WWW 文書集合から自動抽出した意味的関係を用いた大規模な検索用ディレクトリの試作

隅田飛鳥 後河内脩平 三浦二三高 相川昌裕 鳥澤健太郎
北陸先端科学技術大学院大学

1. はじめに

本稿では、大量の Web 文書と Wikipedia (http://ja.wikipedia.org/wiki/メインページ) から、自動的に検索用ディレクトリを構築する現在進行中の試みについて述べる。検索用ディレクトリ（もしくは検索用カテゴリー）はインターネット黎明期から存在し、一時は一般ユーザーのかなりの数が利用していたと言われていたが、現状はキーワード検索に圧倒されているというのが大方の認識であろう。ここで、キーワード検索と比較しつつ、ディレクトリ検索の長所／短所を著者の観点から整理すると以下のようになる。

ディレクトリ検索の長所

1. キーワード検索を使いこなすには、適切な検索キーワードを想起するのにそれなりのスキルが必要とされる。しかしディレクトリの場合、選択肢の提示があるため検索キーワードの想起を行なう必要がない。
2. ディレクトリ検索は人手で作られるのが主流であり、人間が確認済みの情報だけが入手できる。この点でユーザーにとっては安心であるといえる。

ディレクトリ検索の短所

1. 人手で作られていることから、カバーされているカテゴリ、ページ数が少ない。
2. カテゴリの分類が恣意的であり、階層構造の探索には時間を要することもある。

以下では、ディレクトリ検索の長所１を最大限に生かすことを念頭に、短所１を回避すべく、ディレクトリの自動生成を試みる。つまり、検索時に検索キーワードの想起を支援するという目標のもとにディレクトリを自動生成する。ここでは、単に「欲しい検索キーワードがのどまで出かかっているのだけれど、ちゃんと出てこない」というような状況のみならず、そもそも検索を開始した時点では想定していなかった検索キーワードの想起を支援することも含む。例えば、金沢市に観光をするとして、まず「寺社仏閣」を調べるつもりで検索を開始したが、ディレクトリ中の「寺社仏閣」の横に「名産品」というカテゴリが並んでいれば、当初全く想定していなかったにも関わらず、「名産品」の情報を詳しく調べるといったことができる。このような形での情報アクセスはいわゆる「ガイドブック」の利用においては普通に行なわれることであるが、我々は、Web においては文書が書籍に比べて断片的であること、またユーザーがガイドブックを見る場合に比べてより目的志向的であることからあまり行なわれておらず、もし十分なカバレージのあるディレクトリが生成され、それらが多様な対象に関して「ガイドブック」的役割を果たしうるのであれば、大変意味があると考えている。

一方で、ディレクトリを自動生成する以上、長所２に挙げた信頼性を保証するのは非常に難しく、本研究では長所２を実現することは狙わない。また、短所２に関しては、基本的に、カテゴリ間になるべく多くのリンクを張ることで必要なカテゴリに容易に到達させることを狙うが、これに関してもリンクが増えれば、検索時に読まなければいけないリンクのアンカーテキストが増え、使い勝手が悪くなることも容易に想像できる。この点に関する検討は今後の課題とする。

また、ディレクトリの自動生成は Sato らも試みている(Sato, 1999)が、我々の狙いは、Sato らのように特定のドメインに限定するのではなく、より多様な対象に関して、巨大でカバレージの大きいディレクトリを生成することである。

2. 利用したリソース

本研究で利用したリソースは以下の２点である。

金沢市

概要 (括弧内は各アイテムからさらに辿れるアイテム数)

マスメディア(10) 観光(6) 金沢市出身の有名人(5) 金沢市を舞台とした作品(5) 交通(5) 経済(4) 歴史(3) 地理(3) 主な学校(3) 施設(3) 行政(1) 姉妹都市(0) 概要(0)

図 1. ターゲットに関する概要情報

下位語等 (括弧内は各アイテムからさらに辿れるアイテム数)

