Web上のQAデータの構造の抽出と利用

池上 敬明　竹内 孔一
岡山大学工学部情報工学科
{taka, koichi}@cl.it.okayama-u.ac.jp

1　はじめに

Web 上のデータや電子化されたデータから必要とする情報そのものをとりだす質問応答システムの研究が盛んに行われている．特に Web 上のデータは人手により最新の情報が入力されるため，その情報が必要とされる人に提示することができれば大変有効な情報源となる．

多くの質問応答システムでは質問は自然言語による入力を仮定している．このとき質問文のタイプによってその回答の性質がおおきく異なることから質問文のタイプを分類して回答となる文書を探す方法がとられている [3,5,6].

もっともよく扱われている質問文のタイプは名前や場所を聞くような固有表現にまつわる質問応答である [1,2,4]. しかしながら、実際の人間が求める質問の中にはより複雑な状況を聞く場合が多く単純ではない．例えば「なぜか」や「どうすればよいのか」といった理由や方法に関する質問も少なくない．こうした質問は知識を組み合わせなければならず、また的確な回答というのは現段階ではコンピュータ処理だけでは難しい．

一方, Web 上には様々な専門分野や技術分野，製品情報を提供する Web サイトなどで人手により作成された質問と回答のセット (以降，単に QA データと呼ぶ) [1]が掲載され、実際に利用されている．こうした回答は人間が作成しており具体的な個別の質問に対して的確に答えた回答になっている．

そこで本研究では質問応答システムの一つの具体化の方法として Web 上の QA データを利用することで質問の回答を作成するシステムの構築を目指す．以下では，この目標のために Web 上の QA データをどのように抽出できるのか，どの程度の質問と回答のセットが存在するのかについて調べた結果を明らかにして，QA データを利用した質問応答システムの可能性について述べる．

[1]よくある質問や FAQ など．

2　QA データ利用の予備的考察

QA データを抽出して質問応答システムを構築するためにまず，QA データが Web 上でどのように記述されているのか，利用という観点から考察する．これにより後の章で説明する QA データ抽出システムの際どのようにデータを取り出すべかきの方針を整理する．前提条件として Web 上のデータを検索できる検索エンジンが利用できるとする．以下では，QA データを抽出する必要性と QA データはどのような情報が記述されているかの 2 つについて考察する．

2.1　QA データを抽出する必要性

質問者が Web 上で自分の質問を検索する場合にそのままの質問文を検索エンジンにかけることができる (通常はキーワードに分解する)．これは QA データを想定してその回答を直接検索で見たい場合にはひとつの有効な検索手法である．定型的な質問であれば目的とする回答を即座に発見できる可能性が高いが次節でも述べるように多くの QA データは何のための QA データであるのか (テーマ) をページ上のどこかで提示しておき，そのテーマの中での具体的な質問に対して回答を記述する形式が多い．よってほとんどの場合，求める回答が Web 上にあったとしても検索エンジンでは直接探し出すことは難しい．そこで QA データをあらかじめ抽出してデータベース化しておき質問者の質問とマッチングを行う必要性が出てくる。

2.2　QA データにある情報

QA データは大きく分けて，掲示板などの質問と回答の集合文と情報提供者側が必要と思える形に整理しなおして構築されている場合がある．掲示板などの情報も大変有効であるが，必要とする回答がどの範囲までか掲示板の会話的言語データから抽出するのは非常に難しい．よって現段階では整理されている QA データを対象とする．

QA データのほとんどは FAQ であり，いくつかの質

問について回答を書くようになっている．そのため，質問はある分野の特定のことについて，特に専門的内容について作成されるため，質問文の中には分野を特定する言葉は入らず，専門的な内容のエッセンスだけが質問形式として入ることになる．例えば，

Q.「キャリア」とは何ですか?

