ユーザーの選択した観点からの自動分類を利用したWeb検索の支援

伊東 敏章、松本 忠博、池田 尚志
岐阜大学 工学部

1　はじめに

近年、インターネットやブログなどの普及により、インターネット上に情報を発信する人や、またその機会が増えている。それに伴い、人々が知識や情報を得るために、インターネットの Web 検索サービスを利用する機会も増えている。

しかし、既存の検索サービスの多くは、検索語を含む Web ページの一覧を提示するだけなので、目的の知識や情報を得るのに苦労したり、得ることが出来ない場合もしばしば存在する。

この問題の解決のために、最近では検索結果を何らかの方法で分類し提示する検索サービスが開発されている。

- mooter [1]

  検索結果をページの内容でクラスタ化し、一覧表示、もしくはヴィジュアル表示する。

- clusty [2]

  検索結果をページの内容でクラスタ化し、階層表示する。

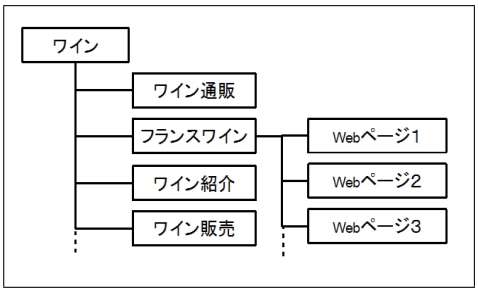

図１：既存のクラスタリング結果の例

しかし、これらの方法は、検索結果をシステムが定めた単一の基準で分類するだけなので、必ずしもユーザーの考える分類と一致するとは限らないという問題がある。

そこで、本研究では検索結果の様々な分け方をユーザーに掲示し選択させ、その分け方により分類することで、この問題を解決できないか検討し、システムの試作を行った。

2　提案システム

我々が試作したシステムでは、ユーザーがキーワードを用いた通常の検索を行うと、システムはユーザーに、自動作成した分類観点の候補を提示する。ユーザーはこの中から分類したい観点を選択し、システムはその観点に重点を置いて検索結果を分類し、サマリーなどと共に提示する。

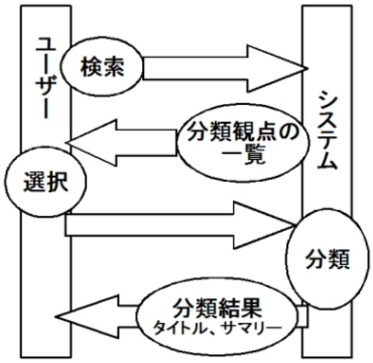

図２：システムの概要

本研究では、上記のシステムを実現するために以下の内容に取り組んだ。

- 分類のための観点の自動取得
- 観点に焦点をあわせた自動分類
- クラスタのタイトルやサマリーなどの作成

3　分類のための観点の自動取得

Web ページを分類する観点には様々なものが考えられる。

・ ページ本文に関連した分類

ページの内容、例えば、見出し語や言葉遣いなど、テキスト中に現れる情報で分類

・ ページ本文以外の情報での分類

ファイルサイズやドメイン、デザインなど、テキスト以外のページ情報で分類

・ ユーザーに依存した分類

ユーザーの好き嫌いや、既読未読など、ユーザーに依存した情報での分類

本システムは知識や情報の取得の支援を目的としているので、ページ本文以外の情報での分類や、言葉遣いでの分類など、書かれている情報に影響の少ない分類はあまり有効ではない、と考えられる。また、ユーザーに依存した分類は、検索サービスであることを考慮すると実現が難しい。よって本稿では、ページの本文による分類を考える。また、検索結果の Web ページは検索語に関連した内容が書かれている場合が多い。そこで、我々は検索キーワードの属性を表す語（以下属性語とする）に着目し、ユーザーに提示する分類観点の候補として使用することとした。

表１：観点の候補の例

「ワイン」で検索
・「産地」で分類 ・「価格」で分類 ・「タイプ」で分類

これを実現するため、以下の手法を利用して検索キーワードから属性語の自動取得を行った。

3.1 Web ページ中の強調表現を利用した属性語の取得

Web ページでは、検索キーワードの属性語が、「○評価」や「＜ｂ＞産地＜/b＞」など、強調表現を伴って表現されている場合が多い [3]。そのことを利用して、検索語の属性語を取得し、分類観点の候補を作成した。

