自 然 言 語 処 理 154－2
（2003．3．6）

Naïve Bayes を用いた効率的な Web 検索システムの構築

西田　絵美子† アントニ・ローレンス‡ ラシキア・ジョージ‡

WWW には膨大な量の情報が溢れている．今や情報自体は容易に入手できる時代となったが，単に情報を得るのではなくユーザの興味に即した情報を得ることが重要である．一般的な情報収集手段として検索エンジンがあるが，望む情報を効率的に得ることは難しい．本論文では学習機能を導入し，ユーザ個々の検索意思を考慮した新たな Web 検索システムを提案する．この手法では，まず既存の複数の検索エンジンより検索結果を取得し，ユーザが必要，不要な情報かを選択する．そのデータを Naïve Bayes の学習知識として格納し，以降の検索結果を適切に識別する．本システムを評価するために，一つの語で複数の意味をもつ多義語をキーワードとした検索を行い，既存の検索エンジンと比較して有用性を確認した．

An Efficient Web Search System using Naïve Bayes

Emiko Nishida† Laurence Anthony‡ George Lashkia‡

The WWW contains a huge quantity of information. However, although information can now be easily obtained, it is important that this matches the interest of the user. The most general means for obtaining information is using a traditional search engine. However, it is difficult to acquire the desired information efficiently. In this paper, we propose a novel web search engine that uses a learning function to enable results to match the user's intention. Using several existing search engines, multiple web pages are first searched, and then the user classifies these in terms of whether or not they are desired results. This data is stored as knowledge for a Naïve Bayes Classifier that subsequently classifies future search results accordingly. To evaluate the system, we performed various searches using English homonyms, and as a result confirmed the validity of the system.

1. はじめに

近年，インターネットの普及により WWW で入手できる情報量は飛躍的に増大し，情報提供の高速化からリアルタイムの情報収集が可能となった．一方，収集可能な情報量が膨大になったため，情報を活用しきれないユーザが増えており，目的とする情報を効率よく得るための手段が必要とされている．代表的な情報収集手段として Google や AltaVista などの検索エンジンを使用するものがあるが，一般ユーザに妥当な検索キーワードを求めることは難しい．目的とする情報を含む検索結果を得た場合でさえ，多量の望まない情報に混在している場合，結果を全て確認することは非効率的であり，ユーザへの負担が大きい．また，汎用検索エンジンの対象は不特定多数であり，ユーザ個々の興味を反映した結果は期待できない．本研究では，システムに学習機能を導入して検索効率を上げる手法を提案する．この手法では，まずユーザが入力した検索キーワードをもとに既存の複数の検索エンジンより数件の検索結果を取得する．取得した検索結果をユーザは必要，不要な情報か選択する．選択された Web ページにはユーザの検索意図を示す情報が存在しているはずである．そのため，本システムではそのデータを Naïve Bayes Classifier (NB)の学習知識として格納し，以降の Web ページの識別を行う．知識はユーザの意思により形成されるため，システムは個人を対象としたものとなり，ユーザが望む情報を効率よく得る検索システムを構築することが可能である．

以下では，まず 2. で提案する Web 検索システムの構成を述べ，3. で Web ページ識別過程について詳しく説明する．そして 4. でその評価実験をし，最後に 5. で考察を行う．

† 岡山理科大学大学院工学研究科情報工学専攻
Department of Information and Computer Engineering, Graduate School of Okayama University of Science

‡ 岡山理科大学工学部情報工学科
Department of Information and Computer Engineering, Okayama University of Science

2. システム構成

本研究では NB を用いた効率的な Web 検索システムの中核部分を PKF (Personal Knowledge Filter)と表現する．PKF は複数の検索エンジンから収集した Web ページをユーザの興味に即して分類する部分である．PKF は重要語集合 KI (Knowledge Information)と Web ページ分類器である NB から構成される．検索システムの構成を図 1 に示す．以下でシステムの概要を説明する．

