WWW新着記事収集・配信システムの開発

遠藤裕英† 藤田義之† 上林彌彦††

企業の技術情報サービス部門ではWWW（World Wide Web）のホームページから新製品情報や新技術情報を取得し，研究者に迅速に伝達することが重要な課題になってきた．そこで，あらかじめ登録したホームページに毎日アクセスし，新着記事のキーワードを電子メールで配信するプッシュ型の技術情報サービスを開発した．ホームページから新着記事を抽出する方式として，ホームページに現れるアンカーのURL（Universal Resource Locator）の一致を照合する「アンカーを手掛かりとした新着記事抽出方式」を採用した．また，新着記事のキーワードにはホットテキストとホットイメージ代替文を用いた．上記方式を実装したWWW新着記事収集・配信システムを開発し，著者らの研究所で運用した．新着記事のキーワードを毎日配信することにより迅速な技術情報サービスが実現できた．また,「アンカーを手掛かりとした新着記事抽出方式」によって新着記事だけを抽出できることと，画像の新着記事を抽出できることが分かった．

WWW New Article Extraction and Delivery System

HIROHIDE ENDOH,† YOSHIYUKI FUJITA† and YAHIKO KAMBAYASHI††

An increasingly important topic in the technical information service divisions of large corporations is the obtaining of new product and technical information from homepages on the World Wide Web, and the quick transfer of this to researchers. In this regard, a "push" technical information service has been developed that accesses pre-registered homepages everyday, and delivers keywords of new articles by electronic mail. This system compares and checks the anchor URL (Universal Resource Locator) that appears on the homepages, using this anchor as the key to extracting new articles. Hot text and altenative text to hot image are also used in the keyword of new articles. A World Wide Web new article extraction and delivery system incorporating the said system has been developed, and is operating at our laboratories. By delivering keywords of new articles everyday, a fast technical information service has been achieved. In using this new article extraction system, it has been found that new articles can be extracted regardless of differences in location of articles or exression of articles, and that even keywords of new images can be extracted.

1. まえがき

タイムリーに製品を開発するには，新製品情報や標準化動向などを逸早く入手し研究開発に反映させる必要がある．従来，情報収集は業界紙など紙媒体に頼ることが多かったが，インターネットの普及によりWWW（World Wide Web）経由で情報提供元から直接入手することが可能になった．このため，企業の技術情報サービス部門ではWWWのホームページから新製品情報や新技術情報を取得し，研究者に迅速に伝達することが重要な課題になってきた．

研究所の研究開発分野はあらかじめ定まっているので，日常的に情報収集を必要とするホームページは特定できる．そこで，技術情報サービス部門のサーバから特定のホームページを定期的にアクセスし，新着記事をキーワード化して電子メールで研究者に配信するプッシュ型技術情報サービスシステムを提案する[1]．利用者は必要に応じてキーワード情報から詳細情報にアクセスする．

本システムはイントラネット対応のサーバ集中型プッシュ技術に位置付けられる[2]．各人が個別に情報収集するのに比べ，収集作業の集中化による効率向上とネットワーク資源の有効的活用がはかれる点に特徴がある．

このようなシステムでは，ホームページから新着記事を抽出し，キーワード化するアルゴリズムの開発が

† 株式会社日立製作所システム開発研究所情報センタ
System Developments Laboratory, Information System Center, Hitachi, Ltd.
†† 京都大学大学院工学研究科情報工学専攻
Department of Information Science, Faculty of Engineering, Kyoto University

課題となる．WWW ホームページの記事には，構造化記述言語 HTML（Hyper Text Markup Language）で記述され，更新される記事にはアンカーのついたテキスト（ホットテキスト）または画像（ホットイメージ）が含まれるという特徴がある．そこで，これらの特徴に着目した「アンカーを手掛かりとした新着記事抽出方式」と「ホットテキスト・ホットイメージ代替文キーワード方式」を提案する．前者はアンカーの URL（Universal Resource Locator）の一致を調べて新着記事を抽出する“ハイパーリンク情報（URL）”ベースの新着記事抽出方式である．記事の配列や記述の表現に影響を受けない点と，文章と画像の新着記事を同一方式で抽出できる点に特徴がある．後者は簡単な処理にもかかわらず高い比率でキーワードを抽出できる点に特徴がある．

提案方式を実装した WWW 新着記事収集・配信システムを開発し，著者らの研究所で運用・評価した．その結果，新着記事のキーワードを毎日配信することで迅速な技術情報サービスが実現できた．また，「アンカーを手掛かりとした新着記事抽出方式」によって，(1)‘新着記事’だけを抽出できる，(2) 画像の新着記事も抽出できることが分かった．一方，「ホットテキスト・ホットイメージ代替文キーワード方式」では，新着記事の内容をかなり高い割合で抽出できることが分かった．

以下，2 章で WWW 新着記事収集・配信システムとその中心技術である新着記事抽出方式の課題について述べ，3 章で著者らが提案する「アンカーを手掛かりとした新着記事抽出方式」と「ホットテキスト・ホットイメージ代替文キーワード方式」について述べる．4 章で提案方式を実装した WWW 新着記事収集・配信システムと運用例を，5 章でシステムの評価結果を示す．

