1U-01

目的特化型 WWW サーチエンジンの開発

福島 俊一　石黒 義英　喜田 弘司　山田 洋志　松田 勝志

NEC

1. はじめに

インターネット上には膨大な数の WWW 情報が発信されており、これらは利用者の様々な問題解決目的に有用な情報源となる[1][2][3]。しかし、多くの WWW サーチエンジンで提供されているキーワード検索機能では、利用者の目的に合った WWW 情報へ的確に絞り込むことが難しい(適合率が低い)。人手によって分類・選別された特定目的向け検索サイトでは、規模・情報量に限界がある(再現率が低い)。

このような問題に対して、目的を特化することで検索精度(適合率&再現率)を高めようというアプローチが考えられる。商品価格を比較する ShopBot、個人ホームページを見つける Ahoy!、FAQ を自然言語で検索する FAQ Finder などの先行研究がある[1]。

筆者らは、同様のアプローチをとりながらも、目的特化型 WWW サーチエンジンとして、より汎用性の高いフレームワークを目指している。本論文では、ページタイプ選別[3]＋情報抽出・分類[4] を共通のフレームワークとし、複数種類の目的特化型 WWW サーチエンジンを試作した事例について報告する。

2. フレームワーク

本論文で提案する目的特化型 WWW サーチエンジンのフレームワークは、ページタイプ選別処理と情報抽出・分類処理を組み合わせたものである(図-1)。ページタイプ選別処理[3]では、個々の問題解決目的に適合したページタイプの WWW 情報のみを検索対象として切り出し、情報抽出・分類処理[4]では、それらを目的に応じた分類軸で検索可能にする。

2.1 ページタイプ選別処理

個々の問題解決目的に対して、多くの場合、それに適合する WWW 情報のページタイプが定められる。例えば、商品購入という目的に対する「商品カタログ」

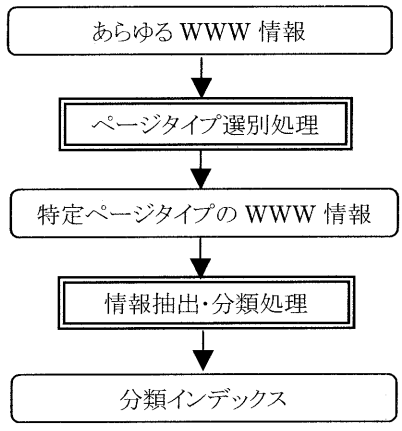

図-1 目的特化型 WWW サーチエンジンのフレームワーク

タイプや「オンラインショップ」タイプ、就職・転職という目的に対する「求人情報」タイプなどである。ここでいうページタイプは、テキストの意味内容に深く立ち入らないと判別できないような分類ではなく、むしろ、WWW ページを一見して得られるような外観的な特徴に基づいて判別できる性格のものである。ページタイプの判定には、ページ内の特徴的なキーワードに加えて、URL 文字列、HTML タグ構造、リンク数、画像サイズなど、スタイル的なファクタも用いる。ページタイプ選別処理では、種々雑多な WWW 情報のなかから、特定のページタイプに該当するもののみを取り出す。

2.2 情報抽出・分類処理

この段階では、特定のページタイプをもつ WWW 情報のみが処理対象になる。このとき多くの場合、そのページタイプに固有のキー項目が存在する。例えば、「イベント情報」タイプに対する開催日や開催地、「求人情報」タイプに対する勤務地や職種などである。これらのキー項目は、利用者の問題解決目的に即した分類軸となる。処理対象の WWW 情報が特定のページタイプをもつことを前提とすれば、辞書とパターンマッチルールから成る対象知識を用意することによって、上記キー項目の高精度抽出が可能となる。情報抽出・分類処理では、特定のページタイプをも

Development of Specialized WWW Search Engines
Toshikazu Fukushima, Yoshihide Ishiguro, Koji Kida, Hiroshi Yamada, and Katsushi Matsuda
NEC Corporation
t-fukushima@cj.jp.nec.com

