日本Mテクノロジー学会事務局

〒260-8677 千葉市中央区亥鼻 1-8-1

千葉大学医学部附属病院企画情報部内　　鈴木隆弘

Tel: 043-226-2346

Fax: 043-226-2373

Email: mta-office@mta.gr.jp

ISSN 2185-4866

Technology
Association
Japan

Mumps

Vol.27,2014 Journal of MTA-Japan

The Official Journal of M-Technology Association-Japan

Mumps Vol.27, 2014

目次 頁

巻頭言

日本Mテクノロジー学会
会長　　土屋　喬義

M テクノロジー学会の前会長の木村一元先生が急逝されてから早や 2 年が過ましました。先生は学会誌編集委員長を 2002 年より、さらに学会長になられてからも続けられ、5 巻の学会誌 MUMPS を発行されております。木村先生は亡くなられる直前まで M テクノロジー学会のホームページの充実、今回発行する学会誌の準備に心血を注いでおられました。先生がこの学会誌に収載すべく作られた遺稿となる論文も収載致しました。木村前会長のご冥福をお祈り致しますと共に、無事学会誌を発行できたことをご報告いたします。

M 言語は 1977 年に米国規格協会（ANSI）標準に、1992 年に国際規格（ISO）標準に、1995 年に JIS 標準に制定されています。この CUI ベースの言語システムは高速、大容量、ネットワークの接続性など優れた特徴を最初から持っていました。私もその一人ですが、マイクロコンピュータの黎明期より、プログラミング言語 MUMPS を使用して仕事、研究に役立つプログラムを作った経験のある人々は多数いました。昨年の医療情報連合大会於いて開催した本学会の企業展示ブースでも M に接したことのある方々から多くの声をかけられました。OS が DOS から Windows に替わっても多くの M ユーザーは CUI（DOS 窓）を使用し続けていました。M には VB に対するインターフェースや独自の GUI ユーザーインターフェースを実装しましたが、当時は GUI 技術が難解で多くのユーザーが離れていったのでした。近年、Web 技術の発展により Ajax、jQery、EWD.js などの様々な外部のライブラリーが M でも使える事が判ってきました。また M 言語を包含する InterSystems 社の Caché では CSP や ZEN という Web アプリケーションのフレームワークが公開されています。これらは使い慣れた HTML＋M 言語で表現できるので驚くほど簡単に状況に応じた動的な Web アプリケーションを生成できます。これは個人又はシステム部でのユーザープログラミングの復活の鍵になるのではないでしょうか。

また、M 言語は SQL では対応できない巨大なデータを高速に取り扱う NoSQL としても注目を浴びています。医療連携、金融システムなどで M 言語は使われており SQL 勢と熾烈な競争が繰り返されています。M の大規模性、高速性、接続性、に注目した巨大プロジェクトの研究、開発、事例発表などが行われています。

この様な中で、M 言語環境を持つ Caché ユーザー数増加してきています、M テクノロジー学会としては魅力的な M の使用技術の提供とコンピュータサイエンスとしての学会活動、学会誌刊行との両立を図る事が大切だと感じております。雑誌 MUMPS が情報学を目指す研究者の発表の場となり、また多くの M 言語に興味を持たれる皆様にアイデア提供の場となるよう願っております。

2014 年 2 月吉日

CSP を利用した入院病歴要約情報からの患者検索システム

Patients Retrieval System in Discharge Summaries by Using Cache Server Pages

土井俊祐 1)2)，木村隆 2)，鈴木隆弘 2)，田村俊世 1)，高林克日己 2)
Shunsuke Doi1)2), Takashi Kimura2), Takahiro Suzuki2), Toshiyo Tamura1), Katsuhiko Takabayashi2)

1）千葉大学院工学研究科，2）千葉大学医学部附属病院企画情報部
1) Graduate School of Engineering, Chiba University
2) Department of Medical Informatics and Management, Chiba University Hospital

千葉県千葉市中央区亥鼻 1－8－1
1-8-1, Inohana, Chuo-ku, Chiba, 260-8677, Japan
TEL:043-226-2346, FAX:043-226-2373
e-mail: s.doi@graduate.chiba-u.jp

要旨

近年、病院情報システムの普及により、膨大な量の医療情報が電子的に保存されるようになった。しかし、秩序のない膨大量のデータから必要な情報を検索することは、非常に困難である。そこで本研究では、千葉大学医学部附属病院の病院情報システムに保存されている入院病歴要約情報を利用し、キーワード検索と類似症例検索を備えた症例検索エンジンを構築したので報告する。開発環境としては MUMPS を搭載する Cache と、インターフェイスには CSP を用いた。本システムを利用することで、より簡単に患者を検索できることが可能になり、院内における診療情報の利活用を促進させることが期待できる。

Abstract

Recently, huge data of medical information have been stored electronically. With these data, we constructed a search engine that enables similar cases as well as keyword retrieval and similar case retrieval from discharge summaries in Chiba University Hospital. We used MUMPS and CSP(Cache Server Pages) to develop this system. Using this system, it is expected that doctors are able to search patients more easily and discharge summaries can be used more effectively.

キーワード：情報検索　入院病歴要約　類似症例検索　Cache Server Pages
Keywords：Information retrieval, Discharge summary, Similar case retrieval, Cache Server Pages

平成 23 年 9 月 9 日受付　　平成 23 年 11 月 2 日受理

4　CSP を利用した入院病歴要約情報からの患者検索システム

1.　はじめに

近年、病院情報システムの普及により、膨大な量の医療情報が電子的に保存されるようになった。その情報量は集積回路の発達によりますます膨れ上がる一方であるが、より多くの情報を”保存すること”にのみ重きがおかれており、保存した情報を”利用すること”は、まだ発展途上の段階である。

保存した情報の利用方法の１つとして、医師が過去の症例情報を参照することがある。多種多様な疾患に対し、自らの経験・知識のみで臨むことは、全ての医師にできることではない。そこで医師は、過去のカルテを参照することで、知識や経験の不足分を補っている。しかし、秩序のない膨大量のデータから自力で的確な患者や症例を検索することは、非常に困難である。大規模病院においては、DWH (DataWare House)等の症例検索装置を導入している施設もあるが、検索にスキルが必要であり、煩雑である。それ故、実際の医療現場では貴重な診療情報を利活用しきれていない現実がある。

そこで本研究では、千葉大学医学部附属病院の病院情報システムに保存されている入院病歴要約を利用し、院内における診療情報の利活用のための患者検索エンジンを構築したので報告する。インターフェイスは使い勝手のよい Web ベースの検索エンジンとし、検索方法としては、通常のキーワード検索に加え、MTA2008 にて発表した自然言語処理に基づく類似症例検索[1]を並列して設置し、ユーザに目的に合わせた検索方法を選択できるようにした。

本論文では、検索エンジン構築にあたり、Cache 並びに MUMPS(以下、M言語)を用いたが、これらが検索においてどのような役割を担っているかを処理方法やインターフェイスと合わせて論ずる。

2.　対象

本研究では、千葉大学医学部附属病院院の病院情報システムに保存されている 1977 年以降の入院病歴要約情報約 34 年分約 28 万件をデータベースとして利用した。そのうち、1999 年以降の約 13 万件については、退院時所見についても全文が電子保存されている。図 1 に当院の入院病歴要約の表示画面例を示す

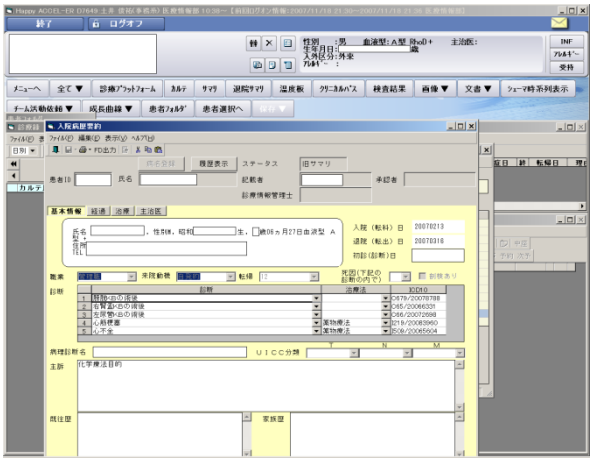


図１　入院病歴要約情報の表示画面例

入院履歴内で実際に利用した情報は表 1 の通りである。これらの情報は、病院情報システムの内において別々の場所に保存されており、本研究ではあらかじめ必要な情報を抽出し、用意したサーバ内に保存した。

表１　利用した入院履歴

保存場所	利用した情報
患者基本情報	患者 ID，年齢，性別
入院病歴要約基本情報	入退院年月日，病名 入院診療科，DPC，ICD，入院時併存疾患， 入院後発生疾患，主訴 家族歴，既往歴，転帰 手術情報，作成者
退院時所見	電子化分のみ全文

3.　方法

症例検索エンジンを開発するために、まず入院履歴を整理し、患者データベースを作成する。退院時所見については、類似症例検索のために自然言語処理によるベクトル化処理を行い、作成した文書ベクトルを文書の類似性評価の指標とした。作成したデータベースをもとに、院内 LAN 上に Web ベースの検索エンジンを構築・運用し、臨床現場に提供する。

3.1 患者データベースの作成

患者データベースの作成プロセスを図 2 に示す。まず、病院情報システムから患者基本情報、入院病歴要約基本情報、退院時要約を抽出する。これらの情報は病院情報システム内にグローバルとして保存されているため、グローバルごとエクスポートすることで取り出すことができる。ここで、検索エンジンにおいて直接病院情報システムに照会することはリスクを伴うため、本研究では、定期的に病院情報システムからグローバルを抽出し、専用のサーバにアップロードすることで、データベースを更新する方式とした。データベースの作成方法として、まず患者基本情報、入院病歴要約情報の中から必要な情報のみを抽出・整理し、1 入院ごとに患者データベースとなるグローバル^CUHDIS のノードに保存した。退院時要約については、類似症例検索に利用するため、1 入院 1 ファイルとしてサーバ内に保存した後、後述する形態素解析・tf*idf 法によるベクトル化の処理をし、作成した文書ベクトルを同様にグローバル^CUHDIS に保存した。

