5W1H情報の在否により結果を分類する情報検索システム

3P－13

池田 崇博　佐藤 研治　奥村 明俊
NEC C&Cメディア研究所

1. はじめに

インターネットの急速な普及とともに、情報の入手が簡単になった反面、無数の情報があふれ、必要な情報だけを選別することが困難になってきている。Web情報を検索するサービスは、すでに数多く提供されているが、入力したキーワードによる条件だけでは候補を絞りきれず、膨大な検索結果の中から、改めて目的のページを探し出さなければならないことも多い。逆に、条件が厳しすぎ、結果が0件になってしまうこともしばしばである。

例えば、宴会の幹事になったユーザが、適当な会場を探すために、「居酒屋」「東京」という2つのキーワードでWeb情報検索を行ったとする。この場合、居酒屋を紹介するページがユーザにとって有益なページということになるが、検索結果には「居酒屋」と「東京」を含むすべての文書が含まれる。例えば、東京の居酒屋に関する経済レポートや、演歌のタイトル一覧など、このユーザにとって不要な情報も数多くヒットし、検索結果の上位には、有益なページがほとんど存在しないこともあるだろう。

この例では、ユーザが求めている本質的な情報は、実際には、居酒屋の電話番号、価格、場所などであり、ユーザが必要としているページは、これらの情報が含まれているページである。キーワードでは、これらが含まれているという条件を的確に指定できないために、十分な絞り込みの条件を与えられず、本来不要であるページまでもがヒットしてしまう。

そこで、本稿では、住所（Where）、日時（When）、価格（How much）といった、5W1Hに関する情報が文書中に存在するかどうかによって検索結果を分類し、検索結果を絞り込む5W1H絞り込みナビゲーションを提案する。

筆者らは、これまでに、文書中から5W1H情報を抽出し、文書検索・文書分類を行うシステム[1]、および、5W1HのWhen情報に着目してスケジュール情報を抽出し、ユーザに配信するシステム[2]を開発してきた。本稿では、特に高精度で抽出可能であり、ユーザが有益な情報として活用できることの多い、住所、電話番号、最寄り駅、日時、価格などの5W1H情報に着目し、検索結果を分類する。

Information Categorization Based on Presence of 5W1H Information
Takahiro Ikeda, Kenji Satoh, and Akitoshi Okumura
NEC C&C Media Research Laboratories

2. 5W1H絞り込みナビゲーション

2.1.5W1H要素の抽出方法

プロトタイプシステムでは、以下の5W1H情報によって検索結果を分類する。

- 住所（例：東京都目黒区）
- 電話番号（例：03-xxxx-xxxx）
- 最寄り駅（例：渋谷駅）
- URL（例：http://www.xxx.co.jp/）
- e-mailアドレス（例：xxx@xxx.co.jp）
- 価格（例：5,000円）
- 日時（例：10月1日）
- 組織名（例：××屋）

これにより、例えば、電話番号が書いてある飲食店情報のページや、価格が書いてある商品情報のページ、日時が書かれているイベント情報のページなどを検索することが容易になる。

このために、あらかじめ、検索対象となる文書から上記5W1H情報を抽出し、各文書に含まれている5W1H情報の種類と内容について、インデックスを作成しておく。5W1H情報の抽出は、以下の3種類の方法で行う。

1. パターンマッチングによる抽出
   抽出対象の5W1H情報は、いずれも特有のパターンで書かれるという特徴を持っているため、基本的にパターンマッチングで抽出する。例えば、「～市～区～」というパターンにマッチする部分は住所情報、「株式会社～」というパターンにマッチする部分は組織名として抽出する。
2. 文書のフォーマットに着目した抽出
   文書中で箇条書きになっている部分に着目し、そのラベルから対応する情報を抽出する。例えば、「店名：　××」と書かれている場合、組織名として、「××」を抽出する。
3. 辞書を用いた抽出
   最寄り駅、組織名については、固有名詞辞書を参照し、上記の方法で抽出されない場合でも、辞書にあれば抽出する。

