

2004年2月5日
日本電信電話株式会社
株式会社NTT-X

ポータルサイト「goo」上で日本語自然文検索サービス 「Web Answers」の共同実験を開始

～話し言葉調の自然な表現による質問を即座に解析し
検索結果を的確に提示する新たな検索手段～

日本電信電話株式会社(以下NTT、本社：東京都千代田区、代表取締役社長：和田紀夫)と株式会社NTT-X(以下NTT-X、本社：東京都千代田区、代表取締役社長：中嶋孝夫)は、本日より、NTTサイバースペース研究所が開発した高度テキスト解析技術を用いた日本語自然文検索サービス「Web Answers」の商用環境での検証・評価のための共同実験を、NTT-Xが運営するポータルサイト「goo」(*1)上で開始します。

「Web Answers」は、例えば「2008年のオリンピックの開催地はどこ？」といった話し言葉調の自然な表現による質問を即座に解析し、

「goo」の検索結果から回答候補(上記の質問例では回答は「北京」)となりうる言葉や表現を高速に解析・抽出し、これらを含むWebページをより上位にランキングし、ユーザに提示します。(別紙参照)これにより、ユーザは知りたい情報をより効率よく取得することができ、日々増えつづけるインターネット上の情報資源を、今まで以上に有効活用することが可能となります。

この「Web Answers」は、昨年10月に「goo」上に開設された実験サイト「gooラボ」(*2)(<http://labs.goo.ne.jp/>)にて展開するもので、既に共同実験中の「InfLead」を活用した総覧型3Dウェブ検索実証実験に続く第2弾として、2004年2月5日～2004年3月31日(予定)までご利用いただけます。

1. 共同実験の背景と目的

「goo」を始めとしたWeb検索サービスは、Web上に存在する多種多様な情報から欲しい情報を取得する手段として、不可欠なものとなっています。しかし、「2008年のオリンピックの開催地はどこ？」といった質問への回答に相当する情報を得ようとすると、現状では「2008年」、「オリンピック」といったキーワードをユーザ自らが選んで入力し、検索結果として示されたこれらキーワードを含むWebページから、回答として適切な情報をユーザ自身が探していく必要がありました。

今回提供する「Web Answers」は、これまでのキーワード入力による検索を超えた自然な表現による高度な検索と検索結果の最適な提示を実現したもので、検索機能を重要な構成要素とするポータルサイトの価値を高めるための有用なツールと考えられます。両社は、一般のインターネットユーザを対象として「gooラボ」にて実験を行うことにより、同機能の技術検証とビジネス性の評価を行います。

2. 各社の役割

(1) NTT

日本語自然文検索サービスのコンセプト提案を行うとともに、本検索サービスに必要な高度テキスト解析機能を提供します。これにより、商用環境における同機能の技術的検証を行うとともに、精度向上のためのデータを得ることにより、今後の開発に反映させます。

(2) NTT-X

実験終了後の商用サービス導入検討に向け、「gooラボ」コーナーにおける総覧型3D Web 検索サービス実験に続く第二弾として、「goo」の利用度向上やサービス効果の検証を通してビジネス性の評価を行います。また、NTTグループのインターネット事業分野を代表するインターネットポータルとしての「goo」の先進性のアピールを行います。

3. 高度テキスト解析技術のポイント

高度テキスト解析技術は、以下の3つの技術で構成されています。特に、高速固有表現(*3)抽出技術は、Web ページを対象とした検索において、ユーザが求める情報を即座に提供するために不可欠なキー技術です。

また、これら技術に基づく回答の精度と技術の更なる向上に向け、質問文への回答内容の満足度をユーザが評価し、その評価をサービスに反映させます。

(1) 質問文のタイプ分類技術

「Web Answers」では、ユーザが、例えば「鉄腕アトムの日付はいつ?」などと入力すると、質問で求められている情報のタイプが誕生日の「日付」であることを即座に理解します。迅速な検索の実現にあたっては、質問文のタイプを迅速に把握・分類することが重要ですが、質問文が多岐にわたるため、人手でそれらを迅速に分類することは困難でした。本技術では、予め用意された質問文のサンプルとそのタイプを元に、質問文のタイプを判定するルールを自動的に生成することで効率的なタイプ判定を可能にします。加えて、単語の意味の分類(「誕生日」の意味属性は「日付」)を、大規模な日本語語彙の知識ベースである日本語語彙大系(*4)と照らし合わせ、「誕生日」と「生年月日」など表現

