社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－MPS－62 (23)
2006－BIO－7 (23)
2006／12／22

# パーソナライズを考慮したWeb検索フィルタリングアルゴリズム

堀田知宏　丸山崇　北栄輔

名古屋大学大学院 情報科学研究科

膨大な情報源であるWeb上において，検索対象分野に関する知識が乏しい場合，ユーザの知識から導かれる検索語だけでは，検索要求を具体的に表現できず，目的の情報に到達することが難しい．更に，現在の検索エンジンは，同一の検索語では，誰が検索を行っても同一の結果しか得られない．そこで本研究では，各ユーザの嗜好に適合するようにURLをソーティングするアルゴリズムを提案する．さらに，認知度という指標により，各ユーザの検索領域に関する知識レベルを表現し，検索結果の妥当性を検証する．その結果，本アルゴリズムは，ユーザの知識不足を補ったURLの提示を行えたとともに，他のユーザの検索結果を利用し，各ユーザの嗜好情報を反映した結果を提示した．

# Web Search Filtering Algorithm by Considering Personalization

TOMOHIRO HOTTA, TAKASHI MARUYAMA and EISUKE KITA

Graduate School of Information Science, Nagoya University

If a Web user does not have enough knowledge, he cannot express search words only from his knowledge and it's difficult to get information of purpose. When we use a present search engine, we get the same result by using the same retrieval word.Then, we propose the Web Search Filtering Algorithm that sorts URL to suit the user's preference. In addition, We express a degree of the knowledge about search field by the index of acknowledgment level and verify the validity of the result. As a result, this algorithm was able to present URL that supplemented user's limited knowledge. And then, We present the result that suited each user's preference information by using other users' retrieval results.

## 1　はじめに

ブロードバンドや常時接続の普及が進み，WWW(World Wide Web) で利用可能な情報量は膨大なものとなっている．しかし，情報量があまりに多すぎるために，利用者が求めている情報の元に辿り着くことが困難になりつつある．それに伴い，情報の探索には様々なツールが使用されるようになってきた．

インターネット利用者が，求める情報を得るために頻繁に使用するサービスが検索エンジンである．しかし，現在の検索エンジンでは，同じ検索語で情報検索を行った場合，誰が検索を行っても同一の結果しか得られない．例えば，同じ検索語で検索を行って得られたURLの順序は人によって異なるべきであるが，実際には，検索結果として表示されるURLの順序に変化がないため，個人の嗜好に対応したURLが必ずしも上位に表示されるとは限らない．さらに，木谷ら [1] は，検索対象分野に関する知識が乏しい場合，検索者の知識から導かれる検索語では検索要求を具体的に表現できず，目的の情報に到達しにくいと報告しており，検索領域の知識が乏しいユーザにとっては，どのURLが自身の欲する情報なのかを適切に判断することは困難である．

本研究では，個人の嗜好を検索結果に反映させるとともに，ユーザの検索語履歴を利用し，各URLにスコアリングをすることで，個々の嗜好に適合するようURLをソーティングするアルゴリズムを提案する．また，スコアリングの際には，ベイジアンネットワークを利用する．グラフとしては，全ユーザ共有用と，各ユーザ用の2つのベイジアンネットワークを作成し，利用者全体，個人，両者の嗜好を反映させている．さらに本研究では，各検索領域に関して，ユーザの知識が豊富か否かを表現するため，各検索語に対して認知度という指標を用いて，各ユーザが有識であるかを表現し，ユーザ間で検索結果の比較を行う．

本論文は以下のような構成になっている．2章では，従来システムとの比較と本アルゴリズムの概要，3章では，本研究のWeb検索フィルタリングアルゴリズムについて詳細を述べる．4章では，実験方法，およびその結果を示し，5章で本研究についてまとめを行う．

## 2 提案アルゴリズムの概要

### 2.1 従来のシステムの問題点

従来から用いられるユーザの検索履歴を情報選択に利用する技術としては，協調フィルタリング [2] があげられる．しかし，時間経過を考えた情報推薦や，グループを更に細分化した個人単位での情報推薦への対応は困難である．また，情報検索のサービスとして検索エンジンがあげられるが，検索領域の分野に関して知識の乏しいユーザは，検索語を適切に表現できないという問題点がある．

### 2.2 提案システムの概要

本システムでは，各ユーザの嗜好情報，検索語および検索結果の履歴を利用し，従来の検索エンジンで得られた各 URL にスコアリングをすることで，ユーザの嗜好に適合するよう，URL をソーティングする．

嗜好情報としては，静的嗜好情報，動的嗜好情報という 2 種類の情報を使用する．静的嗜好情報は，ユーザがあらかじめ興味を抱いている分野を指す．この情報は，ユーザから直接入力される情報であるため，長期的に扱いやすく，情報として確実性が高い．動的嗜好情報は，各時点でのユーザの状況によって変動する嗜好情報である．本研究では，履歴から抽出した複数の検索語から動的嗜好を見出しており，静的嗜好と組み合わせることで，従来のシステムにおける，時間経過によるユーザの嗜好変化への対応という問題点を解決できる．

