平成21年度　茨城大学工学部情報工学科卒業研究論文

One Class SVM を利用した非該当の人名ページの検出

著者：SA MIN SUN(06T4070L)

指導教官：新納　浩幸 准教授

平成21年2月10日

平成21年度　茨城大学工学部情報工学科卒業研究論文

One Class SVM を利用した非該当の人名ページの検出

著者：Sa Min Sun (06T4070L)
指導教員：新納 浩幸　准教授

論文 要旨

本論文では、Web上の人物検索における曖昧性の問題を扱う。一般に人物検索では重要な人物とそれ以外のマイナーな人物が検索されるので、マイナーな人物の検索結果を検索結果集合の外れ値と見なし、外れ値検出の手法であるOne Class SVM を利用して、上記の問題の解決を目指す。

Web 上の人物検索はWeb 検索において重要な地位を占めてきている。このような状況の中、人物の検索に関する問題として人物の同姓同名問題がある。人物の同姓同名問題とは、Web 検索において同姓同名の人物の存在によって検索結果から目的の人物のページを発見することが困難になるという問題である。

この問題に対し、One Class SVM (Support Vector Machine)を利用する。One Class SVMとは、大量のデータの中から他と異なる特徴をもつデータで境界となるサポートベクターを形成し、外れ値検出を行う手法である。ここでは、非該当のページを外れ値とみなすことでOne Class SVMを用いて非該当ページの検出を行う。

One Class SVMを利用するためにライブラリLibSVMを利用する。LibSVMは、サポートベクタ分類器（C-SVC、nu-SVC）、回帰分析（epsilon-SVR、nu-SVR）、分布評価（One Class SVM）のための統合ソフトである。 LibSVMのコマンド「svm-train」では、カーネルを選択し、各パラメータを指定することができる。

実験では、本研究では、韓国人の有名人である同姓同名の検索結果300件（スニペット）をN-gramで表現し、それらの共起関係を求め、One Class SVMを利用した。2個のデータセットに対して、3つのカーネル（多項式、RBF、シグモイド）と、nuパラメータとgammaパラメータをそれぞれ変えながらLibSVMを実行し、適合率と再現率を組み合わせたF値を用いて評価した。その結果、データ1とデータ2のF値の最大値が0.5以下ということもあり、あまりいい結果ではなかった。その原因として、bi-gramで共起関係を取る際のノイズとヒューリスティックに指定したパラメータの設定が考えられる。

目次

図目次

表目次

第1章　序論

１．１　はじめに

Web 上の人物検索はWeb 検索において重要な地位を占めてきている。このような状況の中、人物の検索に関する問題として人物の同姓同名問題がある。人物の同姓同名問題とは、Web 検索において同姓同名の人物の存在によって検索結果から目的の人物のページを発見することが困難になるという問題である。たとえば、Googleで「山口百恵」を検索してみる。当然ながら、元歌手の山口百恵という人物の検索結果が大量に表示される。しかし、Wikipediaには載っているフットサル選手の山口百恵の検索結果は上位10位内には表示されない。今度は、情報処理学会の会長である「佐々木元」という人名で検索してみる。上位10件は、NECの会長、東北大学の教授、プロのBMXライダー、翻訳者の4名の人物で占められている。もしも、探している人物が、フットサル選手の山口百恵だった場合は12位まで、また、1960年代の映画監督の佐々木元だった場合には82位までリストを見続けなければいけない。さらに、特に有名ではない山口百恵や佐々木元を捜している場合には100位以内にも入っていないかもしれない。このように、Web上の人物検索において、字面上同一の人名が現実世界にいる複数の人物を指すことがあるという同姓同名の問題がある。

この問題に対し、One Class SVM (Support Vector Machine)を利用する。One Class SVMとは、大量のデータの中から他と異なる特徴をもつデータで境界となるサポートベクターを形成し、外れ値検出を行う手法である。これにより、非該当の人名ページを外れ値と見なすことができる。

本研究では、韓国人の有名人である同姓同名の検索結果300件（スニペット）をN-gramで表現し、それらの共起関係を求め、One Class SVMを利用することができる「LibSVM」というライブラリで非該当の人名ページの検出を行い、WEB文書中の同姓同名の曖昧性解消を目指す。

１．２　本論文の構成

第2章において、N-gramについて述べる。ここでは、文字N-gramについての概要や、bi-gramを選択した理由、意味のないタームを排除する基準等を述べる。

第3章では本研究のメインである、SVMとOne Class SVMについての概要、およびOne Class SVMの手法を使うためのツールであるLIBSVMについて説明する。

第4章ではOne Class SVM を利用した非該当の人名ページの検出実験を行い、第5章に考察、第6章に結論へと進む。

また、巻末にはプログラムのソースリストを添付する。

第２章　N-gramを用いたスニペットの表現

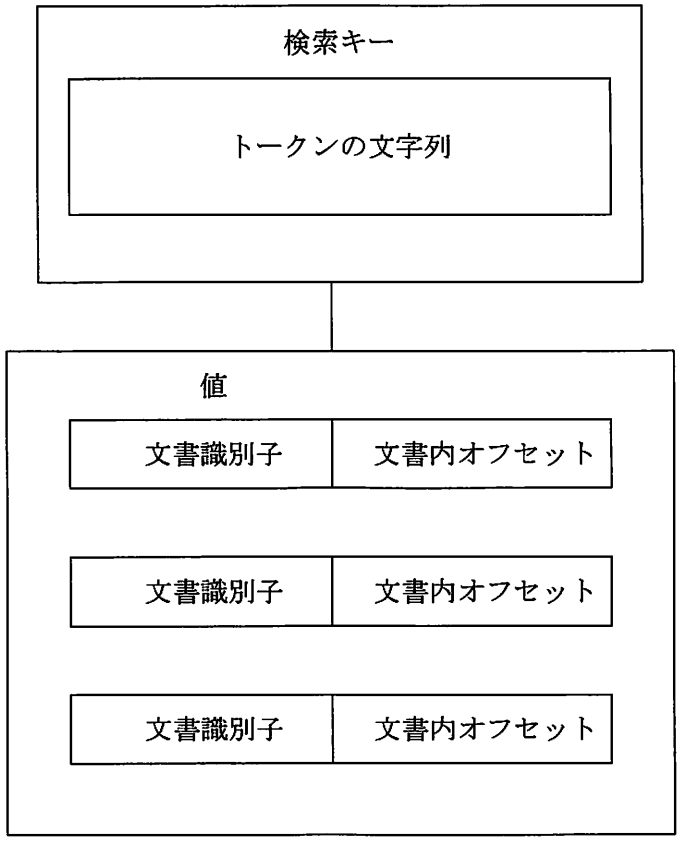
２．１　N-gram

日本語や韓国語の場合は複雑な形態素解析をしなければ単語を切り出すことができない。そこで、形態素解析を使わずにキーワードらしきものを切り出す方法として文字 N-gram がある。　N-gramという概念は、情報理論の創始者として名高いクロード・エルウッド・シャノン（Claude　Elwood　Shannon、1916-2001）が想起したものである。N-gramの定義は、「あるテキストの総体を前から順に任意のN個の文字列または単語の組み合わせで分割したもの」である。

N-gram 法は、全文検索システムにおいて対象文書のテキストからトークンを抽出して転置インデックスを構築するための主要な手法のひとつである。N-gram 法においては、対象文書のテキストから，連続した一定の文字数の文字列をトークンとして切り出す[1]。テキストを1 文字ずつ分割する方法を1-gram（uni-gram）法、2 文字ずつ分割する方法を2-gram（bi-gram）法、3 文字ずつ分割する方法を3-gram（tri-gram）法と言い、N 文字ずつ分割する方法を総称してN-gram 法と言う。トークンを構成する文字が何文字であっても、トークンは1 文字ずつずらしながら重複して切り出される。例えば「形態素解析」という文字列に2-gram 法を適用すると、「形態」「態素」「素解」「解析」というトークンが得られる。切り出された各々のトークンには、その位置情報が付与される。位置情報は，そのトークンを含む文書の識別子と、そのトークンが文章中の何番目に出現したかというオフセットからなる。転置インデックスにトークンの情報を記録する際には、文書内における同一パターンのトークンの情報をまとめてレコードを構成する。レコードはトークンの文字列を検索キーとして転置され、ハッシュ表やB+木などのインデックスを伴ったデータベースに記録される（図2.1）。記録した位置情報は、検索時に、トークンの存在と連続性を調べるために使われる。例えば「形態素」という検索要求に対しては、「形態」および「態素」が同一の文書に含まれるかどうかをまず調べ、次に、「形態」をN 番目としたときに「態素」がN+1 番目にあるものを該当とみなすこ

とになる。

図2.1：N-gram法の転置インデックスの構造



文書内オフセットにおいて、変域を制限せずに効率的に値を表現するためには、可変長のデータ型を用いるのが一般的である[2]。可変長のデータ型においては、小さい値は短いビット長で表現できるが、大きい値を表現するには長いビット長が必要となる。よって、個々の文書のテキストが長くなるほど文書内オフセットのビット長は長くなる。N-gram 法によるトークンの数は文書内の文字数とほぼ等しくなるので、文書識別子の情報量よりも文書内オフセットの情報量の方が支配的になる。したがって、N-gram法の転置インデックスにおいては、対象文書のテキストが長くなるほど、転置インデックスのサイズが肥大化し、すなわち空間効率は悪化する。なお、文書内オフセットのリストは単調増加列であるため、より値の低い差分列に可逆変換することが可能であり、それ

によって情報量を抑制することもできる。文書内オフセットを固定長のデータ型で表現した場合は、転置インデックスの空間効率は対象文書のテキストの長さに依存せずに済む。しかし、固定長データ型において十分な変域のビット長を設定すると、可変長のデータ型を用いるよりも空間効率が悪くなる。N-gram 法の転置インデックスを使った検索においては、検索語の文字数がN 文字でない場合に時間効率が悪化するという問題がある。

