音声言語情報処理　16－11
自然言語処理　119－5
（1997．5．26）

情報検索における音声・言語処理

亀田弘之（東京工科大学）・藤崎博也（東京理科大学）

抄録

来るべき２１世紀には、世界的規模の情報ネットワーク環境にふさわしい高度情報検索システムが必要である。本稿ではまず、現在急速に普及しつつあるインターネットを対象として、インターネットにおける情報の特徴、インターネットの情報提供者と利用者の特徴、さらには、インターネットで提供されている基本機能の観点から、インターネットの現状について概観するとともに、それを踏まえて、高度情報検索システムに必要な諸機能について提案し、その実現方法についても言及した。また、高度情報検索システムに関する主要な大規模プロジェクトについての簡単な紹介を行った。

Spoken & written language processing in information retrieval

Hiroyuki KAMEDA (Tokyo Engineering University)
Hiroya FUJISAKI(Science University of Tokyo)

Abstract

A new advanced information retrieval system adequate for a world-wide information network environment is required to be implemented in the forthcoming 21th century. This paper describes, at first, status of the Internet from the viewpoints of characteristics of information on The Internet, information providers and users, and utilities for the Internet. Essential facilities for the new advanced information retrieval system and approch methods to implement the system are also presented with a brief introduction of a main big project.

1．はじめに

計算機技術の進歩はめざましく、近年、情報システムのネットワーク化とそれにともなう情報の分散的管理化が急速に進んでいる。とりわけ、世界的規模でのインターネットの整備にともない、インターネットを通じての情報公開が積極的に行われるようになっている。しかしながら、いかに多くの量の情報が収集・蓄積され公開されようとも、情報を必要とする人々に対して、必要な情報が常に迅速かつ的確に提供されなければ意味がない。つまり、情報は適切に利用されなければその価値を発揮することはできない。

このような観点から現在のインターネットによる情報公開の状況を詳細に分析すると、政府機関や学術研究者から一般の市民に至るまで、あらゆる環境の人々が情報を公開しており、多様かつ膨大な情報を利用することのできる状態になってはいるものの、それらの中から適切な情報のみを迅速かつ的確に収集する問題に関しては、ｙａｈｏｏ等のいわゆる検索エンジンと呼ばれるものはあるが、真に有効な情報検索を支援する手段は提供されているとはいい難い。

特に、現在の情報検索環境は、１９８０年代に主流となったような、情報提供センターがすべての情報を一括して収集・蓄積・管理・提供する形態から、インターネット等のネットワーク環境を前提として、情報が分散的・個別的に収集・蓄積・管理・提供されるようになってきているが、現在までの電子化された情報の検索手法・システムは、そのほとんどが電子化以前の形態を踏襲しており、電子化されたことの利点を必ずしも十分に享受していない。さらには、ネットワーク化のテンポが余りにも速いため、ネットワーク環境にふさわしい情報検索手法・システムはまだ模索状態にあり、現在まさにこれらは研究・開発し確立しなければならない技術である。

このような観点から、筆者らはインターネットやイントラネット等の情報ネットワークを念頭においた「情報検索における音声・言語処理に関する研究」に従事している。本稿では、そのうち、現在筆者らが構築中の高度情報検索システムについて報告する。具体的には、情報検索環境としてのインターネットの現状、従来の分類・検索方式とインターネットにおける情報検索方式について述べるとともに、高度情報検索システムの概要と関連プロジェクトについて述べる。

2．情報検索環境としてのインターネットの現状

以下ではまず、情報検索環境としてのインターネットの現状を概観する。

2-1．インターネットにおける情報の特徴

インターネットは、先にも述べたように、分散的・個別的な側面があり、その情報には以下のような特徴がある。

(1)分散性：検索すべき情報が、世界的規模で分散的に散在している。

(2)データ形式の統一性：情報のデータ形式が、原則としてHTML(Hyper Text Markup language)に統一されいる。

(3)用語の多様性：情報提供者が多様であるために、用語に統一性がない。同一のことが異なる単語（表記）で記述されることがある。

(4)文体の多様性：政府機関の公表する「形式的な文章表現」とともに、一般市民により提供される「会話風の文章表現」もあり、様々な文章表現が存在する。

(5)多言語性：国際的な情報源であるため、情報記述言語を特定のものに限定することができない。

(6)即時性：様々な情報が時々刻々公開あるいは更新・訂正されるとともに、場合によっては削除されることもある。

(7)質の不均一性：情報の信憑性や最新性（更新の時期・頻度）は、情報提供者に依

存しており、すべての情報を質の観点から均一視することができない。

2-2. インターネットの情報提供者と利用者の特徴

インターネットにおける情報の提供者・利用者は、先にも述べたように、検索の専門家（いわゆるサーチャーと呼ばれている人々）から一般市民（高齢者や学童等）にわたる様々な人々により構成されている。このために、インターネットの情報提供者と利用者には、以下のような特徴がある。

