文書タイプ分類による問題解決のためのWWW検索システム

4T－2

松田 勝志　福島 俊一
NEC ヒューマンメディア研究所

1 はじめに

WWW(World Wide Web)の普及に伴い，問題解決向けの様々な検索サービスがインプリメントされてきている．これらのサービスは従来の汎用のキーワード検索ではなく，特定の分野やタスクに特化した精度の良い検索サービスが多い[1][2][3]．一般ユーザにとって的確なキーワードを決めることは困難なため，汎用のキーワード検索では検索結果にゴミが多くなってしまう．今後はAskJeeves[4]のような特定の分野やタスクに特化した検索サービスが増えてくるであろう．

筆者らが開発している問題解決向けWWW検索システムでは，カテゴリー検索で使われる一般的な概念構造であるカテゴリーとは違い，特定の問題解決に利用できるコンテンツの種類である文書タイプという概念を導入した．このシステムでは，WWWページをあらかじめその文書タイプに固有の構造的な特徴をもとにそれらの文書タイプに分類しておくことによって，検索時に問題解決の種類に応じた文書タイプを指定することで的確な検索ができる．前稿[5]では，システムのアイデアと簡単な評価実験結果について報告した．

本稿では，筆者らが導入した文書タイプという概念について詳しく述べ，またどのような文書タイプがあるのかについて列挙する．また，実験規模と文書タイプを拡大し，本システムの実用性と拡張性を明らかにする．

2. 文書タイプ

WWW検索サービスを利用するユーザは，何らかの問題解決を行うことを目的としていることが多い．例えば，パソコン購入や旅行計画などである．そして実際これらの問題解決に役立つオンラインショッピングやトラベルのサイトへのリンクを集めたポータルサイトがある[6]．しかし，このような問題解決すべてについて質の高いリンクを用意するのはコストの面から非常に困難である．そこで筆者らは，ある問題解決にはその問題解決に応じて要求されるコンテンツのタイプがあり，そのタイプはある種の固有なページのスタイルを持っているのではないかという仮説を立てた．例えば，購入計画という問題解決にはカタログのようなものが要求され，カタログにはそのページがカタログであると同定できる固有なスタイルを持っている，ということである．このタイプが文書タイプである．

文書タイプはある問題解決に使えるページの集合であるため，ユーザがある問題解決に直面した場合，容易に文書タイプを指定することができる．例えば，上記の購入計画ならばカタログであり，就職や転職ならば求人案内である．またパーソナルユースにはプレゼントやチャットなどが文書タイプとしてある．

以下にビジネスユース，パーソナルユースでの文書タイプの例を示す．

ビジネスユース	パーソナルユース
カタログ	
オンラインショップ	
FAQ	
リンク集	
調査報告	料理レシピ
求人案内	プレゼント
事例	教室・講座
イベント情報	アップデートプログラム

表1．文書タイプの例

これらのような文書タイプをあらかじめ用意しておくことによって幅広いユーザの検索要求に対応することができる．

3. 文書タイプ分類

WWWページを文書タイプに分類するには，その文書タイプに固有の構造的特徴を利用する．従来さまざまな文書分類の研究がなされている[7]が，それらは文書中の単語のみに着目したものが多い．しかし，WWWのHTML文書にはさまざまな付加情報(タグ，イメージ，ハイパーリンク等)が内包されている．実際，ユーザはWWWのページを一瞥するだけでそのページがカタログであるか掲示板であるかということが判断できる．これはそれらの文書タイプに応じたデファクト的なページの形式や最低限の項目や要件等が存在するためである．例えば，カタログであれば，製品名が目に付き易い大きさで表現され，その製品の画像があり，仕様や特長を記したページへのリンクがある，などである．

文書タイプへの分類はこのようなページの構造的な特徴をもとに分類する．実際には，各文書タイプ毎に構造的特徴ルールを記述した特徴記述を用意し，WWWページ毎にその特徴記述を満足している度合を調べ，各

WWW Retrieval System for Problem Solving by Document Type Classification
Katsushi MATSUDA and Toshikazu FUKUSHIMA, Human Media Research Laboratories, NEC.

