3D-6

インターネットからのマーケティング情報収集システムに関する開発研究

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

近年，ブログ・ＳＮＳ・掲示板などインターネット（Ｗｅｂ）上での発信文書の数は増加の一途を辿っている。それら発信された情報内容は，例えば，ある製品についての消費者側の率直な意見をタイムリーに反映したものなどが多々存在する。企業側にとって，このような顕在化されたサイレントマジョリティ集団から，有益な製品評判情報等を効率的に収集することが，マーケティング活動の成否を左右する。

本研究では，刻々とＷｅｂ掲示板等に書き込まれる膨大な情報から新着情報を収集するにあたり，１ページ（１ホームページ）中に記載されている複数文書（記事）等から，収集目的の文書（記事）と異なる文書の更新判定・通知を避けるため，文書群を同一話題の適切なサイズにブロック分割する手法を考案した。また，その書込み文章が，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，記述者の著作権を保護しながら合法的かつ効率的に行う手法を考案した。これら手法に関し，プロトシステムを開発し，実用化検証を実施した。

２　Ｗｅｂ文書分割手法

ホームページ中に記述されている文書は，通常，複数の話題が記述されている。このような場合，適切な（話題ごと）に文書をブロック分割する必要がある。

一般的には，ＨＴＭＬタグのシングル改行タブ「
」，またはダブル改行タグ「

」などにより，文書を分割している場合が多い。シングル改行タブ「
」の場合，
で文書が改行されているため，
から
までを一つのブロック単位と認識する。

また，ダブル改行タグ「

」などが存在する場合，文書の前（後）に空行が挿入されることから，空行から空行までを１つの文書（話題）として認識する。また，「●」や「◆」などの行頭文字については，ニュースの見出しなどの行頭に多く用いられるが，この文字を用いることで，ニュースの見出しごとに，文書を分割することができる。

しかし，このシングル改行タブ「
」でブロックに分割する方法では，ニュース見出しのような行単位（箇条書き）ごとに異なる話題が記述されている場合は有効であるが，解説文など文書長が長い文書の行末を整形する場合に用いられる場合があり，この場合，一つの解説文を行単位に分割してしまう可能性がある。

また，ダブル改行タグ「

」でブロックに分割する方法では，分割されたブロックの中に，ニュースの見出しなどが含まれる場合がある。この場合，１つのブロックに異なる話題の文書が混在し，その異なる話題の更新判定結果を通知してしまう可能性がある。ニュース見出しなど箇条書き文書を，行頭文字「●」や「◆」などで判定する方法では，行頭文字は様々な文字が使用されるため，その全てを網羅的に判定することは困難である。

これらを解決するため，ニュース見出しなど箇条書き文書については，ブロックの全文字数におけるハイパーリング（下線）の文字数含有率が高いことに着目する（「パレスチナの武装集団がガザでＡＰ通信のカメラマンを拉致（ロイター） - 17 時 26 分」など）。この特長を利用し，パイパーリンクの文字数含有率が一定値以上の場合は，ニュース見出しなどであると判定し，シングル改行タブで分割する。また，Ｗｅｂサイトの最下位行にハイパーリンクが細切れに羅列している場合が存在する（「アメリカ - カナダ - ブラジル - メキシコ - アルゼンチン - スペイン語 など）。この場合，ニュース見出しなどの場合と同様の条件で分割すると，シングル改行タブでの分割となり，その行中で，意味のことなる単語（文書）を一つのブロックとして分割してしまう。これを回避するため，このようなハイパーリンクの文字数含有率は，先のニュース見出しなどの含有率よりも一般的に高いことに着目し，ハイパーリンク文字数含有率が非常に高い場合は，リンクタグ「<a href>」で分割する。以上の手法により，適切に文書を分割することが可能となる。

３　文書更新判定手法

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，全文比較法，ハッシュ値比較法，時間情報取得法など，多様な手法が存在する。しかし，これら手法は，著作権に留意せず，文書をディスクにコピーしたり，更新判定処理時間が長時間掛かったりするなどの課題があった。

今回の手法（文字抽出法）では，更新判定する文書を全文ではなく，数文字程度の抜き取りサンプリング文字を対象とする。これにより，文書全文をディスクにコピーすることなく（著作権を保護しながら），判定処理することが可能となる。また，少量のサンプリング文字による比較判定処理のため，大量文書の更新判定処理時間を短縮することができる。この文字抽出法を行う際に懸念されるのは，その精度となる。従来の全文比較法による更新判定処理では，全ての文字について更新判定を行うため，その更新誤りの確率は 0%であるが，文字抽出法における誤判定確率は，サンプリング文字のため，少なからず発生すると考えられる。この文字抽出法における誤判定確率*(p)* は，以下のとおり算出することができる。

$$p=(\text{ｽﾗｲﾄﾞ回数})*(\text{抽出文字合致確率})*\text{残文字(一部)非合致確率}$$

これにより，例えば，1 時間周期で大量文書の更新文

Development research about a marketing information collecting system from the Internet

† TOSHIO,Ikeda(ikeda.toshio@a3.kepco.co.jp),
HIROSHI,Ota(ota.hiroshi@b5.kepco.co.jp)
IT Service Research Division,Power Engineering R&D,
The Kansai Electric Power Co., Inc.