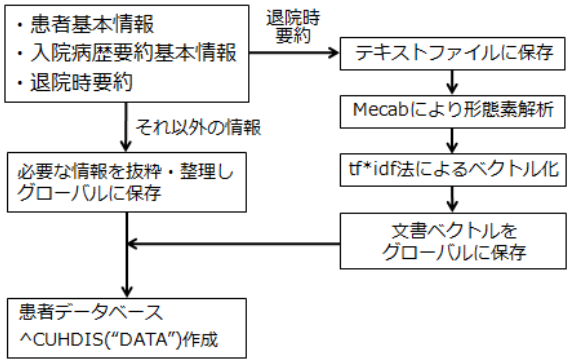


図 2　患者データベースの作成フローチャート

3.1.1 形態素解析

文書の類似性を比較するためには、自然言語処理による文脈理解が欠かせない。しかし、日本語は他の言語と異なり、単語の間に区切りとなるスペースなどがないため、この形態素解析により文章を分かち書きしする必要がある。

形態素解析器には、処理の高速性から京都大学で開発された Mecab ver0.97[2]を利用した。処理については、まず Cache サーバ上から$ZF 関数を用いてバッチファイルを作成・起動し、図 3 に示す出力結果のファイルを OPEN 関数、$PIECE 関数を用いて一般名詞のみを Cache グローバル^CUHDIS に取り込んだ。特に$PIECE 関数によるパターン照合は、M言語の代表的な文字列処理機能の 1 つであり、コード・処理時間ともに大幅なコスト削減を実現している[3]。

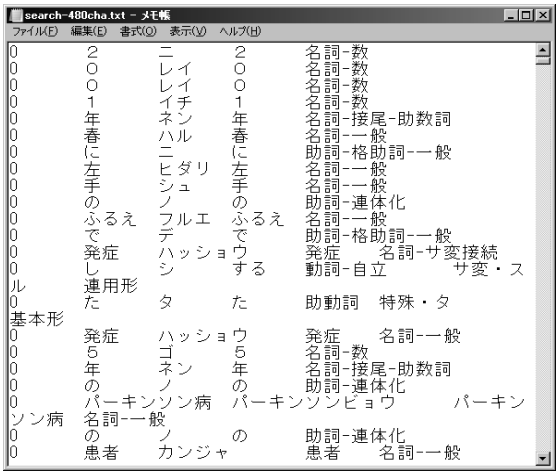


図 3　Mecab の出力結果

※Tab 区切りテキストとして出力させることで、$PIECE 関数での読み取りを容易にしている

3.1.2 *tf-idf*法による文書ベクトル作成

tf-idf 法[4]は単語の頻度をもとに文書中の単語の重要度を算出する手法であり、代表的な情報検索手法の 1 つである。文書中の単語の出現回数で表わされる出現頻度 *tf*(term frequency)と、全文書中の単語の出現割合で表わされる逆出現頻度 *idf*(inverse term frequency)、文書長を補正する正規化係数 *R*(document norm-alization factor)で表わされる。今、文書数が m 件、全文書内に n 種類の単語が含まれている文書集合 *D* があると仮定とする。文書集合 D において任意の文書 *Di* の中の単語 *Wj* の重要度 $\alpha(i,j)$ は(1)式のように表わされる。

$$\alpha(i,j)=\frac{tf(i,j)\times idf(j)}{R(i)} \qquad (1)$$

これを文書中の全単語において計算することで、(2)式に表すような文書ベクトルを作成することがで

きる。

$$D(i)=\left(\alpha(i,1),\alpha(i,2),,,\alpha(i,j),,,\alpha(i,n)\right) \quad (2)$$

これにより、テキストデータであった文書を数値として扱うことができる。作成した文書ベクトルは、患者データベースとなるグローバル^CUHDIS に同様に保存した。類似症例検索では、この文書ベクトル同士の内積距離を文書の類似性を表す指標として用いる。

3.2　検索エンジンの構築

本システムは、院内 LAN という限られたネットワーク上にて運用するものであるが、電子カルテ上から起動する DWH との差をつけるため、院内 Web 上のブラウザから簡単にアクセスできるシステムとする。これを実現させるために、Cache の Web アプリケーションとしての特性を利用する。

3.2.1　開発環境　　用意したサーバの仕様は、Windows7 Professional 64bit、Intel Core i7 960 CPU 3.4GHz、RAM 16.0GB である。Web サーバには Apache2.2.13 を用いた。検索システムのインターフェイスとしては、Cache に標準装備されている CSP(Cache Server Pages)を用いた。CSP はロジック、デザイン、データアクセスの全てを同じプラットフォームで実行することができるうえ、基本的な記述が HTML であるため、Stylesheet や JavaScript 等も組み込むことができる。また、CSP Gateway により Web サーバを介したブラウザへの通信もスムーズに行うことができる[5]。図 4 に本システムのネットワーク構成図を示す。

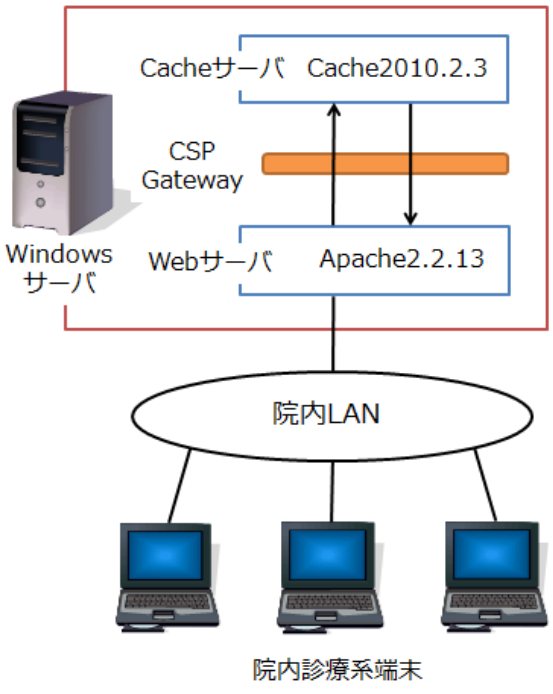


図 4　ネットワーク構成図

3.2.2　検索処理　　図 5 に検索処理のフローチャートを示す。Cache サーバがユーザからの要求を受け取ると、まず全てのデータを%CSP.Session オブジェクトの Data プロパティに格納する。%CSP.Session オブジェクトは同一セッションであればセッション終了まで保持されるため、ページ遷移がある場合でもデータを保持することができる。

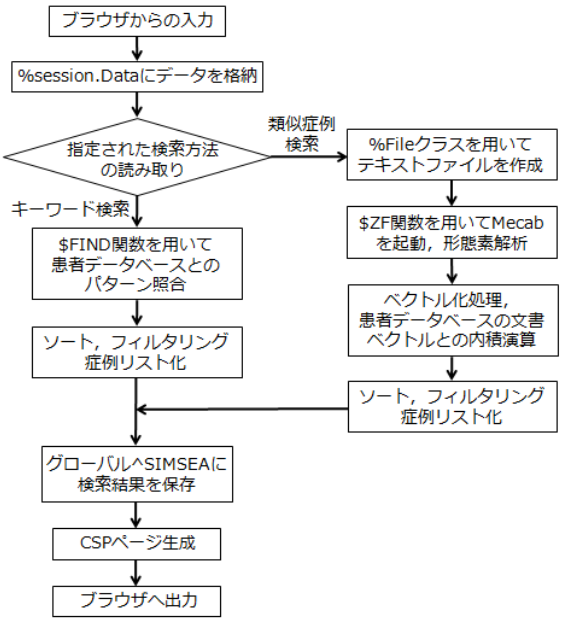


図 5　検索フローチャート

次に、ユーザがキーワード検索と類似症例検索のどちらを選択したかの判別が行われ、それにより処理方法が異なる。キーワード検索の場合は、退院時所見を含めた全データに対し$FIND 関数によるパターン照合が行われる。M 言語には$FIND と$LENGTH の 2 種類のパターン照合が可能な関数があるが、検証の結果速度で勝る$FIND 関数を利用した。類似症例検索の場合は、患者データベース作成の際と同様

