Web コミュニティに基づく情報検索の個人化手法

黄　海湘　　藤井　敦　　石川　徹也
筑波大学大学院図書館情報メディア研究科
{lectas21, fujii, ishikawa}@slis.tsukuba.ac.jp

1. はじめに

World Wide Web は、我々の貴重な情報源になっており、検索エンジンは Web から必要な情報を効率良く取得するためのツールである。しかし、既存の検索エンジンを使うと、ユーザが異なる場合でも、同じ検索キーワードに対して同じ結果しか得られない。この問題を解決するためには、情報検索の個人化（personalization）が必要になる。

本研究は、Web コミュニティを用いた情報検索の個人化手法を提案する。Web コミュニティ[1]とは、リンクで密に結合したページ群であり、現実社会における（ページ作成者の）興味や関心を反映している可能性がある。そこで、Web コミュニティを用いることで、ユーザの嗜好に応じた情報検索を実現することができると考えた。

以下、2 章で先行研究について検討する。3 章で本研究の提案手法について説明し、4 章で評価実験結果について考察する。

2. 情報検索の個人化に関する先行研究

2.1 概要

情報検索の個人化に関する先行研究は、二種類の方式に大分することができる。

一つは、ユーザの明示的な行動履歴（投票履歴、購買履歴など）や暗黙的な行動履歴（ページの閲覧履歴、アクセスログなど）のような個人情報を利用してユーザモデルを構築し、情報を選別する方式である。これを「ユーザモデリング方式」と呼ぶ。

もう一つは、情報源に含まれるデータの内容を解析してモデル化し、様々な「メニュー」を用意する。そして、ユーザに欲しいメニューを選ばせることによって異なるユーザに異なる情報を提供する方式である。この方式は「誰が何故そのメニューを選んだのか」というユーザ情報は考慮しないため、「データモデリング方式」と呼ぶ。

2.2 ユーザモデリング方式

ユーザモデリング方式は、ユーザモデルの構築手法によって、「情報フィルタリング」と「協調フィルタリング」に細分類することができる。

「情報フィルタリング」は、対象とするユーザに関する情報だけを利用する。

MONTAGE [2]では、ユーザのアクセス履歴を分析して、ユーザプロファイルを構築する。これを利用して、ユーザの嗜好を内容別にカテゴリ化し、ユーザがページを閲覧しやすいように専用のスタートページを作る。さらに、そのページはユーザの一日の行動に合わせて変化する。しかし、個人化の効果はユーザの利用頻度に依存する。また、行動傾向が反映されるまでに時間差が生じる。

「協調フィルタリング」は、対象ユーザ以外のユーザに関する情報も利用する。

GroupLens [3]は Usenet ニュースのフィルタリングを行う。ニュースに対する各ユーザの明示的な評価を利用して、対象ユーザと嗜好が似ているユーザを選び、ニュース記事を提供する。しかし、ユーザの評価値が疎であると、適切に機能しない。さらに、誰も読んでない記事には推薦できない。例えば、Amazon.com では大半の情報が未評価である[4]。

Sugiyama ら[5]は、閲覧履歴からユーザの長期嗜好と一日限りの短期嗜好を抽出し、他のユーザに関する情報を併用して、ユーザプロファイルを構築した。しかし、ユーザの嗜好抽出に、Web ページの閲覧時間を利用するので、閲覧途中で席を外すような利用状況に影響されやすい。

ユーザモデリング方式は、ユーザの利用状況に大きく依存することが問題である。また、ユーザの行動履歴を収集するが本質的に困難であるという問題もある。

2.3 データモデリング方式

My Yahoo! [6] は、Web ページを人手でカテゴリに分類し、ユーザにカテゴリを選ばせる。ユーザがキーワード検索を行うと、選ばれたカテゴリの内容だけを提供する。しかし、ユーザの嗜好はアンケートに基づいて獲得するので、ユーザにとって負担になる。また、ユーザが登録した嗜好を自分で変更しない限り状態が変わらない。

Google の Site-Flavored Google(beta)[1] も My Yahoo!と同じような手法をとっている。しかし、Googleでは、アンケートの代わりにURLによってカテゴリを選択することができる。

データモデリング方式は、情報源を人手でカテゴリ分類するため、高価である。また、ページの内容解析に基づく自動分類手法は計算量が膨大になる上、人手による分類に比べると精度が低い。

