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

2006－MPS－61 (7)
2006／9／14

嗜好情報を用いたWeb検索フィルタリングアルゴリズム

堀田知宏　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．更に，現在の検索エンジンは，同一の検索語利用では，誰が検索を行っても同一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．更に，本アルゴリズムと他の解析手法との結果を比較する．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することで，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 Using Preference Information

TOMOHIRO HOTTA 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 compare the results of this algorithm and an other analytical technique. 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，それに伴いインターネットの利用者数は増加傾向にある．それに伴い，WWW(World Wide Web)で利用可能な情報量も比例して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最も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，同じキーワード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，文献[1]では，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く検索語で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，各ユーザの検索語の履歴を利用し，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，変数間の依存関係をグラフ構造で表した確率モデルである，ベイジアンネットワークを利用する．グラフの構造は，検索語をノードとし，同じURLが検索された際に用いられた検索語間にリンクを張って表現する．グラフとしては，全ユーザ共有用と，各ユーザ別用の，2つのベイジアンネットを作成し，利用者全体，個人，両者の嗜好を反映させている．

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，上記のツールを用いた，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，およびその結果を紹介し，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があげられる．協調フィルタリングとは，ユーザの検索，あるいは閲覧履歴から，各ユーザの要求に近いものを選択し，情報選択に利用する技術の総称である [2]．この手法では，アクティブユーザと同一の嗜好を持ったユーザグループを選択し，そのグループ内での履歴を利用して，アクティブユーザに推薦するコンテンツを選択する．しかし協調フィルタリングでは，アクティブユーザに対して一度グルーピングがなされると，ユーザの興味が別のものに変わった際には，対応するのに時間がかかってしまう可能性がある．また，ユーザグループ単位への情報推薦は可能であるが，グループを更に細分化した個人単位での情報推薦に対応するのは困難である．

2.2 提案システムの概要

本研究で示すシステムでは，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であり，ユーザ自身がその分野に興味があることを自覚している情報を指す．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，時々のユーザの状況によって変動する嗜好情報である．これは，ユーザの嗜好の変化を追うことができるため，静的嗜好とは別に，そのユーザの新たな嗜好が発見可能である．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．

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．ベイジアンネットワークとは，事象間の依存関係を非循環有向グラフで表し，各変数に割り当てられた条件付確率の集合を用いて構成される確率ネットワークである [2]．各変数の条件付確率は，CPT(Conditional Probability Table) と呼ばれる条件付確率表に保存される．さらに本研究では，キーワードへの得点付けの際に，全ユーザの静的嗜好情報，キーワードをノードとする共有ベイジアンネット，各々のユーザの静的嗜好情報，キーワードをノードとするユーザ別ベイジアンネットの 2 種類を使用する．共有ベイジアンネットは，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．また，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．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) 各ベイジアンネット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

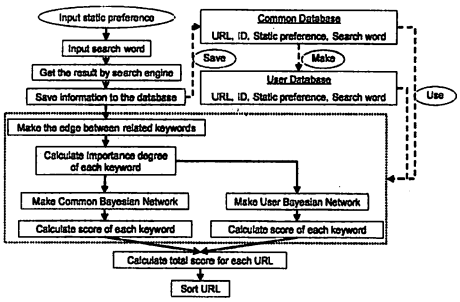

Fig. 1: Web Search Filtering Algorithm

3.1 各情報の保存

ステップ (1)〜(3) において，データベースに，URL，ユーザ ID，静的嗜好情報，検索語を保存する．ユーザ ID は，ユーザ別ベイジアンネットを形成する際に用いる．また，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．さらに，ステップ (3) の過程では，各ユーザ用のデータベースも作成する．

3.2 エッジ関係の生成方法

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．文献 [3] 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．

また本研究では，あるキーワードが，2 語以上で構成されている他のキーワードに含まれている場合，2 語以上のキーワードは，1 語のキーワードから派生したものとし，そのキーワード間にエッジを引いている．図 2 は，本研究におけるベイジアンネットワークの例である．四角の枠は静的嗜好情報，丸い枠は検索語を表している．

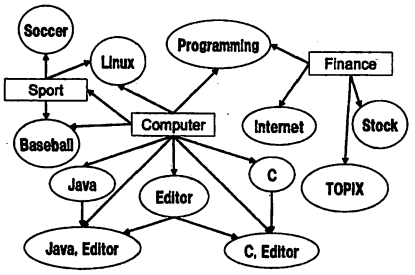

Fig. 2: Bayesian Network of Keywords

3.3 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，次式を用いて計算される．本研究では，ベイジアンネットのノードの属性を true，false としており，この $w(K)$ の値が，属性が true の場合の値となる．false の値は，1 と $w(K)$ の値の差で求められる．この式は，$tf \cdot idf$[4] を基にした計算式である．

$$w(K) = TF(K) \times IDF(K) \quad (1)$$

$$TF(K) = \frac{C(K)}{\sum_{K_i \in KWS} C(K_i)} \quad (2)$$

$$IDF(K) = \log \frac{\sum_{U_j \in URLS} \sum_{K_i \in KWS} C(U_j, K_i)}{\sum_{U_j \in URLS} C(U_j, K)} \quad (3)$$

$TF(K)$ は，データベース上に保存されている全てのキーワード数に対するキーワード K の出現頻度を表しており，$IDF(K)$ は単語の特定性を表す．また，$C(K)$ は，そのキーワードが使用された回数，$C(U, K)$ は，キーワード K が含まれている URL の個数を表している．

ここで，IDF の計算では，扱う情報の種類によって対数の底の値を変化させている．静的嗜好情報として入力される単語は，一般的な語が使用される可能性が高く，IDF の影響は検索語よりも少なく抑える必要があるためである．

