1G3-1

HTMLタグの木構造に着目したWebページのクラスタリング手法

Web Document Clustering Based on HTML Tag Trees

折原 大*1　内海 彰*2
Hiroshi Orihara　Akira Utsumi

*1電気通信大学大学院 電気通信学研究科 システム工学専攻
Department of Systems Engineering, Graduate School of Electro-Communications, The University of Electro-Communications

*2電気通信大学 電気通信学部 システム工学科
Department of Systems Engineering, Faculty of Electro-Communications, The University of Electro-Communications

In this paper, we propose a Web page clustering algorithm based on HTML tag trees. This algorithm clusters Web pages according to their style, rather than classifying them into predefined genres. The algorithm consists of following three steps: (1) construction of the feature vector of a Web page, (2) calculation of a distance, (3) clustering.

The experimental results show that the performance of the algorithm based on HTML tag trees is better than the performance of the algorithm based on HTML tag frequency when Web pages are classified mainly by whether official site or personal site, or classified mainly into news site, weblog and shopping site. The result of overall processing time suggests that the algorithm based on HTML tag frequency is better than the algorithm based on HTML tag trees.

1. はじめに

今日 Web 上には膨大な量の多種多様な情報が存在し，様々な要求を満たす情報を Web 上より得ることが可能である．その膨大な情報の中から必要な情報を探し出すための情報検索技術は必要不可欠となっている．Web 情報検索の一般的な方法として google などに代表される検索エンジンがよく用いられるが，その検索結果には必要な情報とともに不必要な情報も多く得られてしまうことが少なくない．

そこで，検索結果として得られた情報を分類しユーザに提示することで検索支援を行う手法が研究されている．これらの研究は大きく２つに分けて，(a) 文書内の単語から得られる情報に基づく文書クラスタリングを用いた手法 (document clustering)[江口 99, 馬場 07] と，(b) ジャンルと呼ばれる，トピックとは直行する概念へのテキスト分類手法 (text categorization/classification) [Finn 02, 久野 00, Lee 04] がある．(a) では，文書内の単語の分布や共起関係に基づく文書クラスタリングを行う．また (b) では，ユーザが望むカテゴリを予め用意し分類を行う．

現在，(a) に属する研究では文書内の単語情報を用いて内容に基づいて文書クラスタリングを行う手法が主流であり，Web ページのジャンルや形式のような見た目のスタイルに注目した手法についてはあまり研究されていない．しかし，検索要求によっては Web ページのジャンルや形式のようなスタイル別（例えばニュース系サイトとブログ系サイトに分類する，など）に分類する必要がある．

それに対して，(b) に属するジャンルへのテキスト分類の研究では先ほど述べた分類に対応することが可能であるが，それぞれがある特定のジャンル体系に基づく分類を行っており，ユーザのさまざまな検索要求に柔軟に対応することは難しい．例えば，”ショッピングサイト”というカテゴリに分類できても，検索結果のほとんどが”ショッピングサイト”となる場合にはより詳細な分類が必要である．このように，当然のことながら全ての要求に対して事前に分類体系を用意することは不可能であり，さらに新しい分類体系に対応できないなどテキスト分類手法での対応では限界がある．

そこで我々は，Web ページの形式に着目した文書クラスタリング手法として，HTML タグを用いた文書クラスタリング手法を提案した [折原 06]．形式に着目した文書クラスタリングを行うことで，(b) であげたジャンルへの分類も行うことができ，さらに Web ページ群に応じた特定のジャンル体系によらない分類を行うことが可能となる．先行研究では，比較的単純なタグの頻度情報を用いた手法により Web ページの形式によるクラスタリングを行っていたが，本研究では HTML タグの特徴の一つである木構造を用いたクラスタリング手法を提案する．HTML タグの木構造は Web ページの表示に関する情報を多く含んでおり，この情報を用いることでより Web ページの形式によるクラスタリングが行えると考えられる．そして，本研究で提案する HTML タグの木構造 (以下，タグ木と呼ぶ) に基づく手法と，先行研究である HTML タグの頻度情報 (以下，タグ頻度と呼ぶ) に基づく手法との比較を行い，２つの手法の有用性を比較，検証する．

