WWW検索システムにおける 分野別URLデータベースを用いた改良型ナイーブ・ベイズ選定手法

宮城　暖† 獅々堀正幹†† 小泉　大地† 北　研二†††

† 徳島大学大学院 工学研究科
†† 徳島大学 工学部
††† 徳島大学 高度情報化基盤センター
E-mail: †{miyagi,bori,koizumi,kita}@tokushima-u.ac.jp

あらまし WWW 情報検索システムは膨大な WWW 空間から情報を手軽に検索できるが，単純な検索質問に対する検索結果は多岐にわたっており，検索結果からさらに選定と検索を繰り返す手間が必要となる．この問題点に対して，ユーザが求める分野毎に検索結果を分類する URL 選定手法が存在する．URL 選定手法は分野を象徴する基底単語から分野毎の URL データベースを自動構築し，検索結果の URL を照合することで分類を行う．しかし，URL 選定手法ではデータベースに未登録の URL に対する分類精度が低下する問題がある．そこで本稿では，自動構築したデータベース内の URL にリンクする HTML ページのコンテンツを学習データの正事例とし，ナイーブ・ベイズを用いて分類を行う手法を提案する．また，ナイーブ・ベイズ選定手法は学習データ内の単語の出現確率を用いて分類を行うが，学習データ内に存在しない新出単語に対しては，ゼロ頻度値を適用する手法が一般的である．この点に関して，本研究のフィルタリング環境は WWW 検索エンジンを利用できるため，新出単語を WWW 検索エンジンを用いて再検索し，新出単語を登録済の単語集合に置き換えることで，より適切な新出単語の確率値を推定する改良を行う．

キーワード WWW 検索システム，URL データベース，コンテンツ・フィルタリング，ナイーブ・ベイズ

Modified Naive Bayes Classification Method using URL DataBase According to Field in WWW Search Engine

Dan MIYAGI†, Masami SHISHIBORI††, Daichi KOIZUMI†, and Kenji KITA†††

† Graduate School of Engineering, Tokushima University
†† Faculty of Engineering, Tokushima University
††† Center for Advanced Information Technology,Tokushima University
E-mail: †{miyagi,bori,koizumi,kita}@tokushima-u.ac.jp

Abstract WWW retrieval systems can retrieve the required infomation from the WWW space. The search result however, often includes not only true information but also a lot of false information. In order to solue this problem, Koizumi et al. proposed the URL selection method. This method can classify retrieval results of WWW retrieval systems into some fields according to the user's request by using the URL database. This URL database can be constructed automatically based on retrieval results of WWW retrieval systems for some basis keywords that associate each field. However, this method is not effective for the set of unregistered URLs into URL database. In this paper, we propose a new method using the Naive Bayes that learns contents on the HTML page linked with URL in the URL database as a positive case. The Naive Bayes classification method learns the appearance probability of the word in the training data. As for the word that does't appear in the training data, the zero frequency problem is happened. In addition, we have improved assuming more appropriate probabilities from the WWW retrieval result of words that don't exist in the training data, because this filtering environment can use WWW retrieval systems.

Key words WWW Search Engine，URL DataBase，Contents・Filtering，Naive Bayes Classification Method

1. はじめに

WWW 情報検索システムは，膨大な WWW 空間から情報を手軽に検索するツールとして，今日の情報社会において必要不可欠なものとなっている．しかし，単純な検索質問に対する検索結果は多岐にわたっており，検索結果からさらに選定と検索を繰り返す反復的な手間が必要となる．Google [1] [2] を代表とするロボット型 (ページ検索) サーチエンジンにおいてこの点は特に研著となるため，本研究ではロボット型サーチエンジンを WWW 情報検索システムとして扱う．ロボット型サーチエンジンは，WWW 上を自動的に巡回するプログラムによって，WWW 上に存在するほぼ全ての情報に対して索引付けを行うため，検索キーワードを含む情報を大量に検索できる．

この問題点に対して，ユーザの検索要求に適合する情報を効率良く収集する手法として，検索結果に含まれる検索要求に適合する情報から検索質問拡張を行う適合性フィードバック [3] [4]，WWW の個人適応化を行うためにユーザの特性からモデルを構築する手法 [5]，検索結果の内からユーザが求める分野の情報のみを分類する URL 選定手法 [6] が提案されている．URL 選定手法は分野を象徴する基底単語から分野毎の URL データベースを自動構築し，検索結果の URL を照合することで分類を行う．しかし，URL 選定手法ではデータベースに登録されていない URL に対する分類精度が低下する問題がある．

一方，文書内容を判定し，文書集合を正負といった 2 値化分類する手法としてナイーブ・ベイズ分類 [7] [8] が提案されており，実際に迷惑メイルを排除するシステム [9] に適用され，その有効性が確認されている．ナイーブ・ベイズの分類手法は，各分野に属する文書内の単語の出現確率に基づいているため，高い分類精度が得られるが，判別モデル作成のために多くの学習データが必要となるといった問題点がある．

そこで本稿では，自動構築したデータベース内の URL にリンクする HTML ページのコンテンツを正事例とし，学習過程を簡略化したナイーブ・ベイズ分類手法を提案する．また，ナイーブ・ベイズ選定手法は学習データ内の単語の出現確率を用いて分類を行うが，学習データ内に存在しない新出単語に対しては，ゼロ頻度値を適用する手法が一般的である．この点に関して，本研究のフィルタリング環境は WWW 検索エンジンを利用できるため，新出単語を WWW 検索エンジンを用いて再検索し，新出単語を登録済の単語集合に置き換えることで，より適切な新出単語の確率値を推定する改良手法を提案する．

