Unicodeの文字属性を用いた言語非依存型迷惑メール選別手法

A language independent method for spam filters using Unicode character properties

佐々木琢磨　阪口哲男

筑波大学大学院図書館情報メディア研究科

概要

インターネットとともに普及した電子メールは、情報伝達の基盤として様々な用途で用いられている一方で、多量に送りつけられる迷惑メールが増え、英語以外の言語による迷惑メールも増加している。受信者側の迷惑メール対策では、学習型迷惑メールフィルタが広く利用されている。しかし、学習型迷惑メールフィルタでは、メールから単語の抽出を行い、それに含まれる単語の特徴を分析する手法が多く用いられるため、単語を抽出する際の言語への依存性が生じている。言語に依存しない文字列の切り出し規則としては、固定長切り出し等があるが、判定に寄与しない文字列が多量に切り出されたり、英語に基づく記述が多いヘッダ部分を有効に利用することができない。そこで本研究では、Unicodeの文字属性に基づいて、メールの文字列を可変長に切り出し、それらを特徴として判定に用いる手法を提案する。本稿では、Unicodeの文字属性を利用した迷惑メール選別手法、およびその分類実験に基づいて精度について論じる。

1　はじめに

インターネットとともに普及した電子メールは、利用者間のコミュニケーションや連絡手段といった目的だけではなく、利用者の嗜好に合わせた宣伝や広告手段としても広く利用されている。電子メールは送信コストが低く、不特定多数に送信することが容易なことから、利用者の同意を得ない宣伝メールが横行し、近年では実体のない組織やwebサイトへの加入を勧誘したり、金銭の振込みを強制するような詐欺目的での利用も急増している。日本では、このような電子メールを迷惑メールと呼ぶことが一般的である。

迷惑メール対策においては、電子メールの送信者の信頼度を利用してフィルタリングする方法と、電子メールの内容を解析・学習してフィルタリングする方法がある。また近年では、NNIPF(Nearest Neighbor IP address based mail Filter)[1]も着目されている。これは、送信者情報であるIPアドレスを学習および判定に用いる手法である。本研究では、多量の迷惑メールをユーザが判断する代わりに自動選別することを目的とし、内容を解析する方式での迷惑メール対策として一般的な学習型迷惑メールフィルタに着目する。

学習型迷惑メールフィルタは、メールの本文やヘッダから単語を切り出して、それらの単語の出現頻度をそのメールの特徴として学習する。単語の切り出しの際には、特定の言語に依存した処理を行うことが一般的である。

しかし、将来的な電子メールの普及や、日常的に複数の言語を扱う利用者を考えると、特定の言語に依存しない方がよい。そこで本研究では、Unicodeの文字属性を利用した言語に依存しない文字列切り出し規則を提案し、メール選別実験によってその有効性を論じる。

2 学習型迷惑メールフィルタと言語依存性

2.1 Paul Graham 方式

学習型迷惑メールフィルタとは、Paul Graham が「A plan for spam」[2] で提唱したもので、過去に届いた迷惑メール及び非迷惑メールを統計的に学習し、その学習結果に基づいて新たに送られてきたメールが迷惑メールであるかどうかを判断して振り分ける手法である。

Graham により提案されたベイズ理論を用いた迷惑メール確率の計算過程は次のようになる。

まず、過去に受信した迷惑メールと非迷惑メールを準備し学習データを作成する。この学習データに登録される情報は、それぞれのメールのヘッダや本文より切り出された単語 w_i について、非迷惑メール中の出現頻度 g_{w_i}、および迷惑メール中の出現頻度 b_{w_i} である。また、非迷惑メールの学習総数 n_{good}、迷惑メールの学習総数 n_{bad} も登録しておく。次に、判定対象のメールに対しても同様に、ヘッダや本文を単語単位に分割する。こうしてある単語 w_i が得られた時に、w_i を含むメールが迷惑メールである確率 $p(w_i)$ は、学習データの情報を参照し、次の式により計算される。

$$p(w_i) = \frac{\frac{b_{w_i}}{n_{bad}}}{a \cdot (\frac{g_{w_i}}{n_{good}}) + \frac{b_{w_i}}{n_{bad}}} \qquad (1)$$

