7Q-08

HTML内の並列構造を利用したWebページ上のイベント情報抽出

河村　一希†　　竹内　孔一†
岡山大学大学院自然科学研究科†

1　背景，目的

現在，Web上には多数のイベント情報が掲載され，「じゃらん」などの大手イベント掲載サイトも存在する．しかし，規模の小さい地域イベントは大手イベント掲載サイトには掲載されていないことが多い．そのため開催地や開催施設ごとに個別に検索する必要があるが，多数のWebページを人手で検索し情報を入手することは多くの時間と労力を費やす．

このような背景から，本研究では地域イベントの自動抽出を目的とし，(1) HTMLタグ木構造を利用した手法，(2) Support Vector Machineを利用した手法を提案する．

2　地域イベント抽出法

本研究ではイベントをイベント名，開催日時，開催場所，備考の4種類の要素が複数個集まったものと定義し，イベント掲載ページはイベントが複数掲載されているものとする．また，イベント掲載ページはその作成者ごとに自由に記述され，決まったフォーマットは存在していない．さらに，ページ内にはイベントと関係のないテキスト，広告，画像などのノイズが存在する．このような複数のイベント掲載ページからノイズを除去し，イベント単位での抽出を2つの手法で行った．

2.1　HTMLタグ木構造を利用した手法

HTMLタグ木構造を利用した手法について説明する．HTMLタグは入れ子構造を持っており，木構造として考えることができる．本手法でのイベント抽出にあたり，イベント情報掲載Webページのタグ構造の特徴を事前に調査した．

調査結果からイベント掲載ページについての以下の3つの仮説を立てた．

1. イベント掲載ページはイベント情報を主として構成されており，イベント情報に関係するHTMLタグ数が最も多い．
2. イベントは基本的に並列に並んでいる．すなわち複数のイベントが同一の親ノードを持つ．
3. イベントは同一のHTML内で類似したタグ構造のくり返しによって記載されている．

上記の仮説から，方針として，HTMLタグ木の兄弟関係の部分木に対して木構造の類似度比較を行い，類似する構造をもつ複数の部分木を抽出することでイベント情報を得る．

木構造間の類似度を求める手法は，木構造をBinary Branch Vector [1] で記述し，編集距離を利用した計算手法を適用することで求める．これにより，近似的ではあるが効率的な木構造間類似度を求めることが可能である．その詳細な手続きを下記に示す．

(T-i) 全ての兄弟ノード間でその子孫ノードの総数が最大の部分木を獲得する．その際に <script> タグ， タグなどイベント情報が記載されている可能性のないノードは対象外とする．

(T-ii) 得られた最大の部分木と，最大部分木とのノード総数の差が許容誤差 ε_1 以下の部分木を全て2分木に変換する．

(T-iii) 2分木 T から Binary Branch B を定義する．k 番目の Binary Branch B_k はあるノード u，その左ノード u_l，その右ノード u_r より $B_k = uu_lu_r$ と表せる．また，Binary Branch Vector $BRV(T)$ は $BRV(T) = (b_1, b_2, \cdots, b_{|\Gamma|})$ で表され，b_k は B_k の出現回数，$|\Gamma|$ は B の総数を表す．

(T-iv) 最大の部分木とその他の部分木を比較する．比較する部分木 T_1, T_2 は (T-iii) より，$BRV(T_1) = (b_1, b_2, \cdots, b_{|\Gamma|})$, $BRV(T_2) = (b'_1, b'_2, \cdots, b'_{|\Gamma|})$ で表され，その部分木間距離 $d(T_1, T_2)$ は $d(T_1, T_2) = \sum_{i=1}^{|\Gamma|} |b_i - b'_i|$ で求められる．

(T-v) 求めた部分木間距離 $d(T_1, T_2)$, 許容誤差 ε_2 として，$d(T_1, T_2) \leqq \varepsilon_2$ の場合はそれらの部分木をイベントとして獲得する．

(T-ii) で使用したノード総数の許容誤差 ε_1 は，全兄弟ノードの子孫ノード総数の平均を2で割ったものを使用した．また，(T-v) で使用した部分木間距離の許容誤差 ε_2 は5.0に設定した．これらの許容誤差は全Webページに対して最も高い精度を得られる値を実験的に求めた．

