ウェブページにおける非コンテンツ領域の検出に関する研究

中村 達也 (510073)

北陸先端科学技術大学院大学 情報科学研究科

2007 年 2 月 8 日

キーワード: WWW、非コンテンツ領域、チャンキング、学習、情報検索.

近年のWWWの普及により、ウェブページから様々な情報を得る機会が多くなった。また、情報検索やウェブマイニングなどウェブを対象とした研究も多い。しかし、ウェブページには有用な情報を含むコンテンツ領域とそうでない領域 (非コンテンツ領域) が混在している。非コンテンツ領域とは、例えば広告、目次、検索フォームなど、特に有用な情報を含まない領域である。このような非コンテンツ領域は、情報検索、ウェブマイニングなど様々なアプリケーションに対して悪影響を及ぼすことが考えられる。例えば、情報検索において、非コンテンツ領域を検出し、その領域に含まれる単語は索引語としないことで、索引語付けや検索の処理時間を短縮できる。また、ページの内容とあまり関係のない非コンテンツ領域の中のキーワードとマッチすることで不適切なページが検索されることを妨げることにより、情報検索の精度向上が期待できる。このように非コンテンツ領域の自動検出は多くのウェブアプリケーションに対して有益である。

本研究では、様々なウェブアプリケーションに対する前処理として、ウェブページの非コンテンツ領域を検出する手法を提案する。ページのどのような部分を非コンテンツ領域とするかはアプリケーションによって異なるが、本研究では情報検索を想定する。本手法では、HTML タグで分割されたテキストに対して、コンテンツ領域のテキストか非コンテンツ領域のテキストかを判別するラベル付けを行う。一般に非コンテンツ領域は複数のテキストから構成されることから、IOB2 モデルのチャンキングによって非コンテンツ領域を検出する。すなわち、非コンテンツ領域であるとラベル付けされたテキストをまとめ上げることで 1 つの非コンテンツ領域を検出する。ラベル付けを行うチャンキングのモデルは、正解として非コンテンツ領域があらかじめ付与されたウェブページの集合から学習する。学習には、学習アルゴリズムとして Suppor Vector Machine を採用した汎用チャンキングツール YamCha を利用した。

次に、学習に用いた素性について述べる。実験用データとは別に集めた 21 ページのウェブページを調査し、非コンテンツ領域の検出に有効であると思われる手がかりを見つけた。その結果、以下の 8 つを素性とした。(1) 非コンテンツ領域に現れやすいキーワードが含まれるか、(2) テキスト長、(3) テキストに動詞、形容詞が含まれるか、(4) テキスト

が内部リンクか外部リンクかリンク以外か、(5)DOM ツリー上で近傍にある HTML タグ、(6) 直前のテキストと比べたときの DOM のパスの深さの変化、(7)<table>タグ内のテキストの平均長、(8)<table>タグ内のリンクの割合。また、非コンテンツ領域に現れやすいキーワードは学習データから自動的に選別する。具体的には、キーワードの出現回数が多い、キーワードが非コンテンツ領域に現れる確率が高い、キーワードが非コンテンツ領域内に出現するページのドメインの異なり数が多いことという条件を満たす名詞を非コンテンツ領域を示唆するキーワードとして選別する。ドメインの異なり数を条件としたのは、特定のウェブサイトの非コンテンツ領域のみに頻出するキーワードを誤って選択しないようにするためである。

非コンテンツ領域の検出手法の有効性を確認するために、ウェブディレクトリから 781 ページをランダムサンプリングして実験用データとした。これらのウェブページに人手で非コンテンツ領域をマークアップした。このデータを用いて 5 分割交差検定によって学習とテストを繰り返し、提案手法の有効性を確認した。

提案システムのテキストに対するラベルの正解率は 0.769 であった。一方、全てのテキストに対してラベル O(コンテンツ領域) を与えるベースラインシステムの正解率は 0.698 であり、提案手法はベースラインを大きく上回っていることがわかった。非コンテンツ領域検出の精度は、領域単位で約 3 割、テキスト単位で約 7 割であった。このことから、提案システムが、非コンテンツ領域をその範囲まで完全に検出することは難しいが、部分的にはある程度検出できていることが分かる。しかし、これらの精度は十分高いとは言えないので、更なる手法の改良が必要である。また、情報として有用なコンテンツ領域が誤って非コンテンツ領域と誤判定されているテキストの割合は 7% であり、自動検出された非コンテンツ領域をページから除去してもウェブページにおける有用な情報を大きく失わないことがわかった。また、チャンキングに用いた素性の有効性を検証するための実験も行った。その結果、有効な素性は『(1) 非コンテンツ領域によく現れるキーワード』や『(2) テキスト長』であった。一方、有効でなかったのは『(6) 直前のテキストと比べたときの DOM パスの深さの変化』であった。