HTMLファイル分割に基づくユーザの興味の把握とWWW検索

5V－3

大野 潮満　黄瀬 浩一　松本 啓之亮

大阪府立大学　工学部

1　はじめに

現在WWWで普及している検索エンジンにおいて，ユーザが興味を単語で表現する難しさが問題とされている．解決方法の1つとして，ユーザのブラウジングの履歴等を使う方法が提案されている[1]．しかし，複数の内容を含むことが多いHTMLファイル(以下，ファイル)のブラウジングの履歴からユーザの興味を把握することは容易でない．また，検索結果はファイルを単位としているが，ユーザが求める情報はファイルの一部であることも多い．

そこで本稿では，ファイルを話題ごとに分割して，ユーザの興味を把握し，ファイルの中でユーザが求める部分を検索結果として提示するシステムを提案する．

2　処理方法

処理は図1のように大きく(a)ユーザの興味の初期化，(b)WWW検索と興味の更新に分けることができる．システムの起動時に(a)の処理を行ない，その後(b)の処理を繰り返すことでユーザの興味をより正確に把握する．ユーザの興味は事例ベース[2]を用いて表現される．事例ベース内には，分割されたファイルが興味の有無を表すラベルと共に格納される．以下ではまず，(a),(b)の処理の基本となるファイルの分割方法と分割された部分の表現方法について述べる．次に(a),(b)の処理方法について順次述べる．

2.1　ファイルの分割

ファイルをブラウザで見たときに，同じようにレイアウトされた文章が繰り返されている場合，それらの文章は箇条書のように，互いに異なる話題について述べられていることが多い．

レイアウトはファイルのHTMLタグによって指定される．そこで，本手法ではまず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作る．次に，タグ列の先頭から1つずつタグを取り出す．このタグをAとすると，Aから2回目のAが出るまでをタグ列から取り出して部分列Tを作る．同様に$i+1(i \geq 1)$回目のAから$i+2$回目のAまでを取り出し，部分列T_iを作る．Tと各T_iをDPマッチングにより比較して類似度を計算する．類似度がある閾値以上のときT_iをTの繰り返し候補と判断する．そして，T_iが繰り返し候補として連続してn $(n \geq 3)$個続くとき，$T_i \sim T_n$をTの繰り返し列と判断する．繰り返し列を見つけた場合は，タグ列からT_nの次のタグを取り出し，見つからなかった場合は，Aの次のタグを取り出して，上述と同じ処理を繰り返す．

見つけた繰り返し列の各々に対応するファイルの部分と繰り返し列に含まれない部分を合わせてファイルの分割部を作る．タグ列に繰り返しが1つも見つからなかった場合は，ファイル全体を1つの分割部とする．

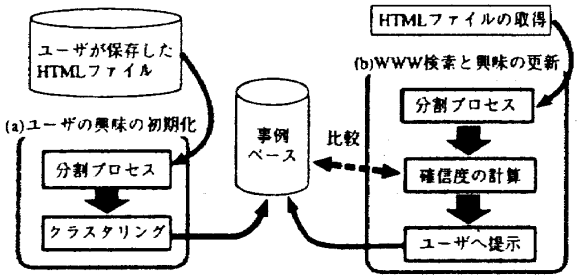

図1　処理のながれ

2.2　ベクトル空間法による表現

得られた分割部の繰り返し列に対応するテキストを形態素解析[3]し，名詞，動詞，形容詞を単語として取り出す．これらの単語をもとに分割部D_iをm次元単位ベクトル$\boldsymbol{v}_i/\|\boldsymbol{v}_i\|$で表す．ベクトル$\boldsymbol{v}_i$は，

$$\boldsymbol{v}_i = (w_{i1}T_1, \ldots, w_{ij}T_j, \ldots, w_{im}T_m)$$

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

で定義される．ここで，mは(a)の処理で出現する全単語数，T_jはD_iにおける単語jの出現(0または1)，w_{ij}はD_iにおける単語jの重み，tf_{ij}はD_iにおける単語jの頻度，df_jは単語jが現れる分割部数，nは全分割部である．また，分割部の類似度にはベクトルの内積を用いる．

2.3　ユーザの興味の初期化

WWWでは1つのファイルに複数の話題が述べられることが多く，一般にユーザはその一部に興味を持つと考えられる．そこで本手法では，ユーザが興味を示すファイル集合を話題ごとに分割し，類似した分割部が多く存在する場合，ユーザがその分割部に興味を持つと考える．

具体的には以下の処理を行なう．ユーザは一部もしくは全体に興味があるファイルを予め保存しておく．保存したファイルの集合を**2.1**の方法で分割し，各分割部を**2.2**の方法で表現する．次に，これらの分割部に対して階層的クラスタリングを行う．ここで，クラスタ間の類似度がt_s未満になるまでクラスタリングを続ける．各クラスタは分割部と同様にベクトル(クラスタ内の分割部のベクトルの重心)で表現される．クラスタリングの結果，t_n個以上の分割部が集まったクラスタを取り出す．これらのクラスタ内の分割部をユーザが興味を持った事

Extraction of user's interests and WWW retrieval based on segmentation of HTML files.

