2007－DD－63 (2)
2007／9／28

情報取得支援のための用語索引システム

遠藤 健一　藏野 昌彦　梅澤 淳子

独立行政法人 国立印刷局

E-mail : seihin01@res.npb.go.jp

現在最も広く使われている情報検索の方法である全文検索方式では、ユーザーは検索のための用語を入力する必要があるが、調べたいテーマについて予備知識を持たないユーザーには適切な用語を発見することは難しく負担が大きい。一方、書籍では、本文中の用語が巻末索引に一覧として示されているため、このような問題がある程度解消されている。そこで、本研究では索引方式、すなわち、あらかじめ用語を提示してユーザーからの選択を求める方式の用語索引システムを試作した。

A Terminology Index for aiding Information Acquisition

Ken'ichi Endo, Masahiko Kurano, Atsuko Umezawa

National Printing Bureau

Full-text search engines are the most popular means of information retrieval. It is necessary for the user of a full-text search engine to input keywords for searches. However, users who lack prior knowledge about a topic are burdened by the difficulty in finding suitable keywords. This sort of problem is to some extent relieved in books that list terminology in the main text in an index at the end. In this paper, we propose a terminology index system which shows the user keywords, and invites the user to choose among them.

1 はじめに

ウェブで情報検索をする際に用いられるGoogleやYahoo! JAPANなどの全文検索サービスでは、調べたいテーマに関係しそうな用語をキーボードで入力すると、検索結果としてヒットしたウェブページのURLや内容の一部などが一覧で示される。

これらを使用していると、検索のために入力すべき用語（以下、検索語という）が思いつかない、自分が記憶していた用語とページ中の用語が一致しておらず期待した検索結果が出ない、思いつきやすい簡単な用語を入力すると必要でないページを含む膨大な検索結果が示される、などの不便な状況にしばしば直面する。

このような状況に対し、ユーザーは検索結果で示されたページに目を通すなどして、よりよい検索語を発見しようと試みるが、この作業は負担が大きいという問題がある。

一方、印刷物を振り返ってみると、専門的な解説を行っている書籍には巻末に索引が付いていることが多い。

巻末索引は、書籍中に現れる用語を五十音順に並べたり、分野別（人名索引等）に分けたりして構成され、用語が現れる本文のページ番号を示すことが主な目的であるが、その書籍に含

表1　検索における行動の例

	①ニーズ	②ニーズの言い換え	③検索語	情報が得られるページ
例1	製品Xについて知りたい。	製品Xの仕様が書かれているページはどこか？	製品X　仕様	製造元のホームページ。 販売店のホームページ。
例2	製品Xの評判を知りたい。	製品Xのレビューが書かれているページはどこか？	製品X　レビュー あるいは 製品X　欠点	ブログ。 クチコミサイト。
例3	製品Xを安く買いたい。	製品Xの価格が書かれているページはどこか？	製品X　価格	販売店のホームページ。 仮想商店街ポータルサイト。 価格比較サイト。
例4	"APEC"の正式名称を知りたい。	"APEC"の説明が書かれているページはどこか？	APEC あるいは APECとは	説明が詳しく書かれたページ。 百科事典サイト。

まれる用語が一覧として示されているため、上記の「検索語の発見」に関する問題が、ある程度解消されている。

そこで、本研究では索引方式、すなわち、あらかじめ用語を提示してユーザーからの選択を求める方式を検討した。

2　検索のニーズと検索機能

2.1　検索における行動

ユーザーは検索を行う際に、自分のニーズを「検索語を含むページがどこにあるか」という形式に近づくように言い換えながら、検索条件（本稿では検索語の列だけを考える）を作り、これを全文検索サービスに渡す(表1の①～③)。

そして同サービスから検索結果を得るが、この時点で検索は終わりではなく、得られたURLが指し示すページに目を通してニーズが満たされれば、一連の検索行動は終わる。しかし、得られた内容でニーズが満たされない場合は検索条件を修正して検索を繰り返すことになる。

