質問応答システムを用いた個人支援 Web 情報検索方式の研究

高橋 英史朗 †　　辻　秀一‡

現在、日本のインターネット利用世帯率は 73%、日本だけでも Web ページ数は 1 億ページを超える。そのような中で、問題視されているのが Web 上から見つけたい「情報」が見つけにくくなっているということである。そこで従来の研究で個人適応検索システムにより個人履歴情報から個人の嗜好に合った検索結果を求める提案がある。しかしこれは従来の検索と同様、「文書」(ファイル)単位で結果を提供するもので必ずしも要求に合った「文書」であるかわからない問題点を抱えている。そこで、要求を満たす「情報」だけをピンポイントで提供できる質問応答システムを利用した個人支援型の「気の利く」検索システムを提案する。本研究では、質問応答システムの質問理解部を中心に提案を行う。

Study of the Personalized Web Information Searching Method using Question-Answering Systems

Eishiro TAKAHASHI　Hidekazu TUJI

In recent years, the Web amount of information exceeds 100 million pages. It has been a problem under such circumstances should arrive at the "information" to search so hard. Then, the "considerate" search engine of the individual support type using the question answering system which can offer the "information" currently searched for at pinpoint is proposed.

１．はじめに

本研究は個人支援型質問応答システムの中の質問理解部に重点を置いている。質問理解部とはユーザ要求に対してユーザに満足した結果を提供するために重要な機能である。自然言語入力により入力されたユーザ要求の意図を理解し、かつ検索 KW を抽出する機能である。もし質問理解部での意図理解に失敗すると、その後の評価部が精密なものでもユーザ要求を満たすことができない。

† 東海大学大学院工学研究科
‡ 東海大学電子情報学部
† Graduate School of Engineering,Tokai University
‡ School of Information Technology and Electronics,Tokai University

２．従来の検索システム

2-1 「気が利かない」検索エンジン

現在 Web 上の膨大の情報から探し物をするときに欠かせないのが検索エンジンである。もっとも主流なのもは「Google」「LYCOS」「goo」などのロボット型検索エンジンである。ロボット型検索エンジンはロボットといわれるインターネット自動巡回プログラムを用い常に Web 上を徘徊してホームページ情報を収集しているため膨大の量の情報を常に維持している。そしてその膨大な情報量からユーザが必要な情報を求めるときにはキーワード(KW)入力にて検索が可能となる。しかし情報量が多いことで不必要なものが検索によりかかってしまうことがあり、更なる KW 入力を必要とされ、また検索結果の山と格闘する場面が多々ある。一般的にユーザが検索時に使用する KW は平均 2～3 単語である言われており、またそれ以上の KW は未知のものを検索するにあたり容易ではなく多く

の場合「検索結果 0 件」になることがある。
そこで、近年注目されてきている技術の 1 つに質問応答システムがある。以下に質問応答システムについて説明する。

2-2 質問応答システム

2-2-1 質問応答システムとは
質問応答システムとは、よく聞かれる質問とその回答をデータベース化し知識ベースとして蓄え、コンピュータが自動的に質問を受け付け、回答する情報検索システムの一種である。
2-2-2 従来技術との比較
以下に質問応答システム [2]を従来の検索システムと比較して説明をする。

『ユーザ要求』
「阪神タイガーズの監督の出身はどこですか？」

［従来の検索システム］
- 検索方法 ⇒ KW 入力（ブーリアン検索）
  [阪神タイガーズ and 監督 and 出身]
- 特徴
  1．ユーザ要求 ＝ 文中により多くの KW が含まれていること。
  2．検索結果が「文書」単位での提供。
- 問題点
  1．KW 集合だけではユーザ要求の意図を把握しきれてはいない。
  2．１の場合、検索結果の「文書」にユーザ要求を満たしている情報が無い場合もある。

［質問応答システム］
- 検索方法 ⇒ 自然言語入力
  ［阪神タイガーズの監督の出身はどこですか］
- 特徴
  1．自然言語入力でのユーザの問い合わせ（Query）
  2．ユーザ要求の意図を理解する。
  例）どこ ⇒ 場所情報
  3．１よりユーザの意図に合った情報だけを提供することが可能。
  4．検索結果は「文書」単位ではなくユーザ要求の範囲の「情報」のみ提供
  例）場所情報 ⇒ 岡山県倉敷市

