研究指導 中澤 真 助教授

ブックマーク情報を利用した
個人の嗜好に対応する検索システム

佐藤 来未

1. はじめに

豊富な情報量を誇るインターネットの浸透は，人々の情報検索・収集の手段を大きく変えつつある．その影響もあり，2004 年のWebページ数は 150 億ページという膨大な数を記録し，2006 年の時点で 330 億以上のWebページが存在していると推測されている[1]．このことは，情報を探し求めているユーザにとって手に入る情報量が増えるという利点があるが，同時に欲しい情報が見つかりにくくなるという欠点もある．この問題に対し，Google社はユーザによって異なる検索結果を出力する新サービス「Googleパーソナライズ[1]検索」を開始した．これはユーザの閲覧履歴を利用してユーザプロファイルを生成し，その情報を基にユーザにとって有用であるだろうWebページを推薦するサービスである．しかし，閲覧履歴が蓄積されるまで検索結果に反映されず，ユーザにとって興味の無いWebページが閲覧履歴として残り，ユーザプロファイルに組み込まれる可能性がある．そうなると，ユーザにとって興味のないWebページが検索結果に反映されてしまい，システムの意味を成さなくなってしまう．

よって本研究では，ユーザにとって興味のない嗜好情報を反映させないことで精度を高めたユーザプロファイル生成方法および，生成したユーザプロファイルを情報検索に対応させるモデルを提案する．なお，ユーザの強い嗜好を示している情報として，個人の記録するブックマークを利用する．

2. パーソナライズ検索システムの現状と問題点

今までの検索エンジンは，どのようなユーザが検索しても出力される検索結果は同一であるのが一般的であった．Web上の情報量が増えてきていることも背景とされるが，これに対しGoogle社は 2005 年 11 月に日本語の「Googleパーソナライズ検索[2]」の提供を開始した．これは無料で提供されているサービスであり，Googleアカウントを取得すれば誰もが利用可能である．主な特徴は，過去の閲覧履歴からユーザの嗜好を学習し，そのユーザにとって重要なWebページを検索結果の上位に反映させるというものである．しかし，閲覧履歴が蓄積されて検索結果に反映されるまで時間がかかるということ，閲覧したWebページがユーザの興味のないものでも，それは閲覧履歴に記録され，検索結果に影響してしまうということが挙げられる．後者においての原因は次のような可能性が考えられる．

（ｉ）ユーザが誤って興味のない Web ページを閲覧してしまう．

（ｉｉ）閲覧してみたが，さほど興味のない内容だった．

このことから，閲覧履歴を用いる手法では，少なからずユーザの嗜好に合わない情報が組み込まれ，検索結果に反映される．それには，ユーザプロファイルの生成方法が重要となってくる．

2.1 ユーザプロファイリング技術

ユーザの嗜好に合った情報を提供するためには，対象となる情報に対してユーザがどの程度興味を持つか判断しなくてはならない．そのために，ユーザプロファイルを生成することが必要となる．このユーザプロファイルとは，ユーザの個人的な嗜好を表した情報のことを意味し，それを生成する技術をユーザプロファイリング技術と呼ぶ [2][3].

この技術には，明示的手法と暗黙的手法の 2 つが存在し，明示的手法は直接ユーザから興味に関する情報を入力してもらい，このデータに基づいてユーザの嗜好を抽出して，ユーザプロファイルを生成する．一方の暗黙的手法は，システムを利用するユーザの行為を用いて嗜好を判断し，それらの情報からユーザプロファイルを生成する．一般的に，明示的手法は暗黙的手法に比べてユーザにかかる負担が大きいという欠点がある．このため Web サービスなどにおけるユーザプロファイリング技術では暗黙的手法がしばし