2.2.絞り込みナビゲーションの方法

ユーザが、いくつかのキーワードを入力し、検索ボタンを押すことで検索が開始される。ユーザには、検索条件に適合する文書のリストとともに、前節の

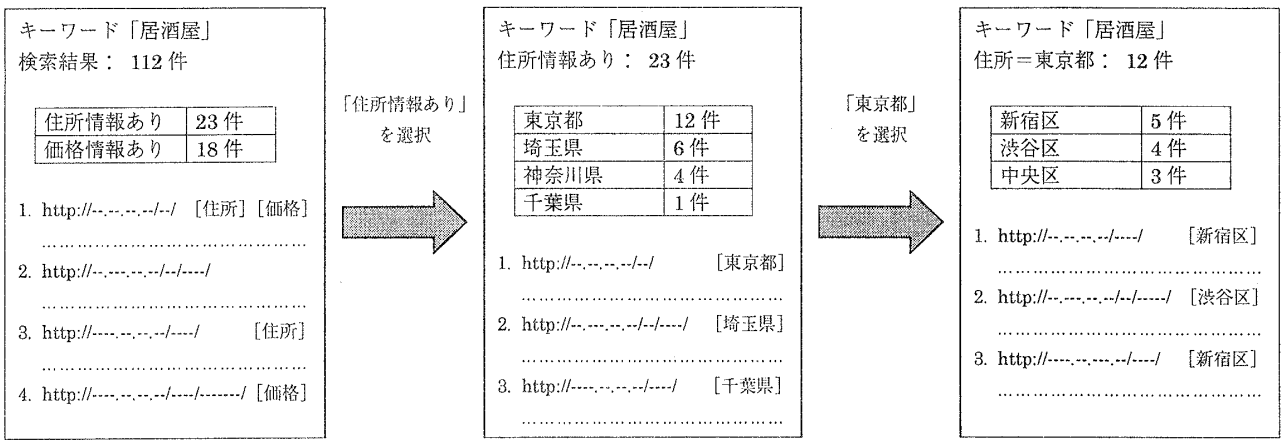

図 1：5W1H 情報による検索結果の絞り込み

ようにして抽出した 5W1H 情報によって文書を分類した結果を提示する。このとき、分類項目の数を適当な数に抑え、ユーザが容易に検索結果を絞り込むことができるように、再帰的に分類を行うインターフェイスを採用する。再帰的な分類のイメージを図 1 に示す。

まず、各文書が前節で挙げた 5W1H 情報のどれを含むかによって検索結果を分類し、各 5W1H 情報を含む文書の件数を表にして提示する。ユーザが、分類結果の特定の 5W1H 情報を選択した場合、検索結果をその 5W1H 情報を含む文書だけに絞り込むとともに、その 5W1H 情報として抽出された内容によって分類した結果を提示する。例えば、ユーザが、住所情報を選択した場合には、住所情報として抽出された情報を都道府県単位で分類して表示する。ユーザが、さらにその中の 1 つ、例えば、「東京都」を選択した場合には、検索結果を住所情報として「東京都」を含む文書だけに絞り込み、市区町村単位で分類して表示する。

5W1H 情報による条件により絞り込まれた検索結果を、文書のリストとして常に下部に表示する。このとき、各文書にどの 5W1H 情報が含まれているか、あるいは 5W1H 情報のどの内容が含まれているかが分かるようにし、目的の文書にユーザを適切にナビゲートする。

3. まとめ

本稿では、住所、電話番号、日時、価格といった、5W1H に関する情報が文書中に存在するかどうかによって検索結果を分類し、順次結果を絞り込む 5W1H 絞り込みナビゲーション手法を提案した。これにより、例えば、飲食店情報のページを検索する際に、電話番号が書かれているページや、価格が書かれているページだけに検索結果を絞り込むことが容易になる。また、絞り込んだ結果を、さらに 5W1H 情報の内容で分類することで、目的に合ったページをユーザが容易に選択できるようになる。

この検索方式をプロトタイプとして実装し、各 5W1H 情報の抽出精度、分類の有効性を検証していく。また、今後、以下の点についても検討を進める。

・キーワードに応じた絞り込み条件の提示
検索の目的によって、絞り込みに有効な 5W1H 情報は異なる。例えば、住所の情報は、飲食店情報を検索する際には有益だが、新商品情報の検索ではあまり意味を持たない。ユーザが入力したキーワードに応じて、5W1H 情報による絞り込み条件を変える必要がある。

・マルチメディア情報抽出
飲食店情報の検索などでは、住所が含まれているページよりも、地図が含まれているページの方が有用であろう。5W1H 情報として、テキストだけでなく、画像も扱えるようにすべきである。

・5W1H 以外の情報抽出
分類に有効な情報は、必ずしも 5W1H に関する情報に限らない。例えば、製品や店舗の評判に関する情報などを分類軸にできるとよい。

参考文献

[1] 池田崇博、奥村明俊、村木一至：MIIDAS：情報の選別と Easy Reading のためのエピソード、情報処理学会第 55 回全国大会、5Q-10 (1997)

[2] 池田崇博、佐藤研治、奥村明俊：非定形文書中の日程情報を自動配信するスケジュールリマインダ、情報処理学会第 57 回全国大会、2H-01 (1998)