

# 学習機能をもつ検索システムの構築

† 西田 絵美子   ‡ アンтони ローレンス   † ラシキア ジョージ  
† 岡山理科大学大学院工学研究科   ‡ 岡山理科大学

## 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。また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。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。その情報を Naïve Bayes Classifier (NB)の学習データとして Web ページの識別を行い、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。

## 2. システム構成

本研究の学習機能をもつ検索システムの中核を 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(図2)されるため、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 KI 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

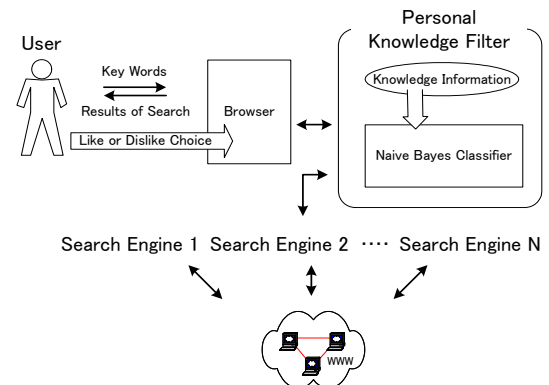


図1. 検索システムの構成図

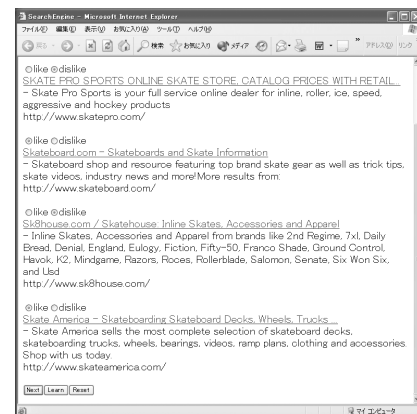


図2. 検索結果

もとに WWW から Web ページを取得する。

## 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 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

Title: A Web Search System with Learning Ability.

† Emiko Nishida, Graduate School of Okayama University of Science

‡ Laurence Anthony, Okayama University of Science  
George Lashkia, Okayama University of Science

可能となる。

- 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[1][2]。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。

- 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  $m$  と偏差  $\sigma$  を計算する
4. 平均  $m$  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、偏差  $\sigma$  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 [2][3][4]を用いる。

- NB 分類法

NB法に基づく Web ページ  $w_i$  の分類クラスを決定する決定関数  $v_{NB}$  は、次式により示される。

$$v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{k=1}^l P(a_k | v_j)$$

ここで、分類すべき  $w_i$  は単語  $a_1, a_2, \dots, a_l$  をもつ。また  $v_j$  は分類クラスを示し、 $V = \{v_1, v_2, \dots, v_n\}$  ( $n$  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、 $V = \{\text{like}, \text{dislike}\}$  とする。

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

| キーワード (目的の意味)   | 正確率   |          |
|-----------------|-------|----------|
|                 | 本システム | 汎用検索エンジン |
| ruby (言語)       | 0.80  | 0.63     |
| Halifax (カナダ)   | 0.80  | 0.75     |
| skate (スケートボード) | 0.56  | 0.29     |
| lighthouse (灯台) | 0.65  | 0.52     |

表 1. 正確率

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[4]、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。本研究では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。

A : 目的の意味に関連するページ数

B : 本システムの検索結果ページ数

C : 汎用検索エンジンの検索結果ページ数

本システムの正確率 =  $A/B$

汎用検索エンジンの正確率 =  $A/C$

表 1 は実験結果の一部であり、検索キーワード ruby, Halifax, skate, lighthouse の正確率を示している。本システムの正確率について説明する。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 に正しく分類するための十分な知識が存在しない場合がある。その際に分類を誤り正確率が下がることを確認したが、この問題は再学習により解決される。結果的に本システムの検索効率は向上したことがわかる。

### 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る。今後の課題として、trip, journey などの同義語の問題を解決する、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上げられる。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] Y. Yang & J. O. Pedersen, "A comparative study on feature selection in text categorization," ICML'97, 412-420.
- [2] T. M. Mitchell, Machine Learning, Boston, MA: McGraw-Hill, 1997.
- [3] David D. Lewis, "Naive (Bayes) at forty: The independence assumption in information retrieval," ECML, 4-15, 1998.
- [4] C. D. Manning & H. Schutze, Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press, 1999.