2. WWW を利用した技術情報サービスシステムの課題

2.1 WWW 新着記事収集・配信システム

インターネットを利用した情報提供にはプル型とプッシュ型とがある[2]．

プル型は情報源（たとえば，WWW ホームページ）にアクセスして情報を取得する．WWW ホームページの情報は開示とともに瞬時に入手でき，サイトが提供する情報を集中的に入手できる．しかし，アクセスしてみないと記事が更新されたかどうかが分からないことや新着記事の入手が遅れるなどの問題点がある．このため，ホームページの更新を自動的に検出するソフト[3]や更新記事を抽出するソフト[4),5]が開発された．しかし，このソフトを起動しておかなければ更新情報が得られないことや更新内容を知るにはファイルを開く必要があるなど，更新情報の収集・伝達機能は十分とはいえない．

プッシュ型は最近注目されている情報提供方式で，情報源へのアクセス形態で中央制御型，エージェント型，サーバ集中型に分けられる[2]．中央制御型では情報提供元や情報種類を指定しておけば該当する情報が自動的に配信されてくる[6)~8)]．中央にトラフィックが集中すること，情報提供元や情報種類の選択巾が限られることなどの問題点がある．エージェント型では，利用者の指定した WWW サーバだけにアクセスして更新情報を収集する．利用者各人にソフトが必要になることや，同一サイトへのアクセスの重複などネットワークの利用効率に問題がでてくる．そこで，特定部署に集中管理サーバを置き，外部情報の収集を集約して行い効率向上をはかるサーバ集中型が注目されている．

そこで，著者らは，プッシュ型でサーバ集中型に位置付けされる新しい技術情報サービスを提案する．技術情報サービス部門のサーバから，メンバ間で共通する情報源（ホームページ）に定期的にアクセスし，新着記事をキーワード化して電子メールで研究者に配信する．ホームページが更新されたことを単に‘報知’するのではなく，更新内容をキーワードの形で直接配信することに特徴がある．利用者はキーワード情報で記事を絞り込んでから必要に応じて詳細記事にアクセスする．このシステムを WWW 新着記事収集・配信システムと呼ぶ．本論文はこのシステムを対象とする．

2.2 新着記事抽出方式

WWW 新着記事収集・配信システムでは，(1) あらかじめ登録したホームページに定期的にアクセスし，(2) 更新されたホームページを取得し，(3) ホームページから新着記事を抽出してキーワード化し，(4) 配信する．本論文の主要課題はこのプロセスにおける新着記事の抽出方式とキーワード化方式である．

2.2.1 新着記事の抽出方式

WWW ホームページの更新記事を抽出するシステムに AIDE（AT&T Internet Difference Engine）がある[5]．AIDE では更新記事の抽出に記事の配列の一致照合と文字列の類似性照合を用いている．

この方式には 2 つの問題点が含まれている．1 つは，新着記事だけでなく‘手直し記事’も抽出されることである．記事の区切りとなるタグの出現順序で記事の配列を照合しているので，記事内容が同一であっても記

表 1 ホームページの記事構成の調査結果
Table 1 Survey results of articles in homepages.

トップページのレイアウトブロック数	413
(I) 'キマリもの'	244
(II) 記事	169
(a) 文章だけの記事	98
(b) 画像だけの記事	13
(c) 文章と画像の記事	58

調査範囲：情報関連企業 21 社のホームページ（日本と米国）
調査方法：レイアウトブロックは目視で切り出し，'記事' か 'キマリもの' かは著者の主観で判定した．

事の配置に変更があると更新記事として抽出される．文字列の類似性照合では記述表現の変更は更新記事として抽出される可能性がある．このような問題点を解決するためには，記事の配列や記述の変更に影響されない新着記事抽出方式が必要になる．

もう 1 つの問題点は，新着記事には記事画像も含まれるが，記事画像の抽出が取り扱われていないことである．**表 1** に示す「ホームページの記事構成の調査結果」から分かるように，記事の 42%（71 件/169 件）に画像が含まれ，7.7%（13 件/169 件）は画像だけの記事である．このことから，画像の記事内容の抽出は研究されるべき課題であるといえる．

一方，Transceive Receiver[4]☆ではテキスト，リンク，画像名に指定値以上の変化があれば更新記事と見なす方式を採用している．テキストの変化を利用するだけでは AIDE と同様の問題が含まれる．リンク，画像名の変化を利用すれば先に述べた問題点を回避できるが，リンク，画像名の変化だけでどれだけの更新記事が抽出できるのか検証する必要がある．

HTML 文書の記事では，記事が更新されればこの記事に含まれるアンカーの URL（リンク）も更新される．著者らは，この HTML 文書の特徴に着目し，「アンカーを手掛かりとした新着記事抽出方式」を採用する．アンカーの URL は記事の配列の変更や記述の変更によって変わることはないので，記事の配列や記述表現の影響を受けずに新着記事を抽出できる．また，アンカーは文章と同様に画像にも付けられているので，アンカーを手掛かりにすれば文章と画像の新着記事を同じ方式で抽出できる．しかし，Transceive Receiver と同様に本方式でどれだけの更新記事が抽出できるのか検証する必要がある．