(1) 式の定数 a は、各単語にかけるバイアスであり、Graham は $a = 2$ を採用している。これは、実際に迷惑メールフィルタを利用した場合に起こりうる選別ミスを考えたときに、非迷惑メールを誤って迷惑メールと判断する場合 (誤遮断, False-Positive) の方が迷惑メールを誤って非迷惑メールと判断する場合 (誤通過, False-Negative) よりも損害が遥かに大きいので、前者が発生しにくいようにするためである。

メールを判定する場合は、(1) 式で算出された個々の単語の迷惑メール確率を結合して算出されるメール全体の迷惑メール確率に基づいて判定する。

この Graham による手法は迷惑メール対策に一定の効果を挙げ、後にこの手法を基に、Gary Robinson が改善方式を提案した。

2.2 Gary Robinson-Fisher 方式

Robinson は「Spam detection」[3] で、Graham 方式において経験則から 0.4 に設定していた、学習不足による確率に着目し、それが自動的に適切な値が計算されるような改善方式を示した。まず、Graham の手法の (1) 式に対して $a = 1$ で適用し、ある単語 w_i を含むメールが迷惑メールである確率 $p(w_i)$ を、

$$p(w_i) = \frac{\frac{b_{w_i}}{n_{bad}}}{\frac{g_{w_i}}{n_{good}} + \frac{b_{w_i}}{n_{bad}}} \qquad (2)$$

により計算する。

Robinson は (2) 式で求めた $p(w_i)$ を、学習不足を考慮した確率補正関数に与えて得られた $f(w_i)$ を用いて迷惑メール確率を求める手法を提案し、Gary Robinson 方式と呼ばれるようになった。

また、後に Robinson は「A statistical approach to the spam problem」[4] において、迷惑メール指標 I を基準にして選別を行う方式を提唱した。これは、逆 χ^2 関数 C^{-1} および、先述した $f(w_i)$ を用いるもので、式の簡略化のために H, S を用いて示すと、次式のようになる。

$$H = C^{-1}(-2\ln\prod_{i=1}^{n} f(w_i), 2n) \qquad (3)$$

$$S = C^{-1}(-2\ln\prod_{i=1}^{n}(1 - f(w_i)), 2n) \qquad (4)$$

$$I = \frac{1 + H - S}{2} \qquad (5)$$

上式で得られた I を用いて迷惑メール選別を行う手法は、Gary Robinson-Fisher 方式と呼ばれ、現在において、迷惑メール確率計算方式と

して広く知られている。本研究でもこの方式を採用する。

2.3 言語依存性の問題

以上のように、Graham が提唱した学習型迷惑メールフィルタの動作には、メールのヘッダと本文を単語単位に切り出す機構が必要であり、このような目的には一般的に形態素解析器が用いられる。しかし、形態素解析器は特定の言語に依存していることが普通である。

これに対し、近年では迷惑メールの記述言語も多様化しているため、正確に単語を切り出せないことが原因となり、フィルタの選別精度が低下している。これに対応するために、可能性のある言語それぞれに対応した形態素解析器を準備することは現実的ではない。従って、様々な言語で書かれたメールを日常的に扱う場合は、特定の言語に依存せずに文字列を切り出す規則が必要である。

そこで本研究では、Unicode の文字属性を用いた言語に依存しない文字列の切り出し規則を提案する。それによって分割された文字列で学習や判定を行い、文字列が適切に切り出せないことが原因となる選別精度の低下が発生しない学習型迷惑メールフィルタの開発を目指す。

3 従来の言語非依存型選別手法

本研究が対象とする言語非依存型選別手法とは、言語に依存する処理を用いずにメールのヘッダと本文から文字列を切り出し、切り出された文字列を学習や分類に用いる手法のことである。本章では、固定長切り出しを用いた先行研究とその問題について述べる。

3.1 阪口・于による研究

[5] では、従来の学習型フィルタが用いていた単語の切り出しを固定長切り出しに変更し、切り出す長さを 1 から 3 に変えながら、学習・選別実験を行っている。固定長切り出しの例を図 1 に示す。

学習および判定に用いる機構としては、既に電子メールの分類において [6] で効果が示されていた SVM を採用している。

N=1：今/日/は/い/い/天/気/で/す/ね/
N=2：今日/日は/はい/いい/い天/
天気/気で/です/すね/ね。
N=3：今日は/日はい/はいい/
いい天/い天気/天気で/
気です/ですね/すね。