Event Extraction from Web Documents Utilizing Parallel Structures in HTML
†Kazuki Kawamura, Koichi Takeuchi, Okayama University

2.2 Support Vector Machine (SVM) を利用した手法

本手法では Web ページのテキスト情報をベースにイベント抽出を行う．SVM によってテキストからイベント名，開催日時，開催場所，備考，ノイズの 5 要素を同定し，イベント情報に関するノイズ以外の 4 要素をイベント単位に区切ることでイベント情報を抽出する．

本研究では多クラス分類を可能にするため LIBSVM* を用いた．また，素性については浅野ら [2] が使用した素性を採用した．イベント名や日付に関する一定の正規表現パターン，CaboCha [3] の固有表現解析結果，品詞，活用型，表層などを素性として利用した．正解データは上記の 5 要素を HTML のテキストノードに対して人手で付与することで作成した．

これらの素性，学習データを用いて SVM を利用したイベント抽出法は以下のようになる．

(S-i) 人手で作成した正解データからイベント情報がどのような要素の組み合わせで構成されているか調査する．1 イベントに対してのイベント要素数，各要素の構成数の分布を算出する．

(S-ii) 学習データから SVM でモデルを作成する．このモデルを用いてテストデータのイベント名，開催日時，開催場所，備考，ノイズの 5 要素を同定する．

(S-iii) SVM でクラスタリングした結果のラベルを HTML のテキストノードに付与する．(S-i) で得られた単体イベントの平均的な構成情報と比較して 1 イベント単位ごとに HTML ノード木を分割しイベント情報を得る．

3 評価実験

地域イベント情報が掲載されている Web ページとして，本研究では岡山県に関するイベント掲載ページを 17 件人手で収集した．収集した Web ページから正解データとしてイベント情報を人手で抽出し，イベントの要素に対してイベント名，開催日時，開催場所，備考の 4 つのタグを付与した．結果，17 件の Web ページから 257 件のイベント情報が得られた．

提案した 2 手法で抽出できた各イベント要素が，正解データのイベント要素を全て含み，かつ抽出したイベント要素数が正解データのイベント要素数の 1.5 倍以内であった場合に抽出したイベントは正しいとする．

SVM を用いた手法に関してはイベント単位での抽出を阻害しないために，Web ページをイベント候補単位で 2 分割し，2 分割交差検定を行った．

*https://www.csie.ntu.edu.tw/~cjlin/libsvm/

4 結果および考察

HTML タグ木構造を利用した手法，SVM を利用した手法によって得られたイベントの適合率，再現率，F 値の 17 件の Web ページの平均を表 1 に示す．

表 1: 木構造手法と SVM によるイベント抽出実験結果

手法	適合率	再現率	F 値
木構造	**0.868**	**0.716**	**0.785**
SVM	0.532	0.484	0.506

表 1 より木構造を利用した手法が SVM を利用した手法より全ての項目で高い結果を得た．木構造手法では節 2.1 で述べた仮説に基づき，木を全探索し類似度比較するのではなく，並列部分のみで類似度を計算しイベント抽出を行った．これにより，計算量を抑えつつ再現率が 0.716 と全体の約 7 割のイベントが獲得できた．

SVM 手法が低い精度となってしまった原因の 1 つとして，SVM のクラスタリングの平均正解率が 0.642 と低いことが挙げられる．これについては，タイトルや開催場所の記述のバリエーションが多く，本質的に機械学習では精度が上がりにくいと考えられる．

5 まとめ

本研究では Web ページからのイベント抽出法として，HTML タグ木構造を利用した手法と SVM を利用した手法の 2 手法を提案し評価した．その結果 HTML タグ木構造を利用した方法では F 値 0.785 と SVM を利用した手法より高い精度を獲得することができ, 木構造ベースのイベント抽出の有用性を確認できた．

参考文献

[1] R. Yang, P. Kalnis, and A.K.H. Tung. Similarity Evaluation on Tree-structured Data. In *Proc. 2005 ACM SIGMOD International Conference on Management of Data*, pp. 754–765, 2005.

[2] 浅野一輝, 竹内孔一. Web ページの HTML 構文構造を考慮した地域イベント情報の抽出. 電子情報通信学会, 言語理解とコミュニケーション研究会, NLC2014-10, pp. 53–58, 2014.

[3] 工藤拓, 松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌, Vol. 43, No. 6, pp. 1834–1842, 2002.