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

2008－DBS－145 (3)
2008－ＦＩ－ 91 (3)
2008／6／19

ODPを利用したユーザプロファイルを用いた個人化検索システム

神原 義明† 大石 哲也† 長谷川 隆三†† 藤田 博†† 峯 恒憲†† 越村 三幸††
†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

抄録. インターネットが急速に普及し，ユーザが目的のページを探すためにサーチエンジンを利用する機会が増えた．しかし，目的のページが検索結果の上位に存在しないことも多い．この問題を解決するための研究として，個人化検索 (Personalized Search) がある．個人化検索にはユーザの興味・関心を表すユーザプロファイルが必要であり，それはユーザの保有する文書などを分析することで生成される．

本論文では，Open Directory Project(ODP) を利用したユーザプロファイル作成手法を提案する．更に，本手法を利用した個人化検索システムを提案する．このシステムはユーザの保有する文書としてブックマークを利用する．実験の結果，ユーザプロファイルは妥当性が示され，システムは精度が向上した．

A Method to Make User Profiles using ODP for Personalized Search

Yoshiaki Kambara†, Tetsuya Oishi†, Ryuzo Hasegawa††, Hiroshi Fujita††,
Tsunenori Mine††, Miyuki Koshimura††
†Graduate School of Information Science and Electrical Engineering, Kyushu University
††Faculty of Information Science and Electrical Engineering, Kyushu University

Abstract. The user often uses search engines to search his intended pages in the Internet. However, the intended pages do not always exist on the higher rank of the results. The Personalized Search is promising approach to solve this problem. In the Personalized Search, User Profiles, which are often created from the documents of the user to present his/her interests, are well used. This paper proposes (1) a method for generating User Profiles based on Open Directory Project (ODP) and shows (2) a Personalized Search system using Book Marks as the documents for profiling. Some of experimental results show the validity of the method for creating User Profiles and the precision enhancement of this system.

1 はじめに

インターネットが急速に普及し，これに伴いWeb空間は年々肥大化してきている．このような状況の中，目的とするWebページを素早く探すためにWebサーチエンジンを利用する機会が多くなっている．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，ユーザの求めるWebページが一般的とはいえない場合，上位に表示されないことがある．検索時に利用する語句の多くは検索結果のページ数が莫大となるため，仮にユーザの目的とするページが下位に存在すれば，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ユーザが欲するページ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．

これらの問題を解決するために，ユーザの「必要としているもの」と「興味のあるもの」に合わせた検索を行う技術，つまり個人化検索（Personalized Search）が研究されている．個人化検索を行うためにはユーザの興味・関心を分析し，分析結果を表現しなければならない．この分析結果を表すものをユーザプロファイル（User Profile, UP）という．

ユーザプロファイルの作成には，ユーザが保有するPersonal Document(文書) から興味・関心を分析する必要がある．このPersonal Documentには様々なものが用いられ，具体的にはE-mailやユーザの過去の閲覧ページ履歴などがある．

本研究におけるユーザプロファイル作成手法ではOpen Directory Project（ODP）[1]を利用する．

ODPはWebページを手動でカテゴリ（話題）に分類しており，各カテゴリの特徴が明確である．ODPを検索システムに利用する研究は現在までにいくつか存在しているが，WebページがODPのどのカテゴリに属すかという情報を利用する研究が多い．そこで，我々は従来の手法に加え，ODPに属するWebページの内容そのものにも注目する．これにより，より細かい情報を取得できる．

本研究では上記のユーザプロファイル作成手法を用いた個人化検索システムとして，Personal Documentにブックマークを用いたシステムを提案する．本システムは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．この検索結果の精度を検証することで，個人化検索システムの有用性を検討する．

[1]Open Directory Project: http://www.dmoz.org/

2 関連研究

主要なサーチエンジンとして Google[2]がある．Google は PageRank[1] という評価システムを基に検索結果のランキングを行っている．PageRank は幾つのページからリンクされているか，質（重要度）が高いページからリンクされているかなどの情報を基に評価が行われている．本研究では，検索結果の各ページについて ODP のカテゴリをインデックスとしたページベクトルを算出し，ユーザプロファイルと比較を行うことでランキングを行う．

[2] や [3] ではクエリを曖昧，半曖昧，明確と分類して研究を行っている．また [4] では，ユーザプロファイルは使用する Personal Document により性質が変化する特徴を持つと述べている．この特徴を Z. Dou らは長期的ユーザプロファイル（日常的な趣味・興味など），短期的ユーザプロファイル（その場限りの調べものなど）と表現している．更に，長期的ユーザプロファイルと短期的ユーザプロファイルは同時に対応することが困難なため，[2] や [3] によるクエリの分類を基にそれぞれの特徴を調べる実験を行っている．その結果，明確なクエリには短期的ユーザプロファイルを，曖昧，半曖昧なクエリには長期的ユーザプロファイルを利用すると効果が高いことを示した．

