6G-7

自己組織化マップを用いたユーザからの興味パターンの抽出

堀 幸雄 † 安達 章 ‡ 水田 明宏 ‡ 中山 堯 §
香川大学 情報基盤センター † 香川大学 工学部 ‡ 神奈川大学 理学部 §

1. はじめに

情報システムを利用するユーザの持つ興味関心は重要であり，情報フィルタリング [6]，ユーザ同士のコミュニティ分析 [3] など様々の観点から研究が行なわれている．この中で特にユーザの持つ興味関心をモデル化することが重要となる．しかしながらこれら研究では blog や BBS などある特定のサービスをもとにユーザプロファイリングを行なっており，これでは該当サービス上での興味関心は把握できるもののサービス外を含めユーザの大域的な行動を分析していないため，ユーザの興味関心を正しく把握できていないのが現状である．またユーザの興味関心の評価にユーザの興味に適合する情報の適合率，再現率を用いている．このような情報検索システムの評価手法を用いた場合，ユーザ，文書といった自由度があり，必ずしもプロファイリングされたユーザの持つ興味関心が評価されているわけではない．

本研究ではユーザの行動パターンが網羅的に記録されたアクセスログからユーザの興味パターンを抽出し，自己組織化マップによりマップ上にモデル化する手法を提案する．そして得られた興味パターンマップがどの程度ユーザの興味関心を現しているのかを，ユーザの検索エンジンに投入するキーワードを用いて評価した．その結果について報告する．

2. 関連研究

ユーザの興味関心に関する情報を分析する研究として，KANSHIN [1] や blogWatcher[2] などがある．これらは blog サイトから RSS を定期的に収集し，記事の解析を行い，興味関心パターンの分析を行なっている．しかしながらこれら研究では blog や BBS などある特定のサービスを用いたものであり，これでは該当サービス上での興味関心は把握できるものの，サービス外を含めユーザの大域的な行動を分析していないため，ユーザの興味関心を正しく把握できていないのが現状である．

また Web サーバ，プロキシなどのログを用いて網羅的にユーザの閲覧した情報を分析する研究もある．橘高らの研究 [5] では閲覧したページを興味のあるページと仮定し，閲覧したページをもとにユーザの興味をモデル化している．しかしこの研究ではモデル化されたユーザの興味関心の評価に，ユーザに適合する情報かという観点において，適合率，再現率が用いられている．これではユーザの興味関心が直接的に評価されたわけではなく，適合する文書の自由度を含めて評価される．

Caputuring user interests via SOM and its quantitative evaluation

†Yukio HORI, Kagawa University, horiyuki@cc.kagawa-u.ac.jp
‡Akira ANDATSU,Aihiro MIZUTA,Kagawa University
§Takashi NAKAYAMA, Kanagawa University

本稿ではユーザの閲覧行動パターンが網羅的に記録された Web アクセスログからユーザの興味パターンを抽出し，自己組織化マップによりマップ上にモデル化する．ユーザの興味関心を 2 次元平面上に射影することで見た目にもわかりやすく抽出することが可能となる．そして得られた興味パターンマップがどの程度ユーザの興味関心を現しているのかを，ユーザが検索エンジンに投入したキーワードを用いて評価する方法を提案する．

3. 興味パターンの抽出手法

本研究での興味パターンの抽出方法を述べる．まず各ユーザの興味パターンを抽出するまでの流れを図 1 に示す．

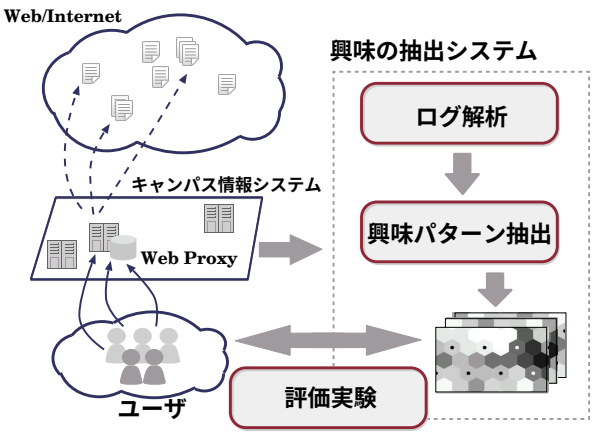

図 1: 興味パターン抽出システムの概要

はじめに各ユーザの Web 閲覧履歴が保存されているログを用いて，各ユーザが閲覧した Web ページ集合 $U_i = (w_1, w_2, \ldots, w_n)$ を求める．$w_j (j = 1 \sim n)$ は閲覧した Web ページの特徴ベクトルとし，その作成方法は次の通りである．

1. HTML から不要なタグやスクリプト，ヘッダ部分を削除する．
2. 残ったテキストを mecab[1] を用いて単語に分割する．その際名詞や動詞以外は削除する．
3. ストップワード処理として，ひらがな，カタカナ 1 文字の単語は形態素解析に失敗している可能性が高いために除外し，全体で出現頻度に閾値を設定する．
4. 次に各単語に重みを付けるために各単語の tf-idf 値を次式により計算する．

$$w(t,d) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t)$$

[1]http://mecab.sourceforge.jp/

$$\mathrm{idf}(t) = \log(N/\mathrm{df}(t))$$

ただし，$\mathrm{tf}(t,d)$ は Web サイト d における単語 t の出現頻度であり，N はユーザの閲覧した総 Web サイト数，$\mathrm{df}(t)$ は単語 t が 1 回以上出現する Web サイト数である．