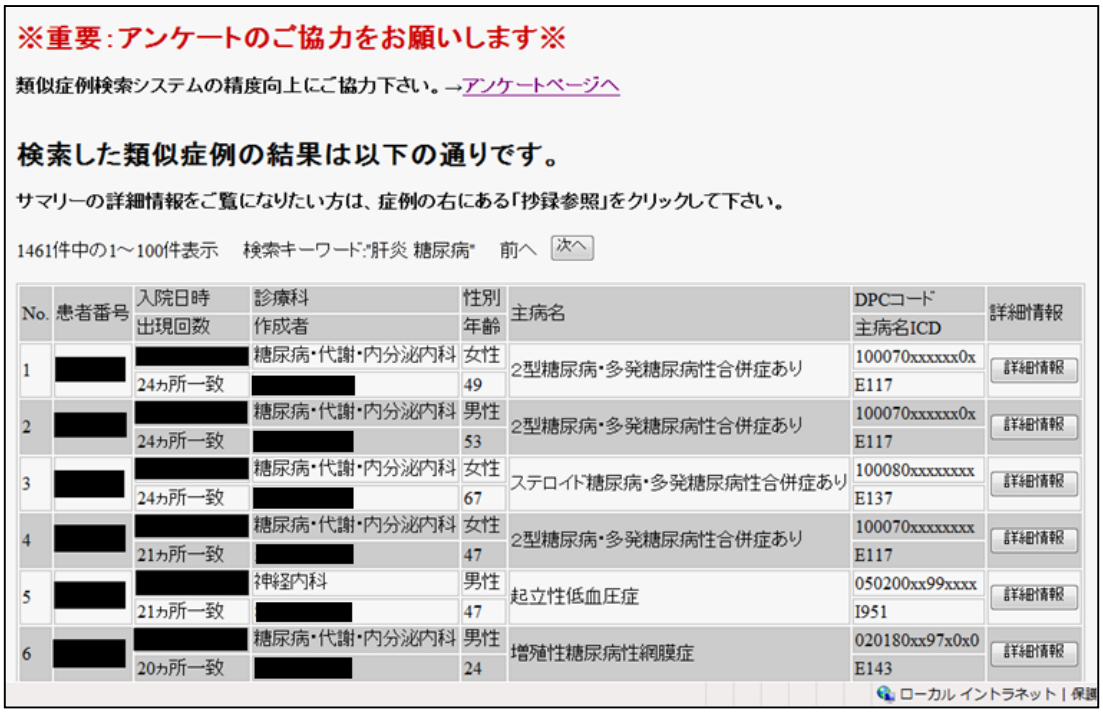
の手法で文書ベクトルを作成し、データベース内との内積距離による類似度を算出する。そして、両検索方法ともに、ユーザの設定した方法に従い、ソート、フィルタリングを行い、症例をリスト化する。ソートの方法としては、検索結果格納用グローバル^SIMSEA のノードにソートに用いる値や文字列を保存し、$ORDER 関数を用いることで、軽快にソートを行うことができた[6]。最後に、リスト化された症例から CSP ページを生成し、ブラウザに出力する。

4. 結果

検索エンジンのトップページを図 6 に示す。ユーザは院内 LAN に接続された端末からのみ本システムにアクセスが可能であり、院内向けホームページにリンクを用意している。 検索の手順としては、トップページのフォームに検索キーワードもしくは類似症例検索をする症例の文書を入力し、ソート，フィルタリングの条件を選択し送信する。類似症例検索については、普段は使わない文書を入力しての検索ということもあり、使い方を説明するページを用意し、トップページにリンクを貼った。送信されたデータは、3.2.2 項で述べた処理が行われ、ブラウザには図 7 に示す検索結果表示ページが出力される。このページでは、キーワードの一致度もしくは文書の類似度が高い入院病歴要約がリスト化され表示される。図 7 の例では、「肝炎　糖尿病」という 2 語を入力した場合のキーワード検索結果を表している。その他に、患者番号、入院日時、情報作成者、性別、年齢、主病名、DPC コード、ICD コードが確認でき、退院時所見等のそれ以外の情報については、各入院患者リスト右端の詳細情報ボタンから図 8 に示す入院病歴要約情報表示ページに遷移することで閲覧することができる。

図 6　検索エンジントップページ

8　CSP を利用した入院病歴要約情報からの患者検索システム



※重要：アンケートのご協力をお願いします※

類似症例検索システムの精度向上にご協力下さい。→アンケートページへ

検索した類似症例の結果は以下の通りです。

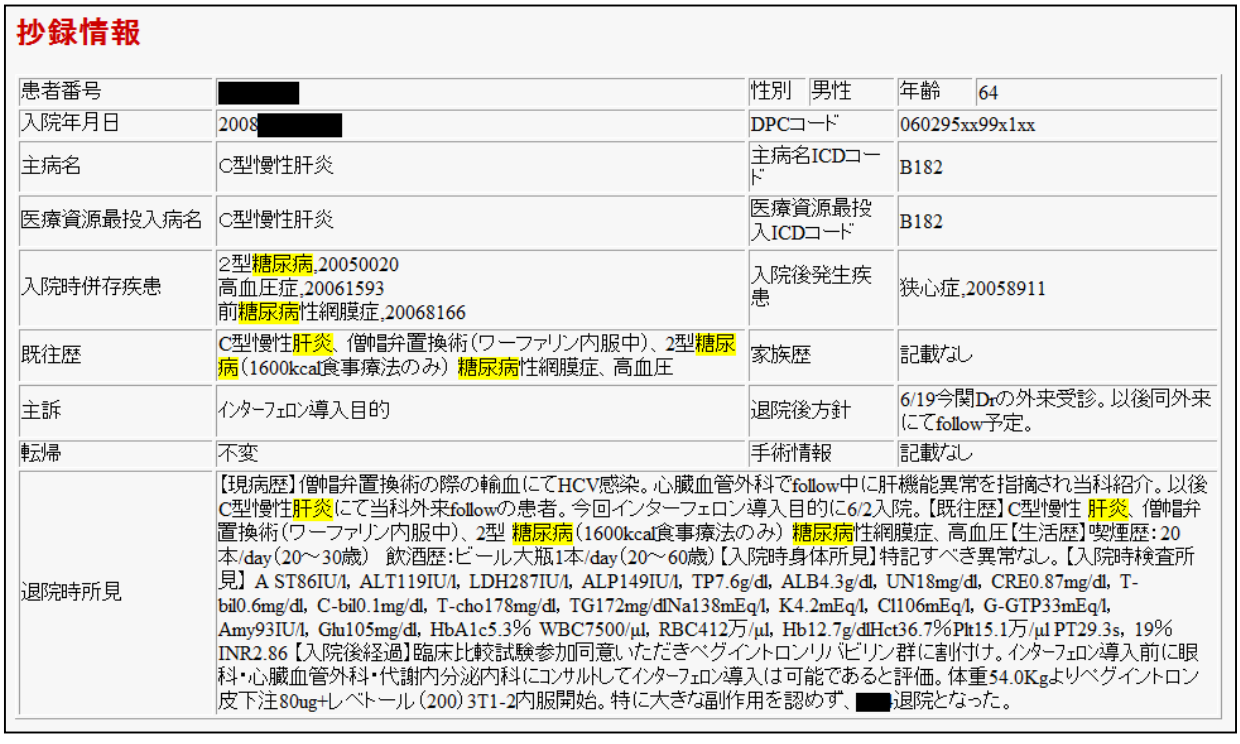
サマリーの詳細情報をご覧になりたい方は、症例の右にある「抄録参照」をクリックして下さい。

1461件中の1～100件表示　検索キーワード:"肝炎 糖尿病"　前へ　次へ

No.	患者番号	入院日時 出現回数	診療科 作成者	性別 年齢	主病名	DPCコード 主病名ICD	詳細情報
1	[redacted]	[redacted] 24ヵ所一致	糖尿病・代謝・内分泌内科 [redacted]	女性 49	2型糖尿病・多発糖尿病性合併症あり	100070xxxxxx0x E117	詳細情報
2	[redacted]	[redacted] 24ヵ所一致	糖尿病・代謝・内分泌内科 [redacted]	男性 53	2型糖尿病・多発糖尿病性合併症あり	100070xxxxxx0x E117	詳細情報
3	[redacted]	[redacted] 24ヵ所一致	糖尿病・代謝・内分泌内科 [redacted]	女性 67	ステロイド糖尿病・多発糖尿病性合併症あり	100080xxxxxxxx E137	詳細情報
4	[redacted]	[redacted] 21ヵ所一致	糖尿病・代謝・内分泌内科 [redacted]	女性 47	2型糖尿病・多発糖尿病性合併症あり	100070xxxxxxxx E117	詳細情報
5	[redacted]	[redacted] 21ヵ所一致	神経内科 [redacted]	男性 47	起立性低血圧症	050200xx99xxxx I951	詳細情報
6	[redacted]	[redacted] 20ヵ所一致	糖尿病・代謝・内分泌内科 [redacted]	男性 24	増殖性糖尿病性網膜症	020180xx97x0x0 E143	詳細情報

ローカル イントラネット | 保護

図 7　検索結果表示ページ



抄録情報

患者番号	[redacted]	性別　男性	年齢　64
入院年月日	2008[redacted]	DPCコード	060295xx99x1xx
主病名	C型慢性肝炎	主病名ICDコード	B182
医療資源最投入病名	C型慢性肝炎	医療資源最投入ICDコード	B182
入院時併存疾患	2型糖尿病,20050020 高血圧症,20061593 前糖尿病性網膜症,20068166	入院後発生疾患	狭心症,20058911
既往歴	C型慢性肝炎、僧帽弁置換術（ワーファリン内服中）、2型糖尿病（1600kcal食事療法のみ）糖尿病性網膜症、高血圧	家族歴	記載なし
主訴	インターフェロン導入目的	退院後方針	6/19今関Drの外来受診。以後同外来にてfollow予定。
転帰	不変	手術情報	記載なし
退院時所見	【現病歴】僧帽弁置換術の際の輸血にてHCV感染。心臓血管外科でfollow中に肝機能異常を指摘され当科紹介。以後C型慢性肝炎にて当科外来followの患者。今回インターフェロン導入目的に6/2入院。【既往歴】C型慢性 肝炎、僧帽弁置換術（ワーファリン内服中）、2型 糖尿病（1600kcal食事療法のみ）糖尿病性網膜症、高血圧【生活歴】喫煙歴：20本/day（20～30歳）　飲酒歴:ビール大瓶1本/day（20～60歳）【入院時身体所見】特記すべき異常なし。【入院時検査所見】A ST86IU/l, ALT119IU/l, LDH287IU/l, ALP149IU/l, TP7.6g/dl, ALB4.3g/dl, UN18mg/dl, CRE0.87mg/dl, T-bil0.6mg/dl, C-bil0.1mg/dl, T-cho178mg/dl, TG172mg/dlNa138mEq/l, K4.2mEq/l, Cl106mEq/l, G-GTP33mEq/l, Amy93IU/l, Glu105mg/dl, HbA1c5.3% WBC7500/μl, RBC412万/μl, Hb12.7g/dlHct36.7%Plt15.1万/μl PT29.3s, 19% INR2.86【入院後経過】臨床比較試験参加同意いただきペグイントロンリバビリン群に割付け。インターフェロン導入前に眼科・心臓血管外科・代謝内分泌内科にコンサルトしてインターフェロン導入は可能であると評価。体重54.0Kgよりペグイントロン皮下注80ug+レベトール（200）3T1-2内服開始。特に大きな副作用を認めず、[redacted]退院となった。		