また，ユーザの特性を見出すために，各検索語がどの程度汎用的なものかを表す指標として，認知度を算出する．検索語間の関連性と，検索回数を基に認知度を計算することで，各分野に対してユーザが有識であるか定量的に表現し，結果比較に利用する．

スコアリングの手法としては，URL が検索される際に使用されたキーワード間に関連を持たせ，各キーワードの重要度を計算し，ベイジアンネットワーク [2] でキーワードの得点を算出後，各 URL に総合スコアをつける．さらに本研究では，全ユーザの情報を利用した共有ベイジアンネットワーク，各々のユーザの情報を利用したユーザ別ベイジアンネットワークの 2 種類を使用する．共有ベイジアンネットワークは，ユーザ全体の検索の傾向をソーティング結果に影響させるために用いる．ユーザ別ベイジアンネットワークには，ユーザ自身のみの情報が使用されるため，ユーザの現在の嗜好を反映したキーワードの得点付けが行える．

## 3 提案する Web 検索フィルタリングアルゴリズム

本研究で提案するアルゴリズムは，以下のような手順である．下記のステップでは，ユーザが検索エンジンに入力した情報を検索語，静的嗜好情報および検索語がデータベースに保存されたものをキーワードと表現する．検索語が 2 語以上入力された場合は，それらをまとめて 1 つのキーワードとして保存する．

また図 1 は，本アルゴリズムのフローチャートであり，点線で囲まれた部分にて，キーワードへのスコアリングを行っている．各ステップの詳細については後述する．

(1) ユーザが静的嗜好情報を入力

(2) ユーザが検索エンジンに検索語を入力し，URL 群を取得

(3) 検索結果として得られた URL を，ユーザ ID，静的嗜好情報，検索語とともにデータベースに保存

(4) 同じ URL を検索する際に使用されたキーワードの間をエッジで結ぶ

(5) 各キーワードの重要度を計算

(6) 重要度，エッジ関係から，共有，ユーザ別のベイジアンネットワークを形成

(7) 全ベイジアンネットワークより各キーワードの得点を計算

(8) (2) で得た URL 群それぞれに総合スコアを付ける

(9) URL 群を降順にソーティング

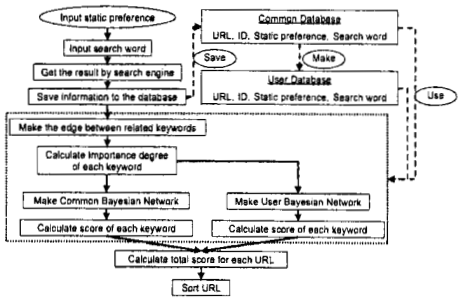


Fig. 1: Web Search Filtering Algorithm

### 3.1 エッジ関係の生成方法

ステップ (1)〜(3) において各情報を保存後，ステップ (4) で関連するキーワード同士を結びつける．例えば，検索語 $A$，検索語 $B$ によって同一の URL が検索された場合，$A$ と $B$ は関連性が高いとみなし，両ノード間にエッジを引く．酒井ら [3] の研究より，対象 URL の分野に関して知識の乏しいユーザの検索結果の中には，その分野の有識者が検索結果として得た URL が含まれるといえるためである．

エッジ関係を構築後，各キーワードの認知度 $CD(K)$ を以下の式を用いて計算する．$C(K)$ はキーワード $K$ の検索回数，$RelatedKWs$ は $K$ と関連するキーワード群，$KWNum$ はキーワードの数を表している．本

研究では検索回数として，2006 年 9 月時点での各検索語のオーバーチュアを使用している．各キーワードの認知度を計算後，ユーザが使用したキーワードの認知度の平均値 $AveCD(K)$ を求め，その逆数 $AL(K)$ を認知度レベルとする．この値が高いほど，そのユーザは検索領域に関して有識であると言える．

$$CD(K) = \frac{C(K)}{\sum_{K_i \in RelatedKW_S} C(K_i) + C(K)} \quad (1)$$

$$AveCD(K) = \frac{\sum_{K_i \in KW_S} CD(K_i)}{KWNum} \quad (2)$$

$$AL(U) = \frac{1}{AveCD(K)} \quad (3)$$

### 3.2 キーワードの重要度計算手法

ステップ (5) における各キーワードの重要度は，式 (4) を用いて計算される．この式は，$tf \cdot idf$[4] を基にした計算式で，$TF(K)$ はキーワード K の出現頻度を，$IDF(K)$ はキーワード $K$ の特定性を表す．また，$KCount$ はキーワード $K$ の検索回数，$TotalKeywordCount$ は全キーワードの検索回数の合計値，$URLCountK$ はキーワード $K$ が含まれている URL の個数，$TotalURLCount$ は全 URL の個数である．