2.2.2 キーワード抽出方式

記事の内容を表す代表的な語彙がキーワードである．一般文書のキーワード抽出には統計的手法を用いる方法と自然言語処理手法を用いる方法とがある．前者の代表的な方法に単語の出現頻度を利用する方法[9]がある．後者にはシナリオや文脈のフレームに沿ってキーワードを抽出する方法がある[10)~12)]．これらの手法は，話題ごとに一定量以上の文章があることを前提にしているが，WWW ホームページでは各話題の文章は短く，これらの手法は適切ではない．

HTML 文書からは統計量や自然言語情報のほかに構造化情報やハイパリンク情報が得られる．特に，ハイパリンク情報は情報間の橋渡しを行う要となる情報である．著者らは，この情報に着目し，「ホットテキスト・ホットイメージ代替文キーワード方式」を提案する．すなわち，アンカーが付けられた文章や画像の画像代替文を新着記事のキーワードとして用いる方式である．

2.3 用語の定義

(ア) ホームページを関連する内容ごとにブロック化し，このブロックを 'レイアウトブロック' と呼ぶ．

(イ) 毎回定位置にあり，それ自身が新しい情報を開示するものでないレイアウトブロックを 'キマリもの' と呼ぶ．ヘッドカバー，目次，タイトルバー，フッター（連絡先，著作権表示など）などである．

(ウ) 情報を開示するレイアウトブロックを '記事' と呼ぶ．'記事' には見出し記事と本文記事とがある．記事を記述した文章と画像をそれぞれ記事文章，記事画像と呼ぶ．

(エ) 前回から変更のあった記事を更新記事と呼ぶ．更新記事には '新着記事' と '手直し記事' とがある．'手直し記事' は記事の内容が同じで，記事の配列を変更した記事や記述を変更した記事である．

(オ) アンカータグが付けられた文章をホットテキスト，画像をホットイメージと呼ぶ．

(カ) 画像を表示できないブラウザ用に画像の代替えとして表示される文字列を画像代替文と呼ぶ．ホットイメージの画像代替文をホットイメージ代替文と呼ぶ．画像代替文が付いている画像の比率を画像代替文付与率と呼ぶ．

(キ) キーワードは記事内容を表す代表的な語彙である．本論文では，新着記事の内容が推測できて，さらに詳細な内容を参照すべきか否かの判断ができる語または文を 'キーワード' と呼ぶ．ホットテキストや画像代替文がキーワードになっている比率をキーワード率と呼ぶ．

上記の定義に基づくホームページの構成例を**図 1** に示す．ホームページは複数個のレイアウトブロックか

☆ Transceive Receiver はカナダ Caravelle 社の商標である．

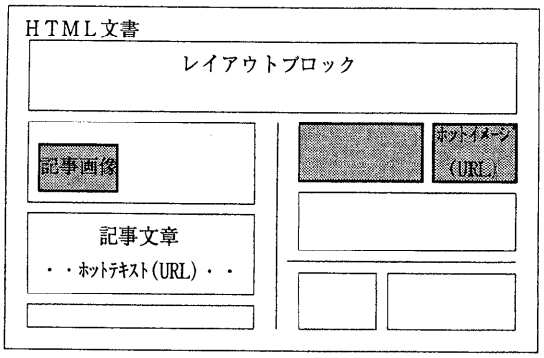


図 1 ホームページの構成例
Fig. 1 Configuration of a homepage.

ら構成される．'キマリもの'と'記事'に含まれる文字列と画像にはアンカーが付けられ，対応するHTML文書にリンクが張られている．

3. アンカーを手掛かりとした新着記事抽出方式

新着記事の抽出アルゴリズムは，(1) 更新アンカーの抽出，(2) ホットテキストの抽出，(3) ホットイメージ代替文の抽出の3ステップからなる．

ステップ1：更新アンカーの抽出

アンカー（⟨A HREF=···⟩ タグ）の URL リストと MAP 画像のエリアタグ（⟨AREA ···, HREF="http://···"⟩）のリンク先 URL（"http://···"）リストを作成する．

次に，各 URL を前回取得したホームページの URL リストと照合する．URL が前回のリストにあれば新着記事でないと判断し，次の URL を調べる．URL が前回のリストになければ新着記事であると判断し，ステップ2に進む．

リンク先 URL の抽出は次のように行う．HTML で記述されるアンカーのシンタックスは，下に示す（ア）と（イ）であるので，(ア) の "URL" と（イ）の "http://···" を抽出する．URL の抽出は ⟨A HREF= の文字列を検出し，= 以降，スペースまたは ⟩ が出現するまでの文字列を URL として取り出す．画像の場合は，⟨AREA ··· HREF= の文字列を検出し，= 以降，スペースまたは ⟩ が出現するまでの文字列を URL として取り出す．

