ブログページ集合からのポスト及びコメントの自動抽出

吉田 光男 †　　乾 孝司　　山本 幹雄

筑波大学大学院システム情報工学研究科
† m.yoshida@mibel.cs.tsukuba.ac.jp

1 はじめに

近年の Web ページの増加は，ブログの普及に一因がある．日本では，2004 年から 2005 年ごろにかけてブログ及びその記事が急増しており，現在も増加傾向が見られる [1]．ブログの普及に伴い，ブログのコンテンツを利用する研究も盛んになってきている [2]．

図 1 のブログページ*1からもわかるとおり，ブログページは，ヘッダ，メニュー，広告，関連記事リストなど不要部分が多々存在しており，ページに占めるコンテンツ（主要部分）の割合が低い．そのため，ブログのコンテンツを利用するためには，コンテンツ抽出が必要になる．従来のコンテンツ抽出（本文抽出，主要部分抽出などとも呼ばれる）は，ポストとコメントを区別せずに抽出したり，ポストのみを抽出したりしていた．しかし，ブログのコメントを利用する研究 [3, 4] が行われ始めるなど，今後，ブログのコンテンツ抽出はポストとコメントを分離抽出することが期待される．

コンテンツ抽出手法の代表例として，人手によって抽出ルールを記述する方法が挙げられる [5]．コンテンツ抽出を行う際は，ブログページごとにコンテンツの位置が異なることに留意する必要がある．同一のブログホスティングサービスを利用しているブログページに関していえば，コンテンツの位置が同一であると考えられるため，ブログホスティングサービスごとに抽出ルールを準備する必要があるものの，人手によって抽出ルールを記述する方法は有効に機能する．しかし，我々による予備調査では，人気のあるブログサイト*2の少なくとも 30% がブログホスティングサービスを利用していないことが判明した．このような状況下では，人手によって抽出ルールを記述する方法には多大な労力を必要とすることは明らかであり，自動抽出手法を検討する必要がある．

ポストとコメントを分離抽出する先行研究として Cao ら [6] があり，我々はその先行研究が抱える問題点のいくつかを解決する手法を提案した [7]．我々が提案した手法は，コンテンツのうち，ポストはブログページ集合全てのブログページに出現するが，コメントはいずれかのブログページにしか出現しないという仮説をもとにする．この手法の中で，HTML のブロックレベル要素をもとに定義した，コンテンツと不要部分のかたまり（ブロック）の他に，要素識別子を用いることで，各ブロックをグループ化する手法を提案した．

本稿では，グループ化されたブロックを用いることで，コンテンツ抽出性能を改善する手法を提案する．その結果，ポスト及びコメントの抽出性能が，それぞれ 87.7%，87.4% に改善した．

図 1　ブログページにはコンテンツではない不要部分が多く存在し，コンテンツはポストとコメントに分かれる．

*1 http://www.100shiki.com/archives/2009/08/gscreen.html (cited 2009-10-23)

*2 livedoor Reader (http://reader.livedoor.com/) 登録者数ランキング上位 1,000 サイトから推定した．

2 提案手法

2.1 コンテンツ及びポスト・コメントの定義

本稿では，ブログページの**不要部分**を除いた主要部分を**コンテンツ**とし，コンテンツをさらに**ポスト**と**コメント**に二分する．ポストは記事本文のほか，それに付随する記事タイトル，投稿日時，著者名，写真・図及びその説明文を指す．コメントは読者によるコメント本文及びトラックバック本文のほか，それらに付随するタイトル，投稿日時，コメント著者名（トラックバック送信元ブログ名）を指す．なお，単にコンテンツと記した場合は，ポストとコメントを区別しない．

2.2 提案手法の概要

本稿で提案するポストとコメントを自動的に分離抽出する手法（以下，**提案手法**と呼ぶ）は，我々が過去に提案した手法 [7]（以下，**従来手法**と呼ぶ）の拡張である．提案手法は以下の 6 過程からなる．