2. HTMLタグを用いたクラスタリング手法

本研究で提案する HTML タグを用いたクラスタリング手法は，次の３つの過程から構成される．

Step.1 ［特徴ベクトルの構成］クラスタリングの対象となる各 Web ページを HTML タグの情報を用いた特徴ベクトルで表現する．

Step.2 ［類似度の計算］Step.1 の特徴ベクトルに基づき，各 Web ページ間の類似度（または距離）を計算する．

Step.3 ［クラスタの生成］Step.2 で求めた Web ページ間の類似度に基づき，クラスタ内の類似度が最大のクラスタ対を逐次結合する．

A: 折原 大, 電気通信大学大学院 電気通信学研究科 システム工学専攻, 〒 182-8585 東京都調布市調布ヶ丘 1-5-1, ori@utm.se.uec.ac.jp

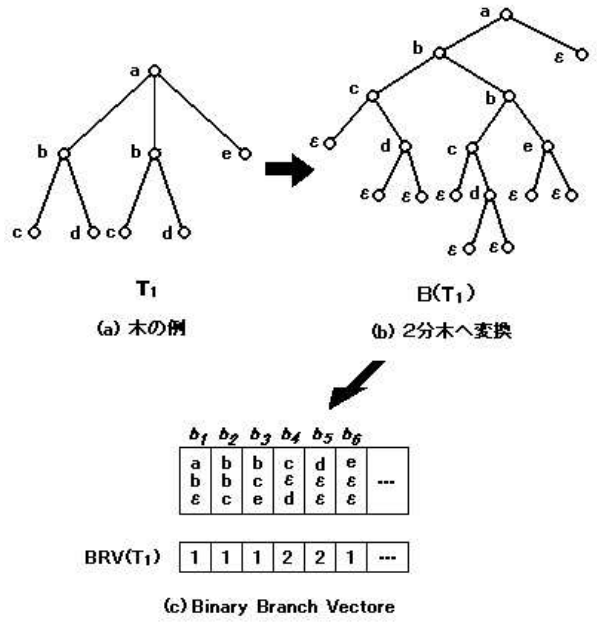

図 1: Binary Branch Vector への変換例

2.1 特徴ベクトルの構成

ここでは，クラスタリングの対象となる各 Web ページに対して特徴ベクトル D_i を構成する．以下に特徴ベクトルの構成方法について述べる．

一般的に木構造間の類似度 (または距離) を求める方法は，主に Bottom-Up 最大共通部分木による類似度, Top-Down 最大共通部分木による類似度, Edit Distance の 3 つがあるが [Gabriel 02]，これらの手法はどれも計算コストがかかる手法である．そこで Rui らは，木構造間の距離を求めるのにより計算コストが小さい手法として，Edit Distance の考え方を元にした Binary Branch Vector 表現を用いた木構造間の距離を提案している [Rui 05]．本研究では，この Rui らが提案する Binary Branch Vector 表現を用いて，タグ木の特徴ベクトルを構成する．

Step.1-1 始めに，任意の木 T_i を 2 分木 $B(T_i)$ に置き換える．2 分木への変換の一般的な手法として，自分の子ノードのうちの左端のノードと自分の兄弟ノードのうちの右隣のノードを用いた表現方法がある．これは，(1) すべての兄弟ノードをリンクで結び，(2) 各ノードの最初の子ノード (子ノードのうちの左端のノード) とのリンクを除く全てのリンクを削除する，ことで求められる．図 1(a) はある木の例 T_1 であり，図 1(b) は木 T_1 を 2 分木 $B(T_1)$ に変換した結果である．さらにここでは木 T_i の全てのノード u について，変換後にどちらかの子ノードを持たない場合，もしくは子ノードを両方持たない場合に ϵ という子ノードを付加している．

Step.1-2 次に，変換された 2 分木に対して Binary Branch を定義する．Binary Branch とは 2 分木のうちの 1 つの階層のみを取り出したものであり，k 番目の Binary Branch b^k は任意のノード u とその左子ノード u_l，右子ノード u_r を使って $b^k = uu_lu_r$ として表される．そして，Binary Branch Vector $BRV(T_i)$ は，Binary Branch b^k を用いて $BRV(T_i) = (b_i^1, b_i^2, \cdots, b_i^{|\Gamma|})$ として表される．ここで b_i^k は，ある 2 分木 T_i における Binary Branch b^k の出現回数とし，$|\Gamma|$ はクラスタリング対象の文書集合中の Binary Branch の総数を表す．図 1(c) は 2 分木 $B(T_1)$ を Binary Branch Vector $BRV(T_1)$ に変換した結果である．