N 文字未満の検索語で検索を行う場合には、その検索語に前方一致する全てのトークンによる検索結果の和集合をとることになる。N 文字を越える検索語で検索を行う場合には、検索語をN 文字からなる複数のトークンに分割し、その各々の検索結果の積集合をとり、さらに位置情報を参照してトークンの連続性を確認する必要がある。

近年、人文学におけるテキスト研究においても、Nグラムが注目されてきている。これまで、テキスト研究における統計的な分析（計量文献学）においては、形態素分析が不可欠の作業とされており、実際、品詞をマークアップしたテキストデータベースによる研究によっていくつかの注目すべき成果が上がっている。しかし、形態素分析に基づくデータ処理には、次のような問題点が指摘されている[3]。

- 1語の単位を認定する基準が一通りではない。
- 複合語や強い共起性のある単語群（連語、慣用句など）の分析に不利。

一方、N-gram の場合、単語の区切りなどを問題にすることなく網羅的に数え上げるため、たくさんのノイズが含まれる反面、上記のような問題点を克服しうるデータを得ることができるのである。本研究では、このN-gram を利用し実験を進めた。上記で述べた「ノイズが含まれる」という問題点を克服するため工夫する。

２．２　bi-gramの利用

Nが大きくなると位置情報が分散し、各索引語に対する位置情報の量は少なくなる。ただし、索引語のパターンが多くなるので転置ファイルは大きくなる。 Nが小さいと、検索単語数が多くなり、位置情報の照会回数が多くなる。 Nが大きいと、Nより短い文字列長の検索質問に対して、前方一致で全ての転置ファイルから検索する必要があるため検索時間が大幅に悪化する。 N-gramのNの値を多くすればするほどより正確に表現できるようになるが、逆に統計的にスパースになるため[4] 、本実験では、Nの値を2にしたbi-gramを利用する。

N-gram では、隣り合った文字列または単語の組み合わせを連続要素と呼ぶ。例として、文字列「健康保険証」の連続要素を以下に示す。

「健康保険証」の連続要素

- 1(uni)-gram「健」「康」「保」「険」「証」
- 2(bi)-gram「健康」「康保」「保険」「険証」
- 3(tri)-gram「健康保」「康保険」「保険証」

表2.1にbi-gramで分解しした例を示す。

表2.1：bi-gramで分解した例

私	は										
	は	茨									
		茨	城								
			城	大							
				大	学						
					学	の					
						の	学				
							学	生			
								生	で		
									で	す	
										す	。

表1のように、N-gramでは「私は」「は茨」「茨城」のようなN個を単位とする文字列の組み合わせを「共起」関係にあるとする。またテキスト全体での任意の「共起」が現れる回数を「共起頻度」と呼ぶ。

本手法では、スニペットにある文字列と全スニペット間において、この連続要素がいくつ共起関係にあるかを調べる。ここでbi-gram の共起関係を例として示す。図2.2 のような検索文字列「今日は良い天気です。」と用例文字列「今日は晴れです。」は5 つの共起(図中の＊) が存在する。

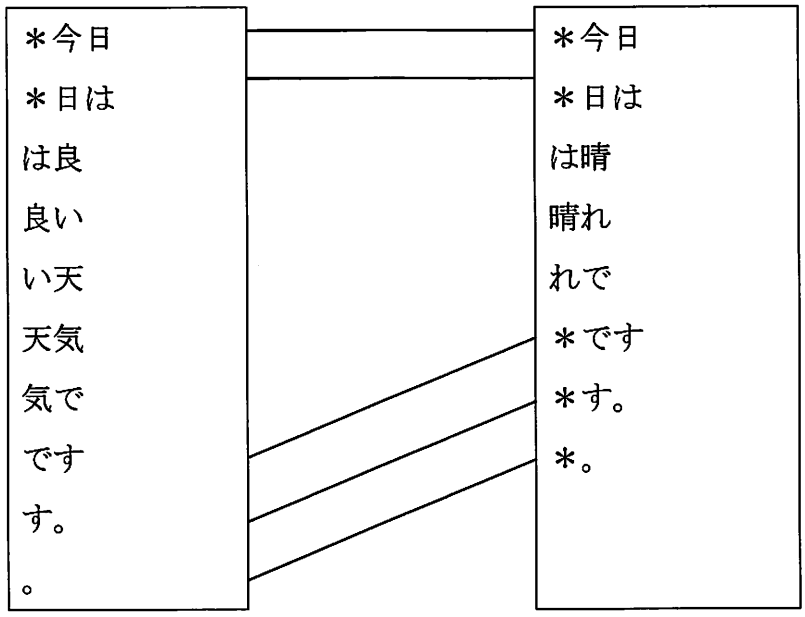


図2.2： 文書間の共起関係

ここでa を検索文字列の長さ，b を用例文字列の長さとする。a、b の連続要素数は、bi-gram のため、それぞれ a-1、b-1 となる。これらの値のうち、大きい方を分母とし、検索文字列との共起関係数を分子とした数値を類似度N と呼ぶ[5]。

$$N = \frac{\text{共起数}}{\max(a,b)-1}$$

図2 の例では、N=$\frac{5}{9}$ である。このN の値が大きくなるほど、検索文字列と用例文字列が類似しているといえる。これにより、検索文字列と用例間の類似度を調べることで、それらが類似しているかどうかが判定できる。

また、N-gram モデルを用いることで用例間の類似度が計算できる。しかし、検索文字列とデータベースに登録されている全ての用例との類似度をそれぞれ計算するには、検索毎に全用例を取得する必要があるため非効率である。この問題を解決するために、転置インデックスデータベースを用いる。一般的に転置インデックスとは、全文検索を行

う対象となる文書群から単語の位置情報を格納するための索引構造をいう。本手法では、データベースに登録されている全用例を予めbi-gram の連続要素でデータベースに格納する。またその連続要素を含む全ての用例を一意に特定できるID を同時に登録しておき、それをリストとして保持する。例として、ID 番号が1 である用例「今日は良い天気です。」とID 番号が2である用例「今日は晴れです。」の転置インデックスを表2.2 に示す。

表2.2： 位置インデックス

連続要素	ID
今日	1、2
日は	1、2
は良	1
良い	1
い天	1
天気	1
気で	1
です	1、2
す。	1、2
。	1、2
は晴	2
晴れ	2
れで	2

これにより類似度計算における共起数を求める際には，検索文字列における連続要素に関連付けられたID を数えるだけでよい。例えば、検索文字列が「明日の天気」の場合、この連続要素である「明日」、「日の」、「の天」「天気」を含むID を数えることになる。この場合、連続要素「天気」を含む文章1 だけ共起関係にあると分かる。このた

め、全用例と直接共起関係を調べなくとも、検索文字列がどの文章と共起関係にあるかが分かる。このように転置インデックスを用いることで、用例対訳の増加に依存せず検索効率を上げることができる。

本実験では、韓国の有名人の同姓同名を検索し、その結果中300個のスニペットを用いて実験を進める。300個のスニペットから2文字づつ取り出し共起関係を作り、この共起関係を同じ300個のスニペットと比較をして共起頻度を求める。だが、2.1節でも述べたように、N-gramで文字列を取る際に、意味の無いターム（ノイズ）が現れる。例えば、表2.1の「私は茨城大学の学生です。」からは「私は」「は茨」「茨城」「城大」「大学」「学の」「の学」「学生」「生で」「です」「す。」 の 11個の文字 bi-gram が得られる。しかし、「は茨」や「の学」などという意味のないタームもでてきてしまう。そこで、これらの文字 bi-gram から何らかの基準や、TF・IDF などを使ってよりよいキーワードを選択することが重要になる。本実験では、韓国語で検索されたスニペットを利用するため、韓国語について少し説明する。文法の順番は日本語と同じであるが、文章を書くとき、助詞以外の各単語を離して書かなくてはならない。例えば、先ほどの例で示すと

「私は　茨城大学の　学生です。」

のように分かち書きをする必要がある。この空白を基準に、前後の文字を取らないことでノイズを減らすことができる。よって、この文章からbi-gramで取り出す文字列は、「私は」「茨城」「城大」「大学」「学の」「学生」「生で」「す。」の8つである。また、文章の終わりを示す、「。」や「、」なども同様にする。意味のないタームを無くす基準を表3に示す。

表2.3：意味のないタームを無くす基準

文字	処理方法
空白	前後の文字は取らない
。（韓国では .）	〃
、	〃
括弧	この文字は取らない
特殊記号 (! @ # $ % ^ & * + / = ? ¥ | ~ ` ' " ; : - _)	〃

この表2.3に従い300個のスニペットからbi-gramで二文字づつ取得した。この共起関係を付録Aに示し、共起関係から300個のスニペットを比較した共起頻度を付録Bに示す。また、これらの動作をするプログラムを作成した。

プログラムの流れを簡単に説明する。　図2.3のように[1]、スニペットから2文字づつ取得し、順番に従って番号を与える。プログラムでは連想配列（ハッシュ）を使い、ここでは文字列が「キー」となり、番号が「値」となる。「キー」は重複してはならないため、すでに取得した文字列には「キー」を与えないようにする。

