Webページの閲覧履歴とブックマーク情報を用いたユーザの興味の獲得*

和田 洋祐† 相場 亮‡
芝浦工業大学 システム工学部 電子情報システム学科§

1 はじめに

ブログなどのツールの発達やニュースサイトの活発化により Web 文書は急速に増加し，今後も増加を続けると考えられている．Web 文書が増加し種類が多様化することで，閲覧したが予想していた内容と異なる情報であったり，多くの情報に埋もれ見逃してしまう情報が出てくる．そのためユーザにとって興味関心ある有益な情報を手に入れる労力は大きなものとなっている．そうした背景からユーザの興味関心を基準としてユーザに合った情報を推薦する仕組みが注目されている．例として Google news ではユーザが指定したキーワードとサイト内から選択したニュースの履歴を用いることで興味に合わせて情報推薦を行っている．しかし，ユーザがキーワードを設定しなければならない，サイト外で閲覧した情報は反映されないなどの問題もある．

本研究ではこれらの問題点を考慮し，プロキシサーバに記録された Web 文書の閲覧履歴とユーザのブックマークを用いることでユーザに負担をかけることなく興味に合わせた情報推薦を行う．

2 先行研究

興味の獲得にはいくつかの手法が提案されている [1]．履歴を用いた興味獲得では，ユーザの閲覧した Web 文書の中に興味があると考え，閲覧ページからキーワードを抽出することでユーザが興味を持つ話題を獲得していく．キーワードの抽出には形態素解析などが用いられ，抽出されたキーワードを特徴ベクトルとし，特徴を統計的に処理していくことで，ユーザの興味情報を獲得する手法が用いられている．

閲覧した Web 文書すべてを平等に扱う場合，閲覧してみたが興味が持てなかった情報に含まれるキーワードまで興味として扱ってしまう問題もある．この問題に対して，ユーザが興味関心を持っていそうな文書やその一部を推定する手法が用いられることがある．これには Web 文書の閲覧時間の長さ [2] や視線の観測による単語抽出 [3] などがあるが，前者は人により文書の読む速度が異なることやブラウザを開いたまま席をはずした場合に正しい推定ができなくなるし，後者の場合には視線を観測するデバイスが必要となるなどの問題点がある．

3 提案手法

興味位置の推定にブックマークを用いるのは，ブックマークがユーザの手によって興味がある Web 文書につけるものであり，興味の対象を明確に示すものだからである．また，ブックマークするという行為は，興味獲得のための特別な作業ではなく，ユーザに特別な労力をかけないという利点も持っている．

興味獲得の流れを図 1 に示す．

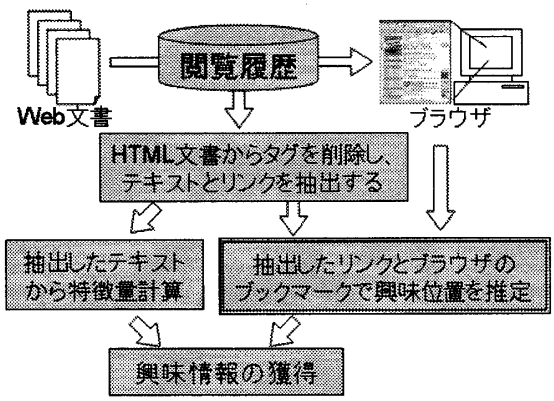

図 1: 興味情報獲得までの流れ

3.1 文書ベクトルの計算

履歴に残された Web 文書集合を D とする ($|D| = N$)．集合 D から形態素解析で名詞，動詞，未知語のいずれかである M 種類の単語を抽出したとき，Web 文書 d_j の特徴ベクトルを以下で定義できる．

$$d_j = \{w_1^j, w_2^j, \ldots, w_t^j \ldots, w_M^j\} (j = 1, 2, \ldots, N) \quad (1)$$

w_t^j は文書 d_j 中の単語 t の重みで tf·idf 法により求める．

---

* Acquisition of user interests on Web page by browsing history and bookmarks
† Yosuke Wada
‡ Akira Aiba
§ Department of Electronic Information Systems, College of System Engeineering, Shibaura Institute of Technology