これは「厚生労働省職業能力開発局キャリア形成支援室」が作成した「キャリア・コンサルティングＱ＆Ａ」の質問文の例である．この場合の「キャリア」とは仕事のキャリアであるが，こうした背景知識はタイトルである「キャリア・コンサルティングＱ＆Ａ」にこめられてしまい質問文そのものでは出現しない．

もし，携帯電話の分野で「「キャリア」とは何ですか?」という質問をすると回答は「携帯電話会社」ということになる．よってQAデータ情報を抽出する際にはこうした質問と回答の集合に対して背景知識が何であったのかという情報を付与しておく必要がある．

さらに，回答についてもWeb上の独特の表現が利用されていて回答はただしくても有効でないものがある．例えば，専門用語の羅列による説明でそれぞれにリンクが張ってあり，リンク先の説明を追わないと回答に成っていない場合である．また，こちらをごらん下さい．という参照先を回答にしている場合である．

例) Q　○○○の設定方法は?
A. こちらのページをご覧下さい．

こうした回答は有効でない場合が多いためQAデータを用いた質問応答システムを作成する際には候補としての優先度を下げる必要がでてくる．またQAデータとして抽出する際にはこうしたリンク情報も保存しておき2次引きできるようにする必要がある．

3　QAデータ抽出システムの設計

Web上のQAデータの抽出システムを設計してその実現方法について考察する．システムの構成は大きくわけて，1) QAのWebサイトを識別する，2)QAのWebサイトからQAデータを取り出す，の2つのステップが必要になる．それぞれのアプローチについて考察する．

3.1　QAのWebサイトを識別する

Web上で検索エンジンを利用できると仮定してQAデータのサイトを探し出す．前節で述べたように人手によるQAデータはまとまって記述されるので表現(言葉による)や体裁に特徴が出てくることが多い．そこで，まず表現についてQAデータを表す言葉はどのようなものがあり，それがどのくらいヒットするのか検索エンジンのgoogleを利用して調べてみた．

表1はQAデータを検索するためのキーワード，その検索ヒット数，さらに右の欄には有効率を示している．これは検索結果の最初の20件のリンク先を確かめてQAデータが存在した割合を示している．これは検索エンジンでヒットしてもQAデータではない場合があるため人手により調べた．有効率について説明する．

表1: QAデータ検索キーワード

検索キー	google ヒット件数	有効率 (上位20件)
FAQ	1,540,000	(11+7)/20
Q&A	803,000	(6+4)/20
QA	186,000	(2+2)/20
よくある質問	1,700,000	(11+7)/20
よくあるご質問	925,000	(7+11)/20
よくある御質問	429	(15+4)/20
よく寄せられる質問	29,200	(15+2)/20
一般的な質問	32,900	(18+0)/20

表1内の括弧の中の値は左側が直接質問のリストがページ内にあった場合で，右側がリンク先に質問リストがある場合である．これはgoogleはページの優先順位をリンクの信頼度から計算しているため，QAデータのトップページを検索するがそこでは質問を分類したリンク情報のみがあり質問と回答のセットはリンク先にあるという場合である．例えば

例) javaについてのFAQの場合
general, network, what's new, mediaの4つの分類

というリンクのあるページを検索してしまう．このため，QAデータのサイトを検索する処理としては正解であるが，のちのQAデータの抽出ではこうしたリンクをたどって質問と回答を抽出する機構が必要になる．

検索キーについて，FAQやQAといった英語の表記では検索ページの半分が英語のページになる．さらに，QAと名のつくもの(本やソフトウェアの名称)などが存在し検索結果として出力されるためノイズが多い．これに対して，日本語の「よくある質問」という検索キーは必要とするQAデータをもつサイトを見つけることができた．この理由は一つはこの表現が特にQAデータの存在を示す名称であることと，googleのランキン

グでシステムにおいてQAデータが高いスコアーを得たためと考えられる．

うまく検索できなかったページとしては「よくある質問をご参照ください」というリンク情報であったページ，また，直接質問をフォームで担当者に送るようになっているページに記載されていた．こうしたページはhtml形式を利用して排除できると考えられる．

こうした考察から「よくある質問」など日本語による検索キーワードが一つの手がかりになること，さらにhtmlの形状にしたがってさらにリンクをたどるシステムを考える必要があることが明らかになった．[2]