金田ら[1]は、情報検索の本来の目的はURLのリストを得ることではなく、知りたい情報(例えば、知りたい情報をまとめたレポート）を得ることである点を指摘し、これを「情報取得」と呼んで伝統的な「情報検索」という用語と区別している。

全文検索サービスは、ページのURLを提示しており、ユーザーの求める情報自体は提示しないが、使い方がシンプルで、情報収集範囲が広く、更新が頻繁であるため、「最初に使うツール」としてユーザーに選ばれている。

製品の価格や評判などのように典型的で多数のユーザーが求めているニーズに対しては、価格比較サイトやクチコミサイトなど、URLではなくユーザーの求める情報そのものを提供するサービスがある。

2.2　全文検索が不得意とする分野

多くのホームページでは外部の全文検索サービスを利用した「サイト内検索」機能を付加しているが、そのページの掲載内容についての予備知識を十分に持たないユーザーは、この機能を前にしても検索語を思いつかず、この機能を活用できない。

このようなケースでは、通常の検索における行動に先立って、「ある組織のホームページ」や「ウェブで公表された報告書」のように限定された対象の中に「どんな用語が現れるか」を知りたいというニーズが発生している。

しかし、検索語を含むページがどこにあるかを提示するという全文検索サービスは、このようなニーズに対応していない。

吉井ら[2]は、全文検索における検索語の選択や、その用語が文章中で実際にどのように表記されているかを知ることがユーザーにとって困難である点に注目し、行政機関のウェブサイトで提供されている行政情報に現れる用語の分析や頻度調査などの基礎的調査を行っている。

3　用語のデータベース化と活用

3.1　索引方式の概要

本研究では、上記のような全文検索が不得意

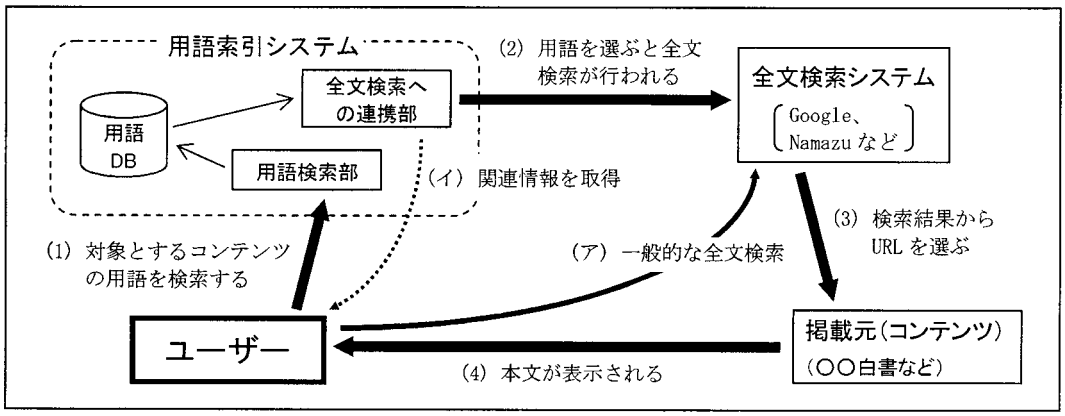

図1　用語索引システムを用いた文書検索の流れ

とする分野でのユーザーの負担を軽減するため、書籍の巻末索引の効果に着目し、これを実現するシステムを試作した。

本研究では、用語を提示しユーザーの選択を求める索引の機能を、①対象とするコンテンツ（ウェブページや文書ファイルの集まり）に使用されている用語を切り出してデータベース化し、②書籍の巻末索引と同等かそれ以上に便利に利用できるように用語データベースの多様な検索方法を用意し、③検索された用語を本文の検索語として利用する機構を設けることで実現した。

索引方式による検索の流れを図1に示す。ウェブ上にユーザー、用語索引システム、全文検索システム、掲載元（コンテンツ）の4者が存在する。

ユーザーは、用語索引システムで検索語の候補を用語データベースの中から検索した後、全文検索システムを利用する。これは図1中では(1)→(2)→(3)→(4)という流れとなる。なお、一般的な全文検索を行う場合は、(ア)→(3)→(4)という流れとなる。