2. URL 選定による分類手法

2.1 分野別 URL データベースの構築

例として「阪神」をキーワードとして検索した場合，一般的な WWW 検索システムでは，数多くの分野の意味をもつ「阪神」に関するページが検索されてしまう．図 1 はその例である．そのため，ユーザは検索結果を取捨選択しながら目的のページをみつけるといった労力におわれる．この問題点に対して，URL 選定手法は分野を象徴する基底単語から分野毎の URL データベースを自動構築し，検索結果にリンクされている URL

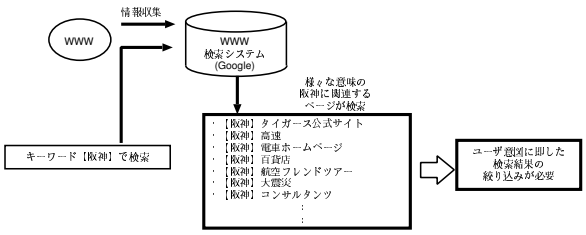

図 1　WWW 検索システムの現状

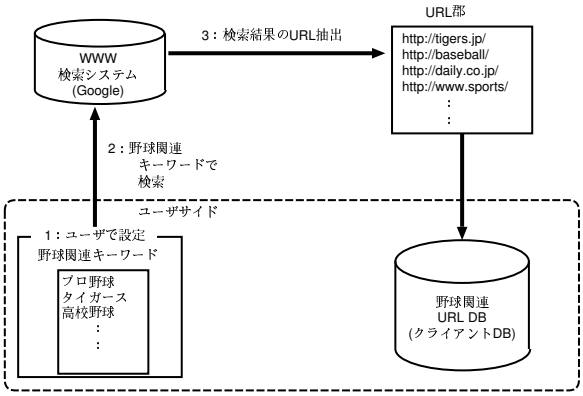

図 2　野球関連の URL データベースの構築例

を部分マッチングによって照合することで分類を行う．有効な URL データベースを構築できれば高精度な選定も可能である．通常のブラックリストやホワイトリスト形式の URL フィルタリングでは，URL データベースに登録された URL しか選定の対象にならないが，小泉ら [6] の手法では URL のパス毎に分野との関連度を付与した URL データベースに対して，URL のパス毎に部分マッチングするため，，部分的に共通する URL についても選定を行うことができ，データベースに登録された URL 数以上の分類精度を発揮することができる．

図 2 は，野球関連の URL データベースの構築の流れである．ユーザは求める分野に関連性の高いキーワードをいくつか設定する．次にそれらのキーワード，基底単語を基に WWW 検索システムの検索結果を収集し，リンクされている URL を抽出する．これらの URL の出現頻度を正規化し，頻度の高いものを URL データベースに登録する．

基底単語から得られた各 URL に対し，WWW 空間中の URL 出現頻度で正規化する．出現頻度の正規化で，基底キーワードと URL における関連性の強弱を判別することができる．これにより，関連性の低い Web サイトの検出を抑え，関連性の高いと思われる Web サイトを特定することができる．以下に正規化の手順を示す．

手順 1：部分 URL 毎の出現頻度の計算

URL データベース内に出現する部分 URL 毎の出現頻度を求める．部分 URL は，“/”を区切りとして分割したものである．例として，“ http://www.tokushima-u.ac.jp/G-life/main.htm ”の URL に対して部分 URL を求めると“www.tokushima-u.ac.jp”と“ www.tokushima-u.ac.jp/G-life ”の 2 つの部分 URL が作成される．これらの部分 URL の各パスの共通部分の頻度を出現頻度とする．

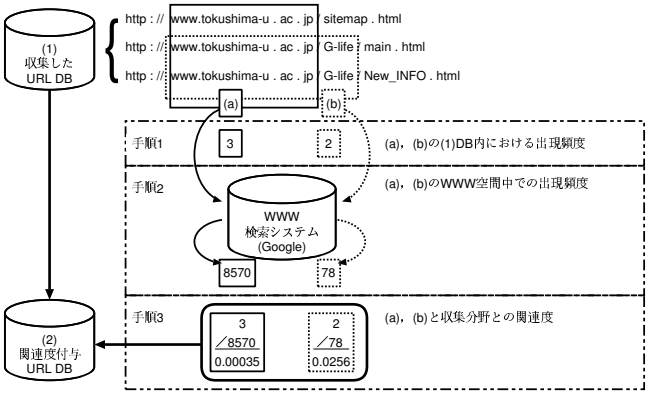

図 3　出現頻度の正規化の例

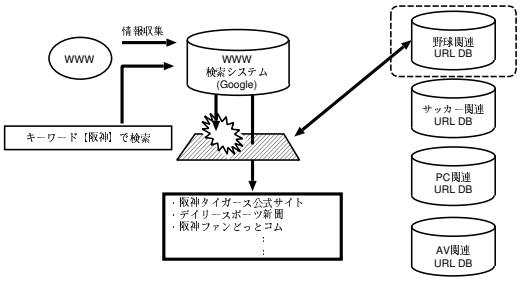

図 4　URL 選定手法

手順 2：部分 URL の大域的頻度の取得
各部分 URL を WWW 検索システムの URL 検索機能に入力し，検索結果内の「検索件数」を部分 URL が WWW 空間中に存在する大域的出現頻度とする．
手順 3：部分 URL の出現頻度の正規化
手順 1 の出現頻度を式 (1) により大域的出現頻度で正規化し，その値を関連度とする．

$$\text{URL の関連度} = \frac{\text{部分 URL のデータベース内での出現頻度}}{\text{部分 URL の WWW空間中での大域的出現頻度}} \quad (1)$$

