統合情報アクセスシステム
―欲しい情報を欲しい形で―

1. 背景

現在、情報検索、質問応答、情報抽出、自動要約など、いくつもの情報アクセスの技術が研究されている。これらの技術は、ユーザの知りたいこと（情報要求）に対して、それぞれのシステムがなんらかの形式で「情報」を出力するという機能で一般化することができる。この機能を「情報アクセス」と呼ぶ。こうした情報アクセスの中でもっとも有名なものに、Google (www.google.com) や Yahoo! (www.yahoo.com) を中心とした情報検索システムがある。これらのシステムはユーザから数種類のキーワードを入力として受け取り、それらのキーワードが記載されている Web ページを、ユーザが必要としている情報としてインターネット上から検索し、それらのページの情報の一部をユーザに提示するものである。こうしたいわゆる検索エンジンは、基本的にキーワードの有無によって Web ページを検索するため、キーワードが含まれてはいるが、必要な情報を記載していないページもまた検索してくることが多い。その結果、検索結果が膨大になることも少なくない。ユーザは検索結果数によって、キーワードを追加・変更を行い再検索して、自分の必要な情報に到達することができる。

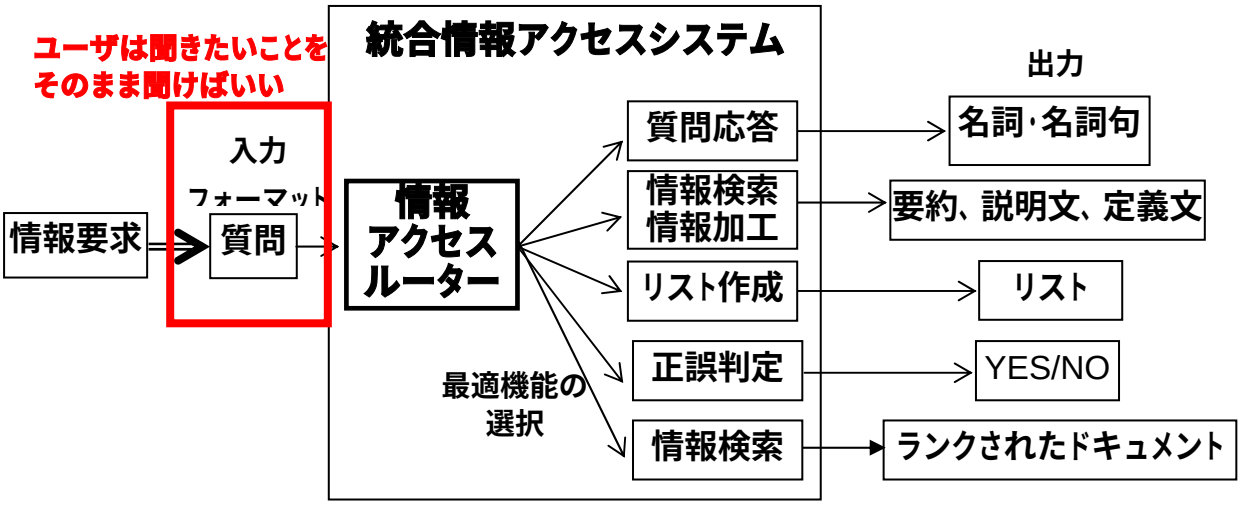
現在、情報アクセスを行う場合、先に挙げたそれぞれの機能ごとにシステムが別々に存在するため、ユーザは2つの技術が必要とされる。まず一つ目に、自分の情報要求に応じたシステムを明示的に選択しなければならない。そして、自分の情報要求を、選択した情報アクセスシステムが要求する検索クエリ、キーワード、質問文といった入力形式に変換することである。

