解　説

自然言語処理技術の応用

3. 情報検索における自然言語処理†

藤澤浩道††　絹川博之†††

1. はじめに

「情報検索」というと，従来は科学技術文献などの文献データベースをサーチャと呼ばれる専門家に検索を委託する，というのが典型的なイメージであった．近年は情報機器の普及にともない，一般の文書，資料，伝票，電子メールなど，さまざまな情報が電子化されつつある．このため，検索したい情報の範囲は広がり，情報検索はわれわれエンドユーザにとっても身近な課題になりつつある．また，増え続ける電子化情報により，その必要性は切迫したものとなっている．そこに新しいシステム概念と新しい技術のニーズが存在する．

ワープロやパソコンで作成した電子化文書はいよいよ机の引き出しから溢れ始めている．電子メールによる通信も欠かせない存在になり，メールボックスの中は何本もの文脈のメールが交錯して格納されている．電子メール経由で配送されるニュースの量も増えている．これに加え，コンピュータの CD-ROM 電子マニュアルや電子化特許公報の CD-ROM サービスは，フルテキストサーチ（全文検索）への期待感を大いに高めている．

今や文書は電子的に作ることが常識であり，広義の文書処理は文書作成の効率向上にとどまらず，組織全体の知的生産性を向上させるものと目されている．最近着目されているグループウェアは，文書情報の蓄積と流通をうまく制御して，組織の知識や情報を共有化する手段として期待されている[1]．

これらの期待に応える重要な技術の一つは，テキスト情報を扱うところの自然言語処理である．本稿では広義の情報検索を上記のような視点からとらえ，最近のニーズと技術の動向について自然言語処理の役割に触れながら論述する．

2. 情報検索の新しい課題

大きな流れの中でとらえると，情報検索はその利用目的や利用形態，あるいはシステム形態を大きく変化させている．多少単純化した見方であるが，**表-1**に示すごとく，扱う情報のマルチメディア化，利用者の一般化，検索システムの部品化などが起きている．その背景には，光ディスクなど記憶装置の高密度化・大容量化や，処理装置の高速化・小型化がある．

（1） 光ディスクを用いた文書ファイリング

光ディスクの出現は，文書をイメージデータとして蓄積する電子ファイリングシステムを生んだ．システムの規模は小さいながらも情報検索システムの各要素を保持している．従来は専門のサーチャやデータベース管理者が行っていた検索や，文書登録におけるキーワード付けを，オフィスの一般ユーザが自ら行う必要性が出てきた．この問題を解決する方法として，「知的ファイリング」や「知的検索」などが提案されている[2)~4)]．キーワードの異表記や同義語の処理に始まり，文脈に応じて検索条件の解釈を変えられる「あいまい検索」や，概念を扱う自然言語処理技術が求められている．

（2） フルテキストサーチと文書管理

文書の本体，すなわち1次情報を保持するデータベースを全文データベース，あるいはフルテキスト・データベースという[5]．これには上記のようにイメージ形式で全文を保持する場合も含まれるが，少なくともテキスト部分は文字コードで保

† Natural Language Processing in the Field of Information Retrieval by Hiromichi FUJISAWA (Software Development Center, Hitachi, Ltd.) and Hiroshi KINUKAWA (Systems Development Laboratory, Hitachi. Ltd.).

†† (株)日立製作所ソフトウェア開発本部
††† (株)日立製作所システム開発研究所関西システムラボラトリ

表-1　情報検索が利用される環境の変化

比較項目		従来の情報検索	新しい情報検索
検索対象情報		2次情報・所在情報 (書誌的事項／テキスト)	1次情報 (全文データ／マルチメディア)
検索操作者		検索の専門家（サーチャ）に依頼	ユーザが自ら操作
情報登録者		管理者	●ユーザ自身 ●外部情報源（電子メールなど） ●パッケージメディア(CD-ROMなど)
システム管理者		あ　り	な　し
システム形態	物理的形態	オンラインシステム	●クライアント／サーバ ●スタンドアロン
	利用形態	独立した専用システム	情報処理環境の一部