・情報提供者の特徴：

(a)政府機関や大学：原則として、統制された用語・表現により情報を提供している。

(b)一般市民：様々な用語・表現を自由に使用して情報を提供している。場合によっては、新たな表現を創造して自らの意思を表現している。

・情報利用者の特徴：

(c)検索の専門家：標準的なシソーラス等を利用して、効率のよい検索をする。

(d)一般市民：自分の概念体系に基づき、思い付いた用語で検索を試みる。

2-3. インターネットの提供する基本機能

以上、インターネットにおける情報、情報の提供者と利用者の側面を概観したが、このような状況下において、インターネットではどのような機能が通常提供されているかを概観する。

(1)**e-mail**：電子メール。電子化された文書の送受信が可能。文書には音声・画像を含めることも場合によっては可能。

(2)**FTP**：ファイル転送機能。情報検索手段として、anonymous FTPのサービス機能がある。大量のデータ送受信も可能。

(3)**rlogin**：リモートログイン機能。情報検索手段として、guest login サービス機能がある。

(4)**telnet**：リモートログインと同様の機能。

(5)**archie**:リソース所在検索機能。通常、FTP可能な情報格納・提供場所（サイト）に関する情報の検索が可能。

(6)**gopher**：情報検索用ブラウザ機能。現在は、次に述べるWWW(World-Wide Web)に取って変わられつつある。

(7)**WWW**：マルチメディア情報源に対する情報検索用計算機環境。WWWは、マルチメディア時代の中心的な情報検索機能として、現在最も注目されている。なお、WWWは、利用のためのインターフェースとして、ブラウザが、また、情報検索の利便を図るために検索エンジンが通常提供されている。

上記の諸機能は、大学等の研究機関では通常そのすべてを利用することができるが、商用プロバイダを介してインターネットを利用している一般のユーザの多くは、これらの内、(1)と(7)を主として利用している。なお、プロバイダによっては、telnetの利用環境も提供している場合もあり、また、FTPはWWW利用のためのソフトウェア（ブラウザ）の一機能として組み込まれて利用されることが多くなる傾向にある。

以上のように、インターネットで提供されている機能は、情報検索に直接利用することのできる機能もあるが、その機能はWWWを除けば、多くの場合素朴なものである。

3．従来の分類・検索方式とインターネットにおける情報検索方式[1]

従来の情報検索システムでは、大きく分けて2つの分類方式が利用されている。1つは、図書館での蔵書整理で従来から利用されている分類表方式（テーマの木を利用する方式）であり、他の1つは計算機を利用した各種データベースで利用されているキーワード方式（キーワードを利用する

方式）である。これらに対応して、検索方式にもテーマ入力によるものとキーワード入力によるものとがある。これらのいづれの方式も、有効に利用されているが、インターネット上での情報検索への適用を考えた場合には、以下のような長短がある。

分類表方式：

・長所：事前に分類整理されているために、キーワード方式の検索よりも一般に検索精度が高い。

・短所：検索精度・再現性を高めるためには、厖大な量の人的作業が必要なため、即時性に欠ける。

キーワード方式：

・長所：事前の分類が必要ないために、分散している種々の情報を広く、かつ、時間遅れなしに検索するのに向いているため、検索の再現性に優れてる。

・短所：異表記同義語や同表記異義語により検索精度が低下する。

現在、インターネット上で利用可能な検索エンジンは、これらのいずれかの機能を有している。しかしながら、先の2-1と2-2で述べた状況を考えあわせると、従来の検索エンジンは真に有用な検索ツールとしては機能的に不十分である。

