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

2005－ＮＬ－168（５）
2005／7／22

構造と特徴選択に着目した電子メールの分類手法の提案
－From フィールドと Jeffreys Perks を用いたナイーブベイズ分類－

川前徳章

ＮＴＴ情報流通プラットフォーム研究所 〒180-8585　東京都武蔵野市緑町 3-9-11
E-mail:　kawamae.noriaki@lab.ntt.co.jp

あらまし　本研究は、E-mail をあらかじめユーザが設定したフォルダに自動分類する E-mail フィルタリング手法について提案と比較検討を行った。著者はナイーブベイズに基づいた E-mail のフィルタリングの検討に加え、E-mail の構造の特徴であるフィールドの違いに着目する手法を提案する。多くの研究でも E-mail を従来のテキストフィルタリングの手法を用いて行ってきたが、E-mail は新聞や Web など従来のテキストフィルタリングで扱われてきたテキストと異なり、（1）一テキストファイルに含まれる単語の数が少ないこと、（2）ファイル内にフィールドの構造を持つといった特徴がある。これらの特徴を考慮し、著者はナイーブベイズに基づいたテキストフィルタリングを E-mail のフィルタリングのために最適化を実現する手法を提案する。その中で、我々は E-mail のフィールド、単語の出現頻度の推定に Jeffreys Perks、単語に対して DF(Document Frequency)で重みを付けてその値で特徴選択を行うことを提案した。データとして ENRON CORPUS を用い、フィールドと単語の重み付けの違いの比較実験を行い、提案手法の有効性を確認した。

キーワード　テキストフィルタリング，E-mail　Filtering，Naïve Bayes，Jeffreys Perks，ENRON CORPUS

E-mail Filtering Focusing on E-mail Field and Feature Selection
－Naïve Bayes Filtering with From Field and Jeffreys Perks－

Noriaki KAWAMAE

NTT Information Sharing Platform Laboratories　3-9-11,Midori-cho Musashino-shi Tokyo 180-8585 Japan
E-mail:　kawamae.noriaki@lab.ntt.co.jp

Abstract　In this paper, we discuss the methods of email filtering, to assign emails to the appropriate folders based on the user's classification strategy. Our contribution is to focus on the E-mail field with comparative study of email filtering in the framework of text filtering based on Naïve Bayes. Although most previous works deal with email by the conventional text filtering method as the same way, email has the different characteristics comparing with other kinds of texts such as news and web pages so on. Characteristics of email are assumed as (1) the small number of terms included in it, and (2) its structure of fields. Considering these characteristics, we propose the methods to deal with email structure, estimate the term probability and the feature selection using the term weighting method to make the text filtering based on Naïve Bayes optimized for email filtering. The experiment on a huge mail archive, the ENRON corpus, shows that the methods using Jeffreys Perks low improved about 10 % of prediction accuracy for variety of known feature selection methods.

Keyword　**Text　Filtering,　E-mail　Filtering,　Naïve Bayes,　Jeffreys Perks,　ENRON CORPUS**

1. はじめに

本研究の対象分野は個人の E-mail を、あらかじめ設定したフォルダに自動分類する E-mail Filtering である。自動分類の目的は分類を正確に、各個人の意図の反映を実現するルールの獲得である。既に E-mail は多くのユーザに用いられ、各個人あたりが送受信する E-mail の数も増えてきていることから、E-mail の処理は個人にとって負担となり、分類の自動化に対してのニーズが高まっている。例えば、迷惑メールのフィルタリングのニーズがある。

E-mail は従来のテキストフィルタリングで対象とされてきた新聞記事や Web 文書とは異なった次のような特徴を持っている。

・一ファイル（メール）当たりに含まれる単語の数が少ないこと。
・フィールドの構造を持っていること。

・フィルタリングのルールがあいまいなこと。
これらの特徴がフィルタリングに対して与える影響を考える。単語の数はフィルタリングの精度に関係があり、少ない場合はそれを低下させる恐れがある。一方で、メールが共通に持つフィールドはフィルタリングの精度を上げる可能性がある。最後の特徴はルールがメールの内容だけでなく、各個人の興味やニーズに依存するために主観的なものとなり、個人間で共有が難しく、一般化できない場合が多い。このようなメールが持つ特徴から既存のテキストフィルタリングの手法では E-mail の自動分類の精度を向上できないと考えられる。