持するほうが検索性能やデータ加工性が優れており，フルテキストというと文字コードで表される情報を指すことが多い．この種のデータベースでは，全文を探索範囲とするところのフルテキストサーチ（全文検索）が着目されている．米国においては，英語文書がすべての単語への索引付けが容易なこともあり，商用の判例データベースなどでかなり以前からフルテキストサーチがサービスされている．

ワープロ・パソコンで作成する電子化文書，電子メール，あるいはCD-ROMによる電子化文書などの氾濫により，オフィス情報環境における文書蓄積・管理のニーズも増大している．

いかに文書を整理・保管するか，いかに経緯を把握しながら電子メールでやり取りするか，いかに膨大な量のニュースから関心のあるものだけを選択するかが現在課題となっている．

こうした背景をもとに文書処理市場では文書管理機能が重視され，最近，フルテキストサーチ機能を有する文書管理ソフトウェアの製品化が米国および国内で相次いでいる[6),7)]．

（3）情報フィルタリング

ネットワークで繋がれたパソコンやワークステーションが一人一台になる環境が常識的になるにつれて，組織全体の知的生産性向上が期待されている．とくに，電子化文書の蓄積・管理と流通をうまく制御することを狙う情報共有型のグループウェアは，組織に散らばっている知識や情報を，組織を横断して共有可能とする情報環境として期待されている．

このようなシステムでは，文書管理・検索の問題のほかに，外部から着信する文書やシステム内で作成される文書をいかに適切に必要とする人に送付するかが問題となる．これを行う情報処理は情報フィルタリングと呼ばれている．これを実現する技術は，情報分類（Categorization）とトピックス抽出（Topics identification）である[8),9)]．技術的には情報検索技術との共通点が多い[10)]．

フィルタリングされて行き先が同定された情報は電子メールで配送されるが，その流れの制御はワークフロー制御と呼ばれ，グループウェアを構成する一つの重要な要素である．

グループウェア・システムでは，情報検索はシステム全体の中の一要素である．従来，情報検索システムは一つの専用システムを構成していたが，これからはオフィス情報環境の一部を構成するサブシステムとなる．

（4）検索支援ユーザインタフェース

情報検索の操作をユーザ自身が行う必要が出てきたことによって，本来の検索の難しさが再認識されている．検索条件の定式化を支援する知的な機能や，あいまいな検索条件から探索したり，自然言語で表現した条件から意味的な検索を行う知的検索が研究されている[11)~14)]．

3. フルテキストサーチ

自動インデクシングを用いたインデクス型検索方式は大規模なデータベースに対しても高速なアクセスを可能とする[15)]．しかし，多様なオフィス情報環境における情報検索を実現しようとすると，シソーラスの構築と保守にかかるコストや，統制された索引語でしか検索できないという制限が問題となる．