2. 目的

こうした背景から、情報アクセス技術はそれぞれ単独では高い精度で情報をユーザに提示できたとしても、利用するユーザは自分の情報要求を入力するためのシステムの選択に精通する必要があり、更に選択した情報アクセスシステムに適した入力を自分の情報要求から生成する必要があるため、必ずしもユーザにとって使い勝手のよいシステムとはいえない。

そこで、情報アクセスに不慣れなユーザの利便性の向上や情報アクセスの新しいパラダイムの形成を目的として、どんなことを聞いても適切な形式で答えてくれるようなシステムの開発を行う。このシステムは、まるで頭に浮かんだ疑問を物知りの人にそのまま聞くと、その場で適切な形式で的確な情報を得られるといったイメージになる。このような統合情報アクセスシステムの実現には、ユーザの情報要求に対して、最適な出力形式を判定し、相当する処理により情報を提示する必要がある。これが本提案で実現したいシステムである。

3. 開発の内容



本プロジェクトでは図1のようにユーザが、欲しい形式で欲しい情報を得られることを目的とて、いくつかの情報アクセスシステムを統合したしステムを開発する。どのようなユーザの情報要求に対しても回答できることが望ましいが、新聞を読んでいるときに思いつく情報要求を対象とする。この情報要求を、ユーザが欲しいと思われる出力形式により分類した結果、18 種類の形式に対してアクセスシステムを構築する必要があることがわかった。これらの 18 種類を情報アクセスタイプと定義した。アクセスタイプは次の通りである。

FACTOID クラス：ORGANIZATION (組織名)、PERSON (人名)、LOCATION (地名)、
PRODUCT (製品名)、NUMBER (数値表現)、TIME (時間表現)、
OTHER_NAME (その他名前)
PASSAGE クラス：DEFINITION (定義)、DESCRIPTION (説明)、OPINION (意見)、
NEWS (ニュース)
クラスに属さないタイプ：LIST (リスト)、YES-NO (イエスノー)、TABLE (表形式)、
OTHER(その他)

そこで、本プロジェクトではまず、アクセスタイプを同定するための情報アクセスルータのプロトタイプを開発する。そして、この 18 種類のアクセスタイプのうち、FACTOID クラス、PASSAGE クラスの一部 (定義、説明)、YES-NO タイプ、LIST タイプに焦点を当てる。これは収集した質問を分析した結果、これらのタイプが 90%以上を占めたため、開発する優先順位が高いと判断したからである。次に示すアクセスタイプに対して開発を行った。

(1) 情報アクセスルータ

情報アクセスルータは入力された質問から、ユーザが要求している情報の形式、つまり質問のタイプを推定する。この推定は、質問で表現される特徴的な単語や文法を捉えることで行う。ルールの記述法などを簡略化することで、開発者にとっての利便性を高めた。また精度として、平均8 2%の同定精度 (Recall) のルータとなった。

(2) LIST タイプに対するアクセスシステム

LIST タイプの同定精度が他のアクセスタイプに比べて最も低いため、提案の時点でルーティングに成功していなかったトレーニングデータの質問を分析し、ルールの改善を行った。また、回答するアルゴリズムは FACTOID クラスに対するものと同じで、回答数によりタイプが FACTOID もしくは LIST に変化することが多いため、回答候補のスコアから上位 N 位を回答とするかの決定行うモジュールの開発を行ったが的確に回答数を導くことは容易ではなく、今後の課題とする。

(3) DESCRIPTION タイプに対するアクセスシステム

これまでの予備実験から、辞典や辞書の利用が有効であることが確認されていることから、まず複数の辞典を用いてアクセスシステムを構築した。辞書項目として記載されているものはカバーできることが確認された。

