# Webページの更新日時推定技術の検討

4X−04

栗島聡哉*　森大二郎†　竹野浩*　稲垣博人*

*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

†株式会社エヌ・ティ・ティ エックス

## 1. はじめに

ロボット型検索エンジン[1]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Webページの更新日時を正確に知ることが重要である。

更新日時を推定する方法としては、Webサーバから取得する方法やWebページの内容の変化から推定する方法が考えられる。しかし、Webページに広告や最新ニュースのヘッドラインなどを自動的に挿入するサイトが最近増加しており、このようなサイトでは単純にWebページを比較するだけでは更新日時の推定を正確に行うことができない。

本研究ではWebページ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、最新ニュースのヘッドラインなどを削除し、Webページの本文部分のみを抽出し、これを過去のものと比較する更新日時推定技術の開発を行った。また更新日時推定の従来手法と本技術の比較検討を行ったので報告する。

## 2. 従来手法による更新日時推定

従来手法として、Webサーバから更新日時を取得する手法と、収集したWebページの内容の変化から更新日時を推定する方法がある。

### 2.1. Webサーバのヘッダ情報から推定

WebサーバからWebページを取得する際に用いられる通信手順であるHTTP[2]では、Webページの最終更新日時を示すLast-Modifiedフィールドをページの内容と共に送信することが定められており、これを用いればWebページの更新日時を知ることができる。しかし、この動作は必須では無いこと、Webサーバが必ずしも正しい値を返す保証が無いことから、信頼性に疑問が残る。そこで、無作為に抽出した76サイトのWebサーバからページを収集し、内容の変化とLast-Modifiedフィールドの関係を調査した。結果を表1に示す。

**表1 ヘッダ情報による更新日時の推定**

| | |
|---|---|
| すべてのページ | 122399ページ |
| Last-Modifiedフィールドが変化している、又は返さないページ | 35476ページ |
| 内容が更新されているページ | 7000ページ |

これより、8割近くのWebページのLast-Modifiedフィールドから得られた更新日時は正確でないことがわかる。

### 2.2. Webページの内容の変化から推定

定期的に同一のWebページを収集すれば、内容の比較を行って差異が生じていれば収集の間で更新が行われたと推定することができる。

しかし、最近では自動的に広告やニュースのヘッドラインなどを挿入するサイトが増加しており、受信データの単純な比較では更新されたかどうかを正確に推定出来ない可能性がある。

そこで、2.1で収集したサイトの中でLast-Modifiedを返さなかった5サイトのWebページからデータが更新されているページ、HTMLタグの部分を取り除いたテキスト部分が更新されているページ、記事自体が更新されているページの数を調査した。結果を表2に示す。

---

Update time estimation method for web pages
Toshiya Kurishima*,Daijiro Mori†,
Hiroshi Takeno*,Hirohito Inagaki*
*NTT Cyber Solutions Laboratories,
NTT Corporation
†NTT-X, Inc.

**表 2 内容の変化による更新情報の推定**

| | |
|---|---|
| データが更新されているページ | 5432 ページ |
| テキスト部分が更新しているページ | 876 ページ |
| 実際に更新されているページ | 359 ページ |

このように、単純に受信データを比較するだけでは 9 割以上、テキスト部分を比較するだけでは 6 割以上の更新情報の推定が出来ないことが判明した。

## 3. 本文抽出技術を用いた更新情報推定

2.1 で示すように多くの Web ページで更新日時が正確に取得できず、また、単純に内容を比較するだけでは更新情報の推定が正確にできないことが判明した。

そこで推定精度を向上させるためには、Web ページから広告や最新ニュースのヘッドラインなどの情報を削除し本文部分を抽出し、抽出した部分を比較することで更新されたか否かを判別する必要がある。

そこで、Web ページ中の自動的に追加された部分を削除し本文部分を抽出する手法[3]を用いて、更新情報の推定を行った。

### 3.1. 実験

本手法による更新情報の推定は以下のようにして行った。

まず、2.1 で収集したサイトの中で Last-Modified を返さなかった 5 サイトの Web ページから本文部分を抽出し、本文部分が変化しているページを抽出する。

2.2 の示した HTML タグの部分を取り除いた部分で比較する手法と、本手法を用いて更新されていると判断したページを比較し調査した。また、更新したと判断したページが正確に識別できているかどうか確認し、以下の結果が得られた。

**表 3 本手法による更新情報の推定**

| | 従来手法 (ページ) | 本手法 (ページ) | 不正解 (ページ) |
|---|---|---|---|
| サイト A | 365 | 102 | 0 |
| サイト B | 120 | 120 | 0 |
| サイト C | 60 | 60 | 0 |
| サイト D | 263 | 22 | 0 |
| サイト E | 65 | 55 | 0 |

以上の結果、更新されているかどうかを高い精度で識別できていることを確認した。

### 3.2. 考察

本手法を用いれば、本文が更新された Web ページを高い精度で識別することが可能なことが判明した。

今回の実験では間違って識別したページはなかったが、ニュースのヘッドラインだけが含まれるような Web ページや同一のページが同じサイト内に含まれるような場合では本手法は正確に更新されたかどうかを識別するのは難しいと考えられる。

また、更新日時の推定を正確に行うためには高い頻度で収集を行う必要性があるが、単位時間あたりに収集できる Web ページの量は決まっているため、高い頻度で収集するためには収集する対象を絞り込む手法や、更新されたページだけではなく新しいページを効率的に収集することが可能な手法を開発する必要があると考えられる。

## 4. 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 ページに適用し、本文部分を抽出し、Web ページが更新されたかどうかを高精度で推定することが可能であることを確認し、高い精度で本手法が更新日時を推定するために有効であることがわかった。

## 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] 栗島, 森, 竹野, 稲垣: "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", 情処第 63 回全国大会, 2V-2(2001), p.3-29－3-30