O-033

リンク特性分析による Web ドキュメント中のノイズデータ除去手法の提案
A Noise Removal Method for Web Mining using Hyperlink Analysis

堤 浩太† 梅澤 猛† 大澤 範高†
Kota Tsutsumi Takeshi Umezawa Noritaka Osawa

1. はじめに

ブログや SNS などの普及に伴い，エンドユーザが発信する CGM(Consumer Generated Media)から集合知(Collective Intelligence)獲得する手段として Web マイニング技術が注目されている．適切なマイニングを行うためには，スパムページなどのコンテンツとして価値のない Web ページを判別し除外するだけでなく，ページの中で本質的に重要な個所(主要部分)を選び出すことが求められる．

一般に，Web ページに含まれる広告などのノイズデータ除去には DOM(Document Object Model)ツリー分析が有効であるが，下位層の処理が粗くなってしまうことが知られている．そこで，本研究では DOM ツリー内のノードにおいてリンク特性分析を行うことで下位層におけるノイズデータ除去を適切に行う手法を提案する．また，教師信号付き機械学習により，ノイズデータにおけるリンク特性モデルを獲得する機能により，ルールベースの既存手法に比べ未知のノイズデータや言語の差異にも適応可能な汎用的手法を目指す．

2. Web ドキュメントのノイズデータ除去

一般に Web ドキュメントからノイズデータを除去する手順は 1）ドキュメントの構造解析によるブロック分割，2）各ブロックの分析評価によるノイズデータ箇所特定，の 2 段階で構成される．Web ドキュメントにおけるノイズデータ除去のイメージを図 1 に示す．

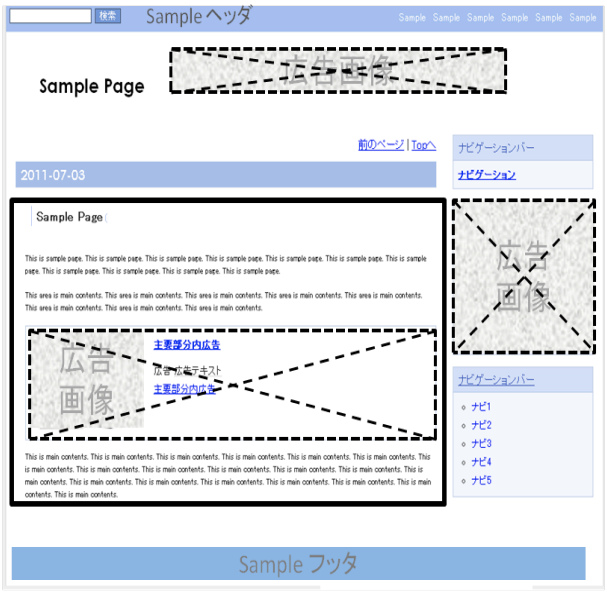

図 1　ノイズデータ除去例

2.1 ブロック分割

ドキュメント全体を構造解析によって分割した各ブロックは，ノイズデータ判別の際に最小単位となるため，適切な粒度で分割する必要がある．分割が粗すぎるとノイズデータが除去し切れずに残ってしまい，細かすぎるとノイズデータを正確に判別できない．

Web ドキュメントのブロック分割手法には，HTML をツリー構造で扱う DOM を利用した DOM-based segmentation[1]や，レイアウト情報によりページを 5 分割（top, down, left, right, center）する Location-based segmentation[2]などがあるが，ノイズデータ除去に適用するためには適切な分割粒度を設定する必要がある．また，DOM に基づいた分割後，フォントサイズや背景色などの類似性を利用して更に分割を行う Vision-based segmentation[3]では，閾値によりブロック分割の粒度を指定できるが，多種多様な Web ドキュメントに適切な固定的設定は困難である．

2.2 ノイズデータ判定

分割された各ブロックに対し，特定の判定基準に従ってノイズデータかどうかの判定を行う．

既存の判定基準としては，ヘッダや振ったなど同一サイト内で共通に記述された部分を利用するもの[4]や，主要部分はページの中央付近に現れるというレイアウト上の性質を利用したもの[5]などがある．しかし，ナビゲーションバーなどの，DOM ツリー階層において主要部分と離れているノイズデータ除去には適しているが，主要部分近辺に入り込んだ広告部分などの除去は困難である．

2.3 ノイズデータ除去の要件

適切なノイズデータ除去には，適切な粒度でのブロック分割と，高い精度でのノイズデータ判定が必要である．細粒度での分割に対応するには DOM ツリーを用いた手法が有効であるが，DOM ツリー解析はトップダウンに分割を進めて途中で打ち切るため，主要部分の間に入り込んだノイズデータの取り扱いに課題が残る．たとえば，図 2 において，主要部分 text1, text2 の間に挿入されたノイズ部分 a を除去する手法が求められている．

主要部分と判定したブロック内に残ったノイズデータ除去に対する試みとしては，広告リストを用意してそれに基づいて削除を行う手法 [6]があるが，未知のノイズデータや言語の違いに弱い．

そこで，本研究では DOM ツリーによるブロック分割を行った後，主要部分を含むブロックに対してリンク特性分析を行うことで高精度なノイズデータ除去を行う手法を提案する．