長期的ユーザプロファイルに特化したシステムとしては [5] がある．これは，ユーザのクリック履歴に基づいて ODP の最上位カテゴリごとに評価を与え，ユーザプロファイルを作成している．

本研究では，ブックマークを利用することで長期的ユーザプロファイルに特化したシステムを提案する．そして，[2] や [3] によるクエリの分類を考慮し，日常的な興味に対応したシステムとしての有用性を検討する．

3 ODP を利用したユーザプロファイル作成

3.1 Open Directory Project

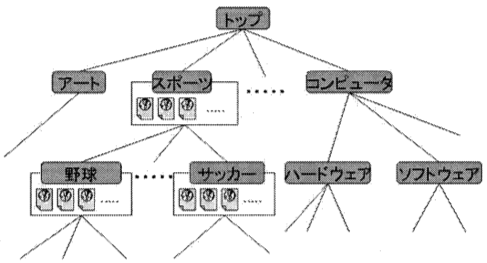


図 1: Open Directory Project

Open Directory Project(ODP) は，Web ページを階層的なカテゴリ（話題）に分類した世界最大の Web ディレクトリである．図 1 に ODP の概要を示す．

[2]Google : http://www.google.co.jp/

ODP では，人手によって Web ページをその内容に該当するカテゴリに分類する．このことから，ODP の Web ページに出現する単語は，Web ページが属するカテゴリ（話題）に関連する可能性が非常に高い．また，人手によって Web ページが分類されているためカテゴリが細かい．また，原則として一つのページは一つのカテゴリにしか登録されない[3]．

本研究で提案するユーザプロファイル作成法は，ODP のこのような性質を利用する．

3.2 ユーザプロファイル作成

まず，ユーザプロファイルについて具体的に説明する．

本研究のユーザプロファイルは，ODP の最上位カテゴリ[4]をインデックスとし，各カテゴリへの興味を数値で表したベクトルで表現される．

表 1: ユーザプロファイル (例)

ユーザ	アート	スポーツ	コンピュータ
A	0.4	0.1	0.5
B	0.2	0.7	0.1

表 1 はカテゴリが「アート」「スポーツ」「コンピュータ」の 3 つであった場合の例である．ユーザ A は「スポーツ」にはあまり興味がないが「アート」と「コンピュータ」には同程度の興味を持っている．一方，ユーザ B は「スポーツ」に大きな興味を持っており，他の 2 つにはあまり興味を持っていない．このようなベクトルがユーザプロファイルである．

次に，提案手法におけるユーザプロファイル作成の手順を以下に示す．

手順 1　ODP の出現単語分析
手順 2　Personal Document の出現単語抽出
手順 3　ユーザプロファイル作成

手順 1　ODP の出現単語分析
・ODP からの単語抽出領域

まず，ODP の各カテゴリから，そのカテゴリに出現する単語を抽出する．単語抽出には形態素解析エンジンである MeCab[5]を用いる．この際，単語の抽出領域として 2 通りが考えられる．一つは，カテゴリに属している Web ページの内容そのものを抽出領域とする．もう一つは，ODP の各カテゴリに付与された，カテゴリに属するページを表すタイトルと紹介文を抽出領域とする．前者は出現する単語数が多くなりより詳しくそのページの特徴を捉えられる．しかし，不要な語や一般的な語などのノイズが増える．後者は，出現単語数こそ少ないが，そのページを端的に表す単語が抽出できる．この理由から，本手法では後者の抽出領域を利用している．

[3]例外 :“ 地域 ” に関する内容の場合，“ 地域 ” とそれ以外の 2 箇所に登録されることがある

[4]「アート」,「オンラインショプ」,「ゲーム」,「コンピュータ」,「スポーツ」,「ニュース」,「ビジネス」,「レクリエーション」,「家庭」,「科学」,「各種資料」,「健康」,「社会」,「地域」の 14 項目

[5]MeCab : http://mecab.sourceforge.net/

・単語の重み計算

次に，抽出した単語 t_i の各カテゴリ c_j に対する重み $W(t_i, c_j)$ を計算する．重みとは、単語の各カテゴリに対する関連度のことである．例えば，単語「野球」を考える．「野球」はカテゴリ「スポーツ」への関連が高いので，$W($“野球”,“スポーツ”$)$ は高い値をとる．しかし，カテゴリ「アート」に対しては関連が低いので，$W($“野球”,“アート”$)$ は低い値となる．本手法では，ODP の最上位カテゴリの Web ページ群から抽出した単語の出現頻度を基に各カテゴリへの重みを算出する．

