WWWの分類・検索システム Crowww

6Z－6

伊藤 篤　Phyllis Anwyl　大野 亜矢子

(株) リコー 情報通信研究所

1 はじめに

WWWの検索サービスの多くは日本や世界中といった広い範囲の検索をカバーする。一方、特定のサイトや組織のWWWページに限って検索できるサービスをそのWWWページの付加価値として行なうことがある。これは、更新を反映する速度や表示構成のカスタマイズの点において前者より優れている。

ディレクトリ型の検索サービス[1]は、検索意図のはっきりしていない情報散策を行なう場合に有効である。ところが、ディレクトリのような分類体系を保守するのは非常に手間のかかる作業である。一方、WWWロボット型のように自動収集を行なう場合は分類体系の構築は困難であり、また分類する必要のないゴミ情報を多く拾うことにもなる。

そこで、これらの問題を解決するために、WWW文書の属性やメタ情報を利用し、あらかじめ定義しておいた分類体系上に自動的に仕分けをし、検索を行なうシステム (Crowww) を作成し、WWWサーバー上に実装した。(URL http://www.ricoh.co.jp/search/)

2 システム構成

図1にシステム構成を示す。収集部は特定の範囲のWWW文書をリンクを辿ることで収集する。解析部は収集したWWW文書を種々の観点から解析し、その文書の属性として付与する。分類部はあらかじめ定義した分類体系に沿って文書を分類し、検索部は分類されたWWW文書の検索を行なう。表示部は、分類されたデータや検索されたデータの表示をHTMLに変換する。

実行画面を図2に示す。

Crowww - Classification and Retrieval on WWW
ITOH Atsushi, Phyllis Anwyl, OONO Ayako
RICOH Co., Ltd.

3 分類

いくつかの観点からの分類を用意することで、ユーザに情報散策を行なう機会を提供することができる。

Crowwwでは木構造状の分類体系をあらかじめ定義[2]しておき、収集したWWW文書を自動的に分類する。

分類のキーには原則としてメタ情報を使用する。これは、HTMLのMETAタグによって書かれるもので、キーワードや注釈文などを書くことができる。メタ情報を使って分類することはWWW文書の作成者側と検索サービスの管理者側の両方で分類作業を分業することであり、作業が効率化される。

メタ情報が文書に付けられていない場合はキーワード抽出や重要文抽出[3]によって仮想的にメタ情報をWWW文書に付与する。

注釈文は、分類された文書の一覧を表示するときにコメントとして使用される。

分類項目には、日付別などのように文書属性によって決められるものと、文書の内容によって決められるものがある。

(1) 属性による分類

文書の属性を使って分類する。ここで属性とは、HTTPヘッダの項目、TITLE、メタ情報および、解析部によって自動的に抽出され付与される任意の属性からなる。これらは体系作成の手間が簡単な割に、ユーザの複数の視点を提供するという点で効果が高い。

Crowwwではデフォルトの分類項目として次のものを用意している。

What'sNew: 最新の文書だけを集めたもの
更新日別: 更新日を年、月で分けたもの
ディレクトリ: 実際のファイルの位置
メールアドレス別: 問い合わせ先で分けたもの
画像ファイル: IMGタグで埋め込まれた画像をファイル形式で分けたもの

(2) 内容による分類

主としてキーワードを基に分類を行なう。キーワードは原則としてメタ情報を利用するが、メタ情報が無い場合にはキーワード抽出によって補う。

内容別の分類体系は、対象とするWWWページに合わせて設計されるべき部分であるが、一度作成すれば仕分けは自動である。

Crowwwでは次の分類項目を用意した。

製品別: 企業の製品に合わせてさらに下位の分類項目が用意される

趣味: 「音楽」「写真」といったキーワードで分類する

4 検索

検索機能には全文検索[4]を使用した。データが事前に分類されていることを利用して、

- 分類の各カテゴリーからの検索
- 検索結果の、分類体系に合わせた表示と絞り込み機能

を実現した。

5 表示

プログラムとページデザインを分離するため、テンプレートを使って、分類状態や検索結果をHTML文書に変換し出力するようにした。テンプレートは変数と制御構造を持ったHTML形式のファイルである。

分類状態は木構造を持つが、構造のどこを指しているかで、テンプレートの評価を変えるようにした。これをコンテキストと呼ぶことにする。コンテキストによって、テンプレート中の変数の値が変化し、分類状態によって、デザインや表示されるデータが変わる。

テンプレートの変更だけでGUIやデザインを変えられ、いろいろなサイトや目的に適応できる。

6 おわりに

1997年10月から12月の約2ヶ月間の、Crowwwへのアクセスログを集計し、評価した。図3に各カテゴリへのアクセスの回数とどこのカテゴリから検索を行ったかを示す。約半数が分類されたページにアクセスし、3分の1以上がカテゴリを選択してから検索を行っていることがわかる。これにより、検索時の分類機能の有効性を確認した。

以上のように、メタ情報や内容により分類・検索を行なうシステムを作り、検索時の分類の有効性を確認できた。

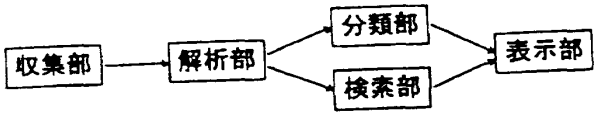

図1 システム構成

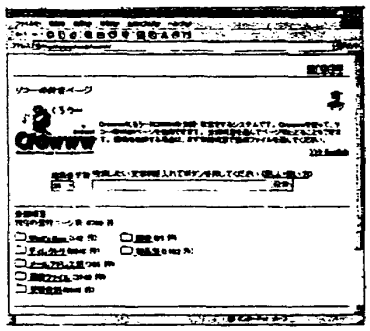

図2 実行画面

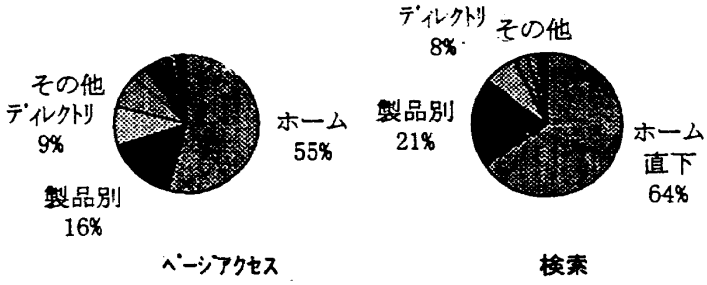

図3 検索アクセスログ集計結果

参考文献

[1] 田村健人. Wwwの検索サービスは何をしているのですか？ 情報処理, Vol. 38, No. 12, pp. 1099–1100, Dec 1997.

[2] 伊藤篤. WWWを用いたNetNewsの分類サービスシステム NETCoW. 情報処理学会第52回全国大会予稿集, Vol. 1, p. 143, Mar 1995.

[3] 亀田雅之. 擬似キーワード相関法による重要キーワードと重要文の抽出. 言語処理学会 第2回年次大会, 1996.

[4] 岩崎雅二郎, 小川泰嗣. 文字成分表による文字列検索の実現と評価. 情報処理学会 研究会報告 (93-DBS-92), pp. 1–10, Mar 1993.