1A3-5

HTMLタグを用いたWebページのクラスタリング手法

Web Document Clustering using HTML Tags

折原 大[*1] 塚田 大介[*2] 内海 彰[*2]
Hiroshi Orihara Daisuke Tsukada 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

This paper proposes a method for web document clustering technique that uses HTML tags. we propose four elements to generate document vectors, (1) the number of partitions that considers position where tag appears, (2) n-gram of tags, (3) whether the frequency of or the presence of tags is used and (4) whether the idf value is considered. This paper also reports an evaluation experiment in which the performance of the proposed methods is compared with that of the existing tfidf-based method.

1. はじめに

今日 Web 上には多種多様な情報が存在し，ユーザはそこから様々な情報を得ることが可能である．それに伴い，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検索エンジンがよく用いられているが，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文書クラスタリングを用いて選別し，ユーザに提示する検索支援手法が研究されている．これらの研究は大きく分けて，(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, 成田 02] 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 (text categorization) 手法 [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っている．また (b) では，予めジャンル体系を用意しその体系への分類を行っている．

現在，文書クラスタリング手法は (a) であげた文書内の単語に基づく手法が主流であり，Web ページの形式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，文書の内容は類似したものであっても，例えばニュース系サイトとブログ系サイトといった Web ページの形式により必要な文書と不必要な文書が存在する場合があり，形式に基づく分類が必要である．

一方 (b) であげたジャンルへのテキスト分類の研究では，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，ある Web ページ群中にオークション系サイトが多くある場合に，ユーザが各社（各サイト）毎に分類されることを要求することも考えられるが，このような要求には予め分類体系を用意する手法での対応は難しい．

そこで本研究では，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する．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．また，提案する手法を (a) の文書クラスタリングで一般的な手法である単語の分布に基づく手法 (Bow:Bag-of-words) と比較し，その有用性を検証する．

連絡先: 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, 0424-42-5258, ori@utm.se.uec.ac.jp

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．

2.1 特徴ベクトルの構成

本研究では，クラスタリングの対象となる各 Web ページを特徴ベクトル D_i で表現し，以下のように構成する．

1. 対象となる Web ページに対して HTML タグを抽出する．ただし，<BODY> タグ範囲外のタグ，コメントタグ範囲内のタグ，要素の終わりを示すタグ (例:</A>) は抽出の対象外とする．

2. 抽出したタグに次の 2 種類の特徴を考慮したものを特徴ベクトルの属性 k とする．

   分割数 m　Web ページ内のどの位置に出現しているかを考慮する要素．抽出されたタグ総数を m 等分し，m 個の範囲に分割したものをそれぞれの属性とする．例えば，$m = 2$ であれば同じ要素のタグであっても前半に出現したタグと後半に出現したタグは別の属性とする．

   n-gram n　特徴的なタグの組み合わせを考慮する要素．抽出されたタグを連続する n 個の組み合わせで 1 つの属性とする．例えば，$n = 3$ であれば連続する 3 つのタグの組み合わせを 1 つの属性とする．

表 1: 評価データの詳細

検索要求	検索クエリ	抽出した形式	抽出ページ数
姉歯建築士に関する情報	"姉歯"	・ニュース系サイト	22
		・ブログ系サイト	39
デジタルカメラの製品情報および価格情報	"デジタルカメラ"& "価格"	・価格比較情報を掲載しているサイト	13
		・ショッピングサイト	25
		・各メーカーのサイト	18
カルボナーラのレシピ	"カルボナーラ"& "レシピ"	・手順が文字情報のみで掲載されているサイト	46
		・手順すべてに写真が掲載されているサイト	12
民主党の代表選に関する情報	"民主党"& "代表選"	・ニュース系サイト	30
		・ブログ系サイト	20