あるカテゴリに偏って多く出現する単語はそのカテゴリへの関連性が高く，カテゴリを特徴付ける単語として有用である．一方，様々なカテゴリに同頻度で出現する単語は各カテゴリへの関連性が低く，特定のカテゴリを特徴付ける単語としては有用でない．

そこで，カテゴリでの単語 t_i の出現の偏りをエントロピーを用いて求め，各単語 t_i がカテゴリを特徴付けるために有用かどうかを評価した基本評価値 w_{t_i} を算出する．式 (1) に単語 t_i のエントロピー H_{t_i} を表す．

$$H_{t_i} = -\sum_j P_{t_i}(c_j) * log(P_{t_i}(c_j)) \qquad (1)$$

- $n(t_i, c_j)$: カテゴリ c_j 内の単語 t_i の出現頻度
- $N_{t_i} = \sum_j n(t_i, c_j)$: 全カテゴリにおける単語 t_i の総出現頻度
- $P_{t_i}(c_j) = n(t_i, c_j)/N_{t_i}$: 全カテゴリにおけるカテゴリ c_j 内の単語 t_i の出現割合

単語 t_i が全カテゴリで等しく出現するとき，即ち $P_{t_i}(c_j) = 1/N_c \quad (j = 1 \cdots N_c, N_c$: カテゴリの総数$)$ のとき，エントロピー H_{t_i} は最大値 $log(N_c)$ をとる．また，単語 t_i があるひとつのカテゴリにのみ出現するとき，このエントロピー H_{t_i} は最小値 0 をとる．エントロピー H_{t_i} を用いて以下の式 (2) の通り基本評価値 w_{t_i} を定義する．

$$w_{t_i} = log(N_c) - H_{t_i} \qquad (2)$$

基本評価値 w_{t_i} は単語 t_i が特定のカテゴリ内で偏って出現するほど大きな値をとる．このことにより，偏りが大きい単語はそのカテゴリを特徴付ける単語として有用である．

以上で求められた基本評価値 w_{t_i} と各カテゴリでの出現頻度を $n(t_i, c_j)$ を基に，カテゴリ c_j に対する単語 t_i の重み $W(t_i, c_j)$ を以下の式 (3) で定義する．

$$W(t_i, c_j) = P_{t_i}(c_j) * w_{t_i} \qquad (3)$$

この式を ODP 内の全ての出現単語 t_i に用いて重みを算出する．

表 2: 単語の出現例

単語	アート	スポーツ	コンピュータ
t_1 サッカー	2	34	1
t_2 本	15	8	13

表 3: エントロピー H_{t_i} と基本評価値 w_{t_i}

単語	H_{t_i}	w_{t_i}
t_1 サッカー	0.48	1.11
t_2 本	1.54	0.05

表 2 に単語の出現例を，表 3 に算出したエントロピーと基本評価値を示す．この例では，単語 t_1（サッカー）はカテゴリ「アート」に 2 回,「スポーツ」に 34 回,「コンピュータ」に 1 回出現している．カテゴリ「スポーツ」に偏った単語 t_1 のエントロピー H_{t_1} は 0.48 と小さい値をとり，基本評価値 w_{t_1} は式 (2) より 1.11 となる．一方，単語 t_2（本）は「アート」に 15 回,「スポーツ」に 8 回,「コンピュータ」に 13 回と，各カテゴリに似通った頻度で出現している．この単語 t_2 の場合，エントロピー H_{t_2} は 1.54 と最大値 $log(3) = 1.59$ に近い値となり，基本評価値 w_{t_2} は 0.05 と小さい値になる．

表 4: 単語の重み

単語	アート	スポーツ	コンピュータ
t_1 サッカー	0.06	1.02	0.03
t_2 本	0.02	0.01	0.02

得られた基本評価値 w_{t_i} を基に式 (3) を用いて計算した重みを表 4 に示す．表 4 から，特定のカテゴリで偏って出現する単語「サッカー」は，それ自身が多く出現するカテゴリに対して重みが大きくなり，そのカテゴリへの関連度が大きくなる．また，どのカテゴリにも均等に出現する単語「本」はどのカテゴリに対しても重みは小さくなる．

手順 2　Personal Document の出現単語抽出

本手法ではユーザプロファイルを作成するために，ユーザの興味が現れる文書（Personal Document）を用いる．ここでは，ODP での単語抽出と同様に，Personal Document に出現した単語（User Term : ut）が抽出される．

