WWW情報フィルタリング・検索システム(FreshEye) ——— サービス概要 ———

3L－5

住田一男[*1]　鈴岡節[*1]　平野宜史[*2]　野上宏康[*2]
*1（株）東芝 研究開発センター　*2（株）東芝 ＩＰ事業推進室

1．はじめに

ここ数年 World Wide Web（WWW）の規模の増大はめざましく、多くの企業や個人がホームページを WWW に設立し、様々な情報を提供している(1)。一方、公開されている膨大な数のページの中には作成されて以来まったく更新されていないページも多く存在している。こののため、ある話題の情報をウォッチすることを目的としている人にとっては、従来の検索エンジンを使用した場合、新規な情報が古い情報の中に埋もれてしまい、必要な情報が得られないという問題があった。このような問題意識から、我々はすでに PC 単独で動作する WWW 情報フィルタリングソフトを開発している(2)。ＰＣソフトであるため、処理の範囲を限定せざるを得ないという問題がある。内容が更新されたページや新規に公開されたページを検出することを中心的な課題ととらえ、サーバタイプの WWW 情報フィルタリング・検索システム FreshEye を開発した。本稿では、FreshEye が提供可能なサービスについて述べる。

2．従来のディレクトリサービスと検索エンジンでのページ の取り扱いと問題点

どこか特定の決まったページのURLを知っている場合を除いて、WWW のユーザは、膨大な数のページの中から自分の関心にあったページの URL を探し出す必要がある。現在のところ、これを大規模にサポートするサービスは、ディレクトリサービスと検索エンジンの2種類のサービスとに分類できる。

通常、ディレクトリサービスでは、カテゴリごとの分類を人手で行うことから、WWW 上に公開されているすべてのページを対象にすることはできず、網羅性に欠ける。また、頻繁かつ大量に情報を更新することも人手がかかり難しい。

一方、検索エンジンは、キーワードをユーザが入力することにより、入力したキーワードを含むページを検索するサービスを提供する。検索エンジンでは、高速なキーワード検索を実現するために、ページをあらかじめ収集しておき、各ページを解析しそこに含まれる単語についてのインデックスを作成しておく必要がある（ページを自動的に収集するソフトウェアはロボットと呼ばれる）。WWW上に公開されているページは膨大な数に上るため、ロボットがすべてのページを瞬時に収集することは物理的に不可能である。このため、順次ページを収集していくことにならざるえない。このため、ページが新規に登録されたり更新されてから、その内容が検索に反映されるまでに時間遅れが生じる。また、作成されてからまったく更新されていないページもインデックスに残されてしまう結果、新規な情報が古い情報の中に埋もれてしまい、必要な情報が得られないという問題があった。

FreshEye では、上記のような問題を考慮し、情報の新規性に焦点を絞り、ディレクトリサービスと、検索エンジンとを組み合わせたシステムとして構築した。

FreshEye : An Information Filtering and Search System for WWW ——— Service Overview ———
Kazuo SUMITA[*1], Takashi SUZUOKA[*1], Takashi HIRANO[*2], Hiroyasu NOGAMI[*2]
*1 Research and Development Center, Toshiba Corp., 1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210-8582, Japan.　*2 Information Provider Division

3．FreshEyeで提供する機能

提供する機能は以下の通りである。

①キーワードによるページの検索

ロボットが収集したページを対象にして、ユーザがキーワードを入力することにより、そのキーワードを含むページを検索、そのページへのリンクをリスト表示する。ロボットの収集から収集したページを検索可能になるまでの遅延の短さが特長である。ユーザが入力可能な条件式は、AND や OR などの論理演算子を含むブール式である。

入力された条件式と関連性のあるトピックを検出し、表示する機能を持つ。これは、検索結果に含まれる語と、各トピックの分類結果に含まれているページに含まれている語との重複の度合いを算出し、関連性を判断している。

検索結果表示としては、1ヶ月以内に取得したページを類似度順に、1週間以内に取得したページを更新日順に表示する2つのモードを用意した。

②トピックによる分類

各トピックごとにプロファイルを用意しており、ロボットが収集したページに対してこれらプロファイルと照合し、各トピックに分類する。

トピックごとの表示についても、検索と同様に月間表示と週間表示の2モードを用意し、それぞれ分類結果を類似度順、更新日順に表示する。なお、分類結果をページとして生成する都合上、現状では収集したページを1日単位でまとめて処理している。

4．動作例

図1に「ワールドカップ」という語を検索条件とした検索結果画面（月間表示）を示す。関連するトピックが動作の詳細については、文献(3)を参照されたい。関連するトピックとして「Jリーグと世界のプロサッカー」が検出されている。

図2に「プロ野球応援団」のトピック分類を示す。内容は日々更新され、[NEW]というマークがその日取得されたページを意味している。

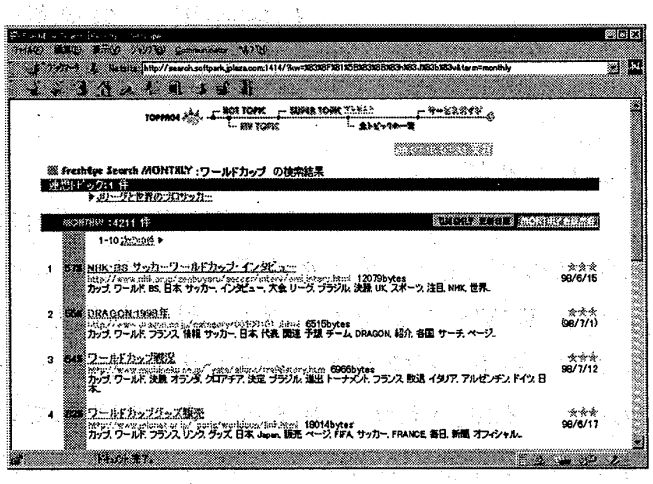

図1 「ワールドカップ」の検索例

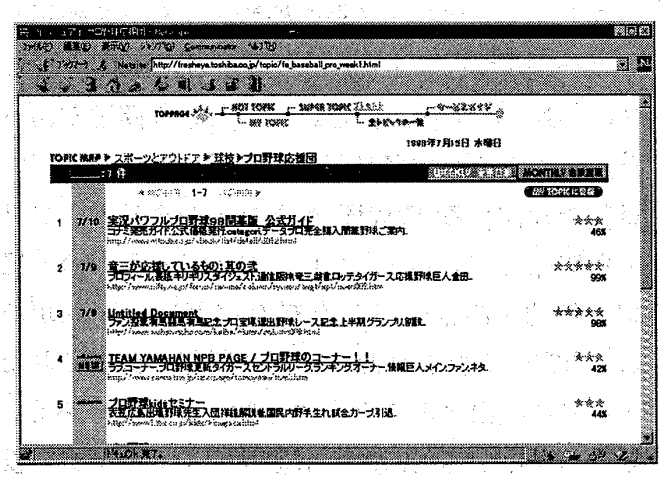

図2 トピック「プロ野球応援団」

5．まとめ

更新されたページや新規に作成されたページに焦点を絞った、オンデマンドな検索サービスとページ分類を可能にするための、情報フィルタリング・検索システムを開発した。本システムは、http://fresheye.toshiba.co.jp で公開している。

参考文献

(1)例えば、インターネット白書'98

(2)住田他: "WWW 上のフロー情報を対象にした情報フィルタ(FreshEye)", インタラクション '97, pp.63-64 (1997).

(3)鈴岡他: "WWW 情報フィルタリング・検索システム—全体システムの構成と動作—", 第 57 回情全大, 3L-06 (1998)