3.2　対象としたコンテンツ

国立印刷局ホームページに掲載されている「白書のあらまし」[3]という書籍コンテンツから「平成18年版外交青書のあらまし」など数タイトル分を用いた。

「白書のあらまし」のオリジナル版では1タイトル分の本文が1個のHTMLファイルとなっているが、索引の効果を高めるために、本文中の見出しを拾って目次を作成するとともに本文を複数のファイルに分割し、用語索引システムと同じサーバーにファイルを掲載してオリジナル

表2　用語データベースの例（「平成18年版外交青書のあらまし」から）

用語	読み	付加情報
愛・地球博	あい ちきゅうはく	[イベント]
アジア・アフリカ首脳会議	あじあ あふりか しゅのう かいぎ	[会議・会合]
アジア欧州会合	あじあ おうしゅう かいごう	
アジア太平洋経済協力	あじあ たいへいよう けいざい きょうりょく	APEC
アジア・大洋州外交	あじあ たいようしゅう がいこう	
アジア通貨危機	あじあ つうか ..き	
麻生外務大臣	あそう がいむ だいじん	[人名]
アナン国連事務総長	あなん こくれん じむそうちょう	[人名]
アフガニスタン	あふがにすたん	[国名・地名]
アフリカ開発会議	あふりか かいはつ かいぎ	[会議・会合]
安保理	あんぽり	
安保理改革	あんぽり かいかく	

版の代わりに用いた。

3.3 用語索引システム

本システムは、以下に述べる3要素で構成される。用語データベースにはMySQLを、用語検索等のプログラムにはPHPを使用し、Apacheベースのウェブサーバー(ホスティングサービス)上にシステムを構築した。

(1) 用語データベース

対象コンテンツの本文から用語を切り出して構築したデータベースである。試作では「白書のあらまし」のタイトル毎にデータベースを構築した。用語データベースに収録される用語の例を表2に示す。

用語の切り出しには、形態素解析システム「茶筌」[4]と、東京大学中川研究室・横浜国立大学森研究室で開発された用語抽出システム(termex.pl)[5]を用いた。茶筌が文章を形態素に分解し、これをtermex.plが受け取って複合名詞を含む名詞を抽出する。

次に、ユーザーが検索に使わないと思われる用語を取り除いた。例えば、「平成18年版外交青書のあらまし」の場合、「国際社会」、「国」、「協力」、「地域」、「関係」などである。

書籍の索引では紙数等の制限や読者自身の手間の点から、収録語数を増やし過ぎない方が使いやすいと考えられるが、用語データベースはコンピュータが活用できることから、語数の制限は設けなかった。

