Support Vector Machine(SVM)とは、2 つのクラスを識別する識別器を構成するための学習法である。学習データを用いて、2 つのクラスを線形分離し、分離超平面と最も近い学習データとの距離が最大になるようなモデルを学習する。この手法により、線形分離できるデータにおいては、試験データにおいても高い識別性能を持つ学習手法の一つである。

文書の特徴量ベクトル $\mathbf{X}=\{x_1,x_2,\ldots,x_n\}$と文書が属するクラス y=1 または y=-1 が与えられた時、SVM の分

離平面は次のような式で与えられる。ここで $\mathbf{W}=\{w_1,w_2,\ldots,w_n\}$は重みベクトル、$b_0$はバイアス定数である。

$$\mathbf{W}\cdot\mathbf{X}+b_0=0$$

分離平面との距離の最小値を 1 とすると分離平面との距離は、次のような式で計算される。ただし、$\|\mathbf{W}\|=\sqrt{(\mathbf{W}\cdot\mathbf{W})}$であり、$y$=1 のとき $g(\mathbf{X})\geqq 1$、y=-1 のとき $g(\mathbf{X})\leqq -1$ である。

$$g(\mathbf{X})=(\mathbf{W}\cdot\mathbf{X}+b_0)/\|\mathbf{W}\|$$

したがって、マージン最大化は次のような次のような問題を解けばよい。

制約条件:$y(\mathbf{W}\cdot\mathbf{X}+b_0)-1\geqq 0$

$G(\mathbf{W})=\|\mathbf{W}\|^2/2$ を最小化する

この問題を解くことによって $\mathbf{W},b_0$ を設定し、識別器を学習する。

6. 実験

実験コーパスには Reuter-21578 テキストコレクションの" ModApte"学習・テストセットを使用し、SVM には SVM の実装のひとつである SVMlight を利用し、トピック分類を行う。

使用文書は、本文のうち TITLE タグと BODY タグのついた文書とし、Reuter-21578 の文書が含まれるトピックの上位 10 トピックにテキスト分類する。ただし、各文書は複数のトピックに属することがある。学習には、学習用文書セットの全 9,603 文書を使用し、テストには学習文書とは異なるテストセットの全 3,299 文書を使用する。表 2 に学習セットおよびテストセットにおける上位 10 トピックの正例の文書数を示す。

前処理として各文書のアルファベット以外の文字を空白に変換し、2 文字以上空白が続く場合は空白 1 文字に変換する。

特徴量に用いる文字列には、条件付確率による特徴量抽出のパラメータは、l=80, h=8000, b=8, p=0.8, q=0.8 (3.3 節参照)として 8,438、反復度による特徴量抽出のパラメータは、l=80, h=8000, a=0.3(4 章参照)として 7,099 文字列を抽出した。反復度パラメータは学習データでの学習が最も良くなる値を調べ決定している。ただし、どちらの条件においても空白で始まる文字列は特徴量に使用する文字列からは除く。

表 2.上位 10 トピックの正例文書数

トピック	学習セット	テストセット
earn	2877	1087
acq	1650	719
money-fx	538	179
grain	433	149
crude	389	189
trade	369	117
interest	347	131
wheat	212	71
ship	197	89
corn	181	56

SVM の特徴量には、抽出した文字列の *tfidf* を使用し、線形カーネルによってテキスト分類を行う。結果の評価には次の 3 つの尺度を利用する。

$$\text{適合率}=\frac{\text{トピックに属すると分類した文書の正解文書数}}{\text{トピックに属すると分類した文書数}}$$

$$\text{再現率}=\frac{\text{トピックに属すると分類した文書の正解文書数}}{\text{コーパス中の正解文書数}}$$

$$F\text{値}=\frac{2\times\text{適合率}\times\text{再現率}}{\text{適合率}+\text{再現率}}$$

以上の実験結果を図 4,5,6、表 3 に示す。図 4,5 に注目すると、トピック上位 2 つの earn,acq および trade については条件付確率を用いた特徴量によるトピック分類のほうが適合率、再現率ともによいが、これら以外の 7 トピックについては、反復度を用いた特徴量によるトピック分類のほうが結果が良くなる場合が多く、特に再現率が大きく改善されるという結果になった。図 6、表 3 の F 値でも同様に上位 2 トピックの earn,acq および trade では条件付確率を用いたほうが F 値が高くなったが、これら以外のトピックでは反復度を用いたほうが改善され、全体平均では 3.74%改善した。