---

1 全員に同じサービスを提供するのではなく，ユーザ個人の嗜好に合わせて最適化したものを提供する手法を指す．

2http://www.google.com/support/bin/answer.py?answer=26651&topic=9005

ば用いられ，ユーザの行動履歴としてWebサイトの閲覧履歴が利用される．次節では，閲覧履歴を用いたプロファイル生成手法について述べる．

2.2　閲覧履歴を用いたプロファイル生成手法

ユーザの行動履歴として，Web の閲覧履歴を用いたユーザプロファイルを生成する手法が研究されている[4][5]．これは「Googleパーソナライズ検索」にも使われている手法であり，ユーザの Web ページ閲覧行動に応じて，該当するページの特徴的なキーワードをシステムが自動的に抽出する．最終的に，抽出されたキーワードをユーザの特徴として，個々のユーザプロファイルが生成される．

2.3　閲覧履歴に基づく特徴抽出

ユーザの嗜好を抽出するためには閲覧したページの特徴となっているキーワードを抽出する必要があり，そのために*tf・idf* 法[6][7]が用いられている．これはキーワードの出現頻度から重み付けする手法で，*tf* は対象となるWebページにおける特定キーワードの出現頻度を，*idf*はWebページ集合の中で特定キーワードが含まれているWebページの数を指す．もし，あるキーワードが一つのWebページ内に頻出するものであり，そのキーワードが出現するWebページ数がわずかな場合であれば，それは*tf・idf* 値が高くWebページの特徴的なキーワードと言える．この手法を用いた特徴抽出について次式を定義とする．なお，ここではWebページ*hp*におけるキーワード*k*の*tf・idf* 値を$w^{hp}(k)$として定義する．

$$w^{hp}(k) = tf(k) \cdot idf(k) \qquad (1)$$

$tf(k)$ = 閲覧したWebページ*hp*におけるキーワード*k* の出現回数

$$idf(k) = \frac{\text{全てのWebページ数}}{\text{キーワード}k\text{ が出現するWebページ数}}$$

ここで，*tf・idf* 値を求めて算出したWebページの特徴となるキーワードを，ベクトル空間モデル[3]を用いて表す[8][9]．つまり特徴をベクトル化することで，その向きや長さで特徴の度合いを測ることができ，これを特徴ベクトルと呼ぶ．あるWebページ*hp*の特徴ベクトル$\boldsymbol{w}^{hp}$は次式のように定義される．

$$\boldsymbol{w}^{hp} = (w^{hp}(k_1), w^{hp}(k_2), \cdots, w^{hp}(k_m)) \qquad (2)$$

k_i = 式(1)より抽出された各キーワード($i = 1, 2, \cdots, m$)

m = Webページ*hp*におけるキーワードの総数

この特徴ベクトルをユーザプロファイルとして用いる．しかし，閲覧したページがユーザの嗜好に合っていなかった場合，ユーザの嗜好と異なる情報も共に抽出され，ユーザの嗜好が適切には反映されない．また，閲読時間を設けてもユーザによって閲覧するスピードが異なるため，この手法は精度の高いユーザプロファイル生成手法とは言えない．そこで次章において，ブックマークを用いることで，ユーザの嗜好だけを抽出した精度の高いユーザプロファイル生成方法，並びに検索システムとの組み合わせを提案する．

3.　ブックマークを用いた個人向け検索システム

情報検索に対応する本研究の提案モデルは，キーワード検索技術と情報フィルタリング技術[4]の両技術で構成されている．基本コンセプトは，検索結果であるWebページが，生成したユーザプロファイルに類似しているほどユーザが望む情報であると仮定して，そのWebページを上位に提示するというものである．例えば，ユーザがあるキーワードを入れて検索した時，Web上に存在する大量のWebページがヒットしたとする．そこで，情報フィルタリング技術と組み合わせることによって，ユーザの嗜好と似ているWebページを抽出し，必要な情報だけに絞り込む．この結果，ユーザは全てのページをチェックすることなく，有益なWebページを手早く得ることが出来る．