図 3 に上記の手順に従い，部分 URL の出現頻度の正規化を行った例を示す．図 3 の基底キーワードから収集した (1) の URL データベースには 3 つの URL から作成される部分 URL が登録されている．部分 URL は (a)www.tokushima-u.ac.jp と (b)www.tokushima-u.ac.jp/G-life の 2 つであり，部分 URL(a) のデータベース内での出現頻度は 3，(b) は 2 である．つぎに各部分 URL を WWW 検索システムの URL 検索機能に入力して検索を行うと部分 URL(a) は 8570 件，(b) は 78 件の検索結果を得る．最後に，式 (1) により正規化した出現頻度を求める．部分 URL(a) は 0.00035，(b) は 0.0256 となる．(1) の URL データベース内の URL にそれらの関連度を付与することで，収集した URL についても収集した分野と関連度の高低を求めることができる (2) の URL データベースを構築する．

2.2　URL データベースを用いた分類

URL 選定手法は URL データベースに登録されている URL と WWW 検索結果にリンクされている URL のそれぞれの部分 URL を照合し，分野との関連度を求め，定めた閾値に基づき選定する手法である．図 4 はいくつかの分野の URL データベースを構築し，野球関連のみを分類する例である．関連度は部分 URL 毎に，WWW 空間全体内と URL データベース内の正規化した出現頻度の割合で定められている．しかし，データベースに未登録の URL に対して分類精度は低下する．

3.　ナイーブ・ベイズの分類方法

ナイーブ・ベイズ (N・B) とは，学習データを用いてベイズの定理に基づき何種類かのクラスへ文書を分類する手法の一つである．ナイーブ・ベイズの分類方法は，それぞれの文書 $\boldsymbol{x}$ が単語の集合 $< a_1, a_2, \cdots, a_n >$ で表され，学習データのクラス集合 $\boldsymbol{V}$ に全ての文書が分類される条件で，ナイーブ・ベイズの分類は学習を行い，モデルを構築する．学習データに含まれない新しい文書をクラスに分類するベイズの方法は，学習データのモデルを設定し，入力文書 $\boldsymbol{x_{in}}$ から単語の集合 $< a_{in1}, a_{in2}, \cdots, a_{inn} >$ を基に各クラスに属する確率 V を求め，最大の確率になる v_{MAP} を決定することである．すなわち，次の値を求めることになる．

$$v_{MAP} = \text{argmax}_{v_j \in V} P(v_j | a_{in1}, a_{in2}, \cdots, a_{inn}) \quad (2)$$

ベイズの定理を使うと，この等式は次のように書き換えられる．

$$\begin{aligned} v_{MAP} &= \text{argmax}_{v_j \in V} \frac{P(a_{in1}, a_{in2}, \cdots, a_{inn} | v_j) P(v_j)}{P(a_{in1}, a_{in2}, \cdots, a_{inn})} \\ &= \text{argmax}_{v_j \in V} P(a_{in1}, a_{in2}, \cdots, a_{inn} | v_j) P(v_j) \quad (3) \end{aligned}$$

今，学習データに基づいて式 (3) のうち，2 つの項を計算する．学習データの中で，単純に個々のクラスに属する文書を数えることによって，$P(v_j)$ を概算することができる．しかし，非常に多くの学習データの集合を持たなければ，それぞれの $P(a_{in1}, a_{in2}, \cdots, a_{inn} | v_j)$ の項を概算することは不可能である．ここで，文書内の各単語の出現確率が独立であると仮定すると，$P(a_1, a_2, \cdots, a_n | v_j) = \prod_i P(a_i | v_j)$ となり，これを式 (3) に代入するとベイズ分類は次のようになる．単語の出現確率が単純に独立していると仮定することから，このベイズ分類はナイーブ・ベイズと呼ばれる．

$$v_{N \cdot B} = \text{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \quad (4)$$

ナイーブ・ベイズ選定手法における確率計算を計算機で実装する際は確率の総積を求めるため，アンダーフローが生じる問題がある．そこで，アンダーフローに対する処置として全単語の確率値から対数をとって得られるスコアの総和を求めて分類計算を行う．

4.　URL データベースを用いた改良型ナイーブ・ベイズ分類手法

4.1　提案手法の概要

本手法では WWW 検索システムにおいて，URL データベース内の URL に対応するコンテンツをナイーブ・ベイズの学習データとして用いて，従来の検索結果を選定する手法を提案する．ナイーブ・ベイズは学習データが適切であれば高い分類精度を得ることができる．しかし，ユーザが求める多く分野の学

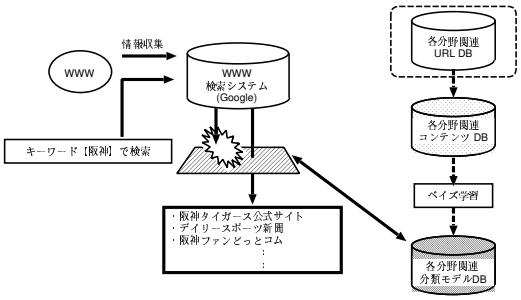

図 5　野球分野の選定例

習データを人手で収集することは時間と労力を要するが，本手法では 2.1 で自動構築した URL データベースにリンクされるコンテンツをダウンロードすることで構築したコンテンツデータベースを解析することで学習過程を軽減し，ナイーブ・ベイズの学習を行う．データベース内の各コンテンツは形態素解析し，効率的に動詞，名詞の単語を抽出する．学習過程では各クラスのコンテンツ数，それらの単語出現頻度に基づいたモデルが構築される．図 5 に例を示す．

URL データベースを用いて，WWW 検索結果をナイーブ・ベイズ選定手法で分類する手順を以下に示す．

準備：分類モデル構築

2.1 の手順にしたがって各分野の URL データベースを構築する．各分野の URL データベース内の関連度の高い URL から順に，コンテンツを WWW からダウンロードし，解析する．その結果の単語の出現頻度から各クラスのモデルを構築する．

