2P-4

HTML 文書から RSS を自動生成する方法の提案*

矢島崇史† 勝野裕文‡
東京電機大学大学院理工学研究科§

1 はじめに

Web には膨大な情報が存在する．その多くは HTML で記述されているが，人によりタグの意味づけ方が違うため，コンピュータが情報の意味を取り扱うのは困難である．本研究では，HTML で記述された Web ページの持つ構造に着目し，コンピュータが情報を取り扱いやすい形式である RSS に変換する方法を提案する．

以下では，2 節では本手法の基礎となる木の編集マッピングについて説明する．3 節で提案手法を与え，4 節でその実験結果と考察について述べる．最後に 5 節でまとめをおこなう．

2 木の編集マッピング

順序木 T_i のノード集合を $V(T_i)$ とするとき，T_1 と T_2 間の編集マッピングとは，$V(T_1) \times V(T_2)$ の部分集合 $M(T_1, T_2)$ で，その任意の要素 (v_1, w_1)，(v_2, w_2) が

1. $v_1 = v_2 \iff w_1 = w_2$
2. v_1は v_2の祖先 $\iff$ w_1は w_2の祖先
3. v_1は v_2の左側 $\iff$ w_1は w_2の左側

を満たし，この条件に関して極大なものをいう [2][3].

3 RSS 生成手法

HTML 文書はタグの入れ子構造に基づいた木（以後 DOM 木と呼ぶ）で表現できる．DOM 木では，テキストを (text) という葉ノードで表す．図 1 は DOM 木の例である．ただし，図 1 では属性情報を省略している．実線で囲まれている部分木が抽出したい記事に対応すると考え，以下この例を使って本手法を説明する．

3.1 共通するパス

DOM 木 T_D の根から葉に至るパスに対応するラベル列の部分列 seq に対して，T_D のノード v で，T_D の根から v に至るパスのラベル列の接尾辞が seq である，という条件を満たすものの集合を $N(seq)$ とする．例えば，図 1 では，seq を <p>とすると，$N(seq)$ は部分木（a, b, c, d）の根からなる．接尾辞は PAT 木を用いると得られることが知られている．[1]

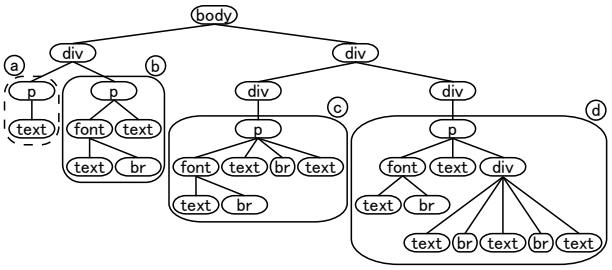


図 1 DOM 木

3.2 クラスタリング

共通するパス seq から定まるノード集合 $N(seq)$ の各ノード n に対し，n を根とする T_D の部分木を $T(n)$ とし，これらの部分木の集合を $\mathcal{T}(seq)$ とする．$\mathcal{T}(seq)$ の各部分木から類似する構造を抽出するために，群平均法と呼ばれる階層的クラスタリング手法を $\mathcal{T}(seq)$ に適用する．

群平均法では，最初は各要素（ここでは部分木）を一つのクラスタと考え，次にクラスタ間の距離が最小となるクラスタ同士を結合し新しいクラスタを作るという操作を繰り返す．クラスタ間の距離 d_c は，クラスタの各要素（部分木）間の距離の平均をとる．

本研究では，部分木間の距離関数として，編集マッピングを利用した以下の関数 d を考案した．

$$d(T_1, T_2) = \sum_{v \in T_1} (a_v p(v)^2) + \sum_{v \in T_2} (a_v p(v)^2)$$

ここで，T のノード v に対して $p(v)$, a_v はそれぞれ重みである．編集マッピング $M(T_1, T_2)$ に参加する，v の先祖 v' がラベルに関するある条件を満たすとき，$p(v)$ は v と v' の間のパスの長さを表す．a_v は $M(T_1, T_2)$ に参加する v の先祖と子孫の属性に関する条件から決まる．

図 2 に考案した距離の計算例を示し，図 3 は図 1 において seq = <p>のとき，$\mathcal{T}(seq)$ に対して群平均法を適用した結果を示す．a, b, c, d, bc, abc, $abcd$ というクラスタが得られている．

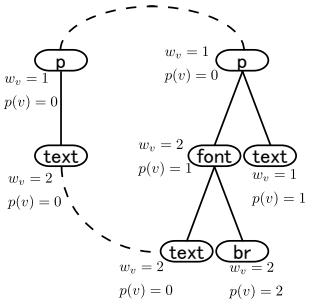
3.3 各クラスタ毎の記事集合の抽出

HTML 文書をブラウザで表示するときに出現する行間は意味的な距離を表すことが多いので，表示で生じる改行の数を基にして文章を群にまとめ，それを本研究では文章群と呼ぶ．

3.2 節で求めた $\mathcal{T}(seq)$ の各クラスタから記事を抽出

---

* An automatic generation method of RSS files from HTML documents using similarities among document structures.
† Takashi Yajima
‡ Hirofumi Katsuno
§ Graduate School of Information Sciences, Tokyo Denki University