この論文の目的は E-mail の自動分類の精度を上げるテキストフィルタリング手法の提案である。この中で我々は、先に述べたテキストフィルタリングの問題点がメールの特徴を反映していないことに着目した。そして
・メールのフィールド情報を活用する
・単語数が少ないため、出現しない単語の推定を行う。
・あいまいなルールでも大局的に安定した重み付けにより、単語を選択する。
ことでこの問題の解決を図った。これらの検証に、ENRON CORPUS [1] を用い、その有効性を確認した。

この論文は次の構成を取っている。第二章では既存のテキストフィルタリングで用いられてきた特徴選択を取り上げ、その問題点について論じる。第三章は Naïve Bayes と MAP(Maximum a posteriori)を用いたテキストフィルタリングと単語の重み付け、第四章では実験の概要とその結果、最後に考察を行い、結論を述べる。

2. 既存研究

本稿に関係ある既存研究は Naïve Bayes のテキストフィルタリングへの応用と E-mail Filtering を扱った分野である。テキストフィルタリングの精度を向上するために特徴選択の利用が考えられる。特徴選択はフィルタリングの精度を損なわずに、分類ルールあるいはモデルを構成する単語の数を減らすことで、分類ルールやモデルを簡潔かつ明瞭、計算コストの低減や予測精度を上げる [2,3] と期待されている。Yang と Pederson [4] はテキストフィルタリングに用いられる種々の特徴選択の比較を行い、E-mail がモデルに追加された数として定義される statistics word's age に着目した頻度が有効 [5] であることを述べている。
既存の E-mail Filtering の研究に関する問題点は、その多くが非常に小さなデータセットに基づいており、E-mail のフィールドの有効な利用についてあまり検討されていなかったことにある。実際、多くの研究が Naïve Bayes に基づいた E-mail を対象とした実験を行っているが、その実験で用いられた E-mail の数は少なく、公には利用できないものである [6]。

本稿は既存研究の問題を踏まえ、E-mail の特徴を考慮したナイーブベイズの特徴選択と単語の重み付け手法の比較検討を一般に利用可能な大規模なデータセットを用いて行う。

3. ナイーブベイズに基づいた分類

3.1. Naïve Bayes

ナイーブベイズ分類は、テキストに出現する単語は文脈と位置に関して独立であるという仮定に基づいた確率モデルである。この仮定により、全てのテキストファイル d_j はパラメタ θ から構成される混合モデルの確率分布として与えることが出来る。混合モデルはいくつかの要素から構成され、その要素がラベル t_k に相当し、パラメタ θ によって構成される。テキストファイル d_j は、まず構成要素である $P(t_k|\theta)$, を選択し、それらの $P(d_j|t_k;\theta)$. との積により生成する。全てのラベルが与えられることで、テキストファイルの尤度関数を次のように定義することが出来る。

$$P(d_j|\theta)=\sum_{k=1}^{O}P(t_k|\theta)P(d_j|t_k,\theta) \quad (1)$$

ここで O はカテゴリの数を示している。ナイーブベイズではテキストファイル d_j をファイルに出現した単語のリストとして考え、更にそれらの単語は同一ファイルの中で独立に出現すると仮定を置いている。その仮定により、ファイルの出現確率の尤度の式（1）は次のように書き換えられる。

$$P(d_j|\theta)=\sum_{k=1}^{O}P(t_k|\theta)\prod_{h=1}^{N}P(w_h|t_k,\theta) \quad (2)$$

ここで N は単語の異なり数、w_h はテキストファイル d_j に含まれる単語を示している。この式を用いて各テキストファイルのラベルに対する条件付き確率はベイズの公式によって 次のように定義することが出来る。

$$P(t_k|d_j)=\frac{P(t_k|\theta)\prod_{h=1}^{N}P(w_h|t_k,\theta)}{\sum_{k=1}^{O}P(t_k|\theta)\prod_{h=1}^{N}P(w_h|t_k,\theta)}\propto P(t_k|\theta)\prod_{h=1}^{N}P(w_h|t_k,\theta) \quad (3)$$