レストラン(3544) 無集配局(842) スナック(572) 駐車場(137) ショッピング(81) 石川県中学校一覧;公立中学校:金沢市:市立中学校(24) 金沢駅(13) 自動販売機(12) 石川県の郵便局一覧;市部:金沢市:集配局(7) 金沢五社(6) 石川県青年会館(6) 天徳院_(金沢市)(6) 湯涌温泉(4) 香林坊(4) 金沢21世紀美術館(4) 金沢市立玉川図書館(4) 第14普通科連隊(4) 兼六園(3) 犀川_(石川県)(3) 金箔(3) 石川県立金沢向陽高等学校(3) 石川県立金沢桜丘高等学校(3) 尾山神社(3) 石川県立金沢二水高等学校(3)

図 2. ターゲットの下位語, もしくはターゲットの一部を指すメロニミー

1. Wikipedia（2006 年 9 月、約 38 万記事）
2. 約 0.7TB の Web 文書(約 4500 万文書)

1 を取得した時点では Wikipedia のトップページには，約 26 万記事と書かれていた。この約 38 万記事というのは約 26 万記事にカウントされていない書きかけの記事も含む。また 2 は我々の研究室で独自にクローリングを行ない、収集したものである。各々の位置づけであるが、Wikipedia は多数の信頼のおける情報が特に上位の概念レベルに関して掲載されているものの、「歴史上の人物」等の特殊なケースを除くと固有名詞がそれほど多くない。特に、我々の日常生活に関する固有名詞（商品名、レストランの名称等）はほとんど掲載されていない。一方で約 0.7TB の Web 文書からは、上位の概念間の階層関係を獲得するのは現時点では難しいものの、隅田らの手法(Sumida, 2006)によって比較的大量の固有名詞とその上位概念の関係が獲得できる。本研究の狙いは、これらの性質の異なる 2 つのリソースを使うことで、大規模なシソーラス的な階層構造を作成し、これを用いて検索ディレクトリを作成することである。

3. ディレクトリの全体像

図 1 ー図 4 に実際に生成されたディレクトリ中の一ページを示す。(ページ自体が比較的大きなサイズであるため、分割して示した。) このページは「金沢市」に関するものであるが、以下では、このような一ページが対応づけられる自然言語表現の

石川線(5)	野町駅(9) 乙丸駅(4) 額住宅前駅(4) 馬替駅(4) 四十万駅(4)
私立(5)	私立金沢高等学校(6) 遊学館高等学校(3) 金沢工業高等専門学校(2) 星稜高等学校(1) 金沢科学技術専門学校(1)
博物館(4)	前田土佐守家資料館(1) 石川近代文学館(1) 藩老本多蔵品館 石川県立伝統産業工芸館

図 3. 他のターゲットの特徴語

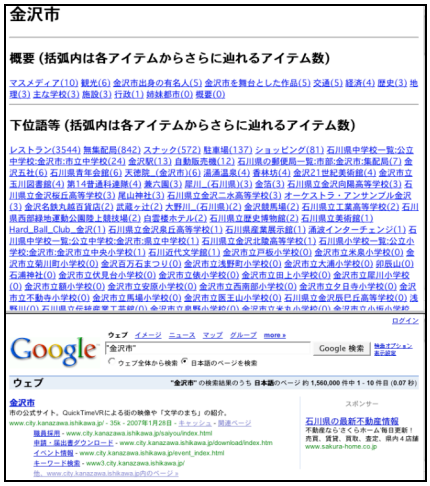

図 4. ターゲットの検索結果

ことを「ターゲット」と呼ぶことにする。つまり、図中のページに関して言えば、ターゲットは「金沢市」である。各ページは以下のように 3 つの部分に分かれる。

1. ターゲットに関する概要的情報（図 1）
2. ターゲットの下位語、もしくはターゲットの一部を指すメロニミー （図 2）
3. 1 、2 から辿れる他のターゲットの特徴語（図 3）

1 は、後述するように Wikipedia のターゲットに関するエントリをパターンマッチングで解析することによって得られており、階層的な構造をしている。2 は、大量の Web 文書から自動獲得されたものの他、Wikipedia の各エントリから抽出された定義文に語彙統語パターンを適用することで抽出されたものや、Wikipedia にもともと存在する category の階層関係から抽出されたものも含む。3 は、Wikipedia から抽出された定義文に現れた名詞を特徴語として提示している。これは例えば、「ロボット」というターゲットから「ソニー」という特徴語を介して「AIBO」「QRIO」といったターゲットに飛ぶことを可能とする。

