WWW情報フィルタリング・検索システム(FreshEye)

3L－6　――― 全体システムの構成と動作 ―――

鈴岡節　澤島信介　上原龍也　住田一男
(株) 東芝 研究開発センター

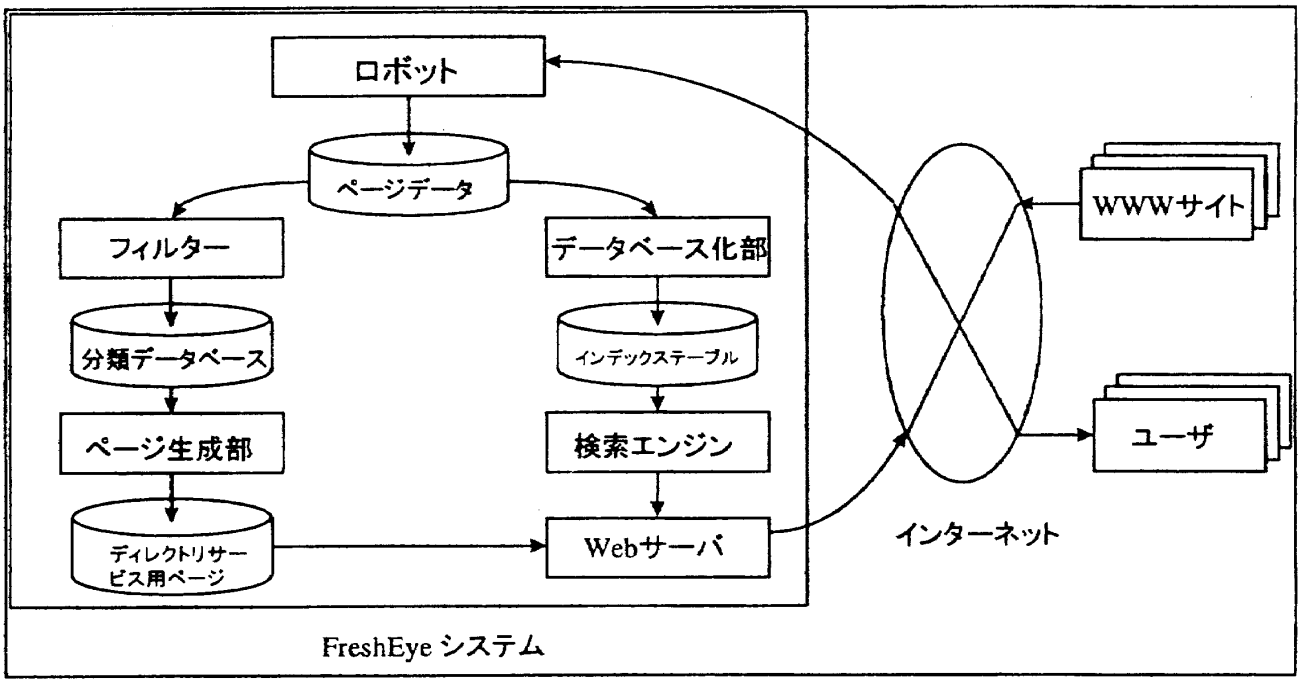
1．はじめに

一般的に人々がまず第一に関心を持つのは、1ヶ月以内のことの方が、1ヶ月以上前のことよりも多いと考えられる。しかし、ロボットを用いた、WWW 検索エンジンや機械的に情報を整理するディレクトリサービスでは古い情報が検索には威力を発揮するが、最近の話題には弱いものが多い。これは国内でもWebページの数は1千数百万あると言われており、このような膨大な数のページのデータベースの管理に手間がかかるからである。しかも、Webページの数は増大の一途を辿っており、問題を放置しておくと、ますます、古い情報が多いデータベースになる傾向がある。外国の例では、netpromote 社(http://www.netpromote.com)の調査によると、データベースの更新には、Altavista や Infoseek など著名な検索エンジンでのデータベース更新には数週間かかるという報告がある。

