統合情報アクセスシステムへ向けたアクセスルータの実現

村上浩司　　関根聡
ニューヨーク大学

1　はじめに

現在，情報検索，質問応答，情報抽出，自動要約などの，ユーザに必要な情報を提供する技術の研究が数多く行われている．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何らかの形式で「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．

現在，情報アクセスを行う場合は図 1 で示されるように，それぞれの機能ごとにシステムが別々に存在する．基本的に，ユーザは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．また，ユーザは自分の情報要求を，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必要がある．しかしながらこうしたシステムの選択やクエリ作成は，情報アクセスに不慣れなユーザにとっては負担となるだけではなく，適切な情報を得ることを妨げる場合がある．

そこで，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，図 2 で示すような，ユーザの情報要求の種類に依存せずにその情報要求に対して適切な形式で答える，統合情報アクセスシステムを提案する．システムへの入力は，ユーザが情報要求を的確に表現できると考えられる自然言語での質問形式とした．統合情報アクセスシステムはユーザの質問から情報アクセスタイプを同定する必要がある．情報アクセスタイプは，ユーザがどんなフォーマットの出力が欲しいのかを示す．この情報アクセスタイプ同定を行う「情報アクセスルータ」は統合情報アクセスシステムにおいて最も重要な機能である．我々はこのアクセスルータをルールベースで構築し，9 人の被験者から収集した 3,400 程の質問を用いて評価した．その結果，情報アクセスルータが高いタイプ同定を高い精度で行うことができるとを確認した．

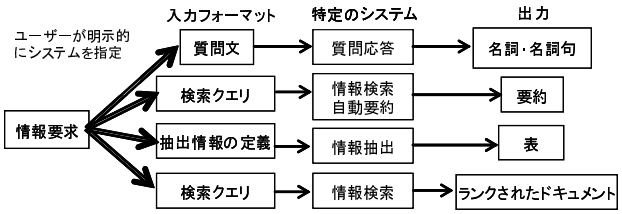

図 1: 現在のシステム

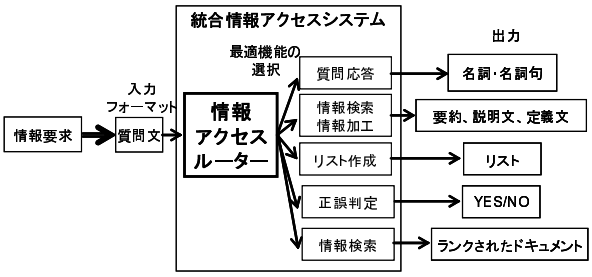

図 2: 提案するシステム

2　ユーザ入力

通常，質問応答や情報検索などの情報アクセスシステムを使うには，ユーザは情報要求をそれぞれのシステムに応じた入力形式に変換する必要がある．情報検索では，本研究で対象とする一般的なユーザが Web 検索エンジンに与えるクエリの長さは，平均して 2 単語程度である [3]．ユーザにとっては数単語のキーワードへの情報要求の変換は必ずしも容易ではなく，不慣れなユーザであるほど適切な情報を得られず，不必要な情報が返されることも少なくない．そのため入力となるクエリには，ユーザがどんな情報を要求するかを特定するため，より多くの情報が必要であると考えられる．

一方，質問応答には自然言語での質問文が入力として使われている．質問形式はユーザの情報要求を最も自然な形で表現でき，数単語のキーワードと比べて重要な情報の欠落が少ないこと，及びシステムを利用するユーザが特別な知識を必要としないことが挙げられる．そこで本研究では，統合情報アクセスシステムに対してユーザから行われる入力のフォーマットは，図 2 に示されるように質問応答同様，自然言語の質問文とする．