手順 3　ユーザプロファイル作成

手順 1 で計算した重み $W(t_i, c_j)$ と手順 2 で抽出した User Term を用いて，以下の式 (4) でユーザプロファイルのカテゴリ値 U_{c_j} を定義する．本手法では，この値をユーザのカテゴリ c_j への興味の度合いとする．

$$U_{c_j} = \sum_i N_{ut_i} * W(ut_i, c_j) \qquad (4)$$

- ut_i : User Term
- N_{ut_i} : 単語 ut_i の Personal Document 内での出現頻度

カテゴリ値 U_{c_j} は ut_i のカテゴリ c_j への重みと ut_i の Personal Documents 内での出現頻度の積を全ての ut_i で足し合わせたものである．

全てのカテゴリに関して式 (4) を適用すると，最終的にユーザプロファイルは以下のベクトル **UP** で表

現される．尚，ユーザプロファイルは $|\mathbf{UP}| = 1$ となるように正規化して用いる．

$$\mathbf{UP} = [U_{c_1}, U_{c_2}, \cdots, U_{c_{N_c}}] \quad (5)$$

表 5: Personal Document の単語出現頻度（例）

	サッカー	本
出現頻度	4	13

手順 1 で利用した例におけるユーザプロファイルの計算例として，Personal Document で「サッカー」と「本」の二つの単語のみが抽出された場合を考える．Personal Document での単語の出現頻度は表 5 の通りである．この場合のユーザプロファイルを式 (4) を用いて求めると $\mathbf{UP} = [0.12, 0.99, 0.08]$ となる．カテゴリ「スポーツ」の値が他に比べて大きいので，このユーザプロファイルを持つユーザは「スポーツ」への興味が大きいことになる．

4 個人化検索システムの検討

先に提案したユーザプロファイル作成手法は多種多様な Personal Document に対して利用可能であり，それを用いた個人化検索システムもいくつか考えられる．本節では，その中でもブックマークに基づいた個人化検索システムについて検討する．

4.1 個人化検索システムの背景と特徴

ユーザプロファイル作成に用いる Personal Document としては，E-mail，Web ページの閲覧履歴，検索履歴など多くのものが挙げられる．これは 2 章で述べたように，長期的ユーザプロファイル，短期的ユーザプロファイルとして分類される．本システムでは，長期的ユーザプロファイルの Personal Document としてブックマークを使用する．ブックマークはブラウジング時に多くのユーザが利用する．加えて，その内容やフォルダ分けなどの状態はユーザ毎に異なり，ユーザの興味・関心をよく現している．

本研究では，ブックマークのフォルダ分けを利用し，各フォルダごとにユーザプロファイルを作成する．このことで，各フォルダのユーザプロファイルがユーザの興味・関心に合わせた特徴を持つことを図る．これは，各フォルダごとに作成されたユーザプロファイルをクエリの目的に合わせて選択するためである．

4.1.1 個人化検索システムの概要

個人化検索システムの概要を図 2 に示し，システムの一連の処理を説明する．

システムの処理手順は以下の通りである．

処理 1 ブックマークに基づくユーザプロファイル作成・保存

処理 2 検索・ReRanking・結果表示

処理 1 ブックマークに基づくユーザプロファイル作成・保存

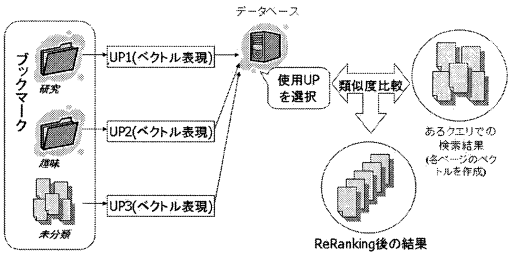


図 2: 個人化検索システム概略図

ブックマーク内のインターネットショートカット一つひとつを対象とし，各フォルダごと（未分類も一つのフォルダとみなす）にユーザプロファイルを作成する．（手法は 3.2 節と同様）

単語抽出領域はインターネットショートカットから読み取った URL のページの内容そのものとする．3.2 節で述べたように Web ページの内容を抽出領域とするとノイズが増える．しかし，この場合はユーザが分けたフォルダごとにユーザプロファイルを作成するため，各フォルダには似通った Web ページが集まる．このため，ノイズはほぼないと考えられる．ユーザプロファイルがどの程度フォルダ内の Web ページの特徴を反映しているかは後述の予備実験で検証する．

表 6: ブックマークの例

ブックマーク	内容
趣味	サッカーのページが 5 つ，音楽のページが 2 つ
研究	コンピュータや科学のページが 4 つ
買い物	オンラインショップが 3 つ
未分類	お店や天気など 4 つ

ここで，ユーザが表 6 で表されるブックマークを持つとする．この場合，フォルダ“趣味”のユーザプロファイルはサッカーと音楽の Web ページが存在するため，「スポーツ」と「アート」の値が高くなる．また，“研究”のユーザプロファイルは「コンピュータ」と「科学」が，“買い物”のユーザプロファイルは「オンラインショップ」のみが高くなる．本システムでは“未分類”のユーザプロファイルも作成され，「ニュース」や「レクリエーション」などのカテゴリが高くなる．