例えば「安保理」、「安保理改革」のように部分的に重複する用語が切り出された場合には、両方とも用語データベースに登録した。

同時に、用語の漢字表記や用語自体をユーザーが正確に記憶していない場合でも用語を検索できるように、データベースに読みがなやその他の情報を付加した。

(2) 用語検索部

用語データベースに対する検索条件を受け付けるユーザーインターフェースである。

例として、「平成18年版外交青書のあらまし」を選んで分野別索引の「会議・会合」をクリックしたときの画面を図2に示す。画面の右側のフレームが用語検索部である。

検索方法として、読みがなから選ぶ「五十音

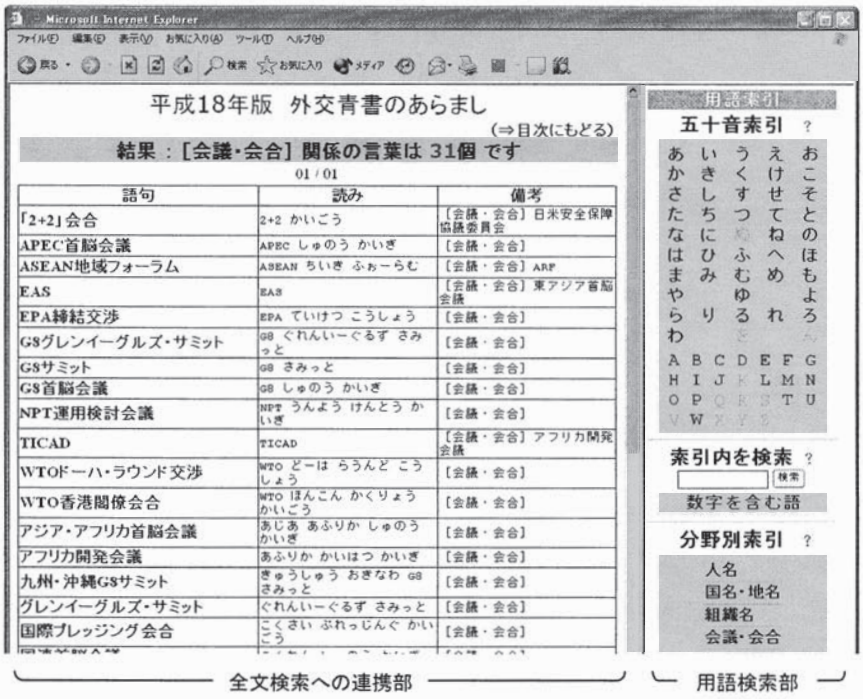

図2 用語索引システムの画面の例

索引」、用語データベース全体のテキスト検索をする「索引内を検索」、付加情報を検索する「分野別索引」を設けた。

(3) 全文検索への連携部

用語検索の結果を表示し、用語の選択によって全文検索を実行するためのユーザーインターフェースである。用語データベースの検索結果から画面表示用 HTML データがプログラムで生成される。

図2に示された画面の左側のフレームが全文検索への連携部である。用語データベースを検索してヒットした用語がここに表示される。この用語が全文検索へのリンクになっており、ユーザーが用語をクリックすると、本システムから全文検索システムに検索条件が渡されてコンテンツの検索が行われる。

3.4 全文検索システム

全文検索システム Namazu[6]用の Perl 版検索プログラムである pnamazu[7]を用いた。

これを用語索引システムと同じサーバー上で稼動させることとし、pnamazu.cgi と、タイトル毎のインデックスファイルを転送した。

4 結果

コンテンツの本文中の用語をデータベース化し、用語索引システムを構築することで以下のような効果が得られた。

① ユーザーは、本文中の主な用語についてはリンクをクリックするだけで検索でき、キーボードを打つ必要がなく、操作が容易である。

② 同一の対象を指す用語が複数ある場合でも（例：「外務大臣」と「外相」）、用語検索結果に表示された用語は、本文中で実際に使われたものであるので、その後の全文検索で必ずヒットする。

③ 書籍の索引よりも多くの用語を収録できる。

④ 用語に付加した関連情報を用いて用語を多面的に検索できる。

五十音索引や分野別索引で用語を表示させることにより、人名等の漢字表記が分からない場合や、用語を正確に記憶していない場合でも検索語を発見できる。

⑤ 「索引内を検索」を用いて、多数のページに目を通さずに、ひとつの用語に関連する情報を得ることができる。

表3は、テスト的に「平成16年版情報通信白書」の本文から用語の切り出しを行い、その中で「携帯」を含む用語を検索した結果である。

この用語検索結果が、よりよい検索語を発見するための参考になれば、図1中の(1)→(2)→(3)→(4)の一部が(1)→(イ)で済み、ユーザーの負担が軽減される。

表3 「平成16年版情報通信白書」に現れる「携帯」を含む用語

GPS 対応型携帯電話端末	携帯端末	携帯電話向け有料コンテンツ利用者
PDA	携帯端末対応	携帯電話モデル
アナログ携帯電話	携帯端末向け放送	携帯電話料金
インターネット対応型携帯電話	携帯電話	携帯電話利用マナー
カメラ付き携帯電話	携帯電話インターネット契約数	携帯電話利用者
携帯インターネット	携帯電話加入者	携帯電話機
携帯インターネット契約数	携帯電話加入数	携帯電話網
携帯インターネットサービス	携帯電話関連消費	携帯型放送
携帯インターネット普及状況	携帯電話契約数	次世代携帯電話
携帯インターネット未利用者	携帯電話サービス	属性別携帯インターネット利用率
携帯インターネット利用格差	携帯電話サービスエリア	第3世代携帯電話
携帯インターネット利用者	携帯電話事業者	第三世代携帯電話
携帯インターネット利用率	携帯電話端末	ブロードバンド・携帯端末対応状況
携帯情報端末	携帯電話等を用いた119番通報のあり方検討懇談会	ブロードバンド・携帯端末対応電子商取引
携帯情報通信端末		