処理 1　ブログページの収集
処理 2　ブロックの抽出
処理 3　コンテンツの抽出
処理 4　位置ラベルの付与
処理 5　**コンテンツの再抽出**（新たに追加する処理）
処理 6　ポストとコメントの分離抽出

従来手法は，コンテンツとして認められるブロックのうち，ブログページ集合全てのブログページの同じ場所に出現するブロックをポストとして抽出し，残りのブロックをコメントとして抽出する手法である．この手法は，上の処理 1 から 4 及び 6 の過程を順次行うことに相当する．提案手法では，コンテンツ抽出を強化する，処理 5 の過程を追加した．その結果，ポストとコメントの分離抽出性能が向上した．

以下，上に示した処理の概要を説明する．詳細については，処理 1 から 3 は文献 [8]，処理 4 及び 6 は文献 [7] を参照されたい．

処理 1 では，コンテンツ抽出の対象となるブログページ集合の準備を行う．このブログページ集合は，(1) 同じブログサイトのページで構成される 2 ページ以上の集合，(2) コメントが付いていないページが含まれる集合，の 2 条件を満たすものとする．一般的にブログページのコメント率は低く，上の条件 (2) は特殊な事例ではない．

処理 2 では，HTML のブロックレベル要素 [9] をもとに，コンテンツ及び不要部分の最小単位である**ブロック**の抽出を行う．ブロックレベル要素をもとに行うことで，ページレイアウト方法の流行の影響を受けずにページを分割することができる [8]．ブロックレベル要素を用いてブロックを抽出する際，ブロックがコンテンツ及び不要部分の最小単位となるよう，入れ子になっているブロックレベル要素を抜いて抽出する．すなわち，図 2 の各ツリー左部分のように，DOM ツリー上の下位ノードにブロックレベル要素が存在しないように抽出する．

処理 3 では，ブロックの一致を判断し，ブログページ集合の中で 1 度だけ出現するブロックをコンテンツとして抽出する．出現回数による閾値を設けないことで，抽出のための閾値を調整する手間を省くことができる [8]．本稿では，このブロックを**コンテンツブロック**と呼ぶ．

処理 4 では，要素識別子をもとに各ブロックに**位置ラベル**を付与し，ブロックのグループ化を行う．2.3 節で解説する．

処理 5 では，付与された位置ラベルと要素名をもとにコンテンツの再抽出を行う．2.4 節で解説する．

処理 6 では，コンテンツのうち，ポストはブログページ集合全てのブログページに出現するが，コメントはいずれかのブログページにしか出現しないという仮説をもとに，ポストとコメントを分離抽出する．まず，位置ラベルのうち，ブログページ集合全てのブログページにおいて，少なくとも 1 つのコンテンツブロックに付与された位置ラベルを準備する（**ポスト位置ラベル**と呼ぶ）．そして，コンテンツブロックのうち，ポスト位置ラベルが付与されているブロックをポストとし，残りのコンテンツブロックをコメントとして抽出する．

2.3 位置ラベルの付与（処理 4）

コンテンツをポストとコメントに分離して抽出するためには，位置と内容に応じてブロックをグループ化する必要がある．しかし，ブロックの内容そのものを捉えるのは困難である．そこで，我々はブロックに付与された要素識別子（Element identifiers）[9] に着目した．要素識別子は要素の内容を示す識別子であることから，内容に応じて付与されていると考えた．

ブロックのグループ化に要素識別子をそのまま用いるとすると，3 点の問題が発生する．1 点目は，全てのブロックに要素識別子が付与されているとは限らないという点である．全てのブロックをグループ化するためには，要素識別子を伝播させる必要がある．2 点目は，要素識別子が与える影響範囲にばらつきが存在するという点である．ポストとコメントの分離抽出では，コンテンツの位置に着目しており，近いブロック同士でグループ化を行いたい．そのため，離れた位置でのグループ化を防ぐ処理を行う必要がある．3 点目は，個々のページ特有の情報を持つ要素識別子が存在する点である．このような要素識別子は，ブログページ集合内で一般化された位置を表現するには適しておらず，除外する必要がある．