作成されたユーザプロファイルは各フォルダ名でデータベースに保存される．

処理 2 検索・ReRanking・結果表示

各フォルダごとのユーザプロファイルから，使用するユーザプロファイルをユーザが選択する．例えば，クエリ [中田] で検索をする際，サッカーの中田英寿選手に関するページが欲しい場合はユーザプロファイル“趣味”を選択する．また，中田英寿選手のグッズが欲しい場合はユーザプロファイル“買い物”を選択する．

次に，クエリを入力し初期検索を行う．検索には Yahoo!デベロッパーネットワーク[6]のウェブ検索を利用する．このとき，検索結果の上位 100 件の各 Web

[6]Yahoo!デベロッパーネットワーク：
http://developer.yahoo.co.jp/

ページごとのページベクトルを作成する．ページベクトルの作成手法は3.2節のユーザプロファイル作成手法と同様だが，抽出領域としてSNIPPET（検索結果に表示される各ページの説明文）を利用する．Webページの内容そのものを抽出領域とすると，検索結果のWebページの数は莫大であるため，検索結果の全ページベクトルを算出する時間は膨大なものになる．このため，本手法ではSNIPPETを利用し，処理時間の短縮を図る．

作成した各ページのページベクトルXと選択したユーザプロファイルYの類似度を以下の式(6)で表されるコサイン類似度で算出する．

$$CosSim(X,Y) = \frac{XY}{||X||||Y||} \tag{6}$$

- 内積　：$XY = \sum_{i=1}^{n} x_i y_i$
- ノルム：$||X|| = \sqrt{\sum_{i=1}^{n} x_i^2}$

ただし，ページベクトルXとユーザプロファイルYは正規化されているため，$||X|| = ||Y|| = ||X||||Y|| = 1$である．よって，実際は以下の式(7)を利用している．

$$CosSim(X,Y) = XY \tag{7}$$

最後に，初期検索結果の上位100件をこの類似度を基にソート，つまりReRankingを行い，この結果をHTMLファイルに出力する．

5　実験

5.1　ユーザプロファイルの特徴反映度検証

5.1.1　目的

ユーザプロファイルは3.2節で説明したように，ODPに登録されているWebページから抽出した単語を基に算出している．しかし，ブックマーク内のWebページはODPに登録されているとは限らない．また，ブックマーク内のフォルダで様々なカテゴリから取得されたWebページが混在することがある．これらのフォルダから作成したユーザプロファイルが，フォルダに取得されたWebページの特徴をどの程度反映するか検証する．

まず，ブックマークのフォルダに取得するWebページの選択基準を以下の2つとする．

(a) ODPから取得する場合

(b) 検索エンジンを利用して取得する場合

この分類を基に，以下の2つの観点から検証を行う．

(A) (a)ODPから取得する場合と(b)検索エンジンを利用して取得する場合の比較

(B) 多数のカテゴリが混在するフォルダから作成したユーザプロファイルの特徴反映度

観点(A)は(a)と(b)を比較することで，ODPに登録されていないWebページからでも該当するカテゴリの特徴を持ったユーザプロフィルが作成されるのかを検証する．観点(B)では，このユーザプロファイルがどのような特徴を持つのか，そしてユーザプロファイルとして個人化検索に利用可能なのかを検証する．

5.1.2　実験手順

手順1　フォルダに目的のWebページを取得し，ユーザプロファイルを作成する

手順2　理想のユーザプロファイルとのコサイン類似度を算出する

手順1　フォルダに目的のWebページを取得し，ユーザプロファイルを作成する

まず，各フォルダにWebページを取得する．観点(A)において，(a)ODPから取得する場合は目的のカテゴリに登録されているWebページをODPから取得する．(b)検索エンジンを利用して取得する場合はユーザが手動で検索を行い，目的のカテゴリに該当すると判断したWebページを取得する．観点(B)では，取得方法は観点(A)の(b)と同様だが，多数のカテゴリから取得したWebページを同一のフォルダに入れる．尚，取得するWebページの数は(a)，(b)共に各カテゴリで5個ずつとする．

手順2　理想のユーザプロファイルとのコサイン類似度を算出する

次に，作成したユーザプロファイルと理想のユーザプロファイルを比較する．ここで言う理想のユーザプロファイルとは，対応するカテゴリの値のみが最大値を取り，他のカテゴリの値は0となるベクトルのことである．例えば，フォルダに「アート」に対応するページのみを取得した場合は「アート」の値が1，他の値は0となる．「アート」と「コンピュータ」に関して取得した場合はそれぞれの値が$1/\sqrt{2}$，他の値は0となる（$|\mathbf{UP}| = 1$で正規化しているため）．上記の例を図3に示す．比較にはコサイン類似度（式7）を用い，算出された値を基に検証を行う．

