HTMLファイルからのトピック抽出に基づく 興味推定とWWW検索

非会員 大野 潮 満 (大阪府立大学)

非会員 前田 英巳子 (大阪府立大学)

正員 黄瀬 浩 一 (大阪府立大学)

正員 松本 啓之亮 (大阪府立大学)

Estimation of User's Interests and WWW Retrieval Based on Topic Extraction from HTML Files

Shiomi Ohno, Non-member, Emiko Maeda, Non-member, Koichi Kise, Member, Keinosuke Matsumoto,
Member (Osaka Prefecture University)

In recent years, many methods which assist WWW retrieval based on user's interests have been proposed. However, it is generally difficult to estimate user's interests directly from HTML files, since they often contain multiple topics some of which may not interest a user. In this paper, we propose a method of estimating user's interests and WWW retrieval both of which are based on topics extracted from HTML files. The characteristics of the method are as follows: (1) Topics in a HTML file are extracted by identifying repetitive sequence of its HTML tags, (2) User's interests are estimated by clustering topics extracted from HTML files which contain user's interesting portions. (3) The accuracy of estimation of user's interests as well as WWW retrieval is improved by incorporating retrieved topics into the case-base as positive or negative cases which are specified by a user. Experimental results for 151 HTML files show that the method improves the precision in both of estimating user's interests and WWW retrieval, compared with a method without the extraction of topics.

キーワード：情報検索, WWW, HTML, ユーザの興味, トピック

1. ま え が き

近年WWW (World Wide Web) をはじめとするインターネットの急激な発展により、扱われる情報量は膨大なものとなっている。ユーザの求める情報は一般にそのほんの一部であるため、必要な情報を探し出すことは困難なこととなっている。そこで、情報検索技術が注目されている。

現在WWWの検索には、比較的操作が容易で高速に検索できることから、主に検索エンジンが用いられている。検索エンジンでは、ユーザは検索要求を単語で表現することが多い。しかし、ユーザが与えた単語が、求めるHTMLファイル(以下、ファイル)の中で用いられていないことや、別の意味に用いられていることから、ユーザの満足する検索結果が必ずしも得られるとは限らない⁽¹⁾⁽²⁾。この問題を解決するため、ユーザに単語を推薦する研究⁽¹⁾⁽³⁾がなされている一方で、ユーザが興味を示したファイルに現われる単語を用いて興味を抽出し、WWW検索する研究がなされている⁽⁴⁾。

ユーザが興味を持つファイルから興味を抽出する場合、1つのファイルに1つのトピックが書かれていることを前提にしているものが多い。しかし、WWWでは1つのファイルに複数のトピックが含まれることもあり、ユーザはその一部に興味を持つことが多い⁽⁴⁾。そのため、複数のトピックを含むファイルから、正確にユーザの興味を抽出することは難しい。また、検索エンジンをはじめさまざまなWWW検索では、検索結果をファイルごとに提示する。しかし、複数のトピックを含むファイルをユーザへ提示しても、ユーザはファイルの中から興味のある部分を探さなければならない。さらに、検索結果に対してユーザからの返答を受け、つぎのWWW検索へ反映させる方法⁽⁵⁾もあるが、複数のトピックを含むファイルに対してこれを行うと、ユーザの興味を取り違える可能性がある⁽²⁾。

以上の問題を解決するため、本論文では、ファイルからのトピック抽出に基づいて、ユーザの興味推定とWWW検索を行う手法を提案する⁽⁶⁾⁽⁷⁾。この手法は、ユーザが興味を持つファイルからトピックを取り出し、トピックを用い

図 1 提案手法の概要

Fig. 1. Overview of the proposed method.