Shiomi Ohno, Koichi Kise and Keinosuke Matsumoto.
College of Engineering, Osaka Prefecture University

例(以下，正事例)とし，それ以外の分割部をユーザが興味を持たなかった事例(以下，負事例)として事例ベースへ格納する．これがユーザの興味に関する最初の事例となる．

2.4 WWW検索と興味の更新

WWWから取得したファイルを話題ごとに分割し，分割部を単位として検索することで，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ができる．

具体的な処理は以下の通りである．まず，WWWから取得したファイルを**2.1**の方法で分割して，分割部を**2.2**の方法で表現する．次に，分割部を事例ベースの事例と比較して確信度を決める．確信度とはユーザが分割部に興味を持つとシステムが確信する度合であり，次のように計算される．分割部 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 t_k 個の事例 $M_1^D \sim M_{t_k}^D$ を得る．分割部 D と事例 M_i^D との類似度を $S(D, M_i^D)$ とするとき，確信度 C は，

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

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．

確信度が高い分割部は，ユーザが興味を持つ分割部である可能性が高い．本手法では確信度が閾値 t_c を越える場合，得られた分割部をユーザへ提示し，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，分割部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 t_c を越えない場合は提示しない．このように事例を増やしてユーザの興味をより正確に把握していく．

3 実験

ファイル分割の有効性を調べるために，本手法と本手法から分割プロセスを除いた手法(以下，分割なし)との比較実験を行った．具体的には，ユーザの興味の初期化，WWW検索と興味の更新について比較した．実験には，学会サイトから得た論文誌，学会誌，研究会に関するファイルの集合，$I_p, D_p, I, D_1 \sim D_4$ を用いた．I_p, I はユーザの保存したファイル各22個の集合であり，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」に関する部分を含む．$D_p, D_1 \sim D_4$ は検索対象のファイルの集合であり，各々19 ~ 26個からなる．実験では I_p, D_p をパラメータ設定用，$I, D_1 \sim D_4$ を評価用とした．

結果の評価には，適合率，再現率を用いた．ここで，ユーザへ提示した全分割部数を X，ユーザが興味を持つ全分割部数を Y，提示した分割部のうちユーザが興味を持つ分割部数を Z とすると適合率は Z/X，再現率は Z/Y で定義される．ただし分割なしでは，検索やクラスタリングの単位をファイルとし，1つのファイルをそのファイルから得られる分割部の集合として扱った．

3.1 パラメータの設定

まず興味の初期化について，I_p を用いて $V =$ (適合率)$\cdot$(再現率) が最大になるパラメータの値を求めた．本手法では t_n =8としたときに t_s =0.0495となった．分割なしでは I_p のファイル全てにユーザは興味をもっているので，クラスタリングのパラメータの値を求める必要はない．次に，本手法と分割なしについて，I_p を用いた興味の初期化で得た事例に基づいて D_p を検索し，V が最大になるパラメータの値を求めた．正事例に $E_i = 1$，負事例に $E_i = 0$ の重みを用いたところ，本手法では $t_k = 24$ で $t_c = 1.51$，分割なしでは $t_k = 9$ で $t_c = 0.899$ となった．

3.2 比較実験

上記のパラメータの値を用いて比較実験を行なった．

興味の初期化の結果を表1に示す．本手法の方が再現率は低いが，適合率は約2倍になっている．適合率が高い方がより正確にユーザの興味を把握できるため，本手法による興味の初期化は有効であるといえる．

WWW検索と興味の更新の実験では，$D_1 \sim D_4$ のうち1つを検索評価用，残りを更新用とし，$D_1 \sim D_4$ を順に検索評価用にして合計4回の実験を行った．各実験では，システムが更新用の集合を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，初期化後と更新後に検索評価用の集合を検索し，適合率と再現率を求めた．実験結果の平均を表2に示す．初期化後，更新後とも本手法は分割なしの手法よりも十分に良い結果を示しており，ファイルの分割の有効性が確認できた．また興味の更新によって，本手法の方は適合率，再現率ともに上昇しており，事例の追加による興味の更新が有効であるといえる．

4 おわりに

本研究ではファイルを分割して，分割部分を用いたユーザの興味の把握とWWW検索の手法を提案した．今後の課題は，ユーザが保存したファイルからより高い精度でユーザの興味を初期化すること，ならびにより効果的な興味の更新方法による検索精度の向上である．

表1 ユーザの興味の初期化の実験結果(単位:%)

	再現率	適合率
本手法	77.1	39.4
分割なし	100.0	21.6

表2 WWW検索と興味の更新の実験結果(単位:%)

		初期化後	興味の更新		
			1回	2回	3回
本手法	再現率	50.6	56.8	58.0	63.0
	適合率	20.8	32.2	30.3	31.7
分割なし	再現率	56.8	19.8	13.6	12.3
	適合率	2.0	3.9	6.2	5.7

参考文献

[1] H.Lieberman: "Letizia:An Agent That Assists Web Browsing", Proc. of IJCAI,pp924-929(1995).

[2] P.Maes : "Agent that Reduce Work and Information Overload", CACM ,vol.37,no.7,pp.31-40(1994).

[3] 黒橋 禎夫, 他: "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).