(4) DESCRIPTION タイプに対するアクセスシステム

説明を回答とする質問タイプで、予備実験からさらに数種類のタイプに分類されることを確認。その中で、技術的に開発が可能と思われる「理由・原因」、「方法」、「属性」に焦点を当て開発する必要がある。「属性」に関しては、現在属性と属性値を抽出する手法を研究している段階であり、実装には至っていない。「方法」に関しては、2004年第2期未踏ソフトウェア、梅村 PM 傘下の三原・馬場 PJ (以下、三原プロジェクト) が開発する「Web を用いたヘルプデスク指向の質問応答システム」と共同して、質問に対する本プロジェクト側でのアクセスルーティング結果が「方法」である場合、三原プロジェクトに処理をフォワードしている。「理由・原因」に関しては、パターンマッチングを用いて回答部分となる表現を抽出するモジュールを開発した。

(5) YES-NO タイプに対するアクセスシステム

予備実験から、表現により3種類の表現タイプ (存在、コプラ、一般動詞) に分類できることが確認され、構文的な言い換え、そして表現の言い換え手法を組み合わせて用いることで回答を得られることが分かっている。3種類の表現タイプのうち、存在に焦点を当てて開発した。構文的な言い換えはルールとして実現化した。表現の言い換えのための7種類のうち1種類を実現化する開発を行った。

4. 従来の技術 (または機能) との相違

一般的に用いられている情報アクセスシステムである検索エンジンでは、自分の言葉でクエリを作成することはできず、また殆どの質問応答システムでは回答が出力される質問のタイプは限定されている。これらの情報アクセスシステムに対して、開発を行った統合情報アクセスシステムは質問応答システムと同様の入力で、出力は名詞や名詞句だけではなく、文で回答する質問や文字列以外の情報媒体を要求する質問などに対しても解答

することができるため、幅広いユーザの情報要求に対して対応できるところが最大の特徴であり、利点である。

また、ユーザは基本的に頭に思い浮かんだ疑問や質問などを入力することでシステムを利用するため、Google のような、浮かんだ疑問をキーワードに変換する必要がないため、少ないキーワードで表現できない情報を捉えることができることも特徴である。

5．期待される効果

このシステムが完成することになれば、情報アクセスシステムのあり方が Google を代表とするドキュメント検索から大きく変革することになる。現在の Google の検索方法は、完全に過去のものとなる。2つのポイントは、

1）ユーザが情報要求をキーワードに変換することなく、情報要求をそのまま入力できるようになる

2）最適な形式で的確な情報を表示できるようになる

という点にある。つまり、ユーザは、システムにあわせて情報要求を変換することなく、そのままシステムに入力し、適切な形の出力を得られることになる。これは、まるで頭に浮かんだ疑問を物知りの人にそのまま聞くと、その場で適切な形式で的確な情報を得られるといったイメージになる。

6．普及（または活用）の見通し

コンピュータが一般家庭まで普及した現在、情報アクセス技術を必要としているユーザの多くは、それほどコンピュータを上手に操作できない人達であり、システムにとって適切な入力形式まで考えている人とはいえない。そのため、情報要求を的確にアクセスシステムの入力形式に変更できないことが多い。例えば情報検索において、何かトラブルに面したとき、それに関するキーワードを何度か変更して情報検索を用いても、望まれる記事やページは見つからないことも少なくない。そしてトラブルは解決しない。だからこそ、自分の言葉で話を聞いてくれる有識者、電話サポートが必要なのである。我々が提案する統合情報アクセスシステムは、前述の2つの問題をユーザに特別な訓練を必要としない方向から解決に近づけることのできるアプローチである。ユーザは自分自身の言葉を入力とすることができ、かつユーザは必要な情報アクセスシステムを選択することなく、求める情報に到達できる。このシステムの実現によって、一般の人々にとって情報アクセスのパラダイムは大きく変わり、情報というものに対する考え方が社会全体においても大きく変革するものと思われる。

7．開発者名　（所属）

開発者　　村上浩司　（ランゲージクラフト研究所）
共同開発者　関根聡　（ランゲージクラフト研究所）

（参考）開発者ＵＲＬ

http://apple.cs.nyu/edu/infotogo (要パスワード)　　現在一般公開を検討中