3.2 ブックマークを用いた興味位置の推定

興味位置の推定にはWeb文書とブックマークとのリンク距離lを用いる．距離lはブックマークされているWeb文書のリンクをたどって何回で対象のWeb文書にたどり着けるかを示すものである．図2のAとBの距離を求める．Aを起点($l=0$)としユーザは4回目のリンクでBにたどり着いたがAからBへユーザがたどらなかったリンクが存在する場合には距離$l=1$となる．

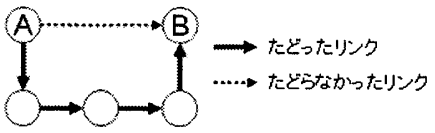

図 2: ユーザのアクセス履歴とリンク

Web文書は多くの場合，他の文書へのリンクを複数持っている．そのためリンクをたどることは複数のリンクの中から一つを選択することでもあり，リンク先の文書にはユーザの興味関心のある情報が掲載されている可能性が高い．この考えを元に文書ごとの重みb_jを以下で定義する．

$$b_j = \begin{cases} 3 & l=0 \\ 1.5 & l=1 \\ \frac{1}{l} & \text{それ以外} \end{cases} \quad (2)$$

3.3 興味ベクトルの作成

ユーザの興味情報をベクトルで表現する．このベクトルを興味ベクトルと呼ぶ．興味ベクトルは閲覧履歴から求めた各文書の文書ベクトルと興味推定における重み付けから以下のように求められる．

$$I = \{w_1, w_2, \ldots, w_i, \ldots, w_M\} \quad (3)$$

$$w_i = \sum_{j=1}^{N} b_j w_i^j \quad (4)$$

3.4 興味ベクトルを用いた情報分類

Web文書d_xに対し，ユーザが興味を持つか否かを興味ベクトルを用いて判定する．Web文書d_xの特徴ベクトルをtf-idf法で計算すると

$$d_x = \{w_1^x, w_2^x, \ldots, w_M^x, w_{M+1}^x, \ldots, w_K^x\} \quad (5)$$

ここで$\{w_{M+1}^x, \ldots, w_K^x\}$はWeb文書の集合$D$になかった種類の単語である．これらの単語の重みは$idf(t) = \log N$として求める．興味ベクトルの方での重みは0として扱う．この様にして求められた興味ベクトルIとd_xの類似性を両ベクトルの余弦を計算することで求めることができる．

$$cos(I, d_x) = \frac{I \cdot d_x}{\|I\|\|d_x\|} \quad (6)$$

この値は[0,1]の範囲の値を取り，1に近いほど類似性が高いと言える．興味の有無は閾値sを用いて，sよりも大きい値の時，興味があると判定できる．

4 評価

ユーザの履歴とブックマークから抽出した興味情報に基づいた情報推薦が最終的目的であるため，ユーザの興味を実際に抽出し，その情報による興味有無の判定とユーザ自身が判定した興味有無を比較することで提案手法の評価実験を行う．

興味判定の対象として二種類のデータを用意する．一つ目はユーザの履歴にある閲覧時間が新しい20件の興味あり文書Y．二つ目はインターネット上のニュースサイトから取得した80件ニュース記事Z．この二つを全集合Uとし，Uを対象とした興味有無の判定を行うことでYとZの分類に関する精度を求める．評価基準には適合率，再現率，F値を用い，ブックマークによる重み付けの有無による違いを調べる．

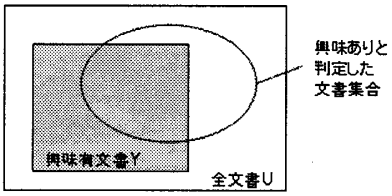

図 3: 閾値 s の時の判定結果

5 おわりに

本稿ではブックマークを興味位置の推定に用いた興味情報の獲得方法とその評価を提案した．今後8名の協力者を対象にした評価実験を行う予定である．

参考文献

[1] 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, 人工知能学会誌 Vol.19 No.3 Page.365-372 (2004.05.01)

[2] 杉山一成，波多野賢治，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的Web情報検索，電子情報通信学会論文誌 D-1 Vol.J87-D-1, No.11, Page975-990 2004.11.01

[3] 大野健彦, IMPACT : 視線情報の再利用に基づくブラウジング支援法, Proc. of the 8th Workshop on Interactive System and Software, 2000