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．

3.4 ベイジアンネットによる各キーワードの得点計算手法

式 (1) で $w(K)$ から各ノードの true，false の値を算出し，ステップ (6) で共有，ユーザ別ベイジアンネットを形成後，ステップ (7) でそれらを用い，各キーワードの得点を計算する．

静的嗜好情報に関しては，ユーザの情報が URL に付随する静的嗜好情報に一致した場合に得点を与え，$SPscore(K)$ の値を得る．検索語については，URL に付随する検索語群の得点の平均値を計算し，$SWscore(U)$ の値を算出する．この計算も，共有ベイジアンネット，ユーザ別ベイジアンネット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，$SPscore^C(U)$，$SPscore^U(U)$ と，2 つの検索語の得点，$SWscore^C(U)$，$SWscore^U(U)$ が得られることになる．なお，添え字の C，U はそれぞれ，共有ベイジアンネット，ユーザ別ベイジアンネットを指す．

3.5 各 URL への総合スコア計算手法

各キーワードの得点を取得後，ステップ (8) にて式 (4)〜(7) で URL の総合得点を計算する．

$$Cscore(U) = SPscore^C(U) + SWscore^C(U) \quad (4)$$

$$Uscore(U) = SPscore^U(U) + SWscore^U(U) \quad (5)$$

$$URLscore(U) = Cscore(U) + \beta(Uscore(U)) \quad (6)$$

$$\beta = \log \frac{\sum_{U_i \in URLS} C(U_i)}{C(U)} \quad (7)$$

$Cscore(U)$ は，共有ベイジアンネット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から算出されたスコアを表し，$URLscore(U)$ は，対象 URL の総合得点である．

ここで，式 (6) において，ユーザ別ベイジアンネットから計算されたスコアには，重み β を乗じている．β は，ユーザが検索した URL の出現確率の逆数によって決めたもので，ユーザ特有の嗜好をより強く反映させるためである．

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

今回の実験では，プログラミングに関する情報検索を例に取り，5名のユーザにはそれぞれ表1のように特徴を持たせている．静的嗜好情報の括弧内の数字は，そのユーザが使用した静的嗜好の比率である．Typeは，そのユーザがプログラミングに関して有識者か，素人かを表したものである．

有識者は，目的のURLを得るために，複数の検索語を用いることができる．そこで本研究では，ユーザ毎に，静的嗜好"コンピュータ"とリンクしている検索語間のエッジの個数を考え，その個数の上位2人を有識者とした．

Table. 1: User Data

UserName	Static Preference	Type
UserA	Computer(8), News(2)	Specialist
UserB	Computer(6), Sport(4)	Specialist
UserC	Computer(5), Sport(5)	Amateur
UserD	Computer(3), Finance(7)	Amateur
UserE	Finance(5), Sport(5)	Amateur

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，プログラミングに関連する語での検索結果をデータベースに登録している．有識者であるUserA，UserBは，"エディタ"，"開発環境"など，専門性の高い単語を使用するが，知識の乏しいユーザは，"Java，入門"などの抽象的な検索語群を使用することとしている．

4.2 実験結果

表2は，ソーティングの結果，上位に得られたURLの番号を表している．rankは，既存の検索エンジン，Googleで得られた各URLの順位を示している．表3は，検索語"プログラミング"での実験結果を表しており，それぞれのユーザに示された，上位5つのURLを示している．全てのユーザの静的嗜好は，そのユーザが最も使用している静的嗜好情報とした．また，ユーザ別ベイジアンネットを使用しない場合のデータと比較するため，表3中のCBNを用意した．

Table. 2: URL Table

URL No.	rank	URLName
URL1	5	http://www.microsoft.com/japan/msdn/student/challenge/
URL2	16	http://d.hatena.ne.jp/brazil/20050829/1125321936
URL3	17	http://www.sm.rim.or.jp/ shishido/
URL4	27	http://www.atmarkit.co.jp/fjava/rensai3/eclipsejava01/eclipse01.html
URL5	29	http://www.hyuki.com/
URL6	30	http://www.hyuki.com/jb/

Googleで得られた結果と比べ，表中のURLの順位が大幅に引き上げられていることがわかる．また特に，有識者であるUserAの本アルゴリズムを使用した検索結果に対し，知識の乏しいユーザでも，有識者とほぼ同等の結果が得られた．また，静的嗜好情報が"コンピュータ"でないUserEでも同様の結果が得られており，動的嗜好の反映が結果に表れている．

Table. 3: Result of This Algorithm

CBN	UserA	UserB	UserC	UserD	UserE
URL1	URL2	URL4	URL1	URL2	URL2
URL2	URL3	URL1	URL2	URL3	URL3
URL3	URL5	URL2	URL3	URL5	URL5
URL5	URL4	URL3	URL5	URL6	URL4
URL6	URL6	URL5	URL6	URL4	URL6

さらに，各ユーザ毎にURL順位が異なっているため，共有ベイジアンネットのみ使用した場合と比較しても，各ユーザそれぞれの嗜好が個々に表れているといえる．

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結び，それぞれのキーワードの得点を計算して，さらにベイジアンネットワークを用いてスコアを出し，各URLの総合スコアを算出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．

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] 大森信行，岡村潤，森辰則，中川裕志，tf・idf法を用いた関連マニュアル群のハイパーテキスト化，情報処理学会研究報告，1997-NL-121，pp.111-118(1997)

[5] Google，http://www.google.co.jp/

[6] 河村晃好，黒武者健一，佐藤亮一，芥子育雄，グループ嗜好モデルと視聴履歴を利用したコンテンツ検索サーバの試作，情報処理学会研究報告，2001-DBS-125，pp.177-184(2001)