概要

ウェブサイト上で長い記事を掲載するときにはページ送りがよく使われる。ウェブにおけるページ送りとは、長い記事をページ番号を付けていくつかのページに分割して掲載することを指す。1 ページにおさまりきらない記事を分割することで、初めのウェブページの読み込み時間を短縮し、ユーザーは最初に表示されたページの内容を見た後、次ページへ遷移して続きを読むかを判断できる。ページ送りはユーザが閲覧する際は便利だが、ウェブから情報を自動的に獲得する際には、複数のページに分割された記事から元の記事全体を復元する必要がある。ページ送りが使われているウェブサイトに対して元の記事を復元する試みとして AutoPagerize がある。AutoPagerize は、8000 件程度のウェブサイトに対してあらかじめ人手で作成された連結規則が登録されているデータベース Wedata に基づいて機能するため、登録されていないウェブサイトについては元の情報を復元できないという問題がある。

本研究は、大量のウェブページから知識を獲得するウェブマイニングのための基礎技術として、ページ送りによって複数のページに分割された記事を自動的に 1 つに連結することを目的とする。AutoPagerize が人手で連結規則を作成するのに対し、本研究は教師あり機械学習によって次ページへのリンクと主コンテンツを自動検出するモデルを学習し、任意のウェブサイトに対応する点に特徴がある。

本研究の提案手法は、「次ページリンク検出タスク」、「主コンテンツ検出タスク」、「連結タスク」を処理する 3 つのモジュールから構成される。「次ページリンク検出」モジュールは、ページ送りのあるウェブページ内から次のページへのリンクを検出する。ウェブページの HTML ソースファイルから同一ドメインへのリンクを抽出し、機械学習されたモデルを適用して、それぞれのリンクが次のページへのリンクに相当するかを判定する。「主コンテンツ検出」モジュールは、ウェブページの HTML ソースファイルと検出された次ページリンクを入力とし、機械学習されたモデルを適用して、個々のタグが主コンテンツに該当するかを判定する。これら 2 つのモジュールは繰り返し適用される。検出した次ページリンクを辿ることで次ページの HTML ソースファイルを取得し、これを新たな入力として次ページリンクと主コンテンツを再起的に検出する。最終的に獲得された複数の主コンテンツを「連結」モジュールで連結する。最後のモジュールは単純な処理であるため、本研究では最初の 2 つのモジュールの開発、特に次ページリンクと主コンテンツを判定する分類器の機械学習に注力する。

次ページリンクを検出する分類器を学習する際には、素性として、(1)「次」もしくは「NEXT」がタグに含まれるか、(2)「ページ」もしくは「PAGE」がタグに含まれるか、(3) リンクテキストが 1 文字であるか、(4) ウェブサイトにおけるリンクの出現回数、(5) リンクテキスト長、(6) リンクテキスト長のウェブページ全体のテキスト長に対する割合、(7) リンクの URL の長さ、(8) リンクの URL の長さのウェブページ全体に対する割合、(9) 近傍のリンクとの類似性 (LinkSimilarity) を用いた。訓練データは、正例である次ページリンクが、負例であるそれ以外の

リンクに対して圧倒的に少ないため、Synthetic Minority Over-sampling(SMOTE) を用いて不均衡データを是正した後、分類器を学習する。学習アルゴリズムとして、決定木、ランダムフォレスト、Gradient Boosting Decision Tree(GBDT) を用いる。

主コンテンツを検出する分類器を学習する際には、素性として、(1) タグの長さ、(2)DOM ツリーにおけるタグの深さ、(3)HTML ファイルにおけるタグの位置、(4)HTML ファイルにおけるタグの相対位置、(5) ブロックレベル要素に該当するか、(6)HTML タグの種類が明らかに主コンテンツにならないものであるか、(7) 兄弟タグ内にあるテキストの長さ、(8) 兄弟タグ内にあるテキストの割合、(9) 兄弟タグ内の句読点の割合、(10) 兄弟タグのテキスト密度、(11) 兄弟タグ数、(12) 子タグ数、(13) ウェブページ全体のタグ数における子タグ数の割合、(14) 次ページリンクタグからの距離を用いた。次ページリンクタグからの距離の素性は、前述のモジュールで検出された次ページリンクタグと判定対象のタグとの DOM ツリー上の距離を値とする。次ページリンク検出タスクと同様に、訓練データでは、正例である主コンテンツのタグが、負例である主コンテンツ以外のタグと比べて圧倒的に数が少ない。そのため、SMOTE を用いて正例を増加させた後、負例をランダムに減少させて、完全に均衡した訓練データを作成した後、分類器を学習する。学習には決定木、ランダムフォレスト、GBDT を用いる。

提案手法の評価実験について述べる。データセットとして Wedata に登録されたウェブサイトを利用する。簡易なルールに基づくベースライン手法と提案手法の性能を比較する。評価基準として精度、再現率、F 値を用いる。次ページリンクの検出モデルについては、3 つの機械学習アルゴリズムのうちランダムフォレストが最も性能がよく、精度は 0.818、再現率は 0.692、F 値は 0.750 であった。ページ送りの特徴を特に考慮した LinkSimilarity 素性によって F 値が 0.027 ポイント向上した。主コンテンツの検出モデルについては、精度は 0.588、再現率は 0.555、F 値は 0.571 であった。また、ページ送りの特徴を特に考慮した「次ページリンクからの距離」の素性を導入することで F 値が 0.07 ポイント向上した。これら 2 つの提案手法の結果は、それぞれ、ベースライン手法よりも顕著に高く、機械学習によってページ送りされたウェブサイトから主コンテンツを検出する提案手法のアプローチが有効であることが確認された。