2008年9月15日

Ｗｅｂ検索結果のページ選択を支援するジャンル分類システム

発表者

相原功昌

1　はじめに

インターネットで検索する
→検索語を含むＷｅｂページを一覧にして返すのが一般的

一覧から必要なＷｅｂページを見つけるのは時間が掛かる
→検索結果に情報を付加して、Ｗｅｂページ選択の補助をするサービスが研究されている

■サービス例

- Yahoo!検索
  一部に検索語を含む、他の検索語を表示する
- Clusty
  検索結果のページの内容でクラスタ化し、階層表示する
- Mooter
  検索結果のページに含まれる検索語以外の単語を、新たな検索語の候補として表示する

■研究の動機

Webページ上の情報の中で

- 「何について」書かれたページか
- 「どのような目的で」書かれたページか

検索者の求めるものが違うことがある
→前者：検索語の変更によって、結果を制御できる場合が多い
→後者：同じ方法で結果を制御することは難しい

■後者の例

「富士山」「登山」というキーワードで検索した場合

- 「富士登山に関する日記や感想」
- 「富士山に登山するための情報」

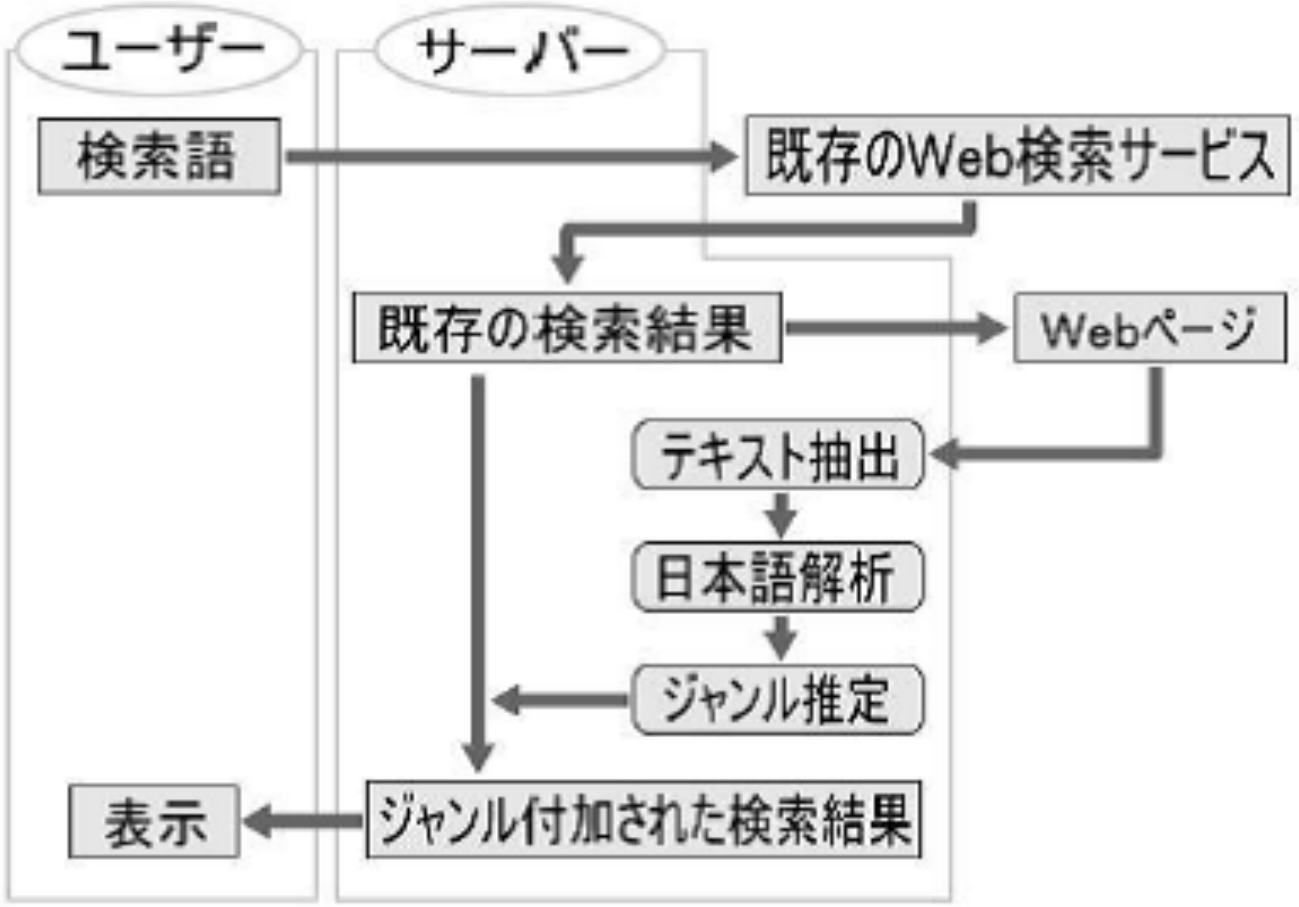
のように、ページを読む目的が違うことがある

■本研究では
検索結果に、「どのような目的で」書かれたという情報を付加
→検索結果の選別を支援

具体的に
「どのような目的で」に関するジャンルリストを用意
どのジャンルに分類されたかを、付加情報とする

2　提案システム

提案するWebサービスの流れ



■検索結果のイメージ

「どのような目的で」に関するジャンルの一覧が左に、
ジャンルが付加された検索結果の一覧が右に表示される

3　ジャンル推定法とジャンル体系

3.1　ジャンル推定

付加する情報を選択するジャンル推定
→ナイーブベイズ推定

$$\arg\max V_{j \in V} P(V_j) \prod_{i=1}^{n} P(a_i|V_j) \qquad (1)$$

■ナイーブベイズ推定