3. 次に，各 Web ページの特徴ベクトル D_i について，各属性 k の属性値 D_i^k を次の 2 つの要素を用いて求める．

属性値のカウント方法 各属性の頻度に基づく方法か，属性の有無に基づく方法かを選択する．

はじめに各 Web ページにおいての各属性 k ごとに頻度 tf_i^k を求める．

次に，頻度に基づく方法の場合は式 (1) のように各属性の頻度をそのまま属性値 D_i^k とし，有無に基づく方法の場合は式 (2) のようにその属性が 1 つでも出現すれば 1 として求める．

$$D_{i\ \text{頻度}}^k = tf_i^k \tag{1}$$

$$D_{i\ \text{有無}}^k = \begin{cases} 1 & (tf_i^k > 0) \\ 0 & (tf_i^k = 0) \end{cases} \tag{2}$$

idf 値の考慮の有無 idf 値を考慮するか考慮しないかを選択する．

idf 値とは Web ページ集合中でその属性 k を含む Web ページ数 df^k と全 Web ページ数 N の比の対数を取ったものであり，この idf 値を用いると少数の Web ページに出現する属性に大きい重みを与えることができる．

idf 値を考慮する場合は，式 (3) のように属性値のカウント方法で求めた値に idf 値をかけ合わせたものを各属性値とし，idf 値を考慮しない場合は，式 (4) のように属性値のカウント方法で求めた値のままとする．

$$D_{i\ \text{あり}}^k = tf_i^k \cdot idf_i^k = tf_i^k \cdot \left(log \frac{N}{df^k} + 1 \right) \tag{3}$$

$$D_{i\ \text{なし}}^k = tf_i^k \tag{4}$$

4. 最後に，各 Web ページごとに抽出されるタグ数が大きく異なる影響をなくすため，各特徴ベクトルの長さ $\|D_i\|$ が 1 となるように正規化する．

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．

$$d(D_i, D_j) = \|D_i - D_j\| = \sqrt{\sum_{l=1}^{n} (D_i^l - D_j^l)^2} \tag{5}$$

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，本研究では 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. 章で述べた手法を実装した評価用のシステムを構築し，本研究で提案する手法の有用性を評価した．

今回の実験では，分割数 m と n-gram n はそれぞれ 1 から 5 までとし，システムが出力するクラスタ数を各正解データのクラスタ数と同数に設定し評価を行った．

また従来の手法との比較を行うために，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) によるクラスタリングも行った．

3.2 評価データ

評価用データとして，次の方法で実際の検索に近い評価データを作成した．まず 4 種類のまったく異なる検索要求を用意し，それらに対する検索クエリを作成し既存の検索システムを用いて検索を行った．その得られた検索結果の上位 30 件から 50 件ほどを概観し，特徴あるジャンルを 2 つもしくは 3 つ設定した．この設定したジャンルに明らかな判断基準を設け，検索結果の上位から判断基準を満たす Web ページのみを抽出した．ここで作成した評価データを表 1 に示す．