図 8　入院病歴要約情報表示ページ

入院病歴要約情報表示ページでは、検索キーワードや、類似症例検索において重要度の高かった単語をマーカーで色分けして表示する工夫を加えた。ページ下には絞り込み検索や、表示した症例の類似症例を検索できるフォームを用意することで、入力の手間の軽減や、利便性の向上を図った。

また、検索速度については 2 語のキーワード

検索の場合で約 3 秒、800 字の文書の類似症例検索で約 10 秒であった。

5. 考察と課題

本システムは現段階で院内に公開され利用できる状態であるが、今後の運用についてはいくつかの課題を残している。

まず、類似症例検索では形態素解析の際に同義語の統制を行っているが、キーワード検索については､現状では M 言語の関数で完全なパターン一致を抽出しているだけであり、複雑な表現形態を持っている医療用語を扱うシステムとしては、改善が必要であると考える。入院病歴要約情報は電子的に入力・保存されているため、入力用辞書によりある程度の標準化が進んでいるものの、年代や個人、診療科等により扱う語に差が出ることは容易に想像できる。方法としては、標準病名集等の同義語を扱う医療辞書を利用し、グローバル上に同義語統制用辞書を作成する方法が考えられる。ただし、読みや漢字の統制のレベルであれば問題ないが、病名等では状況により同義語の数が飽和したり、必要とされる表現形態が存在したりすることが考えられるので、どこまで自動統制するかを慎重に検討する必要がある。

次に、検索速度であるが、キーワード検索の約 3 秒は許容範囲と考えられるものの、類似症例検索の約 10 秒はユーザによっては煩わしさを覚える可能性がある。現在、入院履歴は約 28 万件を数えるが、今後さらに蓄積していくことが想定される。時間コストを抑えつつ運用する方法として、情報が古くなったり、不足したりしている履歴、同一人物の重複している入院履歴を検索対象から除去することが考えられる。入院病歴要約では、入力不足でデータが欠けていたり、同じ患者が何度も入院するとほとんど変わらない情報が記載されていたりするケースが存在する。それらの情報は、検索されてもあまり利用されないことが考えられる。よって、情報量や重複の度合いにより、履歴ごとにある程度の優先度をつけ、優先度の低いものから検索対象からはずすことで、検索速度の向上が期待できる。また、類似症例検索に限れば、自然言語処理分野においても、今後の検索時間コストの増大は喫緊の課題であるため、多くの研究がなされている。例えば、LDA(Latent Dirichlet Allocation)等の文書ベクトルの次元を圧縮する技術を導入すれば、検索速度の向上に期待できる[7]。しかし、恒久に蓄積し続けることを考えると、限られた設備で運用するためには、いずれ対象とするデータの取捨選択は避けられないものと考える。

6. まとめ

病院情報システムに蓄積された入院病歴要約情報を利用し、キーワード検索と類似症例検索の 2 つの検索方法により患者を検索できるシステムを開発した。これらは、M 言語の持つ豊富な文字列処理機能と、CSP の優れた開発プラットフォームにより実現することができた。

本システムにより、従来から導入されている、検索スキルの必要な DWH を使わず、より簡単に患者検索を行うことが可能になった。これにより、院内における診療情報の利活用を促進させることが期待できる。

課題としては、辞書の導入によるキーワード検索の同義語の統制や、LDA 等の検索技術の導入、入院履歴データの取捨選択による検索速度の向上等が考えられた。

参考文献

[1] 土井俊祐, 鈴木隆弘, 藤田伸輔, 高林克日己: Cache を用いたベクトル空間モデルの作成と類似症例検索システム.日本 M テクノロジー学会大会論文集.35 (1) :25. 2008.

[2] Kudo T, Yamamoto K, Matsumoto Y: Applying conditional random fields to Japanese morphological analysis. Proc. EMNLP. 2004(1): 230–237. 2004.

[3] 木村一元, 五十嵐吉彦:CSP での医師国家試験

問題検索システムの構築. Mumps.25(1): 59-65. 2010.
[4] Goldman JA. Chu WW. Parker DS. Goldman RM:　Term domain distribution analysis a data mining tool for text databases. Methods Inf Med. 38(1): 96-101. 1999
[5] 岩本知宏:CSP を用いた Web アプリケーション作成時の留意点.日本Mテクノロジー学会論文集. 31(1): 23-24. 2004.
[6] 日本 M テクノロジー学会: MUMPS 解説. http://www.mta.gr.jp/about/04-02.html [accessed October 31, 2011]
[7] D.M.Blei, A.Y.Ng and M.I.Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research:3(1): 993-1022. 2003

日本語解析システム「ささゆり」の 同義異表現検索機能と日本語の言い換え技術

Search Technology of Synonymous Equivalent Representations in the Japanese and Paraphrasing Technology of Japanese Sentences with the Japanese Analysis System SASAYURI

高橋 亘
Wataru Takahasi
関西福祉科学大学社会福祉学部・大学院社会福祉学研究科
〒 582-0026 大阪府柏原市旭ヶ丘 3-11-1
TEL 0729-78-0088, FAX 0729-78-0377
E-mail takahasi@fuksi-kagk-u.ac.jp

要旨 日本語解析システム「ささゆり」は, M 言語の階層型データの特質をアルゴリズムの中に取り入れて日本語の形態素把握や, 構文解析, 意味解析を行うシステムである.

携帯電話の検索アプリケーションなどに使用されている近年の日本語解析の技術は, 音声認識における進展と気の利いた解答の仕方には目を見張るものがあるが, 日本語文そのものの構文解析や意味解析については危うい点がいくつも観察される. 我々のシステムは, 解析対象としている日本語文の多様性や構文判断の的確性において, こうした技術に比べてはるかに重篤な基礎を保持している.

この論文では, 日本語解析システム「ささゆり」の同義異表現検索機能と難解な日本語の言い換え技術の近年の発展が総説される. 我々のシステムでは, 日本語の意味を担う基礎に知覚連語を取っている. 知覚連語とは, 日本語の使用例の中で, 単語同士が結合して意味的に純粋な状態を呈している連語のことを言う. 知覚連語は, その形成規則に従って日本語解析システム「ささゆり」によって機械的に学習される. 近年, 知覚連語を基礎に日本語の複文を単文化して被修飾名詞の意味を推定する技術が確立された. 同じ用語を含んでいる知覚連語の集合に共通語同値類と名付け, 使用されている語にかかわらず意味的に近い知覚連語の集合に同義性同値類と名付けて, 同義異表現検索機能と難解な日本語の言い換え技術の基礎が確立された.

日本語解析システム「ささゆり」の同義異表現検索機能と難解な日本語の言い換え技術は, 近年次第にその適用範囲を拡げ, 多分野にわたってその有効性を示しつつある. 当初は, 特にコミュニケーション支援という筆者の研究分野における活用に焦点を当て, 聾者への情報保障という分野について, 形式名詞の意味推定やオノマトペの代替表現といった適用が中的心課題であったが, 今日では, 物理, 化学, 動物, 植物, 医学, 文学, 言語, 心理, 美術, 神話, 民族, 文化, 料理などの多様な分野への適用が進展しつつある. この論文では特に神話, 中でもギリシャ神話への適用を例に採ってその有効性が例証される.

適用範囲が拡大されていく中で新たな技術的発展も行われている. この論文では M 言語の大域変数の中間位ノードの修正時における枝葉データのバックアップ技術についても触れられる.

キーワード: 日本語解析システム「ささゆり」, 複文の単文化, 同義検索, 難解文の簡易化と言い換え

平成 24 年 11 月 27 日受付 平成 26 年 1 月 30 日受理

1. はじめに

日本語解析システム「ささゆり」は M 言語の階層型データの特性をアルゴリズムに組み込んで活用した日本語解析システムであるが，語が結合して意味が純粋な連語を構成するという原理を基礎に，日本語文の構文解析や意味解析を行うことが出来るシステムである [1]．我々は，語が結合されて意味的に純粋になった連語が人の知覚を明確に誘発することから，文を構成する要素としての連語を知覚連語と呼んでいる．構文解析については，日本語の複文を，修飾文と，もとの文から修飾文を除去した骨格文(被修飾名詞を含む）と，に分解する機能を備えている．システム名は，システムが複文を修飾文（葉に対応）と骨格文（茎に対応）とに分解した結果のイメージが百合科の植物に似ていることに由来している．

近年，システムに固有な構文解析機能と意味解析機能を活用して，同義異表現検索や日本語を分かりやすい表現に言い換える技術について，いくつかの有効な機能を保持するに至った [2-10]．これらは大別すると，(i) 複文を単文化し，被修飾名詞の意味を推定する機能，(ii) 難解語を含む知覚連語を高速に検索する機能，(iii) 難解語を含む知覚連語に同義で，かつ難解語を含まない知覚連語を検索する機能，の三つであるといえる．これらの機能は，未知の用語や難解な用語を含む表現を既知の言葉で言い換える機能，同義異表現の文章を検索する機能，などとして，キーワード検索とは違う新しい検索技術を提供する．難解な用語として当初想定されていたものは，形式名詞やオノマトペといった聾者に分かりにくい用語に限られていたが，今日では，物理，化学，動物，植物，医学，文学，言語，心理，美術，神話，民族，文化，料理など，各種の分野の専門用語への拡張が試みられている．