フルテキストサーチは本文に現れるすべての単語を探索対象とすることにより，これらの問題を解決する方式である．

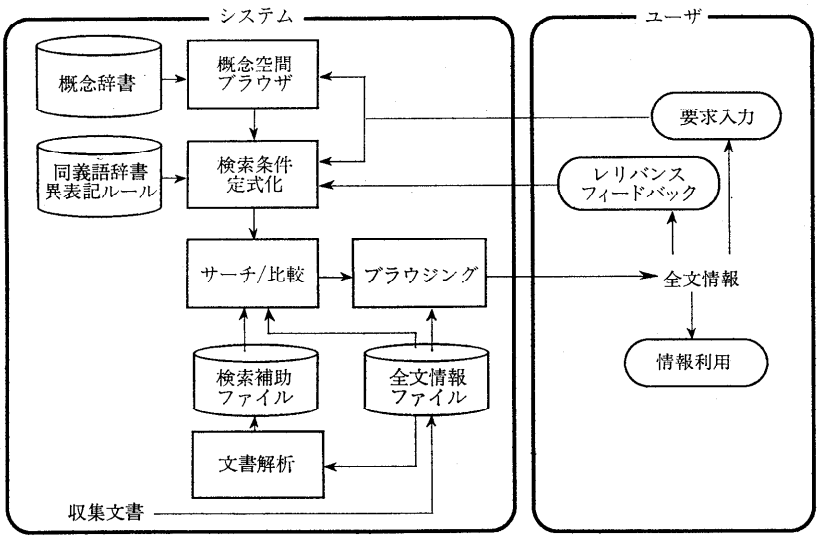


図-1 フルテキストサーチ型情報検索のモデル

（1） 検索システムのモデル

フルテキストサーチ型情報検索では**図-1**に示すごとく，収集された文書は全文情報ファイルに蓄積されるとともに，文書内容を解析することにより検索高速化のための検索補助情報が抽出される．ここでいう検索補助情報は，後述するようにサーチの実現方式に依存して，シグナチャファイル，あるいは全出現単語のインデクスを保持する転置ファイルである．

ユーザの要求はシステムに理解できる条件形式に定式化され，検索補助情報を用いることによって，高速に該当文書をサーチする．検索条件に合致した文書は，本文も含めてユーザに提示される．

フルテキスト型情報検索システムでは，全文情報をもつことによる次のようなメリットを与えることができる．

（a） ブラウジング：検索結果として本文を即座に表示することができるので，検索条件による自動的な絞り込みのみに頼る必要がない．絞り込みに際して，ユーザの検索結果の評価をシステムに伝達して検索条件を自動的に微調整するレリバンスフィードバックと呼ばれる方法が可能になる．すなわち，数件の文書内容を読んで，関係のあるなしをシステムに伝えて検索条件を自動変更するというサイクルを繰り返すことにより，順次欲しいものに絞り込んでいく方法論がある．

（b） 同義語異表記処理：任意の単語で検索することが可能であるので，同義語辞書や異表記展開ルールを用いて，ユーザが指定した検索語を自動的に拡張する方法が有効である．たとえば，「インタフェース」という検索語から，ルールを用いて，「インターフェース」，「インターフェイス」，「インタフェイス」などを自動発生することが可能である[16]．システムによっては1000語までの類義検索語を自動展開して，速度を落とすことなく，一括検索することができる[17]．

さらに発展した形として，語彙を概念的な階層に体系化した概念辞書をもたせて（図-1），抽象的な単語から下位語や関連語，あるいは固有名詞（地名，会社名など）に展開することも可能である．概念を代表する単語集合と各単語の関連度合いを示す重みを定義できるシステムもある[9]．

（2） サーチ方式

フルテキストサーチの実現方式には転置ファイルサーチとシーケンシャルサーチとがある[18),19]．

前者は，テキストデータからすべての単語を切り出し，それらすべての出現位置をインデクスとして保持する方法である．テキストの形態素解析により自立語と付属語の分離を行い，不要語辞書を用いて索引候補の単語や複合名詞句を抽出する．複合名詞句に対しては，切断用の辞書を用いて単語に分解する．たとえば，「複合名詞句」に対しては，「複合，名詞，句，名詞句，複合名詞」に分解する必要がある．

この方法は，切断用辞書の完備率によって検索

の自由度が制限される．したがって，単語レベルのインデクスではなく，単一の漢字をインデクスとする方法も考えられる．この場合，転置ファイルの規模は大きくなるが，自由な複合語での検索が可能になる．

シーケンシャルサーチは，文字どおりテキストを逐次スキャンして検索文字列を探索する方法である．通常，高速化のために，本文テキストをなんらかの形式で圧縮したシグナチャファイルを作成し，階層的に検索対象を絞り込む．また，文字列照合探索に専用のハードウェアを用いることも効果がある[20]．両方の効果を併せてシステム検索速度100 MB/秒を達成している試作機もある[17]．