*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．

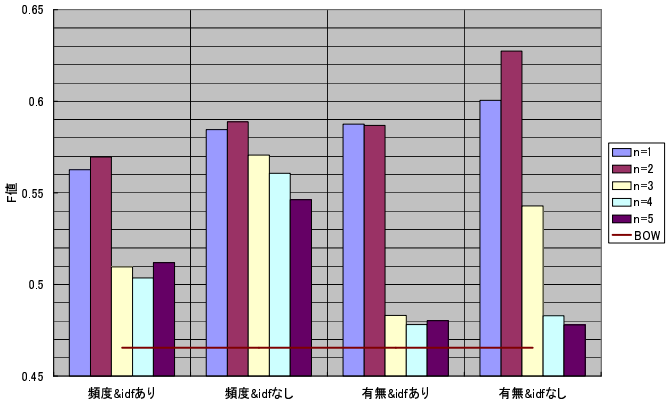

図 1: n-gram に対する F 値

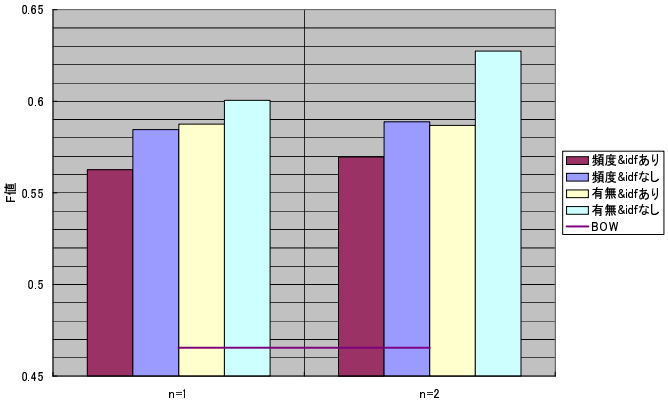

図 2: n=1,2 とした場合の属性値の求め方に対する F 値

表 2: n-gram に対する F 値

n	頻度& idf あり	頻度& idf なし	有無& idf あり	有無& idf なし
1	0.563	0.585	0.587	0.600
2	0.570	0.589	0.587	0.627
3	0.509	0.571	0.483	0.543
4	0.504	0.561	0.478	0.483
5	0.512	0.546	0.480	0.478
Bow				0.465

表 3: n=1,2 とした場合の属性値の求め方に対する F 値

属性値の求め方	n=1	n=2
頻度&idf あり	0.563	0.570
頻度&idf なし	0.585	0.589
有無&idf あり	0.587	0.587
有無&idf なし	0.600	0.627

3.3 評価基準

評価用システムが出力するクラスタ群と正解データのクラスタ群がどの程度近いかの指標として，検索システムの評価で一般的に行われている F 値の考え方に基づく，以下の評価基準を用いる．

まず，F 値とは適合率と再現率の調和平均のことであり，次のようにして求められる．正解クラスタ群を $L = \{L_1, \cdots, L_c\}$，システムが出力したクラスタ群を $S = \{S_1, \cdots, S_o\}$ とする．また，全 Web ページの数を N，正解クラスタ L_r に含まれるページ数を l_r，クラスタ S_i に含まれるページ数を s_i とし，L_r と S_i の両方に含まれるページ数を n_{ri} とする．このとき，任意のクラスタ L_r と S_i との F 値 $F(L_r, S_i)$ は，再現率 $R(L_r, S_i)$ と適合率 $P(L_r, S_i)$ より次のように求める．

$$
\begin{aligned}
F(L_r, S_i) &= \frac{2 * R(L_r, S_i) * P(L_r, S_i)}{R(L_r, S_i) + P(L_r, S_i)} \\
R(L_r, S_i) &= \frac{n_{ri}}{l_r} \\
P(L_r, S_i) &= \frac{n_{ri}}{s_i}
\end{aligned}
$$

この F 値を用いて，次のようにしてクラスタ群対の F 値を求める．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し，残るクラスタ対でのもっとも F 値の高いクラスタ対を逐次決定していき，すべてのクラスタ対を決定する．最後に，決定したクラスタ対の 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，これを評価基準とする．

3.4 評価結果と考察

n-gram の結果と考察

まずはじめに，属性値の求め方ごとに n-gram n を変化させた場合の結果を表 2 および図 1 に示す．なお，F 値は 4 つの評価データの平均値であり，また分割数 m を 1 から 5 まで変化させた場合の平均値である．

この結果より，まずすべての場合において文書中の単語の分布に基づく手法 (Bow) よりも良い結果となった．このことから，Web ページの形式に着目した文書クラスタリングを行うには，単語の分布に基づく手法よりも HTML タグに基づく手法が有用であることが言える．例えば，ニュース系サイトかブログ系サイトかを分類するのにそれぞれ特徴的な単語はあると思われるが，それよりも HTML タグの情報のほうがより Web ページの形式の特徴を表していると言える．

また，それぞれの属性値の求め方において n-gram が 2 の場合のほうが比較的良い結果となっていることから，HTML タグの組み合わせを考慮したほうがより Web ページの形式をとらえることができると言える．しかし，n-gram が 3 以上の場合には明らかに結果が悪くなることから，組み合わせを考慮したほうがよいもののせいぜい隣り合う 2 つまでを考慮したほうがよいと言える．