[1] 実験では、韓国語のスニペットを用いるが、ここでは説明のために日本語のスニペットを利用する

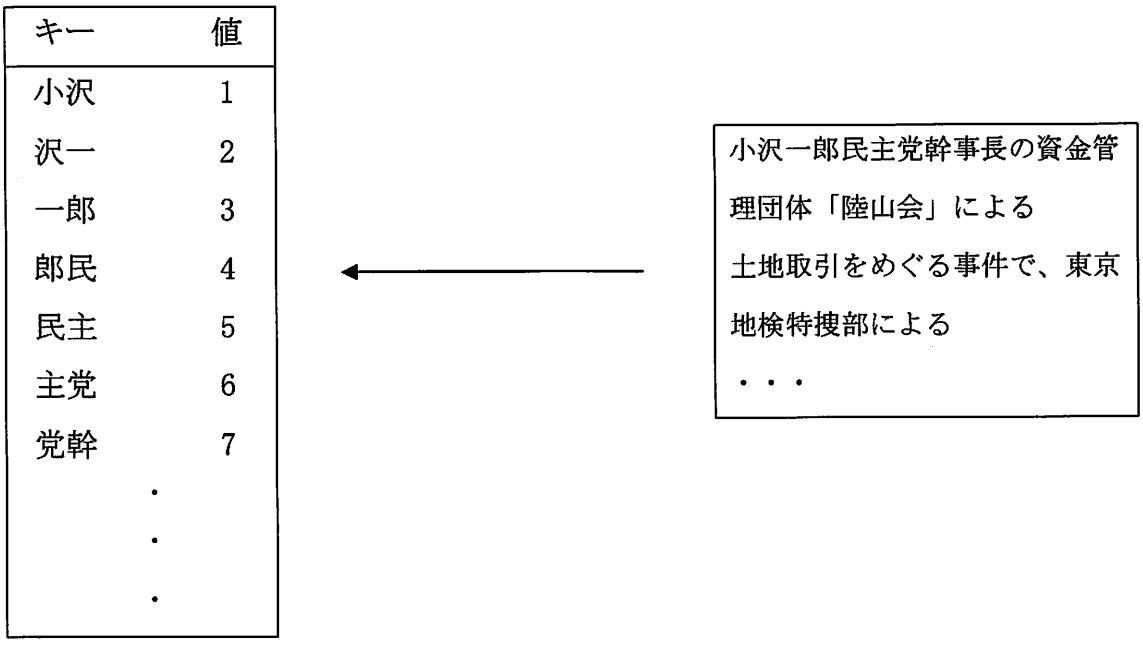


図2.3：共起関係を取るプログラムの流れ

全てのスニペット（300個）から取得をした後は、取得した文字列を図2.4のように全てのスニペットと比較し、各スニペットにどれくらいの共起関係があるかを調べる。例えば、図2.4では、1番目のスニペットには番号が１である「小沢」という文字列が2個ある。そこで二重連想配列を作成し、最初の「キー」にはスニペットの番号である1を、二番目の「キー」には文字列「小沢」を、「値」にはその文字列の個数2を代入する。

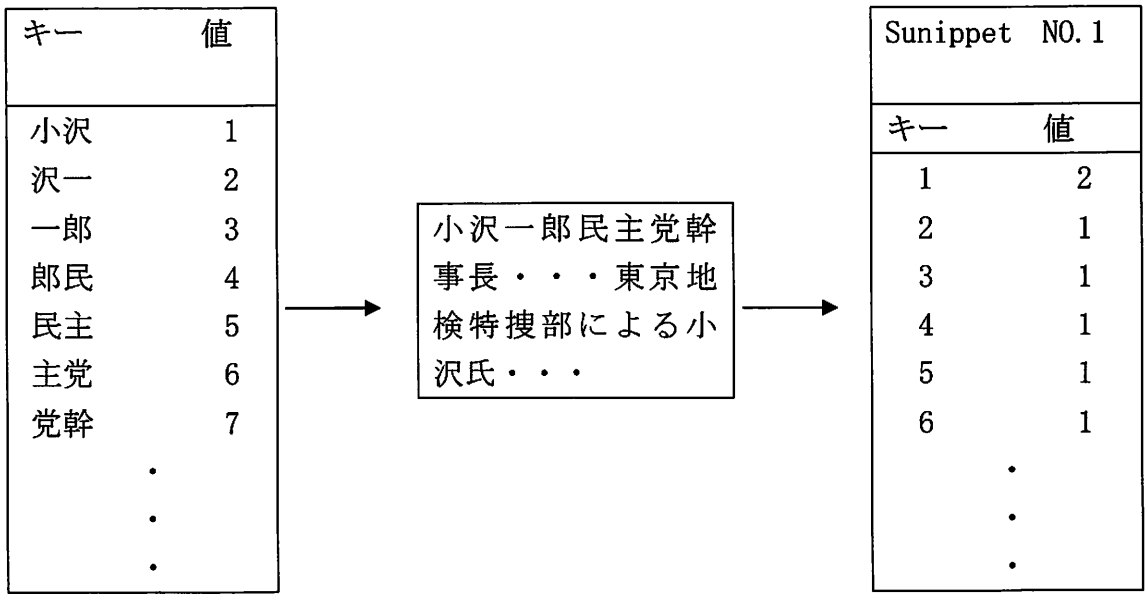


図2.4：共起頻度を求めるプログラムの流れ

次は、この二重連想配列を実験で利用するために、LibSVM用のデータファイルの書式[2]に従って出力し、ファイル化する。図2.5のように、最初のスニペットの共起頻度を1行目に書き、2番目のスニペットを2行目と、順次的に書く。「キー」と「値」の間にはコロン（：）を入れる。

[2] 詳細は3.3節に述べる

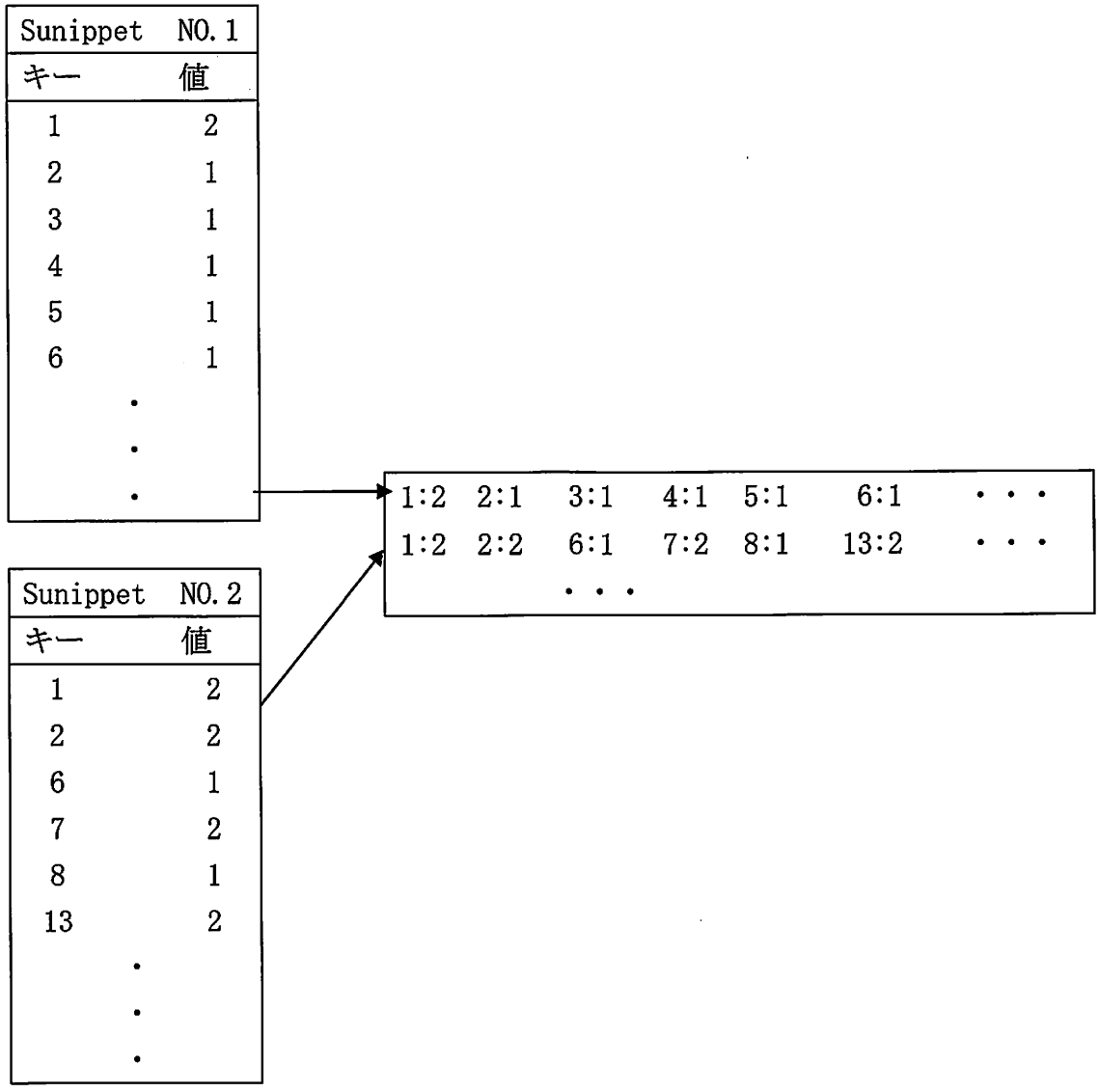


図2.5：二重連想配列をデータファイルの書式に書き込む

上記を実行させるプログラムを付録Cに示す。

第３章　ONE CLASS SVM

３．１　SVM

サポートベクターマシン(Support Vector Machine, SVM) は２クラス判別を行う教師付き学習アルゴリズムである[6]。教師信号 y はクラス1 に所属するデータには1、クラス2 に所属するデータには -1 を与える。SVMは入力データ x を非線形関数 θ によって高次元空間へ写像し、高次元空間上で線形判別を行う。θ を十分な高次元へ写像する非線形関数に決めると、2 クラスの訓練データ $(x_i, y_i),\quad i=1,...,n$ は必ず線形分離可能になり、$\phi(x)$ についての線形関数を $f(x)=\langle w,\phi(x)\rangle+b$ として、

$$y_i f(x_i)\geq 1,\quad i=1,...,n \tag{3.1.1}$$

となるようなパラメータ w、b が存在する。ここで、$\langle\cdot,\cdot\rangle$ は内積を表す。

SVM による判別関数の学習はマージン最大化基準によって行われる。マージンとは超平面 $\langle w,\phi(x)\rangle+b=0$ からデータまでの距離のことであり、マージンを大きくとることによって高い汎化能力が期待される。l_2 ノルムを $\|\cdot\|_2$ とすると、点 $\phi(x_i)$ と超平面との距離は $|f(x_i)|/\|w\|_2=y_i f(x_i)/\|w\|_2$ と計算される。すると式(3.1.1) より、

$$\frac{y_i f(x_i)}{\|w\|_2}\geq\frac{1}{\|w\|_2},\ i=1,...,n \tag{3.1.2}$$

が成り立つ。従って、マージンを最大にするように判別関数を決定するには、式(3.1.1) を満たしながら $\|w\|_2$ を最小化するような w を見つければいいということになる。