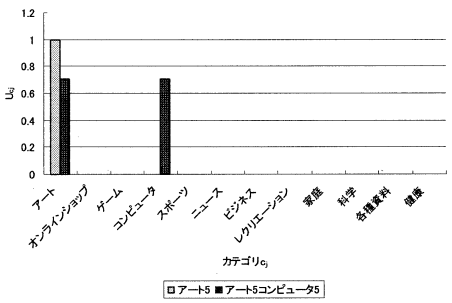
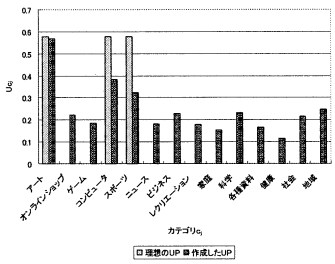
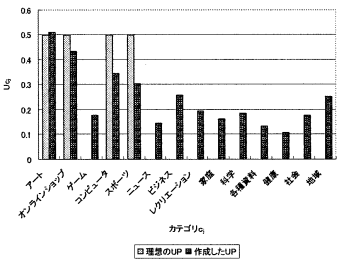


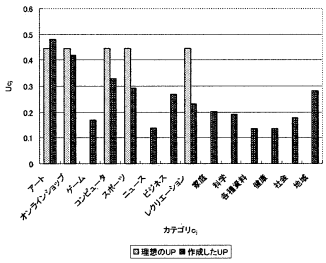
図3: 理想のユーザプロファイル



「アート」「コンピュータ」「スポーツ」



「アート」「コンピュータ」「スポーツ」
「オンライン」



「アート」「コンピュータ」「スポーツ」
「オンライン」「レクリエーション」

図 5: 多数のカテゴリが混在する UP（カテゴリ 3 件，4 件，5 件）

5.1.3 結果・考察

(A) (a)ODP から取得する場合と (b) 検索エンジンを利用して取得する場合の比較

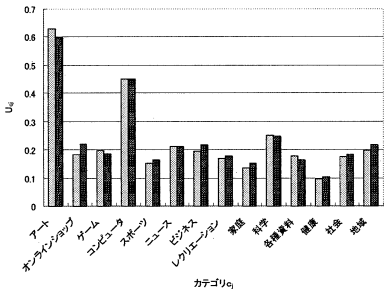


図 4: （左）(a)ODP から取得した UP（右）(b)Web 空間から一般的に取得した UP

図 4 は (a) と (b) の場合のユーザプロファイル (UP) のグラフである.「アート」と「コンピュータ」のカテゴリに対応するページを各 5 個ずつ，各々の基準で取得した結果である.

この結果をみると，ODP に登録された Web ページ以外のインターネットショートカットからでも，有効なユーザプロファイルが作成可能なことがわかる. また，それぞれ「アート」「コンピュータ」の値が高くなっているが，多少のばらつきや他のカテゴリのノイズがある．これは取得してきたページによって出現する単語に違いがあるためである．このノイズを軽減するためには，ODP から抽出した単語の洗練を行う必要がある.

表 7: (A) 理想のユーザプロファイルとの類似度

ユーザプロファイル	(a)	(b)
類似度	0.765	0.743

表 7 は (a) と (b) それぞれと理想のユーザプロファイルとの比較である．それぞれ 7 割以上の類似度となっており，ODP から取得した場合でも検索エンジンを利用して取得した場合でも，ユーザプロファイルはフォルダ内の Web ページの特徴を反映している.

(B) 多数のカテゴリが混在するフォルダから作成したユーザプロファイルの特徴反映度

図 5 に理想のユーザプロファイル (UP) と多数のカテゴリが混在するユーザプロファイル (UP) の比較を行ったグラフを示す．これはカテゴリを 2 件，3 件，4 件，5 件と増やしたユーザプロファイルであり，図には 3〜5 件のグラフを示している．各カテゴリは該当するページをそれぞれ 5 個ずつ取得した．順に結果を見ていくと，追加したカテゴリの値が増加している．また，徐々に該当しないカテゴリの値が増えている．特に，5 件の場合は「レクリエーション」の値がそれほど突出していない．これはノイズの影響が大きくなっていることが原因である．これも観点 (A) と同様に，単語の洗練を行うことで軽減できる.

表 8: (B) 多数のカテゴリが混在する UP の類似度

カテゴリ数	2 件	3 件	4 件	5 件
類似度	0.743	0.734	0.796	0.785

表 8 をみると，理想のユーザプロファイルとの類似度は 7〜8 割を保っていることがわかる．これより多数のカテゴリが混在する場合でも，少数の場合と同様に，ユーザプロファイルはフォルダ内の各 Web ページの特徴を反映しており，個人化検索システムに利用できる.

