WWW上のフロー情報を対象にした情報フィルタ (*FreshEye*)

住田一男† 上原龍也† 小野顕司† 酒井哲也† 池田朋男†† 下郡信宏†††
(株) 東芝 研究開発センター
†情報・通信システム研究所 ††関西研究所 †††システム・ソフトウェア生産技術研究所

1. はじめに

ここ数年 World Wide Web (WWW) の規模の増大は目覚しく、多くの個人や企業が様々な情報をホームページとして公開している。現在、WWWのユーザ数は年毎におよそ倍増していると推定されており、それに呼応して公開されるWebページも増加の一途にある。これらのページを検索するため、いくつかの検索サーバが公開されている。

ところが、これら膨大なページの中には作成時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。このため、ある話題に対して恒常的に新規情報の発生をウォッチすることを目的としている人にとって、公開されている検索サーバを用いた場合、肝心の新規情報が多くの古い情報に埋もれてしまい、必要な情報を得ることができないという問題が生じる。

検索が蓄積された情報を処理対象とするのに対して、情報フィルタリングでは、日々発生する情報を処理対象とする [1]。本稿では、このような情報をフロー情報と呼ぶ。WWW上に存在する情報は、蓄積情報であるとともに、日々多くの情報が創出されており、フロー情報としての側面を持つ。本稿では、WWW上の情報をフロー情報としてとらえ、そこから個人の関心や興味にしたがって必要な情報を抽出する情報フィルタリングシステム *FreshEye* について述べる。

システムは、WWWより新規情報を抽出し、ユーザの関心度の高い順に情報を整理し提供する。また、ユーザの関心度はフィードバック可能であり、これにより情報選択の精度を向上させる機能を持つ。さらに、英語、日本語の両言語についてもフィルタリング処理が可能である。*FreshEye* は、PC上で動作するソフトウェアとして実現した。Webブラウザと連携して動作する。

A Filtering System for Information Streams on the World Wide Web

Kazuo SUMITA, Tatsuya UEHARA, Kenji ONO, Tetsuya SAKAI, Tomoo IKEDA, and Nobuhiro SHIMOGORI

Research and Development Center, Toshiba Corp.
1 Komukai-Toshiba-cho Saiwai-ku Kawasaki, 210, Japan
Tel:044(549)2240, Fax:044(520)1308
E-mail:sumita@eel.rdc.toshiba.co.jp

2. 情報フィルタリングの情報ソースとしてのWWW

WWW上のWebページをフロー情報ととらえると、これらのWebページは新規に創出されたページと更新されたページとに分類することができる (それ以外のページは、蓄積情報とみなし、フィルタリングの対象外となる)。さらに、更新されるページは、更新が定期的に行われるものと不定期であるものに分類できる。定期的に更新されるページの代表的なものとして、新聞社や雑誌社が公開している記事情報をあげることができる。一般のページは、その更新はおよそ不定期に行われる。

超分散的なデータベースであるWWWを対象にした情報フィルタリングにおいては、上記のフロー情報を抽出する必要がある。

検索システムでは、ユーザは検索された結果を見ながら検索条件を修正し再度検索を行うことが可能である。一方、情報フィルタリングシステムでは自律的に処理結果を得る必要がある。このため、検索と比較して検索条件は詳細に設定可能ではあるが、精度の高い処理結果を得ることが求められる。

3. システムの概要

本システムの概要を図1に示す。

FreshEye は、新規ページと不定期更新ページを抽出するために、検索サーバ (http://www.lycos.com/など) と新着サイトリストページ (http://www.ntt.jp/WHATSNEW/index-j.htmlなど) にアクセスする (新規ページ取得サイト)。

ページ取得ゲートウェイを介したページの取得では、検索サーバに対しては検索コマンドを発行することにより関連ページのリストを得る。また、新着サイトリストページからは、直接そのページを読み込むことにより、新規発生サイトのリストを得る。

更新チェックにおいて、前回取得したリストとの差分を求め、新規発生ページや更新されたことで

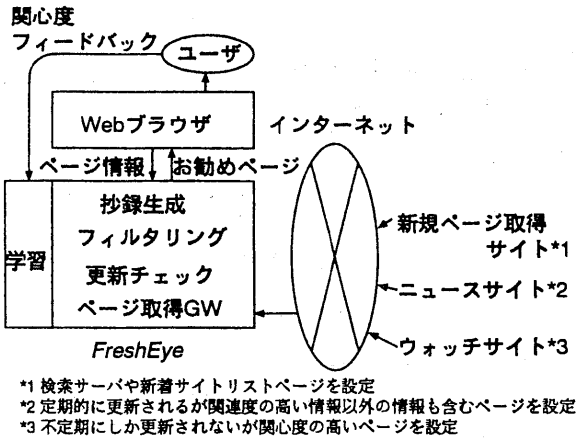


図 1: システム概要

ユーザに関心のある情報を含むようになったページを取り出す。これらの各ページをフィルタリング対象とする。

一方、定期的に更新されるページについては、指定されたページを直接読み込み、そのページ自体に含まれる情報をフィルタリング対象とする (ニュースサイト)。複数の記事が含まれるページについては、記事毎に分割し、これらの記事をフィルタリング対象とする。