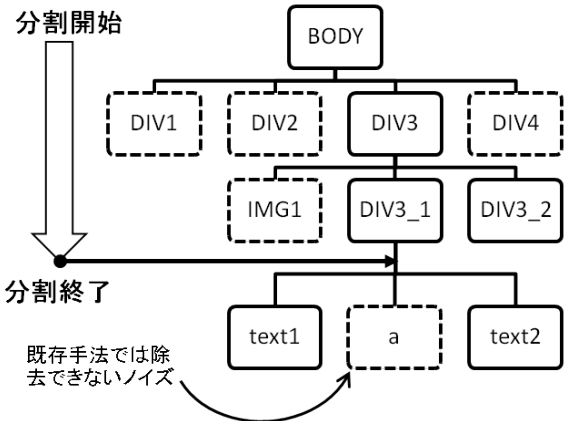

図 2　DOM ツリーの例（点線枠はノイズ）

3. リンク特性分析によるノイズデータ除去

2.3 項で挙げた要件を満たすノイズデータ除去手法を提案する．

ブロック分割には DOM ツリーを利用した手法を採用し，各ブロックに含まれるハイパーリンク記述に関連した特性分析(リンク特性分析)を行うことにより，高い精度でのノイズデータ判定が可能になることが期待できる．また，リンク特性を分析することで，主要部分と判定されたブロックに入り込んだノイズデータ箇所の検出・除去も可能となる．

特性分析にあたっては，教師信号付き機械学習手法を適用することで自動化を図る．予め複数の Web ドキュメントに対して手動で正しい解析結果を与える学習フェーズを設けることにより，運用フェーズではノイズ箇所の判定を自動化することが可能となる．

3.1 リンク特性分析

リンク特性の分析にあたっては，アンカータグ内部の情報だけでなく，その周辺あるいはリンク先 Web ドキュメントの情報までを対象に含める．代表的な特性の例を次に示す．

A) 頻出単語
広告の周辺には「広告」「PR」「Advertise」などのキーワードが記述されていることが多い．ノイズデータに含まれる単語群とその出現頻度をノイズデータ判定の材料とする．

B) URL クエリ
成功報酬型の広告(アフィリエイト)などでは，広告主が広告掲載者を識別するために URL に ID など埋め込むことがある．クエリ変数の文字列や数をノイズデータ判定に活用する．

C) リダイレクト
クリック数カウントを目的として，一旦本来のリンク先とは別ページへと誘導し，その後本来のページへのリダイレクトする手法が多くみられる．ステータスコードやメタタグを解析することで，リダイレクトを検出してノイズデータ判定に活かす．

その他，画像の大きさや開くウインドウサイズなどを基にノイズデータ判別を行う．

3.2 ノイズデータにおけるリンク特性の学習

特性分析を自動化するために，ノイズデータにおけるリンク特性モデルを機械学習によって獲得する．学習フェーズでは，実際の Web ページを閲覧しながらドキュメント中のリンク毎にノイズデータの指定を行い教師信号とする．運用フェーズでは，学習フェーズで得られたリンク特性モデルを用いて，Web ドキュメント中のノイズデータ判別を行う．分類器としては，テキスト分類やスパムメールフィルタに利用される SVM やナイーブベイズ分類器が考えられる．

機械学習を用いることで，画像のノイズデータだけを除去するなどの目的に応じた結果を得るために，学習フェーズにおける教師信号の与え方を工夫することも可能である．また，テキスト情報だけに依存しないリンク特性を利用するため，言語の違いや未知のサイトの広告にも有効に動作することが期待できるなど，事前に広告リストを用意する手法に比べ柔軟で汎用的な適用が可能である．さらに，運用フェーズにおいて，検出に失敗したノイズデータに対する教師信号の付与，学習サンプルとなる Web ページの追加などにより，段階的な精度の向上を図ることもできる．

4. おわりに

本稿では，Web ドキュメントに含まれる広告やナビゲーションなどのノイズデータを除去し，主要部分のみを選び出すために，リンク特性分析によってノイズデータを判別する手法を提案した．DOM ツリーを利用したブロック分割とリンク特性分析を組み合わせることで，従来の DOM ツリーによる分析において課題となっていた主要部分に入り込んだノイズデータの検出・除去を可能とした．実際の分析にあたっては，教師信号付き機械学習により，ノイズデータに関するリンク特性モデルを得ることで実行の自動化とともに判定精度の向上を図った．提案手法は，既存のルールベース手法では困難であった，言語の差異や未知の形式をもったノイズデータに対しても適用可能な汎用性を備える．

今後は，ノイズデータの種類に応じたリンク特性項目や，検出精度の向上に効果的な学習手順などについて，実際に多様な Web ドキュメントに本手法を適用する実験を通して検討していきたい．

参考文献

[1] Chen J, Zhou B, Shi J, Zhang H-J, and Qiu F. “Function-Based Object Model Towards Website Adaptation” Proc. WWW, pp.587-596, 2001.
[2] Kovacevic M, Diligenti M, Gori M, and Milutinovic V. “Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification,” Proc. IEEE International Conference on Data Mining, p.250, 2002.
[3] Cai D, Yu S, Wen J-R and Ma W-Y. “VIPS: a vision-based page segmentation algorithm,” Microsoft Technical Report, MSR-TR-2003-79, p.28 2003.
[4] Yoshida M & Yamamoto M. “Primary Content Extraction from Web Pages without Training Data,” DBSJ Journal. vol.8, No.1, pp.29-34, 2009.
[5] Tsuruta M, Sakai H and Masuyama S. “An Informative DOM Subtree Identification Method from Web Pages in Unfamiliar Web Sites,” IEICE Trans. Information and System, vol.E91-D, no.4, pp.986-989, 2008.
[6] 鶴田雅信, 増山繁, “未知のサイトに含まれる Web ページからの主要部分抽出手法,” 言語処理学会第 14 回年次大会発表論文集, pp.197-200, 2008.