ポストはブログページ集合全てのブログページに出現するが，コメントはいずれかのブログページにしか出現しないという仮説は，レンダリングされた結果をもとに立てた．そのため，グループ化もレンダリングされた結果をもとに行うのが望ましい．そこで，問題点 1 を解

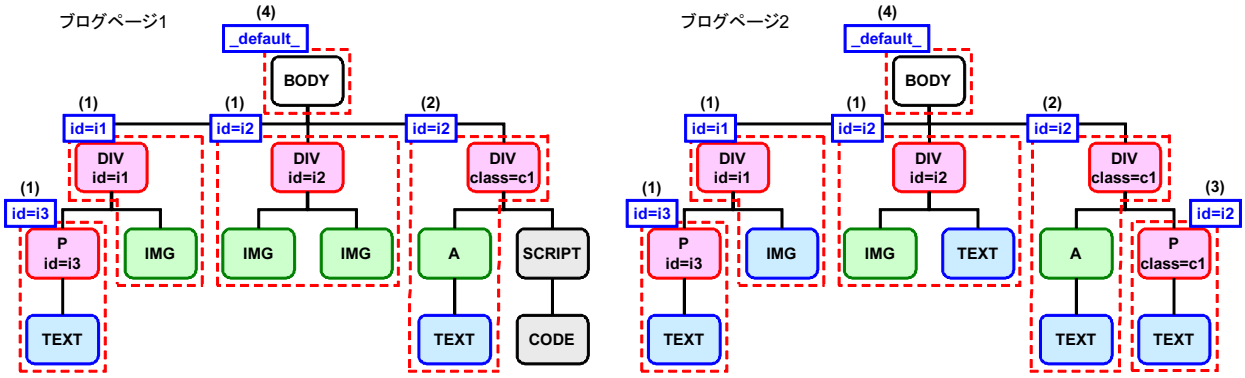

図 2　DOM ツリー（要素識別子は要素内に記載）からのブロック抽出（破線枠部分，計 11 ブロック）及び各ブロックに位置ラベル（実線矩形枠部分）を付与した例．ブロック抽出の際は，SCRIPT, NOSCRIPT, STYLE の 3 要素をブロックに含めない他，BODY 要素は例外的にブロックとして認める．位置ラベル上部の括弧付き数字は，2.3 節に記載した位置ラベル付与規則を表す．たとえば，ブログページ 1 の「class=c1」は，両方のブログページに出現しているが，ブログページ 2 において 2 度出現しているため，規則（1）は適用されず，規則（2）が適用される．

決する伝播手法として，レンダリングされた Web ページにおいて，上から下に伝播させる方法を提案した [7]．すなわち，DOM ツリー上の兄ノードから要素識別子を伝播させる．

以上を踏まえ，処理 4 では，各ブロックにブログページ集合内で一般化された位置ラベルの付与を行う．位置ラベルは，1 ページ中に 1 度のみ出現し（問題点 1 を解決），かつ，ブログページ集合全てのブログページに出現する（問題点 2 を解決）要素識別子を伝播させることにより付与する．位置ラベルの付与規則は，（1）自身の要素識別子を付与する，（2）隣接する兄ノードの位置ラベルを付与する，（3）親ノードの位置ラベルを付与する，（4）標準値（_default_）を付与する，の優先順位をもつものとする．図 2 は，DOM ツリーからブロックを抽出し，位置ラベルを付与した例である．

2.4　コンテンツの再抽出（処理 5）

従来手法では，コンテンツブロックがポストであるかコメントであるかの分離にのみ位置ラベルを利用していた．提案手法では，位置ラベルをコンテンツ抽出にも用いる．

従来手法によるコンテンツ抽出は，ブロックごとにコンテンツ判定を行うため，コンテンツと不要部分を高い粒度で分離できるという特徴を持つ．しかし，各々のブロックの位置を加味せず判断するため，連続するコンテンツブロックの一部を取りこぼすという欠点があった．特に文長の短いコメントはその傾向が高い．図 3 は従来手法による抽出を行ったブログページ[*3]の例であり，コメントの一部を取りこぼしていることがわかる．