対象とするページは、検索キーワードに関する知識について書かれているページが適している。タイトルに「… の詳細」「… の概要」など、「(検索キーワード) の (手がかり表現)」を含む Web ページを検索することで取得する。

表２：使用した手がかり表現

カタログ	ガイド	リスト	紹介
感想	名鑑	図鑑	詳細
概要			

取得した Web ページから、以下の文字や Html タグによって強調されている単語を収集する。

表３：利用した強調文字

●	○	■	□	・	★
◆	◇	☆	◎	＊	…
／	＼	｜	＝	-	
()	【】	〔〕	『』	＜＞	≪≫
「」	[]				

表４：利用した強調タグ

dt	li	td	th	h1	h2
h3	h4	h5	h6	em	strong
tt	i	span	b	big	

実際に、検索語「ワイン」で取得実験を行った。その結果を以下に示す。

表５：「ワイン」で作成した分類観点の候補

ワインの評価	ワイン一覧	ワイングラス
ワインニュース	ワインの詳細	ワインギフト
ワイン生産者	ワイン産地	ワイン雑誌

4 自動分類

本研究では、文書集合に対して、選択された分類観点に基づいた分類を行うこととしている。しかし、一般的なクラスタリング手法では、分類観点を選択することは出来ない。そこで、我々は Fuzzy c-Means クラスタリングの手法を、分類観点を導入できるように拡張して用いることとした。

4.1 Fuzzy c-Means クラスタリング

一般的な Fuzzy c-Means クラスタリングの手法は以下の方法で行われる。

1. c 個のクラスタの中心を任意に設定する。
2. それぞれの文書の各クラスタへの帰属度を計算する。
3. 計算された帰属度を使用して、クラスタ中心を再計算する。

4. (2)(3) を、クラスタの中心がほぼ収束するまで繰り返す。

Fuzzy c-Means クラスタリングでは、結果として分類文書の各クラスタへの帰属度が得られる。Web ページは同じページ中でも内容が多岐にわたる場合も多く、この手法が適していると考える。

表６：クラスタリング結果

ワイン	品種	最大
酸味	ワインバー	フランス
ワイングラス	ヴォーヌ	チーズ
送料無料	印象	カフェ

(各語は各クラスタの重心に近い語)

しかし、この手法で得られる結果には、分類の“ 観点 ”が導入されていない。

4.2 “ 観点“ を導入した Fuzzy c-Means クラスタリング

この手法を、選択された分類観点により分類するように拡張する必要がある。

具体的には、観点に関連性の高い語を予め取得し、クラスタリングの際に、高く重みが付く様に拡張した。

4.2.1 関連語の取得

観点に関連性の高い語を取得するために、検索結果の Web ページを利用して、本文中の単語から、選択された観点を表す語と、同じ文内での共起頻度の高い語を取得して、一覧を作成した。

以下に、実際に作成した「ワイン」の「評価」と関連性の高い語の一部を示す（表７）

表７：ワインの「評価」と関連性の高い語

ワイナリー	獲得	パーカー
満足	非常	フランス
ロバート	世界的	世界

4.2.2 拡張したクラスタリング

Fuzzy c-Means クラスタリングでの各文書のクラスタへの帰属度は、文書を単語ベクトルに見立てて、類似度を計算することで与えられる。

その際に、単語ベクトル中の、観点に関連語の値を大きくすることで、他の単語の影響が少なくなり、観点に関連性の高い語に高い重みが付くよう拡張できる。

5 クラスタのタイトルやサマリーなどの作成

検索の支援を行うためには、分類結果をわかりやすくユーザーに提示することも必要である。

ユーザーの得たい情報は、選択された観点に関する情報である。そこで、ページ中で重要な部分を抜き出した一般的な要約情報よりも、選択された観点に関連する情報を提示する方が、本システムには有効だと考えられる。