2-2-3 問題点（既存システム）

(1)．自然言語での Query といっても Query の意図を理解するものではない。

① ユーザ要求 (Query)入力
「阪神タイガーズの監督の出身はどこですか？」
↓
② 形態素解析による単語抽出
[阪神タイガーズ][監督][出身]
↓
③ キ ーワード検索（AND 検索）

以上の場合、「出身」＝「場所」がユーザ要求になるが、３つの KW の AND 検索の場合、場所情報以外の情報も多数導かれる可能性がある。

(2)．ユーザ要求満足度の評価が動的ではない。

現在の利用例として IT 技術者支援など回答がすでに決まっているような場合に用いられているが、ユーザ要求満足度については常に静的なものでどのユーザ要求に対しても一定の答えが用意されているというシチュエーションに対しての利用が一般的になってきている。

現在このような自然言語検索の研究が進められ、各社から製品も提供されている。しかし既存の質問応答システムには以上の２つの問題点がある。
そこで、本研究では個人適応検索システムの中核として個人支援型質問応答システムにより以上の問題点を解消することを目的として、「気の利く」検索システムへの実現を目指す。

３．個人支援型質問応答システム

3-1 概要

3-1-1 全体構成
本研究で提案するシステム全体図を Fig.1 に示す。

本システムは自然言語による問い合わせの内容をタイプにより理解し検索エンジンから得た「文書」から「情報」を提供するシステムである。この際、ユーザの個人特徴情報を元に評価をするのでユーザごとに「気の利いた」情報を提供することを目指す。

本システムは「質問理解部」「検索部」「個人支援度評価部」の３つのモジュールによって構成される。以下に例題としてユーザ要求が『「ラスト・サムライ」は今どこで上映されているか？』の場合を用いてシステム全体の流れを説明する。

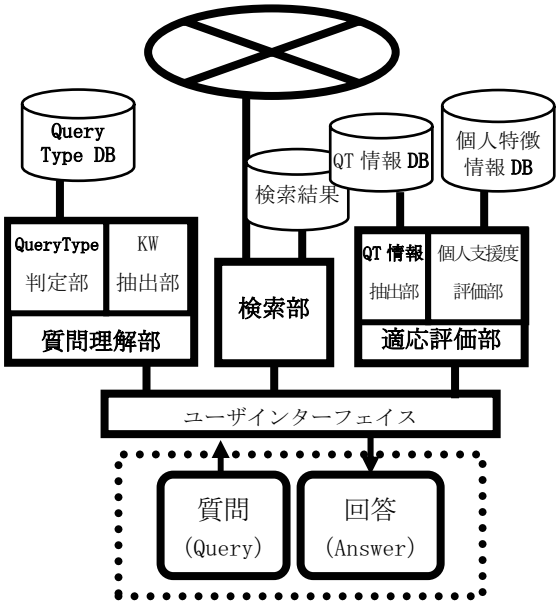

Fig.1 提案の検索処理

① ユーザ要求（Query）入力
『「ラスト・サムライ」は今どこで上映されているか？』
⬇
② 「質問理解部」
a. KW 抽出部にて KW 獲得
[ラスト・サムライ][上映]
b. Query タイプ判定部にて Q タイプ獲得
[どこで] ⇒ Where 型
⬇
③ 「検索部」
a. KW 抽出部で獲得した KW で Web 検索
b. 検索結果をキャッシュする
⬇
④ 「適応評価部」
a. ③で得た検索結果から評価する
b. 個人特徴情報を元に評価する
⬇
⑤ 回答（ Answer）出力
『海老名 TOHO CINEMAS』

3-1-2 有効利用シーン

本提案システムが最も有効に働く利用シーンは個人の好み・趣向・プロファイルが Query に反映される場合である。以下に有効利用シーン例を数例あげる。

- 映画情報 (ジャンル・俳優・監督...)
- 音楽情報 (ジャンル・歌手・年代...)
- グルメ情報 (ジャンル・食材・エリア...)