Step.1-3 特徴ベクトル D_i とその要素 D_i^k は，この Binary Branch Vector $BRV(T_i)$ を使って表す．

$$D_i = BRV(T_i) \quad (1)$$

$$D_i^k = b_i^k \quad (2)$$

最後に，各 Web ページ間でのドキュメントサイズの違い (タグ木の大きさの違い) による影響を抑えるために，特徴ベクトル D_i を正規化する．

2.2 類似度の計算

2.1 節で述べた特徴ベクトルを用いて，各 Web ページ間の類似度を求める．一般的に情報検索では特徴ベクトル間の類似度計算にそれらの成す角のコサイン尺度を用いるが，本研究では後述するクラスタ間の距離（または類似度）の再計算手法として Ward 法を用いるため，多次元ユークリッド空間の距離を計算する．なお距離が近いものほど類似度が高いとみなす．

$$d(D_i, D_j) = \|D_i - D_j\| = \sqrt{\sum_{l=1}^{n}(D_i^l - D_j^l)^2} \quad (3)$$

なお，Rui らは Binary Branch Vector に基づく木同士の距離を定義しているが，予備実験において良い成績とならなかったため，Rui らの手法をとらないこととした．

2.3 クラスタの生成

本研究では，文書集合の個々の文書をクラスタとする初期状態から，一つのクラスタになるまで最も類似する（距離の小さい）二つのクラスタを順次併合していくことによって階層構造を構成する階層的クラスタリング法を用いる．

ここで，併合されたクラスタと他のクラスタとの類似度の再計算手法については，最短距離法，最長距離法，群平均法，重心法，Ward 法などが提案されている．これらの手法にはそれぞれ固有のくせがあり，本研究のように対象データに関する性質が未知の場合には，一般的に Ward 法が最も良いとされている [神嶌 03]．そこで，本研究でも Ward 法を用いてクラスタリングを行う[*1]．

3. 評価

3.1 評価方法

2. 章で述べたタグ木に基づく手法を実装した評価用のシステムを構築し，タグ頻度に基づく手法との比較を行った．また，文書クラスタリング手法で一般的な文書中の単語の分布に基づく手法 (Bow) による手法との比較を行った．

なお，タグ頻度に基づく手法では，各パラメータを文献 [折原 06] で最もよい値となった，分割数 $m = 3$， n-gram $n = 2$，カウント方法は有無，IDF は考慮しない，とした．

表 1: 評価データのページ数と分類グループ数

	ページ数	グループ数		ページ数	グループ数
Data01	55	4	Data12	43	5
Data02	47	8	Data13	51	8
Data03	49	6	Data14	40	3
Data04	44	5	Data15	74	3
Data05	51	9	Data16	93	9
Data06	36	3	Data17	47	4
Data07	46	4	Data18	99	6
Data08	35	5	Data19	68	3
Data09	45	7	Data20	54	3
Data10	44	4	Data21	56	3
Data11	43	6			

3.2 評価データ

評価用データは，19 人の大学生に以下の手順により作成してもらった．

1. 被験者は自由に検索要求を考え，検索エンジン goo[*2] を用いて検索を行う．
2. 得られた検索結果の上位 50 件から 100 件を被験者は全て見る．このとき，PDF ファイルや XML 形式の文書など HTML 形式以外のページは評価データの対象外とし省いている．
3. 被験者に対して実験者が「見た目やスタイルが似ているページに分類してください」と教示を行った上で，被験者は閲覧したページを任意のグループに各ページを分類する．分類グループは，実験者は用意せず被験者に自由に作成してもらう．ここで，被験者が分類が難しいと感じたページは「その他」として分類してもらい，これらのページは評価データの対象外とする．

アンケートにより得られた評価データの各ページ数と被験者が分類したグループ数を表 1 に示す．

3.3 評価基準

評価基準は，文献 [折原 06] と同様の以下に述べる評価基準を用いて評価を行った．

検索結果の評価基準としてよく用いられる F 値 (ここでは説明を省略する) を用いて，次のようにしてクラスタ群対の F 値を求める．始めに，正解データのクラスタとシステムが出力したクラスタ対でもっとも F 値の高いクラスタ対を決定する．次に，残るクラスタの中でもっとも F 値の高いクラスタ対を決定する．これを全てのクラスタ対が決定されるまで繰り返す．なお，システムが出力するクラスタ数は，正解データの分類グループ数と同一としている．最後に，決定したそれぞれのクラスタ対の F 値に対して全文書数に対する正解クラスタ内の文書数での重み付けをした平均を求め，これをクラスタ群対の F 値とし評価基準とする．