$\|w\|_2^2$ を最小化するような w は $\|w\|_2$ も最小化するので、SVM における判別関数の学習は次のように定式化される。

$$\min_{w,b} \frac{1}{2}\|w\|_2^2$$

$$s.t \quad y_i\{\langle w, \phi(x_i)\rangle + b\} \geq 1, \quad i = 1,...,n \qquad (3.1.3)$$

最適化問題(3.1.3) は双対問題を導出して解く。非負のラグランジュ乗数 $\alpha = (a_1,...,\alpha_n)$ を導入し、以下のようなラグランジュ関数を定義する。

$$L(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{n} \alpha_i [y_i\{\langle w, \phi(x_i)\rangle + b\} - 1] \qquad (3.1.4)$$

最適解において、主問題の変数 w と b の偏微分係数が0 になることにより、

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i \phi(x_i) = 0 \qquad (3.1.5)$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \qquad (3.1.6)$$

を得る。式(3.1.5)-(3.1.6) をラグランジュ関数(3.1.4) に適用することで主問題(3.1.3) の双対問題は以下のように表される。

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j \langle \phi(x_i), \phi(x_j) \rangle$$

$$s.t \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \qquad (3.1.7)$$

$$\alpha_i \geq 0, \quad i = 1, \ldots, n$$

高次元空間上の内積$\langle \phi(x_i), \phi(x_j) \rangle$を求めるのは非常に困難であるが、後で述べるカーネルトリックによって、この問題は回避できる。双対問題(3.1.7) の解α^* はデータ点x_iが超平面$\langle w, \phi(x_i) \rangle + b = 1$、もしくは$\langle w, \phi(x_i) \rangle + b = -1$のどちらかにの上に乗っている場合のみ非ゼロとなる。こういった点のことを「サポートベクター」と呼ぶ。一般に、サポートベクターは他の点と比較して非常に数が少ない。そして式(3.1.5) より、その少ないサポートベクターが超平面を構成していることがわかる。主問題(3.1.3) の解をw^*, b^* とすると判別関数f は

$$f(x) = \langle w^*, \phi(x) \rangle + b^*$$

$$= \sum_{i=1}^{n} \alpha_i^* y_i \langle \phi(x_i), \phi(x) \rangle + b^* \qquad (3.1.8)$$

となる。また切片b^* に関しては、サポートベクターが超平面$\langle w, \phi(x_i) \rangle + b = 1$、もしくは$\langle w, \phi(x_i) \rangle + b = -1$のどちらかにの上に乗っているということを利用して、

$$b^* = \sum_{i=1}^{n} \alpha_i^* y_i \langle \phi(x_i), \phi(x_s) \rangle - y_s \qquad (3.1.9)$$

と計算できる。ここでのx_sは任意のサポートベクターであり、y_sはx_sに対応する教師信号である。そして未知のデータxがどちらのクラスに属するかは

$$F(x) = sign(f(x)) \tag{3.1.10}$$

によって決定される。

SVM では高次元空間上での内積$\langle \phi(x_i), \phi(x_j) \rangle$の計算が必要となるが、一般にこれは計算が困難である。そこで、ある関数kを用意し、その関数kがある高次元空間上での内積、$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$になっていないかと逆に考える。どのような関数であれば内積の形で書けるのかは、次のMercer の定理によって与えられる。

定理3.1.1 (Mercer の定理)

$u, v \in X$の関数kが内積の形で書ける必要十分条件は

(i) kが対称関数、つまり、$k(u,v) = k(v,u)$である。
(ii) kが半正定値、つまり、任意の関数gに対して

$$\int_X \int_X k(u,v)g(u)g(v)dvdu \geq 0$$

を満たすことである。

一般にMercer の定理を満たす関数kをMercer カーネルと呼ぶ。そして内積の計算をカーネルの計算に置き換える手法を「カーネルトリック」と呼ぶ。カーネルは計算が容易なものが望ましい。

代表的なものとして、

$$\text{線形カーネル:}\quad k(u,v)=(\langle u,v\rangle) \tag{3.1.11}$$

$$\text{多項式カーネル:}\quad k(u,v)=(1+\langle u,v\rangle)^q \tag{3.1.12}$$

$$\text{ガウシアンカーネル:}\quad k(u,v)=\exp(-\sigma\|u-v\|^2) \tag{3.1.13}$$

$$\text{シグモイドカーネル:}\quad k(u,v)=\tanh(r\langle u,v\rangle-l) \tag{3.1.14}$$

がある。式(3.1.7) の目的関数は、カーネルを用いて次のように書き換えることができる。

$$\sum_{i-1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(x_i,x_j) \tag{3.1.15}$$

また、判別関数(3.1.8) は以下のように書き換えることができる。

$$f(x)=\sum_{i=1}^{n}\alpha_i^* y_i k(x_i,x)+b^* \tag{3.1.16}$$

十分な高次元空間上では訓練データは必ず線形分離可能となり、全てのデータについて正しい判別をする判別関数を求めることが可能である。しかし、データにノイズが含まれている場合や分布がクラス間でオーバーラップしているような場合、全ての訓練データについて正しく判別しようとするとオーバーフィッティングを起こすことがある。この問題を解決するために、「ソフトマージン」を用いるソフトマージン法を導入する。ソフトマージン法では、マージン$\frac{1}{\|w\|_2}$を最大としながら、いくつかのサンプルが超平

面を超えることを許す。その超平面を超えた程度を非負のスラック変数$\xi=(\xi_1,...,\xi_n)$によって表し、その和がなるべく小さくなるようにする。このとき最適な判別関数を求める問題は

$$\begin{aligned}&\min_{w,b,\xi}\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{n}\xi_i\\&s.t.\quad y_i(\langle w,\phi(x_i)\rangle+b)\geq 1-\xi_i\\&\xi_i\geq 0,\quad i=1,...,n\end{aligned}\qquad(3.1.17)$$

という最適化問題になる。ここでのC はマージンの大きさ(第1 項) と超平面からのはみ出しの程度(第2 項) のトレードオフを調整する定数である。最適化問題(3.1.17) は双対問題を導出して解く。非負のラグランジュ乗数$\alpha=(\alpha_1,...,\alpha_n)$, $\beta=(\beta_1,...,\beta_n)$を導入し、以下のようなラグランジュ関数を定義する。

$$L(w,b,\xi,\alpha,\beta)=\frac{1}{2}\|w\|_2^2+C\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i[y_i\{\langle w,\phi(x_i)\rangle+b\}-1]-\sum_{i=1}^{n}\beta_i\xi_i$$

(3.1.18)

最適解において、主問題の変数wとb、ξの偏微分係数が0 になることにより、

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i \phi(x_i) = 0 \tag{3.1.19}$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.1.20}$$

$$\frac{\partial L}{\partial \xi} = C - \alpha - \beta = 0 \tag{3.1.21}$$

を得る。式(3.1.19)-(3.1.21) をラグランジュ関数(3.1.18) に適用し、内積をカーネルに置き換えると、主問題(3.1.17) の双対問題は以下のように表される。

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j k(x_i, x_j)$$

$$s.t \quad \sum_{i=1}^{N} \alpha_i y_i = 0 \tag{3.1.22}$$

$$0 \le \alpha_i \le C, \quad i = 1, \ldots, n$$

ソフトマージン法の場合、超平面をはみ出す点が存在する（以下、外れ点と呼ぶ）。最適なラグランジュ乗数α^* はデータ点x_iが外れ点であれば$\alpha_i^* = C$であり、超平面上の点であれば$0 \le \alpha_i^* \le C$、それ以外の点ならば$\alpha_i^* = 0$となる。外れ点及び超平面上の点はサポートベクターと呼ばれ、これらのサポートベクターが超平面を構成する。