FreshEye[1]では、鮮度の高い情報の提供を第一義とし、この実現のために以下の方法を採った。

1. ロボットにより新規ページをよりよく見つける機能をつけた。
2. データベースの内容の期限を一ヶ月に限定した。期限よりも古くなると、そのページの情報はデータベースから取り除く。
3. アップツーデイトなディレクトリサービスを実現するために、情報フィルタリングの機能を用いて、機械的に即座に情報の分類を行う。

これにより、データベースの更新を10時間にまで縮め、最短で半日前の情報をも検索可能としている。



FreshEye: An information Filtering and Search System for WWW ――― System Configuration ―――

Takashi SUZUOKA, Nobuyuki SAWASHIMA, Tatsuya UEHARA, and Kazùo SUMITA
Research and Development Center, Toshiba Corp., 1 Komukai-Toshiba-cho, Saiwai-ku, Kawasaki, 210-8582, Japan

2. 構成

全体の構成図を前ページに示す。

・ロボット：日本語のページ(教育機関系、新聞社は除く)を巡回し、Webページを収集する。

・データベース化部：1ヶ月以内に生成されたと考えられるページのみをデータベース化し、それ以上古いと思われるデータはデータベースから取り除く。

・検索エンジン：ワードベースの検索エンジン。

・フィルター：[2]での技術を用いた情報フィルタリングで、一日に一回(現状)、三百数十のトピックについてページのふるい分けを行う。

・ページ生成部：フィルタリングされた結果をユーザが見易いようにHTML形式のページとして生成する。

・Web サーバ：http://fresheye.toshiba.co.jpが、検索やディレクトリサービスを対応する。

3. 評価

サッカーのワールドカップ・フランス大会の決勝は日本時間で1998年7月13日に行われたが、その日に「ワールドカップ」で検索した。

評価のポイント

- 1998年フランスで行われたサッカーのワールドカップに直接関係するものが上位30件中いくつとれたかで評価。
- サッカー以外のワールドカップは除外。
- サッカーのワールドカップの話題でも、2002年の日韓共同開催、地区予選、フランス大会前に書かれた情報は除外。
- フランス大会に関するものでも、単に日記やチャットの一部で、サッカーのワールドカップ情報としての重要性が低いと考えられるものは除外。

日本語が扱えるWWW上で公開されている主要な9のロボット系の検索エンジンで試してみたが、上位30位のうち、適合するのは多くても2件であった。このとき、日付順に結果をソートできるものは、新しいもの順で評価した。

これに対してFreshEye searchでは上位30件のうち、16件が内容として適合した。以下に詳細情報を記す。

一ヶ月以内で4289件にマッチ。一週間前で560件マッチ。

7/13　0/1 ... 1件あったが日記であり、ワールドカップとしての情報はなし。

7/12　16/30 ... 30件中16件は情報があった。

7/11　42件 ... 42件マッチ

7/10　79件 ... 79件マッチ

4. おわりに

新鮮な情報の検索を得意とすることには二つの利点がある。一つは、新しい情報が検索できることは明らかである。もう一つは、時事性の高い話題の検索に関しては細かく話題を指定しなくてよいということである。

先の例のワールドカップで言えば、FreshEyeの検索データベースにはワールドカップといえば、1998年のサッカーのフランス大会の情報ばかりであろうから、「1998」、「サッカー」、「フランス大会」、「not テニス」、「not ゴルフ」などの語で検索を補助する必要がない。このような時事性による自然な絞り込みが行われているため、ユーザが検索を行う場合でも、フィルタのプロファイルを作る場合にも役に立つ。

参考文献

[1] 住田他、WWW情報フィルタリング・検索システム (FreshEye) － サービス概要 － 情報処理学会 第57回全国大会, 1998

[2] 酒井他、情報フィルタリングシステムNEATのための検索要求文からのプロファイル生成、情処研会資FI-47-12, pp.83-88, 1997