この確率 $P(t_k|d_j)$ の値を用いて，我々はテキストファイルの所属するラベルを判定することが出来る。これがナイーブベイズを用いたフィルタリングである。

次に式（3）を構成するパラメタを推定する方法を紹介する。本研究ではこの推定に MAP(Maximum a posteriori)を用いる。MAP は次のようにモデル化できる。

$$\hat{\Theta}_{MAP} = \arg\max P(\Theta|D)$$
$$= \arg\max\{L(\Theta;D) + \log P(\Theta)\} \quad (4)$$

ここで D は利用する学習データ、$L(\Theta;D)$ は対数尤度、Θ はこのモデルの未知パラメタであり、Θ = { $P(t_k);k=1,\ldots O$,$P(w_h|t_k);h=1,\ldots,n$}定義する。$P(\Theta)$ はすべてのパラメタの事前分布であり、Dirichlet distribution を用いて次のように定義できる。

$$P(\Theta) \propto \prod_{k=1}^{O} P(t_k)^{\alpha-1} \prod_{h=1}^{n} P(w_h|t_k)^{\alpha-1} \quad (5)$$

ここでαはハイパパラメタである。式の（4）と（5）よりこの推定方法の目的関数は次のように書くことができる。

$$Q(\Theta) = L(\Theta;D) + (\alpha-1)\sum_{k=1}^{O}\sum_{h=1}^{N} P(t_k|\theta)P(w_h|t_k;\theta) \quad (6)$$
$$= \log\left(\sum_{k=1}^{O} P(t_k|\theta)\prod_{h=1}^{N} P(w_h|t_k,\theta)\right) + (\alpha-1)\sum_{k=1}^{O}\sum_{h=1}^{N} P(t_k|\theta)P(w_h|t_k;\theta)$$

EM algorithm の枠組みでは、Q 関数は次のように定義できる。

$$Q(\theta|\bar{\theta})$$
$$= \log(P(\theta)) + \sum_{j=1}^{M}\sum_{k=1}^{O} P(t_k|d_j;\bar{\theta}) \times \log\left(P(t_k|\theta)P(d_j|t_k;\theta)\right) \quad (7)$$
$$= (\alpha-1)\sum_{k=1}^{O}\sum_{h=1}^{N} P(t_k|\theta)P(w_h|t_k;\theta)$$
$$+ \sum_{j=1}^{M}\sum_{k=1}^{O} P(t_k|d_j;\bar{\theta}) \times \log\left(P(t_k|\theta)\prod_{h=1}^{N} P(w_h|t_k;\theta)^{N(w_h,d_j)}\right)$$

ここで $N(w_h,d_j)$はテキストファイル d_j に含まれる単語 w_h の数、λはハイパパラメタを示している。Q 関数の定義に従えば、ラグランジュの未定係数法によって次の関数を得る。

E-step

$$P(t_k|d_j,\theta) = \frac{P(t_k|\theta)\prod_{h=1}^{N} P(w_h|t_k,\theta)}{\sum_{k=1}^{O} P(t_k|\theta)\prod_{h=1}^{N} P(w_h|t_k,\theta)} \quad (8)$$

M-step

$$\hat{\theta}_{w_h|t_k} = P(w_h|t_k)$$
$$= \frac{(\alpha-1) + \sum_{j=1}^{M} P(t_k|d_j,\theta) * N(w_h,d_j)}{(\alpha-1)|N| + \sum_{h=1}^{N}\sum_{j=1}^{M} P(t_k|d_j,\theta) * N(w_h,d_j)} \quad (9)$$

$$\hat{\theta}_{t_k} = P(t_k) = \frac{(\alpha-1) + \sum_{j=1}^{M} P(t_k|d_j,\theta)}{(\alpha-1)|O| + |M|} \quad (10)$$