（3） フルテキストサーチでの自然言語処理

自然言語処理技術は，登録文書のテキスト解析，自然語入力による検索条件の解析，同義語異表記処理，概念語の検索条件への展開処理に用いるほか，検索論理の高度化にも重要な役割をもつ．

現在の検索論理は，文字列探索を基本としてブール論理検索と近傍条件検索の複合条件が可能である．たとえば，「『音声』および『認識』が『知識処理』と8文字以内に隣接している」という条件で検索することが可能である．

今後は，構文解析や意味解析を適用して「知識処理を用いた音声認識システム」と「音声認識を用いて対話できる知識処理システム」を区別する意味的な検索や，ユーザの検索意図を推測して不足情報を補うような検索の実現のために，自然言語処理は重要な役割を果たす必要がある[14]．

また，フルテキストデータベースの構築に不可欠な文書認識・文書理解のための自然言語処理技術も期待すべき応用である．

4. 情報フィルタリングと情報分類

（1） 情報フィルタリングのモデル

情報フィルタリングとは，システムの外部から着信する膨大な量の情報や内部で生成される情報を解析・分類して，それらを必要とする人に配送することをいう（**図-2**）．技術的には情報検索で従来知られている選択的情報提供（Selective Dissemination of Information, SDI）と実質的に同じである[10]．ただし，情報フィルタリングは，SDIが情報提供側に視点を置くのに対して，情報受信側の機能としてみているという点が異なる．

着信する文書は内容の解析，分類，さらに重要度のランク付けが行われて検索補助情報とともに保存される．一方，システムにはユーザの関心事項を表現したインタレストプロファイルが事前に登録されている．新規に文書が到着するたびに，その検索補助情報と各ユーザのインタレストプロファイルが比較され，一致度の高いユーザにその存在が通知される．

ユーザはそれを個人用のファイルに格納したり，その文書が重要なものであることを他の人に伝えることができる．重要である旨を記したメモを作成して，より高い重要度でシステムに登録すればよい．元の文書は，そのメモからハイパーテ

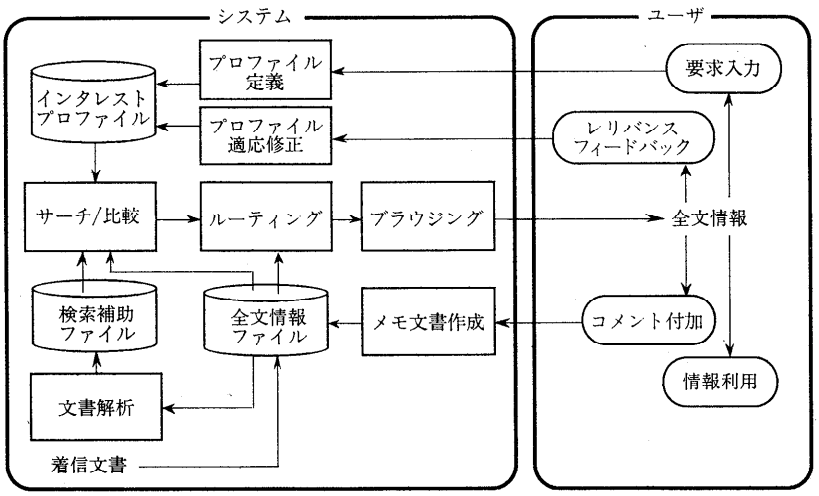


図-2 情報フィルタリングのモデル

キスト機能によって引き出せる.

(2) ルール型情報分類

英語テキストを自然言語処理を用いて自動分類するシステムとしてFRUMP[21]が早くから実用化されている.同様のシステムとしては,TOPIC[9], SCISOR[22], CONSTRUE/TIS[8]があり,一部は製品としても成功している.概念を定義するキーワード集合,ノイズを除去するための文脈条件,および候補概念を確定する条件を表現するルールベースによって,『…円に対するドルの価格…』を「円」のトピックスではなく「ドル」のトピックスと分類するような繊細な分類をも可能としている.近傍条件を含むフルテキストサーチとルール処理による「浅い意味処理」が成功している.

