

Webからの情報抽出における高速文書更新判定アルゴリズムに関する研究

池田 利夫[†] 太田 弘[†][†] 関西電力 研究開発室 電力技術研究所 ITサービス研究室

1 はじめに

膨大なWeb情報において、刻々とWeb掲示板等に書き込まれ表示される膨大な情報から新着情報を収集するにあたり、その書込み文章等が、新規文書であるか、或いは、過去収集・表示済みの文書であるかの更新判定処理を如何に効率的に行うかが重要となる。

RSSによるWeb掲示板等からの新着通知だけではなく、通常のHTMLからの情報収集も可能でなければ精度の高い情報収集はできない。

また、この更新判定処理には、書込者等の著作権に留意し、合法的な処理でなければならない。

従来の文書更新判定処理手法においては、その判定誤り率は非常に低いものであるが、膨大なWebサイト上の大量文章を1時間間隔など短時間周期で、高速に判定処理することは困難であった。また、更新判定処理を行うために、記述文章から、全文あるいは一部文書を利用許諾なしにコピーするなど、場合によっては著作権に抵触する方式も見受けられた。

本研究では、独自に考案した文書更新判定アルゴリズムにより、膨大なWebサイト上の大量文章を高速に且つ、合法的に更新判定を行うプロトシステムを開発し、その検証および評価を実施した。

2 従来の文書比較手法に関する考案

その文書が同一であるかどうかの文書更新判定方法には様々な手法が存在する。以下、特徴を記述する。

2-1 全文比較による判定方法

文章の全文字をそのまま比較する方法である。非常にポピュラーな手法であり、誤判定率（更新文書を更新なしと判定する確率）は「0%」であるが、文書量が増大すると処理速度が遅く、また、著作権に抵触する可能性がある。

2-2 ハッシュ値比較による判定方法

文章の全文字からハッシュ値を取得し、これの比較により文書更新判定を行う方法である。その誤判定率（衝突確率）は、無視できるほど非常に低い。取得したハッシュ値の文字列は一般的に意味がない配列となることから著作物とはならず、また、ハッシュ関数は一方向関数のため、取得したハッシュ値から原文を復元することはできないため、著作権上も合法である。

2-3 形態素解析による判定方法

形態素解析により、品詞レベルで文書の合致度を判定する方法である。柔軟な比較が可能で、誤判定率も非常に低くすることが可能であるが、解析処理に多くの時間を要することから、新着情報の高速判定に利

用することは困難である。また、品詞分解した文書であっても、その文字列が著作物となる場合、著作権に抵触する可能性がある。

2-4 時間情報取得による方法

HTMLの時間情報を取得して更新判定処理を行う方法である。更新判定処理速度は速いが、時間情報を持たない場合は、更新判定そのものが不可能であり、2文書の更新時刻が異なっても同一文章が掲載されている場合もあり、誤判定率は高い。

2-5 暗号文比較による方法

文章を暗号化し、暗号化された文字配列で更新判定を行う手法である。誤判定率は低いが、暗号化処理に多くの時間を要する。また、何らかの方法で復号化可能であれば、場合によっては著作権に抵触する可能性もある。

3 文字抽出による方法

3-1 部分文字抽出法による課題解決

前項で述べた従来の文書比較による方法においては、Web上の大量文書を、著作権に留意し、高速に更新判定処理を行うことは困難であった。その困難さは、以下に起因する。

① 非常に高精度な判定処理

② 判定対象データの著作物性

①の課題解決のためには、可能な限り、少ないデータでの判定処理アルゴリズムを考案する必要がある。全文比較による更新判定処理では、その誤判定率は「0」であるが、文書量が多くなるほど多くの処理時間を要する。実用上、誤判定率「0」である必要性はなく、例えば、Web検索処理において、何万年に1度の誤判定であるならば問題とはならない。従って、更新比較対象データを、全文ではなく、抽出された部分文字とすることで、処理速度を向上することが可能となる。

②の課題についても、①の課題解決方法で解決できる。本文から著作物とならないよう文字を間引き抽出、配列する。

3-2 類似文書判定（裕度の必要性）

また、Web掲示板で記述されている文章やメール文書の特徴の1つとして、前述の文章を引用し、それに対してコメントを記述することがある。

(例)

>昨日はどうもありがとう。25日の打ち合わせの件、御社へ10時に伺います。
>こちらは、私と他担当者1名の2名です。資料は、本日中午にメールで事前送付いたします。
>なにかありましたら連絡願います。
了解しました。

このような場合、>の元文章とそれに対してのコメントを記述した文「了解しました」とは、ある文字数

A study on high-speed judgment algorithm for document update in extracting information from Web

[†] TOSHIO, Ikeda (ikeda.toshio@a3.kepco.co.jp),

HIROMU, Oota (ota.hiroshi@b5.kepco.co.jp)

IT Service Research Division, Power Engineering R&D,
The Kansai Electric Power Co., Inc.