得られる判別関数 f は式(3.1.8) と同様に

$$f(x)=\sum_{i=1}^{n}\alpha_i^* y_i k(x_i,x)+b^* \qquad (3.1.23)$$

となる。ここで切片 b^* は、任意の超平面上の点を x_s、対応する教師信号を y_s として

、

$$b^*=\sum_{i=1}^{n}\alpha_i^* y_i k(x_i,x_s)-y_s \qquad (3.1.24)$$

と計算できる。

3．2　ONE CLASS SVM

SVMは本来、2クラス判別を行う教師付き学習のアルゴリズムである。ところが近年、SVM をデータの高密度領域を推定する領域判別問題に適用した教師なし学習のアルゴリズムであるOne Class SVMが注目されている[7]。One Class SVM では、 入力データを適切なカーネルに対応する非線形写像ϕを用いて高次元特徴空間上に写像し、それらをその特徴空間上で原点から最大のマージンになるような超平面分離する。One Class SVM は、適切なカーネルに対応する写像ϕによって写された先の特徴空間上で、
超平面：$\langle w,\phi(x)\rangle-\rho=0$ によって判別を行う。多くの訓練パターンが判別超平面を挟んで特徴空間上の原点との反対側にあることを維持しながら超平面と原点のマージンを大きくすることで高密度領域の推定を目指す。未知のパターンxが高密度領域に含まれるか否かは、

$$f(x)=w^T\phi(x) \tag{3.2.1}$$

を取ると、サンプルは$f(x^{(1)})$;・・,$f(x^{(n)})$のように1次元のデータとなる。

このデータをしきい値$\rho>0$で二つに分け、$\rho\le f(x^{(i)})$となるサンプルを正常値、

$\rho>f(x^{(i)})$となるサンプルを外れ値に分類する（図3.1）。

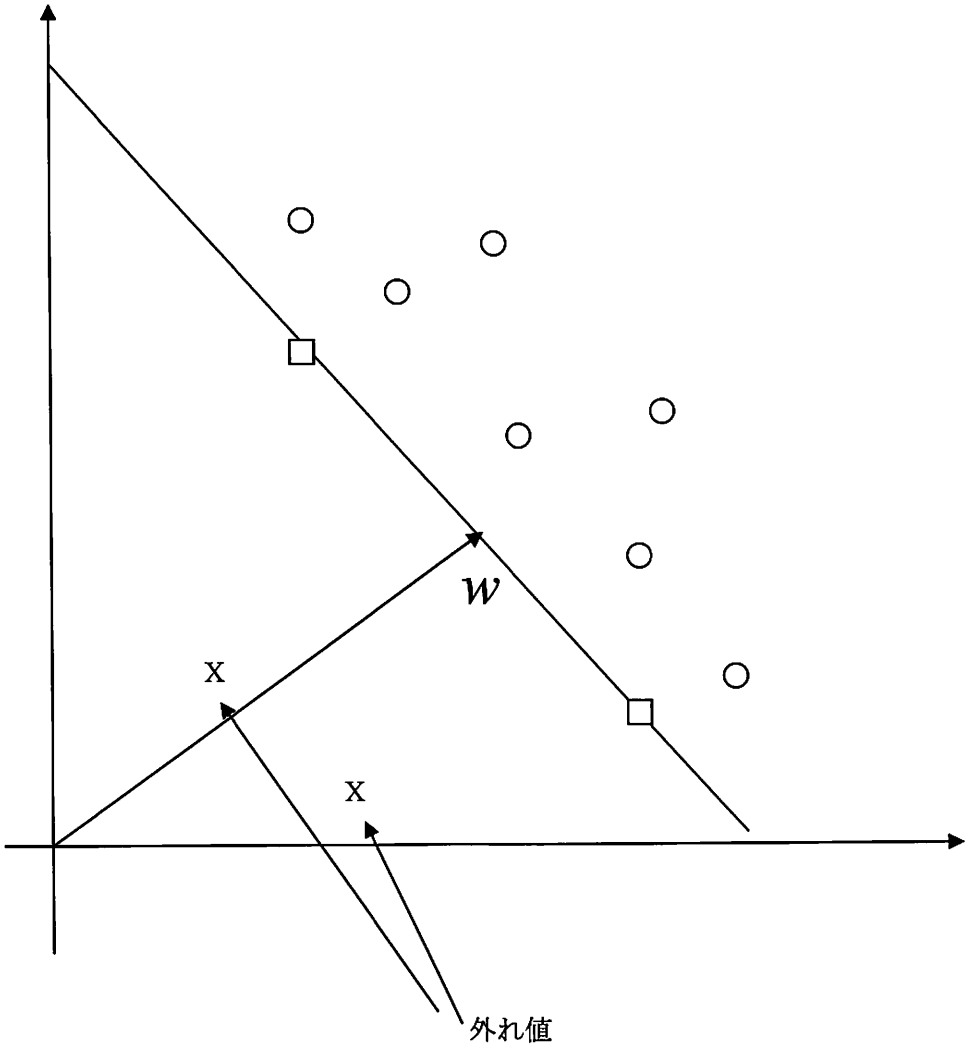


図3.1：１クラスν-サポートベクトルマシンの外れ値検出法の概念図

しきい値のどちら側も正常値と定義することはできるが、wがほぼデータのクラスタの向きに対応すると考えると、クラスタは$f(x)$のある値の周辺に集まるはずであり、それに直交するデータはwとない席を取ると0に近い値を取ると考えられる。したがって、あるしきい値以上の値のときを正常と考えるのが自然である。

式 (3.2.1)は、

$$F(x) = sign(f(x) - \rho) \tag{3.2.2}$$

によって評価し、$F(x)=+1$のとき高密度領域に含まれると判別する。このとき集合$\{x \mid F(x)=+1\}$が高密度領域の推定量となる。

訓練データをx_i, $i=1,...,n$とすると、マージンの最大化とペナルティ項を調整するパラメータ$\nu \in (0,1]$を用いて、上記の問題は以下の2 次計画問題として定式化できる。

$$\min_{w,\rho,\xi} \frac{1}{2} \| w \|_2^2 - \nu n \rho + \sum_{i=1}^{n} \xi_i \tag{3.2.3}$$

$$s.t. \langle w, \phi(x_i) \rangle \geq \rho - \xi_i$$

$$\xi_i \geq 0, \quad i = 1,...,n$$

これは2 クラス判別手法であるν-SVM において全訓練パターンがクラス1 に属し、原点を唯一のクラス2 に属するデータとみなして学習することと同じである。

最適化問題(3.2.3)は双対問題を導出して解く。非負のラグランジュ乗数

$\alpha = (\alpha_1,...,\alpha_n), \beta = (\beta_1,...,\beta_n)$を導入し、以下のようなラグランジュ関数を定義する。

$$L(w,\xi,\rho,\alpha,\beta) = \frac{1}{2} \| \omega \|_2^2 - \nu n \rho + \sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \{ \langle w, \phi(x_i) \rangle - \rho \} - \sum_{i=1}^{n} \beta_i \xi_i \tag{3.2.4}$$

最適解において、主問題の変数 w と ρ,ξ の偏微分係数が0 になることにより、

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i \phi(x_i) = 0 \tag{3.2.5}$$

$$\frac{\partial L}{\partial \rho} = -\nu n + \sum_{i=1}^{n} \alpha_i = 0 \tag{3.2.6}$$

$$\frac{\partial L}{\partial \xi} = 1 - \alpha - \beta = 0 \tag{3.2.7}$$

を得る．式(3.2.5)-(3.2.7) をラグランジュ関数(3.2.4) に適用し、内積をカーネルに置き換えると、主問題(3.2.3) の双対問題は以下のように表される。

$$\max_{\alpha} -\frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

$$s.t. \quad \sum_{i=1}^{n} \alpha_i = \nu n \tag{3.2.8}$$

$$0 \le \alpha_i \le 1, \quad i = 1, \ldots, n$$

問題(3.2.8) の解を α^* とすると、

$$f(x) = \sum_{i=1}^{n} \alpha_i^* k(x_i, x) \tag{3.2.9}$$

として、未知のデータxが高密度領域に含まれるかは

$$F(x) = sign(f(x) - \rho^*) \tag{3.2.10}$$

によって決定される。ここでρ^*は任意の超平面上の点をx_sとして、

$$\rho^* = \sum_{i=1}^{n} \alpha_i^* k(x_i, x_s) \tag{3.2.11}$$

と計算できる。このとき集合$\{x \mid F(x) = +1\}$が高密度領域の推定量となる。

なお、$\nu \in (0,1]$は所与のパラメータで、One Class SVMにおいては次項の定理(ν-property)が成立し、推定された高密度領域に含まれない外れ点の割合はνで押さえられる。つまり、νによって外れ点の割合をコントロールすることが可能となる。

3．3　LibSVM

LibSVMは台湾国立大学のLinらによって作られたSVMのライブラリである。サポートベクタ分類器（C-SVC、nu-SVC）、回帰分析（epsilon-SVR、nu-SVR）、分布評価（One Class SVM）のための統合ソフトである。以下は、開発者のホームページ[8]に記載されているLibSVMの特徴である。

・異なったSVM の計算式が用意されている
・効率的なマルチクラス分類ができる
・モデル選択のためのクロスバリデーション
・可能性予測
・偏ったデータのための、重みつきSVM
・C++ とJava のソースコード
・SVM 分類と回帰分析のGUI デモンストレーション
・Python、R（またはSplus）、MATLAB、Perl、Ruby、Weka、CLISP、LabVIEW のAPI

LibSVMのWindows版は、svm-scale、 svm-train、svm-predictの３つのコマンドからなっており、svm-scaleで訓練セットをスケーリングした後、 svm-trainでmodelを構築し、svm-predictによって予測を行う。

以下に、この3つのコマンドについて説明する。

・svm-scale

作成した訓練セットを引数に従ってスケーリングしてくれるコマンドである。Optionは以下のとおりである。

・-l lower（スケーリングの最小値の指定、デフォルト -1）
・-u upper（スケーリングの最大値の指定、デフォルト 1）
・-y y_lower y_upper
・-s savefile
・-r restorefile

-l、-uオプションはスケーリングの最大、最小値を指定するオプションでデフォルトはそれぞれ、-1、1である。 [0,1]でスケーリングした方が計算時間は早くなるようだが、基本的にはデフォルトのままでよいと思われる。
-s オプションと-rオプションは訓練セットとテストセットを同じようにスケーリングしたいときに、

```
svm-scale -s scaling_parameters traing_data > scaled_traing_data
svm-scale -r scaling_parameters test_data > scaled_test_data
```

のようにする。scaling_parametersには各特徴ベクトルの最大値と最小値が記録されており、これによってテストセットを訓練セットと同じようにスケーリングしてやることができる。このオプションを使わないと、テストセットはテストセットの最大値と最小値に従ってスケーリングされてしまうので注意が必要である。