が異なる語も同じ意味として扱うことが可能になることで、自動学習の更なる効率化を進め、ユーザの質問の意味をより正しく理解することを可能としました。

(2) 高速固有表現抽出技術 (別紙2参照)

「Web Answers」では、入力された「鉄腕アトム」の誕生日はいつ? という質問を受け、まず「鉄腕アトム」、「誕生日」といったキーワードを選び出し、「goo」でキーワードの検索を行います。この時点で検索結果は入力されたキーワードに適合するようにランキングされていますが、求める情報が検索結果の上位にランクされたWebページに含まれていない場合もあります。

そこで、すでに質問の回答となるべき情報は「日付」に関するものであると絞り込んでいることを受け、検索結果で提示されている各Webページの概要文からユーザが求める情報のタイプ(この例では「日付」)に該当する人名や社名、日付などの固有表現の候補を即座に抽出します。Webページ検索は膨大なデータを対象とすることから、高速な処理が不可欠ですが、本技術は、固有表現を構成する単語の並び方のパターンを予め学習し、コンピュータ上にコンパクトな形式で記憶しておくことにより、単語列から高速に固有表現を抽出することを可能としました。

(3) Webページ再ランキング技術

「Web Answers」では、高速固有表現抽出技術により抽出された固有表現(この例では「日付」)について、「検索キーワードと日付に関する表現が近接して現れ、かつその数が多いWebページは、期待する回答を含んでいる可能性が高い」ものとして高い評価点を与え、再ランキングを行うことにより、ユーザが必要としている回答をより上位に表示することを可能としました。

4. 今後の展開

ブロードバンド時代におけるポータルサイトの付加価値をさらに高めるために、NTTでは、インターネット検索サービスの更なる高機能化のための開発を進めていきます。また、NTT-Xでは、本実験で得られたデータをもとに、「goo」でのサービス化も視野に入れビジネス性の検討を図る予定です。

<用語解説>

*1 goo:

NTT-Xが運営している日本を代表するインターネットポータルサイト。サービスの中核である「検索サービス」では、Webページの検索のみならず、画像、動画、音楽、辞書や地図、その他各種実用情報などの多彩かつ膨大なデータベースをもとに幅広い検索サービスを提供している。

*2 gooラボ:

NTT研究所で開発された新技術を活用し、先進的なインターネットサービスの可能性を社会に提示することを目的とする実験サイト。

*3固有表現:

「誰が・いつ・どこで」などの疑問語に対する答えとなるような名詞的表現のことで、具体的には、人名・地名・組織名などの固有名詞や、日付・金額などの数量表現を指す。例えば、「日本電信電話株式会社」という表現は、「日本」、「電信」、「電話」といった単語が組み合わさって構成されているが、これ全体で組織名を表す固有表現となる。

*4日本語語彙大系:

NTT研究所において、日英機械翻訳システムALT-J/E用の辞書として開発を進めてきた日本語意味辞書。40万語規模の日本語単語に対して、その意味的属性などの語彙知識情報を収録している。この日本語意味辞書は、人間向きに再編集され、『日本語語彙大系』として岩波書店より刊行されている。

- ・ [別紙1:日本語自然文検索サービスのイメージ](#)
- ・ [別紙2:高速固有表現抽出技術の概要](#)

〔問い合わせ先〕

日本電信電話株式会社
サイバーコミュニケーション総合研究所
広報担当 定方・山下
TEL : 046-859-2032
e-mail : ckoho@lab.ntt.co.jp

株式会社NTT-X
広報室 鈴木・田畑・栗山
TEL : 03-5224-5500
e-mail : pr@nttx.co.jp

NTT ニュースリリース 