3.2 質問・回答のセット抽出

検索されたリンク構造をもつページから必要な質問と回答のセットを抽出する．ただし2.2節で説明したようにQAデータはその背景知識については質問そのものに記述されないことがほとんどである．そうした背景についてはページ内のどこかに 「○○ついてのよくある質問と回答」というように記述されていることが多く，背景を取り出す部分を言語的なパターンで記述する必要がある．

こうした背景，質問，回答がどのような構成になっているのかを知るために上記のキーワードで検索されたQAデータのパターンをまとめてみた．まず表2は背景知識の記述されている方法について20件分を調べてタイプ分けを行った．表2からhtmlのtitleタグの中に

表2: QAデータのhtmlの構成(背景)

背景のパターン	頻度
titleに「～の質問」	11/20
改めて説明する文がある	11/20
文字以外で記述	4/20
記述無し	1/20

何のQAデータであるか明示しているサイトが約5割あることがわかる．例えば「証券化支援事業Q＆A（よくある質問）」など名詞句で記述されており，こうしたtitleタグの中の表示がQAデータの背景を獲得することができる．

また「改めて説明する文」というのは検索された同じページ内に「ここでは○○のユーザからの質問および回答を掲載しています」と記述があったことを示す．こうした説明文を獲得することで背景情報を獲得することも可能である．

残りの「文字以外で記述」の場合は説明が同一ページにない場合や，タブの中に絵として記述されていて，人間がブラウザーで見れば確認できるがコンピュータで情報を得ることができない場合である．よってこれらは記述無しと同じ状況である．

次に，質問と回答がどのようにhtmlで構成されているか調べた結果を示す.「よくある質問」で検索されたWebページ20サイトについて分類した結果を表3に示す．表中の「同一ページ」上は同じページに質問と回答

表3: QAデータのhtmlの構成(質問と回答)

質問と回答	頻度
同一ページ	6/20
回答は別のページ	5/20
質問の分類	7/20

がhtmlタグで記述されていたWebサイトであり,「回答は別のページ」では質問は検索したページに記載されているが回答がリンク先に記述されていることを示している．両方の場合とも最初に検索されるページに質問が存在するのでそこからhtmlの形式を利用して回答とのセットを抽出するシステムを構築することは可能である．

表中の「質問の分類」とは前節で述べたリンクの先に質問があり，検索されたページでは質問者に対して質問のカテゴリーを問うページになっている．この場合，質問を得るまで次のリンクを繰り返し検索していく必要が生じてくる．しかし，ページのリンクをたどる際にどのリンクが質問に通じるリンクかを計算機に判断させるのは難しい．ページ内の全てのリンクをたどると上位のページに戻ることもあるため，リンクを持つ文章が質問文かどうかを識別するシステムが必要になる。表3はこういう場合が7/20程度あったことを示しておりこの数字は小さくない．

この結果から，検索されたページからhtml形式と言語情報(疑問形かどうかなど)を利用した質問と回答システムの構築が必要であることがわかる．

次の章ではこれらの予備的考察を踏まえてQAデータ抽出システムの具現化を行う．

4 QAデータの抽出システム

QAデータ抽出システムの構築法には大きく分けて次の2つの方法がある．1つはQAのデータをWeb上か

[2]キーワードの違いによる回答のタイプの異なりも観測できた．例えば「一般的な質問」の方がその製品の概要的でわかりやすい説明が多かった．検索キーワードの違いによるQAデータの収集の違いは今後の課題にしたい．

ら集めて巨大な質問応答データベースをあらかじめ作成しておき，質問に対してマッチした応答を提示するシステム．もう1つは，質問者の質問を得られてから，検索エンジンを利用して質問者の質問文と QA サイト検索を同時に行い Web 上の回答となる質問応答セットを取り出す方法である．本研究では，前者のタイプを選択する．理由は QA データの検索にリンク構造を利用して何度も検索して処理に時間がかかるためである．あらかじめ QA データを蓄積しておき質問応答システムとして利用することを仮定している．本研究で提案する Web 上