図 1 「今日はいい天気ですね。」に対する固定長切り出し例

実験に扱うメールとして、自身のメールボックスと公開されている複数のアーカイブから集めた学習用迷惑メール 91 通、学習用非迷惑メール 82 通、評価用迷惑メール 59 通、評価用非迷惑メール 118 通を用いている。また、SVM による分類には SVM^{light} を用いている。実験では、比較対象に bsfilter[7] を取り上げている。

実験によると、3 文字長切り出しの場合に、非迷惑メール・迷惑メール選別精度がともに 100% になり、1-3 の文字列長それぞれの中で最も良い結果が得られている。

一方、bsfilter を用いた比較実験の結果によると、非迷惑メール・迷惑メール選別精度がそれぞれ 100%, 94.9% になり、同じかそれより低い値となっているため、この方式が有効であることが確認されている。

後に、この研究を受けて、石原が「言語に依存しない迷惑メールフィルタの開発」[8] により処理性能の向上をはかっている。

3.2 大福・松浦による研究

大福・松浦は、[9] において、日本語を含むメールに関して、日本語の部分のみを固定長で切り出し、その文字列を学習に用いる機構の学習型迷惑メールフィルタを用いて選別実験を行っている。

この実験では、自身の研究室で受信したメー

ル 1494 通 (非迷惑メール 689 通、迷惑メール 805 通) を選別対象に用いている。

大福らの実験では、2 文字長の場合が一番精度良く選別できたことが示してあり、比較実験に 2 文字長の切り出しを採用している。また、日本語を全く無視した場合と固定長切り出しを適用した場合、および形態素解析を用いた場合は精度に大きく影響しなかったことが述べられている。

日本語の扱いに関わらずにある程度の選別ができた要因としては、ヘッダの大部分を占める英語に基づいた記述部分や、本文に含まれる英単語・URL に含まれる英語などの抽出が実際に効果があったのではないか、と論じられている。

3.3 固定長切り出しの問題点

N-gram 手法による固定長切り出しの選別精度を確認するためにまず、後述する実験で用いるメールコーパスで選別実験を行った。なお、この際の選別実験は、[5] の手法に従った。その結果、非迷惑メール・迷惑メール選別精度がそれぞれ 94.6%, 90.6% と、精度が落ちたことがわかった。一方、bsfilter では 97.8%, 95.6% という結果であり、それよりも精度が劣ってしまった。精度が落ちた原因として、固定長切り出しの場合には、[9] において選別に効果的であると述べられている英語に基づいた記述を適切に抽出できないことが一つの要因と考えられる。

そこで、本研究では固定長切り出しではなく、文字や数字、記号といった文字の種類の変わり目で可変長の文字列を切り出す規則を提案する。

4 Unicode 規格の文字属性を利用した切り出し手法

固定長切り出し規則を用いる際には、マルチバイト文字が文字列中に存在すると、その文字コードにおいての正確な固定長の切り出しができない場合がある。[5] では、この問題を解決するために迷惑メールの記述文字コードを判定し、Unicode に変換した上で、Unicode に基づいて N 文字長の文字列切り出しを実現している。また、Unicode という統一文字コードで処理を進めることは、学習や判定に不要な記号などを除去することが容易であることなど、Unicode への変換は利点が多い。

Unicode Consortium は、UCD(Unicode Character Database)[10] を公開している。UCD は、個々の文字とそれらが持つ属性に関する詳細な情報が記録されたテキストファイルで構成されている。これらは、個々の文字についての情報が書かれたファイルを中心に、約 30 のファイルが存在する。本研究では、UCD の文字分類情報を文字列の切り出し規則に適用する。

4.1 カテゴリ

カテゴリとは、Unicode の文字分類のうち General Category 属性のことである。カテゴリは 2 文字の英文字で表され、表 1 のように 2 階層分類の形式をとっている。

まず第 1 字目の大文字により大きく 7 つに分類され、第 2 字目の小文字でさらに小さく分類される。

カテゴリは、Unicode で規定された全ての文字をその用途や文法上の役割により排他的に分類する規則であり、特定の言語のみに適用される分類ではない。したがって、この文字分類を文字列の切り出し規則に採用することは、言語非依存型選別手法として妥当であると考えられる。