銀行などに常時入電されるテレックス文書の自動分類や,オンラインニュース・データベース構築のためのニュース記事の自動分類に応用されている.2万件のニュースサンプルを用いてジャーナリストと知識エンジニアが9.5人年を掛けてルールベースを構築した,674のカテゴリにニュースを分類するシステム事例がある[8].

(3) 事例型情報分類

テキスト分類手法の一つに記憶ベース推論(Memory-based reasoning, MBR)[23]を応用した教師あり学習型の方法がある.人手で分類した多数のサンプルテキストをすべて記憶しておき,分類したいテキストとそれら全サンプルとのある種の距離を計算して,最も近いサンプルのカテゴリをその分類と決定する.この方法は,最近傍法(Nearest Neighbor Method)としてパターン認識で古くから知られている.MBR は超並列計算機に向いており,特に Connection Machine を用いた研究がある.

MBR の応用研究として,米国の国勢調査の結果分析に適用した例が報告されている[24].13万2千の分類済みサンプルを用いて,約2200万の自然言語による回答を232の業種と504の職業に分類するというタスクである.人手で行ったときの誤り率以下になるように方式を調整したときに,60%以上のテキストが自動分類可能との見通しを得ている.

(4) テキスト分析と情報抽出

自然言語理解の究極の目標の一つは,知識を集約した文書からの概念獲得や知識ベースの構築であろう.これには「深い意味処理」が求められる.この問題の単純化の一つが上述したテキスト分類であるが,テキストの論理構造に着目した研究もある.

論理構造が割合しっかりした「建築法規」のような法規文書を対象にして,テキストから直接知識ベースを構築する試みがある[25),26].守るべき建築基準を遵守しているか否かを確認するエキスパートシステムの構築を狙っている.

(5) 情報フィルタリングでの自然言語処理

情報フィルタリングの多くの機能要素は,前述のように,情報検索のそれに含まれており,自然言語処理の役割も重なる部分が大きい.しかしながら,情報検索はユーザの要求になるべく即座に応えることが求められるのに対して,情報フィルタリングでは,着信文書の分類はユーザの意識しないときに行われる.また,前者では1件の検索条件と膨大な数の文書との比較をするのに対して,後者は1件の文書と多くても1000人程度のユーザのインタレストプロファイルとの比較を行えばよい.

したがって,機能的には類似しているが,要求性能,特に応答性能において異なっており,情報フィルタリングでは,より長い時間をかけることが許される.これにより,意味的な処理に対する期待も大きく,情報フィルタリングは今後自然言語処理技術が大いに活躍できる分野と目されている.

5. 検索支援ユーザインタフェース

情報検索は元来難しい作業である.だれもが容易に実行できるようにするには,検索エンジン部分の高機能化,ユーザとの検索対話の高度化,さらには方法論を含めた新しいシステム概念が求められる[2].フルテキストサーチ,情報分類,トピックス抽出などは検索エンジンに当たる.

検索対話では,検索者にとってもぼやけている要求を具体化させる支援が重要である.概念空間をシステムが積極的に可視化・表示して,検索条件の定式化を支援するアプローチが研究されている[11)~13].認知心理学的な実験からも,関連情報をユーザに提示して刺激することが有効であることが示されている.意味ネットワークを用いる知識ベース技術や,概念空間をブラウジングする

ためのグラフィカルインタフェース技術が求められる．

欲しいものを素直に表現するには自然言語によるアプローチも有効である[27]．関係データベースの検索に初めて適用された自然語インタフェースは，最近は，情報検索のほかにも種々の計算機コマンドの自然言語化を実現する基盤として拡張されつつある[28]．

検索対話における出力の面でも自然言語処理技術は重要である．検索結果は往々にして多数の文書を出力することがある．これらをいかに少ない面積で効率良く対話画面に表示するかは重要な問題である．KWIC (Key Word in Context) と呼ばれる表示方法もあるが，自然言語処理技術を用いる自動要約はその有効な方法である．