3.4 評価結果と考察

評価結果の比較

タグ木に基づく手法とタグ頻度に基づく手法それぞれの評価結果を図 2 に示す．ここで平均 F 値とは，全評価データの F 値の平均を表す．

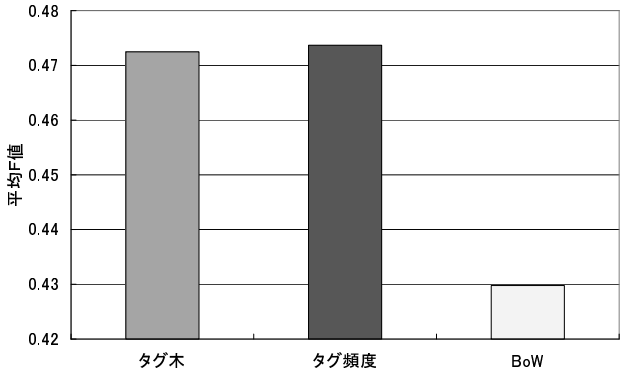

図 2: タグ木とタグ頻度の比較

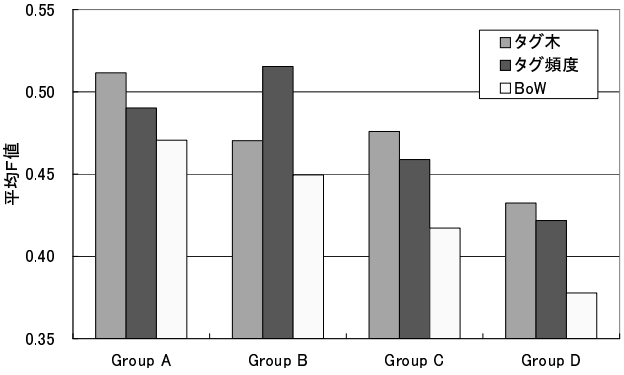

図 3: 分類傾向別に見たタグ木とタグ頻度の比較

タグ木に基づく手法とタグ頻度に基づく手法ではクラスタリングの精度に差がない結果となった．また，単語の頻度情報 (BoW) に基づく手法よりどちらもよい結果となった．この結果から，Web ページの形式によるクラスタリングにおいてはタグ木の情報を利用してもタグ頻度の情報を利用しても精度に差は出ないと言える．また，単語の頻度情報 (BoW) を用いるよりもタグの情報を用いるのが有用であると言える．

そこで，評価データを主な分類傾向に分けてさらに詳しく分析を行った．表 2 に，実験者が抽出した評価データの主な分類傾向と評価データの分類結果を示す．各評価データは被験者が自由に決めた分類グループであり，一意に分類傾向を決められるものではないが，ここでは重複を許さず最もらしい分類傾向に属するものとした．

分類傾向別に見た，タグ木に基づく手法とタグ頻度に基づく手法それぞれの評価結果を図 3 に示す．ここで平均 F 値とは，各 Group ごとの評価データの F 値の平均を表す．

Group B の主にリンクや画像の情報による分類ではタグ頻度に基づく手法がより有用であり，Web ページのタイプ (オフィシャルサイト，ニュースサイトなど) による分類ではタグ木に基づく手法がより有用であることがわかった．この結果より，リンクや画像などはアンカータグのみを利用することで簡単に判断でき，このような場合においてはタグの頻度情報を用いるのみで十分に分類できることが言える．一方，Web ページのタイプは特定のタグのみで判断するには難しく，HTML タグの構造を利用することでより正確に分類できると言える．

*1 これら五つの再計算手法を実装し予備実験を行ったが，直感的に最も良好な結果が得られたのが Ward 法であった．

*2 http://www.goo.ne.jp/

表 2: 評価データを分類傾向で分けた場合

	分類の傾向	
Group A	主にオフィシャルサイトか個人のサイトかに分類している	08, 12, 13, 20, 21
Group B	主にリンクや画像の情報に分類している	03, 05, 09, 11, 17, 19
Group C	主にニュースサイトやブログに分類している	04, 06, 07, 14, 15
Group D	主に内容に基づいて分類している	01, 02, 10, 16, 18