本研究では、1 文字目のみの 7 つの区分, Letter, Mark, Number, Punctuation, Symbol, Separator, Other が前後で変化している部分を区切りとする切り出し規則を採用した。この規則を用いた場合の切り出しの例を図 2 に示す。

図 2 の例では、文字は Letter, スペースは

表 1　カテゴリ値と対応する意味

値	第 1 字目の区分	第 2 字目の区分
Lu	Letter	uppercase
Ll	Letter	lowercase
Lt	Letter	titlecase
Lm	Letter	modifier
Lo	Letter	other
Mn	Mark	nonspacing
Mc	Mark	spacing combining
Me	Mark	enclosing
Nd	Number	decimal digit
Nl	Number	letter
No	Number	other
Pc	Punctuation	connecter
Pd	Punctuation	dash
Ps	Punctuation	open
Pe	Punctuation	close
Pi	Punctuation	initial quote
Pf	Punctuation	final quote
Po	Punctuation	other
Sm	Symbol	math
Sc	Symbol	currency
Sk	Symbol	modifier
So	Symbol	other
Zs	Separator	space
Zl	Separator	line
Zp	Separator	paragraph
Cc	Other	control
Cf	Other	format
Cs	Other	surrogate
Co	Other	private use
Cn	Other	not assigned

Separator、アポストロフィやピリオドは Punctuation、数字は Number であるため、可変長の切り出しができている。

特に、英語などスペースが単語と単語の境界に存在する言語においては、文字の Letter、スペースの Separator であり、それらが隣接しているため切り出しの区切りとなり、結果的に形態素解析器のように動作する。[9] で述べられているように、英語に基づく記述はヘッダなどに高確率で含まれるため、それらを適切に切り出すことは精度の改善に結び付くと考えられる。

○切り出し前
It’s fine until 2 o’clock.
○切り出し後
It/’/s/ /fine/ /until/ /2/ /o/’/clock/./

図 2　カテゴリを用いた可変長切り出し

4.2　コードブロック

カテゴリによる切り出しでは適切に切り出せない場合がある。日本語を例にとると、構成する主な文字種別である漢字・ひらがな・カタカナは、カテゴリ値はいずれも Lo であり、日本語の文章の多くにおいては区切りとなるようなカテゴリ値の変わり目が存在せず、適切な切り出しができないと考えられる。そこで、日本語のように複数の文字種別により構成される言語や、本文やヘッダに含まれる可能性の高い英語の部分を適切に切り出せるように、Unicode 規格のコードブロックを切り出し規則に追加する。

そもそも、Unicode 規格は、各国で標準として規定されている文字セットや実際に使用されている文字を持ち寄って作成された規格のため、元となった文字セットに関する情報も存在する。それらは、図 3 のように対応付けられている。

コードブロックは、Unicode で扱っている文字の種類を区分するものであり、言語に依存しない切り出し規則に利用可能であると考えられる。この規則による切り出しの例を図 4 に示す。

図 4 から、この規則を用いると、文章が複数

```
0000..007F; Basic Latin
...
3040..309F; Hiragana
30A0..30FF; Katakana
...
4E00..9FFF; CJK Unified Ideographs
...
```

図 3　文字セットと文字コードの範囲の対応 (抜粋)

○切り出し前 明日は、13 時から公園で マラソン大会があります。
○切り出し後 明日/は/、/13/時/から/公園/で/ マラソン/大会/があります/。

図 4　コードブロックを用いた切り出しの例

のコードブロックに属する文字で構成される日本語のような言語の場合、可変長の切り出しができることがわかる。

4.3　提案方式の留意点

今回の規則で効果的な可変長切り出しが不可能な例を挙げる。その例として中国語があり、図 5 に示す。

今天是好的天气

図 5　中国語の例

中国語は、Letter, others という同一の Unicode のカテゴリ値が連続し、かつ、CJK Unified Ideographs という同一のコードブロックの文字の連続からなっているため、提案手法では適切な長さに切り出しができない。しかし、[9] で述べられていたように、ヘッダを適切に切り出せたり、本文中に含まれる英語を抽出できるため、ある程度の選別精度を出すことは可能と考えられる。

5　メール選別実験

本実験は、先行研究の言語に依存しない迷惑メールフィルタの処理過程のうち、固定長による切り出し部分を、Unicode 規格の文字属性を用いるものに置き換えて、迷惑メールの選別をする。