$$\arg\max(V_j \in \mathbf{V}) P(V_j) \prod_{i=1}^{n} P(a_i|V_j) \quad (2)$$

ここで、クラスの集合$\mathbf{V}$

$$\mathbf{V} = (V_1, V_2, V_3, \cdots, V_j)$$

各事例a_jはN個の特徴量からなる特徴ベクトル

$$a_i = (a_{i1}, a_{i2}, a_{i3}, \cdots, a_{iN})$$

■文書分類において

- 一般的
  →名詞や動詞
- 我々のジャンル推定
  →名詞や動詞の他に、機能語も重要

■推定の正解率調査

手動でジャンル分類したWebページ：2200件

学習データ：2000件

評価用データ：200件

「日記/日記でない」のジャンル推定に、どのような品詞を用いればよいかを比較

→表1

表1　対象品詞ごとの正解率の比較

品詞	推定の正解率
一般名詞、サ変動詞のみ	85%
記号、助詞、助動詞以外の品詞	90%
記号以外の全品詞	95%

→機能語も重要

本研究では、記号以外の全品詞を用いて推定

3.2　ジャンル体系

- どのようなジャンル体系を設定するかが重要
- ページ選択の手がかりにならない情報は邪魔

3.2.1 ジャンル体系の設定

「どのような目的で」に関する、２階層のジャンル体系を設定

- 親ジャンル：大まかなジャンル
- 子ジャンル：各親ジャンルの、より細かな分類

3.2.2 設定したジャンル体系

- 500のWebサイトについて、どのような目的で書かれているかを検討
  →親ジャンルを設定(表2)
- 親ジャンルについて、2000のWebサイトを検討
  →子ジャンルを設定(表3)

表2　設定した親ジャンル

ジャンル名	説明
紹介	企業、商品などの宣伝、紹介
案内	イベントや事業の案内、活動
知識	物や常識、習慣についての知識
記録	出来事や結果などの記録
私見	個人の感想や意見など
娯楽	Ｗｅｂ上で楽しむページ
形式	サイト運営上のページ
その他	外国語やエラーページ