手順 1：分類対象コンテンツの取得・解析

検索キーワードを入力し，WWW 検索システムの検索結果にリンクされる URL のコンテンツを 1 つずつダウンロードして単語抽出を行う．

手順 2：分類計算

コンテンツの解析結果をモデルに基づき，各クラス毎の分類確率を求め，最大となるクラスに分類する．

4.2　混合型モデルを用いた分類法

ナイーブ・ベイズはモデルを単純に全クラスで構築すると，学習データのクラスに属さない文書も各クラスに属する事後確率が計算され，学習データのいずれかの分野に分類してしまう特徴がある．そこで図 6 に示すように，各分野について，その分野とそれ以外全てを混合した分野として 2 つのクラスを持つモデルを構築する．この混合分野を用いることにより，ある分野かそれ以外の分野かといったの 2 値分類が実現できる．それを全てのモデルに基づいて文書を分類すると，特定の分野だけに分類される場合と，複数の分野に分類される場合，また，全ての分野以外に分類される場合が発生する．1 つの分野に分類された場合は，その分野に属すると定めることができる．複数の分野に分類された場合は，その文書がどのクラスにも属してしまう曖昧性の高い文書であると定め，学習データにない分野の文書として分類する．全ての分野以外に分類された場合も同様に曖昧性が高いと定め，学習データにない分野の文書として分類する．

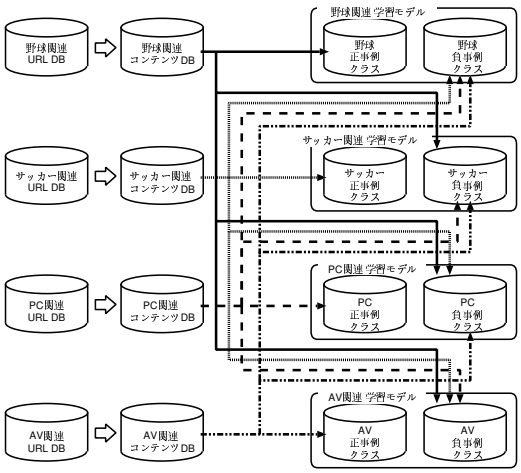

図 6　混合モデル例

4.3　WWW 検索による新出単語の確率推定

4.3.1　新出単語の問題点

ナイーブ・ベイズ選定手法は学習データ内の単語の出現確率を用いて分類を行うが，学習データ内に存在しない新出単語に対しては，ゼロ頻度値 [7] を適用する手法が一般的である．これは，ある分野で観測できなかった単語の出現確率がゼロになり，分類計算においてその単語を含むコンテンツが分類される事後確率もゼロになる問題に対し，その単語の出現確率をゼロ頻度値というゼロではない小さな確率値として適用することで，分類するコンテンツに含まれる新出単語の影響を小さくし，学習データ内で観測できる既存単語の確率値によって正しく分類する試みである．この点に対して，本研究のフィルタリング環境は WWW 検索を利用できるため，新出単語を入力とした WWW 検索結果から新出単語に関連する情報を収集し，それらから新出単語の確率値を推定することで，新出単語も分類計算において有効とする改良を行う．新出単語に対する WWW 検索結果には，学習データ内で観測される単語が含まれているため，それらを基に新出単語の確率を推定することにより，新出単語が学習データ内の各分野との関連性を反映することができる．WWW 検索結果から確率を推定する際は，その検索結果に対しナイーブ・ベイズを適用するが，その検索結果内に含まれる新出単語についてはゼロ頻度値を適用するものとする．

図 7 に示すように，新出単語の多くは構築した学習データ内の分野に属さない未知分野を象徴する単語か，もしくは学習データ内の特定分野における新しい流行語であると考えられる．意義が異なるこの 2 種類の新出単語の WWW 検索結果も，それぞれの分野に関連する単語を多く含むという推察から，未知分野の新出単語から推定される各分野に対する確率値はあまり高くないものになり，学習データ内の特定分野の流行語は，その分野のみに高い確率値を推定することが出来ると考えられる．これにより，従来の新出単語にゼロ頻度値を適用する分類計算より適切な分類結果が得られると考えた．