7. 考察

前節の実験において、学習文書とテスト文書に同じ文書集合を用いてみると、F 値の全体平均は、条件付確率を用いた特徴量では 95.17%、反復度を用いた特徴量では 92.87%となり、条件付確率による特徴量のほうが全体的に F 値が高くなる。学習文書とテスト文書に異なる文書集合を用いる本来の評価では、前節で説明したように反復度による特徴量のほうが F 値が高いことから、条件付確率を用いた特徴量では、反復度を用いた場合に比べ、過学習してしまう傾向があるのではないかと考えられる。また、学習セットとトピックに含まれる文書数に注目すると、上位 2 つのトピック

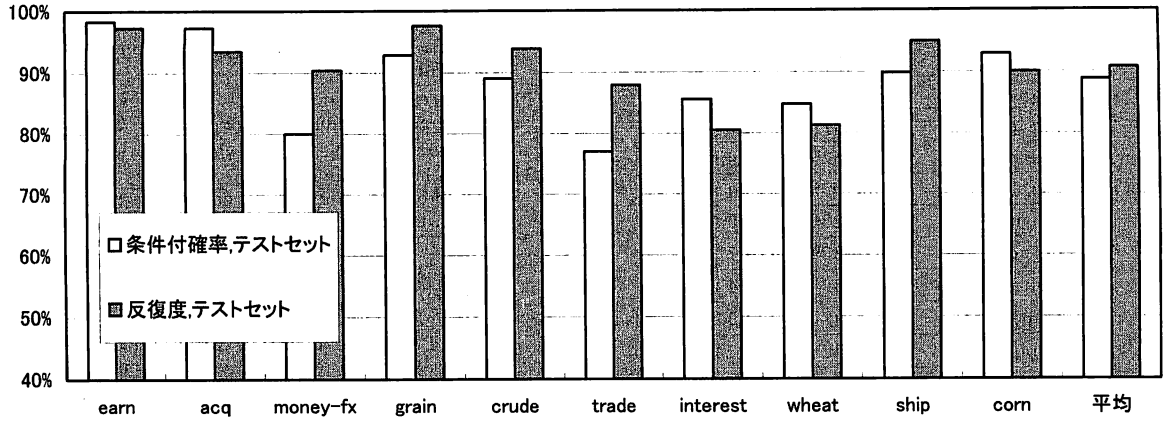

図 4.適合率

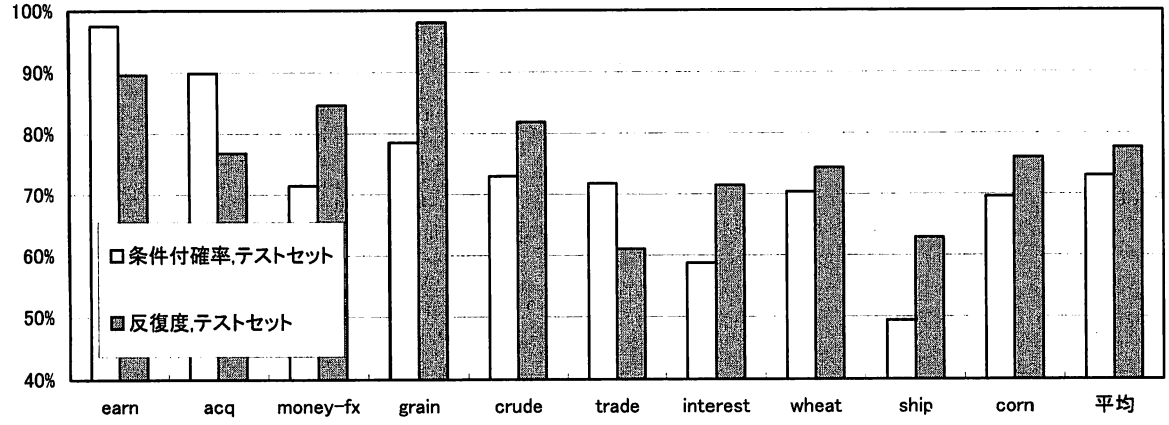

図 5.再現率

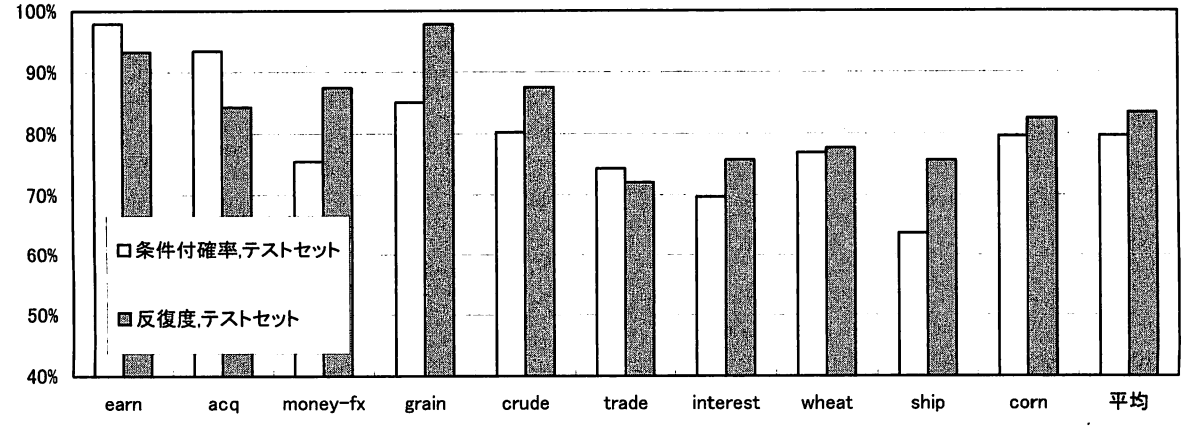

図 6.*F* 値

表 3.F 値

	テストセット	
	条件付確率	反復度
earn	97.92%	93.27%
acq	93.42%	84.32%
money-fx	75.52%	87.44%
grain	85.09%	97.84%
crude	80.23%	87.46%
trade	74.33%	72.07%
interest	69.69%	75.74%
wheat	76.92%	77.68%
ship	63.77%	75.68%
corn	79.59%	82.44%
平均	79.65%	83.39%

earn,acq においては条件付確率による特徴量のほうがテストセットによる分類において F 値が高く、これら以外のトピックにおいては、trade を除いて反復度による特徴量のほうが F 値が高い。したがって、反復度は学習文書が少ない場合に分類に有効な特徴量と抽出することができる統計量であるといえる。

8. まとめ

テキスト分類において、条件付確率を用いて文書の特徴を抽出する代わりに、反復度を用いて抽出した。テキスト分類の結果を比較することで、反復度を用いた特徴量の方が条件付確率を用いた特徴量抽出よりもテキスト分類においてよりよい結果になることを報告した。これにより、反復度がテキスト分類の文書の特徴を抽出する上で有用な統計量であることを示した。

9. 今後の課題

Reuter-21578 テキストコレクション以外のコーパスにおいても反復度を用いた特徴量が有効であるかを調べ、テキスト分類において反復度が文書の特徴量抽出に有用な統計量であるかを調べる必要がある。また、単語の区切りのない日本語のような言語においても、テキスト分類において反復度による部分文字列の特徴量抽出が有効かについても調べたいと考えている。

謝辞

この研究は住友電工情報システム(株)との共同研究の成果であり、戦略的情報通信開発推進制度(SCOPE)の課題「実空間情報処理のためのインターユビキタスネットワークの研究」に使用する予定です。

文　献

[1] C. Manning and H. Schutze, Foundations of Statistical Natural Lnaguage Processing, MIT Press, Cambridge, 1999

[2] D. Gusfield, Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology, Cambridge University Press, USA, 1997

[3] D. Zhang, and W. S. Lee, "Extracting Key-Substing-Group Features for Text Classification, "In Proceedings of the 12th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, no. 1150455, pp.474-483, Philadelphia, USA, Aug.2006