その場合に本提案システムはユーザごとに個人特徴情報を元に個人の趣向に近い情報かつ、Query を満たす情報を提供することができる。

以下の節で、各モジュールについて記す。

3-2 質問理解部

3-2-1 分かち書き

Query に含まれる KW を日本語分かち書きソフト「ChaSen」[5]により抽出する。日本語分かち書きソフトとして有名なものとして「KAKASI」と「ChaSen」がある。「KAKASI」は文書を分かち書きする機能に留まるが「ChaSen」は単語の文法を理解する形態素解析のソフトである。本システムでは後者の形態素解析に有効性があるということで「ChaSen」を利用することにした。

『「ラスト・サムライ」は今どこで上映されているか？』
⬇

「ラスト・サムライ」	名詞 ―一般
は	助詞―係助詞
今	名詞―副詞可能
どこ	名詞―代名詞―一般
で	助詞―格助詞―一般
上映	名詞―サ変接続
さ	動詞―自立
れ	動詞―接尾
て	助詞―接続助詞
いる	動詞―非自立
か	助詞―副助詞
？	記号―一般

以上の ChaSen での形態素解析の結果で名詞の単語を検索 KW とする。

3-2-2　QueryType

Query がどのような内容についての Query か認識する。大きく分けて２つの Query がある。１つは Query に対する応答(Answer)が単語である場合。単語とは例えば人名を問われた場合は人名、というように固有のものである。１つは Answer が文書である場合である。文書とは物事の原因や推移などの説明文のことである。前者の場合は Query タイプによって決まった形で提供することができるが、後者は説明文となるので Query タイプに応じて文書を構成する技術が必須とされてくるので難しい。以下に、Query タイプを記す。

・Answer が「単語」タイプ
　・Who 型 ： 人名を問う
　・When 型 ： 時を問う
　・Where 型 ： 場所を問う
　・Can 型 ： 可能・不可能を問う
　・There 型 ： 存在を問う
・Answer が「文書」タイプ
　・What 型 ： 内容を問う
　・How 型 ： 方法を問う
　・Why 型 ： 物事の原因を問う

以上の Query タイプはあらかじめ単語を用意しておく。Who 型なら「だれ・人」と 3-2-1 で分かち書きした単語が一致した場合 Query タイプを決定付ける。

以上の２つの処理によりユーザ要求の意図理解を容易にして従来の**問題点(1)**を解決することができる。

3-3 検索部

3-2-1 で分かち書きしたことで得た KW により既存の検索エンジンにて検索する。このときに使用する検索エンジンには情報量の多いもの、質問拡張機能がないもの、以上の点を考慮して本システムは「Google」を用いる。

「Google」の検索時の KW 入力変数が q なので検索式は以下のようになる。

例）KW = IC and search
http://www.google.co.jp/search?q=IC+search

検索結果の HTML 文書上位 10 件を評価対象文書とする。3-2-2 で得た Query タイプの情報を抽出し検索結果 DB へキャッシュする。

3-4 適応評価部

本部文は「QT 情報抽出部」「個人支援度評価部」の２部から成り立つ。

まず 3-2 の質問理解部の出力の QT と検索 KW 群を元に 3-3 の検索部で取得されたファイルから QT 情報を抽出し QT 情報 DB に格納する。またこの情報はユーザが利用後に個人特徴情報にフィードバックされる。この処理により次に個人特徴情報 DB と QT 情報 DB の QT の中の情報を照らし合わせ一致したものを Answer とする。もし該当するものがない場合は QT 情報 DB から Answer を提供する。

以上の処理により個人特徴情報が動的であるほどユーザ適合度の評価が動的なものとなりリアルタイムにユーザに情報を提供することができ従来の**問題点(2)**を解決することが出来る。

４．試作実験

4-1 試作システム環境

今回は質問理解部について試作を行った。ソフトウェア環境については以下の Fig2 に示す。形態素解析には ChaSen を用いた。ChaSen には品詞を組み分ける機能があるために今回用いることにした。よって QTDB の内容（Fig.3）は ChaSen の性能から決定した。また開発言語には様々な正規表現の利用可能な Perl を用いた。全てフリーソフトウェアを用いた構築とした。

項目	ソフトウェア
OS	Windows XP
開発言語	Perl 5.6.1
DBMS	MySQL 4.0.13 – win
Local Server	Apache 1.3.27 Win32
形態素解析ソフト	ChaSen 2.3.3

Fig2. 提案システムの試作ソフトウェア構成

id	QT	q_t
1	who	誰 だれ
2	where	どこ 何処 場所
3	when	いつ 時間 何時
4	what	何 なに

Fig.3　QueryType DB

4-2 試作システム概要

本試作システムは対話型のインターフェイスによりユーザの質問意図をより理解することを目指した（Fig.4）。対話型にすることで従来の KW 検索での KW を考える必要性がなくなると同時に素朴な疑問についてもそのまま人に問いかけるように質問をすることが可能となる。