（ア）⟨A HREF="URL"⟩ ··· ⟨/A⟩
（イ）⟨IMG SRC="URL"
USEMAP="#mapdata"⟩
⟨MAP NAME="mapdata"⟩
⟨AREA SHAPE=···, COORDS=···,
HREF="http://···"⟩
⟨/MAP⟩

ステップ2：ホットテキストの抽出

ホットテキストのシンタックスは下に示す（ア）～（ウ）であるので，⟨A HREF= 以降，/A⟩ までに現れる ⟩ から，次の ⟨ までの text をホットテキストと見なして抽出する．

（ア）⟨A HREF="URL"⟩ text ⟨/A⟩
（イ）⟨A HREF="URL"⟩ text
⟨IMG SRC="URL" ··· ⟩ ⟨/A⟩
（ウ）⟨A HREF="URL"⟩
⟨IMG SRC="URL" ··· ⟩ text ⟨/A⟩

ステップ3：ホットイメージ代替文の抽出

ホットイメージ代替文のシンタックスは下に示す（ア）～（ウ）であるので，⟨IMG SRC= 以降の ALT= を抽出し，= からスペースまたは ⟩ までの title を画像代替文と見なして抽出する．

（ア）⟨A HREF="URL"⟩⟨IMG SRC="URL"
ALT="title"⟩⟨/A⟩
（イ）⟨A HREF="URL"⟩ text
⟨IMG SRC="URL" ALT="title"⟩ ⟨/A⟩
（ウ）⟨A HREF="URL"⟩⟨IMG SRC="URL"
ALT="title"⟩ text ⟨/A⟩

4. WWW 新着記事収集・配信システム

4.1 システム構成

WWW 新着記事収集・配信システムのシステム構成を図2に示す．

WWW 新着記事収集・配信システムは Proxy サーバ，WWW 新着記事収集・配信サーバ，クライアント端末，ネットワークシステムよりなる．

Proxy サーバには日立製パソコン Flora 3100SP☆（Pentium☆☆ 90 MHz，メモリ 64 MB，ハードディスク 5 GB）を使用している．WWW 新着記事収集・配信サーバは Proxy サーバと兼用している．クライアント端末（パソコン）約 500 台が LAN に接続されており，Windows 95☆☆☆上でメールソフトとブラウザソフトが稼動する環境を構築している．ネットワークシステムはインターネットとの接続が 6 Mbps，ファイアウォールとサーバ間が 384 Kbps，LAN はバックボーンが 100 Mbps の FDDI，フロア LAN が 10 Mbps の Ethernet である．

☆ Flora は（株）日立製作所の登録商標である．
☆☆ Pentium は米国 Intel 社の商標である．
☆☆☆ Windows 95 は米国 Microsoft 社の登録商標である．

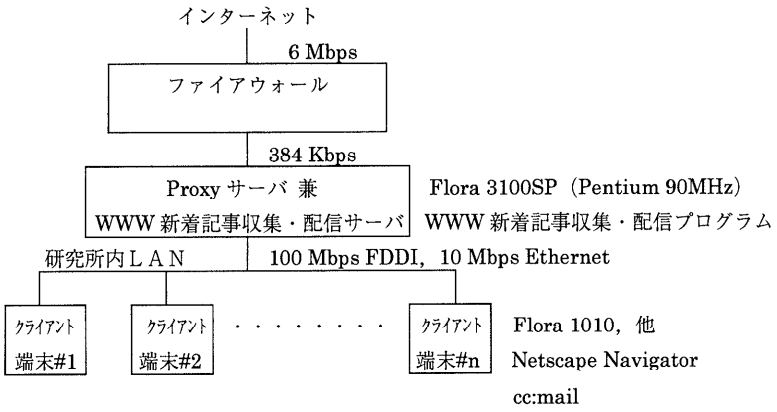


図 2 WWW 新着記事収集・配信システムの構成☆
Fig. 2 World Wide Web new article search and mail system.

4.2 ソフトウェア

WWW 新着記事収集・配信システムはサーバ上の WWW 新着記事収集・配信プログラムと，クライアント端末上の電子メールソフトとブラウザソフトとからなる．

WWW 新着記事収集・配信プログラムは Proxy サーバ上に Perl で作成した更新ページ取得プログラム（0.3 K ステップ），新着記事抽出プログラム〔新着記事のホットテキスト・ホットイメージ代替文抽出，新着記事のキーワード編集（0.4 K ステップ）〕，キーワード配信プログラム（0.1 K ステップ），データベース化プログラム（0.2 K ステップ）から構成される．

図 3 に WWW 新着記事収集・配信プログラムのフローを示す．

（1）更新ページ取得プログラム

URL の登録されたホームページをアクセスし最終更新日を取得する．前回の最終更新日と照合し，最終更新日が更新されていれば更新ページを取得する．WWW サーバから最終更新日を取得できない場合には無条件にページを取得する．

