ブログ検索支援のためのユーザが興味を自己認識するきっかけとなるキーワード抽出法

狩野 憲和　倉本 到　渋谷 雄　辻野 嘉宏
京都工芸繊維大学

1.研究背景

近年，ブログを利用している人が増えている．ブログを読む際，ユーザは何か興味を引くエントリを読みたいが，明確に読みたいものが決まっていないという場合がある．このような場合，新着表示や話題性のあるキーワードを使用した検索が利用される．しかし，これらの方法は，個々のユーザの興味を考慮していないため，興味に合致したエントリを見つけることは難しい．そこで，個々のユーザの興味を抽出することが考えられるが，このような抽出の方法には，明示的な方法と暗黙的な方法がある．

明示的な方法とは，ユーザに対するアンケートにより興味を抽出したり，エントリに対する興味の有無をユーザ自身が判定し興味を抽出したりする方法である．しかし，これらの行為はユーザにとって負担である．

暗黙的な方法とは，ユーザが閲覧したWebページの文章を基に，ユーザの興味に関する情報を自動的に取得する方法である[1]．しかし，このような方法は，ユーザが明示的に興味の対象を示す場合と比較して，興味の判定精度が低下する．そこで，マウスの挙動や視線を利用して，自動的に興味のある部分を特定する研究もなされている[2][3]．

本研究では，ユーザの興味がありそうなキーワードをユーザ自身の書いたブログから自動的に収集，選別，提示し，そこからユーザに明示的に興味あるキーワードを選択させ，それをもとに興味を引くエントリを検索する手法を提案する．この手法では，ユーザが評価するのは数個のキーワードだけでよく負担が少ない．また，ユーザの判断が入ることで興味あるキーワードの自動抽出の精度低下を抑えることが期待できる．

2.提案手法

本手法では以下の手順でユーザの興味あるエントリを検索する．各段階の詳細は2.1節以降で説明する．

1. ユーザが興味を持ちそうなキーワードをシステムが自動的に抽出．
2. 抽出したキーワードのリストをユーザに提示．
3. ユーザは，リストの中から興味のあるキーワードを1つ選択．システムは，その選択されたキーワードに関連するキーワードを抽出．
4. システムはユーザが選択したキーワードを用いて対象となるブログエントリ群内を検索し検索結果を提示．また，この提示と同時に，選択したキーワードと関連する3.で求めたキーワードを絞込み用キーワードとして提示．
5. ユーザは，絞込みリストから任意の数のキーワードを選択．
6. システムは，選択された絞込み用のキーワードを追加して検索を行い，検索結果を提示．

このように検索キーワードの選択を2段階にすることで，一度に多量のキーワードについて興味の有無を評価する必要がない．また，1つのキーワードを用いたときの検索結果よりも，複数のキーワードを使用することで，検索結果のエントリの中でユーザの興味を引くものが占める割合が高くなることが期待できる．

2.1 ユーザの興味を引くキーワードの抽出

抽出の対象として，ユーザの書いたブログエントリの文章を利用する．ユーザが「書く」という行為を行ったということは，少なからずそのブログで対象とした話題について興味があると考えられるからである．

また予備調査により，エントリの中で「タイトル」「小題」「段落のはじめの文」「その他」の順で興味を引くキーワードが出現しやすいことがわかった．

2.1.1 抽出するキーワードの品詞

興味を引くキーワードとして，そのキーワードだけで対象をイメージできるものであることが望ましい．そこで抽出するキーワードは，「固有名詞」「一般名詞」「サ変名詞」とした．茶筌[4]を用いた形態素解析により，これらのキーワードをユーザのブログエントリより抽出する．

2.1.2 長期的・短期的興味

ユーザの興味は短期的なものと，時間がたってもあまり変わらない長期的なものに分かれると考えられる．

短期的なものとは，ある短い期間に話題の対象となったものであり，あるエントリだけに頻出するキーワードがそれを表すと考えられる．長期的なものとはそのユーザのブログエントリ全体を通して頻繁に話題の対象となっているものであり，多くのエントリに出現するキーワードがそれを表すと考えられる．

2.1.3 興味を引くキーワードの算出

形態素解析し抽出した各キーワードに対して以下の式により重みを計算する．

$$tf' = a_t tf_t + a_s tf_s + a_p tf_p + a_o tf_o \quad (1)$$

$$W_S = tf' * \log(N/df) \quad (2)$$

$$W_L = tf' * df \quad (3)$$

W_S, W_L：重み（短期，長期）

N：抽出対象の全エントリ数

df：キーワードが出現したエントリ数

tf_t, tf_s, tf_p, tf_o：エントリ内の「タイトル」「小題」「段落のはじめの文」「その他」での出現回数

Extracting bootstrap keywords for self-recognizing users' interest in blog searching
Norikazu KANO, Itaru KURAMOTO, Yu SHIBUYA, Yoshihiro TSUJINO
Kyoto Institute of Technology

a_t, a_s, a_p, a_o ：エントリ内の出現位置による重み（今回の実験では，順に 5，3，1，0.5）

W_S は個々のエントリで，W_L は全エントリを通して計算し，それをもとにキーワードの順位付けを行う．

2.2 検索候補の提示

ユーザが書いたエントリの内，最新の 7 件からそれぞれ W_S が最大となるキーワードを 1 つずつ取り出す．さらに，全エントリを通して計算した W_L の上位 3 個のキーワードを取り出す．ユーザにはこれら 10 個のキーワードを提示し，興味があるものを選択させる．