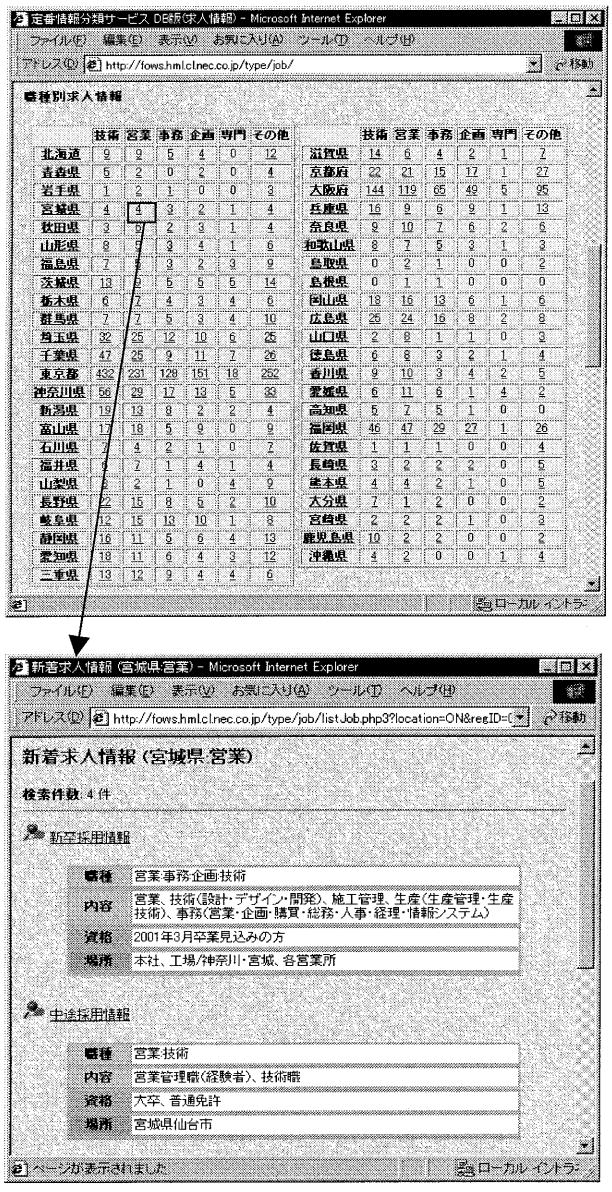

図-2 求人情報サーチエンジンの画面例

つ WWW ページから、キー項目に該当する個所を抽出し、そのキー項目の内容(値)で各 WWW ページを分類する。結果は分類インデックスに書き込む。

3. 試作システム

本論文で提案するフレームワークに基づき、求人情報、イベント情報、モバイルコンテンツの各々に特化した WWW サーチエンジンを試作した。

(1) **求人情報サーチエンジン**:「求人情報」タイプのページを選別し、勤務地や職種で分類。

(2) **イベント情報サーチエンジン**:「イベント情報」タイプのページを選別し、開催日や開催地で分類。

(3) **モバイルコンテンツサーチエンジン**:「i-mode」タイプのページを選別し、場所で分類[5]。

図-2 に求人情報サーチエンジンの画面例を示す。利用者は、勤務地と職種の条件を分類表(図-2 上)から選択すると、その条件に該当する WWW ページの抜粋情報(図-2 下)が得られる。この画面からさらにオリジナルの WWW ページへジャンプすることもできる。これらの画面は、図-1 の分類インデックスに格納された情報に基づいて自動生成している。

4. 精度評価

「求人情報」タイプと判定されたページのうち、新しい順に 300 ページを評価対象とした。「求人情報」タイプ以外が 6 ページあり、タイプ選別の適合率は 98%となった(数件のキーワード検索結果との比較によって擬似的に求めた再現率は 66%)。残りの 294 ページに出現した勤務地 224 件と職種 261 件の抽出精度は、勤務地の再現率 57.6%、適合率 100%、職種の再現率 47.5%、適合率 92.5%となった。イベント情報についてもほぼ同程度の精度が得られている。モバイルコンテンツについては[5]で報告する。

5. おわりに

ページタイプ選別＋情報抽出・分類という目的特化型 WWW サーチエンジンのフレームワークを提案した。また、このフレームワークに基づき、求人情報、イベント情報、モバイルコンテンツの各々に特化した WWW サーチエンジンを試作し、アプローチの有効性を確認した。今後、ページタイプ選別や情報抽出・分類の精度改善に取り組んでいくと同時に、このフレームワークを様々な問題解決目的に適用・実用化を進めていく予定である。

参考文献

[1] O. Etzioni, The World-Wide Web: Quagmire or Gold Mine? CACM, Vol.39, No.11, 1996.

[2] 藤本・他、DSIU システム: Decision Support for Internet Users、人工知能学会誌、15 巻 1 号、2000 年。

[3] K. Matsuda, et al., Task-Oriented World Wide Web Retrieval by Document Type Classification, Proc. of CIKM'99, 1999.

[4] 山田・他、Web ページからのタイプ別情報抽出・分類方式、情処 60 全大、1N-6、2000 年。

[5] 喜田・他、モバイル位置指向サーチエンジンの開発、情処 61 全大、1U-2、2000 年。