（2）新着記事抽出プログラム

前回取得したページと今回取得したページの全アンカーの URL リストを作成する．前回の URL リストと今回の URL リストを照合し，追加された URL と削除された URL のリストを作成する．追加された URL のホットテキストとホットイメージ代替文を抽出し，URL との対応表を作成する．そして，URL ごとのホットテキストとホットイメージ代替文をキーワードとする．

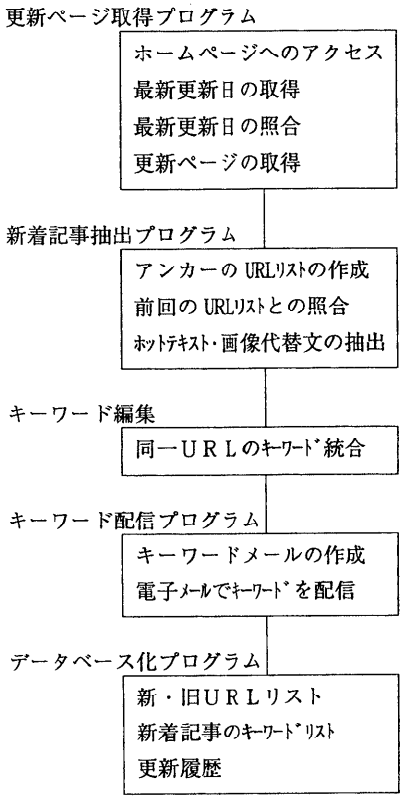


図 3 WWW 新着記事収集・配信プログラムの流れ
Fig. 3 Flow of the World Wide Web new article search and delivery program.

（3）キーワード配信プログラム

登録してある宛先に電子メールでホームページ名と新着記事のキーワードを配信する．メールのヘッダー（宛て先，件名など）を作成し，キーワード情報の先

☆ cc:mail は米国 Lotus Development 社の商標である．Netscape Navigator は米国 Netscape Communications 社の商標である．

頭に挿入してメールテキストを作成する．メールプログラムにメールテキストを渡し配信する．

(4) データベース化プログラム

トップページに追加・削除されたURLリスト，新着記事のキーワードリスト，更新日履歴データをHTMLに変換しWWWサーバのデータに登録する．また，更新ページ（トップページ）もWWWサーバのキャッシングデータとして登録する．

4.3 システムの運用

WWW新着記事収集・配信システムでは，更新ページの取得，新着記事のホットテキストとホットイメージ代替文の抽出，キーワードの編集，キーワードの配信，更新履歴のデータベース化などが行われる．全処理時間は約5分である．

著者らの研究に関連のある日本と米国の28サイトのホームページに，毎日アクセスして新着記事を収集している．WWWホームページへのアクセスは，日本と米国のインターネットのトラフィックが比較的少ないと予想される時間帯（日本時間で午前7時，米国太平洋時間で午後2時）に行っている．ホームページへのアクセスは1日1回としている．米国に所在するサイトであれば，日本の勤務時間中にホームページを更新することは少ないので，1日1回で問題ないと考えている．

WWWホームページの新着記事のキーワードは，毎日，研究所内の電子メールシステムで研究管理職を中心に51人に配信している．**図4**(a)にキーワードの配信例を示す．

各ホームページのトップページをProxyサーバにキャッシングし，詳細記事へのアクセス時間が短くなるようにしている．また，更新によって削除されたURL・キーワードリストと追加されたURL・キーワードリスト（図4(b)），トップページの更新履歴をデータベース化している．図4(a)，(b)で［…］は画像代替文を示す．

4.4 システムの利用

WWW新着記事収集・配信サービスを利用した情報収集は次のようになる．電子メールを開くと，業務上関連の深いサイトのWWWホームページの新着記事がキーワードで配信されている．このメールの内容を一覧して，関心のあるキーワードが含まれていれば，(1) キャシッシングされているホームページ（トップページ）をアクセスするか，(2) 新着記事のURL・キーワードリストからリンク先の情報にアクセスする．トップページだけをキャシッシングしているので，トップページ以外のハイパリンク先はリアルタイムでリンク先にアクセスすることになる．

5. 評　価

5.1 WWW新着記事収集・配信システムの評価

WWW新着記事収集・配信システムを利用した情報収集は，(1) メールで配信される新着記事のキーワードを見る，(2) キャッシングされたホームページを見る，(3) ハイパリンク先をアクセスするという手順になる．ここでは，本論文の対象とした(1)と(2)の効果を検証する．

(1) キーワード配信サービスの効果

関連するサイトの新着記事をキーワードの形で毎朝配信することにより，情報を迅速に伝達できる技術情報サービスシステムを実現できた．

表2はキーワードの配信サービスをした週（5日間）としなかった週の代表的ホームページ（A，B，C，D）へのアクセス回数を調べた結果である．キーワード配信サービスによってホームページへのアクセス回数が1.7倍（900件/519件）に増え，情報収集活動が活発化していることが分かる．ホームページの利用が十分進んでいる環境では，キーワードの配信によって新着記事があるかどうかが分かり閲覧記事も絞りこめるため，アクセス回数は減少することが予想される．著者らの調査した環境では，まだ，ホームページの活用が不十分で，キーワード配信サービスによる刺激でホームページへのアクセスが増加したと考えられる．

