3N-3

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

長内亘†，高山毅†，池田哲夫†，畠山恭佑†，金子大輔††
岩手県立大学ソフトウェア情報学部† 三菱スペース・ソフトウェア株式会社††

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進行している．

1. 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，「分類」[1]
2. 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，「クラスタリング」[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]．本研究ではページタイプの選定や識別においてこれをさらに拡張し，実験評価により有効性を示す．

2. Web 文書のページタイプでの適応的分類[3]

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 検索の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[4]の検索結果上位 100 件を手動で分類した結果から，以下の五つのページタイプを分類カテゴリとして使う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
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：シラバス

2.3 検索目的による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，検索目的をユーザが選択することによって，分類カテゴリを間接的に選択する方式である．具体的には，マッピング関係を

学習目的 ⇒ カテゴリ 1，3，5
買い物目的 ⇒ カテゴリ 3，5

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 ページを，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．

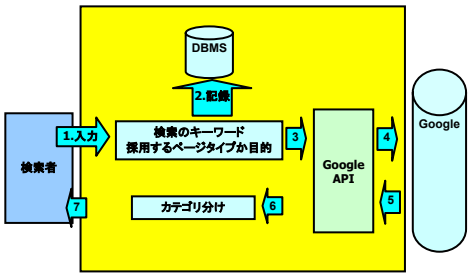

図 1 文献 3)のシステム・アーキテクチャ．

3. 本稿での機能拡張

以下の三つの拡張を行なう．

3.1 カテゴリとするページタイプの再検討

文献 3)での知見から，カテゴリ 5 を外す．代わりに，検討結果に基づき「カテゴリ 6：ニュースサイト」を単独のページタイプとして採用し，「Blog」をカテゴリ 2 に追加してカテゴリ 2’とする．

3.2 検索目的とページタイプのマッピング関係の再検討

2.3 項で述べたマッピング関係には不可解な部分がある．また，ページタイプの入れ替えによって必然的にマッピング関係を作り直す必要もある．そこで，表 1 の再アンケートの結果を基にマッピングを再構築する．

表 1 マッピング関係のアンケート結果

＼検索目的 選択すべきカテゴリ	勉強目的	買い物目的
カテゴリ 1	36.0%	76.0%
カテゴリ 2’	36.0%	20.0%
カテゴリ 3	68.0%	36.0%
カテゴリ 4	88.0%	0.0%
カテゴリ 6	24.0%	52.0%

3.3 分類における，適合率と再現率の向上

文献 3)では，ページタイプの識別での問題

Extension and Evaluation of the Adaptive Classification of Web Documents with Page Type
†W.Osanai, T.Takayama, T.Ikeda and K.Hatakeyama
† Faculty of Software and Information Science, Iwate Prefectural University
††D.Kaneko ††Mitsubishi Space Software Co., Ltd.

点として，以下二つがあげられる：

1. Blog で書かれた日記サイトが識別できない
2. 書籍のページがショッピングサイトに含まれる

本稿では URL のドメイン名やパターンによる識別を追加することで，適合率と再現率の向上を図る．なお向上によって処理時間がかかり過ぎ，応答のリアルタイム性が損なわれないように注意しつつ行なう．

4. 評価

4.1 被験者による主観的評価

学習目的の例として大学入試および Web ラーニングの問題 6 問，買い物目的の例として指定した条件に見合う商品を探す問題 6 問を用意し，被験者に以下の 6 つの検索エンジン使用して回答させる．

1. 文献 3)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指定方式を導入した Google(以降，「新ペ直」と略)
3. 文献 3)のページタイプ間接指定方式を導入した Google(以降，「旧ペ間」と略)
4. 拡張型ページタイプ間接指定方式を導入した Google(以降，「新ペ間」と略)
5. Yahoo[6]
6. Google

そして，以下 4 つの尺度で 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 ページがカテゴリを横断しておらず，一つのカテゴリに集中していて見つけやすいか

尺度 1，4 は，カテゴリに関する評価なので，Google 使用時と，Yahoo 使用時でカテゴリが出現しなかった場合は評価不能とする．

実験結果は図 2，3 のとおりである．[学習目的]では尺度 1，2，4 で新ペ直，新ペ間双方が有効な評価を得ている．[買い物目的]も，微差ではあるが新ペ直，新ペ間双方が有効な評価を得ている．

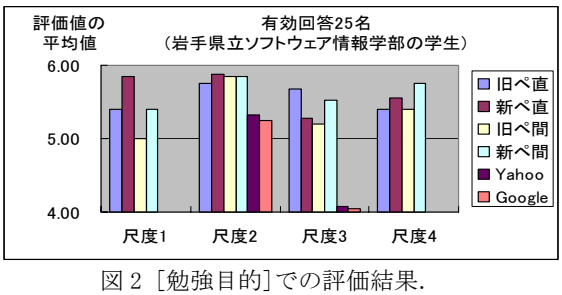

図 2［勉強目的]での評価結果.

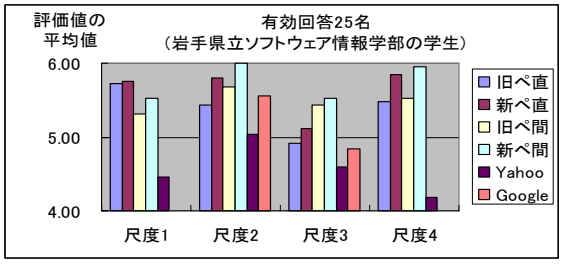

図 3［買い物目的]での評価結果.

4.2 定量的評価

検索用キーワードを計 60 個用意し，旧ペ直と適合率/再現率を比較する．

実験結果は図 4，5 のとおりである．ほぼすべてのカテゴリで，適合率/再現率双方とも，新ペ直が有効な結果を得ている．また，新規導入したカテゴリ 6 の適合率/再現率はともに 90%以上となっている．

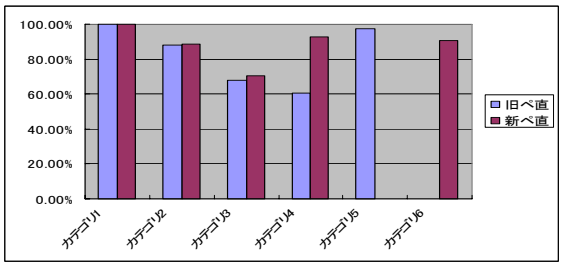

図 4 適合率の評価結果.

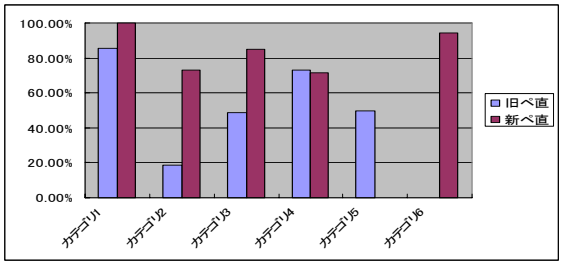

図 5 再現率の評価結果.

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) 買い物目的で，より難しい問題を用いた場合の評価，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: ”Gooots - 検索目的に沿ってカテゴリ名を取捨選択する検索エンジン”，第 67 回情処全大，4U-1 (2005)．

[4] Google, http://www.google.com

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

[6] Yahoo, http://www.yahoo.co.jp