

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

伊東 敏章、池田 尚志
岐阜大学 工学部

1 はじめに

近年、インターネットの普及により多数の企業サイトや個人サイトが作られ、それに併せ、様々な検索サービスが提供されている。現在のところ、検索語を含む Web ページの一覧を返すのが一般的である。しかし、検索者がこの一覧から必要な Web ページを選ぶには、一つずつ、タイトルや説明文を読んで決めていくしかない。この選択には少なからぬ時間が掛かるのが現状である。

この問題を解決するため、検索結果に何らかの情報を付加して、必要な Web ページを選ぶ補助とする様々なサービスも研究されている。

- Yahoo!検索 [1]
一部に検索語を含む他の検索語を表示する。
- Clusty[2]
検索結果のページの内容でクラス化し、階層表示する。
- Mooter[3]
検索結果のページに多く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。

Web ページ上の情報の中で、「何について」書かれているページなのか、ということが最も重要だと考えられるが、これについては、検索者自身が検索語を変更することで、結果を制御できる場合が多いと考えてよい。

しかし、たとえば「富士山」「登山」というキーワードで「富士山の登山について」のページを検索したとき、『富士登山に関する日記や感想』について読みたい場合や『富士山に登山するための情報』を読みたい場合、検索語の変更で、このような検索意図を実現するのは難しい。(図 1)

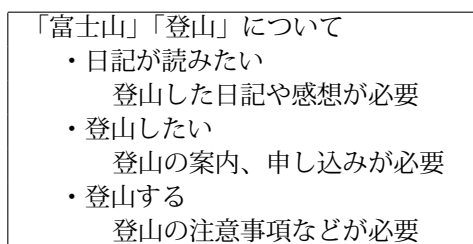


図 1：何について／どのような目的で

そこで本研究では、検索語の変更での制御が難しいと考えられる、「どのような目的で」書かれているのか、という情報を付加することで、検索結果の選別を支援する方法について検討した。

具体的には、検索された Web ページを、あらかじめ用意しておいた「どのような目的で」に関するジャンルリストの中のいづれかに分類し、その情報を付加する方法をとることとした。

以下では、このジャンルの体系についての提案と、ジャンル推定の方法、および、これらを用いて試作した Web サービスについて述べる。

2 提案システム

本研究で提案する Web サービスでは、検索者は Web ブラウザ上から検索語を入力し、我々の試作したサーバーへ送信する。これを受け取ったサーバーは、既存の Web 検索サービスで検索を行い（本研究では Yahoo!検索 Web サービス [4] を利用）、その結果の一覧から、個々の Web ページをダウンロードし、文書解析、ジャンル推定して、結果を従来の検索結果に付加する形で、検索者のブラウザへ返信する。(図 2)

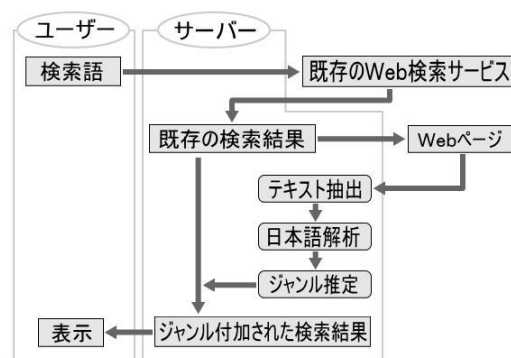


図 2：処理の流れ

サーバーから結果を受けると、検索者のブラウザ上に、図 3 のように「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。



図3：検索直後の画面

次に、ジャンルの一覧から表示したいものを選択すると、そのジャンルに該当すると推定されたページだけが右側に表示される。(図4)
 検索者はその中から目的のページを探す。



図4：「商品販売、宣伝」というジャンルを選択した結果

3 ジャンル推定法とジャンル体系

3.1 ジャンル推定

本研究では、付加する情報を選択するジャンル推定に、ナイーブベイズ推定を使用している。

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

一般的に、文書分類では文書中に登場した名詞や動詞などが使用され、助詞や助動詞などの機能語は使用されないことが多いが、我々のジャンル推定では、機能語も重要であると考えた。

- ・～たい : 食べたい、遊びたい
- ・～か : いこうか、ですか
- ・～とか～ : 映画とか音楽とか

表1は、「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した結果である。我々の手でジャンル分類したWebページ2200件を用意し、2000件を学習データ、200件を評価用データとして、推定の正解率を調査した。

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	96 %

この結果から、少なくとも「日記／日記でない」を判断するには、機能語も重要であると考えられる。よって、本研究では、記号以外の全品詞を用いて、推定することとした。

3.2 ジャンル体系

我々の方法では、どのようなジャンル体系を設定するかということが大変重要である。

ジャンルによる絞込みが閲覧ページ選択の手がかりにならなければ、役に立たないだけでなく、邪魔な情報として、他の情報（タイトル、説明文など）を読み難くすることになる。

以下に、我々の提案するジャンル体系について述べる。

3.2.1 ジャンル体系の設計

ユーザーが何かを買うために検索を利用した場合に、クチコミ情報の載ったサイトを読みたい場合もある。また、メーカーの商品ページを読みたい場合もある。また、商品の説明が載ってさえいれば、どのようなページでもいい場合もあり得る。

このことを踏まえ、本研究で設定するジャンル体系は、「どのような目的で」に関する、大まかなジャンルを親ジャンルとして設定し、そのそれぞれに、いくつかの細かな分類を子ジャンルとして設定する、2階層のジャンル体系を設定した。