次の節で，閲覧履歴ではなくブックマーク情報を採用する利点を述べる．

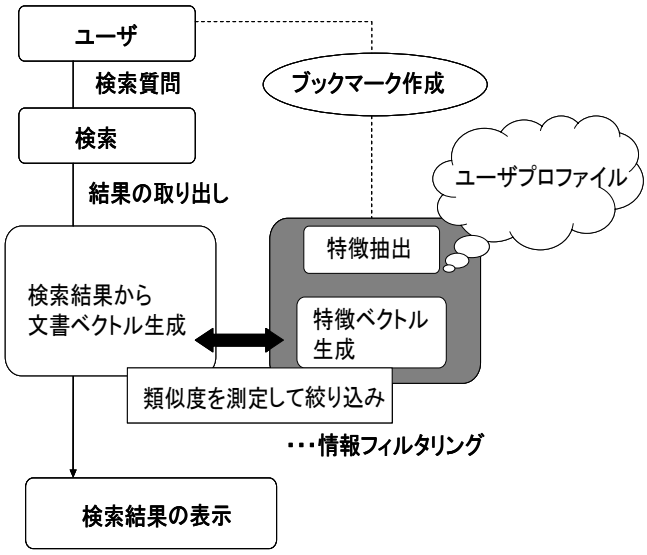

図1：提案する個人向け検索システムのモデル

3.1　検索技術におけるブックマークの利点

ブックマークを作成するユーザの意志とは，「今後もこの Web ページを閲覧したい」という強い興味の表れである．したがって，ブックマークはユーザの嗜好が的確に反映されたものだといえる．ユーザの関心が薄れ，嗜好が変わった場合には，ブックマークを削除するこ

3　情報検索分野で幅広く利用されている検索モデル．検索質問やWebページをベクトルで表現し，ベクトルの向きによって内容を判断する．

4　ユーザの嗜好に基づいて，関心のあるものを提示したり，優先度を与えたりして情報を選別する技術．

とで常にユーザの嗜好を反映するユーザプロファイルを維持することが可能になる. 以上の利点から, 本研究ではユーザプロファイルの生成にブックマークを利用する.

3.2　ブックマークを用いたプロファイル生成方法

3.2.1　特徴抽出

ブックマークからユーザを特徴付けている特定キーワードを抽出するために, 2.3 で述べた *tf・idf* 法を利用する. しかし, ブックマークは閲覧履歴と異なり, ユーザの嗜好が反映された結果として, 偏ったページの集合となっている可能性が高い. そのため, ユーザの選別したブックマークの中には, 内容の類似した Web ページが多数存在しても珍しくない.

このような場合, 特定のジャンルに関わるキーワードの頻出度が上がり, *tf* 値が上昇する. しかし, それらのページは似通っている内容のため, 特定キーワードはほとんどの Web ページに出現することになり, 逆に *idf* 値は減少する. これでは, そのキーワードに対して *tf・idf* 値が低く算出されてしまうため, ユーザの嗜好に合致しているはずのキーワードが抽出されないことになる.

そこで, ユーザのブックマークに存在するフォルダを利用し, フォルダ単位でブックマークの特徴抽出を行う. カテゴリ別に分類して似通ったページを一括りにすることで, *tf・idf* 値が偏ることのない特徴抽出を可能にする.