[User]：ユーザは自分の興味を反映する検索キーワードを入力する．また PKF を通過した検索結果は like (dislike)のボタン付きでブラウザに表示される(図 2)ため，必要(不要)とする情報のボタンをチェックする．

[Personal Knowledge Filter]：ユーザからの Web ページ分類データを活用し K I を形成する．NB によって検索エンジンが新たに収集した Web ページを分類する．

[Search Engine]：ユーザに与えられたキーワードをもとに WWW から Web ページを取得する．単一の検索エンジンでは，カバーする情報に限界があるため，複数の検索エンジンを利用する[1]．

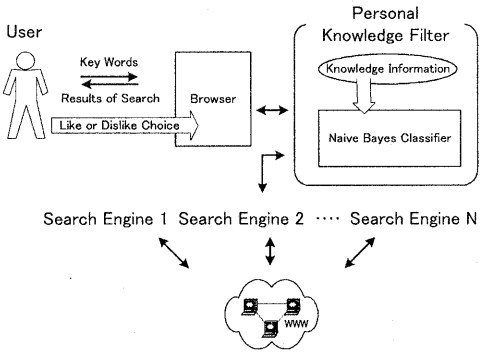


図 1．検索システムの構成図

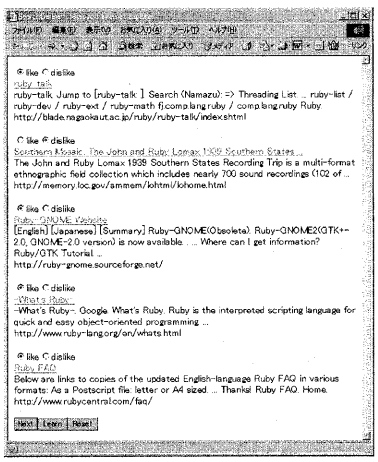


図 2．検索結果

3. Personal Knowledge Filter (PKF)

収集された Web ページが PKF を経ることで，ユーザに不要な情報は排除される．ユーザはブラウザに表示される検索結果を like，dislike クラスに分別し評価を繰り返すだけで，検索エンジンによる検索結果を全て確認することなく，効率的に情報収集を行える．

3.1 Knowledge Information (KI)

ユーザにより分類された Web ページは最終的に KI という形で，知識として格納される．KI は NB によって Web ページを分類する際に利用する重要語集合である．この集合は Web ページから取得した単語集合から情報量を使用して抽出され，過学習，計算時間の問題を避ける．また，本研究では，語と語の関連性を考慮し，任意の単語の組み合わせである**チャンク**を重要語集合に含む．KI はユーザが Web ページを分類する毎に更新されるため，再学習により分類性能の向上をはかることが可能となる．

3.1.1 チャンク

KI を形成する重要語集合は 3 種類のチャンク chunk 1，chunk 2，chunk 3 の組み合わせ (chunks 1 ~3)を要素にもつ．表 1 にチャンクの構造を示す．

元の文		
Ruby is the object-oriented language.		
チャンク		
chunk 1 (1 word)	chunk 2 (2 words)	chunk 3 (3 words)
ruby is the object oriented language	ruby is is the the object object oriented oriented language	ruby is the is the object the object oriented object oriented language

表 1．チャンクの構造

3.1.2 情報量

重要語の抽出には情報量を使うことが有効であると確認されている[2][3]．情報量ではクラスの特徴的な単語が大きな値を得る．情報量を利用することで，the，is，at など出現頻度は高いが Web ページの特性を表さないノイズとなる単語を除去することができる．しかし，ここで重要語とする情報量のしきい値が問題となる．本研究ではしきい値を求めるために，情報量の平均と偏差

を使用する.