(2) キャッシングの効果

トラフィックの混雑しない時間帯に相手サイトのホームページを取得しキャッシングしている．キャッシングによってトップページへのアクセス速度が速くなる．

著者らの職場のLAN環境は10 Mbpsであり，電子メールの運用にはトラフィック上問題はない．しかし，WWWサイトへのアクセスについては就業開始直後の時間帯，昼休み前後の時間帯，終業直前から終業後の時間帯が混雑することが知られている．**図5**は米国C社のホームページ（トップページ）をProxyサーバにキャッシングするのに要した時間を時刻別に表したものである．キャッシングされたデータを利用すれ

表2 トップページのアクセス回数
Table 2 Number of accesses to top homepages.

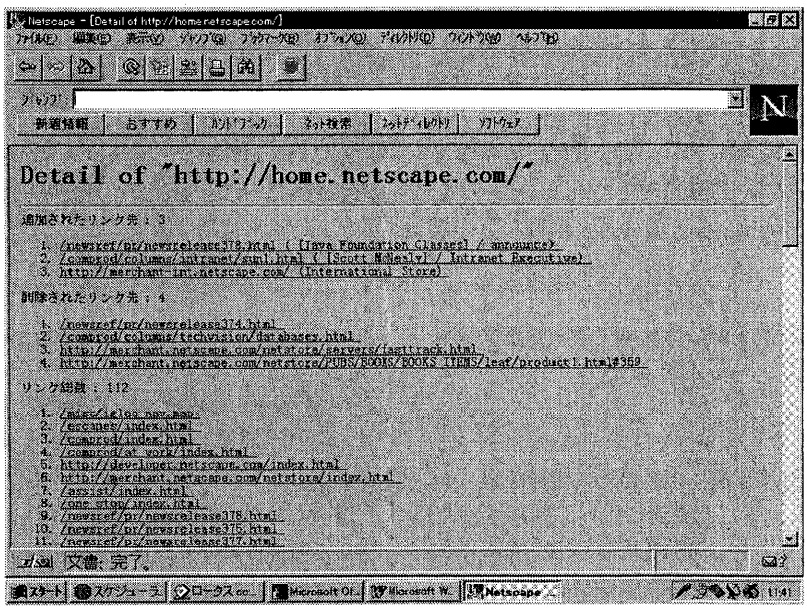
	サービス有	サービス無
A	271	126
B	486	326
C	74	40
D	69	27
合計	900	519

WWWホームページ新着記事案内サービス

WWWホームページの更新（4月3日）

Netscape	([Java Foundation Classes] / announce)
	([Scott McNealy] / Intranet Executive)
	(International Store)
Sun Microsystems	(It's Here!)
	(JavaOne Today)
	(Sun @ NAB)
	(Sun and the Year 2000)
	(JavaOne Announcements)
Intel	([Join the MMX(tm) Owner's Club!] / Join the MMX Technology Owner's Club)

(a) 新着記事のキーワード配信例



(b) アンカーの URL リスト例

図4 システムが提供するデータ例
Fig. 4 Example of data provided by the system.

ばアクセス時間が2〜37秒短縮されることが分かる．混雑する時間帯におけるキャッシングの効果が大きい．

5.2 「アンカーを手掛かりとした新着記事抽出方式」の評価

本論文の目的の1つは記事の配列や記述表現の影響を受けないで新着記事を抽出することと，画像の新着記事を抽出することである．これを検証するため，更新記事に「アンカーを手掛かりとした新着記事抽出方式」を適用して新着記事の抽出結果を調べた．

実験サンプルは情報関連企業19社（日本と米国）のホームページの更新記事44件である．更新記事は'新着記事'39件と'手直し記事'5件とからなり，21件の記事画像を含む．'手直し記事'はホームページ内での記事の配置変更，記事内での文章と画像の配置変更，記述表現の変更が行われた記事である．記述表現の変更には，異なる単語で言い換えたもの，複数の記事を要約して統合したものがある．また，記事画像はすべてホットイメージで，すべての記事に画像代替文が付

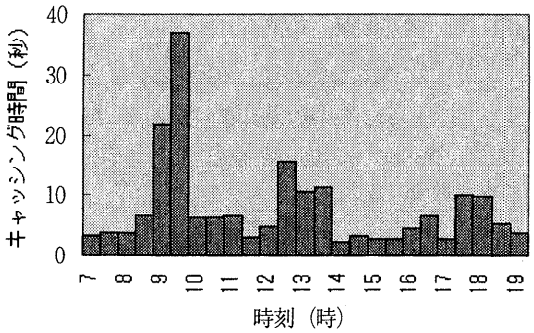


図 5 時刻別キャッシング時間
Fig. 5 Caching time of a World Wide Web site by time of the day.

表 3 実験に用いた記事サンプル
Table 3 Article samples for evaluation.