5.2 評価実験

5.2.1 目的

提案した個人化検索システムの性能を評価する．各クエリの検索結果において ReRanking 前後の精度を比較・考察する．これにより長期的ユーザプロファイルに特化したシステムとして，Personal Document にブックマークを用いた個人化検索システムの妥当性を検証する.

表 9: 評価検索結果（全クエリ（100 件）の平均）

	ReRanking 前			ReRanking 後			
	適合率 [%]	再現率 [%]	MAP	適合率 [%]	再現率 [%]	MAP	MAP の向上率 [倍]
上位 10 件	42.40	15.09	0.192	53.80	27.95	0.452	2.35
上位 20 件	40.75	27.95	0.407	51.50	41.99	0.702	1.72

表 10: 評価検索結果（適合率により分類）

分類 (クエリ件数)		ReRanking 前			ReRanking 後			適合率の向上率 [倍]
		適合率 [%]	再現率 [%]	MAP	適合率 [%]	再現率 [%]	MAP	
明確なクエリ	上位 10 件	94.29	14.69	0.143	91.91	14.10	0.254	0.98
(21)	上位 20 件	87.38	26.75	0.134	90.71	27.62	0.257	1.04
半曖昧なクエリ	上位 10 件	47.37	17.63	0.115	58.95	21.15	0.177	1.24
(38)	上位 20 件	44.34	30.79	0.161	57.11	39.82	0.278	1.29
曖昧なクエリ	上位 10 件	11.22	12.94	0.054	29.51	32.11	0.080	2.63
(41)	上位 10 件	13.54	25.94	0.211	26.22	51.36	0.266	1.94

5.2.2 実験手順

以下の処理を無作為に選んだ100件のクエリに対して行う．使用するユーザプロファイルは，5.1 節で作成したもの（約 50 件）から目的に合わせて選択する．

まずクエリを入力し，検索を行う．次に検索結果の上位 100 件を ReRanking する．そして ReRanking 前後での上位 10 件，20 件において，ユーザが必要としている目的のページの順位を調べる．また，上位 100 件に含まれる目的のページ総数（正解数）も調べ，適合率，再現率を算出する．尚，正解の判断はユーザ（検索者）の基準で行う．

- 正解数：上位 100 件に存在する目的のページ数
- 適合率 $=$ $\dfrac{\text{上位 } x \text{ 件中の目的のページ数}}{\text{上位 } x \text{ 件}}$
- 再現率 $=$ $\dfrac{\text{上位 } x \text{ 件中の目的のページ数}}{\text{正解数}}$

加えて，システムの性能を測るため，Mean Average Precision(MAP) と MAP の向上率を算出する．

MAP は各クエリにおける Average Precision(AP) を求め，その平均をとることで算出できる．クエリ i における AP_i は以下の式（8）で定義される．

$$AP_i = \frac{1}{N_i} \sum_{k=1}^{M_i} \frac{k}{r_{i,k}} \tag{8}$$

- N_i：クエリ i における正解数
- M_i：正解検出数
- $r_{i,k}$：クエリ i の k 個目の正解が検出された順位

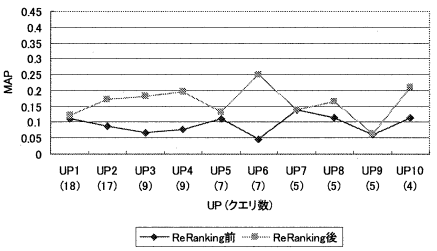
以上の分析結果をみることで，システムの特徴を考察する．

5.2.3 結果・考察

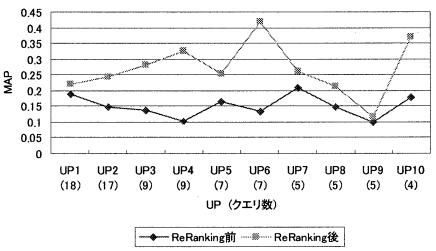
評価実験における検索結果の分析結果として表 9 を示す．

MAP の向上率をみると，上位 10 件では 2.35 倍，上位 20 件では 1.72 倍である．これは，ユーザの目的のページが上位に集中したことを示す．

また，表 10 に各クエリの適合率において 70%超を明確なクエリ，20%超 70%以下を半曖昧なクエリ，20%以下を曖昧なクエリとした分析結果を示す．これをみると，クエリが曖昧である程 ReRanking 後の結果が良くなっている傾向がわかる．これは，明確なクエリに関しては元々Web サーチエンジンの性能が良いことからほとんど変化が見られないためである．