て興味を抽出する。トピックごとに処理を行うので、ファイルごとの処理に比べ、より正確に興味を抽出できる。また、WWW検索ではトピックごとに検索することで、ファイルの中でユーザーが求めるトピックを提示できる。さらに、検索結果に対してトピックごとにユーザーから返答を受け、つぎのWWW検索へ反映させることで、ファイルを用いる場合より正確にユーザーの興味を更新できる。

2. 提案手法

〈2・1〉 処理の概要 本手法の処理は図 1 のように大きく (a) トピック抽出, (b) 興味の初期化, (c) WWW検索と興味の更新の 3 つの処理に分かれる。(a) の処理では、与えられたファイルからトピックを抽出する。(b) の処理は、本手法が初めて用いられるときに実行する。この処理では、ユーザーが予め保存しておいたファイルからトピックを得、それを用いて興味を抽出する。ファイルの保存は、ユーザーがそのファイルに興味を持っていることがわかる最も明示的な行動⁽⁸⁾とされており、一部もしくは全体に興味のあるファイルが保存される。また (b) の処理によって生成されるユーザーの興味は、事例ベース⁽⁹⁾を用いて表現される。事例ベースには、ファイルから抽出された個々のトピックが興味の有無を表すラベルとともに格納される。(c) の処理はWWW検索のときに行われる。この処理では、WWWのファイルから得たトピックの中から、ユーザーが求めるトピックを見つけてユーザーへ提示する。また検索結果に対してユーザーから返答を受け、事例ベースを更新することによって、より正確に興味を把握していく。

本手法では図 1 のように、(b), (c) の処理の前に、(a) のトピック抽出によってファイルからトピックを抽出する。以下では、まずトピックの抽出方法と表現方法について述べる。後に (b), (c) の処理方法について順次述べる。

図 2 複数のトピックを含む HTML ファイル：ブラウザ上

Fig. 2. HTML files which contain multiple topics : on a browser.

図 3 複数のトピックを含む HTML ファイル：ソース

Fig. 3. HTML files which contain multiple topics : sources.

〈2・2〉 トピック抽出 ファイルからトピックを抽出するには、一般にはファイルの論理構造を知る必要がある。ファイルで用いられる HTML は、文書型定義 (DTD: Document Type Definition) を与えることで論理構造を表す。しかし実際には、HTML タグは単にレイアウトを表すタグとして扱われることが多い。これは、人間がレイアウトからある程度論理構造を認識することができるからである⁽⁴⁾。人間はファイルをブラウザで見たとき、文章を読まなくてもレイアウトからある程度の論理構造を知ることができる。特に図 2 のように複数のトピックがリスト状に書かれている場合、人間は類似したレイアウトが繰り返されていることからトピックの境界を判断できる。そこで本手法では、類似したレイアウトの繰り返しの着目し、レイアウトをもとにファイルからトピックを抽出する。

レイアウトはファイルの HTML タグによって指定される。図 2 のようにトピックがリスト状に並んでいる場合、図 3 のようにソースファイルでは類似した HTML タグ列が繰

り返されていることが多い。そこで、類似したタグ列の繰返しを見つけることで、ファイルからトピックを抽出する。トピック抽出のアルゴリズムを、図4を用いて以下に述べる。

- (1) 出現した順番でファイルからタグを取り出してタグ列を生成する。
  - (2) タグ列の先頭から1つずつタグを選択する。選択されたタグを t (図4の場合は、 $\langle \text{HR} \rangle$) とすると、 t から2回目の t の直前までのタグをタグ列から取り出して、部分タグ列 T とする。
  - (3) 同様に $i+1$ ($i \geq 1$) 回目の t から $i+2$ 回目の t の直前までを取り出し、部分タグ列 T_i とする。
  - (4) T と各 T_i の距離を求める。距離がある閾値以下のとき T_i を T の繰返し候補と判断する。そして、 T_i が繰返し候補として連続して k ($k \geq 2$) 個続くと、 $T_1 \sim T_k$ を T の繰返し列と判断する。
  - (5) 繰返し列を見つけた場合は T_k の次のタグを選択し、見つからなかった場合は t の次のタグを選択して(2)へ戻る。
  - (6) 見つけた繰返し列の各々に対応するファイルの部分をつまびくとして抽出する。一方、タグ列に繰返しが1つも見つからなかった場合は、ファイル全体を1つのつまびくと判断する。
- (4)において、部分タグ列 T と T_i との距離計算にはDPマッチングを用いる。このことで、タグ列の部分的な差異を加味した距離を計算できる。DPマッチングでは、タグは1対1対応であるとして、タグの脱落を許した対応づけを行う。脱落したタグを記号*で表すとき、タグ t と t' との距離 $d(t, t')$ を次のように定める。

$$d(t, t') = \begin{cases} 0 & \text{if } t = t' \\ d_1 & \text{if } t \neq t' \wedge t \neq * \wedge t' \neq * \\ d_2 & \text{if } t = * \vee t' = * \end{cases} \quad (1)$$