本実験で用いる迷惑メール確率計算には、Gary Robinson-Fisher 方式を用いる。これは、bsfilter など既に実用化されているフィルタの標準方式である。

選別実験用の非迷惑メールと迷惑メールには、SpamAssassin[11] という迷惑メールフィルタの開発組織が公開しているメールコーパスのうち、2002 年 10 月から 2003 年 2 月分である迷惑メール 4575 通、非迷惑メール 10117 通を用いた。そのうち 200 通程度が非英語のメールである。それらに対し、(a) 最低限度の学習下による選別、(b) クロスバリデーション (leave-one-out) による選別、クロスバリデーションを変形した、(c)leave-”two”-out による選別の 3 種類を行う。

最低限度の学習下での選別とは、コーパス中からそれぞれ 100 通を無作為に抽出した 200 通のセットを 3 つ作成し、そのうち 2 セットで学習し、1 セットで判定する実験である。400 通の学習量とは、Paul Graham 方式のベイジアンフィルタの最低限度の学習量であり、[12] で示された。

クロスバリデーションは、コーパス全体を 3 つのセットに分け、2 セットで学習、1 セットで判定する実験である。同様に 1 セットで学習、2 セットで判定したものを leave-”two”-out とする。

5.1　実験の手順

本実験の処理の流れを図 6 に示す。

まず、学習用の非迷惑メール・迷惑メールをそれぞれ学習させ、学習データベースを作成する。各メールごとに、Unicode の文字分類のう

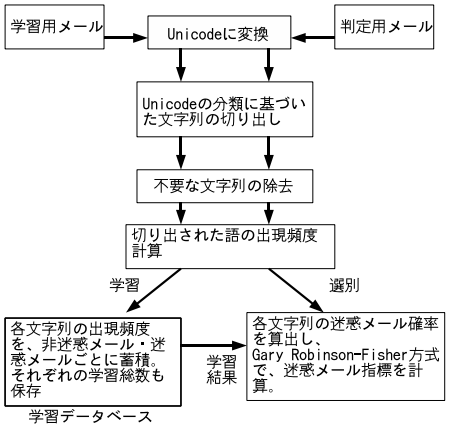

図 6　本実験の処理の流れ

ち、カテゴリ・コードブロックの変わり目で、文字列の切り出しが行われる。今回の実験で適用した規則は以下の通りである。

(i) カテゴリの値 (表 1) の第 1 字目が前後で異なる場合、そこを区切りとする。

(ii) コードブロックが前後で異なる場合、そこを区切りとする。

(iii) 非 Letter 要素のみから成る文字列の除去。(Mark, Number, Punctuation, Symbol, Separator, Other のみで構成される分割文字列は除外する。

この切り出し規則を適用した場合に切り出される文字列の例は、図 7 の通りである。

○例文 明日は、13 時から公園でマラソン大会があります。
(i):カテゴリ値による切り出し 明日は/、/13/時から公園で/ マラソン大会があります/。
(ii) コードブロックによる切り出し 明日/は/、/13/時/から/公園/で/ マラソン/大会/があります/。
(iii) 非 Letter 要素の削除 明日/は/時/から/公園/で/マラソン/ 大会/があります

図 7　文章の切り出し例

(iii) の、非 Letter 要素の削除という処理は、記号だけで構成される文字列やスペース 1 字の文字列は、迷惑メールの判定計算に寄与しないと考えられるからである。

上記の規則により切り出された文字列を出現頻度とともに学習データベースに格納する。

次に、判定用の非迷惑メール・迷惑メールを学習データベースに基づき判定する。判定用のメールも学習用メールと同様の切り出し規則に基づき文字列を抽出する。

最終的に各メールごとに、迷惑メール指標 I が (5) 式により求められる。学習型迷惑メールフィルタは、一定のしきい値に従って判定をするが、今回は、指標 I の分布を見て、選別精度が最も良くなるようにしきい値を設定する。

5.2　実験結果

表 2　各手法の各平均精度

	FP-rate(%)	FN-rate(%)
(a):bsfilter	8.00	9.00
(a):本手法	4.00	9.00
(b):bsfilter	0.16	0.64
(b):本手法	0.14	3.65
(c):bsfilter	0.87	4.55
(c):本手法	0.42	8.52