この式においてMはテキストファイルの総数である。

3.2. 単語の重み付けと特徴選択

次に今回の特徴選択において利用する単語の重み付けの式を示す。

RIDF (Residual IDF)

$$RIDF_h = \log\frac{M}{M_h} + \log\left(1 - e^{-\frac{F_h}{M}}\right) \quad (11)$$

ここで M_h は単語 w_h を含むテキストファイルの数であり、F_h は単語 w_h が全ファイル中で出現する数である。

MI (Mutual Information)

$$MI_h = \sum_{k=1}^{O}\sum_{x_h=0,1} P(x_h,t_k)\log\frac{P(x_h,t_k)}{P(x_h)P(t_k)} \quad (12)$$

ここで x_hは単語 w_h, の存在を示すフラグ、$P(x_h|t_k)$ はラベル t_k での x_h の出現確率、 $P(x_h)$は x_h の出現確率である。

TM (Term Entropy)

$$TM_h = 1 - \frac{H(w_h)}{\log|M_h|} = 1 + \frac{\sum_{j=1}^{|M_h|} P(w_{hj}|w_h)\log P(w_{hj}|w_h)}{\log|M_h|} \quad (13)$$

ここで $P(w_{hj}|w_h)$ は単語 w_h のファイル d_j での出現確率、$|M_h|$ は単語 w_hを含むファイルの総数である。

DF (Document frequency)

$$DF_h = \frac{M}{M_h} \quad (14)$$

IG (Information Gain)

$$IG_{kh} = P(x_h)\sum_{x_h=0,1} P(x_h,t_k)\log\frac{P(x_h,t_k)}{P(x_h)P(t_k)} \quad (15)$$

IG を全てのラベルに対して定義できるように以下のように二通りに拡張する。

$$IG\_e_h = \sum_{k=1}^{O} P(t_k)IG_{kh} \quad (16)$$

$$IG\_\max_h = \max_{k=1}^{O}\{IG_{kh}\} \quad (17)$$

RE (Relative Entropy)

$$RE_h = KL\left(P(t|w_h)\middle\|P(t)\right) = \sum_{k=1}^{O} P(t_k|w_h)\log\frac{P(t_k|w_h)}{P(t_k)} \quad (18)$$

Cross Entropy

$$CR_h = \sum_{k=1}^{O} P(t_k, w_h)\log\frac{P(t_k|w_h)}{P(t_k)} = P(w_h) * RE_h \quad (19)$$

Odds Ratio

$$Odds_{kh} = \log\frac{P(w_h|t_k)\left(1-P(w_h|\neg t_k)\right)}{P(w_h|\neg t_k)\left(1-P(w_h|t_k)\right)} \quad (20)$$

Odds Ratio を全てのラベルに対しての定義できるように、IG 同様に次の二通りに拡張する。

$$Odds\_e_h = \sum_{k=1}^{O} P(t_k) Odd_{kh} \quad (21)$$

$$Odds\_\max_h = \max_{k=1}^{O}\{Odd_{kh}\} \quad (22)$$

Relative Risk

$$RR_h = \log\frac{P(t_k|w_h)}{P(t_k|\neg w_h)} \quad (23)$$

Relative Risk を全てのラベルに対しての定義できるように、Odds *Ratio* 同様に次の二通りに拡張する。

$$RR\_e_h = \sum_{k=1}^{O} P(t_k) RR_{kh} \quad (24)$$

$$RR\_\max_h = \max_{k=1}^{O}\{RR_{kh}\} \quad (25)$$

Weight of EvidTxt

$$WE_h = \sum_{k=1}^{O} P(t_k) * P(w_j) * \left|\log\frac{P(t_k|w_h)\left(1-P(t_k)\right)}{P(t_k)\left(1-P(t_k|w_h)\right)}\right| \quad (26)$$

