2V－2

Webの本文部分抽出技術を用いたコンテンツの更新日時推定の検討

日本電信電話株式会社　NTTサイバーソリューション研究所

栗島聡哉　森大二郎　竹野浩　稲垣博人

1. はじめに

インターネットの発達に伴い急速に増大した情報の検索を可能にするために、ロボット型検索エンジン[1]と呼ばれるインターネット上の情報検索サービスは常時インターネット上のコンテンツの収集を行っている。

ロボット型検索エンジンが新鮮な情報の選別やトレンド分析を行うためには、収集したコンテンツの更新日時を正確に知ることが重要である。

本研究ではコンテンツの更新日時の推定を精度よく行うために、メニューや広告、最新記事へのリンクなどを削除し、コンテンツの本文部分の抽出を行う技術の開発を行った。またこの抽出技術の精度の評価、更新日時の推定の検討を行ったので報告する。

2. 更新日時推定の従来手法

更新日時の推定を行う手法としてはWebサーバが返す更新日時[2]の情報から推定する手法がある。しかし、更新日時の情報を返さないWebサーバは数多く存在するので、すべてのWebコンテンツの更新日時をこの方法で推定することは不可能である。

そのため周期的に収集を行い収集データが変化していると収集の間に更新されたと推定する手法が考えられる。

しかし、最近は図1で示すような広告やメニュー、最新のニュースへのリンクなどを自動的に追加するサイトが増加しており、これらの中には本文部分は変化しなくても広告などが動的に変化するものが多いため、単に収集データの差分を取るだけでは更新日時の正確な推定を行うことは不可能である。

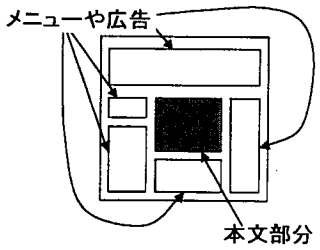

図1　動的に作成されるコンテンツ

3. 本研究でのアプローチ

自動的に広告や最新ニュースなどの情報を追加されるようなWebコンテンツの更新日時の推定を行うためには、本文部分を抽出する必要がある。

本研究ではコンテンツ中の自動的に追加された部分を識別して削除することにより本文部分を推定し、コンテンツの更新日時の推定を行う。

3.1. 本文部分抽出手法

ニュースなどのコンテンツ系サイトでは、サイト内の複数のページに、同一のメニューや広告が反復して現れると考えられる。本手法ではそのような同一サイト内での重複部分を検出し削除することで本文部分の抽出を行う。

重複部分の検出を行う手順は以下に示す。

1. ブロックに分割
⇒ 収集してきたhtmlファイルを「<td>○○○○</td>」等のようにタグで囲まれているものや、「○○
○○」の様にタグで区切られている文字列を一つのブロックとして抜き出す。(図 2(a))
2. ブロックの出現回数をカウント
⇒ 同じWebサーバ上から収集したすべての

Update time estimation method for web pages using body part extraction
Toshiya Kurishima,Daijiro Mori,
Hiroshi Takeno,Hirohito Inagaki
NTT Cyber Solutions Laboratories,
NTT Corporation

html ファイルをブロックに分割し、その文字列の出現回数をカウントする。(図 2(b))

3. 出現回数が多いブロックを削除

⇒ 出現回数が二つ以上のブロックで、二つ以上連続しているブロックを削除する。(図 2(c))

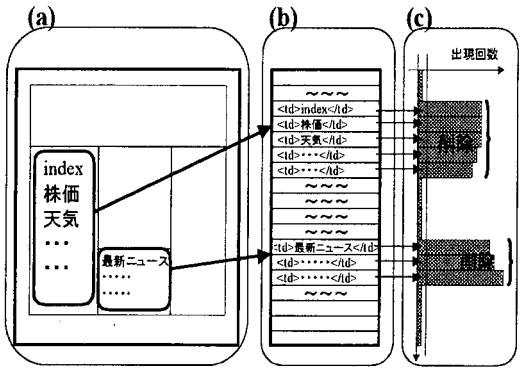

図 2 重複部分を検出

4. 実装、評価

3.1 で示した手法を実現するプログラムを作成し、これを実際の Web コンテンツに適用することにより正確に本文部分を抽出できたかどうかの評価を行った。

本文以外の情報を抽出すると更新していなくても更新をしたと間違う可能性がある。逆に抽出した中に本文が全て含まれていなければ更新したのに更新されていないと間違う可能性がある。

そこで、前者を適合率、後者を再現率という尺度で評価し更新時間の正確な推定が可能か検討を行う。

再現率と適合率は以下のように定義する

$$再現率 = \frac{抽出結果の中で正解に含まれる文字数}{正解全体の文字数}$$

$$適合率 = \frac{抽出結果の中で正解に含まれる文字数}{抽出結果の全文字数}$$

4.1. 本文抽出精度

無作為に抽出した 5 サイトとコンテンツ系の 3 サイトを 50 ページずつサンプリングし、本文だと認識した正解部分と、本文抽出技術により抽出を行った部分との比較を行った。

表 1 にサイト毎に評価を行った適合度と再現率のデータを示す。

表 1 本文抽出技術の精度

	適合率	再現率
サイト 1	99.9%	97.7%
サイト 2	99.8%	99.6%
サイト 3	99.9%	95.2%
サイト 4	99.9%	98.3%
サイト 5	100.0%	98.0%
コンテンツサイト 1	99.5%	82.9%
コンテンツサイト 2	99.6%	91.6%
コンテンツサイト 3	99.4%	97.7%

4.2. 考察

本文抽出の適合率は 99%を越えている。これにより、動的に作成される部分を誤って抽出することで古い情報を新しい情報だと誤認することなく更新時間の推定を行えることがわかった。

また、再現率は 80%から 90%の精度で高い数値を得られているが適合率ほどではなく、更新された情報を古い情報だと誤認する可能性があると考えられるが、実際に抽出されなかった部分を調査すると、抽出されなかったブロックは記号や「index」や「動作環境」等の文字列で記事の本筋とは関係のない部分がほとんどであり、更新時間の推定への影響が小さいことがわかった。

5. まとめ

本研究では、動的に内容が変化するようなコンテンツ系サイトの更新日時の推定を行うために、本文部分を抽出し更新日時を推定が行える手法を考案し実装した。本手法を実際の Web コンテンツに適用し、高精度に本文部分が抽出することが可能で、コンテンツの更新日時を推定することが可能であることを確認した。

参考文献

[1] Oliver A. McBryan: "GENVL and WWWW: Tools for Taming the Web","Proceedings of the first International World Wide Web Conference",1994

[2] R.Fielding, J.Gettys, J.Mogul, H.Frystyk, L.Masinter, P.Leach, T.Berners-Lee:"Hypertext Transfer Protocol -- HTTP/1.1", RFC2616, 1999

[3] 大久保ほか: "話題が混在するテキストからの話題セグメントの抽出方式", 情処第 57 回全国大会, 1V-5(1998), p.3-211－3-212

[4] 上坂, 尾関: パターン認識と学習のアルゴリズム、文一総合出版, 1990