表 2 は、分類実験の結果である。評価精度は、FP-rate と FN-rate である。FP-rate とは、非迷惑メールを誤って迷惑メールと判定してしまった割合で、FN-rate とは、迷惑メールを誤って非迷惑メールと判定してしまった割合である。また、(a) は最低限度の学習下、(b) はクロスバリデーション、(c) はクロスバリデーションの変形の leave-”two”-out である。各パターンを通して、FP-rate に関しては本手法が bsfilter よりも小さいことがわかる。しかし、FN-rate は、学習量の増加に伴い bsfilter が大きく精度を上げていることがわかる。

また、FP 判定されたメールの内容に着目す

ると、提案手法において、FP メールは、各パターンを通して計 14 通あった。HTML メールはそのうち 13 通を占め、1 通が通常のプレーンテキストのメールであった。これらの記述言語は英語であったため、言語に依存せずに抽出された特徴は、学習や判定の単位として有効であると言える。それらの HTML メールを分析すると、HTML タグとしてよく使われ、HTML メールではない迷惑メールに頻出する特徴 ”body”, ”head”などが高い迷惑メール確率を持っていること、本文自体が短かいなど、他に特徴となりうる文字列が抽出できなかったために FP 判定されたと考えられる。

6 おわりに

本研究では、言語に依存しない文字列の切り出し手法として、Unicode 規格の文字分類を採用し、切り出された文字列をメールの特徴として、Gary Robinson-Fisher 方式により選別したところ、bsfilter に比べ、特に非迷惑メールの判定精度の改善ができた。

ここで、今後の展望について述べる。現時点の処理においては、アポストロフィや記号「ー」は、区切りとみなしている (図 2) ため、文字列の一部として利用する形にはしていない。数字や一部の記号についても同様に、学習や判定に利用していない。

しかしながら、図 8 のような特定の記号を含む表現について「文字列の 1 部とみなす」等の処理を行うことにより、判定に寄与しうる文字列に切り出すことができると考えられる。

○英語などの文章で見られる表現 $と数字を含む値段などの表現 アポストロフィを含む 1 単語
○日本語の文章で見られる表現 「ニュース」などに含まれる記号「ー」

図 8　記号の処理について

このように、判定に効果的な部分を精度よく切り出したり、判定に寄与しない文字列の除去手法を検討し、選別精度を上げるだけではなく、判定に大きく寄与する文字列のみを利用することを目指し、処理時間や計算量の低減の側面からも検討していきたいと考えている。

参考文献

[1] 和田俊和. NNIPF. 迷惑メール選別ツール http://vrl.sys.wakayama-u.ac.jp/~twada/NNIPF.html

[2] Paul Graham. A plan for Spam. 2002. http://www.paulgraham.com/spam.html

[3] Gary Robinson. Spam detection. 2002. http://radio.weblogs.com/0101454/stories/2002/09/16/spamDetection.html

[4] Gary Robinson. A statistical approach to the spam problem. 2003. http://www.linuxjournal.com/article/6467/

[5] 阪口哲男, 于家富. 言語に依存しない迷惑メール選別手法. 情報知識学会誌, 2005, vol.15, No.2, pp53-56.

[6] 米倉正和, 堀幸雄, 後藤英一. Support Vector Machine を用いた電子メールの自動分類. 情報処理学会研究報告, NO.083-004, pp.19-26, 2003.

[7] Bsfilter 迷惑メール選別ツール http://bsfilter.org/

[8] 石原幸輔. 言語に依存しない迷惑メールフィルタの開発. 筑波大学, 2005, 修士論文.

[9] 大福泰樹, 松浦幹太. ベイジアンフィルタによる日本語を含むメールのフィルタリングについての考察. 2006 年 暗号と情報セキュリティ・シンポジウム (SCIS2006) 予稿集 (CD-ROM), 2006.

[10] Unicode 5.0. The Unicode Consortium. http://www.unicode.org/Public/UNIDATA/

[11] SpamAssassin. 迷惑メール選別ツール http://spamassassin.apache.org/

[12] Bart Massey, Mick Thomure, Raya Budrevich, Scott Long. Learning Spam: Simple Tecniques For Freely-Available Software. USENIX Annual Technical Conference, FREENIX Track, USENIX, pp63-76(2003).