差内であれば、同一文章と判定するよう裕度を持たす。
元文書(>)ブロックをコメントを含めた全文の中で、スライドしながら、先に述べた部分文字抽出法により更新判定処理を行う(図1)。

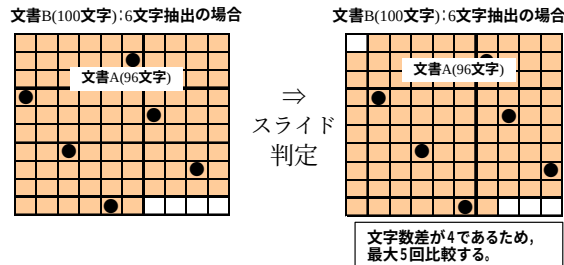


図1：類似文書判定方法（裕度）

3-3 文字抽出アルゴリズムと抽出文字パターン

実際の研究システムでのアルゴリズム(図2)では、裕度(カバー率):95%(または、30文字数差以内)とし、抽出文字配列の比較を行っている。

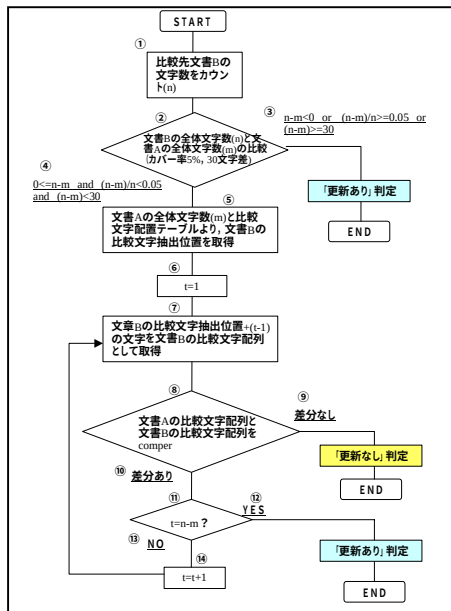


図2：更新文書判定アルゴリズム

また、抽出文字パターン(図3)については、数文字程度の文字抽出を行うが、抽出文字配列が著作権を保持しないよう間引き抽出し配列する。

全体文字数	抽出文字数	文字抽出位置(●:抽出)
1	1	1●(全文字抽出)
2	2	1●2●(全文字抽出)
3	3	1●2●3●(全文字抽出)
4	4	1●2●3●4●(全文字抽出)
5	4	1●2●3●4●(全文字抽出)
6	4	1●2●3●4●(全文字抽出)
7	4	1●2●3●4●(全文字抽出)
8	5	1●2●3●4●5●(全文字抽出)
9	5	1●2●3●4●5●(全文字抽出)
10	6	1●2●3●4●5●6●(全文字抽出)
11	6	1●2●3●4●5●6●(全文字抽出)
12	6	1●2●3●4●5●6●(全文字抽出)
13	6	1●2●3●4●5●6●(全文字抽出)
14	6	1●2●3●4●5●6●(全文字抽出)
15	6	1●2●3●4●5●6●(全文字抽出)
16	6	1●2●3●4●5●6●(全文字抽出)
21	6	1●2●3●4●5●6●(全文字抽出)
26	6	1●2●3●4●5●6●(全文字抽出)

図3：文字抽出パターン

3-4 誤判定率

判定と事象の関係(図4)で、実用上影響が大きいと考えられるのは、(判定):文書A=文書B、(事象):文書A≠文書BのERR(誤判定)である。これは、実際は新規文章であるにもかかわらず「更新なし」と判定する確率であり、重要文章が通知されない可能性が発生する。

判定 事象(実際)	文書A=文書B	文書A≠文書B
文書A=文書B	—	ERR
文書A≠文書B	ERR	—

図4：判定と事象

この場合の誤判定率 p は以下のとおり容易に計算できる。

$$p = (\text{スライド回数}) \times (\text{抽出文字合致確率} \times \text{残文字(一部)非合致確率})$$

これにより、例えば、1時間周期で大量文書の更新文書判定処理を行った場合、実用上、無視できるほどの小さな誤判定発生率となる(1回/(1.9×10^{-6} 年))。

また、文字数の少ない文章であっても、多い文章であっても、数文字程度の抽出では、殆ど誤判定率(10^{-16} 程度)に差がなく、ほぼ同一精度での大量文書の高速判定処理が可能となる。

4 検証結果

著作権上問題ない2方法—文字抽出法およびハッシュ値比較法について、処理速度の比較検証実験を実施した。その結果(図5)、文字抽出法はハッシュ値比較法に対して非常に高い性能結果を示した。

	文字抽出法 (ms)	ハッシュ値比較法 (ms)	速度向上比 (倍)
10文字	10.8	458.1	42.4
100文字	15.3	481.8	31.5
1,000文字	24.0	642.1	26.8

(1,000回繰返判定の3回平均値)

図5：検証結果

5 まとめと今後の課題

本研究により、Webサイト上の大量文章からの新着情報収集において必要となる、高速文章更新判定処理を効率的に、かつ、合法的に実現することが可能となり、非常に短時間サイクルで、大量の新着情報を通知することが可能となる。

今後は、実運用において、検索キーワード数、検索サイト数等を考慮した、最適な巡回検索サイクル等を検討する必要がある。

参考文献

- [1] 山田誠二, “Web更新モニタリング,” 情報処理学会誌 Vol.44 pp.713-719, July, 2003.
- [2] 井上俊一他, “Yahoo!Search Technology(YST)と、検索分野におけるYahoo!JAPANの戦略,” 情報処理学会誌 Vol.46 pp.988-994, September, 2005.