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/(2.4*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，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　評価システムの構築と検証

上記に述べてきた，Ｗｅｂ文書分割手法と文書更新判定手法のアルゴリズムを組み込んだプロトシステムを構築した。このシステムは，マーケティング情報を収集したいＷｅｂサイト，検索キーワード，類語等を登録しておけば，巡回ロボットが 1 時間ごとに 24 時間，登録Ｗｅｂサイトを巡回監視し，文書更新が発生した場合，パソコン画面やメールで更新内容確認を行う。また，検索情報は蓄積され，統計処理により検索傾向をグラフ化する（図 1）。システム環境は，サーバに当システムを組み込み，インターネットで接続されたパソコンから，各々が欲しい情報を取得するための設定をブラウザで行う。パソコン側に特別なソフトは必要無い（図 2）。

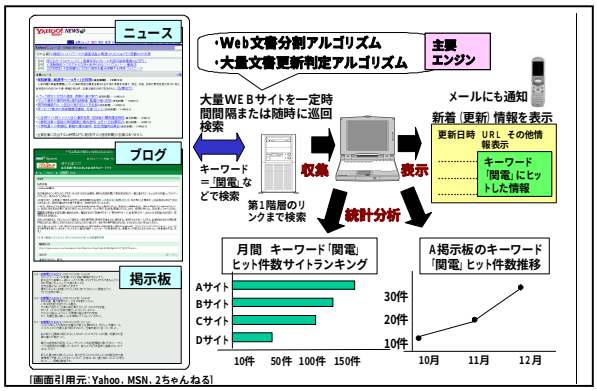

図 1：システム概要図

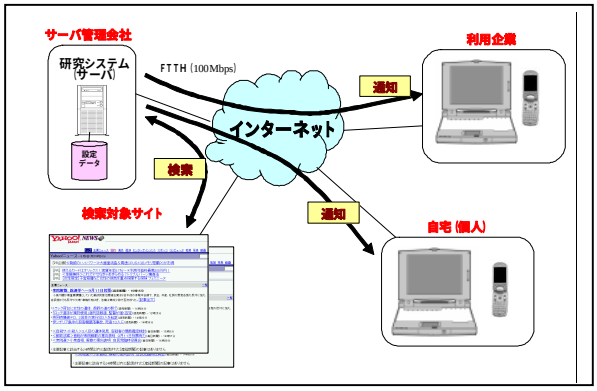

図 2：システム環境構成図

検索条件は，ＵＲＬごとに条件式登録する。

検索結果は，更新日時，ヒットしたキーワード，ヒットしたＵＲＬ（サイト名）などが，一覧表示される（図 3）。検索結果は，Ｗｅｂ上での表示，パソコンメール，携帯メールへの通知が可能であり，通知時間は 1 時間ごとに 24 時間まで設定可能としている。また，検索結果の傾向については，時系列統計（キーワード検索，比較サイト指定），サイト別統計，キーワード別統計により把握することができる（図 4）。

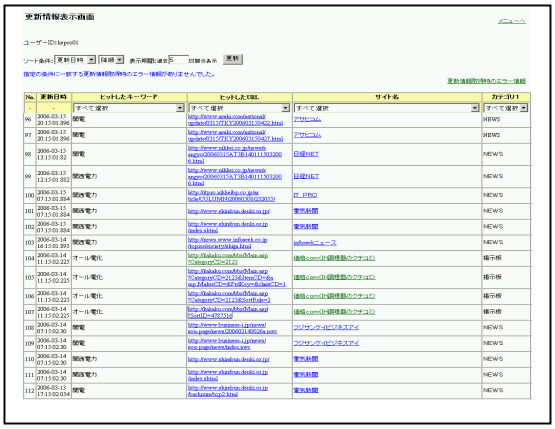

図 3　検索結果画面

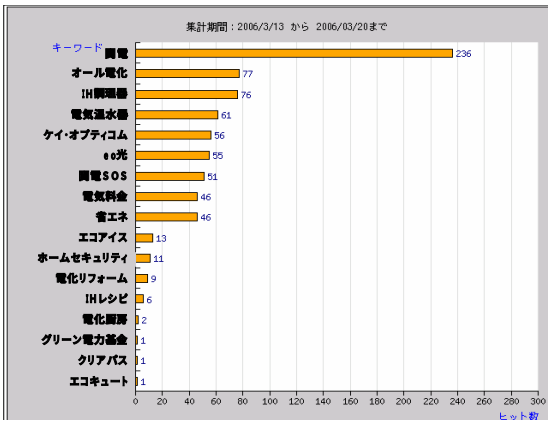

図 4　統計画面

実際に当システムを運用評価した結果，1 時間間隔での巡回処理で，登録 ID(利用者)数 15，登録ＵＲＬ数 196 に対して，取得した HTML は 7,156(page)，処理時間 688(秒)であった。回線速度やサーバ処理能力にもよるが，概ね，当システムにおける性能は，毎秒 10 ページ程度の処理能力を保有することが分かった。

5　まとめ

本研究により，インターネットからのマーケティング情報（新着情報収集）を，高精度・効率的かつ著作権保護を図りながら実現することができた。実運用上においても，十分な処理能力を保有し実際の被験者による主観的評価結果も良好であった。今後は，当システムの機能性・操作性，デザイン等の改良を検討し，商品化へ向けた取り組みを行う予定である。

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.