$d(T_1, T_2) = 2 \cdot 1^2 + 2 \cdot 2^2 + 1 \cdot 1^2 = 11$

図 2　距離の例

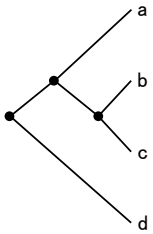


図 3　クラスタリングの結果

する．今，クラスタ C が $T(n_1), \ldots, T(n_r)$ という部分木を持つとし，各 $T(n_i)$ 毎に記事のタイトルにあたる文章群 $G_1(n_i)$ と，本文にあたる文章群 $G_2(n_i)$ を求める．

$G_1(n_i)$ と $G_2(n_i)$ の選択では，$G_1(n_i)$ と $G_2(n_i)$ を含む，$T(n_i)$ の最小な部分木 $T_{1,i}$ と $T_{2,i}$ を考える．そして，各 j（$1 \le j \le r$）に対して $T(n_i)$ と $T(n_j)$ の間の編集マッピング $M_{i,j}$ を考えたときに，$T_{1,i}$ と $T_{1,j}$，$T_{2,i}$ と $T_{2,j}$ の間の対応に整合性があるもののみを考える．

記事候補 $\{(G_1(n_i), G_2(n_i)) \mid 1 \le i \le r\}$ の妥当性を判断するために，タイトル，本文が通常持つ条件として，タイトルは単一の文からなる，本文はタイトルより長い，などを考える．これらの条件を記事候補が満たすかどうかを検査し，最も多く条件を満足する記事候補をクラスタ C の記事集合とする．

3.4　取りこぼした記事の回収

3.3 節の手法では，図 1 の部分木 *a* は文章群を 1 個しか持たないので，*a* を含むクラスタ *abc*，*abcd* は記事候補を持たない．しかし，部分木 *d* は，*b*，*c* のタイトル，本文と共通する構造を持つので，*d* の内部にも記事が存在すると考えられる．そこで，クラスタ C に属さない $\mathcal{T}(seq)$ の要素から C の記事集合と同じ構造をもつ場所を編集マッピングを用いて探し，記事の回収をおこない，C の記事集合に追加する．

3.5　HTML 文書の記事集合の決定

クラスタ C の記事集合において，タイトルの単語の 8 割以上を本文に含むか，タイトルに日付表現を含む記事数を $s(C)$，記事に含まれる文章の総バイト数を $t_s(C)$ とする．記事集合が満たす 3.3 節の条件の数を $c(C)$ とする．このとき，評価関数 F を

$$F(C) = s(C) \cdot t_s(C) \cdot c(C)$$

として $F(C)$ を最大にするクラスタ C の記事集合を共通するパス *seq* の記事集合とし，その最大値を *seq* の評価値とする．最後に，評価値が最大になる *seq* の記事集合を HTML 文書の記事集合にする．

4　実験結果と考察

本研究で考案した手法を用いて，実際の Web ページに対し RSS の生成をおこなった．独自に収集した，記事を持つ Web ページ 10,340 ページからランダムに 500 ページを選び，人手で評価した．

4.1　実験結果

各ページ毎の抽出結果を表 1 に示す．ページ内に含まれるすべての記事が抽出できたページ以外は不備の原因別に集計した．本来よりも大きな記事が抽出された場合の多くは，記事集合の決定の際に DOM ツリー上で本来の記事よりもより根ノードに近いものが選択されたためであった．抽出された記事の数は表 2 に示す．再現率は 79.5%，適合率は 88.1% であった．

表 1　ページ毎の抽出結果の内訳

結果	ページ数	割合
正しく抽出できた	326	65.2
抽出漏れの記事があった	29	5.8
本文が途中で切れていた	47	9.4
本来よりも大きな記事が抽出された	42	8.4
記事ではないものが抽出された	56	11.2

表 2　記事の抽出精度

人手で得られた記事の総数	5245
記事として抽出した数	4738
実際に記事であった数	4172

4.2　考察

すべての記事が抽出できたページ数の割合はあまり高くなかった．これは，ほとんどの記事は正しく抽出できているにもかかわらず，1，2 記事ほど抽出が不完全であるページが多かったためである．特に，本文が途中で切れている例が目立ち，文章に行間を持たせるために多めに改行している記事に対してよく発生していた．タイトルと本文の改行は区別して重み付けをすることでこの問題を解決できると考える．

5　まとめ

今後の課題として精度の向上が上げられるが，抽出の誤りに特定の突出した原因は見られなかった．したがって，現在の手法を大きく変えずに精度を大幅に上げることは難しいと考えられる．今後，文書の構造のみならずテキストの意味処理に関する処理もあわせておこなう必要があると考えられる．

参考文献

[1] Chang C.H.,Lui S.C.; IEPAD: information extraction based on pattern discovery. Proceedings of the 10th international conference on World Wide Web. 2001.

[2] A. Nierman, H. Jagadish; Evaluating structural similarity in XML documents. Proceedings of 5th International Workshop on the Web and Databases, 2002.

[3] 久保山哲二, 宮原哲浩; 木の編集距離を用いた半構造データからの情報抽出. 第 18 回人工知能学会全国大会講演論文集 3F2-05.2004.