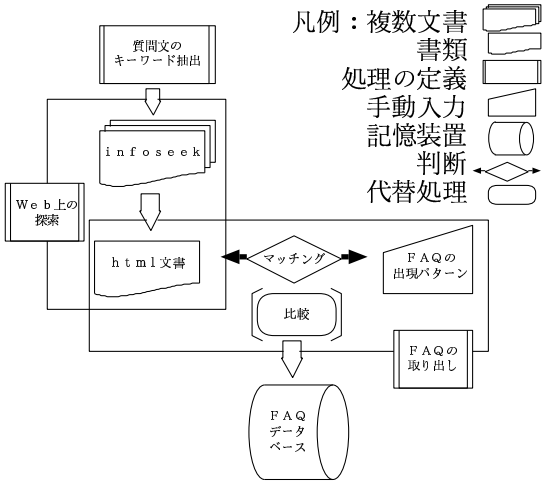

図 1: QA データ抽出システム

の QA データを抽出するシステムの全体像を図 1 に示す．検索エンジンは infoseek を通して google の結果を取り出すシステムを仮定している．図 1 の左上から QA データを取り出すための検索キーワード (「よくある質問」) などを投入する．得られた html ページから QA データの html パターン，ならびに，質問文のパターンなどを利用してリンク構造を利用しながら，背景知識，質問文，回答文のセットを集めてデータとして保存していく．

5 おわりに

実験結果に代えて本手法の見通しについて記述する．QA データの抽出がうまく成功したとして，これを利用した質問応答システムが成功するには次の 2 つの点がうまく機能する必要がある．

a. 質問者の質問と QA データの質問とのマッチング

b. 上記の QA データで集められる回答と質問者の質問の違い

まず **a** については 2.1 節で説明した通り，質問と質問のマッチングが必要になる．これは従来の質問応答システムでのマッチングに比べると同質の文を検索するためよりよい成果が期待できる．しかし，背景も含めて処理する必要があり，質問文の入力方法も含めて検討する必要がある．

また **b** については QA データでまとめられているものがどの程度の範囲かが問題となる．現段階で人手で確認するとほとんどが企業の商品情報に関する FAQ であり，一般的な概念についての QA データはあまり多く存在しない．さまざまな捉え方に基づく専門知識に関する QA データが散らばっており，質問者に有益な質問と回答のセットが Web 上にどの程度存在するのか，またそれが検索可能なのか，本システムを通して確認してみたい．この部分については見通しに関しては現段階では言及できないと考えている．

参考文献

[1] 遠藤哲哉, 福本淳一 (立命館大). 詳細化された質問タイプによる質問応答システム. 情報処理学会研究報告 2003-NL-159, pp. 25–30, 2004.

[2] 加藤恒昭 (東京大学), 福本淳一 (立命館大学), 桝井文人 (三重大学), 神門典子 (国立情報学研). 質問応答技術は情報アクセス対話を実現できるか. 情報処理学会研究報告 2004-NL-162, pp. 145–150, 2004.

[3] 黒橋禎夫, 清田陽司, 木戸冬子. 自動質問応答システム・ダイアログナビの現状と課題. 情報処理学会研究報告 2002-SLP-43, pp. 19–24, 2002.

[4] 市村由美, 齋藤佳美, 酒井哲也, 國分智晴, 小山誠 (東芝). 質問応答と，日本語固有表現抽出および固有表現体系の関係についての考察. 情報処理学会研究報告 2004-NL-161, pp. 17–24, 2004.

[5] 清田陽司, 黒橋禎夫, 木戸冬子. 大規模テキスト知識ベース基づく自動質問応答. 自然言語処理,Vol.10,No.4, pp. 145–175, 2003.

[6] 清田陽司 (京都大学), 黒橋禎夫 (東京大学), 木戸冬子 (マイクロソフト). 大規模テキスト知識ベースに基づく自動質問応答システム～ダイアログナビ～. 言語処理学会第 8 回年次大会発表論文集, pp. 271–274, 2002.