筆者らは、これらの諸側面を分析・検討した結果にもとづき、インターネット上での高度検索システムには、以下の機能が必要であるとの結論に達した。。

(1)キー概念検索機能
(2)未知語処理機能
(3)検索意図推定機能
(4)知識獲得機能

以下、これらの機能を有する高度検索システムについて述べる。

4．高度情報検索システムの概要[2-4]

4-1.高度情報検索システムの諸機能

(1)キー概念検索機能

インターネットにおける情報検索のための検索エンジンは、先にも述べたように、分類表方式型とキーワード方式型とに大別されるが、インターネットにおける重要な特徴の１つである即時性を考えるならば、検索方式としては、キーワード方式の方が望ましい。実際、分類表方式型の検索エンジンの代表であるyahooでも、キーワードによる検索機能を補助機能として提供し、有効な検索を実現している。

しかしながら、高度情報化社会においては、検索の専門家（サーチャ）・高齢者・学童等の多岐にわたる人々が、自ら情報検索を行うようになること、さらには、情報提供自体も、このような多岐に渡る人々により散在的不定期的になされること等、を考え合わせるならば、検索システム設計者の予想の範囲を越えるキーワードによる検索や、必ずしも標準的ではない用語による情報提供がなされ得るために、すなわち、同表記異義語や異表記同義語が一般には存在し得るために、キーワード方式による検索では、検索の精度や再現性の低下が発生する。この問題を回避するための一方法として、キーワードが担っている意味に着目する検索方法、すなわち、キー概念検索が有効である[1]。なお、キー概念検索とは、キーワードの意味に着目する検索方法であり、具体的には、同義語の自動生成により検索を行う。

(2)未知語処理機能[4-8]

情報検索システムにとって未登録なキーワード（未知語）が入力として呈示されても、情報検索システムが、それを未知語と判断し、その意味を文脈等から推定することにより、先に述べたキー概念検索を行うことができる。この機能により、情報検索者は自由なキーワードにより検索することができるとともに、情報提供者も用語や表現を

特に統制することなく、自由に自分の意思を記述することができる。この機能は、未知語獲得システムを利用することにより実現する。

(3)検索意図推定機能

情報検索システムで最も重要な機能に、情報検索システムが検索者の検索要求を適切に理解し、かつ、要求に応じた検索結果を探し出し、さらには、検索要求に適った形式・順序で情報を呈示することがある。高度情報検索システムはこの機能を、音声による対話により実現することを目指しており、現在予備的な研究を推進している[4]。

(4)知識獲得機能[2]

より高度な検索を実現するためには、上記の機能の他に、情報源の所在に関する知識（検索者の要求を最も満足する情報がどこにあるのか）、情報源の提供する情報の質に関する知識（どのサイトの情報がより質がよいのか）等の、情報源に関する知識や、検索者に関する知識（例えば、検索者の検索履歴に関する知識、検索者の嗜好に関する知識等）を、検索システムが保持していることが望ましいが、これらの知識はインターネットの性質上、事前に網羅的に情報検索システムに付与することは不可能である。従って、これらの検索に有効に利用することのできる知識を、自動的に発見・収集・利用する機能、すなわち、知識獲得機能が必要である。この機能は、実際の検索処理を通じて、情報のリンクをたどった履歴情報や、あるいは、検索者がどのような場合に検索結果に満足したか、不満であったか等の補助情報を収集しつつ獲得する。この機能は、検索技能の向上に直接貢献する。

(5)多言語対応機能

インターネットは世界規模のネットワークであるために、そこで提供される情報は、必ずしも1つの言語で表記されておらず、また、1つの言語に限定した場合にでも、情報提供者の裁量・嗜好により、例えば、日本語文章中に外国語綴りで用語を表記する場合もあり得る。従って、このような状況に適切に対応し、インターネットが世界規模である利点を最大限に利用するためには、多言語対応機能は不可欠である。筆者らは、この機能を、先に述べた未知語獲得システムにおける1つのモジュールとなっている、異表記同義語処理に関する知識ベースを拡張することにより、実現する計画であり、予備的実験によりその基本的有効性を確認している。

4-2. 高度情報検索システムの動作

まず比較のために、インターネットの検索エンジンyahooにおいて、キーワード方式により検索する場合を考える。いま例えば、オーストラリアの首都ウィーンに関する情報を入手するために、キーワードとして「ウィーン」、「ウイン」、「Wien」、「Vienna」を個別に入力してみると、それぞれ１８件（内正解１６件）、２７８件（内正解１件）、４件（内正解４件）、６件（内正解２件）の情報が検索結果として得られる（平成９年５月７日現在）。このように、一般には、キーワードが異なれば検索結果が変わる。