広汎な同義検索や日本語の言い換え技術を可能にしている階層型データを管理しているインターフェイスは次の 5 つである．

(1) 知覚連語辞書: 日本語解析システムが最初に行う処理はテキスト文を対象にした知覚連語の機械学習である．知覚連語の形成文法にしたがってシステムは知覚連語にふさわしいと判断される品詞列を抽出し，知覚連語の候補として知覚連語辞書にリストする．リストされた品詞列には，単語の列として意味的に知覚連語としてふさわしくないものが含まれるから，これを利用者が視覚確認をして廃棄する．結果，残された品詞列が知覚連語としてデータベースに蓄積される．知覚連語の保存状態は知覚連語辞書が一括管理しているが，これまでに保存されている知覚連語は 91 万個を越える．知覚連語は，専門用語を，単語もしくは熟語の形の構成要素として，内包するが，これらはデータ中に品詞列として既に同定ずみであるから，知覚連語と専門用語の対応関係は一意的に決定される．

(2) 概念辞書: 概念辞書は知覚連語辞書と双対的に構成され，知覚連語に割り当てられる意味要素を管理している．知覚連語に割り当てられる意味要素はこの辞書で総合的に管理されるが，意味要素と知覚連語の相関関係も機械的に登録され特定の意味要素を保持する知覚連語の検索は即時的である．

(3) 専門用語辞書: 日本語解析システムは専門用語辞書を内包しており，専門用語かどうかの判断はこの辞書との照合によって行われる．専門用語として登録された語と知覚連語の相関関係は知覚連語辞書の操作で機械的に登録される．このため，一つの専門用語を含む知覚連語の検索は即時的である．

(4) 共通の用語を含む表現の検索: 専門用語と対応づけられた知覚連語を検索し，用語の使用例をリストするインターフェイスである．補助機能として概念辞書のデータに知覚連語の保持する意味要素を追加する機能をもっている．このインターフェイスには検索用語の他に補助キーを指定する機能が用意されているので補助キーを活用して用語の使用例を共通の意味を持つものに絞り込みながら概念辞書に未登録の意味要素を追加することが出来る．

(5) 同じ意味を持つ言い換え表現の検索: 課題文に対して言い換え表現を表示するインターフェイスはつぎのようなものである．専門分野を指定し，課題文と検索文との意味的距離の範囲を指定して，課題文を入力すると，システムは課題文を構文解析・意味解析して，そこに含まれる専門用語を抽出し，課題文から指定範囲にある知覚連語を言い換え表現の候補として検索する．もし，課題文が複文であれば，構文解析結果の表示欄に被修飾名詞（形式名詞の場合もある）の，修飾文によって限定された意味要素と代替表現としての内容名詞が表示される．

第 2 節と**第 3 節**で，日本語解析システム「ささゆり」の構文理解の仕方と同義異表現検索の基本的原理を総合的にレビューする．**第 2 節**は，その中でも複文を単文化する技術について，**第 3 節**は，意味的に近い知覚連語を検索する技術についてである．**第 4 節**では，近年の進展を代表的に例証するために，専門分野を特に神話，中でもギリシャ神話を例に採って，日本語解析システム「ささゆり」の保持する同義異表現検索機能と日本語の言い換え技術の機能的側面を紹介する．上記の 5 インターフェイスの適用例を示す形式で，検索技術の有効性を示して行きたい．**第 5 節**では，近年，特に M 言語の方法として進展のあった M 言語の大域変数の中間位ノードの修正時における枝葉データのバックアップ方法についても触れておきたい．この方法は，知覚連語辞書の更新時に即時的に知能データを更新することを可能にするので，今後の知覚連語データの効率的蓄積を保証することが期待される．最後の**第 6 節**はまとめと展望に割り当てられる

2. 日本語解析システム「ささゆり」の同義異表現検索の原理（その 1；複文を単文化する技術）

最初に，日本語解析システム「ささゆり」の近年の進展を代表する基礎技術として同義異表現の検索機能と日本語の言い換え技術に関係する 2 機能を総合的にレビューしておきたい．2 機能というのは，複文を単文化して意味解析する機能と同義異表現の知覚連語を検索する機能である．

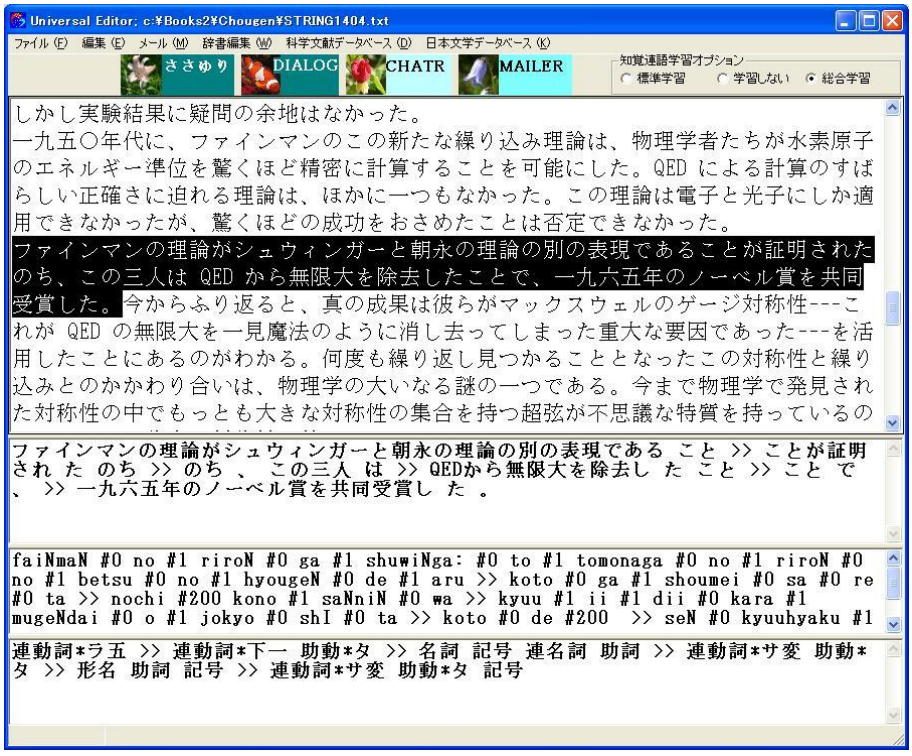


図 1 ユニバーサル・エディター（電子図書館）

この節で，まず，複文を単文化して意味解析する技術について述べ，次の節で同義異表現の知覚連語を検索する機能について述べることにする．

我々の日本語解析システムが開発された動機は，音声ガイドや電子図書館で，視覚障害者が情報機器を利用しやすいようにテキスト文を読み間違いなく読み上げることにあった [1]. 我々のシステムでは，日本語文をテキストとして知覚連語を機械学習するが，機械学習の機会は日本語文を文単位に選択し，選択された文を音声合成装置で読み上げさせるタイミングである．

解析システム「ささゆり」の適用された伝統的アプリケーションであるユニバーサル・エディターを**図 1** に示す．これは，名前の示すとおり，視覚障害の人がテキスト文書を編集するために設計されたエディターであるが，機能面から電子図書館としても活用できる．日本語解析システムは Active X コントロールとして配置されており, Visual M を通して Caché サーバと交信し, M 言語で構築された日本語解析システムの機能を提供する．特に日本語解析に必要なデータの管理について各種のインターフェイスを提供する．ユニバーサル・エディターには, Active X コントロールとして視覚障害のあるユーザーにも一般のユーザーにも利用しやすいファイル操作のダイアログボックスを提供するための DIALOG，音声合成装置の使用を可能にする CHATR，編集された文書を E メールとして送受信する機能を提供する MAILER が配備されている．図では知覚連語学習オプションとして英語部分と日本語部分を総合的に学習するオプションが選択されており，リッチテキストボックスによる編集画面で一文が選択されている．この一文が，システムが判断した CHATR の発音記号と知覚連語の連語範疇列として下段の 2 つのテキストボックスに表示されている．こうした解析結果の表示とともに選択文は音声合成装置 CHATR[1] によって読み上げられる．(CHATR の発音記号は基本的に日本語の発音をヘボン式ローマ字で表現したものであるが，# 記号に引き続く整数でミリ秒単位の休止を与える．#0 は単に音節が切れていることを示すだけの効果しかない．200 ミリ秒以下の休止が知覚的に意味があるかという疑問があるが，#0 や #1 のような記号でも CHATR には微妙なプロソディー変化が生じる)

日本語解析システムは，日本語文を読み上げさせるタイミングで，機械的に学習対象の知覚連語をリストする．学習対象の知覚連語は，知覚連語の形成規則にしたがってリストされる．知覚連語の形成規則は，文の構成要素の数個の文法的範疇を観察して，文法的範疇の並びに対応して知覚連語の候補をピックアップするものであるが，チョムスキーの生成文法に似た構造を持っている．システムは文法的範疇列によって候補を提示するので，候補の中には知覚連語として意味的に不適切なものが含まれるからこれを人の眼で確認して廃棄する．文法的にも意味的にも整合性のあるものが知覚連語としてシステムに記憶される．

我々の日本語解析システムは，既に学習している知覚連語が含まれない初見の文に対しては，単語（形態素）単位で日本語文を分割し，知覚連語の学習が進むにつれて，知覚連語単位で分割するようになる．知覚連語の形成規則は，動詞文による名詞の修飾を知覚連語とみなすことを原則的に禁止しているから，自然に動詞文による修飾文を孤立させるようになる[2,4,10].

結果として，日本語解析システムは複文を修飾文（修飾子）と修飾文を除去した残余（骨格文）とに分解する．修飾文は植物の葉に譬えられ，骨格文は茎に譬えられる．修飾文が修飾している骨格文中の被修飾名詞を接合名詞と呼ぶことにすれば，接合名詞の意味は修飾子との対応関係と骨格文中の知覚連語の構成関係によっ

[1] 音声合成装置 CHATR は ATR 社による波形接続型音声合成装置.