なお、協調フィルタリングは、ユーザの評価によって類似した情報を間接的にモデル化することができる点ではデータモデリング方式と見なすこともできる。しかし、データの内容解析は行わないため、ユーザモデリング方式として扱った。

3. 本研究で提案する個人化手法

3.1 概要

本研究は、ユーザモデリング方式とデータモデリング方式を両方利用する。ユーザが指定した URL を利用してユーザモデルを作り、Web コミュニティによってデータモデルを作る。

Web コミュニティは、Web 上のページ間のリンク構造に従いページ間の関連が計算され、集められた URL の集合である。同じ Web コミュニティに属する URL の間には何らかの共通点がある。また、ページの内容解析を行わないので、計算コストが低いという利点がある。

ユーザが示した URL はユーザの嗜好を反映しており、その URL が属する Web コミュニティを発見すれば、その Web コミュニティに属する他の URL もユーザにとって有益な情報であるという仮説を立てる。

本手法は、提示する URL は一つしかない場合でも、ユーザモデルを構築することができる。

また、ユーザが明示的に嗜好を示すので、閲覧時間のような利用状況による嗜好判断の影響を改善できる。さらに、同じユーザでも、視点（仕事や趣味など）によって、ユーザプロファイルを容易に変えることができる（例えば、仕事と趣味で異なるホームページを持つユーザがいる）。

また、自分のホームページがなければ、ブックマーク中の URL や、自分と嗜好が似ている他人のホームページを指定することもできる。

図 1 に本システムの概要を示す。a、b、c、d はユーザプロファイルを構築する事前処理である（詳細は 3.2 を参照）。

[1] http://www.google.com/services/siteflavored.html

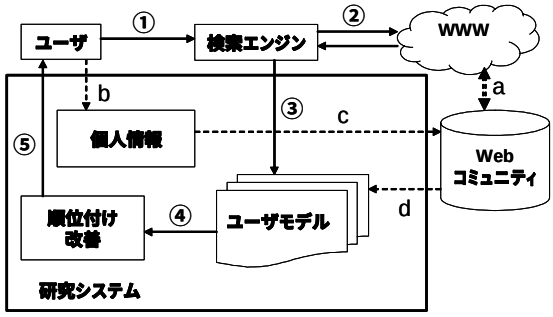

図１：システム概要図(破線は事前処理)

ユーザは検索キーワードを検索エンジンに入力する(①)。検索エンジンは WWW から検索結果を得る(②)。既存の検索システムでは、ここで結果をユーザに返す。しかし、本システムは、構築したユーザモデルと照合し(③)、順位付けを改善して(④)、検索結果を提示する(⑤)。

本システムの実現には、1）Web コミュニティの発見、2）ユーザモデルの構築、3）検索結果順位付けの改善が必要である。

Web コミュニティは、companion- [7]で構築されたデータを利用する。companion-は HITS [8]を利用して、Web ページ群から Web コミュニティを抽出する。得られた Web コミュニティでは、一つの URL は一つのコミュニティにしか属さない。以下、2）と 3）について 3.2 と 3.3 でそれぞれ説明する。

3.2 ユーザモデルの構築

図１の処理 b でユーザのブックマークと個人ページなどの個人情報から全ページの URL を取得する。

次に、処理 c で、上記の URL 群を Web コミュニティ中の URL と照合し、一致する URL が属する Web コミュニティを発見する。

処理 d では発見したコミュニティを一つのユーザモデルとして登録する。

ただし、Web コミュニティに存在しない「未知 URL」に対しては、内容解析に基づく類似度計算によって類似コミュニティを検索する。原理的には、ベクトル空間法などの最良一致（best match）検索手法ならば、どのような手法でも利用することができる。本研究では以下の手法を用いる。

1. 索引語の抽出

各コミュニティに属する全ての URL および未知 URL の HTML コンテンツを収集し、形態素解析を行い、索引語として名詞だけを対象に、そのコミュニティと未知 URL の索引語の集合を取得する。

2. 類似度の計算

上記の索引語の集合を用いて、未知 URL と各 Web コミュニティ間の類似度を式(1)の部分一致係

数（Overlap coefficient）[9]で計算し、類似コミュニティを検索する。

$$sim(U_i, W_j) = \frac{(|X_{U_i} \cap Y_{W_j}|)}{\min(|X_{U_i}|, |Y_{W_j}|)} \qquad (1)$$