3.2.2 設定したジャンル体系

どのような検索目的が考えられるか、実際に500のWebサイトについて、それぞれどのような目的で書かれているのかを検討し、以下の8ジャンルを設定した。

表2：設定したジャンル

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

また、上記8ジャンルについて、2000のWebサイトを調べ、調節し、以下の子ジャンルを設定した。

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」 「商品販売、宣伝」 「おすすめ紹介、特集」 「クチコミ、まとめ情報」 「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	「データ、ソフト配布」 「創作物、イラスト展示」 「ゲーム、占い、テスト」 「コンテンツ配信」 「交流サイト、掲示板」

4 評価

4.1 ジャンル推定の精度評価

設定したジャンル体系に対するジャンル推定の評価を行った。

我々の手でジャンル分類したWebページ、約2500件を用意し、無作為に選んだページ473件を評価用データ、それ以外の約2000件を学習データとして、推定結果の評価を行った。(表4)

表4：ジャンル推定の評価結果

正解	245
不正解	228
正解率	52 %

また、完全な正解ではないが、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にも正解とした評価も行った。(表5)

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

本方法では、前述のとおり、どのようなジャンル体系を設定したかが重要となる。ここでは、設定し

たジャンル体系があいまいさの少ない適切な体系であるといえるか、評価を行った。

表6は、任意に選んだ50件のWebページが、我々が提案したジャンル体系の中のいずれかに、誰もが同じようにジャンル分類できるかどうか、5人の人に判断してもらった結果である。

表6：ジャンル体系の実験結果

	ページ数
過半数の人が同じジャンルへ分類したページ	39
半数未満しか同じジャンルへ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や、どのジャンルにも分類できないWebページが多く存在すると、人によって別々のジャンルに分類してしまうWebページが多くなる。

表6の実験結果を見ると、78%のWebページについて、過半数の人が同じジャンルへと分類していることから、我々の設定したジャンル体系は、あいまいさの少ない適切な体系であると考えられる。

4.3 作成したWebサービスの評価

試作したWebサービスの評価として、キーワードを与えるだけの既存のWeb検索サービスと比べて、目的に必要なWebページを探すのに、どの程度の時間の短縮になったのか、という評価実験を行った。

4.3.1 評価方法

検索の目的を仮定して、Yahoo!検索と、我々のWebサービスとで、設定した目的に必要なと考えられるWebページを10件集めるまでの、以下の2点を比較した。

1. 検索結果のページタイトル等(ページタイトルやその下のテキスト)を読んだ数
2. 開いて内容を確認したWebページ数

評価者は、Yahoo!検索と、我々のWebサービスに対して、それぞれ適切と考える検索語を入力して検索する。検索結果の中で、開くWebページは、ページタイトル等を読んで決定し、実際にページを閲覧して、目的に必要なかどうかを決める。

4.3.2 評価結果

10件の検索目的に対して、著者の一人が4.3.1で述べたような操作をして評価を行った。

以下に、評価結果の一部を示す。

表7：検索目的を「富士山の登山の感想を読む」とした場合の評価結果

	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	—	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：検索目的を「とんかつの作り方を調べる」とした場合の評価結果

	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	—	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：検索目的を「PS2のゲームを比べる」とした場合の評価結果

	Yahoo!検索	本研究
検索語	ゲーム PS2	ゲーム PS2
ジャンル	—	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

以上の実験の結果を集計し、平均の読んだ項目数と開いたページ数を得た。(表10)

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

実際に開いて内容を確認したページ数に大きな変化はないが、ページタイトル等を読んだ数は大きく異なる。

ページタイトル等を読んで、ページを開くのか決定するのに1秒、実際にページを開き、必要かどうか決定するのに5秒かかると仮定して、時間の平均を比較した。(表11)

表11：時間の比較

	Yahoo!検索	本研究
平均時間 (秒)	100	73

結果、27%の時間を短縮できた。

5 今後の課題

今後の課題としては、以下のような点があげられる。

1. このようなジャンル分類では支援しがたい場合

今回の研究では、多くの検索目的を考慮してジャンル体系を作成しているが、それでも全てを網羅しているとは言い切れず、作成したWebサービスでは閲覧ページ選択を支援できない場合も存在する。例えば、「富士山が見えるペンションを探す」場合には、「どのような目的で書かれているか」という観点からは、閲覧ページを絞り込むのは難しい。

2. ジャンル推定

本研究でのジャンル推定の精度は、あまり高いとはいえず、また、テキストが短いページの推定はあまりできない。特に、読みたいページは決まっているが、URLを知らず、検索して探す場合には、候補の絞込みができる反面、推定間違いが起こった場合に見つけることができない。

3. テキスト以外の情報

Webページには、そのページのコンテンツのほぼ全部が、画像やFlash、スクリプトなどで書かれている場合も、少なからず存在する。今回の方法では、そういったページは対象にすることができない。

6 終わりに

本研究では、既存の検索結果に「どのような目的で」書かれたページか、という情報を付加することで、検索結果の選別にかかる時間を短縮できるWebサービスを作成した。しかし、この情報だけで、十分に結果を選別できるわけではない。今後は、さらにより多くの検索目的に対応するために、選択の手がかりに利用できそうな他の情報や、ページ中のテキスト以外のコンテンツも利用して、より便利な検索サービスの作成を目指したい。

7 参考文献

- [1] Yahoo!検索
[Http://www.yahoo.co.jp/](http://www.yahoo.co.jp/)
- [2] Clusty
<http://clusty.jp/>
- [3] Mooter
<http://www.mooter.co.jp/>
- [4] Yahoo!検索 Web サービス
<http://developer.yahoo.co.jp/search/>