て規定される．このことを図式的に表現すると次のようになる．

(i) 修飾子 ⇔ 接合名詞

(ii) 骨格文（接合名詞を含む; 後続子）

また，例文として“今年も、桜の花が咲く季節になりました”に対して上記の比喩を図によって表現すれば**図 2** のようになる．

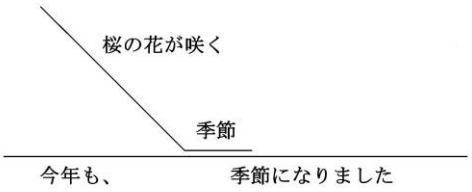


図 2 「ささゆり」の構文理解

上記 (i), (ii) の 2 関係から，接合名詞の意味を限定する原理が最初に示されたのは，2008 年の論文 [2-5] である．ここで，その要旨をまとめると次のようになる．我々の日本語解析システムでは，知覚連語を単位にして意味要素の組が割り当てられる．上述の例に対しては，構成要素である知覚連語，“今年”，“桜の花が咲く”，“季節になります”に現時点で割り振られている意味要素の組は次の通りである．

今年 = [今年/時期/期間]

桜の花が咲く

= [季節/春/時期/桜花/植物/花弁/開花]

季節になります

= [冬/到来/変化/夏/季節/春/時期/秋]

季節 = [冬/夏/季節/春/時期/秋]

Mumps Vol. 24 掲載の論文 [2,3] には，接合名詞の意味は，上述の対応関係 (i), (ii) の双方から規定されることが述べられた．(i) からは，図 2 に示されるように修飾子“桜の花が咲く”と接合名詞“季節”の双方の意味要素の共通部分に限定される．

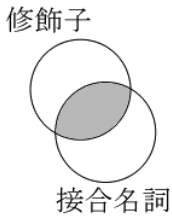


図 3 修飾子と接合名詞の対応関係

(ii) からも，**図 4** に示されるように後続子“季節になります”と接合名詞“季節”の双方の意味要素の共通部分に限定される．

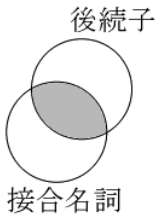


図 4 後続子の形成関係

結果として，接合名詞の意味は 3 つの集合の積集合 (**図 5**) に限定され，

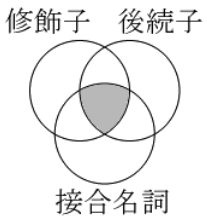


図 5 接合名詞のこの文脈での意味

日本語解析システムは次のような解析結果を表示する．

(1) 桜の花が咲く ⇔ 季節[季節/春/時期]

今年 も 、 >> <季節>(1)になりまし た

ここで (1) は修飾子と接合名詞の対応関係を表現し，<季節>(1) は，後続子のなかで“季節”という接合名詞が (1) の修飾関係で修飾されていることを表現している．修飾関係 (1) の“季節”の後ろの [ ] 内は限定された意味を示している．

こうした，複文を単文化し，被修飾名詞の意味推定する技術が必要とされた分野は，聴覚障

害者のコミュニケーション支援を目指す分野であった．聾者にとって，複雑な構文を持っている文や助詞や形式名詞などの形式的な言葉を含む文などの理解が困難なことが多い．日本語解析システム「ささゆり」の日本語文単文化技術はこうした聾者のコミュニケーション支援の技術として期待されるものである．中でも，それ自体形式的で，修飾関係によってのみ意味が明確になる形式名詞の意味推定の技術は，“分かりやすい手話”を目指す手話通訳の参考としても十分意味のあるものである．

複文を機械的に単文化する技術は，文中の修飾関係を論理的に把握するものであるから，日本語を使用する一般の人々にとっても，文の論理的構造を意識的に捉える機縁を与えてくれるものである．この意味で我々の日本語解析システムの構文判断は示唆的であると言える．その具体的な実例は，以下に述べるこの節末の議論で明らかになると思われる．

日本語解析システム「ささゆり」の構文解析機能は，品詞同定機能や構文判断機能の整合性が進化するにつれて，より正確なものになっていくが，その一方で日本語の構文判断のより本質的な問題点も明確になっていく．この節の最後として，文献で実際に使用された文例を挙げて，こうした問題を整理しておきたい．ここで取り扱う三例文は，ミチオ・カク，ジェニファー・トレイナー共著，久志本克己訳，広瀬立成監修の『アインシュタインを超える---超弦理論が語る宇宙の姿---』[11] に掲載されている文である．

最初に挙げる例文は骨格文に含まれる被修飾名詞の個数が単純に増加したものである．

[例文 1]

夕暮れ時の太陽は水平線に近く、沈みゆく太陽の光はわれわれの目に届くまで水平にやってこなければならないため、比較的多量の大気中を通過する

[構文解析結果]

(1) 沈みゆく ⇔ 太陽の光[夕日/太陽]

(2) われわれの目に届くまで水平にやってこなけれ ば ならない ⇔ ため[根拠/理由]

[骨格文]

夕暮れ時の太陽 は >> 水平線に近く 、 >> <太陽の光>(1) は >> <ため>(2) 、 比較的多量の大気中を通過する

骨格文中“太陽の光”という名詞と“ため”という形式名詞が修飾されているのみで，副詞句の始まり（副助詞“は”）と単文の切れ目に“>>”マークが挿入されている．多くの日本語文はこの形をしており，構文判断に特に問題はない．

次に揚げる例は，修飾文中の名詞が修飾を受けているものである．

[例文 2]

超弦理論が首尾一貫し、あらゆるものを包含する自然の全体像を描くことができるのは、ヴァイオリンの弦があらゆる音色や音調和の規則を一つのものに「統一」できるのとよく似ている

[構文解析結果]

(1) あらゆるものを包含する ⇔ 自然の全体像[万物/包含/基礎/真理]

(2) <自然の全体像>(1)を描く ⇔ こと[事実]

(3) <こと>(2)ができる ⇔ の[事実]

(4) ヴァイオリンの弦があらゆる音色や音調和の規則を一つのものに「統一」できる ⇔ の[事実]

[骨格文]

超弦理論が首尾一貫し 、 >> <の>(3) は 、 >> <の>(4)とよく似ている

骨格文中で修飾を受けているものは二つの形式名詞“の”（(3) と (4)）で，修飾関係 (1) は (2) の修飾文“自然の全体像を描く”の名詞“自然の全体像”を修飾していることを示しており，修飾関係 (2) は (3) の修飾文“ことができる”の形式名詞“こと”を修飾していることを示している．この例文は修飾文中の名詞にさらに修飾文がつく例で，重層的な修飾関係を示しているが，文献には時折見かけられるものである．この例文についても解析システムは正確な判断

をしており，意味解析に支障のあるものではない．

三つ目の例は構文的に句の切れ目が判断しづらい部分を含む例文である．

[例文 3]

このことが、強い力が電磁力に支配されている通常爆薬とは比較にならないほどのエネルギーを生み出せることを、まざまざと見せつけている

この文は“核力”などのことを言う“強い相互作用”という素粒子物理の専門用語を“強い力”と平易に言い換えており，このことばが“ほどのエネルギーを生み出せる”の主語になっているが，この“ほど”という形式名詞を“電磁力に支配されている通常爆薬とは比較にならない”という複文が修飾をしているという難物である．このような文に対し我々のシステムは構文判断が仕切れず，次のような解析結果を提示してしまう．

[構文解析結果]

(1) 強い力 が 電磁力に支配されている ⇔ 通常爆薬[原理/爆発/爆薬]

(2) <通常爆薬>(1)とは比較にならない ⇔ ほどのエネルギー[エネルギー/程度/計量/高エネルギー]

(3) <ほどのエネルギー>(2)を生み出せる ⇔ こと[生成]

[骨格文]

このこと が 、 >><こと>(3) を 、 >> まざまざと見せつけている

その結果“強い力”は，主語として“電磁力に支配されている”につながり，“ほどのエネルギーを生み出せる”という句にはつながらなくなってしまうのである．こうした文は人間が読んでも初見ではとまどいを覚えるものである．この意味で修飾句の割り込みによって主語と述語文が引き離されるような例では我々のシステムは未だ完全ではない．

しかし，もし元の文を判読しやすい表現に改めることが許されるならば，解析システムは正確な判断を下すことが可能である．修正文は次のようなものである．

[例文 3 の修正文]

このことは、強い力が、電磁力に支配されている通常爆薬とは比較にならないほどのエネルギーを生み出せることを、まざまざと見せつけている

“このことが”を“このことは”に修正し，“強い力が”の後に句点を挿入している．この修正文に対する我々のシステムの構文解析は次の通りである．

[構文解析結果]

(1) 電磁力に支配されている ⇔ 通常爆薬[原理/爆発/爆薬]

(2) <通常爆薬>(1)とは比較にならない ⇔ ほどのエネルギー[エネルギー/程度/計量/高エネルギー]

(3) <ほどのエネルギー>(2)を生み出せる ⇔ こと[事実/真実/真理]

[骨格文]

このこと は 、 >> 強い力 が 、 >><こと>(3) を 、 >> まざまざと見せつけている

こうした解析結果は，主語や目的語の後に，それ自体に後続可能な句を挿入して，句の後の名詞を修飾させるような文を書くときには，挿入句の前に句点を打った方が人にも日本語解析システムにも分かりやすい表現になることを物語っていると言える．

3 日本語解析システム「ささゆり」の同義異表現検索の原理（その 2；意味的に近い知覚連語を検索する技術）

2009 年頃，複文を単文化して意味解析する技術に引き続いて 2 つの知覚連語の意味的距離を測定する方法が議論された [6-9]．我々が提案した知覚連語間の意味的距離は次のようなものであった．