ただし、 $d_1 > d_2$ である。 $d(t, t')$ に基づいてDPマッチングで求められる T と T_i との距離を $DP(T, T_i)$ 、 T のタグの数を $l(T)$ とすると、 T と T_i との距離 $d(T, T_i)$ は次式で計算される。

$$d(T, T_i) = \frac{DP(T, T_i)}{d_1 \min\{l(T), l(T_i)\} + d_2 |l(T_i) - l(T)|} \quad (2)$$

$DP(T, T_i)$ の値は部分タグ列の長さに依存するため、上式では $DP(T, T_i)$ がとり得る最大値(分母)によって正規化している。 $d(T, T_i)$ が閾値 α 以下であれば、 T_i を T の繰返し列の候補とする。

〈2・3〉ベクトル空間法による表現 得られたトピック $H_1, \dots, H_n$ ($n \geq 1$) のテキストを形態素解析⁽¹⁰⁾する。本手法では、以下の3つの規則を全て満たすものを単語として取り出す。

- (a) 品詞が名詞(数詞を除く)、動詞、形容詞のいずれかである。

図4 トピック抽出
Fig. 4. Topic extraction.

- (b) 単語の長さが2文字以上である。
- (c) 単語に1つ以上の漢字を含む。

ただし、人名詞については(b)、(c)を、片仮名については(a)、(c)を適用せずに取り出す。これらの単語をもとに、トピック H_i をベクトル D_i で表現する。WWWのトピックは定型文書と違い、長さが一定ではない⁽⁴⁾ため、それぞれのトピックのベクトルの長さが大きく異なる。長さが一定ではない文書を検索に扱う場合、長い文書の方が検索されやすい特徴がある⁽¹¹⁾。そこで、ベクトル D_i を以下のように正規化したものとする。

$$D_i = v_i / \|v_i\| \quad (3)$$

v_i は m 次元ベクトルであり以下のように表現される。

$$v_i = (w_{i1}F_1, \dots, w_{ij}F_j, \dots, w_{im}F_m) \quad (4)$$

ここで、 m は(a)の処理で出現する全単語数であり、 F_j は H_i において単語 j が出現するなら1しないなら0をとる。 w_{ij} は H_i における単語 j の重みであり、次式で定義される。

$$w_{ij} = \log(tf_{ij} + 1) \cdot \log(n/df_i) \quad (5)$$

図5 クラスタリングによる事例の生成

Fig. 5. Identification of topics by clustering.

ここで、 tf_{ij} は H_i における単語 j の頻度、 df_j は単語 j が現れるトピック数、 n は全トピック数である。

(2.4) 興味の初期化 ユーザの保存したファイルからトピックを抽出すると、ユーザが興味をもつトピックとそうでないものが得られる。したがって興味の抽出のためには、得られたトピックの中から、ユーザが興味を持つものを抽出する必要がある。ここで、ユーザがいくつの興味を持っているかが予めわからないため、興味の数を限定して抽出することはできない。そこで、本手法ではユーザが興味をもつトピックは互いに類似しており、それ以外のものは互いに類似していないと仮定する。すなわち、類似したトピックが多く存在する場合、ユーザがそれらのトピックに興味を持つと考える。逆に、多く集まらなかったトピックには興味を持たないと考える。