TF (Term Frequency)

$$TF_{h|j} = N(w_h, d_j) \quad (27)$$

4. 実験

4.1. 実験の概要

著者は今回の実験で ENRON CORPUS[1]を用いた。著者が利用した版は、ユーザの数が 149 名、フォルダの数が 2085、そしてメールの数が 32,140,492 通あった。この実験の目的はテキストフィルタリングの精度向上であり、E-mail のフィールド、推定方法及び特徴選択手法を大規模な E-mail のベンチマークのデータに対して比較及び検証することである。これが ENRON CORPUS を用いた理由である。著者はこの実験データからまず各ユーザの各フォルダからランダムにメールを一通選択してテストデータ、残ったメールを学習データとした。ここで[1]に倣い、各ユーザの "all_documents", "discussion_threads" そして "notes_box"のフォルダは両データから除外した。全てのメールをパージングして単語と共にユーザ ID,メール ID,フォルダ ID、メールの構造と出現回数のリストで管理する。これらのリストを生成した後に、学習データから単語の出現確率を推定し、先に示した各手法で単語に重み付けを行う。

これらのデータを用いた実験の評価には Macro Average を利用した。3.1 に示したナイーブベイズに基づいたフィルタを作成し、そこで用いる単語は利用する単語の重み付けやメールのフィールドを変化させて比較を行う。ここで用いたメールのフィールドは" From", "To", "Subject", "CC" と"Text field"である。この結果、" From"には 38178 種の単語、"Subject"には 230978 種の単語、"Text field"には 2358471 種の単語があった。

4.2. メールのフィールドと推定手法の違いに着目したメールの分類精度の比較

表 1 にメールのフィールドと推定手法を組み合わせ毎の予測精度の macro average を示す。macro average は各ユーザの各カテゴリに正確に分類されたメールの割合である。テストデータ内の各メールに含まれる単語はその単語が含まれていたフィールドの違いと特徴選択によって選択され、ナイーブベイズモデルに取り込まれ、出現確率の推定値を用いて算出された事後確率の値により分類され、その分類精度を測定したものである。今回の実験で三つの異なったフィールドの組み合わせ、推定方法として従来手法、Laplace と Jeffreys の三種類の推定方法を利用したので、計 21 通りの組み合わせにおける分類精度の結果が得られた。

表１．フィールドと推定方法の組み合わせ毎のナイーブベイズの macro average

Email フィールド			推定方法	推定方法	
From	Subject	Text	従来手法	Laplace	Jeffreys
○	-	-	27.1%	27.3%	27.9%
-	○	-	15.0%	14.8%	18.3%
-	-	○	22.1%	22.1%	25.0%
○	○	-	11.8%	11.9%	14.9%
○	-	○	22.5%	22.5%	25.3%
-	○	○	22.4%	22.4%	25.2%
○	○	○	22.7%	22.7%	25.4%

この表の結果から、メールのフィールドのうち”From”のみに含まれる単語を用いた分類が他のフィールドの組み合わせに含まれたものを用いるよりも全ての推定方法でいずれも高い精度を示したことが分かる。一方で” From”と“Subject”を組み合わせが最も精度が低いことも分かる。この結果はヘッダ情報のみを用いたものが一番精度が低いという Diao の結果［7］と一致する。ヘッダ情報は” From”と“Subject”を組み合わせたものである。推定方法に関してこの表の結果からは、Jeffreys による推定が他の推定方法を用いるより分類精度が高いことを示している。組み合わせとして”From”から抽出した単語、つまり送信者のメールアドレスとそれらの推定に Jeffreys を用いたものが 27.9%の精度を示したのに対し、同じ推定方法を用いても、メール内の全てフィールドから抽出した単語を用いた場合の精度は 25.4%であった。従って、この表の結果から” From”と Jeffreys による推定が最も精度が良いことが分かる。

4.3. 特徴選択と推定方法の違いによる分類精度の比較

次の特徴選択手法の比較実験では、3.2 に示した重み付けの式の値によって単語を選択し、ナイーブベイズモデルを構築し、その分類精度を比較した。

表２は 13 種の重み付けの式を用いた特徴選択に対してナイーブベイズの予測精度の macro average を比較したものである。特徴選択は 4.2 の結果を受け、まず” From”からメールアドレス、それ以降は “To”, “Subject”, “CC” と“Text field”のフィールドの区別を考慮せずに、単語の重みによって選択を行う。分類精度は抽出した単語によって構成したナイーブベイズモデルの事後確率によって分類した時のものである。この表では各特徴選択手法で最も高い値を示している。この実験で我々は学習データを各特徴選択の値によって 20 に分割し、それらを利用する単位を 1 から 19 まで増やす。言い換えれば学習データら利用するデータの割合を 0.05 から 0.05 ずつ増やし、最終的には全体の 0.95 を利用する。この表でも先の実験同様に Jeffreys が他の推定方法よりも高い精度を示していることが分かる。