[4] F. Peng, D. Shuurmans and S. Wang, Augmenting Naïve Bayes text classifier with statistical language models, Information Retrieval, Volume 7, Numbers 3-4, pp.317-345, Sept.2004

[5] J. Goodman, A bit progress in language modeling, extended version, Technical report, Microsoft Research, MSR-TR-2001-72, pp.403-434, Oct.2001

[6] K. W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography, " In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, no.church89word, pp.76-83, Vancouver, Canada, 1989

[7] Kenneth W. Church, Empirical Estimates of Adaptation: The chance of Two Noriegas is closer to p/2 than p^2, In Proceedings of 18th International Conference on Computational Linguistics, Volume 1, pp.180-186, 2000

[8] N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. Word-sequence karnels. Journal of Machine Learning Research, Volume 3, pp.1059-1082, 2003.

[9] N. Cristianini and J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Camridge, 2000

[10] S. Dumais, J. Platt, D. Hecherman and M. Sahami, "Inductive learning algorithms and representations for text categorization, "In Proceedings of the 7th ACM International Conference on Information and Knowledge Management, no.288651, pp.148-155, Bethesda, USA, Nov.1998

[11] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, In Proceedings of the 10th European Conference on Machine Learning Research, no.joachims98text, pp.137-142, Chemnitz, Germany, Nov.1998

[12] T. Mitchell, Machine Learning, Eric M. Munson, McGraw Hill, 1997

[13] Y. Takeda and K. Umemura, "Selecting indexing strings using adaptation, "Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, no.TakedaU02, pp.11-15, Tampere, Finland, Aug.2002