・svm-train

訓練セットから予測のためのモデルを生成するコマンドである。Optionは以下のとおりである。

・-s svm_type : SVMタイプの指定（デフォルト 0）

0 -- C-SVC

1 -- nu-SVC

2 -- one-class SVM

3 -- epsilon-SVR

4 -- nu-SVR

・-t kernel_type : カーネル関数の指定（デフォルト 2）

0 -- 線形(linear)： $u'\times v$

1 -- 多項式(polynomial)： $(gamma\times u'\times v+coef0)^{degree}$

2 -- RBF(radial basis function)： $\exp(-gamma\times|u-v|^2)$

3 -- シグモイド(sigmoid)： $\tanh(gamma\times u'\times v+coef0)$

・-d degree : カーネル関数のdegreeの指定（デフォルト 3）

・-g gamma : カーネル関数のgammaの指定（デフォルト$\frac{1}{k}$，kは訓練セットのインデックスの最大の値、つまり入力ベクトルの次元）

・-r coef0 : カーネル関数のcoef0の指定（デフォルト 0）

・-c cost : コストパラメータの指定（C-SVC, epsilon-SVR、 and nu-SVRで使用、デフォルト 1）

・-n nu : nuパラメータの指定（nu-SVC、 one-class SVM、 nu-SVRで使用、デフォルト 0.5）

・-p epsilon : set the epsilon in loss function of epsilon-SVR（デフォルト 0.1）

・-m cachesize : 使用キャッシュメモリサイズの指定(単位MB、 デフォルト 40)

・-e epsilon : 終了の閾値の設定（デフォルト 0.001）

・-h shrinking: whether to use the shrinking heuristics, 0 or 1（デフォルト 1）

・-wi weight: クラスiに対するコストパラメータの重み、$weight \times C$（デフォルト 1）

・-v n: n-foldのクロスバリデーションを実行

用意されていないストリングカーネルなどが使いたいときは svm.cppのKernelクラスを修正する必要がある。-vオプションでn-foldのクロスバリデーションが自動的でできるので性能の評価は手軽に行う。学習がなかなか終わらないという際には、-cオプションのコストパラメータを小さくしてみるか-eオプションで終了条件を緩めてやればよい。

・svm-predict

予測を実行する。

svm-predict test_file model_file output_file

オプションは特にない。

LibSVMのデータファイルの書式は以下のようにする。

```
クラスラベル 番号:数値 番号:数値 ...
クラスラベル 番号:数値 番号:数値 ...
クラスラベル 番号:数値 番号:数値 ...
クラスラベル 番号:数値 番号:数値 ...
・・・
```

本研究で使うOne Class SVMでは、クラスが1であるため、クラスラベルは全て1となる。番号は共起の位置（ID)を示す番号であり、数値は共起がどのぐらい起きたかを示す共起頻度となる。

第４章　実験

４．１　実験概要

本実験では、韓国の異なる二人の有名人の同姓同名を検索し、上位300件の検索結果のスニペットをデータとして扱った。表4.1に一人目の同姓同名のデータを示す。また、表4.2に二人目の同姓同名のデータを示す。

表4.1：データ1-　同姓同名のデータ

検索した名前	박명수　(パク・ミョンス)	
利用した検索エンジン	Google	
検索をした日付	2009年9月12日	
同姓同名の職業	件数	比率
芸能人	264	88%
プロゲーマ	28	9%
一般人　or スニペットではわからない	7	3%

表4.2：データ2-　同姓同名のデータ

検索した名前	이영애　　（イ・ヨンエ）	
利用した検索エンジン	Google	
検索をした日付	2009年12月17日	
同姓同名の職業	件数	比率
芸能人	254	85%
プロゲーマ	46	15%

データ1とデータ2のスニペットの内容を、それぞれ付録Dと付録Eに示す。

４．２　LibSVMの結果

3.3節で述べたように、LibSVMのコマンド「svm-train」では、カーネルを選択し、各パラメータを指定することができる。そこで、3つのカーネル（多項式、RBF、シグモイド）それぞれを使い実験することができた。本実験で、使用したパラメータはnuパラメータとgammaパラメータである。nuパラメータはデフォルトの0.5から0.001まで変更しながら実行した。また、gammaパラメータはデフォルトの0.0033[3]から0.01まで変更しながら実行をした。LibSVMの結果は、「accuracy」と「classification」が出力される。

「accuracy」とは、正答率、すなわち正常値と外れ値を分類した比率である。「classification」とは、正常値と外れ値を分類した個数を示す。例えば、LibSVMの結果で300個のスニペットから270個が正解（30個が外れ値）と判定した場合、「Classification」は270/300となり、「accuracy」は90%となる。

データ1とデータ2に関するLibSVMの結果をそれぞれ表4.3と表4.4に示す。

[3] 正確には$\frac{1}{k}$（300個の入力次元なので　$\frac{1}{300}=0.0033\cdots$）

表4.3：データ1のLibSVMの結果

kernel	nu	gamma	accuracy	classification
多項式	0.5 (default)	0.0033	48.3%	145/300
		0.001	49.6%	149/300
		0.01	48.3%	145/300
	0.01	0.0033	86.7%	260/300
		0.001	94.3%	283/300
		0.01	94.3%	283/300
	0.001	0.0033	94.3%	283/300
		0.001	100%	300/300
		0.01	100%	300/300
RBF (default)	0.5 (default)	0.0033	52%	156/300
		0.001	48.3%	145/300
		0.01	53%	159/300
	0.01	0.0033	92.3%	277/300
		0.001	92.7%	278/300
		0.01	83.3%	250/300
	0.001	0.0033	**89%**	**267/300**
		0.001	90%	270/300
		0.01	79.7%	239/300
シグモイド	0.5 (default)	0.0033	48.3%	145/300
		0.001	47%	141/300
		0.01	0%	0/300
	0.01	0.0033	91%	273/300
		0.001	29%	87/300
		0.01	0%	0/300
	0.001	0.0033	55%	165/300
		0.001	0%	0/300
		0.01	0%	0/300

表4.4：データ2のLibSVMの結果

kernel	nu	gamma	accuracy	classification
多項式	0.5 (default)	0.0033	6.33%	19/300
		0.001	7%	21/300
		0.01	46%	138/300
	0.01	0.0033	0.667%	2/300
		0.001	0.667%	2/300
		0.01	48%	144/300
	0.001	0.0033	0.333%	1/300
		0.001	0.333%	1/300
		0.01	0.333%	1/300
RBF (default)	0.5 (default)	0.0033	50%	150/300
		0.001	48%	144/300
		0.01	49.7%	149/300
	0.01	0.0033	96.7%	290/300
		0.001	97.3%	292/300
		0.01	93.7%	281/300
	0.001	0.0033	91.3%	274/300
		0.001	95.3%	286/300
		0.01	93%	279/300
シグモイド	0.5 (default)	0.0033	50.7%	152/300
		0.001	50.7%	152/300
		0.01	49.7%	149/300
	0.01	0.0033	49%	147/300
		0.001	73%	219/300
		0.01	**81.3%**	**244/300**
	0.001	0.0033	0.667%	2/300
		0.001	35.7%	107/300
		0.01	73%	219/300

LIBSVMのマニュアルではRBFカーネルを推奨しているが、本実験の場合もデータ1やデータ2に対してRBFカーネルが最善の結果となった。多項式カーネルやシグモイドカーネルの場合は、パラメータによるが、0%に近いと判定したり、100%と判定している。表4.3と表4.4の「accuracy」と「classification」に太文字をした部分があるが、表4.1と表4.2に基づき、データ1とデータ2それぞれに対して、一番正解に近づいている値を太文字にした。

４．３　評価

LibSVMの結果は、LibSVMが正常値と判断したスニペットには１を、外れ値と判断したスニペットには-1が出力される。そこで、実際にこれらの結果が正しいかを評価する。

評価は、適合率と再現率を組み合わせたF値を用いて評価する。

適合率（精度、precision）は検索結果として得られた集合中にどれだけ検索に適合した文書を含んでいるかという正確性の指標であり、再現率（recall）は検索対象としている文書の中で検索結果として適合している文書（正解文書）のうちでどれだけの文書を検索できているかという網羅性の指標である。

本実験では、適合率は

$$\text{適合率}=\frac{\text{その内、実際に外れ値の個数}}{\text{LibSVMで外れ値と判定した個数}}$$

となり、再現率は、

$$\text{再現率}=\frac{\text{その内、実際に外れ値の個数}}{\text{全体の外れ値の個数}}$$

となる。

適合率を上げれば再現率が下がり、再現率を上げれば適合率が下がる傾向にあるため、F値（F-measure, F-score）を用いる。F値とは、再現率と適合率の調和平均であり、

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

によって求められる。

表4.3と表4.4のそれぞれの結果に対し、適合率、再現率及びF値を求めた。これらの結果を表4.5と表4.6にそれぞれ示す。

表4.5 ：データ１の評価

kernel	Nu	gamma	accuracy	precision	recall	F-measure
多項式	0.5 (default)	0.0033	48.3%	0.123	0.542	0.2
		0.001	49.6%	0.119	0.514	0.194
		0.01	48.3%	0.123	0.542	0.2
	0.01	0.0033	86.7%	0.175	0.2	0.187
		0.001	94.3%	0.176	0.086	0.115
		0.01	94.3%	0.176	0.086	0.115
	0.001	0.0033	94.3%	0.118	0.057	0.077
		0.001	100%	-[4]	-	-
		0.01	100%	-	-	-
RBF (default)	0.5 (default)	0.0033	52%	0.125	0.514	0.201
		0.001	48.3%	0.129	0.571	0.211
		0.01	53%	0.121	0.486	0.193
	0.01	0.0033	92.3%	0.174	0.114	0.138
		0.001	92.7%	0.182	0.114	0.140
		0.01	83.3%	0.14	0.2	0.165
	0.001	0.0033	89%	0.21	0.2	0.206
		0.001	90%	0.167	0.143	0.154
		0.01	79.7%	0.115	0.2	0.146
シグモイド	0.5 (default)	0.0033	48.3%	0.126	0.543	0.2
		0.001	47%	0.119	0.543	0.196
		0.01	0%	0.117	1	0.206
	0.01	0.0033	91%	0.111	0.086	0.097
		0.001	29%	0.127	0.771	0.218
		0.01	0%	0.117	1	0.210
	0.001	0.0033	55%	0.126	0.486	0.2
		0.001	0%	0.117	1	0.210
		0.01	0%	0.117	1	0.210