4.3.2　新出単語に対する改良手法

本研究の改良手法の流れを図 8 と共に説明する．まず，ページから抽出した単語集合を学習データ内に存在する既存単語集

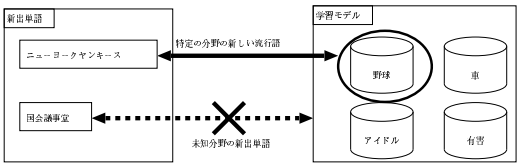

図 7　特徴的な新出単語の例

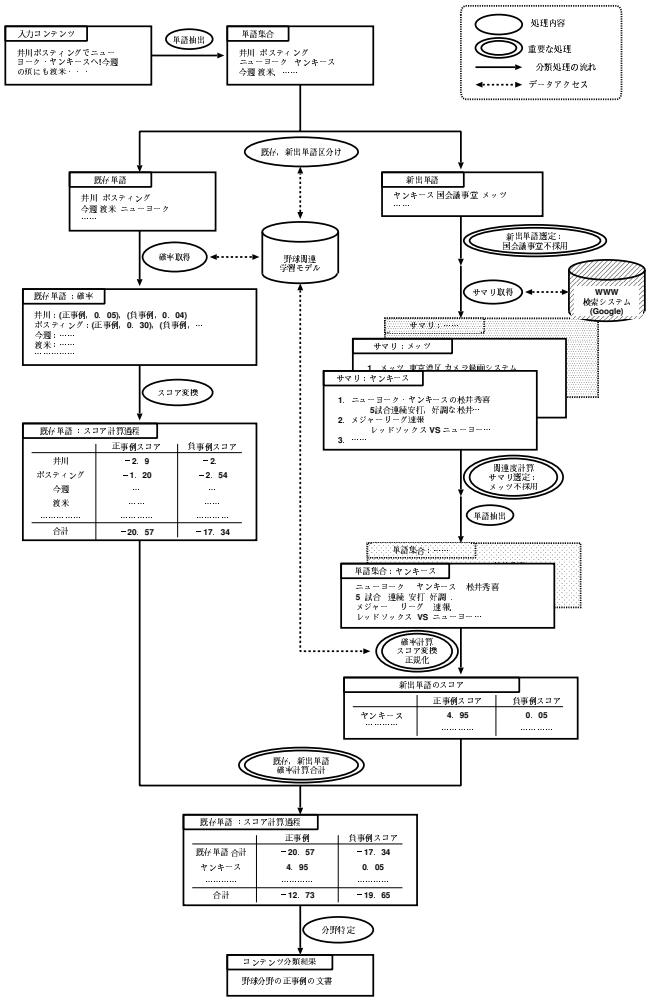

図 8　改良型ナイーブ・ベイズ選定手法の例

合と未登録の新出単語集合に分ける．既存単語については，従来のナイーブ・ベイズを適用して確率値を推定する．一方，新出単語については，新出単語を入力とした WWW 検索結果に含まれる単語のうち，学習データ内に存在する単語集合から得られる事後確率を新出単語の確率値とする．最終的に，既存単語集合について従来のナイーブ・ベイズを適用した確率値と新出単語集合の確率値を統合した値を用いて分類を行う．新出単語の WWW 検索結果から確率値を推定するためには，新出単語に関連する文書情報が必要である．本研究では新出単語の文書情報に WWW 検索結果のサマリを用いることで，各コンテンツを取得するより高速に文書情報を取得している．

本改良手法により高い分類精度を得るためには，新出単語集合内から該当分野の特徴的な新出単語のみを選択し，適切なサマリを取得する必要がある．そこで今回は，まず最初に新出単語集合の出現頻度を計算し，出現頻度が上位の新出単語を採用した．ただし，本手法では，単語抽出に形態素解析を適用しているため，解析失敗の文字列が新出単語として検出される可能性がある．これらの単語は意味が特定できず，分類に有効なサマリも取得することができないので対象外にする必要がある．そこで本手法では，解析失敗した文字列の多くが過分割して解析される傾向に着目し，ある特定の長さ以上の文字列のみを新出単語としている．

また，曖昧な意義を持つ新出単語から取得したサマリも同様に曖昧な内容を含み，分野を特定できない不適切な確率値を推定してしまう問題点がある．そこで，サマリと元のコンテンツとの関連度を式 (5) に基づいて計算し，関連度が閾値を越えるサマリからのみ確率を推定する条件でも実験を行った．

$$\text{関連度} = \frac{\text{サマリとコンテンツの共通単語数}}{\text{サマリの単語数} + \text{コンテンツの単語数}} \times 100 \quad (5)$$

新出単語の確率推定は以上の流れで行われるが，新出単語のサマリから推定した確率値を適用する際に正規化することで適切な分類を試みる．次項では，その点について詳しく説明する．

4.3.3　新出単語の推定確率に対する正規化

信頼性の高い既存単語の確率値による計算結果の優位性を保ちつつ適切に分類計算を行うために，新出単語から推定する確率値を正規化する必要がある．

まず，確率計算においてコンテンツの既存単語の確率値の対数によるスコアの総和と，新出単語に対するサマリ内の既存単語のスコアをそれぞれ求める．サマリの総スコアとコンテンツの総スコアは，それぞれの単語量に比例するので，短いコンテンツ内の 1 つの新出単語のスコアとして，そのサマリの総スコアを適用すると，コンテンツの数少ない信頼性の高い既存単語のスコアを全て無視することになってしまう．そこで，サマリの総スコアを既存単語の平均値を数倍した値で正規化することで，コンテンツの既存単語の優位性を保ちつつ，新出単語にある程度の重みを与えている．図 8 の例におけるスコアを正規化する計算過程の例を図 9 に示す．

出現確率が大きいほど，スコアも大きな値をとる．サマリのスコア合計値を確率値に逆変換し，正事例と負事例のどちらかに分類するそれぞれ 0 から 1 までの確率を得る．コンテンツの既知単語のスコア合計についても大きな値をとる事例へ分類する確率が高くなるので，サマリの確率値を全単語のスコア差の平均程度として図 9 では 5 を倍率として，正規化を行っている．スコアから確率値へ逆変換する際の計算は式 (6) によって行う．

$$\text{確率値} = \frac{\exp(\text{その事例のスコア})}{\exp(\text{正事例のスコア}) + \exp(\text{負事例のスコア})} \quad (6)$$

5.　評　　価

本手法の有効性を確認するため，解析した URL データベースの URL のコンテンツ情報を学習したナイーブ・ベイズ選定手法によって，WWW 検索システムの検索結果に対する選定実験を行った．以下に実験条件，評価基準，実験結果，考察を述べる．

