5T-5

Web 文書のページタイプを用いた適応的分類の拡張と評価

畠山恭佑†，高山毅†，村田嘉利†，池田哲夫††，浅沼直人†
岩手県立大学ソフトウェア情報学部† 静岡県立大学経営情報学部††

1. はじめに

検索エンジンの検索結果をユーザが理解しやすいようグループ分けして表示する研究が広く行なわれている．

・分類：既定のカテゴリ階層の一部を抜き取りカテゴリ名として採用する[1]
・クラスタリング：検索結果と検索キーワードからカテゴリ名を動的に生成する[2]

等である．しかし，カテゴリの分け方に対するユーザの満足度は充分とは言えない．著者らはこれまでに，出現率が相対的に高いページタイプ五つの中から，ユーザが指定したもののみをカテゴリとして採用し適応的に分類を行なう手法を提案し，有効との評価を得ている[3][4]．本稿では新規ページタイプの追加や識別方法の変更などの拡張のほか，検索結果を分類した上で検索結果の補集合を更に分類する補集合分類という概念を取り入れ，評価実験によりその有用性を示す．

2. Web 文書のページタイプでの適応的分類[4]

2.1 検索エンジンの使用目的

検索エンジンの使用目的として，以下の二つに注目している．

(1) [学習目的]：仕事や学習のための検索
(2) [買い物目的]：商品購入のための検索

2.2 検索キーワードを問わず頻出するカテゴリ

学習目的と買い物目的で，Google の検索結果上位 100 件を手動で分類した結果から，頻出する以下の五つのページタイプを分類カテゴリとして使用することを提案している．

カテゴリ 1：特定のアプリケーションを必要とするページ(pdf，xls，など)
カテゴリ 2：掲示板，日記，チャット，Blog
カテゴリ 3：書籍の紹介
カテゴリ 4：ショッピングサイト
カテゴリ 5：ニュースサイト

2.3 カテゴリの取捨選択

「ページタイプ直接選択方式」と，「ページタイプ間接選択方式」の二つを提案している．前者は，カテゴリとして採用するページタイプをユーザが直接選択する方式である．また後者はユーザが検索目的を選択すると，カテゴリが間接的に選択される方式で，具体的には，マッピング関係を

学習目的 ⇒ カテゴリ 3，4
買い物目的 ⇒ カテゴリ 1，5

としている．

2.4 分類の実現方法

各ページタイプの識別アルゴリズムは経験則により設定している．GoogleAPI[5]を用いて検索を行ない，識別アルゴリズムを用いて Web 文書を，選択したページタイプからなるカテゴリに登録している．

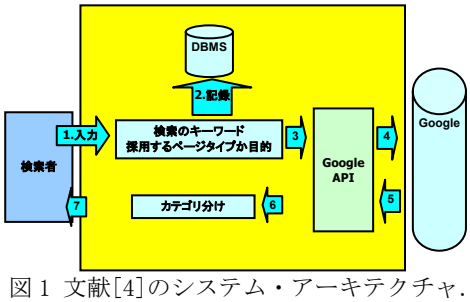

図 1 文献[4]のシステム・アーキテクチャ．

3. 本稿での機能拡張

以下の六つの拡張を行なう．

3.1 カテゴリとするページタイプの追加

以下の五カテゴリを追加する：「カテゴリ 6: 公的機関ページ」「カテゴリ 7: wiki を含む辞書ページ」「カテゴリ 8: リンク集」「カテゴリ 9: トップページ」「カテゴリ 10: DeepWeb」．

3.2 検索目的とページタイプの再マッピング

ページタイプの追加により，必然的にマッピング関係を作り直す必要がある．アンケート結果を基に以下のマッピングへ変更する．

学習目的 ⇒ カテゴリ 3，4，8
買い物目的 ⇒ カテゴリ 1，6，7，8

3.3 分類における，適合率と再現率の向上

文献[4]では Web 文書を排反で分類していたため，本来入るべきカテゴリに属さない場合が発生するという問題があった．本稿では，一つの Web 文書が複数のカテゴリに重複して分類されることも可能とする．

3.4 二段階分類の提案

3.1 項で，ページタイプを 10 に増やしたため、結果としてその選択が煩雑となった．そこで，「検索キーワードとページタイプを選択し，それらにより分類を行なう検索結果分類」と「分類された検索結果の補集合を残りのカテゴリで更に分類する補集合分類」という，二段階分類を提案する．

Extension and Evaluation of the Adaptive Classification of Web Documents with Page Type
†K.Hatakeyama, T.Takayama, Y.Murata and N.Asanuma
†Faculty of Software and Information Science, Iwate Prefectural University
††T.Ikeda ††School of Administration & Informatics, University of Shizuoka