3.1 自己組織化マップを用いた興味マップの作成

各ユーザが閲覧した Web ページベクトル U_i は多次元ベクトル集合として現わされる．このままではユーザの興味関心の直接的把握が困難なため，適切なクラスタリングを行ない，次元圧縮する必要がある．ここで自己組織化マップ [4] を用いてユーザの閲覧した Web ページベクトル $U_i = (w_1, w_2, \ldots, w_n)$ を入力として相互の距離関係を可能な限り保持した状態でこの特徴ベクトルを 2 次元平面上に写像する．これにより 2 次元平面に射影されたユーザの興味関心を参照することが可能となる．こうして得られたユーザの興味パターンマップが図 2 である．なおこのユーザは「名古屋 観光」,「名古屋駅 アクセス」,「万博 EXPO シャトル」といった検索キーワードを投入しており，それに関連した「愛 地球」,「ホテル 愛知」,「JR きっぷ」などの単語がマップ上にも出現していることがわかる．

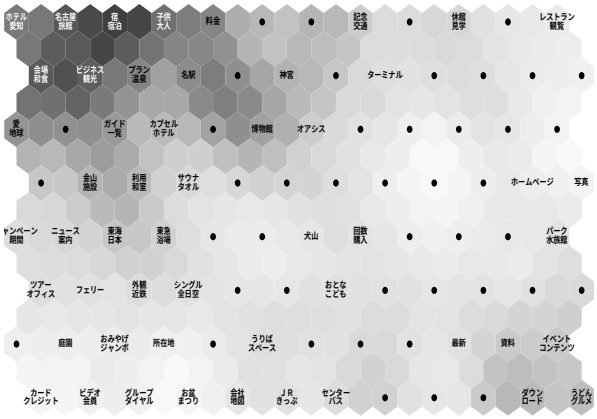

図 2: 得られたユーザの興味パターンマップ

4. 評価実験

ここでは得られた興味パターンマップがどの程度ユーザの興味を現しているのかを評価する．まず各ユーザの興味パターンマップを抽出した．抽出したデータは表 1 の通りである．

表 1: 実験データ

実験期間:	2005 年 8 ～ 9 月
興味パターン抽出タイミング:	1 日単位
総ユーザ数:	2,719 人
ユーザ当たりの平均 PV:	314

得られた興味パターンマップがどの程度ユーザの興味関心を現しているのかという評価方法は次の手順により行なう．

1. 興味パターン抽出日時のユーザの検索クエリ[2] K_i を保存する．
2. 検索クエリを抽出した興味マップ上に当てはめ，該当する単語を W_k とする．
3. シソーラス [8] を用いて K_i と W_k の類似度 S を求める．類似度の計算方法は次の式を用いる [7]．ここで $\overrightarrow{K_i}, \overrightarrow{W_k}$ は各単語のシソーラスにおけるカテゴリ情報をベクトル化したものである．

$$S = \frac{\overrightarrow{K_i} \cdot \overrightarrow{W_k}}{|\overrightarrow{K_i}||\overrightarrow{W_k}|}$$

類似度 S は 0 ～ 1 の間で現される．しかし検索キーワードがマップ上に存在していない可能性もあるため，再現率 R = マップ上に当てはまる単語がある場合/総検索キーワード数とし，類似度，再現率両方を用いてユーザの興味関心が正しく得られたのかを評価する．その評価結果が表 2 である．

表 2: 評価実験結果

総検索キーワード数:	102,034
類似度 S:	0.77
再現率 R:	0.81

5. おわりに

本稿ではアクセスログを用いて，ユーザの大域的な Web 閲覧行動から自己組織化マップを用いて興味パターンマップを抽出した．そしてこの興味パターンがユーザの興味関心を現しているのかをユーザが検索エンジンに投入したキーワードを用いて評価した．その結果見た目にも分かりやすく，ユーザの興味関心とも近い興味パターンマップが取り出せることを示した．

今後は抽出したユーザの興味パターンの遷移や他者との違いの表現方法をどのようにするのかといった問題が残されている．また得られた興味パターンを他のシステムと連携して応用することも今後の課題である．

参考文献

[1] 福原知宏，村山敏泰，中川裕志，西田豊明: ウェブログ記事を用いた関心解析システム，人工知能学会 第 19 回全国大会, 2C2-04, 2005.

[2] Nanno, T., Suzuki, Y., Fujiki, T., and Okumura, M.: Automatic collection and monitoring of Japanese Weblogs., *WWW 2004 Workshop on the Weblogging Ecosystem*, 2004.

[3] 谷口 智哉，松尾 豊，石塚 満，Blog コミュニティの抽出と分析，人工知能学会，第 6 回セマンティックウェブとオントロジー研究会，SIG-SWO-A401-08, 2004.

[4] Kohonen, T.: *Self-Organizing Maps, 3rd Edition*,Springer-Verlag, 2001.

[5] 橘高博行, 佐藤直之, 鈴木英明, 曽根岡昭直: パーソナライズ情報提供方式の提案と評価，情報処理学会論文誌，Vol.40,No.1,pp.175-187, 1999.

[6] 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, pp.365-372, 2004.

[7] 川島 貴広, 石川 勉: 言葉の意味の類似性判別能力に関するシソーラスと概念ベースの性能比較, 人工知能学会全国大会 2D2-10,2004.

[8] 池原 悟，他: 日本語語彙体系，岩波書店, 1997.

[2] Google, Yahoo!, MSN など代表的な検索エンジンに投入したキーワードをアクセスログから抽出する