このような考え方に則って、分類結果のクラスタに対して、以下を作成した。

- 分類結果クラスタのタイトル
- 分類結果クラスタのサマリー
- 個々の Web ページのサマリー

5.1 分類結果クラスタのタイトル

分類結果の各クラスタにタイトルを付与することによって、一目でその内容を大まかに把握できる。ここでは、クラスタのタイトルとして、観点と関連性が高く、かつ、クラスタ中で重要度の高い単語１つを提示することとした。具体的には、クラスタリング時に作成した観点の関連語から、クラスタの中心に最も近い単語を提示している。

以下に、「ワイン」で検索し、分類観点に「ワインの評価」を選択した際の例を示す（表８）

表８：「ワインの評価」でのタイトル一覧

非常に	パーカー	ワイナリー
シャトー	世界	フランス
獲得	日本	イタリア

5.2 分類結果クラスタのサマリー作成

また、分類結果のクラスタごとに、その特徴を現すサマリーとして、各 Web ページの全ての文の中から、クラスタの中心に近い文を選択して提示する。

例：

- ワイン＞ワインの評価＞非常に
  非常に複雑でボディのしっかりした素晴らしい赤ワイン。全てが素晴らしい非常に完成度の高いワイン。フランス国内でも非常に人気が高い為、日本への輸入量はごく僅か。
- ワイン＞ワインの評価＞パーカー
  パーカーはフランスのボルドーへ降り立った。ロバート・パーカーさんも最高評価の５ッ星。ロバート・パーカーが 90 点をつけたワイン。
- ワイン＞ワインの評価＞ワイナリー
  新酒の試飲や見学が楽しめる注目のワイナリー。プレミアムワインを生産するアメリカ屈指のワイナリーです。品質を保つことが出来るのも、ひとつの特徴です

5.3 Ｗｅｂページのサマリー作成

個々のＷｅｂページにも、観点に着目したごとにサマリーを表示させる。

既存の検索サービスのサマリーは、その Web ページ中で重要と考えられる文を抜きだして提示している。しかし我々のシステムでは、ユーザーが知りたい情報は選択された観点に関する情報であると考えられる。そこで、クラスタリングの際に使用した関連語を使用して、TF による文書要約の手法において、関連語の多い文にも高いスコアを設定するよう拡張した手法を使用している。

例：

- ワイン＞ワインの評価＞非常に
  - フランス国内でも非常に人気が高い為、日本への輸入量はごく僅か
  - …これは、非常にお買い得な逸品なのでは？
  - このクリマが、法的にも非常に特殊な位置付けをされていることが良く判ります
- ワイン＞ワインの評価＞パーカー
  - ロバート・パーカーさんも最高評価の５ッ星
  - リリース時のパーカーの評価は 99 点でした
  - ロバート・パーカー氏による評価は 90-95 点が「傑出」、80-89 点は点数別に「並み以上」「優良」「極めて良い」などとランク付けされている

6 今後の課題

- 処理速度
  現在のシステムでは、これら一連の処理を行うために、少なからぬ時間がかかり、実用的とは言いがたい。
- 少数意見
  検索結果には、他の Web ページにはないユニークな情報が書かれている Web ページなども存在する。現在のクラスタリング手法では、それらのページが他の多くの Web ページに紛れてしまい、有効活用するのが難しい。
- 他の分類方法
  ページの内容に関する分類でも、見出し語やその属性での分類以外にも、様々な分類方法が存在する。本システムが提示する分類観点の候補は、分類方法の一部でしかない。今後は他の分類方法にも注目したい。

7 終わりに

本研究では、Web 検索結果の分類方法を複数提示することで、インターネット上からの検索を利用した知識や情報の取得を支援するシステムを作成した。しかし、このシステムを利用しても、やはり見つけるのが難しい情報は多く存在する。今後は、より多くの情報を取得しやすくなるようなシステムの作成を目指したい。

8 参考文献

[1] mooter
URL http://www.mooter.co.jp/
[2] clusty
URL http://clusty.jp/
[3] Web からの具体物の属性・属性値情報の自動獲得, 吉永直樹 (JSPS), 鳥澤健太郎 (JAIST), 言語処理学会第 13 回年次大会論文集,pp.887-890,2007