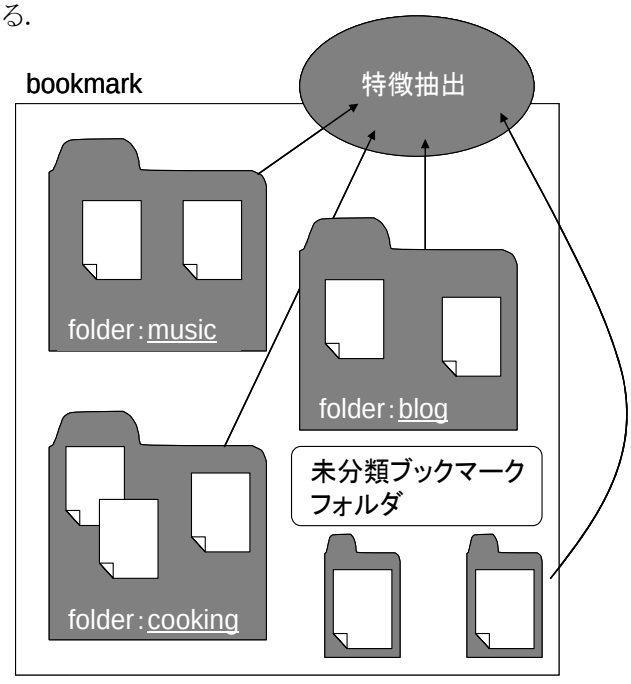

図 2:ブックマークフォルダを用いた特徴抽出手法

提案するシステムでは図 2 のようなモデルを想定し, *tf・idf* 値によるブックマークからユーザの特徴を示すキーワードを抽出する. まず, ブックマークフォルダ *j* におけるキーワード *k* の *tf・idf* 値 $bf_j(k)$ を次のように定義する.

$$bf_j(k) = tf_{bf}(k) \cdot idf(k) \qquad (3)$$

$tf_{bf}(k)$ = *BF* 内におけるキーワード *k* の出現回数

$$idf(k) = \frac{BF}{bf_j \text{ 内でキーワード} k \text{ が出現するWebページ数}}$$

BF = ブックマークフォルダ内の全 Web ページ数

tf・idf 値が高いキーワードほどそのブックマークフォルダの特徴となるキーワードと判断し, 抽出したキーワードの集合をユーザプロファイルとする. しかし, この手法はフォルダ単位で計算しているため, どのフォルダにも属さない Web ページがある場合や, フォルダを一つしか作成していないユーザが存在する場合には特徴抽出をしても効果が見られない可能性がある. そこで, フォルダに分けられていないブックマークを一つのフォルダと仮定し, 通常のフォルダと同様に *tf・idf* 値を算出する. つまり, これらのブックマークは Web ページ単体に対して *tf・idf* 値を求めていることになる. これにより, フォルダのカテゴリにとらわれない特徴抽出が可能となる.

このようにして, 各カテゴリのブックマークフォルダに *tf・idf* 値が最大のキーワードをひとつずつ抽出する. 次項において, この抽出したキーワードを用いた特徴ベクトル生成方法について述べる.

3.2.2　特徴ベクトル生成

検索結果におけるWebページと類似度を測るため, ユーザのブックマークから抽出した特徴をベクトルで表す. 前章で述べた特徴抽出方法を用いて, ブックマークフォルダの特徴となっているキーワードを抽出した. したがって, *tf・idf* 値の高かったブックマークフォルダ $bf_j(k)$ を対象文書として特徴ベクトル $\boldsymbol{f}^{BF}$ を生成する. ブックマークフォルダの集合 *BF* における特徴ベクトル $\boldsymbol{f}^{BF}$ は次式のように示す.

$$\boldsymbol{f}^{BF} = \left(bf_1(k_1^*), bf_2(k_2^*), \cdots, bf_m(k_m^*) \right) \qquad (4)$$

k_i^* = 抽出されたキーワード($i = 1,2, \cdots, m$)

m = ブックマークフォルダの総数

このようにして生成した特徴ベクトルは, ユーザがキーワード検索をした際に出力された Web ページと, どれくらい似ているのか判断するのに利用される. そのためには比較対象である Web ページもベクトルで表現する必要があり, 次にそれらのベクトル生成方法を述べる.

4.　文書ベクトル生成

文書ベクトルとは, キーワードの出現頻度に基づいて, 検索結果である Web ページを一つずつベクトルで表現したものである. 本システムでは, 前述した特徴ベク

トルと文書ベクトル間の類似度を測り，検索結果を出力する．

次式で文書doc_h（$1 \leq h \leq l$）の文書ベクトル$\boldsymbol{d}_h$を次式で表す．

$$\boldsymbol{d}_h = (w_{h1}, w_{h2}, w_{h3}, \cdots, w_{hm}) \quad (5)$$

この文書ベクトルと先に述べた特徴ベクトル間の類似度を測り，ユーザの特徴ベクトルに対し類似度が高いと判定されたWebページを検索結果の上位に出力する．これは，検索対象であるWebページを表す文書ベクトルが，個々で生成された特徴ベクトルに近いほどユーザにとって興味のあるWebページであると判断した．類似度には，2つのベクトルのなす角度を用いたコサイン尺度を用いる[10]．ベクトル間の類似度を$\cos\theta$とした場合，特徴ベクトル$\boldsymbol{f}^{BF}$と文書ベクトル$\boldsymbol{d}_h$の類似度は次のようにして求める．この値は1に近いほど類似度が高いことを意味する．