表3　設定した子ジャンルの一部

ジャンル	子ジャンル一覧
紹介	「見出し、トップページ」「クチコミ、まとめ情報」 「おすすめ紹介、特集」「商品販売、宣伝」「ウェブサービス」
案内	「イベント案内、メンバー募集」「理想、理念」 「活動、活動説明」「生活情報、アドバイス」
知識	「知識、常識」「歴史、年表」「使い方、作り方」
娯楽	「ゲーム、占い、テスト」「創造物、イラスト展示」 「交流サイト、掲示板」「コンテンツ配信」「データ、ソフト配布」

4　評価

4.1　ジャンル推定の精度評価

- 手動で分類したWebページ：約2500件
- 評価用データ：無作為に選んだ473件
- 学習データ：残りの約2000件

→推定結果の評価（表4）
推定されたジャンルが対象Webページの一部を表している場合を正解とした評価
→緩めの評価結果（表5）

表4　ジャンル推定の評価結果

正解	245
不正解	228
正解率	52%

表5　緩めの評価結果

正解	344
不正解	129
正解率	73%

4.2　ジャンル体系の評価

ジャンル体系が適切かどうかを、実験で評価

・任意に選んだWebページ：50件

・5人の人が、我々の提案したジャンルへ分類

→実験結果（表6）

表6　ジャンル体系の実験結果

過半数の人が同じジャンルへ	ページ数
分類したページ	39
分類しなかったページ	11
過半数の得られた割合	78%

- 区別しにくいジャンル
- 分類できないWebページが多く存在

→人によって異なるジャンルへ分類してしまう

4.3　作成したWebサービスの評価

評価実験：キーワードを与えるだけのWeb検索サービスと比較

→目的に必要なWebページ発見までに、どの程度の時間短縮になったか

4.3.1 評価方法

- 検索の目的を仮定
- Yahoo!検索と我々のWebサービスで検索
- 必要なWebページを10件集めるまでに
  1：検索結果のページタイトル等を読んだ数
  2：開いて内容を確認したWebページの数

を比較

4.3.2 評価結果

- 検索目的を10件用意
- 著者の１人が実験

→検索目的別の評価結果の一部（表7〜表9）

表7　「富士登山の感想を読む」での評価結果

	Yahoo!検索	本研究
検索語	富士山、登山、感想	富士山、登山
ジャンル	ー	日記、感想
読んだページタイトル等の数	40	12
開いたページ数	17	12

表8　「とんかつの作り方を調べる」での評価結果

	Yahoo!検索	本研究
検索語	**とんかつ、レシピ**	**とんかつ、レシピ**
ジャンル	**ー**	**作り方、使い方**
読んだページタイトル等の数	13	10
開いたページ数	12	10

表9　「PS2のゲームを比べる」での評価結果

	Yahoo!検索	本研究
検索語	**ゲーム、PS2**	**ゲーム、PS2**
ジャンル	**ー**	**クチコミ、まとめ情報**
読んだページタイトル等の数	69	13
開いたページ数	15	13

表10　評価結果の平均

	Yahoo!検索	本研究
読んだページタイトル等の数	30	13
開いたページ数	14	12

読んだ項目数と開いたページ数の平均
→（表10）
ページタイトル等を読んだ数が、大きく異なる

時間の比較

- ページタイトル等を読んで、ページを開くのかを決定する時間：１秒
- ページを開き、必要かどうかを決定する時間：５秒

と仮定して、時間の平均を比較
→比較結果（表11）

表11　時間の比較

	Yahoo!検索	本研究
平均時間（秒）	100	73

→27%の時間を短縮

5　今後の課題

1. **ジャンル体系**
   設定したジャンル体系で、全てを網羅してるとは言えない
2. **ジャンル推定**
   推定の精度が高いとは言えない
3. **テキスト以外の情報**
   画像やFlash、スクリプトなどで書かれたページは対象にできない

6　終わりに

「どのような目的で」書かれたページかという情報を付加
→検索結果の選別にかかる時間を短縮
→ページ選択を支援

他の情報やテキスト以外のコンテンツも利用し、より便利な検索サービスの作成を目指す