文書タイプへのタイプ適合度を求める．タイプ適合度は0～100までの数値として算出する．

検索時には，ユーザが入力したキーワードでの検索結果集合とユーザが指定した文書タイプがある一定値(タイプ閾値)以上の検索結果集合との論理積が検索結果となる．

4. 評価実験

4.1 特徴記述作成と分類速度

前稿では文書タイプを 2 種類(カタログ，リンク集)としたが，本評価実験では文書タイプを 4 種類(調査報告，求人案内，プレゼント，アップデートプログラム)追加し，計 6 種類とした．追加した文書タイプの特徴記述の作成は，チューニングも含めてそれぞれのべ 5 時間程度であり，作成した特徴記述は 400～1,000 バイト程度であった．

1,000 ページ約 7.5M バイトのデータを文書タイプ 6 種類に分類するのに費やした時間は，EWS4800/460 (R10000, 200MHz)で約 140 秒であった．

このように少ないコストで文書タイプを追加することが可能であり，また，十分実用的な速度で分類することが可能であることがわかった．

4.2 分類精度

6 種類の文書タイプで分類精度の評価実験を行った．カタログとリンクの文書タイプについては NETPLAZA[8]の自動収集データを用い，その他の文書タイプについては別に収集した約 15 万件の実データを用いた．すべての文書タイプについてタイプ閾値を 50 以上とした．

実験では，文書タイプ毎に 1 単語のクエリー(基本キーワード)を用意し，その基本キーワードと文書タイプ指定，基本キーワードと文書タイプに代わる 1 単語キーワードの双方について検索結果の先頭 20 個での分類の適合率を求めた．例えば，基本キーワード"LaVie"と文書タイプ「カタログ」の論理積と，基本キーワード"LaVie"とキーワード"カタログ"の論理積である．分類の適合率とは，検索結果中に含まれる正しい文書タイプであるページの割合である．

実験結果を図 1 に示す．この結果からも明らかなように，ある文書タイプのページを検索する場合，キーワードのみより本システムの文書タイプ指定の方が精度が良い．すなわち，構造的な特徴に着目する文書タイプ指定検索が問題解決のための検索に有効であることがわかった．各文書タイプについてそれぞれ 7 種類の基本キーワードで実験を行い，その平均値をプロットしているが，すべてにおいて文書タイプ指定の検索手法の方が分類の適合率が良かった．また，実験で用いた検索システムでは，検索結果を単に文書登録の逆順で表示するようにしたため，先頭 20 件の比較は平均的な分類精度の比較に相当する．検索結果をタイプ適合度でソートすれば，先頭 20 件の精度は更に向上させることが可能である．

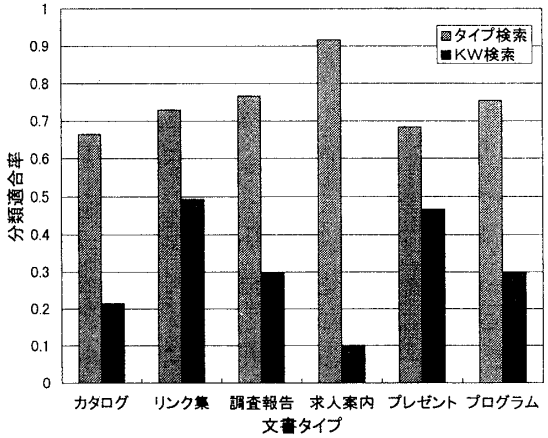

図 1. 実験結果

5. おわりに

さまざまな問題解決に利用できる文書タイプという概念を導入した WWW 検索システムについて述べた．今後は，ビジネスからパーソナルの幅広い問題解決に応じた検索サービスが要求されるであろう．本稿で述べた WWW 検索システムは，少ないコストでさまざまな分野やタスクに特化した質の高い検索サービスを提供することができる．評価実験の結果，文書タイプと同一キーワードを付加した検索より十分高い分類の適合率を達成することができ，ある文書タイプのページを検索するという問題解決のための検索に有効であることがわかった．

本検索システムは NETPLAZA で既に実用化されている．現状は文書タイプが 2 種類であるが，徐々に増やしていく予定である．

参考文献

[1] 冨田ほか:HTML 文書からの商品情報抽出方式の提案，情報処理学会第 56 回全国大会予稿集(3), pp.79-80, 1998.

[2] J. Shakes, et al: Dynamic Reference Sifting: A Case Study in the Homepage Domain, In Proceedings of 6th WWW, pp.189-200, 1997.

[3] R. Burke, et al: Question Answering from Frequently Ask Question Files: Experiences with the FAQ Finder System, Univ. of Chicago, Dept. of CS, TR-97-05, 1997.

[4] http://www.askjeeves.com/

[5] 松田，福島:インターネット多角的検索システム OTROS-構造的特徴量によるタイプ分類と検索-，情報処理学会第 57 回全国大会予稿集(3), pp.145-146, 1998.

[6] http://www.excite.com/

[7] H. Schutze, et al: A Comparison of Classifiers and document representation for the routing problem, In Proceedings of 18th SIGIR, pp.229-237, 1995.

[8] http://netplaza.biglobe.ne.jp/keyword.html