具体的な処理は以下の通りである。まずトピック抽出によって、ユーザの保存したファイルの集合からトピックを抽出し、ベクトルで表現しておく。ここで得られたトピックについては、単語が4種類以上含まれるものを以降の処理に用いる。次に、これらのトピックに対して階層的クラスタリングを行う。各クラスはクラスに含まれるトピックのベクトルの重心で表現する。またクラス間の類似度には、ベクトルの内積を用いる。クラスタリングの結果、図5のように類似度が β 以上で集まったクラスを取り出す。これらのクラスのうち、トピックが γ 個以上集まってできたクラスに含まれるトピックをユーザが興味を持った事例(以下、正事例)とする。逆に、それ以外のクラスに含まれるトピックをユーザが興味を持たなかった事例(以下、負事例)として、それぞれ正事例、負事例のラベルとともに事例ベースへ格納する。これらがユーザの興味に関する最初の事例となる。

(2.5) WWW 検索と興味の更新 本手法では、WWWより取得したファイルからトピックを抽出し、トピック単位で検索する。検索単位をトピックとすることで、ユーザはファイルの興味のある部分を検索結果として得ること

表1 ファイル集合の詳細

Table 1. Details of file sets.

ファイル集合	I_p, I	$S_p, S_1 \sim S_3$
ファイル数	19.0	28.3
正解のトピック数	349.5	697.2
興味のあるトピックの形態素数	2575.5	1399.3
全形態素数	11421.0	19522.5

ができる。また、トピックごとにベクトルを生成することで、検索対象をよりの確に表現し、検索精度を上げることが期待できる。

具体的な処理は以下の通りである。まずトピック抽出によって、WWWから取得したファイルのトピックを得て、ベクトルで表現する。次に、トピックを事例ベースの事例と比較して確信度⁽¹²⁾を求める。確信度とはユーザがトピックに興味を持つと確信する度合であり、次のように計算される。トピックを表すベクトル D と事例ベース内の事例の類似度を計算し、類似度の大きい δ 個の事例 $M_1^D \sim M_\delta^D$ を得る。ベクトル D と事例 M_i^D との類似度を内積で表現し $S(D, M_i^D)$ とすると、確信度 C は、

$$C = \sum_{i=1}^{\delta} E_i \cdot S(D, M_i^D) \dots\dots\dots (6)$$

で与えられる。ここで、 E_i は事例 M_i^D が正事例であるか負事例であるかにより変わる重みである。

本手法では確信度が閾値 ϵ 以上の場合、得られたトピックをユーザへ提示し、興味があるかをユーザに尋ねる。そしてユーザの返答を受けて、トピックを正事例もしくは負事例として事例ベースへ追加する。逆に確信度が ϵ 未満の場合は提示しない。このように事例を増やして興味をより正確に把握していく。

3. 実験と考察

(3.1) 実験方法 実験対象として、情報処理学会、人工知能学会、電子情報通信学会、計測自動制御学会、電気学会、ソフトウェア科学会の各学会サイトから論文誌、学会誌、研究会に関するファイルを151個収集した。これらのファイルは、トピックがリスト状に並んでいることが多く、本手法の検証に適している。収集したファイルから $I_p, I, S_p, S_1 \sim S_3$ の6つのファイル集合を作成した。 I_p, I は図1における「ユーザの保存したファイル集合」として用いるものであり、ユーザの興味「エージェント」、「ニューラルネット」、「遺伝アルゴリズム」に関するトピックを含んだファイルからなる。 $S_p, S_1 \sim S_3$ は検索対象、すなわち図1の「WWWのファイル集合」として用いるものである。各ファイル集合の詳細として、ファイル集合あたりのファイル数、予め作成した正解のトピック数、興味のあるトピックの形態素数、全形態素数を表1に示す。

実験ではまず、「トピック抽出」、「興味の初期化」、「WWW検索と興味の更新」について、それぞれ単体での精度を調べた。すなわち、それぞれ入力が理想的な場合(誤りのな

い入力を与えられた場合)について実験を行った。つぎに、これらの処理を図1のように接続し、本手法の精度を調べた。最後に、本手法と本手法から「トピック抽出」を除いた手法(以下、比較手法)との比較を行い、「トピック抽出」の効果を検討した。なお、比較手法ではトピック抽出をしないため、処理はすべてファイル単位となる。

