情報ワーク支援システムのための情報抽出と構造化文書検索

西野 文人，伊吹 潤
富士通研究所
{Nishino,ibuki.jun}@jp.fujitsu.com

1　はじめに

情報検索は，疑問の解決や資料・原稿の作成といった時に利用されるが，それらの作業の中では，ただ一回の情報検索で完結するのではなく，検索結果を加工したり，クエリーを変えて再検索したりなどの作業が繰り返される．そこで，単に文書検索システムを提供するだけでなく，情報ワーク（目標を持った人が，情報源にアクセスして，その情報を活用して，目標を達成するまでの一連の作業）を支援するシステムが求められている．このような情報ワーク支援システムでは，情報の要求に対して精度の良い検索が行えることはもちろんのことだが，検索結果としてクエリーに関連する文書を回答するだけではなく，その中から必要な情報だけを抜き出して，情報を加工したり比較が容易にできるようにすること，そして必要な情報の要求を自然な形で表現できることが重要である．

我々は情報ワーク全体を支援するシステムに向けて情報要求範囲を限定した概況調査を目的としたシステムを試作した．このシステムでは文書に対してあらかじめ情報抽出　[1]　を行って，構造化文書 (XML) として格納しておく．ユーザが自然言語で質問（情報要求）をすると，システムは適切な XML 検索式を生成して XML 検索 [2] を行い，必要な情報 (XML 部分構造) のみを取り出し，さらに出力を再構成してユーザに適切な形式で出力する．自然言語による質問応答 (Q&A 検索) は，TREC をはじめ [3] 日本でも研究開発が進みつつある [4, 5] が，これらは分野を限定せずにユーザの自然言語での質問に対して，固有名詞や数値，名詞句などの答えを返す，あるいはそれらを含む文を返す．これらに対して我々の目指すものは，答が一意に決まるものではなく，概況を調査するためのものであり，質問の内容に応じて適切な情報を整理して提示しようというものである．

我々はまずは新聞という情報源から企業活動情報の調査に限定して，システムを実際に構築し，実験を行っている．本稿ではそのシステムの概要を示す．なお，質問文解析・検索式合成の詳細と評価は別稿 [6] にて示す．

2　システムの概要

我々の情報ワーク支援システムでは，以下のような方策をとった．

- 構造化文書 (XML) 検索をベースとした質問応答システムとする [7].
- 質問は，概況調査タイプのものを想定した．すなわち，答えが一つに決まるというものではなく，複数の候補があり得る．
- 質問文を解析した結果は，検索式に反映するとともに，出力形式の指示に反映する．

システムの全体構成は図　1のようになっている．以下では個々のコンポーネントの説明と設計にあたっての考え方を示す．

2.1　情報抽出・構造化文書検索

情報検索と情報抽出との結合方法は色々あり [8]，一般の平文に対する検索を行った後，検索結果に対して情報抽出を行ってマッチする情報を取り出すことも可能であるが，我々はあらかじめ情報抽出を行って構造化した文書に対して構造化文書検索を行う方針をとった．それは情報抽出は時間のかかる処

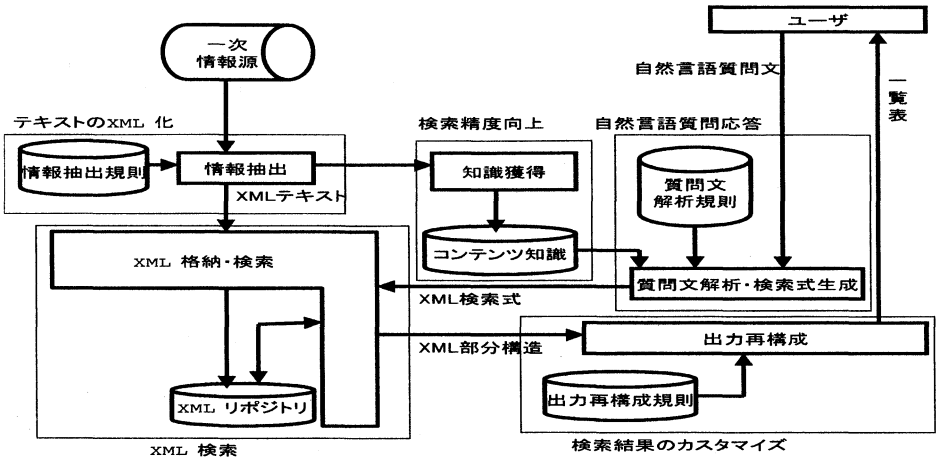


図 1: システム構成

理なので効率の問題もあるが，情報抽出結果をグローバルに見て，全体を分析して，その結果を検索に役立てることができる効用もあるからである．

2.2 自然言語質問応答

構造化文書 (XML) 検索では，単にキーワードを指定するだけではなく，キーワードの出現位置（役割）をタグで指定することで，検索結果の絞り込みができる．しかし，エンドユーザに直接検索式を記述させるのは，構文的な問題もあるが，文書がどのように構造化されている（タグ付けされている）かを意識する必要もあるという点からも得策ではない．そこで自然言語による質問文を受け付け，システムがそれを解釈して検索式を生成するようにした．質問文解析・検索式合成は以下のステップから成り立っている．

