ポータルサイト自動作成の試み

白井清昭　菅井俊介　平野健児　星正人

北陸先端科学技術大学院大学 情報科学研究科

{kshirai,s-sugai,hiraken,m-hoshi}@jaist.ac.jp

1 はじめに

ポータルサイトとは，ここでは，ユーザがあるテーマに関する情報を調べるとき，最初に訪れる入口となるべきサイトを指す．例えば,「言語情報処理ポータル[1]」は，言語処理に関する会議案内，製品ニュース，用語集，関連機関へのリンク集などを集約したポータルサイトである．ウェブから何か情報を得たいとき，例えば言語処理に関する情報を得たいときには，このようなポータルサイトがあれば便利である．我々は，ウェブでの情報探索を支援することを目的に，あるテーマが与えられたとき，そのテーマに関する情報をウェブから収集し，ポータルサイトを自動的に構築することを目指している．

一般に，ポータルサイトのコンテンツとして，関連リンク集，用語集，FAQ 集，掲示版などが挙げられる．これらのうち，本論文は関連リンク集と用語集を取り扱う．本論文の目的は以下の 2 つである．1 つ目は，ポータルサイトのテーマをキーワード 1 語で与えたとき，そのテーマに関連したリンク集を自動的に作成することである．特に，単にページを収集するだけではなく，リンク集に記載する個々のウェブページの説明をいかに記述するかを重視する．2 つ目は，同じくテーマを与えたとき，そのテーマに関連する用語とその用語説明を集めた用語集を作成することである．特に，用語が複数の意味を持つとき，ポータルサイトのテーマにあった説明を自動的に選別することに焦点を当てる．

2 リンク先ページに関する情報の獲得

2.1 概要

関連リンク集の自動生成は，大きく分けて次の 2 つの処理が必要である．

1. リンク集に掲載するウェブページの選別
2. 掲載ウェブページに関する情報の記述

本研究は 2 の処理に焦点を当てる．リンク集において，掲載ページに関する情報が果たす役割は重要である．ユーザは，リンク集があったとき，リンク先ページの説明を読み，自分が知りたい情報がそのリンク先ページに存在するかどうかを判断するからである．

[1] http://www.kc.t.u-tokyo.ac.jp/NLP_Portal/

本研究では，ウェブページに関する情報はウェブから自動的に抽出する．また，ウェブページに関する情報は，そのページそのものからではなく，そのページにリンクを貼っているページ (以下，参照ページと呼ぶ) から獲得する [7]．参照ページには，そのページ自身からは得られないような，第三者によるページの客観的な説明や，第三者の主観に基づく評価などが存在するからである．また，ウェブページに関する情報といっても，単なるページの説明から，そのページに対する意見や評価など，様々な種類のものがある．リンク集に掲載する際には，これらを無秩序に列挙するよりは，その種類に応じて自動的に分類し，整理して提示した方が望ましい．

ここでは，リンク集に掲載するウェブページが与えられたとき，その参照ページからウェブページに関する情報を取得し，さらにこれらをいくつかのカテゴリに自動的に分類してからユーザに提示する [1]．以下，2.2 項では参照ページからの情報の抽出について，2.3 項では得られた情報の自動分類について述べる．

2.2 ウェブページに関する情報の抽出

ここでは，リンク集に掲載するページを対象ページと呼ぶ．先ほど述べたように，対象ページに関する情報は，対象ページにリンクを貼っている参照ページから獲得する．まず，対象ページが与えられたとき，検索エンジン goo を用いてその参照ページの集合を収集する．次に，以下の 2 つの手続きを行う．