5.1　実 験 条 件

既存の WWW 検索システムに Google Image Search [10] を

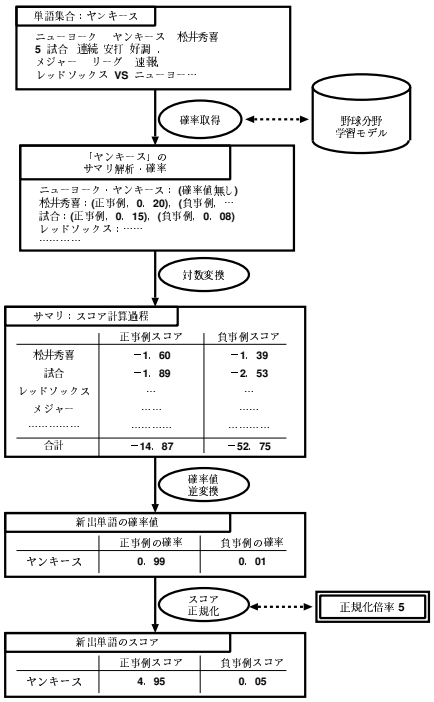

図 9　正規化を行うスコアの計算過程

用いて，「野球」，「アイドル」，「有害」，「車」の分野に対して URL DB と評価用データを作成した．まず，各分野 10 件の基底単語を WWW 検索システムに入力し，各分野の URL DB を構築した．それらを用いた URL 選定手法 (URL) を比較対象として，URL DB の上位 200 件ずつの URL のコンテンツを収集し，ナイーブ・ベイズの学習を改良手法と従来手法 (Simple NB) の両方で行った．ナイーブ・ベイズの改良手法については，混合型モデルを用いたもの (Mix NB) と，それに加えて新出単語の確率推定を行った手法で実験を行った．新出単語は 3 文字以上の長さを持つものに対して，サマリ 30 件を取得し，そこから推定を行ったもの (Mix NB +NW) それと元コンテンツとの関連度が 0.25 以上のものだけから推定を行った (Mix NB + NW(relate))．サマリの取得件数と関連度については，予備実験を行った上で最適な値をとった．2 文字以下の単語は，形態素解析を失敗した単語が多く含まれるために排除した．

次に，各分野の情報が検索される可能性がある「野球」分野に対し「井川」，「松坂」，「松井」，「阪神」等 5 件，「アイドル」分野に対し，「後藤」，「モデル」，「加藤」等 5 件，「車」分野に対し「マールボロ」，「センチュリー」，「ワゴン」，「三菱」等 5 件といった計 28 個の評価用キーワードで検索を行い，検索結果上位 40 件の URL 計 600 件を評価用データとした．

更に，評価用データ中の各分野の分類を人手で判別し，「野球」関連の検索結果 200 件のうち 92 件が「野球」分野，「アイドル」関連の検索結果 200 件のうち 30 件が「アイドル」分野，「車」関連の検索結果 200 件のうち 69 件が「車」分野の情報として得られた．

5.2 評価基準

選定精度の評価尺度には，再現率・適合率から計算する F 尺度 [11] を用いた．F 尺度は，適合率と再現率が共に高いときに 1 に近付く情報検索における指標の 1 つである．ここで，情報が分野に選定されることを TRUE と定義し，分野に選定されないことを FALSE と定義する．評価用データに対して，URL データベースの学習データを用いて選定を行い，式 (7)，(8) に示す各分野の情報の再現率 R_{true} と適合率 P_{true} を求めた．R_{true} は評価用データ中の情報が各分野に正しく選定できた割合を表し，P_{true} は選定した情報の中で本当にその分野に分類されるべき情報であった割合を表す．

$$R_{true} = \frac{\text{TRUE かつ正しく選定された情報数}}{\text{その分野に属する情報数}} \tag{7}$$

$$P_{true} = \frac{\text{TRUE かつ正しく選定された情報数}}{\text{TRUE だった情報数}} \tag{8}$$

また，選定された情報の再現率・適合率を求めると同時に，式 (9)，(10) に示す，不選定の情報の再現率 (R_{false})，(P_{false}) も併せて求めた．R_{false} は評価用データ中の分野に属さない情報に対して，分野に選定する割合を表し，P_{false} は選定されなかった情報の中で本当にその分野に適合していなかった情報の割合を示す．

$$R_{false} = \frac{\text{FALSE かつ正しく選定された情報数}}{\text{その分野に属さない情報数}} \tag{9}$$

$$P_{false} = \frac{\text{FALSE かつ正しく選定された情報数}}{\text{FALSE だった情報数}} \tag{10}$$

TRUE と FALSE ともに適合率と再現率から式 (11)，(12) に示す，F 尺度を求めた．

$$F_{true} = \frac{2}{\frac{1}{R_{true}} + \frac{1}{P_{true}}} \tag{11}$$

$$F_{false} = \frac{2}{\frac{1}{R_{false}} + \frac{1}{P_{false}}} \tag{12}$$

比較データとして同様の評価データに対して URL 選定手法と単純型と混合型モデルのナイーブ・ベイズと，そして新出単語の確率推定を行う改良型ナイーブ・ベイズを用いて再現率・適合率を求めた．

また，従来の URL 選定手法では，URL データベース構築時に準備したいずれかの基底キーワードが存在するページしか選定することができなかった．この問題点に対して，本手法ではコンテンツを解析しているため，基底キーワードを含んでいないページでも適切に分類することができる．この有効性を評価するため，各キーワードに対する試験データ 40 件について，検索要求分野の基底キーワードを含むかどうかを調べ，それらが適切に分類されているかを評価する．

5.3 実験結果

TRUE 選定と FALSE 選定の結果について各検索キーワード毎の F 尺度を，「野球」分野を図 10，11，「アイドル」分野を図 12，13，「車」分野を図 14，15 に示す．また，TRUE 選定と FALSE 選定の各分野毎の F 尺度の平均値を表 1，2 に示す．表 3 は各分野の試験データ 200 件に対して，URL 選定手法と提案手法の Mix NB+NW(relate) がそれぞれ TRUE 選定において正解したコンテンツのうち，基底単語を含んでいないコンテンツの試験データ 200 件に対する割合を示す．