※ 本表に「PDA」が現れるのは、「PDA」の付加情報に「携帯情報端末」があり、これが検索でヒットしたことによる。

⑥ 略語の正式名を知りたい場合など、用語検索の段階で分かる場合がある（例：表2・「APEC」参照）。この場合、図1の(1)→(イ)の流れでニーズが満たされる。

5 終わりに

情報検索に最も広く用いられている全文検索サービスは、「最新版の○○白書」のように対象が限定されている状況でのニーズには対応していない。

本研究では、このような状況に対応する情報取得のアプローチ方法として、用語索引システムを試作した。

書籍コンテンツ以外のウェブページに索引を付けることも有意義と考えているが、本文から機械的に切り出された語から検索に使われそうにない語を取り除く作業の効率化が課題である。

また、表3を見ると、対象とした書籍には「第3世代携帯電話」と「第三世代携帯電話」、あるいは、「携帯情報端末」、「携帯情報通信端末」、「携帯端末」などがあり、用語がばらついていることが分かる。編集中の書籍を対象とした用語データベースを構築することで、用語の乱れをチェックするなど、編集者に対する支援として技術を利用できる可能性もある。

試作では、白書などの専門的な書籍コンテンツを対象としてシステムを構成したが、対象とする情報の性質や、システムの狙いに応じて実装方法を変えることができる。

① 用語索引（用語提示）

試作では多数の用語を様々な切り口から検索できるようデータベースを用いたが、用語数が少ない場合は、データベースを用いずに、検索語と検索範囲を含む全文検索へのリンクを並べた「検索キーワード集」を作る方法もある。

また、サーバーにデータベースをインストールできない場合などでも、用語数が多くなければ「検索キーワード集」を用いて用語索引を構築することができる。

② 全文検索

試作ではNamazuを用いたが、ウェブ上の全文検索サービスに検索語と検索範囲を渡す方法もある。

Namazuでは、ページの上部や左右に付加されている様々な情報と中央部の本文情報の区別ができないため、本来上位にランクされるべきページが正しく評価されない場合がある。全文検索サービスではこれらが改善される場合がある。

全文検索にGoogleを利用した場合のリンクの形式を以下に示す。

```
<a href="http://www.google.co.jp/
search?q=%22検索語%22&sitesearch=
検索範囲">見出し語</a>
```

社内LAN等にある文書ファイルのように非公開で外部の全文検索サービスを利用できない場合は、全文検索システムを自分で用意することになる。

上記①②の組み合わせでは、ウェブ上の全文検索サービスを利用した「検索キーワード集」が用語索引の最も簡単な実装方法となる。

参考文献

[1] 金田晃征、野村浩郷：″情報検索と情報集約による情報取得システム″，情報処理学会研究報告 2007-NL-179 pp. 31-36, 2007

[2] 吉井隆明、石塚英弘：″行政情報に頻出する特徴語に関する基礎的調査″，第2回情報プロフェッショナルシンポジウム予稿集，2005

[3] 国立印刷局：″白書のあらまし″，http://www.npb.go.jp/ja/books/whitepaper/index.html

[4] 奈良先端科学技術大学院大学　松本研究室：″形態素解析システム 茶筌″，http://chasen.naist.jp/

[5] 東京大学　中川研究室、横浜国立大学　森研究室：″専門用語自動抽出システム″，http://www.forest.eis.ynu.ac.jp/Forest/ja/term-extraction.html

[6] Namazu Project：″全文検索システム Namazu″，http://www.namazu.org/

[7] 古川令：″pnamazu″，http://www01.tcp-ip.or.jp/~furukawa/pnamazu/