Fig.4 提案システムインターフェイス

始めに質問したい内容をインターフェイスに自然言語入力により問いかける。すると、ChaSen による形態素解析により単語を品詞分けする。このとき、例えば映画のタイトルのような固有名詞を含めた質問がある場合にはそのフレーズを「」で囲むことをユーザとの約束とする。そのことで、「」で囲まれたフレーズに関しては最重要検索 KW として認識される。また従来ではフレーズが複数の単語の組み合わせの場合、個々の検索 KW として認識されていたが本試作システムでは「」で囲まれたフレーズを最重要 KW として認識する。そしてそれ以外は名詞のみを抽出する。

Fig.5 試作システム画面

ここで抽出された名詞と QueryTypeDB との参照にて QueryType を判定する。もし、ここで抽出された名詞の中に QueryTypeDB と一致する名詞がない場合はあいまいな質問とみなしシステムがユーザに質問を再度要求する。ここで抽出される検索 KW の数について今回は 1 ～ 4 個とした。結果は Fig.5 のようになった。

以下にこの試作システムの評価について記す。

５．評価

5-1 評価方法

被験者 10 名にそれぞれ 1 問づつ質問をしてもらいその質問の意図を正確に理解しているか被験者に判定してもらう。正確な場合には 1 点をもらうことができ、10 点満点で採点する。尚、被験者は 10 代～30 代の流行に敏感な人々である。

以下に質問内容を示す。

- Q0：今月上映している映画には誰が出ていますか？
- Q1：「ラスト・サムライ」は今どこで上映されていますか？
- Q2：「HEAD PORTER」の店の場所を教えてください
- Q3：豆腐が美味い和風の店は何処か知っていますか？
- Q4：ビナウォークまで何キロありま

すか？

- Q5：「丸井」の閉店時間は何時ですか？
- Q6：「牛角」のラストオーダーは何時ですか？
- Q7：春にオススメの映画は何？
- Q8：オススメの腕時計は何？
- Q9：貧血時にオススメの料理はなに？

5-2 評価結果と考察

今回の試作システムの評価結果については 10 問中 9 問正解で 90%の質問理解度であった（Fig.6）。質問の理解に失敗した Q4 については「何」という KW の広義に対応していなかったことで QueryTypeDB に準備できていなかったので被験者の質問意図を満たすことが出来なかった。しかしそれ以外については質問意図を正確に理解した。被験者の中であいまないな質問をしたものもいたがシステムが聞き返すことで解決できた。よって上記の質問内容は最終的な質問である。

Q	0	1	2	3	4	5	6	7	8	9
点	1	1	1	1	0	1	1	1	1	1

Fig.6 試作システム評価結果

６．課題

今回の提案での課題は Query タイプの文書を構成する技術である。これに関しては単語の共起情報や重みの高さなどから質の高い文書を構成する必要がある。また適応評価部にてより個人支援度の高い Answer の提供を実現するためには新しい技術を用いる必要がある。インターフェイスに関してはより個人の嗜好を引き出すような会話を行うことのできるものが必要となる。

７．結論

本研究ではより個人に適応した「情報」の提供をするため、Query の意図の理解と個人特徴情報を用いることで Answer の適合度の評価方法をより明確にする質問応答システムを提案し、質問理解部を試作した。今後は適応評価部での評価方法にリコメンド技術を用いて質の高い Answer の提供を目指す予定である。

参考文献

[1] 高橋英史朗：「個人履歴情報を用いた Web 情報検索方式の提案」，情報処理学会第 65 回全国大会，ユーザモデルを用いた検索-3，2003/3.

[2] 情報処理, Vol.44 No.6,情報化社会に役立つ情報検索の技術動向,pp.618-621,2003/6.

[3] 馬場肇：「改訂 Namazu システムの構築と活用」，ソフトバンク，2003/7.

[4] 北研二・津田和彦・獅々堀正幹：「情報検索アルゴリズム」，共立出版，2003/3.

[5] 松本裕治・北内啓・山下達雄・平野善隆・松田寛・高岡一馬・浅原正幸：形態素解析システム『茶筌』version 2.3.3 使用説明書,奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座,2003/8.

[6] 大沼･池野,沖電気工業：「HTML 文書を対象とした質問応答システムにおける回答抽出方法」，情報処理学会第 63 回全国大会, 2001/3.