3U－7

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

前田 英巳子　大野 潮満　黄瀬 浩一　松本 啓之亮
大阪府立大学 工学部 情報工学科

1 はじめに

現在, World Wide Web(WWW) には文書やデータベースなどの様々な情報が無数に存在しており, 利用者が必要な情報を捜し出すには利用者の代わりに情報を検索するシステムの利用が不可欠である. 我々は, 利用者が保存した Hyper Text Markup Language(HTML) ファイルから利用者の興味を抽出し, その結果を用いて検索を行なうシステムについて研究を進めている [1].

WWW の HTML ファイルの特徴として, 1 つのファイルに複数のトピックを含むものが多いことがあげられる [2]. HTML ファイルから利用者の興味を正確に抽出するにはファイルをトピックごとに分割する必要がある.

本稿では, HTML ファイルのレイアウトの特徴を利用したファイルの分割手法を提案する. 本手法の特徴は, HTML タグをレイアウトを示すものとしてとらえ, タグの並びを利用してファイルを分割すること, タグづけされていない部分にタグを挿入することである.

2 HTML ファイルの分割

HTML ファイルをトピックごとに分割するにはファイルの論理構造を知る必要がある. HTML タグは文書型定義 (DTD: Document Type Definition) を与えることで論理構造を表すことができる. しかし, 実際にはレイアウトを示すタグとして扱われることが多く, この点がトピックを抽出する上で問題となる.

人間は様々な形式を用いられていてもレイアウトからある程度論理構造を認識できる [3]. とくに, 複数のトピックが並列的に書かれている場合, レイアウトの類似性からトピックの境界を発見していると考えられる. そこで本手法では, HTML ファイルからタグの並びをいくつか取り出し, その類似度を計算することによってファイルを分割する. 類似度の計算には, タグ列の伸縮を加味できる DP マッチングを用いる.

さて, HTML タグには整形済みのデータを表す <pre>, </pre> がある. <pre> と </pre> で囲まれた部分には HTML タグが含まれていないためにタグ列を利用した分割ができない. この問題を解決するため, 本手法では, <pre> と </pre> で囲まれた部分のレイアウトを解析し, レイアウトを表現する HTML タグを適切な箇所に挿入することによって他の部分と同等に扱う.

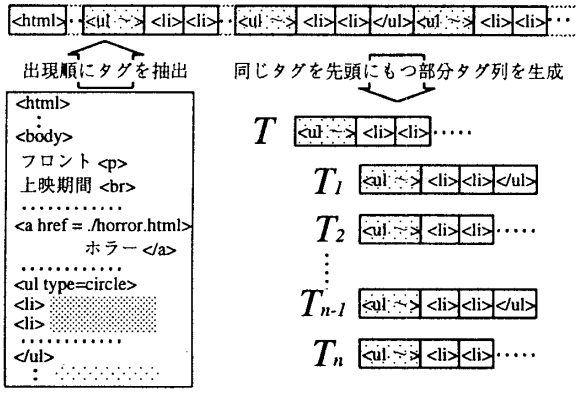

図 1　タグ列の生成と部分タグ列の生成

3 分割手法

具体的な手順を以下で示す.

3.1 タグの挿入

まず, <pre>, </pre> で囲まれたタグを用いていない部分のレイアウトをタグで表現する. レイアウトを表現するタグセットは特別に用意するのではなく, HTML タグを利用する. これは HTML タグでレイアウトの大部分を表現できると考えられるからである. 本手法では, センタリングを表現するために , , 空白行を表現するために <p>, 左寄せの行にインデントされた行が続く場合には <dt>, <dd> を用いる.

空白行は <p> に置き換える. , と <dd>, <dt> タグの挿入は行頭からの空白文字数によって判断する. ファイル f の行長の最頻値を l_f, i 行目の行長を l_i, i 行目の先頭からの空白文字数を s_i とすると, , の挿入は, 以下の式を満たす場合に行なう.

$$s_i > l_f/6$$
$$l_f - 3 \leq l_i + s_i \leq l_f + 3$$

また, $s_i = 0$ である左寄せされた行に

$$s_{i+1} = s_{i+2} = \cdots = s_{i+m} = a(a > 0)$$

となる行が続けば, i 行目に <dt>, $i+k(k = 1, 2, \cdots, m)$ 行目に <dd> を挿入する.

3.2 タグ列と部分タグ列の生成

図 1 のように HTML ファイルからタグを出現順に抽出してタグ列を生成する. 次に先頭のタグから順に 1 つずつタグを取り出す. 取り出したタグと同じタグがタグ

Segmentation of HTML Files into Topics Based on Their Layout
Emiko Maeda, Shiomi Ohno, Koichi Kise and Keinosuke Matsumoto
Department of Computer and Systems Sciences, College of Engineering, Osaka Prefecture University

