ODPを利用した個人化検索システムの比較と効率化

伊美　裕司†　神原　義明††　大石　哲也†††
長谷川　隆三†††　藤田　博†††　越村　三幸†††

†九州大学工学部電気情報工学科
††九州大学大学院システム情報科学府
†††九州大学大学院システム情報科学研究院

1　はじめに

インターネットの普及とともに，ユーザが目的のページを検索するためにサーチエンジンを利用する機会が増えた．Webサーチエンジンの多くは，多数のユーザが求めるWebページを上位に表示している．

しかし，それぞれのユーザによって求められたページが上位に存在しないことも多く，その場合，ユーザが目的のページを探し出すのは困難である．また，ユーザがある目的に関連するページだけを大量に探したい場合，たとえ上位に目的のページが存在しても必要数に満たないなど不十分なことがある．ユーザはこれを補うために下位をみていくが，下位であるほど不必要なページが多く混在し，目的のページを探し出すのは困難になる．

この問題を解決するための研究の一つとして，検索にユーザの興味を反映させる個人化検索 (Personalized Search) がある．個人化検索を行うには，E-mailやユーザの過去の閲覧ページ履歴などのPersonal Document(文書) に基づいてユーザの興味・関心を分析し，ユーザプロファイル（User Profile, UP）として表現しなければならない．

本研究では個人化検索として研究されているもののうち，Open Directory Project（ODP）を利用する検索手法に着目する．ODPとは，Webページをその話題ごとに階層的なディレクトリ (カテゴリ) に分類した世界最大のWebディレクトリである．ODPを利用した個人化検索システムとして，既存の手法 [1] を用いたシステムと，我々の提案手法 [2] を用いたシステムについて実験，比較する．どちらのシステムもユーザプロファイルを基に，Webページの検索結果のReRanking（ソート）を行い，ユーザの目的とするページを検索結果の上位に集中させる．これらの検索結果の精度を検証・比較することで，個人化検索システムの有用性を検討し，今後の課題について考察する．

2　システム概要

本節では，本研究で比較するODPを利用した2つの個人化検索システムの概要を説明する．

2.1　既存手法

ODPのWebページ群はそのページが属するカテゴリに深く関連しており，カテゴリとそれに直接繋がったサブカテゴリには強い関連性がある．ODPはカテゴリをノードとし，カテゴリからサブカテゴリへの有向辺を持つ有向非巡回グラフ G とみなせる．既存の手法では，ODPのこのような性質を利用して，新たな有向非巡回グラフ G' を作成し，G' をユーザプロファイルとする．

まず，ユーザが訪れたページが属するカテゴリをODPを利用して判別し，ブラウジング履歴をモニタする．モニタした履歴をPersonal Documentとみなし，ユーザプロファイルとしてユーザの興味・関心があると思われるカテゴリ群をノードとする有向非巡回グラフ G_{sub} を作成する．作成された G_{sub} はODPのカテゴリ構造を持つ部分グラフである．このとき，各カテゴリノード v には関連重要度 $\beta(v)$ が割り当てられる．この関連重要度 $\beta(v)$ はユーザのカテゴリ v に対する優先傾向を示し，モニタした履歴に応じて初期値が与えられる．各カテゴリノードをつなぐエッジにも重み $d(v_i, v_j)$ が割り当てられており，これはカテゴリ v_i と v_i のサブカテゴリ v_j の関連性の高さを表す．

ユーザは検索の際，実際の検索クエリにかかわらず，検索結果からいくつかのページを選択するが，それらのページは必ずしも同じカテゴリに属しているわけではない．このことは，ユーザ独自の選択において，それらのページが属するカテゴリに潜在的な関連性が存在することを示している．そのため，これらのカテゴリが直接的に繋がっていない場合，カテゴリ同士を相関させる必要がある．その実現として，繋がっていな

い $\beta(v)$ の高いカテゴリ同士を相関させる仮想カテゴリノードを作成する．

ここで，重要なノードに繋がっているノードはやはり重要なノードであるという PageRank に類似したアルゴリズムを適用する．最終的なカテゴリ v に対するユーザの関連重要度 $\beta(v)$ は，$\beta(v)$ に直接繋がっているカテゴリの関連重要度 β とそのカテゴリを繋ぐエッジの重み d の積の総和を加算する処理を $\beta(v)$ の状態が平衡するまで再帰的に繰り返すことで求められる．