表 3: 処理時間の比較

	ページ数	サイズ (MB)	タグ総数	全処理時間 (sec)	
				タグ頻度	タグ木
Data01	55	2.6	42188	9.9	44.1
Data02	47	1.7	23591	6.0	18.6
Data03	49	2.6	36612	17.0	46.0
Data04	44	3.8	9230	3.3	7.3
Data05	51	11.1	35109	94.7	121.5
Data06	36	1.8	18128	4.3	15.3
Data07	46	2.3	28414	6.5	23.0
Data08	35	0.8	9972	2.4	9.9
Data09	45	2.4	29317	8.4	38.5
Data10	44	1.9	25531	7.2	22.8
Data11	43	12.4	135507	132.4	1706.1
Data12	43	1.4	18952	2.9	9.8
Data13	51	2.4	31883	9.2	21.4
Data14	40	1.5	16400	2.8	9.3
Data15	74	5.5	61177	10.3	33.1
Data16	93	2.4	32521	7.3	37.3
Data17	47	2.3	15859	2.6	8.6
Data18	99	3.0	46411	8.7	27.3
Data19	68	1.9	18620	3.7	10.0
Data20	54	3.5	33871	5.5	22.9
Data21	56	1.6	16210	2.5	8.4

処理時間の比較

ここでは，さらにタグ頻度に基づく手法とタグ木に基づく手法の処理時間の比較を行った．各評価データに対するクラスタリング処理時間を表 3 に示す．表中のサイズとは各評価データごとの HTML 文書の総データサイズを表し，タグ総数とは各評価データごとの HTML 文書に出現するタグの総数を表す．

この結果から，タグ木に基づく手法はタグ頻度に基づく手法に比べ，おおむね約 3 倍の処理時間を要していることがわかる．タグ頻度に基づく手法は簡単な処理で実現できるが，タグ木に基づく手法はタグの木構造を解析する処理に時間がかかるため，このような結果となった．両手法間でクラスタリング精度はほぼ同等であることから，処理時間を考慮した場合においてはタグ木に基づく手法よりもタグ頻度に基づく手法のほうがよいと言える．

4. おわりに

本研究では，HTML タグの木構造の情報を用いた Web ページのクラスタリング手法を提案し，その評価を行った．HTML タグの頻度情報を用いた場合と比較した結果，全体ではほぼ同等のクラスタリング精度であるが，Web ページの分類傾向により手法を使い分けるとクラスタリング精度が向上することを示した．さらに，処理時間を考慮した場合においては，タグの頻度情報に基づく手法のほうがタグの木構造に基づく手法よりもより効率的であることを示した．

今後は，本手法を実装した検索支援システムを構築し，検索支援システムとしての有用性について評価を行うことを考えている．

参考文献

[馬場 07] 馬場 康夫, 笹田 鉄郎, 新里 圭司, 黒橋 禎夫. 構造的言語処理による大規模ウェブ情報のクラスタリング. 言語処理学会第 13 回年次大会発表論文集 (NLP2007), pp.562-565, 2007.

[江口 99] 江口 浩二, 伊藤 秀隆, 隈元 昭, 金田 彌吉. 漸次的に拡張されたクエリを用いた適応的文書クラスタリング法. 電子情報通信学会論文誌 (D-I), J82-D-I, 1, pp.140-149, 1999.

[Finn 02] Aidan Finn, Nicholas Kushmerick and Barry Smyth. Genre Classification and Domain Transfer for information filtering. Proceedings of the 24th BCS-IRSG European Colloquium on IR Research, pp.353-362, 2002.

[Gabriel 02] Gabriel Valiente. *Algorithms on Trees and Graphs*. Springer-Verlag, Berlin, 2002.

[神嶌 03] 神嶌 敏弘. データマイイニング分野のクラスタリング手法 (1). 人工知能学会誌 Vol.18, No.1 pp.59-65, 2003.

[久野 00] 久野 高志, 安形 輝, 石田 栄美, 上田 修一. Web ページのタイプ判別法. 2000 年度日本図書館情報学会 春季研究大会発表要綱, pp.55-58, 2000.

[Lee 04] K.-J. Lee : "Document Genre Classification for User Interface of Web Search Engine", IEICE Transactions on Information and Systems, E87-D, 7, pp.1982-1986, 2004.

[折原 06] 折原 大, 塚田 大介, 内海 彰. HTML タグを用いた Web ページのクラスタリング手法. 第 20 回人工知能学会全国大会論文集, 1A3-5, 2005.

[Rui 05] Rui Yang, Panos Kalnis and Anthony K. H. Tung. Similarity Evaluation on Tree-structured Data. *SIGMOD Conference*, pp.754-765, 2005.