処理 5 では，付与された位置ラベルと要素名をもとにコンテンツの再抽出を行う．処理 3 で不要部分と判断

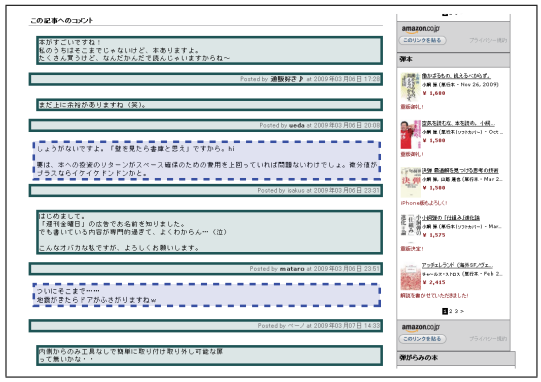

図 3　従来手法には連続するコンテンツの一部を取りこぼすという欠点が存在する（破線矩形部分）．提案手法ではコンテンツブロック（実線矩形部分）の情報を用いて，コンテンツとして取りこぼしたブロックを再抽出する．

されたブロックのうち，コンテンツブロックと同一の位置ラベルと要素名をもつブロックを新たにコンテンツブロックとして抽出する．位置ラベルはブロックをその内容に応じてグループ化するものであり（2.3 節），コンテンツブロックの要素名のみを用いて再抽出を行うよりも，適合率を落とさず再現率を上げられると考える．

3　実験と考察

使用したデータセットは，livedoor Reader[*4]の登録数ランキング上位からブログ形式の 9 サイト（内 6 サイトがブログホスティングサービスを利用していない）を選定し，2009 年 4 月 11 日に収集した計 206 ページである（文献 [7] で用いたデータセットと同様）．ポスト及びコメントの候補となるブロックは 35,216 ブロック存在し，人手によってポストと認められるブロックは

[*3] http://blog.livedoor.jp/dankogai/archives/51185176.html

[*4] http://reader.livedoor.com/

表 1　平均抽出性能の比較（%）

		適合率	再現率	F 値
従来手法	コンテンツ	92.7	88.2	90.4
	ポスト	90.6	82.2	86.2
	コメント	78.5	86.4	82.2
提案手法	コンテンツ	91.1	91.5	91.3
	ポスト	88.9	86.7	87.7
	コメント	83.8	91.3	87.4

提案手法によってコンテンツとして抽出されたブロックの数を N，抽出されたコンテンツのうち正解データに適合していたブロックの数を R，正解データに含まれるコンテンツのブロックの数を C とすると，コンテンツの適合率は R/N，再現率は R/C，F 値は $2 \cdot R/(N+C)$ と計算する．ポスト，コメントの性能も同様に計算する．

2,932 ブロック，コメントと認められるブロックは 973 ブロックであった．使用したデータセットでは 38% のブログページにのみコメントが付いていた．

表 1 に示した実験結果より，提案手法は従来手法に比べて全体的に性能が向上していることがわかる．提案手法によるコンテンツ抽出性能は，従来手法に比べ，適合率には少々の低下が確認できたが，再現率は大幅に向上し，抽出性能（F 値）が向上していることが確認できた．さらに，コンテンツ抽出性能の向上に伴い，コメントの抽出性能が 82.2% から 87.4% に大きく改善した．

図 4 は提案手法による抽出を行ったブログページの例である．このページは，図 3 で用いたページと同一であり，提案手法により従来手法での取りこぼし部分が適切に抽出できていることが確認できる．

4　おわりに

本稿では，従来手法に位置ラベルを用いたコンテンツ再抽出処理を加えることで，ブログのポスト及びコメントの抽出性能を改善する手法を提案した．さらに，日本語ブログサイトを対象とした実験により，提案手法の有効性を示した．特に，コメントの抽出性能の大幅な改善が確認できた．