そのため，作成した部分グラフの全カテゴリについて，関連重要度の算出を複数回行い最終的な β の値を求める．このようにしてカテゴリ v に対するユーザの関連重要度 $\beta(v)$ を算出し，ユーザのカテゴリ優先傾向を表したグラフ G' をユーザプロファイルとする．

検索クエリに対して検索エンジンから返された検索結果から，各 Web ページの PageRank を抽出する．重要性の高い Web ページほど PageRank が高く，検索結果の上位に出現しやすい．この各 Web ページの PageRank とその Web ページが属するカテゴリ v の関連重要度 $\beta(v)$ の平均をとり，ユーザプロファイルを反映させた PageRank' を算出する．この PageRank' の高い順にソートすることで Reranking を行い，その結果をユーザに提示する．

2.2 我々の提案手法

ODP の Web ページ群はそのページが属するカテゴリに深く関連しており，さらに，ODP の Web ページ中に出現する単語はそのページが属するカテゴリ に関連した単語である可能性が非常に高い．我々は，ODP のこのような性質を利用してユーザプロファイルの作成を行う．

提案手法では ODP の最上位カテゴリ中，ユーザの最も興味・関心のあると思われるカテゴリを 1 つ選定する．さらにその下位にある各カテゴリをインデックスとしてユーザの関心度を表すユーザプロファイルベクトルを作成する．この最上位カテゴリは，検索クエリに対して提示された初期検索結果から選択された，ユーザの目的とする Web ページを Personal Document とみなし，選定される．

検索によって提示された各 Web ページについても同様に，ユーザプロファイルと同じインデックスを持つ Web ページのページベクトルを算出する．

このページベクトルとユーザプロファイルベクトルとのコサイン類似度を求めることにより，ユーザが目的とするページと各 Web ページがどの程度類似しているかを数値化する．算出されたコサイン類似度が高い程，その Web ページは，ユーザが Personal Document として選択した Web ページに類似していることになる．したがって，初期検索によって提示された各 Web ページを，このコサイン類似度の高い順にソートすることで Reranking を行い，その結果をユーザに提示する．

3 実験方法

任意のクエリをシステムに与え，検索を行う．この初期検索によって提示される件数は，検索結果の上位 100 件とする．このとき，ユーザに提示された検索結果における適合ページの割合を示す尺度として，適合率 (precision) を算出する．適合率は以下の式 (1) で定義される．

$$\text{適合率} = \frac{\text{上位 } x \text{ 件中の適合ページ数}}{\text{上位 } x \text{ 件}} \times 100\,[\%] \quad (1)$$

本研究において x は，$x = 10$, または $x = 20$ とする．

各システムの評価は，初期検索結果での適合率と Reranking 後の適合率の比較をすることによって行う．具体的には，以下の式 (2) によって向上率を算出する．

$$\text{向上率} = \frac{\textit{Reranking} \text{ 後の上位 } x \text{ 件の適合率}}{\textit{Reranking} \text{ 前の上位 } x \text{ 件の適合率}} \times 100\,[\%] \quad (2)$$

各システムの特徴を考察するため，検索するクエリによって Reranking 精度にどのように差が生じるか，向上度を比較して検証する．各個人化検索システムの有用性を検討し，今後の課題について考察する．

4 おわりに

実験結果の予想としては，既存手法より我々の提案手法のほうが良い結果になると考えている．既存手法では Web ページが属する ODP のカテゴリによって，そのページのカテゴリを判断するが，ODP に属していない Web ページに関してはカテゴリの判断ができないため，目的のページが ODP に属さない場合，検索結果の上位に出現することはないと考えられるからである．今後の予定として，この比較結果を基に，2 つの手法を組み合わせた新たな手法を提案する．

謝辞

本研究は科研費 21500102 の助成を受けたものである．

参考文献

[1] Ch.Makris, Y.Panagis, E.Sakkopoulos, A.Tsakalidis “Category ranking for personalized search” in the Data and Knowledge Engineering Journal (DKE), Elsevier Science, Vol. 60 , No. 1, pp. 109-125.

[2] 中村 徹，大石 哲也，越村 三幸，藤田 博，長谷川 隆三 “ODP を利用したユーザプロファイル作成と個別化検索システムの評価” 第 15 回 WI2 研究会