更新記事総数	44 件
(1) '手直し記事'	5
(a) ホームページ内での配置変更	4
(b) 記事内での配置変更	1
(c) 記述表現の変更	3
(2) '新着記事'	39
ホットイメージ数	21

表 4 更新記事の抽出結果
Table 4 Extraction results for new articles.

	対象件数	抽出件数
更新記事数	44 件	39 件
'手直し記事'	5 件	0 件
'新着記事'	39 件	39 件
ホットイメージ代替文	21 件	21 件

与されている．実験サンプルのプロファイルを**表 3** に示した．

WWW 新着記事収集・配信プログラムで実験サンプルから新着記事を抽出したところ，**表 4** に示す結果が得られた．

(a) '新着記事' 39 件はすべて抽出された．

(b) ホームページ内での配置変更，レイアウトブロック内での配置変更，記述表現に変更のあった '手直し記事' 5 件は，アンカーのリンク先 URL は同じであるため，すべて '新着記事' としては抽出されなかった．

(c) 新着の記事画像（および画像代替文）21 件はすべて抽出された．

ここでは，新着記事に少なくとも 1 つのホットテキストかホットイメージが含まれ，その URL が抽出されたとき，'新着記事' が抽出されたと見なしている．

以上の結果から，「アンカーを手掛かりとした新着記事抽出方式」が '手直し記事' を排除して '新着記事' だけを抽出できる抽出方式であること，新着の記事画像（画像代替文）を抽出できる方式であることを実証できた．

表 5 キーワード分析の結果
Table 5 Evaluated results for the keyword.

記事総数	169
(1) ホットテキスト数	202
ホットテキストがキーワード	160
(2) ホットイメージ数	40
(a) ホットイメージ代替文あり	36
(ア) 代替文がキーワード	28
(イ) 代替文がキーワードでない	8
画像のほかに記事文章あり	8
(b) ホットイメージ代替文なし	4
画像のほかに記事文章あり	3

5.3 「ホットテキスト・ホットイメージ代替文キーワード方式」の評価

「ホットテキスト・ホットイメージ代替文キーワード方式」のキーワード率を検証する．

分析対象サンプルは情報関連企業（日本と米国）21 社のホームページの記事 169 件である．この記事に含まれるホットテキストやホットイメージ代替文がキーワードであるか否かを分析する．キーワードであるか否かの判定は 2.3 節（キ）の定義に従い著者の主観で行った．分析結果は**表 5** に示すとおりで，次のことが分かる．

(a) ホットテキストのキーワード率は 79%（160 件/202 件）である．

(b) ホットイメージの代替文付与率は 90%（36 件/40 件）である．

画像代替文が付与されてない画像 4 件は ISMAP 画像，バナー（見出し画像），顔写真，製品写真で，この中の 3 件には記事文章が併記されている．

(c) ホットイメージ代替文のキーワード率は 70%（28 件/40 件）である．

キーワードでない 8 件は，画像代替文が空文（ALT=“ ”）の場合と新着記事の内容に無関係な画像代替文（ロゴ名や NEW! など）が付けられたものであった．この 8 件にはすべて記事文章が併記されている．

(d) '画像だけの記事'（13 件）の画像代替文付与率は 92%（12 件/13 件）であり，画像代替文が付与されてない 1 件は ISMAP 画像であった．キーワード率は 100%（12 件/12 件）であった．

ホットテキストとホットイメージ代替文のキーワード率はそれぞれ 79%と 70%である．表 5 から 169 件の記事に対しホットテキストが 202 件，ホットイメージ代替文が 36 件あるので，単純に加重平均すると 1

記事あたり109%になる．このことから，ホットテキストとホットイメージ代替文で記事内容を伝達できる割合はかなり高いといえる．

ISMAP/USEMAP画像には画像代替文が付与されていない場合がある．このような画像では画像中に記事が埋め込まれているため，リンク先からキーワードを抽出する必要がある．

5.4 キーワード率向上方法の検討

5.3節の結果から，ホットテキストのキーワード率は79%で，ホットイメージ代替文のキーワード率は70%であった．どのような語句や文をホットテキストや画像代替文にするかは記事作成者に依存するので，後処理的方法でキーワード率の向上をはかる必要がある．ここでは，ホームページから収集したデータをもとにキーワード率を向上させる方法を検討する．

キーワードでないホットテキストには次のようなタイプが含まれる．

(**a**) 見出し記事（ホットテキスト）が一般用語で，本文記事にホットテキストが含まれないタイプ．

（例）Join the Beta Program!, Put the Web to Work, Network Computer News, など．

見出し記事がキーワードであるか否かの判定は難しい．新着記事であることは判定できるので，見出しと本文記事全体をブロックとして抽出する方法が考えられる．

(**b**) 本文記事で一般用語がホットテキストになっているタイプ．

（例）developed, opened, announcements, unveiled, reported, now available, delivers, additional features, fastest laptop computer, price reductions, など

ホットテキストの前後に記事の特徴を表すキーワードが出現する可能性が高い．用語辞書を設け，ホットテキストが該当用語であった場合は，ホットテキストを含む単文，または，記事全体を抽出する方法が考えられる．