3　情報アクセスタイプ

情報要求は様々な「場」で存在するが，一般的な新聞を読んでいる場と法律に関する本を読んでいる場では，情報要求の種類が異なる．我々が本研究で適用す

表 1: 情報アクセスタイプと代表的な質問例

アクセスタイプ		質問例
大分類	小分類	
FACTOID	ORGANIZATION(組織名)	アメリカの携帯電話の最大手企業は何処ですか？
	PERSON(人名)	自衛隊が創設されたときの日本の総理大臣は誰ですか？
	LOCATION(地名，施設名)	横浜市で人気のある観光名所は何処ですか？
	PRODUCT(製品名)	日本高価な日本酒の名前は？
	TIME(時間表現)	固定電話が発明されたのはいつですか？
	NUM(数字表現)	トライアスロンの日本の競技人口はどれぐらいですか？
	OTHER_NAME(その他の名詞句)	週刊文春はどのようなジャンルの雑誌ですか？
PASSAGE	DEFINITION(定義。辞書，百科事典などの説明で 十分)	衆院構成労働委員会とは何ですか？
	DESCRIPTION(説明。定義だけでは不十分で，情報の加工が必要)	携帯電話の解約の仕方を教えてください。
	OPINION(主観的な意見)	議員年金について国会議員はどう考えていますか？
	NEWS(過去の記事)	千と千尋の神隠しについて書かれた新聞記事を見せてください。
MEDIA	PICTURE(絵，写真)	鳥インフルエンザにかかった鳥の写真を見せてください。
	AUDIO(音，音楽)	ポルトガルの国歌を聞かせてください。
	VIDEO(動画)	小泉首相の演説シーンを見せて
ALTERNATIVE(YES/NO 質問)		民主党は与党になったことはありますか？
TABLE(表形式)		スタジオジブリの作品とその公開年を教えて下さい。
LIST(名詞，名詞句のリスト形式)		SPEED のメンバーを教えて下さい。
OTHER(その他)		二酸化炭素の化学式は？

る「場」は，一般的なユーザが新聞記事を読んでいる場である．新聞記事には，その記事で伝えたい出来事に関して選択された情報が記されている．従って，その記事を読んだユーザは，様々な疑問を思い浮かべることがある．例えば,「台風 21 号，四国関西地方に大きな被害」というタイトルの記事を読んだ際に「台風 21 号はどこで発生したか」「大きな被害のあった市町村名のリストが欲しい」「台風 21 号についての最初の報道記事を見たい」などのた疑問 (情報要求) が思い浮かぶ．こうした情報要求はこれらの例にあるように，様々な情報アクセスタイプが含まれる．我々のアクセスタイプでは順に，LOCATION，LIST，NEWS になる．

Broder は，WEB 検索の情報要求はユーザの意図により，特定のトピックに関する WEB ページの獲得を要求する情報指向（Informational)，ある特定の WEB サイトへの到達を要求するナビーゲーション指向（Navigational) そしてオンラインショッピングなどのインタラクションを伴う WEB サイトへの到達を要求するトランザクション指向の 3 種類に分類されるとした [1]．しかしこの分類では，ユーザがどんな形式の情報を必要なのかを分類することはできない．我々はユーザが受け取る情報，つまり情報アクセスシステムの出力の種類からこの分類を考える．

我々は，ユーザが新聞記事を読むときに出てくる情報要求はどのような種類の出力が必要なのかを実際に質問を作成して調査した．ここでは，約 400 の質問を 20 の新聞記事を読むことで生成し，それぞれの質問が必要とする出力の種類で分類した．その結果，18 種類の情報アクセスタイプとなった．表 1 にそれぞれのアクセスタイプ及び代表的な質問例を示す．

FACTOID クラスは，望まれる出力が名詞もしくは名詞句である質問である．このクラスは多くの質問が含まれるため，関根の拡張固有表現階層 [5] を用いて意味的に再分類した．その結果，ORGANIZATION, PERSON, LOCATION, PRODUCT, TIME, NUMBER, OTHER_NAME の 7 つとなった．この階層では FACILITY は独立しているが，LOCATION と統合した．また，これら 7 種類以外のタイプは，すべて OTHER_NAME とした．

PASSAGE クラスは FACTOID クラスとは異なり，得られる情報が文形式であるアクセスタイプが属する．DESCRIPTION タイプは DEFINITION タイプと異なり，ある名詞句やその関係について，辞書や事典から得られる説明文をそのまま表示するだけでは不十分で，情報源から多段階的に情報を取捨選択，結合や加工をしなければ情報要求を満たせないものとする．具体的なユーザの要求としては，ある名詞句の属性や物事の違い，事典に記載されていない事柄や出来事の説明などがある．OPINION タイプは主観的な意見，NEWS タイプは過去の新聞記事をそれぞれ呈示することによりよりユーザの情報要求が満たされるものである．

MEDIA クラスは，テキスト以外の情報媒体の形式となる 3 種類のアクセスタイプが属する．AUDIO タイプは音声や音楽，PICTURE タイプは画像や絵，VIDEO タイプは動画となる．

またこれらのクラスには属さない，YES/NO を要求する ALTERNATIVE タイプ，名詞や名詞句の表形式となる TABLE タイプ，それ以外の OTHER タイプがある．TABLE タイプは LIST タイプとは異なり，2 種類以上の属性の属性値を出力する．また OTHER タイプは，数式や化学式，記号のようなこれまでのタイプではカバーできないものが含まれる．