$$w(K) = TF(K) \times IDF(K) \quad (4)$$

$$TF(K) = \frac{KCount}{TotalCount} \quad (5)$$

$$IDF(K) = \log \frac{TotalURLCount}{URLCountK} \quad (6)$$

以上の計算を，静的嗜好情報，検索語それぞれについて行う．そして，ユーザ別データベースについても行い，各キーワードについての重要度を算出する．

### 3.3 各 URL の総合スコア計算手法

ステップ (6) で共有，ユーザ別ベイジアンネットワークを形成後，ステップ (7) でそれらを用いて各キーワードの得点を計算する．

$SPscore(U)$ は，ユーザの静的嗜好情報が URL に付随する静的嗜好情報に一致した場合の得点，$SWscore(U)$ は，URL に付随する検索語群の得点の平均値で，式 (7) で計算される．$TotalSWscoreinURL$ は，注目 URL に付随するキーワードの得点の合計値，$SWNum$ は付随するキーワードの個数を表す．

$$SWscore(U) = \frac{TotalSWscoreinURL}{SWNum} \quad (7)$$

これらの計算も，共有，ユーザ別，それぞれにおいて行うため，結果として，2 つの静的嗜好情報の得点，$SPscore_C(U)$，$SPscore_I(U)$ と，2 つの検索語の得点，$SWscore_C(U)$，$SWscore_I(U)$ が得られることになる．なお，添え字の $C$，$I$ はそれぞれ，共有ベイジアンネットワーク用，ユーザ別ベイジアンネットワーク用であることを指す．各キーワードの得点を取得後，ステップ (8) にて式 (8)〜(11) で URL の総合得点を計算する．$Cscore(U)$ は，共有ベイジアンネットワークから算出されたスコア，$Uscore(U)$ は，ユーザ別ベイジアンネットワークから算出されたスコア，$UserTotalURLCount$ は，ユーザ個人が検索した URL の総数を表しており，$URLscore(U)$ は，対象 URL の総合得点である．ここで式 (10) においては，ユーザ特有の嗜好をより強く反映させるため，$Uscore$ に重み $\beta$ を乗じている．

$$Cscore(U) = SPscore_C(U) + SWscore_C(U) \quad (8)$$

$$Uscore(U) = SPscore_I(U) + SWscore_I(U) \quad (9)$$

$$URLscore(U) = Cscore(U) + \beta(Uscore(U)) \quad (10)$$

$$\beta = \frac{TotalURLCount}{UserTotalURLCount} \quad (11)$$

最後に，算出された総合得点に応じて，ステップ (9) にて URL をソーティングする．

## 4 実験

### 4.1 実験環境

プログラミングに関する情報検索を例に取り，5 名のユーザ，A から E に，それぞれ表 1 のように特徴を設定して実験を行った．静的嗜好情報の括弧内の数字は，そのユーザが検索を行った際に使用した静的嗜好情報の比率である．AL は，ユーザのコンピュータ分野に関しての認知度レベルであり，高いほど有識であることを示している．

Table. 1: User Data

| User | Static Preference | AL |
|---|---|---|
| A | Computer(100%) | 90.79 |
| B | Computer(81%), Sport(19%) | 29.29 |
| C | Computer(65%), Physics(35%) | 19.50 |
| D | Computer(36%), English(64%) | 14.09 |
| E | Sport(57%), Physics(27%), English(16%) | 0.0 |

また，本アルゴリズムを利用する前に，検索エンジン Google[5] を使用し，コンピュータ，スポーツ，物理，英語の 4 分野に関する検索語を用いて，その検索結果をあらかじめ登録してある．ここで，コンピュータ分野はさらに，GUI プログラミング，エディタ，ホーム

ページ，Linux 日本語化の 4 分野に細分化して登録を行った．コンピュータ分野に関して認知度レベルの高い UserA，UserB は，"Linux GUI プログラミング"，専門性が高く，認知度が低い検索語を使用するが，知識の乏しいユーザは，"java"など，認知度が比較的高めの検索語を使用している．

### 4.2 実験結果と考察

表 2 は，検索語"プログラミング"での実験結果で，それぞれのユーザに示された上位 10 個の URL を示している．表中の，URL-$n$($n$ は正の整数) は，Google で検索した場合，上位から $n$ 番目に表示された URL であることを示す．各ユーザの静的嗜好情報は，各々のユーザが最も使用しているものを設定した．