二つの知覚連語があって，それぞれの知覚連語が意味要素の組 A, B を持っていたとする．A, B の要素数がそれぞれ n_A 個，n_B 個であり，A, B 共通のものが $n_{A\cap B}$ 個あったとすれ

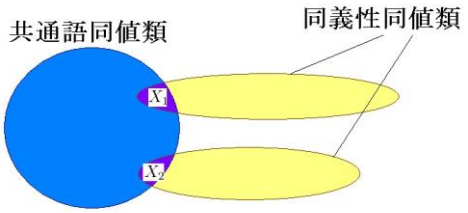
ば，共通性のない意味要素の個数は $n_A + n_B - 2n_{A\cap B}$ である．この要素数が二つの知覚連語の意味的隔たりの程度を表していることは容易に理解できる．これに対して共通の意味要素の個数 $n_{A\cap B}$ 個は二つの知覚連語の意味的共通性の程度を表していることもまた然りである．両者はともに個数を単位にしているので物理的次元は等しい．数学や物理学では物事の特性を示す数量を定義する際にしばしば単位系によらない無次元数として定義することに習えば，知覚連語の間の意味的距離を次の式で定義するのが妥当であるというのが我々の提案であった．

$$d_{AB} = \frac{n_A + n_B - 2n_{A\cap B}}{n_{A\cap B}}.$$

この定義では，A と B が集合として等しい場合に意味的距離が 0 となり，A と B に共通する意味要素がないときに無限大となるという意味でも意味的距離の性質をよく保持していると考えられ，また，この定義は，隔たりの程度と共通性の程度の比であることから，隔たりの程度の大きいものに対して値が大きく，共通性の程度が大きいものに対して値が小さくなるという特性からも知覚連語の間の意味的距離の性質をよく保持していると考えられる．

知覚連語の意味的距離が最初に使用されたのは，**第 2 節**で述べられた，複文で修飾を受けている形式名詞（名詞）に対して意味限定された意味要素の組に対して最も近い内容名詞を検索することであった．こうして形式名詞（名詞）の例文中の意味に最も近い内容名詞が検索されれば，形式名詞（名詞）の代わりに，その内容名詞を使用することで複文は完全に単文化される．第 2 節の例文 “今年も、桜の花が咲く季節になりました”を例にとれば，名詞“季節”の文中の意味要素は［季節/春/時期］であったから，この意味要素の組に最も近い内容名詞“春”が検索される．結果として例文は“桜の花が咲く。今年も、春になりました”のように述べても文意は変わらないことになるのである．こうした単文化は手話通訳が通訳時に直感的に行っている代替表現でもある [3,5].

知覚連語の間の意味的距離を判断基準として，我々の日本語解析システム「ささゆり」の同義異表現の知覚連語を検索する機能はさらに展開される．我々のシステムでは知覚連語を共通語同値類と同義性同値類との 2 種類の同値類によって分類し，日本語の言い換え技術のバックボーンとしている．共通語同値類というのは同じ単語を含んだ知覚連語の集合のことであり，同義性同値類とは使用されている語に関わりなく意味的に近い知覚連語の集合のことである．共通語同値類には，同じ語が使用されていると言っても，使われ方によっては異義的なものが数多く存在する．したがって共通語同値類の個々の要素にはそれと同義の知覚連語の集合，つまり同義性同値類が存在するわけだから，共通語同値類は同義性同値類によって類別されることになる．このような関係性を図示すれば**図 6** のようになる．



X_1, X_2 は共通語同値類に属する知覚連語.

図 6 共通語同値類と同義性同値類

その単語を使用せず同じ意味をもつ知覚連語の集合は，同義異表現の言い換え文の集合でもあるので，こうした知覚連語の検索技術が，難解語を別の言葉で表現する日本語解析システム「ささゆり」の技術の基本的なスキームを与える．この技術はキーワード検索とは異なる新しい検索技術であり，日本語解析システム「ささゆり」の新機能である．

知覚連語の共通語同値類と同義性同値類の相補的活用技術が最初に導入されたのは，聴覚障害のある人のオノマトペの理解を促進すること

が動機であった [6,7,9]. 音声に対する表象性を保持したオノマトペが聾者には, 表象性が無く, 理解しにくいのであるが, 理解の困難さを大きくしているのは, 一つのオノマトペが多様に使用されることにある.

実例を挙げてみると, 例えば, “カンカン”というオノマトペについて, これまで我々の日本語解析システムが知覚連語として学習しているもの, つまり共通語同値類の知覚連語には次のようなものがある.

カンカンと踏切が鳴る
カンカンと鐘を鳴らす
カンカンに怒る
カンカン鳴きはじめる
カンカン鳴っています
カンカン鳴っている
カンカン鳴る
小さな蝉などもカンカン鳴きはじめたりします
小さな蝉などもカンカン鳴きはじめたりする
小さな蝉などもカンカン鳴きはじめる
母親がカンカンに怒る

これらの例では, 同じ“カンカン”というオノマトペが少しずつ異なった意味で使用されている. 最初の例文“カンカンと踏切が鳴る”がどのような意味を持っているのかというときに, 我々のシステムは, 同義性同値類として次のような知覚連語を検索する.

かんかんと踏切が鳴る;連動詞*ラ五
カンカンと踏切が鳴る;連動詞*ラ五
甲高い音;連名詞
甲高い音が鳴る;連動詞*ラ五
甲高い音で鳴る;連動詞*ラ五
踏切が鳴る;連動詞*ラ五
踏切の警報音;連名詞
踏切の警報音が甲高い音で鳴る;連動詞*ラ五

(ここで“;”の後に記されているのは知覚連語範疇である) 表記の違うもの“かんかん”や表現の違うもの“甲高い音”, 意義的に違うもの“踏切の警報音”などをリストしていることに注目されたい. “カンカン鳴っています”という例文は, 宮沢賢治の『風の又三郎』[12] から学習されたものであるが, このような稀少な例文についても, 次のような知覚連語を検索している.

かんかん鳴っています;連助動*マス
かんかん鳴っている;連動詞*上一
かんかん鳴る;連動詞*ラ五
そのこちらを薄い鼠色の雲が速く速く走っています;連助動*マス
そのこちらを薄い鼠色の雲が速く速く走っている;連動詞*上一
カンカン鳴っています;連助動*マス
カンカン鳴っている;連動詞*上一
カンカン鳴る;連動詞*ラ五
甲高い音;連名詞
甲高い音が鳴る;連動詞*ラ五
空がまっ白に光る;連動詞*ラ五
薄い鼠色の雲が速く速く走っています;連助動*マス
薄い鼠色の雲が速く速く走っている;連動詞*上一

これらの知覚連語は風の又三郎の舞台となって岩手県の気象や風土をよく物語っていると思われる.

知覚連語の共通語同値類と同義性同値類の相補的活用技術はオノマトペに対して適用可能なばかりでなく, 一般的に科学や文学の難解語の理解を促進する技術としても有効であり, 物理学用語に適用した場合 [9] や医学用語に適用した場合 [10] の実効性が既に報告されている. こうした技術は, 近年の知識検索技術で多く見られる特定の質問形式の構文に対する解答を用意する種類のものではなく, 純粋に日本語文を構文解析し, 意味解析して, これと同義の表現を行うものであるから, 日本語解析システムの知覚連語学習の進展にともない, 当該分野の広汎な知識を習得する技術として注目される.

我々の検索技術が, 以下に述べるような連鎖的検索を可能にすることにも注目されたい. あ

る用語を用いた表現 a_1 から意味的距離 r_1 を指定して同義検索を行うと半径 r_1 内の同義異表現の知覚連語 a_2 が検索される. a_2 は, a_1 とは意味的に近いと言っても別表現であるから独自の意味要素を保持している. したがって, a_2 を起点とし, 意味的距離 r_2 を半径として再検索をかけると, a_1 とは異なり, a_1 から意味的距離 r_1 の半径内に見つからなかった別の表現 a_3 が見つかるといった具合に, ある経路をたどりながら関連する知識を検索することを可能にするのである.**(図 7)**

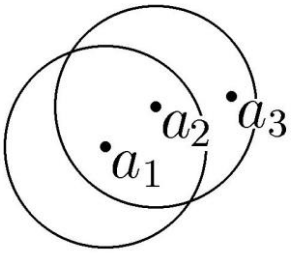


図 7 連鎖的同義検索

物理学の例を挙げて, このような検索法の拡張性を見てみたい. 現在我々のシステムに登録されている知覚連語から“ニュートリノ”という素粒子の名前を起点にして意味的半径 1.5 を臨界として検索をかけると, 代表的なもののみを例示することにして, 次のような表現が見つかる.

W 粒子の交換によって起こる
イタリア語で「小さな中性のもの」を意味する
エンリコ・フェルミが新粒子にニュートリノの名前を付ける
タウ・ニュートリノが発見される
ニュートリノがひどく捉えにくい
ニュートリノと衝突する
ニュートリノの名前を付ける
ニュートリノの存在
ニュートリノの実用的な使い道を考える
ニュートリノの物理
ニュートリノの質量
ニュートリノの質量の有無
ニュートリノの質量の有無を実験で調べる
ニュートリノは非常に貫通力が強い
ニュートリノ望遠鏡
ニュートリノ爆弾
ニュートリノ研究
ニュートリノ線
ミューオン・ニュートリノ
ミューオン自身にも別の相棒「ミューオン・ニュートリノ」がいる
ミューニュートリノ
一九三三年にエンリコ・フェルミが新粒子にニュートリノの名前を付ける
地球さえも簡単に貫通してしまう
天然のニュートリノに質量がある
弱い相互作用をする
素粒子ニュートリノの質量の有無を実験で調べている
電子がニュートリノと衝突する
電子とニュートリノのあいだの力
電気力学と弱い相互作用
非常に貫通力が強い

検索結果は素粒子“ニュートリノ”自身の属性や用途, 関連する素粒子の情報を与えているが, 中でも“弱い相互作用をする”は少し包括的な情報であるから, この知覚連語を起点にしてさらに検索をかける. 臨界半径を 2.0 にして検索すると, 既に検索されているものの他に, 新たに次のようなものが見つかる.

