検索結果クラスタリングのためのページタイプの自動識別*

姜 ナウン 新里 圭司 黒橋 禎夫

京都大学大学院 情報学研究科†
{naun, skeiji, kuro}@nlp.kuee.kyoto-u.ac.jp

1 はじめに

現在，ウェブは情報の収集と発信の重要な対象となっており，その情報の量は爆発的に増加している．ウェブから情報を得るため，ユーザは検索エンジンを使用しているが，既存の検索エンジンでは入力する検索語によって大量のページが検索結果として表示され，求めている情報を探しにくいなどの問題がある．このような問題を解決するために，検索結果をクラスタリングして提供する方法が提案されている [2]．検索結果をクラスタリングすると，すべての検索結果をまとめてクラスタ単位で表示するため，検索結果全体の俯瞰が可能になる．またユーザ自身も思いつかないクラスタが生成されることもあるため，検索語とそのクラスタの意外な関連性が発見できるなどの利点がある．

既存のクラスタリングシステムの多くは，メタサーチを行って複数の検索エンジンから200件程度のウェブページを収集し，それらを分類するに留まっている．そのため，クラスタリングシステムを用いることで，検索結果の下位に埋もれている有用なページを得ることは難しい．このような問題を解決し，検索結果の下位にあるページをユーザへ提示するシステムとして馬場らのシステムがある [2]．馬場らのシステムは，検索エンジン基盤 TSUBAKI[1] を用いることで，数千件のウェブページをクラスタリングすることが可能であり，検索結果の下位に埋もれている有用なページをユーザへ提示することを可能にしている．しかし，数千件のページをクラスタリング対象とする場合，リンク集や画像だけからなるページといったテキスト情報が少ないページやスパムページなど，既存の検索エンジンの結果では順位が低くユーザの目につかないページもクラスタリングの対象となるため，最終的に生成されるクラスタの質の低下を招くという問題がある．このような問題の解決には，事前にページのタイプを識別しておくことが重要となる．

本研究では，クラスタリングだけでなく，今後さまざまな研究で使えるページタイプを提案する．以下，2節では関連研究について紹介し，3節では我々が提案するページタイプについて述べる．

2 関連研究

金子ら [1] は，ページタイプに関する既存の研究では，

- 得られたページタイプをどうやって活用するか
- 定義されたページタイプの妥当性およびその必然性

について言及されていない点を指摘した．そして，それらを解決するため，まずインターネットの利用目的を調査し，調査結果から得られた利用目的と関連があるキーワード集合を作成した．そのキーワード集合をGoogleで検索し，各検索結果の上位100ページを収集し，人手でページタイプの分類を行った．その結果，アプリケーション，ショッピングサイト，書籍の紹介，掲示板・日記・チャット，シラバス，補集合の6タイプが得られ，これらをページタイプとして提案した．しかし，特定のキーワードで検索した結果を基に，ページタイプを定義すると，ページタイプの範囲が検索結果ページのみに限定されてしまう恐れがある．また，金子らは，ユーザの立場からのページタイプしか定義しておらず，発信者の立場は反映されていない．

3 ページタイプの提案

3.1 ページタイプの調査

ウェブページのタイプとして，どのようなものがあるかを調査した．まず，検索エンジン基盤TSUBAKIを使い，「捕鯨問題」や「アンチエイジング」などの20クエリについて1,000件ずつ検索結果を取得した．そして得られた検索結果の中からランダムに20件ずつ

*Page Type Detection for Web Page Clustering.
†Naun Kang, Keiji Shinzato, Sadao Kurohashi. Graduate School of Informatics, Kyoto University.
[1]http://tsubaki.ixnlp.nii.ac.jp/index.cgi

表 1: ページタイプの分類

分類基準	個人		団体						
	主観的	客観的	公共			ビジネス			
			国家機関		公共機関	検索エンジン，ポータルサイト	ショッピングサイト	企業のページ	
			教育機関	教育機関以外				マスコミ	マスコミ以外
共通ページタイプ		○	○	○	○	○	○	○	○
特化ページタイプ	ブログ		シラバス			辞書	商品販売ページ	新聞記事	

ブログ, 新聞記事, 検索エンジンの結果, 商品販売ページ, SEO 対策用ページ, Q&A のページ, 辞書, 住所録, プロフィール, アクセス, B B S, ゲーム, シラバス, サイトマップ, リンク集, ほとんど画像だけのページ, その他

図 1: ウェブページを調査することで得られたタイプ

選びだし，計 400 件のページセットを準備した．このページセットに全てのページタイプが網羅されているとは限らないため，上記の 400 ページに加え，以下の 54 サイトについて，そのサイト中のページも調査対象とした．

- goo や Yahoo!などのポータルサイト 4 サイト
- 各ポータルサイトのトップページからリンクが張られている 20 サイト
- 総務省や厚生労働省，グリンピースなどの国家機関や NGO などのサイト 10 サイト
- ブログや Wikipedia などの辞書サイト 20 サイト

以上のページを調査した結果，図 1 に示す 17 種類のページタイプが得られた．

3.2 ページタイプの分類

3.1 節で得られたページタイプを，ページを発信する側に着目して分類した．その結果を表 1 に示す．まず，得られたページタイプを，発信者が個人であるか団体であるかという観点から分類した．ついで，「個人」に分類されたページタイプを「主観」,「客観」という観点から分類した.「主観」に分類されるページとしては個人のブログ,「客観」に分類されるページとしては個人のプロフィールが挙げられる．

一方，団体に分類されるページタイプであるが，サイトが提供しているサービスの種類や発信者という観点から公共の利益のために活動している団体が発信しているページのタイプを「公共」に，ビジネスを目的で作られたページのタイプを「ビジネス」にそれぞれ分類した．さらに「公共」は,「国家機関」,「公共機関」に,「ビジネス」は「検索エンジン，ポータルサイト」,「ショッピングサイト」,「企業のページ」に細分類される．

図 1 に挙げたページタイプのうち，表 1 に示した複数のカテゴリに含まれるページタイプ，および，ある特定のカテゴリだけにしか含まれないページタイプがあることがわかった．前者のページタイプを共通ページタイプ，後者を特化ページタイプと呼ぶ．共通ページタイプには，BBS，サイトマップ，Q&A，プロフィール，アクセスが含まれる．各カテゴリに特化したページタイプを表 1 に示す．

3.3 ページタイプの自動識別

図 1 に示したページタイプのうち，リンク集，商品販売ページ，SEO 対策ページ，ほとんど画像だけのページについては，ウェブページの DOM 構造，アンカーテキスト情報，ページ中のテキスト情報を手がかりに自動識別する手法を開発した．残りのページタイプについては，今後の課題とする．

4 おわりに

本稿では，17 種類のページタイプを提案し，それらをページ発信者の立場から分類した．今後の課題としては，提案したページタイプを自動的に識別する方法の開発，クラスタリングシステムにおけるページタイプ識別の効果を調査する予定である．

参考文献

[1] 金子 大輔 高山 毅 池田 哲夫 長内 亘. Web 文書のページタイプを用いた適応的分類と試作システムの評価. 知能情報ファジー学会論文誌, *Vol.18, No.2, pp.319-336*, 2006.

[2] 馬場康夫 新里圭司 黒橋禎夫. 検索エンジン基盤 TSUBAKI を用いた大規模ウェブ情報クラスタリングシステムの構築. 情報処理学会 研究報告 *2008-NL-183*, 2008.