〈3・2〉トピック抽出の精度 収集した全てのファイルを対象に「トピック抽出」を行い、その精度を調べた。パラメータは $d_1=2$, $d_2=1$, $\alpha=0.21$ に設定し、評価には以下のように定められる抽出率を用いた。

$$\text{抽出率} = (\text{一致数}) / (\text{正解のトピック数}) \dots\dots\dots (7)$$

一致数は、本手法で得られたトピックと正解のトピックが一致する数であり、それぞれのトピックに含まれる形態素列が同一であることを一致とした。「トピック抽出」の結果は、本手法で生成したトピックが4072個、正解のトピックが3487個、抽出率は57.7%であった。本手法で生成したトピック数は、1ファイル集合あたり、 I , I_p については254.0個、 $S_1 \sim S_3$ では895.3個となった。

この結果より、「トピック抽出」は単純な処理であるが、抽出率が60%程度あることがわかった。また、「トピック抽出」の精度を下げている原因は主に次の2つであった。

- レイアウトがHTMLタグを使って表現されていない場合があり、トピック抽出ができない。
- 1つのタグを連続的に使用してレイアウトを指定する場合があり、トピックを過度に分割してしまう。

これらの問題への対処は、今後の課題である⁽¹³⁾。

〈3・3〉興味の初期化の精度 「興味の初期化」の精度を調べるために、正解のトピックを用いて「興味の初期化」を行った。ファイル集合には I_p , I を用いた。

まず、 I_p の正解のトピックを用いてパラメータの設定を行なった。ここで、式(5)の n を I_p の正解トピック数として計算した。評価には再現率 R , 適合率 P を用いた。事例ベースへ格納されたトピックの形態素数を X , ユーザが興味を持つトピックの形態素数を Y , ユーザが興味を持ちかつ事例ベースへ格納されたトピックの形態素数を Z とすると、 R , P は以下のように定義される。

$$R = Z/Y \dots\dots\dots (8)$$

$$P = Z/X \dots\dots\dots (9)$$

パラメータ β , γ を評価値 $V = R \cdot P$ が最大になるように求めたところ、 $\beta=0.219$, $\gamma=15$ となった。

つぎに、得られたパラメータを用いて、 I の正解トピックを用いて「興味の初期化」を行なったところ、結果は $R = 82.7\%$, $P = 38.5\%$ となった。「興味の初期化」の適合率が約40%と低くなったのは、一般的な単語を含むトピックどうしがクラスタリングによって集まり、興味のあるトピックとして判断されたためである。本手法では、式(5)で単語に重みを付けたり、長さが極端に短いものをク

図6 WWW検索と興味の更新の実験
Fig. 6. Experiment of WWW retrieval and update of user's interests.

ラスタリングしないという対応をとっている。しかし各トピックの長さが短く、保存したファイルの数も少ないため、単語に与えた重みが必ずしも適切とはいえない。この問題に対処するためには、シソーラスを用いて単語の数を増やすことなどが必要であろう。

〈3・4〉WWW検索と興味の更新の精度 「WWW検索と興味の更新」の精度を調べるために、正解のトピックを使用し、かつ「興味の初期化」が正しく行われた場合について、「WWW検索と興味の更新」を行った。ファイル集合には、 I_p , S_p , I , $S_1 \sim S_3$ を用いた。

実験ではまず、 I_p , S_p を用いてパラメータの設定を行なった。「興味の初期化」が理想的に行われたとするために、 I_p に含まれる正解のトピックのうち興味のあるものを正事例、それ以外のものを負事例として事例ベースに格納した。そして、得られた事例ベースを用いて S_p を検索し、評価値 $V = R \cdot P$ が最大になるパラメータ δ , ϵ の値を求めた。ここで、式(5)の $\log(n/df_i)$ には、「興味の初期化」で得られた値を用いた。また、式(8), (9)において、検索されたトピックの形態素数を X , ユーザが興味を持つトピックの形態素数を Y , ユーザが興味をもちかつ検索されたトピックの形態素数を Z として R , P を計算し、評価に用いた。式(6)については、正事例に $E_i = 1$, 負事例に $E_i = 0$ の重みを用いた。得られたパラメータは、 $\delta = 19$, $\epsilon = 0.90$ であった。