3.5 ページタイプ直接選択方式における，ページタイプの割り振りのカスタマイズ機能

10 種のページタイプを，検索結果分類と補集合分類のどちらで選択できるようにするかを、ユーザが自由に変更できるようにする．

3.6 ページタイプ間接選択方式における，マッピングのカスタマイズ機能

3.2 項のマッピングをカスタマイズできるようにする．

4. 評価

4.1 被験者による主観的評価

学習目的，買い物目的の検索を 8 問用意し，被験者に以下の 6 つの検索エンジンを使用して回答してもらう．

1. 文献[4]のページタイプ直接選択方式を導入した Google(以降，「旧ペ直」と略)
2. 拡張型ページタイプ直接選択方式を導入した Google(以降，「新ペ直」と略)
3. 被験者がカスタマイズを行なった拡張型ページタイプ直接選択方式を導入した Google(以降，「新ペ直カ」と略)
4. 文献[4]のページタイプ間接選択方式を導入した Google(以降，「旧ペ間」と略)
5. 拡張型ページタイプ間接選択方式を導入した Google(以降，「新ペ間」と略)
6. 被験者がカスタマイズを行なった拡張型ページタイプ間接選択方式を導入した Google(以降，「新ペ間カ」と略)

評価項目は以下の 4 尺度で，7 段階(1:とても悪い～7:とても良い)の主観的評価を行なう．

尺度 1. カテゴリの分け方が妥当で，わかりやすいか

尺度 2. 検索結果から得られた情報は，満足するものか

尺度 3. 検索目的とは関係のないページタイプがどのくらい目についたか

尺度 4. 必要な Web 文書が一つのカテゴリに集中していて，そこさえ見れば必要な情報を充分得られるか

実験結果（図 2，3）より，[学習目的]では尺度 3，4 で，[買い物目的]では全ての尺度で新ペ直，新ペ間双方が有効な評価を得ている．

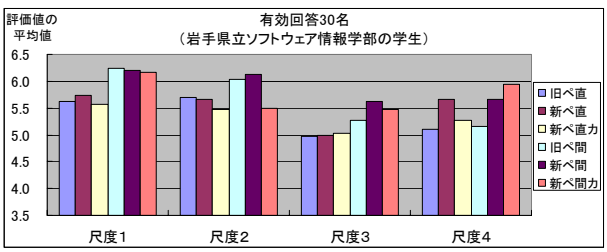

図 2 [勉強目的]での評価結果.

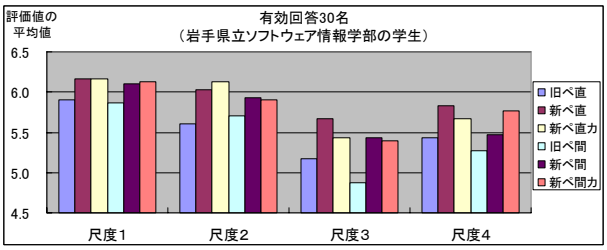

図 3 [買い物目的]での評価結果.

4.2 定量的評価

検索用キーワードを計 60 個用意し，旧ペ直と新ペ直で適合率/再現率を比較する．

実験結果（図 4，5）より，適合率ではカテゴリ 1，2，5 で，再現率では全てのカテゴリで新ペ直が有効な結果を得ている．

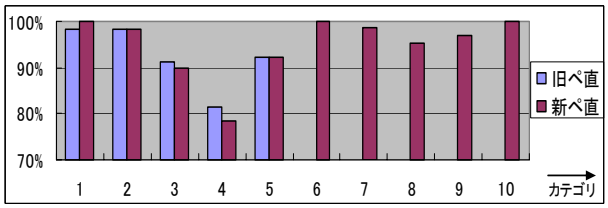

図 4 適合率の評価結果.

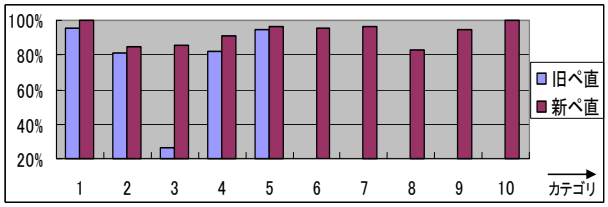

図 5 再現率の評価結果.

5. 結論と今後の展望

本稿では Web 文書のページタイプを用いた適応的分類の拡張と評価を行なった．評価実験の結果，その有効性を確認できた．今後の展望として，i) 補集合以外のカテゴリを分類する機能の追加，ii) 学習目的，買い物目的以外の検索目的の追加等があげられる．

参考文献

[1] 安形ほか:「WWW ページの自動分類:NDC の分類体系と Yahoo のカテゴリを使った分類」，情処研報，FI-54 (1999).

[2] 成田ほか:「階層的クラスタリングを利用したメタ検索エンジンの提案」，情処研報，DBS-128-50 (2002).

[3] 金子ほか: ”Web 文書のページタイプを用いた適応的分類と評価”，日本知能情報ファジィ学会誌，Vol.18，No.2，pp.319-336 (2006).

[4] 長内ほか: ”Web 文書のページタイプを用いた適応的分類の拡張と評価”，第 68 回情処全大，3N-3 (2006).

[5] GoogleAPI，
http://www.google.com/apis/index.html