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ナイーブ・ベイズ選定手法

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

宮城 暖 † 小泉 大地 † 獅々堀 正幹 ‡ 柘植 覚 ‡ 北 研二 §
Dan Miyagi Daichi Koizumi Masami Shishibori Satoru Tsuge Kenji Kita

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．この問題点に対して，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 [1] 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 [2][3] 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．

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

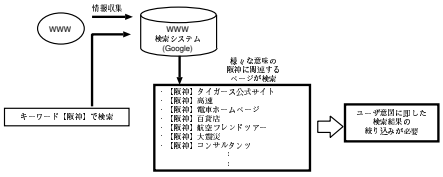

図 1: WWW 検索システムの現状

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 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．URL の部分マッチングにより分野毎の URL を完全に網羅しなくても，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．

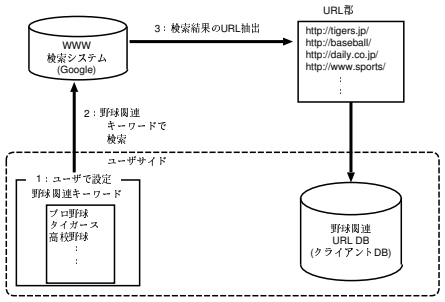

図 2: 野球関連の URL データベースの構築例

2.2 URL データベースを用いた分類

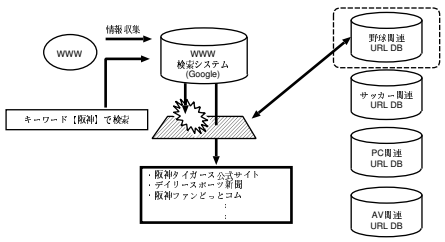

図 3: URL 選定手法

†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

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 3 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．これらのことから，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．

$$v_{MAP} = \mathrm{argmax}_{v_j \in V} P(v_j | a_{in1}, a_{in2}, \cdots, a_{inn}) \qquad (1)$$

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．

$$\begin{aligned} v_{MAP} &= \mathrm{argmax}_{v_j \in V} \frac{P(a_{in1}, a_{in2}, \cdots, a_{inn} | v_j) P(v_j)}{P(a_{in1}, a_{in2}, \cdots, a_{inn})} \\ &= \mathrm{argmax}_{v_j \in V} P(a_{in1}, a_{in2}, \cdots, a_{inn} | v_j) P(v_j) \end{aligned} \qquad (2)$$

今，学習データに基づいて式 (2) 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 (2) 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．

$$v_{N \cdot B} = \mathrm{argmax}_{v_j \in V} P(v_j) \prod_i P(a_i | v_j) \qquad (3)$$

4. URL データベースを用いたナイーブ · ベイズ分類手法

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習データを人手で収集することは時間と労力を要する．そこで，本手法では 2.1 で自動構築した URL データベースのコンテンツ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．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 4 に例を示す．

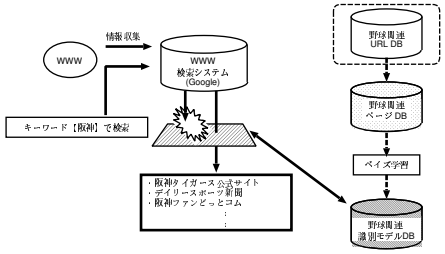

図 4: 野球分野の選定例

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 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 データベースの上位 200 件の URL のコンテンツを 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 検索キーワードを入力し，WWW 検索システム Google Search の検索結果にリンクされる URL のコンテンツ 1 つずつダウンロード，解析する．

手順 2 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．

また，本稿では各 URL データベースの学習データを混合するモデルの構築手法を提案する．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 5 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．

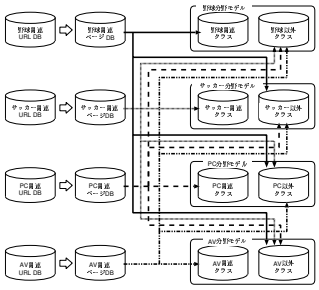

図 5: 混合モデル例

5. 実験・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[4]を用いて，「野球」，「アイドル」，「有害」，「車」の分野のURL DBと評価用データを作成した．まず，5件の基底単語を定め，各分野のURL DBを構築した．URL DBの上位200件ずつのURLを基に学習を行いモデルを構築した．次に各分野毎に実験を行うために，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」，「鈴木」，「松井」，「楽天」等9件，「アイドル」分野に対し，「鈴木」，「モデル」，「握手会」等10件，「車」分野に対し「マルボーロ」，「センチュリー」，「ワゴン」，「三菱」等9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1120件を評価用データとした．更に，評価様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360件のうち210件が「野球」分野，「アイドル」関連の検索結果400件のうち142件が「アイドル」分野，「車」関連の検索結果360件のうち162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・適合率を用いた．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FAULT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(4)，(5)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．