X_{Ui} は未知URL iの HTML コンテンツの中から抽出した索引語の集合である。Y_{Wj} は Web コミュニティ j の索引語の集合である。

また、ユーザモデルを重要度によって順番を付与する。重要度は、そのプロファイルの中に含まれた個人情報中の URL の個数である。個数が多いほどそのモデルはユーザの嗜好を強く反映し、重要度が高い。

3.3 検索結果の順位付け改善

ユーザモデルを利用して、既存の検索エンジンから得られた検索結果に対し、以下の手順で順位付けを改善する。

1. 重要度に従いすべてのユーザモデルを順位付けする。
2. 検索結果の中に、ユーザモデルと一致する URL があれば、一致する URL を全て抽出する。一致する URL がなければ、元の検索結果をそのままユーザに提示し、処理を終了する。
3. 抽出された URL は属するモデルごとに分類する。そして、各モデルの順位関係に従い、分類した URL 群を表示の優先順位に並べ替える。
4. 同じモデルに属する URL 間の順位は元検索結果の順位に従う。
5. 3.と 4.に従い、一致する全 URL を順位付けて、ユーザに提示する。

4. 評価実験

4.1 実験データ

本実験では、検索システムとして Goo を利用した。使用した Web コミュニティは 2001 年に収集した日本語の Web ページから構築された。コミュニティ数は 674,491、含まれる URL 数は 1,452,394 である。

実際には、計算コストの都合上、674,491 組からランダムに抽出した 6,714 組と、その中に含まれる 33,484 の URL を使用した。

被験者は情報学を専攻する大学院生 10 人（日本人 7 名、外国人 3 名）である。図１の「個人情報」として各被験者のブックマークに含まれる URL を利用した。

4.2 評価基準

以下の 2 点を測定した。

(a) 同じ検索キーワードに対し、異なるユーザに異なる検索結果を提示できるかどうか
(b) 提示した検索結果の適合性

(a)は、被験者の検索結果と元の検索結果を比較し確認した。

(b)の適合性を測るために本システムと Goo のそれぞれの上位 10 位までの検索結果だけを対象として精度を求めた。

具体的には、本システムから得られた上位１０位までの URL と Goo の検索結果上位 10 位の URL を統合し（重複 URL を一つにする）、被験者に適合と思われる結果を選ばせ、各自の精度を計算した。ただし、適合するかどうかは被験者の判断に任せた。

4.3 結果と考察

被験者 10 人から得られた回答数は 74 件（順位付けの変化がない結果は含まれてない）の中に、評価の対象として、順位付けの改善を見られた 41 件があった（順位変動したのはすべてユーザのブックマーク中の URL の場合は除いた）。被覆率は約 55.4%であった。

また、評価対象になる結果は一つも得られなかったのは三人だった。その理由は二つある。使用した Web コミュニティには既に存在しない URL がある。このため、検索結果に現れなく、順位付けの改善はできない。もう一つは、この三人のブックマークから生成したユーザプロファイルの中の URL 数は他の方と比べると、少なかったためである。

表 1 は被験者 B と F が「情報検索」について検索した上位 10 位までの結果を示している。各部分の左側の数字は、その結果が Goo の検索結果の中での順位である。被験者 B の場合はほとんど書籍に関連するサイトであるに対し、被験者 F の場合は地図や、線路や、乗り換えに関するサイトであった。つまり、ユーザプロファイルの違いによって、得られた検索結果も異なる。

各被験者の検索結果の精度は、表 2 で示したように、各被験者の精度には少しばらつきがある。A のように、変わらない被験者もいれば、F のように、精度が結構良くなった被験者もいた。しかし、平均では Goo より 10 ポイント精度が向上し、本システムの有効性を検証することができた。

ただ、精度の絶対値はまだ低い。これは次のようなことと考えられる。まず、今回の Web コミュニティのデータが少し古く、より最新のデータでもう一

表1 同じ検索キーワードに異なる利用者による異なる検索結果の例(検索キーワード: 情報検索)