$$\cos\theta(\boldsymbol{f}^{BF}, \boldsymbol{d}_h) = \frac{\boldsymbol{f}^{BF} \cdot \boldsymbol{d}_h}{|\boldsymbol{f}^{BF}||\boldsymbol{d}_h|} \quad (6)$$

このように，ベクトル空間モデルを利用してユーザの嗜好情報を表す特徴ベクトルと対象 Web ページを表す文書ベクトル間の類似度を算出させることで，類似度の高い Web ページから順に結果を表示する．

5.　むすび

本研究では，ブックマーク情報を用いたユーザプロファイル生成手法を提案した．この手法はブックマークの機能を普段から利用しているユーザにとって，ユーザの負担は皆無であり，ブックマークに多数の登録をしているユーザは，より精度の高い検索結果を得ることが可能となる．例えば，個々のユーザが「アップル」と検索した場合，パソコン関係の Web ページをブックマークしていたユーザなら「アップル社」関連の Web ページが上位に表示され，料理関連の Web ページをブックマークしていたユーザなら「アップル(りんご)」関連の Web ページが上位に表示される．

さらに，ブックマークフォルダによってカテゴリが構成されているため，ユーザの嗜好を従来の方法より正確に抽出することを可能にした．

しかし，ブックマークに登録される Web ページの増加量は閲覧履歴のページ数よりもかなり緩やかである．そのため，個人の嗜好が反映されるまでに時間を要するという問題もある．よって，今後はブックマーク量が増加するような機能の追加など，本システムの改善に向けて検討していきたい．

Web 上の情報量が年々増加していく現代において，ユーザ個人の嗜好に対応する検索システムは不可欠なものだと言える．今後，Google社が提供している「Googleパーソナライズ検索」のように，個人に対応するシステムが一般的なサービスとなるであろう．

参考文献

[1] http://rblog-media.japan.cnet.com/0058/2006/11/googlemapagearan_c27d.html CNET Japan, インターネットの理解, 2006. 11. 6.

[2] 土方嘉徳, ”情報推薦. 情報フィルタリングのためのユーザプロファイリング技術”, 人工知能学会, 19 巻 3 号, pp.365-372, 2004. 5.

[3] 堀 幸雄, 今井 慈郎, 中山 堯: “個別化する Web : ―検索キーワードの個別化―”, 情報知識学会誌, Vol. 16, No. 4, pp.4_33-4_40, 2006.

[4] 杉山一成・波田野賢治・吉川正俊・植村俊亮,”ユーザからの負担なく構築したプロファイルに基づく適応性 Web 情報検索”,電子情報通信学会, No.11 ,pp.975-99 ,2004. 4.

[5] 中村雄一,菊池浩明,”閲覧履歴に基づく情報検索の相互支援”, 情報処理学会研究報告. DPS,マルチメディア通信と分散処理研究会報, 126, pp. 25-30, 20060316, (ISSN 09196072).

[6] http://www.ntts.co.jp/products/knowledgeocean/detail/qanda.html NTT ソフトウェア.

[7] http://nlp.nagaokaut.ac.jp/~sekiguti/doc/pdf/tfidf/tfidf/ TFIDF.

[8] 福島 俊一. “Web サーチエンジンの基本技術と最新動向(上)基本技術”. 情報管理. Vol. 46, No. 6, (2003), 363-372.

[9] 大谷紀子,”情報検索におけるベクトル空間モデルの応用”,武蔵野工業大学環境情報学部研究論文 3-6.

[10] 小松香爾, ”情報システムにおける言語処理技術の利用”, 経営論集, pp.105-114, vol.13, No.1,2003. 12.