検索後の処理も，広い意味では検索支援に含まれるべきである．共有のデータベースなどから情報検索で得られた情報や知識は，利用された後，個人の情報検索システムに格納されることになる．この場合は，ユーザ個人の概念空間にマッチした格納の仕方が取られてしかるべきである．個個人にとって重要な概念を概念のネットワークに体系化して，各概念ノードに文書などの情報塊を連結するハイパーテキスト的なアプローチが提案されている[29]．ユーザが情報検索によって獲得した知識をいかにシステムに記憶させ，情報検索システムよりもいかに自然に，かつ適切に取り出せるようにするかは，これからの課題である．

6. むすび

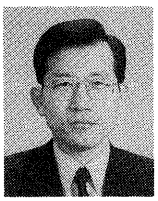
「記憶」から情報を取り出すところの情報検索は，人間の記憶のメカニズムが「理解」のメカニズムと切り離せない関係にあることからも分かるように，きわめて奥の深い問題である．究極の姿は，自然言語理解による深い意味処理を行い質問応答を行うシステムであろう．しかしながら，当面は，工学的なバランスを保ちつつ「浅い意味処理」を行う自然言語処理が成功の近道であると思われる．

参考文献

1) Lai, K.Y., Malone, T.W. and Yu, K.C.: Object Lens: A "Spreadsheet" for Cooperative Work, ACM Trans. Office Information Systems, Vol. 6, No. 4, pp. 332-353 (Oct. 1988).
2) 拝原正人：日本語文献データベースへの知的アクセス，電子情報通信学会誌，Vol. 72, No. 7, pp. 797-806 (July 1989).
3) 絹川博之：自然言語および AI 技術を利用した新しい情報検索システム，電気学会誌，Vol. 110, No. 2, pp. 113-118 (Feb. 1990).
4) 藤澤浩道，畠山 敦他：概念ネットワークを用いた知的ファイリングシステム，電子情報通信学会オフィスシステム研究会資料，OS 86-48 (1987-3).
5) 根岸正光：フルテキスト・データベースの応用動向，情報処理，Vol. 33, No. 4, pp. 413-420 (Apr. 1992).
6) Cote, R.G. and Diehl, S.: Searching for Common Threads, BYTE, pp. 290-305 (June 1992).
7) 日経 BP 社（北郷達郎）：急増する全文検索システムの動向を探る，日経インテリジェントシステム，172 号，pp. 16-21 (1993-2).
8) Hayes, P.J. and Weinstein, S.P.: CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories, Proc. 2nd Ann. Conf. Innovative Applications of Artificial Intelligence, Georgetown University, Washington, DC (May 1990).
9) McCune, B.P., Tong, R.M. et al.: RUBRIC: A System for Rule-Based Information Retrieval, IEEE Trans. Software Engineering, Vol. SE-11, No. 9, pp. 939-945 (Sep. 1985).
10) Belkin, N.J. and Croft, W.B.: Information Filtering and Information Retrieval: Two Sides of the Same Coin?, Comm. ACM, Vol. 35, No. 12, pp. 29-38 (Dec. 1992).
11) Tou, F.N., Williams, M.D. et al.: RABBIT: An Intelligent Database Assistant, Proceedings AAAI-82, Pittsburgh, American Association for Artificial Intelligence, pp. 314-318 (1982).
12) 木内伊都子，畠山 敦他：知的ファイリングシステムのビジュアルインタフェース，情報処理学会文書処理とヒューマンインタフェース研究会資料，21-3 (1988-11).
13) McMath, C.F., Tamaru, R.S. et al.: A Graphical Thesaurus-Based Informasion Retrieval System, Int. J. Man-Machine Studies, Vol. 31, pp. 121-147 (1989).
14) 杉山健司，秋山幸司他：自然言語理解に基づく情報検索システム IRIS，情報処理学会自然言語研究会資料，58-8 (1986-11).
15) 諸橋正幸：自動索引付け研究の動向，情報処理，Vol. 25, No. 9, pp. 918-925 (Sep. 1984).
16) 畠山 敦，川口久光他：自由語検索のための同義語異表記展開方式，第 39 回情報処理学会全国大会論文集，p. 1077 (1989).
17) 加藤寛次，藤澤浩道他：大規模文書情報システム用テキストサーチマシンの開発，情報処理学会情報学基礎研究会，14-6 (1989).
18) 菅野祐司，安藤 敦他：フルテキストデータベースの技術動向，電子情報通信学会データ工学研究会報告，DE 90-34, pp. 31-40 (1990).
19) 小川隆一，菊地芳秀他：フルテキスト・データ