(1) User B		(2) goo		(3) User F	
310	http://ocl.city.okayama.okayama.jp/	1	http://www.jicoo.co.jp/	353	http://www.coneco.net/
745	http://www.nikkei.co.jp/	2	http://www.hellowork.go.jp/	250	http://www.bestgate.net/
83	http://webopac2.ndl.go.jp/	3	http://www.hellowork.go.jp/kensaku/servlet/kensaku?pageid=001	185	http://www.do-map.net/
836	http://bookweb.kinokuniya.co.jp/	4	https://search.npb.go.jp/	43	http://www.chizumaru.com/
71	http://www.trc.co.jp/trc-japa/index.asp	5	http://webfront2.nii.ac.jp/	277	http://www.e-hon.ne.jp/bec/EB/Top
634	http://www.kosho.or.jp/	6	http://clearing.rsa.go.jp/kasmkin/index.pnp	94	http://www.ipdl.jpo.go.jp/homepg.ipdl
137	http://www.ndl.go.jp/	7	http://www.tele.soumu.go.jp/j/musen/	20	http://itp.ne.jp/
49	http://ss.cc.affrc.go.jp/ric/opac/opac.html	8	http://search.sme.ne.jp/	15	http://transit.yahoo.co.jp/
27	http://www.books.or.jp/	9	http://www.mhlw.go.jp/search/	14	http://www.mapion.co.jp/
46	http://webcat.nii.ac.jp/	10	http://oppo.jp/	453	http://www.jorudan.co.jp/

表2 被験者ごとの検索精度

被験者	Goo	本手法
A	40.0%	40.0%
B	55.0%	61.7%
C	52.7%	62.7%
D	23.3%	27.8%
E	40.0%	40.0%
F	36.7%	62.2%
G	45.0%	55.0%
平均	**41.5%**	**52.2%**

度検証すべきである。次は、ユーザモデルを構築するとき、コミュニティに存在しなかった URL に対して、類似コミュニティの検索を行う際に生じた誤差が響いた。

5. おわりに

従来の検索結果の個人化に関する研究では、ユーザの嗜好と行動傾向の分析対象は主に個人情報から得られたページの内容だけである。

本研究では、情報検索結果の順位付けを改善するために、ユーザのブックマークを基に、Web コミュニティを利用して、ユーザモデルを構築するシステムを提案した。実際の検索システムを利用し評価した結果、平均精度は既存の検索エンジンより高かった。

しかし、精度の絶対値は未だ低い。この原因の一つはユーザのブックマークのみの利用によるものと考えられる。このため、今後はユーザの様々な個人情報を取り入れて評価することが必要になる。

謝辞

Web コミュニティデータを提供して頂いた喜連川優先生と豊田正史先生（東京大学生産技術研究所）に深謝致します。

参考文献

[1] R.Kumar, P.Raghavan, S.Rajagopalan, A. Tomkins, “Trawling the Web for Emerging Cyber-Communities”, In Proc. 8th WWW conference, 1999.

[2] Corin R. Anderson, Eric Horvitz, “Web Montage: A Dynamic Personalized Start Page”, Appearing in Proceedings of the 11th World Wide Web Conference (WWW’02), 2002.

[3] M.Claypool, A.Gokhale, T.Miranda, P. Murnikov, D.Netes, M.Sartin, “Combining Content-Based and Collaborative Filters in an Online”, In Proc. ACM SIGIR ’99 Workshop on Recommender Systems: Algorithms and Evaluation, Berkeley, California, 1999.

[4] G.Linden, B.Smith, J,York, “Amazon.com Recommendations: Item-to-Item Collaborative Filtering”, Proc. IEEE Internet Computing, 2003.

[5] K.Sugiyama, K.Hatano, M.Yoshikawa, “Adaptive Web Search Based on User Profile Constructed without any Effort from Users”, Proc. 13th International World Wide Web Conference (WWW’04), pp.675-684, 2004.

[6] U.Manber, A.Patel, J.Robison, “Experience with Personalization on Yahoo!”, Commun. ACM, Vol.43, No.8, pp.35-39, 2000.

[7] M.Toyoda, M.Kitsuregawa, “Creating a Web Community Chart for Navigating Related Communities”, In Proc. Hypertext 2001, pp.103-112, 2001.

[8] J.Kleinberg, “Authoritative Sources in a Hyperlinked Environment”, In Proc. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[9] P.Resnick, N.Iacovou, M.Suchak, J.Riedl, P.Bergstorm, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews”, Proc. ACM 1994 Conference on Computer Supported Cooperative Work(CSCW ’94), pp.175-186 1994.