(**c**) ボタン代用（ホットテキスト）などのタイプ．

（例）Click here! など

同一のURLにリンクしている記事文章があるので，記事文章のホットテキストを利用する．用語辞書を設けてこの種のホットテキストはキーワードから除外する．

(**d**) 新着記事と無関係な画像代替文であるタイプ．

（例）［NEW!］，［NEWS RELEASE］など．

(c)と同様の方法が考えられる．

6. まとめ

インターネット時代のプッシュ型企業内技術情報サービスとして，登録したホームページから新着記事を毎日収集し，キーワード化して研究者に配信するWWW新着記事収集・配信システムを開発した．

開発したWWW新着記事収集・配信システムは，1996年7月から現在まで，著者らの研究所の技術情報サービス部門で運用され，迅速な技術情報サービスと新着ページへのアクセスの高速化を実現している．

本システムでは新着記事抽出方式として「アンカーを手掛かりとした新着記事抽出方式」を，キーワード化方式として「ホットテキスト・ホットイメージ代替文キーワード方式」を採用した．これらの方式は処理が簡単で，記事の配列や記述表現の影響を受けないで新着記事を抽出できること，文章と画像の新着記事を同一方式で抽出しキーワード化できるなどの特徴を持つ．

日本と米国の情報関連企業21社のホームページを対象に，「アンカーを手掛かりとした新着記事抽出方式」を適用した結果，新着記事はすべて抽出された．新着記事の抽出では，'手直し記事'は抽出されず，'新着記事'だけが抽出された．一方，新着記事の内容を要約する「ホットテキスト・ホットイメージ代替文キーワード方式」によるキーワード化では，79%のホットテキストと70%のホットイメージ代替文からキーワードが得られることが分かった．

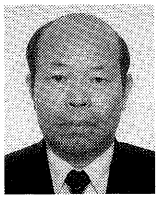
WWWを利用した情報流通は急速に拡大しており，ハイパリンク情報ベースの情報抽出やハイパリンク情報のキーワードとしての利用方法は研究されるべき価値があると考える．

参考文献

1) 藤田義之，遠藤裕英，武藤英男，松尾 修，小川和美：WWW更新ページ報知サービスシステムの開発，第54回情報処理学会全国大会論文集，4-315 (1997).
2) WWWの情報洪水を解消するPush技術，日経コンピュータ，No.414, pp.216–229 (1997).
3) FirstFloor Software Inc.: http://www.firstfloor.com/
4) Caravelle Inc.: http://www.caravelle.com/
5) Douglis, F. and Ball, T.: Tracking and Viewing Changes on the Web, *Proc. 1996 USENIX Technical Conference*, pp.165–176, San Diego (1996).
6) PointCast Inc.: http://www.pointcast.com/
7) Lanacom Inc.: http://www.lanacom.com/
8) Marimba Inc.: http://www.marimba.com/

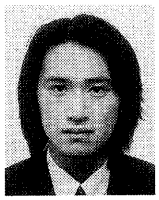
9) Luhn, H.P.: The Automatic Creation of Literature Abstracts, *IBM Journal*, Vol.2, No.4, pp.159–165 (1958).
10) 猪瀬 博，斎藤忠夫，堀 浩一：シナリオを用いる論文抄録理解・作成支援システム，情報処理学会論文誌，Vol.24, No.1, pp.22–29 (1983).
11) 石橋弘義，重永信一，新納浩幸，福重貴雄，安川秀樹：英文要約システム「DIET」，第38回情報処理学会全国大会論文集，6D-9, pp.239–240 (1989).
12) 安藤真一，土井伸一：新聞記事からの情報抽出システム—指定情報の抽出と多言語文章による提示，人工知能学会全国大会，24-3, pp.671–674 (1994).

(平成9年 5 月8日受付)
(平成9年10月1日採録)



遠藤 裕英（正会員）

1941年生．1965年京都大学工学部電子工学科卒業．同年（株）日立製作所入社．中央研究所主任研究員，マイクロエレクトロニクス機器開発研究所部長を経て，現在システム開発研究所情報センタ長．この間，制御用計算機，パターン認識装置，ワープロ，技術情報サービスなどの研究開発に従事．



藤田 義之

1973年生．1991年香川県立多度津工業高校電子科卒業．同年（株）日立製作所入社．1995年日立京浜工業専門学院研究科卒業．現在，システム開発研究所情報センタにて研究情報システムの構築を担当．インターネット利用技術に興味を持つ．

上林 彌彦（正会員）

1970年京都大学大学院博士課程修了．イリノイ大学リサーチアソシエイト，京都大学助教授などを経て、1984年九州大学教授．1990年より京都大学教授．この間，マッギル大学，武漢大学，クウェート大学客員教授．情報処理学会データベースシステム研究会主査，電子情報通信学会コンピュテーション研究会委員長，情報処理学会理事，ACM日本支部副支部長などを経て，現在，情報処理学会理事．データベースや協調処理の研究を行っている．1995年ACM SIGMOD Contribution Award受賞．1996年より文部省科学研究費重点領域研究「高度データベース」代表．