図 1 は特徴選択手法の macro average の上位四件の利用データの割合における変化を示している。ここでの推定方法はこれまでの実験で最も分類精度の良かった Jeffreys を用いている。この図は DF が他の手法よりも利用データの利用割合が小さい段階で高い値を示している。この結果は”上位 2～5%の単語の利用が最適”という結果［15］と一致しないが、同じ報告の”テキストフィルタリングにおいて DF が最も効果的”という結果と一致する。この図では我々は利用データの割合を更に増やしたときの統計的な違いを認めることができなかった。

表２．特徴選択と推定方法の組み合わせ毎のナイーブベイズの macro average

重み付け	推定方法		
	従来	Laplace	Jeffreys
RIDF	39.2%	39.1%	41.4%
MI	32.6%	32.7%	35.7%
TM	39.4%	39.3%	42.5%
DF	41.4%	41.4%	44.2%
IG_e	37.1%	37.2%	39.2%
IG_max	37.8%	37.6%	41.1%
RE_EN	38.3%	38.3%	41.2%
CR	40.2%	40.4%	42.4%
Odds_e	37.4%	37.3%	41.1%
Odds_max	38.7%	38.8%	41.8%
Rr_e	37.5%	37.4%	40.8%
Rr_max	36.2%	36.3%	39.7%
We	40.3%	40.3%	43.5%

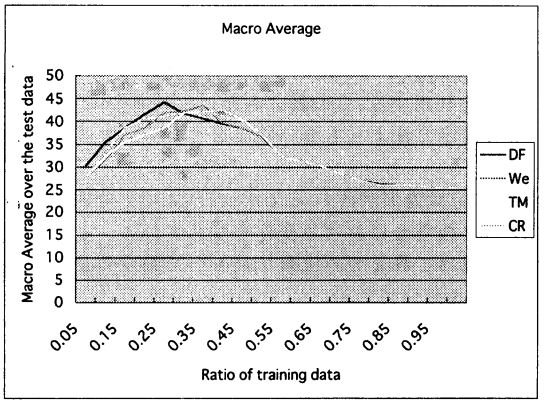

図1．各特徴選択手法の利用データの変化の利用割合に伴う macro average の変化

5. 考察

E-mail のフィールドの利用に関して、" From"のみを用いたものが他のフィールドの組み合わせよりもメールの分類精度が高かった。推定方法に関しては、Jeffreys による推定が最も分類精度が高かった。フィールドの組み合わせで比較すると" From"のみを用いた場合が最も分類精度が高かった。この理由は各フィールドに含まれる単語の異なり数が関係あると考えられる。4.1 に示したように、"Subject"と"Text field"に含まれる単語の異なり数は両者ともに" From"の四倍近くある。 これはそれだけ単語の出現確率の推定値が0に近くなることを意味する。それゆえ単語の出現回数に1を加える Laplace は 0.5 を加える Jeffreys よりも誤差が大きくなったと考えられる。

E-mail と他のテキストフィルタリングに用いられるテキストとの大きな違いは一ファイルあたりに含まれる単語の数が非常に少なく、フィルタリングに役立つ単語の数も少ないということがある。それゆえ他のテキストのフィルタリングよりも特徴選択と単語の出現頻度の推定方法がより重要になる。

6. 結論

本論文は E-mail を対象としたテキストフィルタリングの精度向上をメールのフィールドの利用、単語の推定方法、単語の特徴選択によって比較検討を行った。メールのどのフィールドを利用し、どのように単語を推定し、どのような特徴選択を用いれば良いかという課題に対し、我々はそれぞれ、フィールドとして" From"を含め、推定には Jeffreys、特徴選択に DF を用いることをこの論文で提案し、ENRON CORPUS を用いて検証を行い、その有効性を確認した。

文　献

[1] K. Bryan and Y. Yiming: The Enron Corpus: A New Dataset for Email Classification Research. ECML 2004: 217-226, 2004.