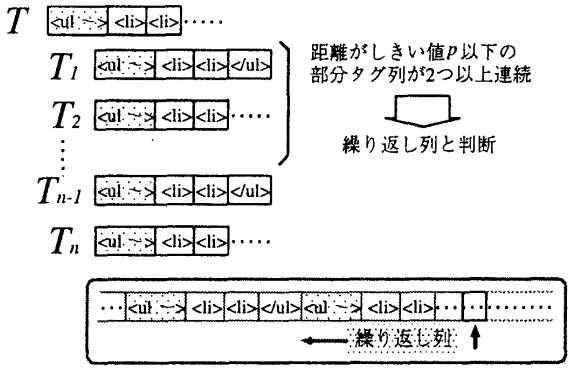

図 2　繰り返し列の判定

列にあれば, 部分タグ列を生成する. 図 1の場合はまず, <ul> が繰り返し使われていることが分かるので <ul> で始まる部分タグ列を生成することになる. 1 つめの <ul> から次の <ul> が出るまでのタグを抽出し, 部分タグ列 T を作る. 同様に, $i+1(i \geq 1)$ 回目の <ul> から $i+2$ 回目の <ul> までのタグを抽出して部分タグ列 T_i を作る. から始まる部分タグ列を全て作り終えたら DP マッチングを用いて T と T_i の距離を計算する.

3.3　DP マッチングを用いた距離計算

部分タグ列 T と類似したタグ列が繰り返されているかどうかを調べるために T と T_i との距離を計算する. 距離計算にはタグ列の伸縮を加味できる DP マッチングを用いる. DP マッチングにはタグ列の伸縮を 2 つまでとする傾斜制限を設ける. また, 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なう. 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.

$$d(t,t') = \begin{cases} 0 & (t = t' \text{のとき}) \\ d_1 & (t \neq t',\ t \neq *,\ t' \neq * \text{のとき}) \\ d_2 & (t = * \text{ または } t' = * \text{のとき}) \end{cases}$$

ただし, $d_1 > d_2$ である. $d(t,t')$ に基づいた DP マッチングで求められる T と T_i との距離を $\mathrm{DP}(T,T_i)$, T の長さを $l(T)$ 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.

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

$\mathrm{DP}(T,T_i)$ の値は部分列の長さに依存するため, 上式では $\mathrm{DP}(T,T_i)$ がとり得る最大値 (分母) によって正規化している. $d(T,T_i)$ がしきい値 p 以下であれば, T_i を T の繰り返し列の候補とする.

3.4　繰り返し列の判定

図 2のように距離 $d(T,T_i)$ がしきい値 p 以下である T_i が連続して $n(n \geq 2)$ 個続けば, $T_1 \sim T_n$ 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て **3.2**, **3.3**で説明した手順を繰り返す. 一方, 全ての T_i に対して $d(T,T_i) > p$ である, あるいは, 先頭のタグ (図 2の場合は <ul>) から始まる部分列がない場合には, 先頭の次のタグを取り出して **3.2**, **3.3**で説明した手順を繰り返す.

最終的には以上の処理で抽出された T,$T_1 \sim T_n$ を個々のトピックに対応する分割部とする. 分割部は HTML ファイルから各部分タグ列に当たる部分を取り出して生成する. なお, タグ列に繰り返し列が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.

4　実験および考察

情報処理学会, 人工知能学会, 電子情報通信学会, 電気学会のサイトから得た HTML ファイル 156 個に対して, 本手法と <pre> と </pre> で囲まれた部分にタグを挿入しない手法との比較実験を行なった. 各パラメータは, $d_1 = 2, d_2 = 1, p = 0.21$ に設定した. 評価は人手で作成した分割部 3551 個に対する抽出率を用いた.

本手法で生成した分割部は 4516 個, 抽出率は 63.9% であった. 一方, タグを挿入しない手法で生成した分割部は 4110 個, 抽出率は 57.5% であった. この結果から, 以下のことが分かる.

- 複数のトピックが並列的に書かれたファイルを対象とすれば, 本手法のような単純な処理でも 60% 程度のトピックを抽出できる.
- <pre> と </pre> で囲まれた部分へのタグの挿入により, 分割の精度を向上できる.
- タグに基づくファイル分割は, 過度に分割する傾向がある.

本手法では, 部分タグ列間の距離計算に DP マッチングを用いているので, 部分タグ列間の多少の違いを吸収できる. しかし, 改行を示す
 や空行を表示する <p> が連続して用いられると, 1 つのトピックが複数行にわたっていても各行を分割部としてしまう. この問題に対処するためには, タグ以外の特徴である文字列中の文字種の規則性などを考慮した分割が必要となるだろう.

5　まとめ

本稿では, HTML タグをレイアウトを示すタグをしてとらえ, タグの並びを利用してファイルをトピックごとに分割する手法を提案した. 今後の課題としては, レイアウトだけでなく, 文字種の違いも利用してファイル分割の精度を向上させることがあげられる.

参考文献

[1] 大野潮満 他: "HTML ファイル分割に基づくユーザの興味の把握と WWW 検索", 情報処理学会第 57 回全国大会,5V-3,pp.3-279–3-280(1998).

[2] 野本豊裕 他: "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).

[3] 大久保雅且 他: "話題が混在するテキストからの話題セグメントの抽出方式", 情報処理学会第 57 回全国大会, 1V-5, pp.3-209–3-210(1998).