なお、ディレクトリの階層を辿って、あるター

ゲットに対応したページに到達すると、同時にターゲットを検索エンジン(現在はGoogle)に投げ、検索結果も同時に出力される。(図４) 現在は、単にターゲットとなっている日本語表現をクエリーとして検索を行なっているだけだが、これは、最終的なシステムの全体像をデモンストレートするためにそうなっているだけで、今後は現在注目しているターゲットにいたるまでのディレクトリ探索の経緯を考慮したクエリーの生成や、言い換え等を考慮した柔軟な検索の導入を考えている。

4. 作成手法と現状のディレクトリ

以下では本ディレクトリの生成手法をより詳しく説明する。まず、一般の Web ドキュメントからの上位下位関係の獲得手法は、Hearst の語彙統語パターンに依るアプローチ(Hearst, 1992)を拡張したものであり、名詞列のパターンである、

上位語「下位語」 e.g., 映画「ジョーズ」

上位語 下位語 e.g., 映画ジョーズ

の二つによって、大量の上位下位関係の獲得を行なうものである。特に、二番目のパターンは、上位語と下位語の間の境界が明示的に判別できないため、一番目のパターンから獲得され、かつ人手のチェックを経た上位語を用いること、検索エンジンから得られるヒットカウントを用いてフィルタリングを行なうことなどによって、高い精度で上位下位関係を獲得する。詳細は(Sumida, 2006)を参照されたい。

ついで、Wikipedia の解析であるが、まず、ターゲットに関する概要的情報(図１)については、Wikipedia のソースから単純なパターンマッチングによって取得する。図 5 に「金沢市」に対する Wikipedia のエントリーのソースの一部を示すが、まず、見出し等を示す`===', `*'等の Wikipedia 特有の修飾記号に対してパターンを用意する。ついで、そのパターンをマッチさせて見出しを特定し、次に修飾記号間の階層的関係に従って、見出し間の意味的関係を抽出する。例えば、図 5 の例で行けば、まず、「金沢市」と最も上位の見出しである「観光」の間の関係が認識され、ついで「観光」と「名産品」、「名産品」と具体的な名産品の名称である「森八長正殿」の間の関係が認識される。なお、このようにして獲得された関係は階層的構造を持つが、 その最も下位のもの (e.g., 森八長生殿)を除き、「観光」等の見出しは、様々な

```
==観光==
== 名産品 ===
*森八「長生殿」([[和三盆]]を用いた上質な[[落雁]])
*[[かぶら寿司]] (熟れ鮨の一種)
*[[ゴリ]]料理　([[佃煮]]、[[天ぷら]]など)
```

図 5. 「金沢市」に対する Wikipedia のエントリのソースの一部

エントリーに出現するものと考えられるため、実際には、「金沢市:観光」のように、ターゲット「金沢市」に特化したカテゴリと見なす。これにより、例えば、金沢市と東京都の「観光」カテゴリが混同されるのを防ぐことができる。

前述したように図 2 の「下位語等」の項目にはメロニミーが多数含まれているが、これらの多くは、この Wikipedia の階層構造の解析から得られるものである。例えば、Wikipedia の「石川県小学校一覧」というページの「金沢市」の項目の下に「金沢市立西小学校」があることから、「下位語等」の項目に「金沢市立西小学校」が列挙されている。(上述した、「観光」のような見出しは他のターゲットに依存したカテゴリとするという制約はこの場合弱められている。)

また、Wikipedia から抽出した定義文からも上位下位関係を抽出する。まず、ある語に関する Wikipedia エントリの最初の一行は、その語からはじまる簡潔な定義文であるという性質を積極的に利用する。(e.g., **金沢市** (かなざわし) は、北陸地方の西部、石川県のほぼ中央に位置する都市で、同県の県庁所在地である。)ついで、これらの定義文には「X は..＜X の上位語＞である。」等の頻出するパターンが存在することに注目し、特に文末から＜X の上位語＞に至るまでの表現を合計１２５個、手作業で抜き出した。これにより上位語の位置を特定し、X とその上位語の関係を抽出した。上の例では「金沢市」の上位語として「県庁所在地」が獲得されることになる。この手法で獲得された上位下位関係の中から、ランダムに抽出した１００個のうち、８２%が適切な上位下位関係と見なせた。