表 1，2 から，URL 選定手法に比べて提案手法の F 尺度が

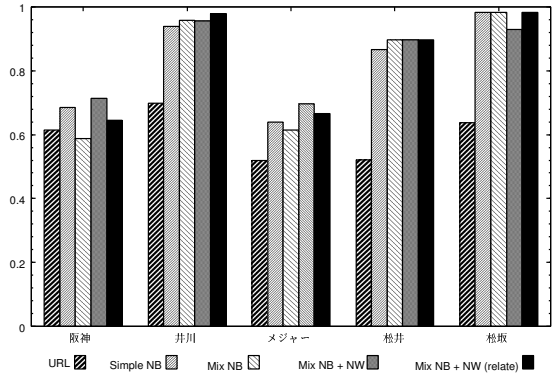

図 10 「野球」分野の TRUE 選定の F 尺度

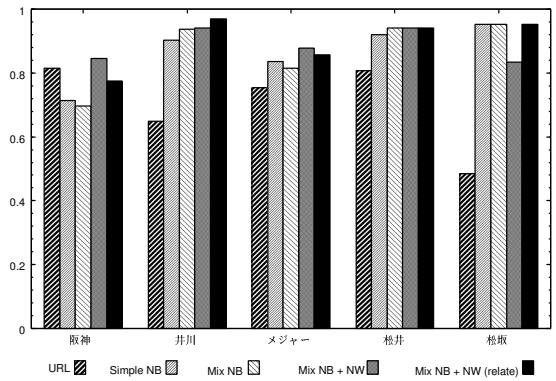

図 11 「野球」分野の FALSE 選定の F 尺度

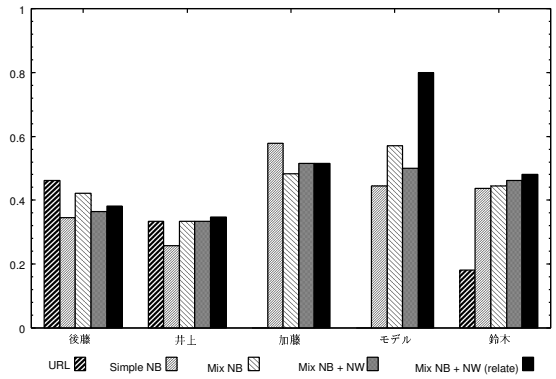

図 12 「アイドル」分野の TRUE 選定の F 尺度

表 1 各手法における TRUE 選定の平均 F 尺度

手法/分野	野球	アイドル	車
URL	0.598	0.195	0.407
Simple NB	0.823	0.413	0.576
Mix NB	0.808	0.453	0.591
Mix NB + NW	0.808	0.435	0.599
Mix NB + NW(relate)	0.834	0.505	0.605

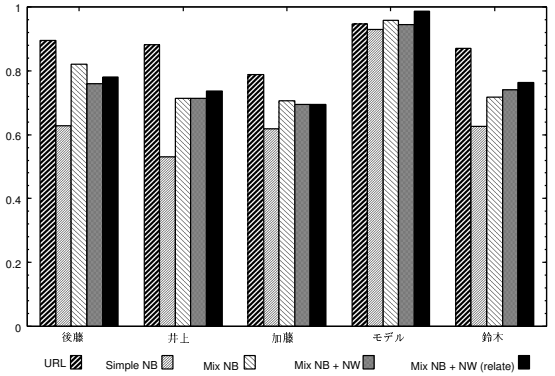

図 13 「アイドル」分野の FALSE 選定の F 尺度

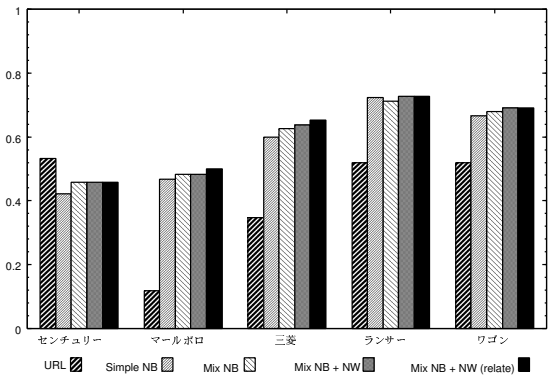

図 14 「車」分野の TRUE 選定の F 尺度

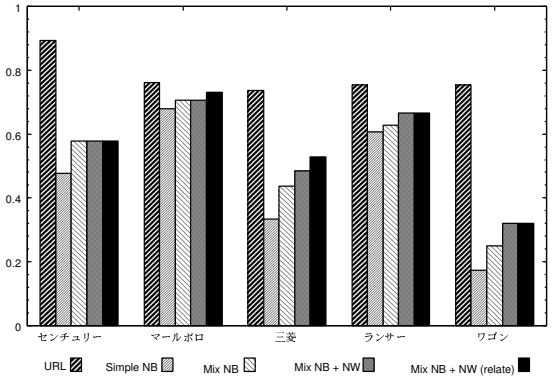

図 15 「車」分野の FALSE 選定の F 尺度

表 2 各手法における FALSE 選定の平均 F 尺度

手法/分野	野球	アイドル	車
URL	0.702	0.877	0.780
Simple NB	0.865	0.666	0.454
Mix + NB	0.868	0.783	0.520
Mix NB + NW	0.888	0.770	0.551
Mix NB + NW(relate)	0.899	0.792	0.565

大きく上回っていることから，URL データベースの URL コンテンツに基づいて学習を行うナイーブ・ベイズ分類が有効であるといえる．ナイーブ・ベイズの分類精度は学習データの精度に依存するため，自動構築した各分野の URL データベースの情報が適切であったことを示している．「アイドル」分野の「加藤」，「モデル」の実験結果について，URL 選定手法で問題であった未登録の URL に対する分類精度の低下を解決できているといえる．

「アイドル」，「車」分野の FALSE 選定では提案手法が URL 選定手法を下回る結果になっている．これは URL 選定手法において，TRUE 選定が極端に低いことと関連して考えて，TRUE 選定された件数が少な過ぎることと，収集した実験データ内に