1. リンクを手がかりとした情報の抽出
   参照ページのリンクの周辺には，対象ページに関する情報が記述されていることが多い．ここでは，板橋らの手法 [7] により，HTML タグなどを手がかりとし，リンク周辺にある文または文章を対象ページに関する情報として抽出する．
2. サイト名を手がかりとした情報の抽出
   対象ページに関する記述の中には，そのページのサイト名が含まれることが多い．そこで，参照ページ内にある対象ページへのリンクを探し，そのアンカータグ内の文字列を対象ページのサイト名とみなす．次に，参照ページの中からサイト名を探し，そのサイト名を含む文または文章を対象ページに関する情報とみなして抽出する．

2.3 ウェブページに関する情報の自動分類

2.2 項で抽出した情報を以下の 5 つのカテゴリに自動的に分類した．なお，それぞれのカテゴリに属する情報の例は先に示す図 1 を参照されたい．

- 説明:記述
  次に示す「説明:機能」以外のウェブページに関する客観的な記述．
- 説明:機能
  ウェブページが提供している機能 (検索，予約など) に関する記述．
- 評価:情報量
  紹介している物件の数など，ウェブページに掲載されている情報の量に関する記述．
- 評価:利便
  ウェブページがユーザにとってどのように便利か，あるいは役に立つのかに関する記述．
- 評価:その他
  「評価:利便」「評価:情報量」以外で，ウェブページを主観的に評価している記述．

上記 5 つのカテゴリへの自動分類を行うために，それぞれのカテゴリに属する記述に頻出する定型表現パタンやキーワードのリストを人手で作成した．例えば,「評価:利便」に属する記述には「便利だ」「使いやすい」などのキーワードがよく現われる．また,「説明:機能」に属する記述に頻出する定型表現として「X ができる」「X は可能」などがある．ただし，X は「検索」「予約」「印刷」などウェブページの機能を表わす単語である．最終的に，ウェブから評価表現の抽出を行う先行研究 [3, 4] で提案された抽出パタンなどを参考に，18 個の定型表現パタンと約 100 語のキーワードリストを作成した．カテゴリの自動分類は，これらの定型表現パタンやキーワードとの照合により行った．

2.4 実行例

提案手法により,「MapFan Web」に関する情報の抽出・分類を行った．結果を図 1 に示す．最終的に，図 1 のような情報を提供するリンク集の作成を目指す．

MapFan Web (http://www.mapfan.com/)
説明:記述
- インクリメント P 株式会社が運営する、インターネットユーザーの為の地図を利用した情報配信サービス。
- インターネット地図検索サービス MapFanWeb

説明:機能
- 日本全国の地図 ⇒ 郵便番号・住所で検索可能
- ジャンルやスポット名、住所、郵便番号、駅名で検索し、グルメ、ショップ、宿泊など周辺のスポットをさがせる。

評価:情報量
- 日本全国の地図はもとより、約７５０都市の美しいタウンマップが無料で閲覧できます。

評価:利便
- 表示したい地図の住所を入力すると即座に地図が画面に出てくるので、ドライブに便利！
- 地図をメールで送れてべんり。

評価:その他
- 見易い地図が魅力的!!

図 1: リンク集掲載サイトの例

3 用語集の自動作成

本節では用語集の自動作成について述べる．まず，ポータルサイトのテーマを与えたとき，そのテーマに関連した用語の集合を獲得する (3.1 項) [2]．次に，各用語の説明をウェブから獲得する (3.2 項) [6]．

3.1 関連用語の獲得

ポータルサイトのテーマと関連する用語の集合をウェブから自動的に獲得する．なお，ここでの手法は佐藤らの手法 [5] に多少変更を加えたものである．まず，ポータルサイトのテーマをクエリとして検索エンジンに与え，テーマと関連のあるウェブ文書の集合 D を獲得する．次に，複合名詞または「A の B」という名詞句のうち，HTML タグ，記号，句読点で囲まれているもののみを用語候補として抽出する．これは，不必要な用語候補の抽出を妨げるためであり，用語集に加えるべき用語は見出しのように他の語句とは独立に現われやすいという観察に基づいている．さらに，用語候補に対してスコア付けを行い，上位 100 語の候補を得る．スコア付けは以下の 2 通りの方法で行う．