1. 質問文パターンの認定

   質問文パターンと照合して，質問文パターンを認定する．

2. 質問詳細の解析

   質問文の詳細部を解析して，条件や対象の属性等を明らかにする．

3. 質問拡張の処理

   各属性に対して，あらかじめ抽出してあるコンテンツ知識（表記のゆれ，略語，同義語等）に基づいて質問拡張を行う．

4. XML タグの同定

   ユーザが質問の中で与えた属性名は文書中のタグと一致しているとは限らない．文書中のどのタグに相当するのか同定する．

5. XML 検索条件式の生成

   XML 検索の条件式を生成する．

6. XML 検索回答式の生成

   検索した結果の XML 文書の中でどの部分を取り出すかを指定する．取り出すものは回答に関する部分のみならず，回答や質問に関連する付随情報も取り出す．

2.3 検索戦略

検索では以下のような戦略をとった．

- システムの対象とする範囲外の質問の場合はその旨を通知する．
- 条件が緩い質問に対しては，ケース別に検索を行う．例えば，「富士通の活動について知りたい」に対しては，開発事象，販売事象，合併・

提携事象などを順に検索し，それぞれ提示する．

- 回答がうまくえられなかった場合には，条件を順に緩くすることで，なるべく多くの質問に答えるようにする（最終的には通常のキーワード検索レベルでの対応となる）．

2.4 知識獲得

検索の高精度化を図るために，出現する情報を分析して，その結果を保持する．ここでは，文書群の中にある，表記のゆれや，同義語対を見つけ出す．この処理は，情報抽出によって付与されたタグに基づいて，同一カテゴリの中での処理にとどめることで確実性を増している．

2.5 出力の再構成

情報の重要度やユーザの関心に応じて提示方法をカスタマイズする．すなわち，質問の答のみあるいは答を含む文を返すのではなく，関連情報を含めて一覧表の形に整理してユーザに提示する．構造化文書検索の結果は，文書の部分構造の集合であるが，出力再構成規則にしたがって一覧表の形にまとめる．例えば，「カメラを販売している会社について知りたい」との質問文に対して会社名を答えるだけでは不十分である．その会社がどのような会社なのか，販売しているカメラはどのような種類のものかという情報も同時に提示することが望まれる．このような付加的情報も同時に提示することによって，多義の場合や検索誤りの場合など，システムが提示した結果が自分の意図したものかどうか判断できる．

3 実際例

我々は企業活動情報に関する質問応答システムを構築した．対象は日本経済新聞 1990 ～ 1997 年の 8 年分で約 140 万文書である．情報抽出はすべての記事に対して適用され，企業活動に関する文に対して組織情報（業種，組織名，所在地，代表者名など），活動事象に関する情報（新製品販売，新事業，新技術，合併など），個々の活動に関する情報（新製品の種類，名称など）を認定して階層的なタグ付けを行っている．

例えば，「冷凍のすしを開発したメーカは？」という質問を与えたとする．質問文を解析し，活動情報として「開発」，要求が組織情報であることを認定する．そして，製品クラスとして「すし」，製品補足情報として「冷凍」であることを解析する．また，自動的に獲得した同義語辞書から「すし」が「ずし」と濁る可能性のあることを獲得し，検索式としては，

```
select 開発情報 where <開発情報  /*> <開発内容> 冷凍 &(すし | ずし)</ 開発内容>
```

を得る．また，出力する情報は，質問項目が組織名なので，まずは組織名，続いて，組織に付属する情報（所在地，代表者），また注目項目として開発内容を出力することを指示する．属性名は文書内のタグ名を直接表示するのではなく，ユーザの質問文に合わせている．図 2は，検索結果である．

4 おわりに

我々は情報ワーク支援の一部として，概況の調査を行うためのシステムを試作した．今後解決すべき課題としては以下のものがある．

- 質問受け付け範囲の明確化

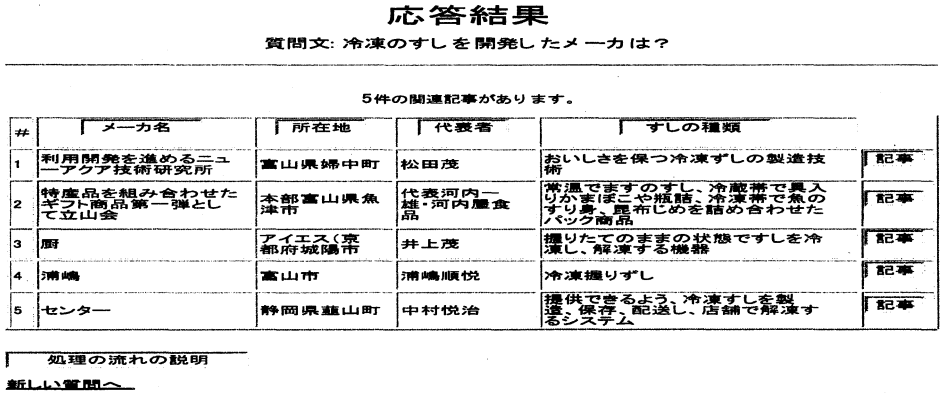
  ユーザの質問文入力に対してユーザの意図した結果が得られない最大の原因は，システム側が想定していない範囲の質問文が入力されることである．システムが対応できる範囲はどこまでなのかを明示すること，システムが対応する範囲をより適切なものにすることが重要である．