3.1.3 重要語抽出アルゴリズム

1. Web ページに含まれる単語を取得する
2. 各単語の情報量 IG を求める
3. IG の平均（m）を計算する
4. m 以上の IG をもつ単語を重要語とする
5. 重要語が規定数に満たない場合，IG の偏差（σ）を計算する
6. $m-\sigma$ 以上 m 未満の IG をもつ単語を利用して重要語の数を増やす

3.2 Naïve Bayes Classifier (NB)

本研究では複数の検索エンジンから収集した Web ページを分類する際に NB [3][4][5]を用いる．以下で NB 分類法について説明する．

NB 法に基づく Web ページ w_i の分類クラス決定関数 υ_{NB} は，次式により示される．

$$\upsilon_{NB} = \arg\max_{\upsilon_j \in V} P(\upsilon_j)\prod_{k=1}^{l} P(a_k \mid \upsilon_j)$$

ここで，分類すべき w_i は単語 $a_1, a_2, \ldots, a_l$ をもつ．また υ_j は分類クラスを示し，$V = \{\upsilon_1, \upsilon_2, \ldots, \upsilon_n\}$（$n$ は分類クラス数）である．本研究では Web ページを like, dislike の 2 クラスに分類するため，$V = \{like, dislike\}$ とする．$P(\upsilon_j)$ は分類クラスの確率であり，今回は 0.5 とした．NB では単語集合 KI を用いて Web ページ分類を行う．KI は個人の関心に合わせて形成されるため，ユーザの興味を反映した分類を行うことが可能となる．

4. 評価実験

提案した手法の有効性を確認するため，本システムを実現して評価を行った．このようなシステムでは評価が主観的になるため，今回は検索キーワードとして一つの語で複数の意味をもつ多義語を用い[5]，目的の意味に関連したページと，他の意味に関連するページとを比較した．

4.1 重要語の推移

表 2 は検索キーワード mouse（目的の意味：遺伝学の実験動物）での学習回数による重要語の推移を示している．学習 1 回目では大部分のページに現れる this，it，when などの語が重要語として認識されている．学習を重ねることでノイズとなる単語が排除され，目的の意味に関連した Web ページの特徴を表現する単語が重要語となることがわかる．さらに chunk 2，chunk 3 が重要語に含まれていることから，語と語の関連情報の重要性が明らかとなった．

	学習回数		
	1 回	5 回	10 回
1	you	genome	genome
2	genome	human	mouse genome
3	genes	research	gene
4	sequence	human genome	sequence
5	mouse genome	genes	genes
6	whole	gene	genetic
7	this	mouse genome	human
8	some	project	research
9	it	data	human genome
10	when	genetic	the mouse genome

表２．学習回数と重要語
（検索キーワード mouse）

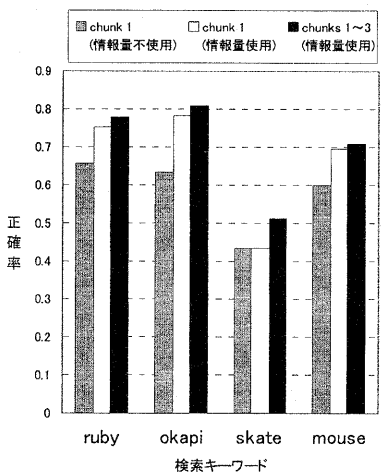


図３．KI の構成要素と正確率

4.2 正確率による比較

本研究では正確率を用いて比較実験を行う．本システムと汎用検索エンジンの正確率を以下のように定義する．

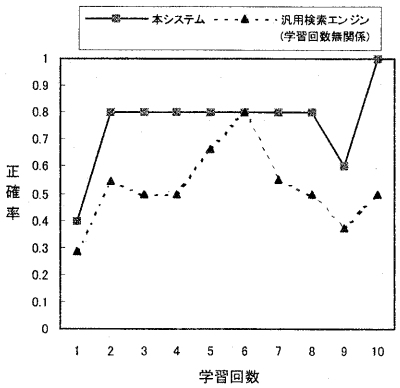
n ：本システムの検索結果ページ数
m ：本システムの検索結果ページ数（n）のうち目的の意味に関連するページ数
n' ：汎用検索エンジンの検索結果ページ数
m' ：汎用検索エンジンの検索結果ページ数（n'）のうち目的の意味に関連するページ数