しかしながら、高度情報検索システムでは、キーワード「ウィーン」のみが入力されても、キー概念検索機能により、キーワード「ウィーン」の意味を推定し、その意味を担い得る異表記同義表現「ウイン」、「Wien」、「Vienna」を自動生成し、検索を行うことにより、検索の再現性が向上する。また、キーワード「ウイン」が入力されても、これが「ウィーン」の異表記同義語である可能性があると推測し、さらには、多言語対応機能により、「Wien」、「Vienna」もキーワードとして検索することが必要に応じて可能となる。このように、高度検索システムは、より漏れの少ない（再現性の高い）検索を行うことができるとともに、さらには、意図推定機能や知識獲得機能によ

り、より効率よくノイズの少ない（精度の高い）適切な情報を検索・呈示することが可能となることが期待される。

5．今後の計画

以上では、知的情報検索システムについて述べたが、この知的情報検索システムは、日本学術振興会の援助により未来型プロジェクト「音声言語による人間－機械対話システムの研究」として、平成８年８月２３日から平成１３年３月３１日の５年間にわたり、音声入出力機能、音声対話による意図推定機能等の、さらに真に高度な機能を備え持つものとすべく、集中的・意欲的に研究・開発がなされている。検索対象は、情報が十分に蓄積整備されているとともに、実用性の側面から、学術文献を中心とする文献情報に設定されている。

なお、平成８年度の研究組織は、以下の通りであり、この分野におけるわが国の第一線の研究者から構成されており、高度情報検索システムの実現が強く期待されている。

・研究代表者：藤崎博也（東京理科大学）
・研究分担者：
古井貞熙（東京工業大学）
中川聖一（豊橋技術科学大学）
堂下修司（京都大学）
樽松明　（電気通信大学）
広瀬啓吉（東京大学）
白井克彦（早稲田大学）
板橋秀一（筑波大学）
大野澄雄（東京理科大学）
原田哲也（東京理科大学）
伊丹誠　（東京理科大学）
天野明雄（日立製作所）
中島邦男（三菱電機）
山崎泰弘（ＡＴＲ音声翻訳通信研究所）
渡辺隆夫（日本電気）
新田恒雄（東芝）

6．おわりに

来るべき高度情報化社会においては、情報の公開・共有化等が重要であり、この点において、インターネットをはじめとする情報ネットワークにおける高度情報検索システムが求められている。本稿では、インターネットの現状と、高度情報検索システムの概要について紹介した。併せて、関連の大規模プロジェクトについて簡単に紹介した。

＜＜参考文献＞＞

[1]亀田・藤崎：“テーマ・キー概念・キーワード間の階層構造を利用する新聞記事情報の分類・検索システム，”情報処理学会論文誌，Vol.28, No.11, pp.1103-1111(1987).

[2]藤崎・大野・伊東・阿部・佐久間・亀田：“知的エージェントを用いるインターネット上の情報検索システム，”電子情報通信学会総合大会講演論文集，D-8-12, pp.186(1997).

[3]藤崎・亀田・大野・阿部・伊東・佐久間：“キー概念の抽出と未知語の処理に基づく情報検索方式の高度化，”情報処理学会第54回全国大会講演論文集，4K-1, pp.23-24(1997).

[4]藤崎・亀田・田島・大野：“対話による高度情報検索システムの構築，”言語処理学会第３回年次大会発表論文集，pp.261-264(1997).

[5]亀田・藤崎：“高次辞書データベースのための語彙知識自動獲得システム，”「公開シンポジウム人文科学とデータベース」予稿集，pp.75-82(1995).

[6]横田・亀田・藤崎：“日本語の文法および未知の認知単位の自動獲得のための一方法，”言語処理学会論文誌，Vol.3, No.4, pp.115-128(1996).

[7]久保村・桜井・亀田：“未知語獲得アルゴリズムの評価，”電子情報通信学会技報，TL96-6, pp.21-30(1996).

[8]亀田：“未知語獲得システムとその評価，”言語処理学会第２回年次大会発表論文集，pp.277-280(1996).