不定期にしか更新されないものの、ユーザにとって関心の高いページも存在する。*FreshEye* では、このようなページをウォッチサイトとして登録する。登録されているページが更新されたか否かをチェックし、更新されていた場合これをユーザに通知する。

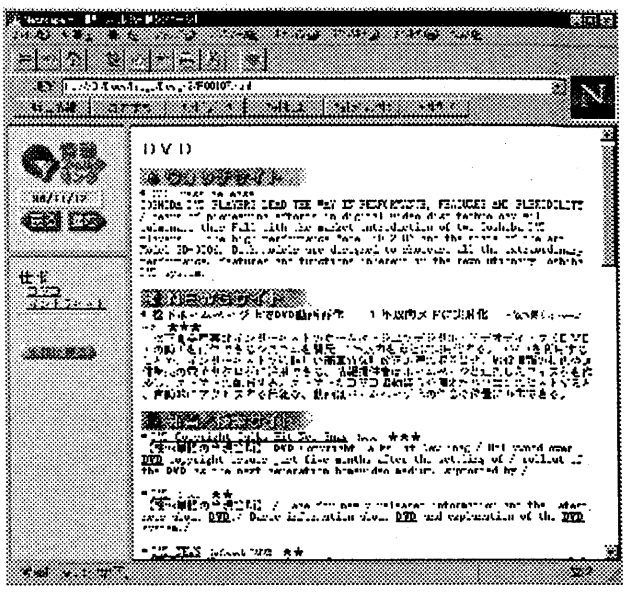
フィルタリング処理では、ユーザの関心や興味を記述したプロファイル (検索条件) との類似度を算出し、類似度の順に情報をランキングする。類似度の算出では、ベクトル空間法をベースに語の頻度情報、見出しや段落情報、係受け情報等を解析し高精度にランキングを行う [2]。

フィルタリング処理で選択されたページについて抄録を生成する。ここでは、プロファイル中の単語を含む部分テキストを抽出する単純な方式を採用した。

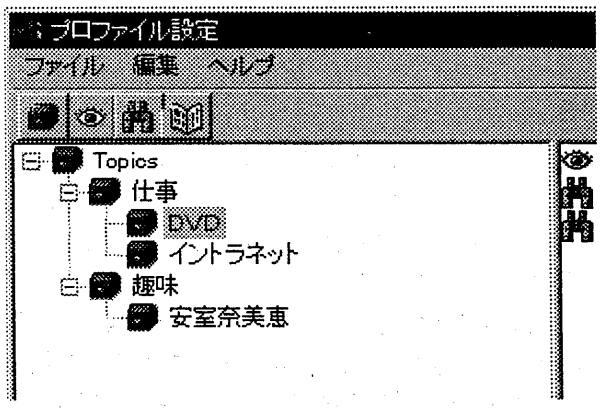
ユーザは Web ブラウザで任意の文書を参照時にその情報に関心があるかないかをフィードバック可能である。現在、複数のプロファイルが存在する場合、ユーザが参照しているページと最も類似度の大きいプロファイルの内容が更新される。

4. 動作例

図 2に動作例を図示する。(a) はフィルタリング結果の表示例、(b) はユーザプロファイルを設定するユーザインタフェース部の表示例を示している。



(a) フィルタリング結果の表示



(b) プロファイル設定のユーザインタフェース

図 2: 動作例

5. おわりに

WWW 上に公開されている情報を対象にした情報フィルタリングシステムを開発した。PC 上で動作するシステムであるため、情報ソースの設定をユーザが自由に行えるメリットがある反面、フィルタリング可能な情報量に制限がある。一部機能のサーバ化を行う予定である。

参考文献

[1] Belkin, N.J. and Croft, W.B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin?, *ACM Comm.*, Vol.35, No.12, pp.29–37 (1992).

[2] 住田一男, 三池誠司: 情報フィルタリング技術, 東芝レビュー Vol.51, No.1, pp.42–44 (1996).