本システムの正確率 $= m/n$

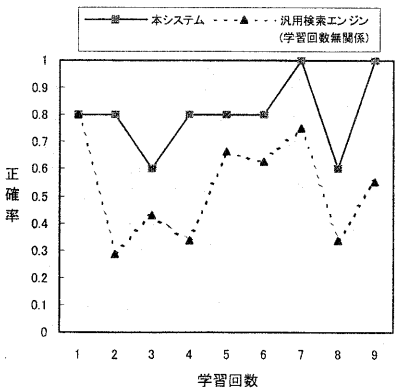
汎用検索エンジンの正確率 $= m'/n'$

4.2.1 KI の構成要素

図 3 では本システムの正確率を重要語の違いごとに示す．重要語として chunk 1（情報量不使用），chunk 1（情報量使用），及び KI の形成要素である



(a) ruby



(b) okapi

図４．学習回数と正確率

chunks 1~3 (情報量使用)を用いる．chunk 1 (情報量不使用)は単語の出現頻度のみを考慮したものである．図から chunk 1 (情報量不使用)の正確率が最も低いことがわかる．情報量を使用することによって，重要語集合のノイズを除去でき正確率が上がる．さらに，chunk 1 より語と語の関連性をもつ chunks 1~3 の正確率が高いこともわかる．これは chunks 1~3 が KI の形成要素として妥当であることを示している．

4.2.2 本システムの有用性

本システムと汎用検索エンジンの正確率を比較する．図４(a)は検索キーワード ruby の学習回数による正確率の推移を示す．図から学習を重ねることで本システムの正確率が向上することがわかる．図４(b)では検索キーワード okapi の場合を示す．KI を形成した Web ページと全く異なったページが初めて分類されるとき，KI に十分な知識が存在しない場合がある．その際，分類を誤り本システムの正確率が下がる(学習３，８回目)ことを確認した．しかし，この問題は再学習により解決される．表３は実験結果の一部であり，検索キーワード ruby，lighthouse，mouse，keyboard，skate，okapi の正確率を示す．本システムと汎用検索エンジンの正確率を比較した結果，本システムの検索効率が向上していることがわかる．

キーワード (目的の意味)	正確率	
	汎用検索エンジン	本システム
ruby (言語)	0.53	0.78
lighthouse (灯台)	0.50	0.53
mouse (遺伝学の実験動物)	0.35	0.71
keyboard (音楽)	0.29	0.56
skate (インラインスケート)	0.34	0.51
okapi (動物)	0.44	0.81

表３．正確率

5. 考察

本論文では Naïve Bayes を用いた学習機能を実装するシステムの構築について述べた．本システムの特徴はユーザが検索結果を必要，不要の２クラスに分類し，そのデータをもとに KI を形成する点にある．KI はユーザの検索意思を表す重要語集合であるため，検索システムは不特定多数ではなく個人を対象としたものとなる．また KI は再学習により洗練されるため，より効率的な情報収集を行うことが可能となった．今後の課題として，trip, journey などの同義語の問題を解決し，知識の共有を行いユーザの検索作業の更なる軽減を行うなどが挙げられる．また，実際に被験者を用いた実験を行い，本手法の有効性を評価する．

参考文献

[1] S. Lawrence, and C. L. Giles, “Accessibility of information on the web,” Nature, 400, 107-109, 1999.

[2] Y. Yang & J. O. Pedersen, “A comparative study on feature selection in text categorization,” ICML’97, 412-420.

[3] T. M. Mitchell, Machine Learning, Boston, MA: McGraw-Hill, 1997.

[4] D. D. Lewis, “Naive (Bayes) at forty: The independence assumption in information retrieval,” ECML, 4-15, 1998.

[5] C. D. Manning & H. Schutze, Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press, 1999.