[2] D. Mladenic and M. Grobelnik: Feature Selection for Unbalanced Class Distribution and Naive Bayes, ICML 1999: 258-267, 1999.

[3] M Zaffalon and M Hutter: Robust Feature Selection by Mutual Information Distributions, Proceedings of the 18th International Conference on Uncertainty in Artificial Intelligence (UAI-2002), 2002.

[4] Y. Yang and J. Pedersen: A Comparative Study on Feature Selection in Text Categorization, 14th International Conference on Machine Learning (ICML-1997), 1997.

[5] J. Rennie: ifile: An application of machine learning to e-mail filtering, Proc. KDD Workshop on Text Mining, 2000.

[6] I. Androutsopoulos, J. Koutsias, K.V. Chandrinos, G. Paliouras and C.D. Spyropoulos: An Evaluation of Naive Bayesian Anti-Spam Filtering, in Proc. of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000), pp. 9-17, May 2000.

[7] Y. Diao, H. Lu, and D. Wu: A Comparative Study of Classification Based Personal Email Filtering, in Proc. PAKDD-2000, pp. 408-419, Apr. 2000.

[8] D. Mladenic, J Brank, M Grobelnik and N. Milic-Frayling: Feature selection using linear classifier weights: interaction with classification models, SIGIR 2004: 234-241, 2004.

[8] G. I. Jeffreys: The Estimation of Probabilities: An Essay on Modern Bayesian Methods, MIT Press, 1965.

[9] T. Joachims: Text Categorization with Support Vector Machines: Learning with Many Relevant Features, Proceedings of the European Conference on Machine Learning (ECML),Springer, 1998.

[10] S. Kiritchenko and S. Matwin: Email Classification with Co-Training, Proceedings of CASCON 2001, 2001.

[11] G. Manco, E. Masciari, M. Ruffolo, and A. Tagarelli, Towards an Adaptive Mail Classifier, AIIA 2002, Sep. 2002.

[12] A. McCallum and K. Nigam: A Comparison of Event Models for Naive Bayes Text Classification, AAAI-98 Workshop on Learning for Text Categorization, 1998.

[13] D. Mladenic and M. Grobelnik: Feature selection for classification based on text hierarchy (uncompressed) Working notes of Learning from Text and the Web, Conference on Automated Learning and Discovery CONALD-98, 1998.

[14] D. Mladenic and M. Grobelnik: Feature Selection for Unbalanced Class Distribution and Naive Bayes, ICML 1999: 258-267, 1999.

[15] D. Mladenic, J Brank, M Grobelnik and N. Milic-Frayling: Feature selection using linear classifier weights: interaction with classification models, SIGIR 2004: 234-241, 2004.

[16] H. Ng, W. Goh, and K. Low: Feature selection, perception learning, and a usability case study for text categorization, Proceedings of SIGIR97, 1997.

[17] K. Nigam, A. McCallum, S. Thrun and T. Mitchell: Using EM to Classify Text from Labeled and Unlabeled Documents, Technical Report CMU-CS-98-120, Carnegie Mellon University, 1998.

[18] V. V. Raghavan and H. Sever: On the Reuse of Past Optimal Queries. In Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR'95), Seattle, WA, USA, July 1995.

[19] J. Rennie: ifile: An application of machine learning to e-mail filtering, Proc. KDD Workshop on Text Mining, 2000.

[20] Y. Yang and J. Pedersen: A Comparative Study on Feature Selection in Text Categorization, 14th International Conference on Machine Learning (ICML-1997), 1997

[21] Y. Yang and X. Liu: A re-examination of text categorization methods, in Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.

[22] M Zaffalon and M Hutter: Robust Feature Selection by Mutual Information Distributions, Proceedings of the 18th International Conference on Uncertainty in Artificial Intelligence (UAI-2002), 2002

[23] Z. Zheng, X. Wu and R. Srihari: Feature selection for text categorization on imbalanced data, SIGKDD Explorations 6(1): 80-89, 2004