1. 造語能力に基づくスコア付け
   造語能力とは，単名詞の複合語の構成しやすさを表わす統計的尺度であり，これが大きい名詞を含む候補に高いスコアを与える [5].
2. 相対出現頻度に基づくスコア付け
   用語候補を t，その構成語数を n とするとき，t のスコアを式 (1) で与える．

$$\frac{t \text{ の出現頻度}}{\text{語数を } n \text{ とする用語候補の平均出現頻度}} \quad (1)$$

   基本的には，D 中に頻出する用語候補に高いスコアを与える．ただし，出現頻度をスコアとした場合，短い用語候補に高いスコアが与えられる．これを補正するため，同じ長さの用語候補の平均出現頻度との比をとった．

最後に，ポータルサイトのテーマとの関連度にしたがって用語候補の順位付けを再度行い，上位 20 語を関連用

語として選択する．関連度は，ウェブにおけるテーマと用語候補の共起性であると定義し，検索エンジンのヒット件数をもとに算出する [5].

上記の手法を用いて，ポータルサイトのテーマとして，「自然言語処理」,「シルビア」,「クラシック音楽」など 20 のテーマを用意し，関連用語を抽出した．我々は，造語能力に基づくスコア付けと相対出現頻度に基づくスコア付けの 2 つの手法を試したが，後述するように，それぞれの手法によって抽出された用語は異なる性質を持ち，一概にどちらがよいと決めることはできなかった．したがって，本論文は，ポータルサイトの用語集を作成するという目的において，関連用語を抽出する最適な手法は何かという結論を得るに至っていない．ここでは，我々が予備実験で得られた知見をいくつか紹介し，それに基づく今後の研究方針について述べる．

まず，造語能力と相対出現頻度によるスコア付けで抽出された用語はかなり異なる性質を持つ．造語能力に基づく手法ではテーマを部分文字列として含む用語を数多く抽出するのに対し，相対出現頻度による手法では人名，製品名などの固有名詞を多く抽出した．例えば，テーマを「クラシック音楽」としたとき，造語能力による手法では「クラシック音楽情報センター」や「クラシック音楽教室」など,「クラシック音楽」を含む用語がほとんどであった．一方，相対出現頻度による手法では,「マーラー」や「ベートーヴェン」などの人名がよく抽出された．人名や製品名を用語集に加えるべきかどうかは，ポータルサイトのテーマによって異なると考えられる．例えば,「クラシック音楽」の場合は音楽家の説明を掲載してもよいだろうが,「自然言語処理」がテーマのときに研究者の名前を掲載するのはふさわしくない．このように，スコア付けの手法によって抽出される用語の性質が異なり，またテーマによっても用語集に載せるのにふさわしい用語の性質は異なる．以上から，ポータルサイトの自動作成という立場からは，用語候補に対する唯一の絶対的な順位付けの基準は存在しないと考えられる．したがって，いくつかの抽出手法を用意し，テーマに応じて適切な手法を選択する必要があろう．その具体的な手段を探究することは今後の重要な課題である．

3.2 用語説明の獲得

用語集に掲載する用語に対して，その説明をウェブから獲得する．本研究では，特に用語が多義のとき (複数の意味を持つとき) の取り扱いに焦点を当てる．例えば，ポータルサイトのテーマが「プロ野球」で，それに関する用語として「エージェント」があるとする．ところが，エージェントには (1) 選手のために球団と交渉を行う代理人，(2) 自律的に情報を収集するソフトウェア，などの意味がある．この場合，プロ野球と関連があるのは (1) の意味なので，その意味を説明した文章をウェブから獲得する．