ベースの技術動向，情報処理，Vol. 33, No. 4, pp. 404-412 (Apr. 1992).
20) 高橋恒介：機能メモリのアーキテクチャとその並列計算への応用—文字列照合処理への応用，情報処理，Vol. 32, No. 12, pp. 1268-1275 (Dec. 1991).
21) Dejong, G.: An Overview of the FRUMP System, *Strategies for Natural Language Processing*, Lehnert, W. G. and Ringle, M. H. Eds., Lawrence Erlbaum, Hillsdale, New Jersey, pp. 149-176 (1982).
22) Jacobs, P. S. and Rau, L. F.: A Friendly Merger of Conceptual Analysis and Linguistic Processing in a Text Processing System, 4th IEEE AI Applications Conference, Washington, D. C, pp. 351-356 (Mar. 1988).
23) Stanfill, C. and Waltz, D. L.: Toward Memory-Based Reasoning, Comm. ACM, Vol. 29, No. 12, pp. 1213-1228 (Dec. 1986).
24) Creecy, R. H., Masand, B. M. et al.: Trading MIPS and Memory for Knowledge Engineering, Comm. ACM, Vol. 35, No. 8, pp. 48-63 (Aug. 1992).
25) Moulin, B. and Rousseau, D.: Automated Knowledge Acquisition from Regulatory Texts, IEEE Expert, pp. 27-35 (Oct. 1992).
26) Reimer, U.: Automatic Knowledge Acquisition from Texts: Learning Terminological Knowledge via Text Understanding and Inductive Generalization, Proc. 5th Banff Workshop on Knowledge Acquisition for Knowledge-Based Systems (1990).
27) 西山敏雄，大山芳史：データベース検索における協調的な自然語対話処理と評価，電子情報通信学会論文誌，Vol. J 73-D-II, No. 4, pp. 625-632 (1990-4).
28) 絹川博之：表階層モデルに基づく自然語インタフェース処理方式，情報処理学会論文誌，Vol. 27, No. 5, pp. 499-509 (May 1986).
29) 藤澤浩道，山崎直子他：概念ブラウザと個人情報環境，情報処理学会情報メディア研究会資料，7-6 (1992-7).

(平成5年4月30日受付)



藤澤 浩道（正会員）

1946年生．1969年早稲田大学理工学部電気工学科卒業．1974年同大学院博士課程修了．同年(株)日立製作所入社．中央研究所にて文字認識，文書理解，テキストサーチマシン，知的ファイリングなどの研究に従事．1980年～1981年カーネギーメロン大学客員研究員．1992年～1993年同社ソフトウェア開発本部副技師長．現在中央研究所主管研究員．早稲田大学非常勤講師．電子情報通信学会，IEEE Computer, AAAI, ACM 各会員．工学博士．

絹川 博之（正会員）

1947年生．1970年東京大学理学部数学科卒業．同年(株)日立製作所入社．以来，漢字・日本語情報処理システム，仮名漢字変換，自動インデクシング，日本語文書処理，自然言語処理などの研究開発に従事．1986年度情報処理学会論文賞受賞．理学博士．現在同社システム開発研究所関西システムラボラトリ長．電子情報通信学会，計量国語学会，ACL 各会員．