4 情報アクセスルーター

本論文における主眼は，入力された質問文からユーザの求める情報アクセスタイプを判定し，そのタイプにより適切な処理を選択する機能をもつ情報アクセスルーターを実現することである．情報アクセスルーターでは，まず入力された質問文を形態素解析ツール JUMAN，および構文解析ツール KNP および NE タガー [5] を用いて解析する．次に，情報アクセスタイプの判定を行う．TREC の QA-Track[6] に参加するグループの多くは，固有表現階層を質問のタイプ同定に用いており [4, 2]，質問応答の精度向上に大きく貢献している．

```
RULE start
RULEID DOKO-ORGANIZATION
MATCH ^(どこ|何処)
TYPE ORGANIZATION
SCORE 100.0
RULE end

RULE start
RULEID DOKO-LOCATION
MATCH ^(どこ|何処)
TYPE LOCATION
SCORE 100.0
RULE end

RULE start
RULEID ITSU-TIMEX
MATCH ^いつ
TYPE TIME
SCORE 1000.0
RULE end

RULE start
RULEID DARE
MATCH ^(だれ|誰)
TYPE PERSON
SCORE 1000.0
RULE end

RULE start
RULEID TOWA-NANI-DESU
MATCH とは$
NEXTBUNSETSU
MATCH ^(何|なん)で(す|した|しょうか)
TYPE DEFINITION
PROIRITY 8
SCORE 1000.0
RULE end

RULE start
RULEID YOUYAKU-SURU
MATCH ^要約(し|する)
TYPE DESCRIPTION
SCORE 1000.0
RULE end

RULE start
RULEID <NUM>-NIN-WA-DARE
MATCH ([1234567890]+).+と?は$
NEXTBUNSETSU
MATCH ^(だれ|誰)
TYPE LIST_PERSON
SCORE 1000.0
RULE end

RULE start
RULEID XXX-MEI-WO-OSHIE-PRE
MATCH (名|名称)を$
NEXTBUNSETSU
MATCH ^(教|おし|答|にた|述|の|挙|あ)
TYPEOF PRE 1
PRIORITY 4
RULE end

RULE start
RULEID NANI-DESU
MATCH と?は$
NEXTBUNSETSU
MATCH ^(何|なん)で(す|した|しょう)か
TYPEOF HEAD 1
PRIORITY 8
RULE end
```

図 3: パターンルール例

我々はこの判定に質問パターンルールを適用する．パターンルールは，それぞれのアクセスタイプ特有の表現を正規表現化したものであり，このパターンマッチングにより，アクセスタイプを判定する．

ここで用いるパターンルールの例を 3 に示す．同定するタイプに 2 つ以上の候補がある場合，図中の左上 2 つのルールのように複数ルールを用意する．ルール中の MATCH の項目が質問文中のそれぞれの文節とマッチングされる．また，NEXTBUNSETSU は次の文節へマッチングが続くことを意味する．ルールによるタイプの同定方法には 2 種類ある．ひとつは図 3 の左側の 3 つのルールで示されるように，質問文中の特定の表現からアクセスタイプが直接判定できるパターンである．そしてもうひとつは図中右側 2 つのルールで示すように，パターンが質問文中の，ある名詞（以後，フォーカスワード）に着目して，そのフォーカスワードからアクセスタイプを判定するパターンである．

この後者のタイプである図中右側上のルールは，例えば「野田聖子衆議院議員の属する政党名を教えてください」という質問に対して適用される．まず始めに，ルールとのマッチングにより「政党」がフォーカスワードとして抽出される．フォーカスワードとなる名詞は，対応するアクセスタイプと対の形でセンターワード辞書に登録されている．このフォーカスワード辞書を用いて,「政党」と対応付けされている ORGANIZATION アクセスタイプが返される．ここで用いるフォーカスワード辞書は，独自に構築したもの，約 15,000 単語が登録されている．

また，質問パターンルールは他にも幾つかの属性を持つ．PRIORITY はルールを適用する優先度を，SCORE はアクセスタイプへの尤度をそれぞれ表す．

5 評価

5.1 情報要求の収集とルータの評価実験

新聞記事を読んだ際に思いつく情報要求を，統合情報アクセスシステム構築のための基礎データとする．我々は，9 人の日本人に 20 の新聞記事を読んでもらい，その過程で出てきた情報要求を質問文で書きとめてもらった．その結果，合計 3379 の情報要求が得られた．このとき被験者には，要求する情報アクセスタイプも記述してもらった．図 2 の中央に収集した質問の分布を示す．

3379 の質問データ中，2702 質問をトレーニングデータとし，残りの 677 質問をテストデータとして評価した．我々はトレーニングデータとなる質問文から各アクセスタイプ特有の表現を抽出し，それらを元に 289 種類の質問パターンルールを作成した．

5.2 実験結果

表 2 の右 2 列に，それぞれのアクセスタイプの分類結果を再現率および適合率で示す．再現率は平均で 82%と高い結果を示した．主要アクセスタイプ（10 以上の質問数）の中では，50%を下回るのは LIST タイプだけであった．