- 質問背景の推測

  ユーザのあいまいな質問文に対して，あるいはユーザの望むぴったりした情報が得られなかった時，より適切な情報提示を行うためには，ユーザの質問背景を知ることが必要である．一連の情報要求の中からそれを推測したり，あるいはユーザに問い合わせるようなことが必要である．

- 条件緩和の戦略

  「松本市にあるゲームのソフトを開発している会社について知りたい」というような質問文に対して，マッチする検索結果がなかった時に，



応答結果

質問文: 冷凍のすしを開発したメーカは？

5件の関連記事があります。

#	メーカ名	所在地	代表者	すしの種類	
1	利用開発を進めるニューアクア技術研究所	富山県婦中町	松田茂	おいしさを保つ冷凍ずしの製造技術	記事
2	特産品を組み合わせたギフト商品第一弾として立山会	本部富山県魚津市	代表河内一雄・河内屋食品	常温でますのすし、冷蔵帯で具入りかまぼこや瓶詰、冷凍帯で魚のすり身、昆布じめを詰め合わせたパック商品	記事
3	厨	アイエス(京都府城陽市	井上茂	握りたてのままの状態ですしを冷凍し、解凍する機器	記事
4	浦嶋	富山市	浦嶋順悦	冷凍握りずし	記事
5	センター	静岡県富士山町	中村悦治	提供できるよう、冷凍すしを製造、保存、配送し、店舗で解凍するシステム	記事

処理の流れの説明

新しい質問へ

図 2: 検索結果例

条件を緩和してそれに類似する結果を提示するようにしているが，どのように条件を緩和するのがよいかは大きな問題である．

- 情報の統合

  現在は一つの事象記述から一つの回答を返しているが，複数の情報源からの関連情報を統合して提示することが必要になってくる．

- テストコレクションの充実

  音声と結びつけないならば，自然言語文として入力させなくても，典型質問パターンの穴埋め式でも良いかもしれない．ではその時，どのような質問パターンを用意すべきか．典型質問パターンを探る上でも，システムの評価をする上でも，質問例を充実させる必要がある．

- 次の作業へのリンク

  情報ワークを支援するという観点から，検索の後の作業を支援することが必要である．検索結果の詳細を調査する（例えば，企業情報をもっと詳しく調べるなど），結果をまとめる，傾向を分析するなどがある．具体的作業項目の洗い出し，他情報源へのリンクが課題である．

- 実用に向けて

  今回は情報源として新聞データを利用したが，本来の情報ワークでは種々のデータベースの結合，そして Web 情報の利用が不可欠である．また，このシステムでは質問文を自然言語の文で入力するようになっているが，実用的にはキーボードからの自然言語入力は面倒であり，音声認識との結合も必要である．

謝辞　本研究では，日本経済新聞　CD-ROM　版 (1990 ～ 1997) を利用させていただいております．関係者の方々に感謝いたします．

参考文献

[1] 西野文人，落谷亮，木田敦子，乾裕子，桑畑和佳子，橋本三奈子：トップダウンなパターン解析に基づく情報抽出，情処研報，NL124-13, pp. 95–102 (1998).

[2] 井形伸之，難波功：大規模な構造化文書データベースにおけるインデクシングと検索の手法，情処研報，FI57-2, pp. 9–16 (2000).

[3] Voorhees, E. M.: The TREC-8 Question Answering Track Report, in *The Eighth Text REtrieval Conference(TREC 8)*, pp. 77–82 (2000).

[4] 村田真樹，内山将夫，井佐原均：類似度に基づく推論を用いた質問応答システム，情処研報，NL135-24, pp. 181–188 (2000).

[5] 佐々木裕，磯崎秀樹，平博順，廣田啓一，賀沢秀人，平尾努，中島浩之，加藤恒昭：質問応答システムの比較と評価，信学技報，NLC 2000-24, pp. 17–24 (2000).

[6] 伊吹潤，西野文人：質問文からの検索条件と提示情報の決定，言語処理学会第 6 回年次大会 (2001).

[7] 難波功，井形伸之，小櫻文彦，山根康男：大規模 XML 文書の検索と格納技術の開発，情処研報，DD27-3 (2001).

[8] Smeaton, A. F.: Information Retrieval: Still Butting Heads with Natural Language Processing?, in Pazienza, M. T. ed., *Information Extraction*, pp. 115–138, Springer (1997).