[4] 外れ値がないと結果が出たため、適合率、再現率及びF-値を求めることができない

表4.6：データ2の評価

カーネル	nu	gamma	accuracy	precision	recall	F-measure
多項式	0.5 (default)	0.0033	6.33%	0.15	0.915	0.262
		0.001	7%	0.154	0.915	0.264
		0.01	46%	0.142	0.489	0.220
	0.01	0.0033	0.667%	0.158	1	0.272
		0.001	0.667%	0.158	1	0.272
		0.01	48%	0.147	0.489	0.227
	0.001	0.0033	0.333%	0.157	1	0.272
		0.001	0.333%	0.157	1	0.272
		0.01	0.333%	0.157	1	0.272
RBF (default)	0.5 (default)	0.0033	50%	0.22	0.70	0.335
		0.001	48%	0.218	0.723	0.335
		0.01	49.7%	0.219	0.702	0.333
	0.01	0.0033	96.7%	0.5	0.106	0.175
		0.001	97.3%	0.375	0.064	0.109
		0.01	93.7%	0.474	0.191	0.273
	0.001	0.0033	91.3%	0.385	0.213	0.274
		0.001	95.3%	0.357	0.106	0.164
		0.01	93%	0.381	0.170	0.235
シグモイド	0.5 (default)	0.0033	50.7%	0.216	0.681	0.328
		0.001	50.7%	0.196	0.617	0.297
		0.01	49.7%	0.205	0.660	0.313
	0.01	0.0033	49%	0.183	0.596	0.28
		0.001	73%	0.235	0.404	0.297
		0.01	81.3%	0.25	0.298	0.272
	0.001	0.0033	0.667%	0.154	0.979	0.267
		0.001	35.7%	0.192	0.787	0.308
		0.01	73%	0.160	0.277	0.203

第5章　考察

データ1の評価は、カーネルがシグモイドの時、nuパラメータが0.01でgammaパラメータが0.001の時にF値が0.218と一番高かった。また、データ2の評価では、カーネルがRBFの時、nuパラメータがデフォルトの0.5でgammaパラメータがデフォルトの0.0033と0.001の時にF値が0.335と一番高かった。

実験結果から、本研究で行った外れ値の抽出はデータ1とデータ2のF値の最大値が0.5以下ということもあり、あまりいい結果ではなかった。この原因として、bi-gramで共起関係を取る際のノイズが考えられる。表3で意味のないタームを減らす基準を示したがこの基準が不十分だと思われる。例えば、空白や「、」の文字列が来ると前後の文字は取らないようにしたが、括弧や特殊記号の場合はこれらの文字は取らないが、前後の文字は取るようにした。

他の反省点としては、パラメータの設定がある。LibSVMでは「grid.py」という最適なパラメータを見つけてくれるファイルがあるが、One Class SVMは教師なし学習であるため、このファイルを使い最適なパラメータを探すことはできなかった。したがって、ヒューリスティックにパラメータを指定するしかなかった。筆者が表6と表7で指定したnuパラメータとgammaパラメータの他にもいろいろなパラメータ、いろいろな数値で実行すべきであった。

また、LibSVMの結果である「accuracy」が高ければF値が高いとは限らないことが分かった。表4.6では、「accuracy」が48～50%であるF値が、「accuracy」が90%であるF値より高かった。外れ値検出を行うという概念では「accuracy」が低くてもF値が高いと良い結果だが、正常値を検出するという概念では「accuracy」が低いのは良くない結果であった。

第6章　結論

本研究では同姓同名の曖昧性を解消を目指し、One Class SVMを用いて非該当の人名ページの検出を行った。

2人のスニペットの文書データをbi-gramで表現し、One Class SVMが使えるライブラリLibSVMを利用して外れ値検出を行った。その評価として、適合率と再現率を組み合わせたF値を求めたが、F値の最大値が0.5以下であり、よい結果を得ることができなかった。その原因として、bi-gramでスニペットから文字列を取る際のノイズとヒューリスティックに指定したパラメータの設定問題が分かった。今後、これらの点を改良した外れ値検出を行う。

参考文献

[1] C.E. Shannon: “Predication and Entropy of Printed English”, The Bell System Technical Journal, 1951.

[2] P. Elias: “Gamma Code, Delta Code: Universal codewords sets and representations of the integers”, IEEE Trans. on Information Theory, IT-21(2), pp.194-203, 1975.

[3] 近藤みゆき「n-gram統計による語形の抽出と複合語　—平安時代語の分析から—」(『日本語学』Vol.20、2001年8月号)

[4] http://nlp.nagaokaut.ac.jp/n-gram

[5] 内山将夫, コーパスベースの機械翻訳, http://www2.nict.go.jp/x/x161/members/mutiyama/saitama-u.html

[6] 赤穂昭太郎, カーネル多変量解析 - 非線形データ解析の新しい展開, 2008年

[7] Scholkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. “Estimating the Support of a High Dimensional Distribution,” *Neural Computation*, 13(7): 1443-1471, 2001.

[8] C. C. Chang, C. J. Lin, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

付録A
データ1の共起関係

공지 => 0001
인사 => 0002
사이 => 0003
이드 => 0004
거성 => 0005
박명 => 0006
명수 => 0007
갤러 => 0008
러리 => 0009
통합 => 0010
업러 => 0011
러됴 => 0012
개그 => 0013
그맨 => 0014
수에 => 0015
관련 => 0016
련된 => 0017
사진 => 0018
진과 => 0019
...

라마 => 3601
밥줘 => 3602
줘를 => 3603
코믹 => 3604
버전 => 3605
디헸 => 3606
기리 => 3607
영중 => 3608
선우 => 3609
고액 => 3610
논란 => 3611
란과 => 3612
불만 => 3613
만을 => 3614
토로 => 3615
로헸 => 3616
황금 => 3617
금어 => 3618
어장 => 3619
무릎 => 3620
릎팍 => 3621
팍도 => 3622
도사 => 3623
사도 => 3624
구현 => 3625
현하 => 3626
연헸 => 3627
길이 => 3628
휠씬 => 3629

付録B
データ2の共起関係

이영 => 0001
영에 => 0002
프로 => 0003
로필 => 0004
대장 => 0005
장금 => 0006
출연 => 0007
연작 => 0008
소개 => 0009
수록 => 0010
사진 => 0011
자료 => 0012
료실 => 0013
운영 => 0014
영화 => 0015
기사 => 0016
사를 => 0017
보니 => 0018
니까 => 0019
예전 => 0020
…

분과 => 3801
일해 => 3802
별히 => 3803
있나 => 3804
상도 => 3805
전혀 => 3806
분에 => 3807
회관 => 3808
소회 => 3809
의실 => 3810
실에 => 3811
북한 => 3812
인권 => 3813
론회 => 3814
회가 => 3815
황우 => 3816
우여 => 3817
원과 => 3818
주최 => 3819
최로 => 3820
열렸 => 3821
가졌 => 3822

付録C
プログラムソース

```
1.  #!perl
2.
3.  $no = 0;
4.  $n = "001";
5.
6.  while (<>) {
7.          if(($no+2) % 3 == 0){
8.                  $_=~s/[₩w,₩n,₩',₩!,₩/,₩%,₩(,₩),₩,,₩@,₩^,₩",₩*,₩-,
    ₩_,₩?,₩",#₩+,₩=,₩<,₩>,₩`,₩~,₩[,₩],₩{,₩},₩;,₩:,₩&,₩|,₩#!]//g;
9.                  $word .= $_;
10.                 $_=~s/[₩.,₩s!]//g;
11.                 $list_sni{$n++} = $_;
12.                 }
13.         $no++;
14. }
15.
16. $word =~s/ /꿲/g;
17. $word =~s/₩./꿲/g;
18.
19. $len = length($word);
20. $n="0001";
21.
22. for ($i = 0; $i < $len; $i = $i+2){
23.         if(substr($word,$i+2,2) eq "꿲" || substr($word,$i,2) eq "꿲"){
24.                 break;  #文字列が空白か「. 」なら前後の文字列は取らない
25.         }
26.         else{  #2文字づつ取る
27.                 $bi = substr($word, $i, 4);
28.                 unless(exists($list_bi{$bi})){
29.                         $list_bi{$bi}=$n++;
30.                 }
31.         }
32. }
33.
34. #共起関係を求める
```

```
35. foreach (sort{$list_bi{$a} <=> $list_bi{$b}}(keys %list_bi)) {
36.         print "$_ => $list_bi{$_}¥n";
37. }
38.
39. #foreach $_ (sort keys %list_sni) {
40. #         print "$_ => $list_sni{$_}¥n";
41. #}
42.
43. while(($key1, $value1) = each % list_sni){
44.         while(($key2, $value2) = each % list_bi){
45.                 if($list_sni{$key1} =~/$key2/){
46.                         $list_i{$key1}->{$value2}++;
47.                         while($' =~/$key2/){
48.                                 $list_i{$key1}->{$value2}++;
49.                         }
50.                 }
51.         }
52. }
53.
54. #各スニペットの共起頻度を表示
55. #foreach $key1 (sort keys( %list_i)) {
56. #         print "snippet NO: ",$key1,"¥n";
57. #         foreach $key2 (sort keys( %{$list_i{$key1}} )) {
58. #                 print $key2,"¥t",$list_i{$key1}->{$key2},"¥n";
59. #         }
60. #         print "¥n";
61. #}
62.
63. #LibSVMのデータファイル用に出力
64. foreach $key1 (sort keys( %list_i)) {
65.         print "1 ";
66.         foreach $key2 (sort keys( %{$list_i{$key1}} )) {
67.                 print $key2,":",$list_i{$key1}->{$key2}," ";
68.         }
69.         print "¥n";
70. }
```