用語説明をウェブから獲得する試みは過去にもいくつか行われている．ここでは，多義の用語の取り扱いに関して本研究との違いを述べる．桜井らは，獲得した用語説明の集合に対してクラスタリングを行い，異なる意味の説明を弁別することを試みている [9]．これに対し，本研究では全ての意味を取り扱う必要はない．テーマにふさわしい意味だけを取り扱えばよいので，クラスタリングよりも単純な処理を行うだけで十分である．一方，藤井らは，あらかじめ 19 の分野コーパスを用意し，用語説明がある分野に出現する確率モデルを学習することにより，得られた用語説明がどの分野に属するかを判定している [8]．この手法は，用語説明が分野とどれだけ関連が深いかを測るという点で本研究に近い．しかし，本研究における分野とはポータルサイトのテーマであり，これはユーザが自由に入力することを仮定している．したがって，あらかじめ分野コーパスを用意しておくことはできない．その代わり，ポータルサイトのテーマに関連のあるウェブ文書を分野コーパスとして動的に獲得する．

以下，用語説明獲得の詳細について述べる．

3.2.1 用語説明の抽出

ここでは用語を t と表わす．まず,「t とは」と「t は」をクエリとし，検索エンジンを用いてこれらの文字列を含むウェブ文書を獲得する．次に，以下の 2 つのパタンに分けて t の用語説明を抽出する．

説明文パタン

以下のように,「t とは」や「t は」で用語説明文が始まるパタンである．

$$t \text{ とは，} (t \text{ の説明}) \langle \text{br} \rangle \text{ しかし，...}$$

まず,「t とは」や「t は」という文字列を検出し，次の HTML タグまでの文字列を用語説明として抽出する．また，それに続く文が接続詞 (しかし，など) や指示詞 (その，など) で始まる場合，用語の説明はさらに続くとみなして，次のセグメント (HTML タグで囲まれた部分) も用語説明として抽出する．

見出しパタン

以下のように,「t とは」や「t は」が見出しとなり，その後に t の説明が続くパタンである．

$$\langle \text{h1} \rangle\ t \text{ とは?} \langle /\text{h1} \rangle$$
$$(t \text{ の説明}) \langle \text{br} \rangle$$

まず，t を検出し，次の HTML タグまでの文字列を用語説明として抽出する．また，説明文パタンと同様に，次の文が指示詞や接続詞で始まる場合はその文も抽出する．

3.2.2 用語説明の選別

得られた用語説明 E に対し，ポータルサイトのテーマ x との関連度を求める．まず，x をクエリとし，検索エンジンを用いて x を含むウェブ文書を収集し，x の分野コーパスとする．さらに，E と x の関連度 $Rel(E,x)$ を式 (2) と定義する．

$$Rel(E,x) = \frac{1}{M}\sum_{n\in E} RDF_x(n) \tag{2}$$

式 (2) において，n は E 中の名詞，M は n の総数である．一方，$RDF_x(n)$ は x の分野コーパスにおける n の相対文書頻度であり，式 (3) で与えられる．

$$RDF_x(n) = \frac{df_x(n)}{N} \tag{3}$$

$df_x(n)$ は n が出現する文書数，N は x の分野コーパスの文書の総数である．$RDF_x(n)$ が大きい名詞ほど，それは x と関連が深い単語であり，そのような名詞を多く含む用語説明もまた x と関連が深いとみなす．最終的に，$Rel(E,x)$ の高い上位 10 件の用語説明を出力する．

3.2.3 予備実験

「動物」と「証券」の 2 つのテーマについて，それぞれ 10 語の用語を用意し，提案手法によって用語説明を獲得した．その結果，上位 10 個の用語説明のうち，用語説明としてふさわしい説明の数の平均は 6.5 であった．また，適切な用語説明が最初に抽出された順位の平均は 1.6 であった．抽出された用語説明の例を以下に挙げる．

- テーマ=動物，用語=ワシントン条約，順位=**1**
  ワシントン条約は絶滅の危機に瀕している動物のみを保護するものだ。動物の福祉と保護という面で言えば、ＥＵ協定 36 条とＧＡＴＴ 20 条が、動物と人間の健康と生命を守る方策を採ることを定めている。