抽出された定義文は特徴語を抽出するのにも用いた。この特徴語は現状では、定義文中の名詞句、または動詞のうち上位語以外のものをすべて抽出し、その中から、ターゲットの概要、もしくは、下位語等から辿れる他のターゲットに二つ以上共通して現れるものをすべて、ページ中に列挙して

いる。もちろん、このような単純な方法ではあまり意味のない特徴語も大量に取得されるため、今後、さらなる検討／改良を加える予定である。

また、最後にもともと Wikipedia に備わる category と呼ばれる階層構造がある。これは、あるターゲットに関するエントリ中に執筆者がターゲットの上位語に相当する category をタグによって挿入しているものであり、どの category をふるかは執筆者の恣意によるため、大きなカバレージは期待できないものの、精度は高い。

実際のディレクトリは以上のように抽出された二項関係の集合を、ルートとして与えられたいくつかのターゲットから網羅的な探索をしていくことによって生成される。実際にはループも存在するため、探索の深さには制限を加えている。

以上の手法を用いて作成した、現状のディレクトリは､約１５９万語のキーワードを含んでおり、「金沢市:観光」のような他のターゲットに依存したターゲットまで含めると、約２６０万個のターゲットからなっている。このうち Wikipedia 以外の文書から取得した上位下位関係は約２８万語（約３８万個の関係）であり、残りの約１３０万個のキーワードは何らかの形でWikipediaから抽出されたものである。このキーワード数はディレクトリ作成時の Wikipedia の記事数(約３８万)よりかなり多いが、これは Wikipedia の記事中で使われている表現で、記事の見出しになっていないものが多数存在するためである。

5. 今後の課題

今後検討すべき課題は多数あるが、まず、今回作成したディレクトリは、我々が計画しているディレクトリの一部でしかないことを強調しておきたい。以後は、ターゲットに関わる人の行動を反映した選択肢を加えて行くつもりである。例を言えば、「金沢市」に「行き方」「住環境」「予想されるトラブル」等の項目を追加していく予定である。２の計画については本大会併設ワークショップにおいて発表予定である。（鳥澤, 2007）

次に今後検討すべき課題の一部を列挙する。1）「下位語等」の項目の下に非常に多くの項目が列挙されるが、これらは何らかの基準により分類して表示すべきである。これに関しては faceted search(Hearst,2002) の考え方を導入すべきであり、それに関しては Web 上からの属性/属性値のマイニング(吉永, 2007)が有効であろうと考えている。また、上位下位関係とメロニミーとの区別も重要な課題であり、これらを正確に認識することがディレクトリ全体の精度向上につながるであろう。2）単なる上位下位関係とは異なり、階層的な構造を生成しているが故の問題が存在する。例えば、「パン」というのは食物であるだけでなく、「伝説の生物」でもある。このような状況で単純な探索により階層構造をくみ上げると､「伝説の生物」>> 「パン」>> 「菓子パン」といった奇妙な階層構造が出来上がる。これは多義性の問題でもあり、興味深い研究課題である。 3）本稿で生成したディレクトリはこの種のものとしては巨大である。しかしながら、例えば、数百件のラーメン屋のリストは含まれているが、近所にあるラーメン屋の名前が抜け落ちていたりする。これには、より大量の Web 文書を利用する、あるいは、新たな上位下位関係の獲得手法を開発することで対応する予定である。

6. まとめ

本稿では、検索の支援を行なうための大規模検索ディレクトリの試作について報告し、今後の検討課題、拡張について述べた。

参考文献

Asuka Sumida, Kentaro Torisawa, Keiji Shinzato. 2006. Concept-Instance Relation Extraction from Simple Noun Sequences Using a Search Engine on a Web Repository, Proc. of the Web Content Mining with Human Language Technologies workshop

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Proc of COLING.

Marti Hearst, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ping Yee. 2002. Finding the Flow in Web Site Search, Communications of the ACM, 45 (9).

鳥澤健太郎. 2007. 検索要求の想起支援に向けて. 第１３回言語処理学会年次大会併設ワークショップ「大規模Web研究基盤上での自然言語処理・情報検索研究」

吉永直樹, 鳥澤健太郎. 2007. Webからの具体物の属性・属性値情報の自動獲得. 第１３回言語処理学会年次大会.

Satoshi Sato and Madoka Sato. 1999. Toward Automatic Generation of Web Directories. Proc. of ISDL'99.