表 3　各選定手法による TRUE 選定の基底キーワードを含まない正解コンテンツの割合

	野球	アイドル	車
URL	16.5	0.5	9.5
Mix NB + NW(relate)	30.5	9.5	29.0

各分野に関連する文書が少なかったことが考えられる．URL 選定手法は未登録である URL について一切対処できないため，ある一定の関連度を越え TRUE 選定された文書が少な過ぎることと，FALSE 選定すると正解となる文書が実験データ内に多く含まれていることから，URL 選定手法において FALSE 選定結果の F 尺度が高くなった．提案手法では，コンテンツの内容をもってして判別するため，TRUE，FALSE 選定それぞれでの誤検出を生じる可能性があるため，結果 URL 選定手法を下回る結果になった．しかし，総合的に TRUE 選定と FALSE 選定の精度評価を行うためにそれぞれの平均を取ると，URL 選定手法に比べ，改良型ナイーブベイズ選定手法は平均 F 値が野球分野で 0.214，アイドル分野で 0.113 増加し，車分野では 0.09 減少する結果となった．このことから全体として分類精度は向上したと考えられる．

また，本手法で提案した混合型のモデル構築手法は，単純型のモデル構築手法に比べて F 尺度が向上している．このことから，いくつかの曖昧性の高い文書が混合型モデルによって，学習データに属さない分野として分類できたことがわかる．

そして，新出単語の確率推定を適用したことにより，未知の分野の特徴的な単語から得られた確率値が各分野の混合モデルにおいて負事例へ分類するように働いたため，未知の分野の文書を検出できる．また，同様に学習データで学習できなかった目的分野の単語についても適切な確率推定により混合型モデルにおける正事例へ分類するように働き，目的分野の文書を検出できている．

新出単語の確率推定に関連度を加えたことで，適切に分類するように働くサマリとそうでないサマリの検出ができたため，他の手法に比べ最も高い F 尺度を示した．また，新出単語に対する取得するサマリの件数や関連度を変化させて実験を行うと分類精度が変化した．このことは現在の確率推定手法で適切に確率を推定できたサマリだけを，コンテンツとの関連性の高さから検出することで提案手法に適用していることを意味する．より高精度な分類を実現するためには，関連度を適用せずに新出単語から得られるサマリを全て利用する必要がある．より適切な確率を推定するために，サマリ内に含まれる全単語を確率推定に用いずに，サマリにおいて高頻度で出現する単語のみで確率を推定する手法，またはサマリに含まれている学習データ内において高頻度で出現する単語のみで確率を推定する手法，これら二つの確率推定を組み合わせる手法などが考えられる．

表 3 からわかるように，提案手法は URL 選定手法の問題点である基底キーワードを含まないコンテンツの URL に対して再現率が低下する点を，基底単語と高頻度で共起する各分野の単語によって適切に分類することで改善することができた．しかし，同時に「アイドル」,「車」分野において，FALSE 選定とすべき検索要求に不適合な基底単語を含まないコンテンツに対して TRUE 選定とした不正解例が増加することとなった．これは提案手法が，URL 選定手法の問題点を解消するために生じる問題である．そこで，この問題点については URL 選定手法と提案手法間で，TRUE 選定として検索要求に適合する情報のみを出力するように適合性を高める際は，提案手法に重きをとり，URL 選定手法に重きをとることで再現率を高めるようにユーザの検索要求に応じて，調整をすることが可能になっている．

6. ま と め

本稿では，数個の基底となる各分野を象徴するキーワードを準備するだけで自動構築する URL データベースの情報を用いて，既存の WWW 検索システムの検索結果に対しユーザが求める各分野の情報を分類するのに有効なナイーブ・ベイズ分類の学習の簡略化を行う手法を提案した．また，ナイーブ・ベイズが学習データに存在しない曖昧な分野の文書を検出できない点について混合モデルを適用，新出単語について対応できない点について，WWW 検索を利用した確率推定を行う改良を適用した．評価実験では，URL 選定手法に比べ，分類精度を向上することができた．今後は，WWW 検索を利用した確率推定を用いてより高精度の分類を実現するために，適切な確率推定手法の確立に取り組む予定である．

謝辞　本研究の一部は，科学研究費補助金基盤研究 (B)(17300036)，科学研究費補助金基盤研究 (C)(17500644) を受けて行われた．

文　献

[1] Google. http://www.google.co.jp/.
[2] TaraCalishain(著)，RaelDornfest(著)，山名早人 (訳)，田中裕子 (訳)：GOOGLE HACKS，オライリー・ジャパン，2003.
[3] Kenji Yanai：Image Collector Ⅱ：A System for Gathering More Than One Thousand Images from the Web for One Keyword, *In Proc. of IEEE International Conference on Multimedia and Expo*，volume I，pp.785〜788，2003.
[4] 獅々堀正幹，小泉大地，柘植覚，北研二：画像知識データベースを用いた WWW 画像検索システムの開発，電子情報通信学会論文誌，VOL. J87-D-I NO.2，pp154〜163，2004.
[5] 三浦信幸，高橋克巳，島健一：個人適応型 WWW におけるユーザモデル構築法，情報処理学会論文誌，VOL.39 NO.5,pp1523〜1535，1998.
[6] 小泉 大地，獅々堀 正幹，中川 嘉之，柘植 覚，北 研二：WWW 画像検索システムにおける有害画像フィルタリング手法，情報処理学会 論文誌 Vol. 47 No. SIG8 P147〜156，2006.
[7] 北 研二 著：言語と計算− 4 確率的言語モデル，東京大学出版会，1999.
[8] Nigam K，McCallum A，Thrun S，Mitchell T：Learning to classify text from labeled and unlabeld documents，Proceed of the 15 National Conference on Artifical Intelligence，1998.
[9] Paul Graham：A Plan for Spam. http://www.paulgraham.com/spam.html.
[10] Google Image Search. http://images.google.co.jp/.
[11] 北 研二，津田 和彦，獅々堀 正幹 著：情報検索アルゴリズム，共立出版，2002.