$$R_{true} = \frac{\text{TRUE かつ正しく選定された情報数}}{\text{その分野に属する情報数}} \quad (4)$$

$$P_{true} = \frac{\text{TRUE かつ正しく選定された情報数}}{\text{TRUE だった情報数}} \quad (5)$$

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(6)，(7)に示す，不選定の情報の再現率(R_{fault})，(P_{fault})も併せて求めた．R_{fault}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_{fault}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．

$$R_{fault} = \frac{\text{FAULT かつ正しく選定された情報数}}{\text{その分野に属さない情報数}} \quad (6)$$

$$P_{fault} = \frac{\text{FAULT かつ正しく選定された情報数}}{\text{FAULT だった情報数}} \quad (7)$$

これらの再現率･適合率を求め，それらをプロットし，再現率・適合率曲線[5]を求めた．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・ベイズを用いて再現率・適合率を求めた．

5.3 結果・評価

「野球」，「アイドル」，「車」の分類結果の再現率・適合率曲線をそれぞれ図6，7，8に示す．また，各分野について，各手法におけるTRUE選定とFAULT選定の平均適合率を表1，表2に示す．

URL選定手法に比べ，提案手法の再現率・適合曲線を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

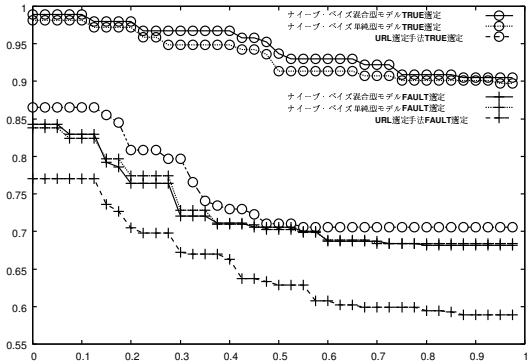

図 6: 「野球」分野の再現率・適合率

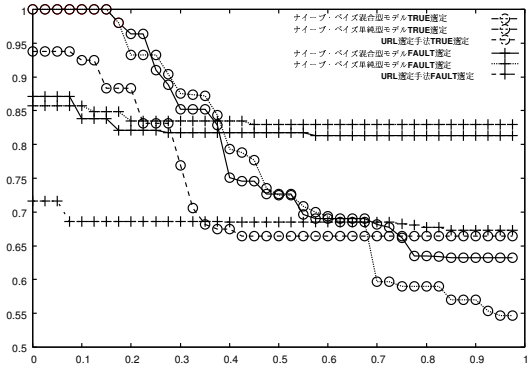

図 7: 「アイドル」分野の再現率・適合率

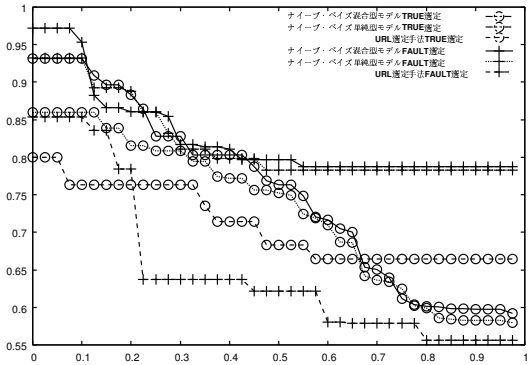

図 8: 「車」分野の再現率・適合率

表 1: 各手法における TRUE 選定の平均適合率

	N・B 混合型モデル	N・B 単純型モデル	URL 選定手法
野球	0.95	0.93	0.75
アイドル	0.78	0.77	0.74
車	0.76	0.74	0.71

表 2: 各手法における FAULT 選定の平均適合率

	N・B 混合型モデル	N・B 単純型モデル	URL 選定手法
野球	0.73	0.73	0.65
アイドル	0.82	0.84	0.69
車	0.83	0.82	0.65

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．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．

また，本手法で提案した混合型のモデル構築手法は，若干ながら単純なモデルより適合率が向上した．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．適合率に大きな変化がみられなかったことから，単純型と混合型の学習データに同じものを用いたため，曖昧性の低い文書は混合型で 1 つのみ分類された分野と単純型で分類された分野が同じものになったと考えられる．

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．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，より多くの分野の URL データベースを構築することで，より高精度な分類が行えると考えられる．混合型モデルについても，多くの分野の学習データを用いることで，曖昧性の高い文書をより高精度で分類できると考えられる．

謝辞：本研究の一部は，財団法人 電気通信普及財団，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．

参考文献

[1] 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，情報科学論文フォーラム講演論文集 2005 年 9 月 P45〜46

[2] 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)

[3] 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)

[4] ：Google Image Search. http://images.google.co.jp/

[5] 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)