また表 3 は，表の左端のユーザ，即ち対象ユーザに対する注目ユーザの再現率，適合率を表したものである．各要素の左側が再現率，右側が適合率である．再現率 $R$ は式 (12)，適合率 $P$ は式 (13) によって求められる．$N$ は推薦された URL 数，$T$ はテスト URL 数，$M$ は $N$ と $T$ のうち，一致した URL の数を表す．本研究では，$N$ を対象ユーザに提示された検索結果のうち上位 10 個，$T$ を注目ユーザに提示された検索結果のうち上位 20 個の URL として計算した．

$$R = \frac{M}{N} \tag{12}$$

$$P = \frac{M}{T} \tag{13}$$

表 3 では，本アルゴリズムの検証として，知識の乏しいユーザの検索結果が有識者のそれにどれほど近づいているかを調べる必要があるため，認知度レベルの高いユーザに対する，認知度の低いユーザの再現率，適合率のみを示した．

Table. 2: Search Result

| | UserA | UserB | UserC | UserD | UserE |
|---|---|---|---|---|---|
| 1 | URL-04 | URL-97 | URL-02 | URL-26 | URL-07 |
| 2 | URL-02 | URL-02 | URL-70 | URL-80 | URL-02 |
| 3 | URL-70 | URL-70 | URL-46 | URL-07 | URL-70 |
| 4 | URL-87 | URL-13 | URL-13 | URL-04 | URL-97 |
| 5 | URL-49 | URL-18 | URL-18 | URL-25 | URL-87 |
| 6 | URL-13 | URL-87 | URL-97 | URL-65 | URL-13 |
| 7 | URL-18 | URL-99 | URL-88 | URL-74 | URL-18 |
| 8 | URL-97 | URL-46 | URL-87 | URL-31 | URL-46 |
| 9 | URL-46 | URL-45 | URL-04 | URL-59 | URL-41 |
| 10 | URL-76 | URL-38 | URL-86 | URL-01 | URL-03 |

Table. 3: Recall Rate / Precision Rate

| | UserB | UserC | UserD | UserE |
|---|---|---|---|---|
| UserA | 1.0/0.5 | 0.8/0.4 | 0.4/0.2 | 0.7/0.35 |
| UserB | - | 0.7/0.35 | 0.2/0.1 | 0.8/0.4 |
| UserC | - | - | 0.3/0.15 | 0.7/0.35 |
| UserD | - | - | - | 0.1/0.05 |

まず表 2 から，URL-97 や URL-26 が最も上位に提示されるユーザが存在するなど，Google での検索で得られた URL 順位と比べ，順位が大幅に引き上げられている URL が存在することがわかる．さらに，表 3 より，認知度レベルが最も高い UserA に対しての各ユーザの再現率を見ると，コンピュータ分野に関して知識が乏しく，静的嗜好に"Computer"を持たないユーザとして設定した UserE に対して，再現率 0.7 と比較的高い値を得られた．これは，共有ベイジアンネットワークによるスコアリングが影響し，UserA の動的嗜好が UserE にも反映されたためであり，その結果，有識者と類似した URL を上位で得られている．

また，UserD に関しては，再現率と適合率が低い結果となった．これは，UserD が使用した検索語として，コンピュータ分野の中でもプログラミング以外の分野の検索語を多く設定したため，UserD 自身の嗜好の方が強く反映されたためだと言える．さらに，再現率が高いユーザ間でも，ユーザ毎に URL 順位が異なっている．即ち，ユーザそれぞれの嗜好が，ユーザ毎の結果に強い影響を与えていることがわかる．

## 5 おわりに

本研究では，静的嗜好情報，動的嗜好情報を用いて，ユーザが求めている情報が得られるよう，URL をソーティングするアルゴリズムを提案した．その手法は，同一の URL が検索された際に使用されたキーワード間を関連づけ，キーワードの重要度を計算して，さらに，ベイジアンネットワークを用いてスコアを算出し，各 URL の総合スコアを計算してソーティングするというものである．さらに，共有，ユーザ別という 2 種類のベイジアンネットワークを用意し，個人の嗜好がスコアリングに与える影響を個々によって変化させる手法を取り入れた．そして，検索領域に関する知識の豊富さを表現する基準として，各検索語に関して認知度という指標を設け，結果比較に用いた．

## 参考文献


[1] 木谷強，高木徹，木原誠，関根道隆，フルテキストと抽出キーワードを利用した情報検索，情報処理学会研究報告，1996-NL-115，pp.129-134(1996)

[2] 藤本和則，本村陽一，松下光範，庄司裕子，知の科学 意志決定支援とネットビジネス，オーム社，pp. 61-92(2005)

[3] 酒井浩之，大竹清敬，増山繁，絞り込み語提示による一検索支援手法の提案，言語処理学会第 7 回年次大会，pp. 185-188(2001)

[4] Karen Sparck Jones，A statistical interpretation of term specificity and its application in retrieval(1972)

[5] Google，http://www.google.co.jp/