つぎに、これらのパラメータを用いて「WWW検索と興味の更新」を行なった。この実験では、 $S_1 \sim S_3$ をそれぞれ評価用、更新1, 2回目用として、順序を入れ換えることで合計6回の実験を行なった。各実験では図6のように、更新用の集合を1つずつ順に検索し、その度に検索されたものについてユーザからの返答を受けて興味を更新した。また、興味の初期化後と更新後に評価用の集合を検索し、 R, P を求めた。ここでユーザからの返答については、検索された興味のあるトピックを正事例、それ以外を負事例として事例ベースへ格納した。6回の実験の平均は表2のように

表 2 WWW検索と興味の更新の結果
(「トピック抽出」,「興味の初期化」とともに理想的な場合)

Table 2. Experimental results of WWW retrieval and update of user's interests (Topic extraction and initializing user's interests are ideal).

	初期 化後	興味の更新	
		1 回後	2 回後
再現率 R	81.7%	88.4%	89.5%
適合率 P	51.4%	30.3%	30.0%

表 3 WWW検索と興味の更新の結果 (本手法)
Table 3. Experimental results of WWW retrieval and update of user's interests (Proposed method).

	初期 化後	興味の更新	
		1 回後	2 回後
再現率 R	60.2%	61.7%	61.1%
適合率 P	26.0%	30.3%	32.4%

なった。
この結果より初期化後においては、適合率、再現率ともに 50%を越える検索精度が得られている。しかし興味を更新することで、再現率は 7.8 ポイント上がるものの、適合率が大きく下がっている。これは、現在のパラメータが I_p , S_p のみを用いて設定されており、興味の更新を考慮していないためと考えられる。

〈3・5〉本手法の精度 本手法では、〈3・2〉によって得られたトピックを用いた。まず「興味の初期化」について、 I_p からトピックを抽出し、パラメータを求めた。本手法のパラメータは、 $\beta=0.0254$, $\gamma=14$ となった。

得られたパラメータを用いて、 I に対して「興味の初期化」を行った。結果は $R=74.4\%$, $P=44.5\%$ となった。これは、〈3・3〉と同程度の結果である。よって、本手法の「トピック抽出」はそれほど高い精度ではないが、「興味の初期化」への影響は少なく、有効な方法であるといえる。

つぎに、「WWW検索と興味の更新」についてパラメータを求めた。パラメータ設定では、 I_p に対して理想的に「興味の初期化」を行った場合の事例ベースを用いて、 S_p を検索した。パラメータはそれぞれ、 $\delta=13$, $\epsilon=0.94$ となった。

これらのパラメータを用いて、〈3・4〉と同様に $S_1 \sim S_3$ を検索した。ここで、本手法はトピック抽出が理想的でないため、複数のトピックを含んだまま 1 つのトピックとして抽出されたものが存在する。そこでユーザの返答では、ユーザが検索されたトピックの大部分 (形態素数で 8 割以上) に興味がある場合に正事例、それ以外のものを負事例として事例ベースへ格納した。6 回の実験の平均を表 3 に示す。表 3 と表 2 について、初期化後の検索精度を比べると、本手法の方が再現率、適合率ともに 20 ポイント以上低い値となっている。〈3・4〉では本手法と異なり、「トピック抽出」,「興味の初期化」をいずれも理想的に行っていることから、本手法では少なくとも一方の処理が検索精度を下げていると考えられる。

そこで、〈3・3〉で得られた事例ベース (「トピック抽出」

表 4 WWW検索と興味の更新の結果
(「トピック抽出」のみ理想的な場合)

Table 4. Experimental results of WWW retrieval and update of user's interests (Topic extraction is ideal).

	初期 化後	興味の更新	
		1 回後	2 回後
再現率 R	88.8%	80.8%	85.7%
適合率 P	15.5%	26.0%	26.9%

表 5 WWW検索と興味の更新の結果 (比較手法)
Table 5. Experimental results of WWW retrieval and update of user's interests (Comparative method).