次に，用語が複数の意味を持つとき，提案手法がテーマにあった意味をどれだけ正しく選択できるかを評価した．意味を複数持つ 5 個の用語に対し，その意味に応じたテーマを 2 つずつ与え，用語説明を獲得した．結果を表 1 に示す．表 1 中の A はテーマに関連する説明の最上位の順位，B はそれ以外の説明の最上位の順位を表わす．実験に用いた 10 語中の 7 語について，A の順位が B の順位よりも高いことがわかった (表 1 中の「A<B?」の列に Yes とある語)．これは，複数の意味の用語説明があったとき，テーマにあった意味に対する用語説明をある程度うまく選択できたことを表わす．

以上から，小規模な実験ながら，提案手法は用語説明を獲得する手法として，またポータルサイトのテーマにあった意味の用語説明を優先的に選択する手法として有望であるとの見通しを得た．

表 1: 複数の意味を持つ用語の説明の獲得

テーマ	用語	A	B	A<B?
神話	キメラ	1	5	Yes
医療	キメラ	7	6	
交通	バイパス	1	–	Yes
医療	バイパス	1	2	Yes
ヘリコプター	アパッチ	9	6	
民族	アパッチ	4	10	Yes
情報技術	エージェント	2	6	Yes
プロ野球	エージェント	3	6	Yes
Perl	ハッシュ	7	4	
暗号化技術	ハッシュ	2	7	Yes

4 おわりに

あるテーマに関するポータルサイトのコンテンツとして，関連リンク集と用語集を自動構築することを試みた．本論文はポータルサイトの自動作成に向けた最初のステップであり，課題はいくつも残されている．まず，本論文で提案した手法についての大規模な評価実験を行う必要がある．関連リンク集については，リンク集に掲載するべきページの選別や順位付けを行わなければならない．用語集については，関連用語の抽出方法についてはまだ検討の段階であり，最適なスコア付け方法などを探究する必要がある．最後に，本論文での提案手法や上記の未解決の問題に対する手法などの要素技術を統合し，ポータルサイト自動作成システムとして実現させたい．

参考文献

[1] 平野健児. 第三者による解説・評価を含む Web 関連リンク集の自動生成. Master's thesis, 北陸先端科学技術大学院大学, 3 2004.

[2] 星正人. Web からの関連用語の自動獲得. Master's thesis, 北陸先端科学技術大学院大学, 3 2004.

[3] 小林のぞみ, 乾健太郎, 松本裕治, 立石健二, 福島俊一. テキストマイニングによる評価表現の収集. 情報処理学会情報処理学会自然言語処理研究会, Vol. 2003, No. 23, pp. 77–84, 2003.

[4] 村野誠治, 佐藤理史. 文型パターンを用いた主観的評価文の自動抽出. 言語処理学会第 9 回年次大会, pp. 67–70, 2003.

[5] 佐藤理史, 佐々木靖弘. ウェブを利用した関連用語の自動収集. 情報処理学会情報処理学会自然言語処理研究会, Vol. 2003, No. 4, pp. 57–64, 2003.

[6] 菅井俊介. ポータルサイト自動作成のための用語説明獲得. Master's thesis, 北陸先端科学技術大学院大学, 3 2004.

[7] 板橋英夫, 望月源, 白井清昭, 奥村学. 参照ページからの情報を利用した Web 探索支援. 言語処理学会第 8 回年次大会, pp. 471–474, 2002.

[8] 藤井敦, 石川徹也. World Wide Web を用いた事典知識情報の抽出と組織化. 電子情報通信学会論文誌 D-II, Vol. J85-D-II, No. 2, pp. 300–307, 2002.

[9] 桜井裕, 佐藤理史. ワールドワイドウェブを利用した用語説明の自動生成. 情報処理学会論文誌, Vol. 43, No. 5, pp. 1470–1479, 2002.