再現率と比較して，タイプの候補の複数出力を許すため適合率はそれほど高くない．例えば,「どこ」を含む質問は ORGANIZATION と LOCATION の両タイプに分類される．しかしながら平均 69.4%，主要タイプ中 50%を下回るのは 3 タイプのみに抑えられた．PRODUCT タイプは，フォーカスワード辞書中の単語からのみタイプが同定されるため，辞書に未登録単語が多かったことが原因であると考えられる．実験の結果から，FACTOID クラス，PASSAGE クラスは，おおよそ分類の精度が高いため，それぞれのアクセスタイプに対して特有の表現があると考えられる．

表 2: 情報アクセスタイプの分布と分類実験結果

アクセスタイプ		分布 (%)		分類実験の結果	
大分類	小分類			正解数 (再現率)	適合率
FACTOID	ORGANIZATION	19 (2.8%)	352 (52.0%)	15 (78.9%)	23.0%
	PERSON	38 (5.6%)		36 (94.7%)	71.4%
	LOCATION	65 (9.6%)		61 (93.8%)	73.0%
	PRODUCT	11 (1.6%)		9 (81.8%)	20.5%
	TIME	55 (8.1%)		54 (98.1%)	58.6%
	NUM	154 (22.7%)		151 (98.1%)	80.3%
	OTHER_NAME	10 (1.5%)		5 (50.0%)	41.6%
PASSAGE	DEFINITION	68 (10.0%)	181 (27%)	56 (82.3%)	100%
	DESCRIPTION	90 (13.3%)		76 (84.4%)	65.4%
	OPINION	16 (2.4%)		13 (81.2%)	76.4%
	NEWS	7 (1.0%)		6 (85.7%)	100%
MEDIA	PICTURE	12 (1.8%)	20 (3.0%)	9 (75.0%)	100%
	AUDIO	3 (0.4%)		2 (66.7%)	100%
	VIDEO	5 (0.7%)		5 (100%)	100%
ALTERNATIVE		69 (10.2%)		69 (100%)	25.3%
TABLE		10 (1.5%)		8 (80.0%)	88.8%
LIST		43 (6.4%)		18 (41.8%)	43.8%
OTHER		2 (0.3%)		0 (0%)	0%
合計		677 (100%)		593 (82.0%)	69.4%

6 考察

アクセスタイプの分類では主要アクセスタイプ中，LIST のみが低い精度（再現率:41.8%，適合率:47.1%）となった．そこで我々は LIST タイプのトレーニングデータを分析したところ，分類を誤った質問のうちの 50%においては，LIST タイプと FACTOID クラスを質問そのものから見分けることは人間にも難しいことが分かった．例えば,「北海道で生産される日本酒の銘柄を教えてください」という質問では，銘柄が 1 つなのか，もしくは 2 つ以上なのかによって，FACTOID もしくは LIST となる．こうした質問に対しては，アクセスタイプを判定するために情報リソース側にある知識を使う必要がある．

また，他の 40%については，ルールをより精密化ことで対応可能であることを確認した．

そして残りの 10%の質問は、例えば「NEDO の活動内容を教えてください」のように DESCRIPTION タイプと同じ表現が使われてる質問であった．こうした質問に対しては，質問そのものを分析するだけではタイプを決定することは難しいが，こうした質問に対しても情報リソース側の知識を使うことが有効であると考える．

7 まとめ

統合情報アクセスシステムの実現に向けて，入力となる質問文からユーザが要求している情報の種類である 18 種類の情報アクセスタイプを定義し，それらを同定するために情報アクセスルータのプロトタイプを作成して，評価した．その結果，全体で約 82.0%の再現率，69.4%の適合率が得られた．アクセスタイプの中で分類精度が高くなかった LIST タイプについて考察を行い，誤りの傾向を分析した．

今後は，より高い精度での分類を目指し，アクセスルータを改善していくだけではなく，それぞれのアクセスタイプに対する情報アクセスシステムを構築する予定である．

参考文献

[1] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[2] S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. Answer mining by cobining extractin techniques with abductive reasoning. In *The 12th TREC*, pages 46–53, 2003.

[3] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: A study and analysis of users on the web. *Information Processing and Management*, 36(2):207–227, 2000.

[4] B. Katz, J. Lin, D. Loreto, W. Hildebrandt, M. Bilotti, S. Felshin, A. Fernandes, and a. F. M. G. Marton. Integrating web-based and corpus-based techniques for questin answering. In *The 12th Text REtrieval Conference(TREC12)*, pages 472–480, 2003.

[5] S. Sekine and C. Nobata. Definition, dictionaries an tagger for extended named entity hierarchy. In *In proc. of Language Resources and Evaluation Conference(LREC2004)*, pages 1977–1980, 2004.

[6] E. M. Voorhees. Overview of the trec 2003 question answering track. In *The 12th Text REtrieval Conference(TREC12)*, pages 54–68, 2003.