属性値の求め方の結果と考察

次に，図 1 より，属性値の求め方に依らず n-gram が 1,2 の場合が明らかに良い結果となったことから，この場合において属性値の求め方に対する 2 つの要素について考察する．その結果を表 3 および図 2 に示す．

n-gram が 1,2 のどちらの場合においても，属性値の求め方は属性の有無に基づき idf 値を考慮しない場合が良い結果となった．これは，文書中の単語に基づく手法では一般的に頻度の情報を用いているのに対して，本研究で提案する HTML タグに基づく手法では頻度を考慮せず単純な有無のほうが良いと

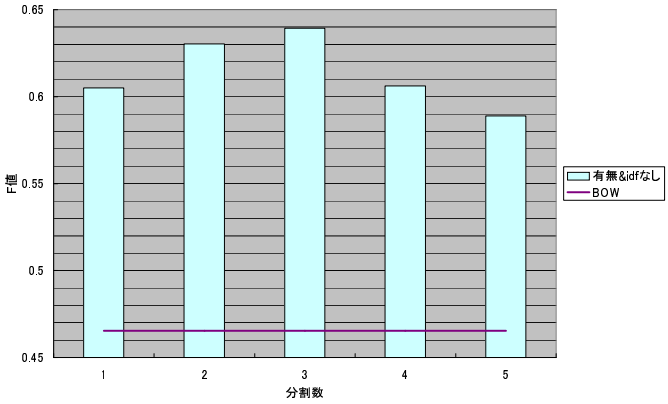

図 3: $n = 1, 2$, 属性値の有無, idf なしの分割数に対する F 値

表 4: $n = 1, 2$, 属性値の有無, idf なしの分割数に対する F 値

	1	2	3	4	5
有無&idf なし	0.605	0.630	0.639	0.606	0.589

いう結果となった.

理由として，HTML タグは属性の種類が有限であるためではないかと考えられる．また，例えば <A> タグのように頻出しやすいタグと <FORM> タグのような頻出しにくいタグがあるため，頻度を考慮するよりも有無を考慮するほうがより Web ページの形式をとらえられると考えられる．idf 値の考慮に関しても同様に属性が有限であり，特定の Web ページにのみ出現する HTML タグが必ずしも Web ページの形式を表しているとは言えないと考えられる.

分割数の結果と考察

最後に，属性値の求め方が属性の有無に基づき idf 値を考慮しない場合において，分割数 m を変化させた場合の結果を表 4 および図 3 に示す．なお，F 値は n-gram が 1,2 の場合の 4 つの評価データの平均値である.

Web ページ内の出現位置をまったく考慮しない場合 ($m = 1$) よりも考慮した場合 ($m = 2, 3$) のほうが良い結果となった．Web ページ内での HTML タグの出現位置を考慮することで，ブラウザ上での見た目をより反映され，より Web ページの形式をとらえることができたと言える.

さらに，$m = 3$ の場合が最も良好な結果となった．これは n-gram が 1,2 の場合をそれぞれ個々に見た場合でも同様の結果となっていた．このことから，HTML タグの出現位置を「前，中，後」程度に分けることが最も Web ページの形式をとらえることができると言える.

4. おわりに

本研究では，HTML タグを用いた 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．Web ページの形式で文書クラスタリングを行う場合は，単語の分布に基づく手法よりも HTML タグに基づく手法が有用であることを示した．また，HTML タグを用いた文書クラスタリングでは，属性の有無を属性値とし idf 値は考慮しないほうが有用であること，HTML タグの組み合わせや出現位置を考慮したほうが有用であることを示した.

今後の課題としては，HTML タグの出現位置を HTML 総数を等分しているだけであるが，テキスト分割を行うなど，より見た目を考慮した Web ページの分割方法の検討する必要がある．また，評価データを実験者以外の人に作成してもらうなどして，さらにさまざまな検索要求に関する場合での評価を行うことを考えている.

参考文献

[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).

[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).

[成田 02] 成田 宏和, 太田 学, 片山 薫, 石川 博 : "階層的クラスタリングを利用したメタ検索エンジンの提案～METAL～", 情報処理学会研究報告, DBS128-50, pp.375-382 (2002).

[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).