付録D
データ1のスニペット

http://gall.dcinside.com/list.php?id=pmyeongsu
공지, DC인사이드 거성 박명수 갤러리 통합 공지 Ver.03 (完) [8], 업러됴¹. 2009/07/25, 430. 공지, 개그맨 박명수에 관련된 사진과 내용을 올려주세요. ...

http://news.mk.co.kr/se/view.php?sc=30000023&cm=%EB%AC%B8%ED%99%94%C2%B7%EC%97%B0%EC%98%88%20%EC%A3%BC%EC%9A%94%EA%B8%B0%EC%82%AC&year=2009&no=477342&selFlag=sc&relatedcode=&wonNo=&sID=507
지난 8일 방송된 MBC 에브리원 '지금은 꽃미남시대'에서 박명수는 모발이식 의사들로부터 러브콜을 받았다. 이날 방송에서는 진행자 박명수를 비롯해 FT아일랜드의 ...

http://epg.epg.co.kr/star/profile/index.asp?actor_id=2235
,본명, 박명수, ,성별, 남자. ,생일, 1970년 8월 27일, ,국적, 한국. ,분야, 코미디언,가수, ,소속그룹. ,데뷔일, 1993, ,데뷔동기, MBC 개그콘테스트 ...

http://star.mt.co.kr/view/stview.php?type=1&no=2008030509015078494
2008년 3월 5일 ... 거성' 박명수의 결혼소식이 알려진 가운데 그의 마음을 사로잡은 여의사 한 모씨(30)에 대한 네티즌들의 관심이 집중되고 있다. ...

http://sports.hankooki.com/lpage/entv/200705/sp2007051810522158390.htm
2007년 5월 18일 ... 박신혜는 '웃기거나 슬프거나'를 주제로 진행된 이날 녹화에서 지난해 박명수와 함께 CF에 출연한 적이 있다. 처음에는 '호통개그'로 유명한 박명수 ...

http://kr.dir.yahoo.com/Entertainment/Humor__Jokes__and_Fun/Comedians/Park_Myung_Soo/
박명수 갤러리 [현재창]: 개그맨, 팬 커뮤니티, 디시인사이드 제공. http://gall.dcinside.com/list.php?id=pmyeongsu ? MBC FM4U 두시의 데이트 박명수 입니다@ ...

http://kafuri.tistory.com/338
2009년 8월 30일 ... 모든 맴버들의 성적이 비슷했으나 가장 마음을 짜안하게 만든 것은 ' 하찮은' 박명수의 성적표였습니다. 특히 그의 아내(한수진)가 학교로(제작진에게 ...

http://www.imbc.com/broad/radio/fm4u/date/index.html
방송 : 오후 2시~4시 진행 : 박명수. 연출 : 양시영 구성 : 전희주, 이영희, 장소영. [사연안고 뮤직콜] 깨알같은 사연 받아요 ? 멋진 라이브와 함께 선물 쏩니다! ...

http://focus.chosun.com/people/peopleView.jsp?id=1952
2009년 8월 5일 ... 경력 : 사무용소프트웨어연합(BSA) 청소년 저작권 홍보대사. MBC 두시의데이트 라디오 진행. MBC FM4U 박명수의 funfun라디오 진행 ...

http://ko.wikipedia.org/wiki/%EB%B0%95%EB%AA%85%EC%88%98_(%ED%94%84%EB%A1%9C%EA%B2%8C%EC%9D%B4%EB%A8%B8)
치열했던 경기는 박명수의 2:1 승리로 마무리되었으며, 이후 박명수는 대테란전에 가장 강력한 저그로 각광받은 한편 염보성은 번번이 개인리그 16강을 넘지 못하는 ...

http://pncp.textcube.com/52
2009년 6월 20일 ... 유재석이 요즘 박명수 구하기에 열중하고 있습니다. 국민MC가 2인자 박명수 구하기에 열중인 이유가 무엇일까 매우 궁금해 지네요. ...

http://tvexciting.com/723
무한도전을 통해 가장 큰 인지도를 얻은 사람이 있다면 바로 박명수가 아닐까 싶다. 예전에 이승철 흉내를 낼 때는 "우쒸~!" 외에는 인지도가 없었던 박명수였지만, ...

http://labstal.tistory.com/315
2009년 6월 21일 ... 아마 모르긴 몰라도, 100에 90 정도가 유재석, 박명수 콤비를 최우선으로 선택할 것이다. 토크쇼 놀러와를 통해 진행자와 패널로 첫 만남을 가진 이후 ...

...

付録E
データ2のスニペット

이영애 프로필, 대장금 등 출연작 소개, CF 수록, 사진 자료실 운영.

2009년 11월 18일 ... Daum 영화 > 이영애. 기사를 보니까 예전에 심은하씨랑 결혼 하기로 하고 ... 친자포기다 내가 알기론 불륜상태에서 여성이 남성에게 결혼에 조건으로 ...

이영애의 법률대리인 법무법인 동인 측은 25일 보도 자료를 내고 이영애가 24일 미국에서 재미교포 정 모씨와 가족들만 참석한 가운데 미국에서 결혼식을 했다며 ...

이영애 결혼, 이영애 남편은 정호영 사실일까?,파이어폭스 부가기능,윈도우7,무료아이콘,파이어폭스,트위터,구글,유틸리티,Window,검색엔진최적화,미드,lovedweb, ...

2009년 8월 27일 ... 이영애의 결혼을 놓고 화제가 끊이지 않고 있습니다. 이영애라는 스타의 위치를 생각하면 당연한 일입니다. 그렇게 비밀 유지를 위해 온 정성을 기울 ...

2009년 8월 25일 ... 이영애는 법률대리인 법무법인 동인을 통해 지난 24일 미국에서 미국 교포인 정모씨와 가족들만 참석한 가운데 결혼식을 했다고 25일 밝혔다. ...

톱스타 이영애의 깜짝 결혼 소식이 여러 가지 뒷 얘기를 낳고 있다. 최대의 관심은 그녀의 남편이 누구인가 하는 사실. 그러나 철저하게 베일 속에 가려져있다. ...

2009년 8월 26일 ... 허나 오랜 시간 '신비주의' 전략을 사용하며 자신의 위치를 지켜온 이영애의 행보로 보자면 이번 비밀 결혼도 그 테두리 안에서 해석할 수 있을 듯도 ...

한류스타 이영애가 지난 24일(한국 시간) 미국에서 극비리에 결혼했다는 소식이 알려진 뒤 미국의 어느 교포 사이트에 하와이 카할라 호텔에서 결혼식 당일 이영애를 ...

탤런트 이영애(38)가 24일(현지시간) 미국에서 사업가인 재미 교포 정 ... 이영애의 법률대리인 법무법인 동인 측은 25일 보도자료를 내고 이영애가 24일 미국에서 ...

이영애씨가 지난 8월 24일 미국에서 깜짝 결혼을 했습니다. 연예인들은 속성상 열애설이 돌고 결혼상대에 대한 이런 저런 뉴스가 나온 후 결혼날짜가 잡히고 동료 ...

2009년 2월 5일 ... 4일 오후 6시 현재 총 1만75명이 참가한 설문조사에서 이영애는 21.9%의 지지를 얻어 1위에 올랐다. 일본 네티즌은 "이영애는 가장 한국적인 미인 ...

2009년 2월 11일 ... 이영애와 장서희는 지난 7일 오후 서울 강남구 청담동의 모 음식점에서 오랜만에 만나 회포를 풀었다. 두 사람은 점심 식사를 함께 하고 격의 없는 ...

이영애의 법률자문을 맡은 법무법인 동인은 이영애가 일리노이 공대를 졸업하고 미국계 IT 업종에서 일하는 교포 정모 씨와 결혼식을 올렸다고 25일 말했다. ...

2009년 8월 26일 ... 팬들 충격 .. '극비결혼' 이영애 쉿 남편은 비밀이에요 정씨 궁금증 고조 ... 비밀결혼 후 , 인천공항을 통해 따로따로 귀국한 이영애와 정모씨. ...

이영애는 잠정초등, 정신여중, 잠실여고, 한양대 독어독문과를 졸업하고 배우로 활동하다 중앙대 대학원 연극영화과에 입학했다. 14살때 주니어 잡지 표지모델로 데뷔 ...

2009년 8월 25일 ... 톱스타 이영애와 24일 미국에서 결혼한 정모씨는 이영애가 데뷔할 때부터 ... 이날 이영애의 결혼을 공식 발표한 법무법인 동인은 정씨에 대해 미국 ...

스크린과 브라운관을 넘나들며 오랫동안 전국민적 사랑을 받아온 배우 이영애. 깨끗하고 아름다운 모습과 함께 당당함을 잃지 않는 이미지로 꾸준한 사랑을 받고 있는 ...

이영애 국회의원 이영애 (李玲愛) Young A. Lee (1948/09/21) 자유선진당 / 비례대표 초선 (18대) 서울대학교 법학과 졸업 당 최고위원 김성수, 박영미 박훈신 김동철, ...

2009년 8월 28일 ... 이처럼 떡밥난무의 시기에 던져진 최신-최대의 떡밥은 이영애의 결혼 발표임이 분명하다. 얼마전까지 'G드래곤-표절'이란 제목의 게시물들로 도배 ...

2009년 8월 25일 ... 이영애의 남편 정씨는 미국 교포로서 미국 일리노이 공대를 졸업하고 현재 ... 이영애 측은 남편에 대한 상세한 신상 및 사진 등을 사생활침해의 우려 ...

...