検索結果の上位 10 件における MAP



検索結果の上位 20 件における MAP

図 6: 各 UP における MAP

図6は各ユーザプロファイルにおけるMAPの値を示したグラフである．ユーザプロファイルは使用したクエリ数により並び替え，上位10件をグラフに起こした．ReRanking前後で結果がよかったユーザプロファイルはUP4，UP6，UP10であり，それぞれオンラインショップ，ゲーム，本に関するフォルダである．結果が悪かったUP1，UP7，UP9は，それぞれレクリエーション，科学，ニュースに関するフォルダである．平均的にはReRanking前のMAPが低いほどReRanking後の結果がよくなる傾向がみられた．UP9は例外で，正解がほとんどないクエリを含んだためにこのような結果となった．

これらの結果より，Personal Documentにブックマークを用いた個人化検索システムは，検索結果のReRankingを行うことで上位にユーザの目的のページを集め，長期的なユーザプロファイルに特化したシステムとして有用であることが示せた．

最後に，結果が向上しなかったクエリや向上率が低かったクエリについて考察する．個々の検索結果をよくみると，検索結果のページベクトルにおいて，高い値をもつカテゴリが分散している場合は，ReRankingにより結果が良くなる．しかし，多くのページベクトルにおいて高い値をもつカテゴリが偏っていると，ReRankingの結果がうまく反映されないことがわかった．これは現在のユーザプロファイルのインデックスがODPの最上位カテゴリ14項目のみで作られていることが主な原因である．

ODPにおいてカテゴリ「アート」の下位層には「コミック」,「アニメーション」,「ビジュアルアート」,「音楽」，他数種の下位カテゴリが存在する．例えば，カテゴリ「アート」の値が高いユーザプロファイルを用いて個人化検索を行う．このとき，初期検索結果の上位100件においてカテゴリ「アート」の値が高いページが多数存在した．ユーザの目的のページが下位層の「音楽」に関係するものであった場合でも，ReRankingは単純に使用するユーザプロファイルとの類似度が高いWebページ，ここではカテゴリ「アート」の値が高いものを上に持ってくる．このため，結果的には「コミック」や「ビジュアルアート」に関するユーザの意図しないページが上位にくる可能性が存在する．

この問題の解決方法としては，3.2節のユーザプロファイル作成手法においてODPの下位層にも対応していくことが考えられる．下位層まで拡張したユーザプロファイルを実装すると，より詳細な興味・関心を表したユーザプロファイルが作成可能となる．そうなれば，ユーザのより細かな目的に答える個人化検索システムを作成できると思われる．

6 おわりに

本研究ではODPを利用したユーザプロファイルの作成の妥当性の検証，およびブックマークを基に作成したユーザプロファイルを利用した個人化検索システムの検討を行った．

ユーザプロファイル作成では，多少偏りはあるが，Webページの取得方法に関わらずブックマーク内のフォルダ毎でWebページの特徴が反映された．また，多数のカテゴリが混在しているPersonal Documentからでも各カテゴリの値が高くなったユーザプロファイルが作成された．

個人化検索システムにおいては，ユーザプロファイルを利用したReRanking後の検索結果は平均してユーザの目的のページが上位に集中した．ただし，クエリと使用するユーザプロファイルの組み合わせによってはReRankingの結果が悪化することもある．この問題の主な原因はユーザプロファイル作成時にODPの最上位カテゴリのみをインデックスとしているためである．

また，今回の研究を通して出てきた課題を以下に示す．

- ユーザプロファイルの拡張
- ODPから抽出された単語の洗練
- 短期的ユーザプロファイルへの対応
- 実験データの増加
- クラスタリング技術の利用

これらの項目を中心に，システムの更なる発展・改良を目指す．

謝辞

本研究の基礎は九州大学（現：九州日本電気ソフトウェア株式会社）の永田 廣人により行われた．

参考文献

[1] L.Page, S.Brin, R.Motwani, and T.Winograd :
「The PageRank Citation Ranking: Bringing Order to the Web」,
http://google.stanford.edu/~backrub/pageranksub.ps, 1998.

[2] P.A.Chirita, C.Firan, and W.Nejdl :
「Summarizing local context to personalize global web search.」, In Proc. of CIKM '06, 2006.

[3] P.A.Chirita, W.Nejdl, R.Paiu, and C.Kohlschütter :
「Using odp metadata to personalize search」, In Proc. of SIGIR '05, 2005.

[4] Z.Dou, R.Song, and J.Wen :
「A Largescale Evaluation and Analysis of Personalized Search Strategies」, WWW 2007, 2007

[5] F.Qiu, and J.Cho:
「Automatic identification of user interest for personalized search」, WWW 2006, 2006

[6] 永田 廣人 他：
「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提案」, JAWS2007, 2007