今後，要素識別子が適切に付与されていないブログページ集合にも適用できるよう，より柔軟に位置ラベルを付与する方法を検討する．また，抽出ルールの自動獲得，ルール自動選択による抽出を検討し，より高速に分離抽出する手法を検討する．そして，本研究の成果をモジュール及びソフトウェアの形で公開し[*5]，Web ページを利用する研究の標準的な手法となることを目指す．

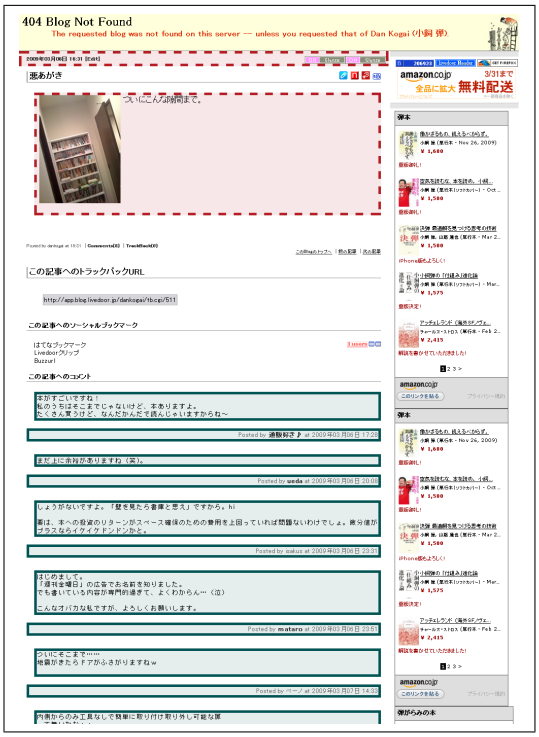

図 4　図 3 と同じブログページを提案手法によって分離抽出したところ，コメント部分の取りこぼしが適切に抽出できた（破線矩形部分がポスト，実線矩形部分がコメントを示す）．

参考文献

[1] 総務省情報通信政策研究所（IICP）．ブログの実態に関する調査研究．http://www.soumu.go.jp/iicp/chousakenkyu/data/research/survey/telecom/2009/2009-02.pdf (cited 2010-01-12), 2009.

[2] 中崎寛之, 川場真理子, 山崎小有里, 宇津呂武仁, 福原知宏．共起語分布の言語間差異を手がかりとする日英対照ブログ分析支援．言語処理学会第 15 回年次大会発表論文集, pp. 701–704, 2009.

[3] Archana Bhattarai, Vasile Rus, and Dipankar Dasgupta. Characterizing comment spam in the blogosphere through content analysis. In *Computational Intelligence in Cyber Security, 2009. CICS '09. IEEE Symposium on*, pp. 37–44, 2009.

[4] Meishan Hu, Aixin Sun, and Ee-Peng Lim. Comments-oriented document summarization: Understanding documents with readers' feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 291–298, 2008.

[5] Kevin Hemenway, Tara Calishain, 村上雅章（訳者）. Spidering hacks—ウェブ情報ラクラク取得テクニック 101 選．オライリー・ジャパン, May 2004.

[6] Donglin Cao, Xiangwen Liao, Hongbo Xu, and Shuo Bai. Blog post and comment extraction using information quantity of web format. In *Information Retrieval Technology: 4th Asia Information Retrieval Symposium*, pp. 298–309. 2008.

[7] 吉田光男, 乾孝司, 山本幹雄．ブログ記事集合を用いたポストとコメントとの自動分離抽出手法の提案．情報処理学会研究報告（データベースシステム研究会）, Vol. 2009-DBS-149, No. 20, pp. 1–8, 2009.

[8] 吉田光男, 山本幹雄．教師情報を必要としないニュースページ群からのコンテンツ自動抽出．日本データベース学会論文誌, Vol. 8, No. 1, pp. 29–34, 2009.

[9] Dave Raggett, Arnaud Le Hors, and Ian Jacobs. The global structure of an html document. In *HTML 4.01 Specification*. World Wide Web Consortium (W3C), 1999.

[*5] http://www.mibel.cs.tsukuba.ac.jp/~m.yoshida/ExtractUniqueBlock/