	初期 化後	興味の更新	
		1 回後	2 回後
再現率 R	95.2%	87.2%	86.9%
適合率 P	7.0%	8.3%	8.7%

は理想的であるが「興味の初期化」は処理誤りを含むもの)を用いて「WWW検索と興味の更新」を行い、検索精度を調べた。パラメータに〈3・4〉と同じ値を用いたところ、結果は表 4 となった。表 3 と表 4 を初期化後について比べると、検索精度に大きな差は見られない。このことから本手法では、「トピック抽出」の精度は検索に対してあまり影響がなく、「興味の初期化」が検索精度を下げる原因となっていることがわかる。よって「興味の初期化」の精度を改善することで、本手法の精度は向上すると考えられる。

〈3・6〉「トピック抽出」の効果 「トピック抽出」の効果を調べるため、本手法と比較手法について、 I_p , S_p , I , $S_1 \sim S_3$ を用いて比較実験を行った。

まず、比較手法について「興味の初期化」を行った。比較手法では、 I_p の全てのファイルにユーザが興味を持つトピックが含まれているので、 I_p の全てのファイルが正事例として事例ベースへ格納される。よって、パラメータを求める必要はない。 I に対して興味の初期化を行ったところ、結果は $R=100\%$, $P=23.0\%$ となった。

この結果は、〈3・5〉の本手法の結果 ($R=74.4\%$, $P=44.5\%$) と比べて、再現率は高くなるものの、適合率が約 1/2 倍となっている。したがって、本手法は「トピック抽出」の後に「興味の初期化」を行うことで適合率が上がり、よりユーザの興味を正確に抽出できたといえる。

つぎに、「WWW検索と興味の更新」についてパラメータを求めた。比較手法では、 I_p のファイル全てを正事例として格納した事例ベースを用いて S_p を検索したところ、パラメータは $\delta=1$, $\epsilon=0.10$ となった。これらのパラメータを用いて、〈3・4〉と同様に $S_1 \sim S_3$ を検索した。ここでユーザの返答では、検索結果のファイルのうち興味のあるトピックを含むものを正事例、それ以外のファイルを負事例として事例ベースへ格納した。6 回の実験の平均を表 5 に示す。

まず、表 3 と表 5 を初期化後について比較する。本手法は比較手法に比べて再現率は約 0.6 倍と低くなるものの、適合率は約 3.7 倍と大幅に上回っている。つぎに、初期化後

から更新2回後までの適合率、再現率の変化を比べる。本手法は再現率がわずかに向上し、適合率も6.4ポイント上昇するのに対して、比較手法の方は再現率が8.3ポイント下がり、適合率はわずかに向上するだけである。したがって、ファイルからトピックを抽出しトピック単位で処理を行うことで、初期化後の検索精度が向上し、より正確に興味の更新ができたといえる。

本手法の精度を調べた(3・5)から、本手法では「興味初期化」の精度に問題が残るために、検索精度を下げていることが分かった。しかし、本手法は比較手法に比べ、各処理で再現率が多少下がるものの適合率が大きく向上している。以上から、トピック単位で処理することで、処理の精度を上げることができたといえる。

4. む す び

複数のトピックを含むファイルはトピックごとに分けて処理を行う必要があるという観点から、本論文では、トピック抽出に基づいて興味推定とWWW検索を行う手法を提案した。本手法の特徴は、レイアウトを指定するHTMLタグの並びに着目してファイルからトピックを抽出すること、興味のあるファイルから得られるトピックの集合に対してクラスタリングを行なって興味を抽出すること、検索単位をトピックとしてWWW検索を行い興味のあるトピックをユーザへ提示することである。

本手法において、個々の処理と全体での処理の精度を比較した結果、「興味初期化」に改善の余地があることがわかった。しかし本手法は、「トピック抽出」を行わない手法に比べて適合率が大幅に向上するため、有効な手法であるといえる。

現在の問題点は、ユーザの興味を含むファイル数が少なく、トピックの長さが短いために、単語に適切な重みを与えられないことである。今後の課題としては、シソーラスを用いて単語の数を増やし、この問題点に対処することや、トピック抽出の方法を複数のトピックがリスト状に並んでいるもの以外にも適用可能とし、さまざまなファイルに対して検討したい。

(平成11年2月25日受付)