ここで取り出すキーワードは，ユーザに検索対象のイメージをより明確に持たせやすいものが望ましいと考え，「固有名詞」を優先して取り出す．

2.3 絞込み候補の提示

2.2 節で選ばれたキーワードのうち W_S をもとに取り出したものについては，絞込みの候補としてそのキーワードが出現したエントリ内の残りのキーワードから，Ws の大きい順に 10 個を絞り込み用に提示する．一方，選ばれたキーワードが W_L をもとに取り出したものであれば，絞込みの候補としてそのキーワードと同じエントリに出現したことがあるキーワードのうち，W_L の大きい順に 10 個を提示する．

同じエントリ内に共起するキーワード同士は関係がある可能性が高いと考えられる．提示した中からユーザが 2.2 で選んだキーワードと関係があるキーワードを選択し，ともに用いて検索することで，より興味あるエントリに絞り込めると考えた．

3.実験

提案手法におけるユーザにかかる負担は、キーワードを選ぶという容易な行動であるため少ないと考えられる。したがって本実験の目的は，提案手法による興味あるエントリの検索精度を評価することのみとする．ユーザの書いたブログには mixi[5]の日記を用いる．

3.1 比較システム

提案手法，類似日記一覧の提示，mixi の新着日記一覧，mixi の日記キーワードランキングによる検索の 4 システム間の比較を行った．類似日記一覧の提示とは，ユーザが書いた日記と出現するキーワードが似ている日記を提示する手法である．また，mixi の日記キーワードランキングによる検索とは，mixi が発表しているランキングのキーワード上位 10 個を提示し，そのキーワードで検索を行う手法である．

3.2 被験者

被験者は，mixi を日常的に使用している学生 6 人である．実験期間中の 1 ヶ月の日記数は平均 12.8 エントリ（9～20）であった．

3.3 実験方法

各システムを用いて被験者に興味あるエントリを探索させた．興味あるエントリとは，そのエントリの内容が読みたいと思うかどうかの度合いを 5 段階（1 読みたくない，2 あまり読みたくない，3 どちらでもない，4 少し読みたい，5 読みたい）で判断したとき 4 か 5 となるものとした．

各システムを用いた各々の興味あるエントリの探索について，興味あるエントリの探索終了時に提示されていたエントリのリストの上位 5 件についての興味の度合いと，提案手法とキーワードランキングで検索候補として提示されたキーワードの興味の度合いを 5 段階で判定させた．評価するエントリ数の 5 件とは，エントリのリストが提示されたとき画面内に見える件数である．また，提案手法に関しては，絞込みを行う前の検索結果の上位 5 件に対しても評価させた．

各被験者に対して比較 4 システムでの探索を行わせ，日を変えて計 4 回行った．

4.結果

実験の結果，各システムで提示されたエントリのリストの上から 5 件の興味度の評価値は図 1 のようになった．二元分散分析の結果，システムと被験者の間に交互作用があり，多重比較の結果，提案手法は他のシステムより有意に評価値が高かった（$p<0.05$）．

また，キーワードランキングと提案手法で提示したキーワードは，t 検定の結果，提案手法でのキーワードの方がユーザの興味を引くものであった（$p<0.05$）．

絞込みに関して，絞込みありとなしでの検索結果の上位 5 件の興味度評価を t 検定した結果，絞込みありの方が，検索結果に興味あるエントリが多く出現することが分かった($p<0.05$)．

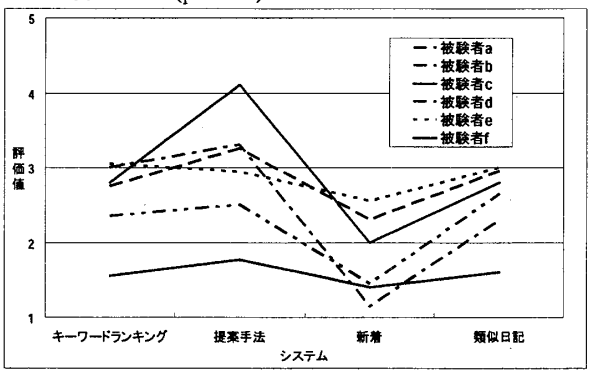

図 1. 興味度評価の平均

5.まとめ

実験結果より，提案手法ではキーワードランキングに比べ各ユーザの興味にあったキーワードを提示できたといえる．このことからユーザ自身が書いた文章に出現するキーワードは，そのユーザの興味をよりよく表すと考えられる．また，絞込み候補として提示された一覧からユーザによって選択されたキーワードを併せて検索することによって，興味ある文書を検出する度合いが高くなることが分かった．

参考文献

[1]Morita,M. and Shinoda,Y :Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval,Proc.of SIGIR '94,pp.272-281(1994).

[2]土方,青木,古井: マウス挙動に基づくテキスト部分抽出方式と抽出キーワードの有効性に関する検証，情報処理学会論文誌，Vol. 43，No. 2，pp.566-576(2002).

[3]大野: 視線情報の再利用に基づくブラウジング支援法,Proc.of WISS2000,pp.137-146(2000).

[4]茶筌:http://chasen-legacy.sourceforge.jp/

[5]mixi：http://mixi.jp/