文 献

- (1) 砂山 渡, 野村 勇治, 大澤 幸生, 谷内田 正彦:「Web ページ検索における改善検索キー指示システム」, 人工知能学会研究会資料, SIG-FAI-9801-9(5/29), pp.49-54(1998)
- (2) 杉本 雅則:「情報収集システムにおけるユーザモデリングと適応的インタラクション」, 人工知能学会誌, vol.14, no.1, pp.25-32(1999)
- (3) 大澤 幸生, 須川 敦史, 谷内田 正彦:「ユーザの変化する興味を理解し表現する文獻検索システム Index Navigator」, 人工知能学会誌, vol.13, no.3, pp.461-469(1998)
- (4) 野本 豊裕, 松田 憲幸, 平嶋 宗, 豊川 順一:「文脈情報を利用したブラウジング支援—Context-Sensitive Filtering—」, 人工知能学会研究会資料, SIG-FAI-9702-16(9/30), pp.91-96(1997)
- (5) 江口 浩二, 伊藤 秀隆, 隈元 昭:「ユーザへの適応性を考慮した適合フィードバックによるWWW情報検索」, 電気学会論文誌C, vol.117-C, no.11, pp.1643-1649(1997)

- (6) 大野 潮満, 黄瀬 浩一, 松本 啓之亮:「HTMLファイルの分割部を用いた興味抽出とWWW検索」, 平成10年電気関係学会関西支部連合大会, G15-1, p. G335
- (7) 前田 英巳子, 大野 潮満, 黄瀬 浩一, 松本 啓之亮:「複数のトピックを含むHTMLファイルの分割手法」, 平成10年電気関係学会関西支部連合大会, G15-2, p. G336
- (8) Henry Lieberman:「Letizia:An Agent That Assists Web Browsing」, Proc. of IJCAI, pp.924-929(1995)
- (9) Craig Stanfill, David Waltz:「Towards Memory-Based Reasoning」, CACM, vol.29, no.12, pp.1213-1228(1986)
- (10) 黒橋 禎夫, 長尾 真:「日本語形態素解析システム JUMAN version 3.5」, 京都大学工学部大学院工学研究科 (1998)
- (11) Amit Singhal, Chris Buckley, Mandar Mitra:「Pivoted Document Length Normalization」, SIGIR, pp.21-29(1996)
- (12) Pattie Maes:「Agent that Reduce Work and Information Overload」, CACM, vol.37, no.7, pp.31-40(1994)
- (13) 前田 英巳子, 大野 潮満, 黄瀬 浩一, 松本 啓之亮:「レイアウト情報に基づく複数のトピックを含むHTMLファイルの分割」, 第58回情報処理学会全国大会, 3U-07(1999)

大 野 潮 満 (非会員) 1975年6月29日生。98年大阪府立大学工学部情報工学科卒業。現在、同大学大学院工学研究科博士前期課程電気・情報系専攻在学中。マルチエージェント, WWW情報検索などに関心を持つ。情報処理学会会員。

前 田 英 巳 子 (非会員) 1974年2月22日生。97年大阪府立大学工学部情報工学科卒業。99年同大学大学院工学研究科博士前期課程電気・情報系専攻修了。同年、(株)日立製作所入社。現在、同社システム開発本部に勤務。ソフトウェアエージェント, WWW情報検索に関心を持つ。情報処理学会会員。

黄 瀬 浩 一 (正員) 1963年10月17日生。88年3月大阪大学工学研究科通信工学専攻博士前期課程修了。1989年後期課程退学。1990年大阪府立大学工学部助手。現在、同学部情報工学科助教授。工博。パターン認識, 画像理解, 情報検索に関する研究に従事。電子情報通信学会, 情報処理学会, 人工知能学会 各会員。

松 本 啓 之 亮 (正員) 1952年7月15日生。78年3月京都大学大学院工学研究科精密工学専攻修士課程修了。同年4月三菱電機(株)入社。のち、96年大阪府立大学工学部情報工学科教授となり、現在に至る。主に、電力システムへの知識工学の適用に関する研究に従事。工博。電気学会, 情報処理学会, IEEE 各会員。