スティーヴン・ワインバーグとアブダス・サラム
ワインバーグ=サラム理論
ワインバーグとサラムの弱い力と電磁力の統一理論にはSU(2)×U(1)ゲージ対称性がある
ワインバーグ・サラム理論
光子を弱い相互作用の W 粒子と統一する弱い力
弱い力と強い力
弱い力と電磁力

弱い力と電磁相互作用
弱い相互作用と強い相互作用
弱い相互作用と電磁相互作用
弱い相互作用と電磁相互作用を統一する
強い力と弱い力と電磁相互作用
強い相互作用と弱い相互作用
繰り込みと弱い相互作用
電気力学と弱い相互作用
電気力学と弱い相互作用と強い相互作用
電磁相互作用と弱い相互作用

検索結果には，“ワインバーグとサラムの弱い力と電磁力の統一理論”についての知見がいくつもリストされており，こうした統一理論の中心に“SU(2)×U(1)ゲージ対称性”があること，それと“光子を弱い相互作用の W 粒子と統一する”こととが関連すること，さらには“電磁相互作用”，“弱い相互作用”と強い相互作用”の三つを対象とした統一理論を示唆する情報が検索されている．このような情報の系統的な系列は，専門分野の知識の正確な認識を促進するものと考えられる．

4. 日本語解析システム「ささゆり」の同義異表現検索技術の神話ジャンルへの拡張とシステムの操作性を象徴するインターフェイス群

第 2 節，**第 3 節**では，日本語解析システム「ささゆり」の構文解析機能や複文を単文化する機能，難解な用語を含む日本語の表現を，難解な用語を含まない同義異表現の文や句で代替して理解させる言い換え技術について一般的に述べた．この節ではこうした技術を新しく神話のジャンルへ拡張していくのと平行する形で近年のシステムの発展を紹介して行きたい．方法論的には，日本語解析システム「ささゆり」の知識データを処理するインターフェイス群の機能性と近年の改良点を紹介することによって，その有効性が広汎な発展性を保持していることを例証していくことになる．

(1) 知覚連語辞書

第 2 節では，日本語解析システム「ささゆり」がユニバーサル・エディター (電子図書館) などのアプリケーションでテキスト文を読み上げるタイミングで知覚連語の候補を抽出することに触れたが，こうした知覚連語に関連する知識データを最も基礎的に操作するインターフェイスは，**図 8** に示される知覚連語辞書であり，この辞書は知覚連語に関する基本情報を登録・削除・変更し，知覚連語の機械学習結果を評価して登録もしくは廃棄する機能を持っている．この辞書で編集されるのは知覚連語の基礎データを記憶している大域変数 ^NWDIC であり，この大域変数は 3 階層の添字，知覚連語，連語範疇，特性を保持している．(特性と呼ばれる添字もしくはノードは，漢字の読みや，文法範疇に多義性がある場合に，これらの異種判断を区別するためのものであるが，ディフォルトでは 1 が採用される) 大域変数 ^NWDIC の値は発音記号，補足 (音便や名詞形の有無などの情報)，単語構成，品詞構成，手話記号 (ビデオクリップに対応)をセパレータをはさんで接合したものである．

図 8 は，学習結果を表示した場合の図であり，例文として，ギリシャ・ローマ神話の古典『変身物語』[13] の文章をテキストとして初見で機械学習させた後，[学習結果] ボタンを押して知覚連語の候補を表示させた場合の図である．テキストとした文章は次のようなものである．

> デイアネイラという名前を、きっとお聞きになったことがあるでしょう。そう、デイアネイラというのは、ひとむかしまえ、世にも美しい乙女だったのです。大勢の求婚者たちの望みのまとでもあり、彼らのあいだに競争心をまきおこしていました。

表示内容は，編集可能で，左上部 5 つのボタンは，編集時にグリッドを行単位で編集するためのものである．切り取り (CutR)，コピー (CpyR)，貼り付け (PstR)，削除 (DelR)，挿入 (InsR) を行単位で行うことができる．

図 8 知覚連語辞書

文法繋がりを判断基準として機械的にリストされる知覚連語の候補には，意味的統合体として不適切なものがしばしば含まれる．例の場合，“あいだに競争心”，“きっとお聞きになる”の［特性］が 2 のものや“望みのまとでもある”などには，構成要素分解の不適切性もしくは単語間の意味的規定関係の不適切性や不充分さが観察される．こうした不適切な候補を行削除（CutR）ボタンを押すことによって削除した後，［新規登録］ボタンを押すことで知覚連語の候補が正式に知覚連語として辞書データに登録される．

既に登録されている知覚連語辞書の検索を行うには，右上のオプションを設定し，検索語やその他の情報を入力して（特定しないものは空白でよい），［連語検索］ボタンを押すことで辞書に登録されている知覚連語のリストが表示される．登録内容に変更を加える場合は，学習結果の場合と同様に編集した後，［新規登録］ボタンもしくは［更新登録］ボタンを押すことで辞書データ任意の編集が完了する．

知覚連語辞書の本来的役割は，編集結果を日本語解析システム「ささゆり」の人工知能のための各種データに反映させること，つまり日本語解析システム「ささゆり」が知能の基盤とするいくつかの大域変数（代表例として知覚連語の構成文字に相似した階層構造を保持する大域変数 ^NWTREE などがある．これから先の議論では，これらの大域変数群を知能データと呼ぶことにする）の登録・修正・削除を行うことである．この手続きは，高次の階層構造をもつ大域変数の中間ノードの削除を伴うため，旧来の方法では，更新を完全に行うときには，操作に慎重を期して，大域変数 ^NWDIC をソート順に総てを手繰る形式で行われていた． このためのボタンが［知能更新］ボタンである．新規に追加された辞書データに対する知能データの更新（大域変数の中間ノードの削除は無い）は［知能暫定］ボタンによって行われ，即時対応性は改良されていたが，暫定的更新には不完全さを免れ得なかった．今般，［新規登録］ボタン，［更新登録］ボタンで，知覚連語辞書の更新を知能データの更新に即時的対応させる機能が新たに考案されたが，この新機能について

は，木構造を保持する階層型データベースの中間ノードの即時的編集技術を含んでおり，汎用的な活用が期待されるので，節を改めて**第 5 節**で述べたい．

(2) 知覚連語-意味要素相関辞書（概念辞書）

日本語解析システム「ささゆり」の意味解析の基盤を与えるのは，知覚連語と意味要素との相関関係であり，この情報に関する基本的データは大域変数 ^NCDIC が記憶している．この大域変数を定義・編集するインターフェイスは知覚連語-意味要素相関辞書（概念辞書とも呼ぶ）であり，**図 9** に例示されるものである．**図 9** では，検索語として“ヘラクレス”が指定され検索オプションとして［含む］が選択された後，［概念検索］ボタンが押されて，“ヘラクレス”という記述を含む知覚連語に割り当てられている意味要素のセット（“/”をセパレータとして区切られている）をデータとして表示されている．

画面左上に行削除（CutR）ボタンを筆頭に並ぶ 5 個のボタンの働きはグリッド編集に関するものであり，各ボタンは知覚連語辞書のものと同様の効用を保持する．表示例について既に言及した［概念検索］ボタンは概念辞書データ（大域変数 ^NCDIC）の検索を実行するものである．検索語，連語範疇，特性などに与えられたテキスト項目の入力事項に従って条件に匹敵する大域変数 ^NCDIC の情報をグリッドに表示するようになっている．基本的には，表示された内容を画面上で編集し，［更新登録］ボタンを押すことで，大域変数 ^NCDIC の情報が更新される．

大域変数 ^NCDIC は知覚連語辞書のデータ（大域変数 ^NWDIC）と双対的に定義されるものであるが，両者はいずれも 3 階層で，知覚連語，連語範疇，特性の 3 添字を持っている．この両者の関連づけを行うボタンが［連語検索］ボタンである．このボタンは上記と同様の入力事項に応じて知覚連語辞書の大域変数 ^NWDIC を検索し，その添字に応じて対応する概念辞書の大域変数 ^NCDIC の記憶している意味要素のリストをグリッドに表示させる．つまり［連語検索］ボタンを押すと大域変数 ^NWDIC を検索し，これに対応する大域変数 ^NCDIC のデータの意味要素のリストを画面に表示される．表示内容に画面上で編集を加えて，［新規登録］，もしくは［更新登録］のいずれかのボタンを押すと大域変数 ^NCDIC のデータが編集されるという仕組みである．

画面最上段のテキストボックス［一括補意の意味要素］は検索語に共通の意味要素の追加が想定される場合の意味要素一斉追加編集機能を与えるためのものである．テキストボックスに必要な意味要素を“/”区切りで書き込み，［一括補意］ボタンを押すことで意味要素のリストに一括追加される．

形式名詞，接合名詞の意味推定や与えられた意味要素の組に意味的距離が最も近い知覚連語の検索では知覚連語と意味要素を一対一に対応づける 2 階層の大域変数 ^NWCSAMP とその逆対応を定義する大域変数 ^NWCISAMP が検索効率の向上に重要な役割を担う［10］．こうした大域変数の登録には今までは，^NCDIC の編集に対する応答時間を懸念して，^NCDIC の編集とは独立したボタン（［C-S 相関］ボタン）を用意していたが，今般実験的に，先述の［新規登録］，［更新登録］の 2 ボタンの操作に即時対応するように設計変更がなされた．実験結果として，2 ボタンに対するデータ更新の必要時間に，作業の流れを支障するほどの遅延は観察されなかった．このため，最終的に［C-S 相関］ボタンが概念辞書から除去されることになった．

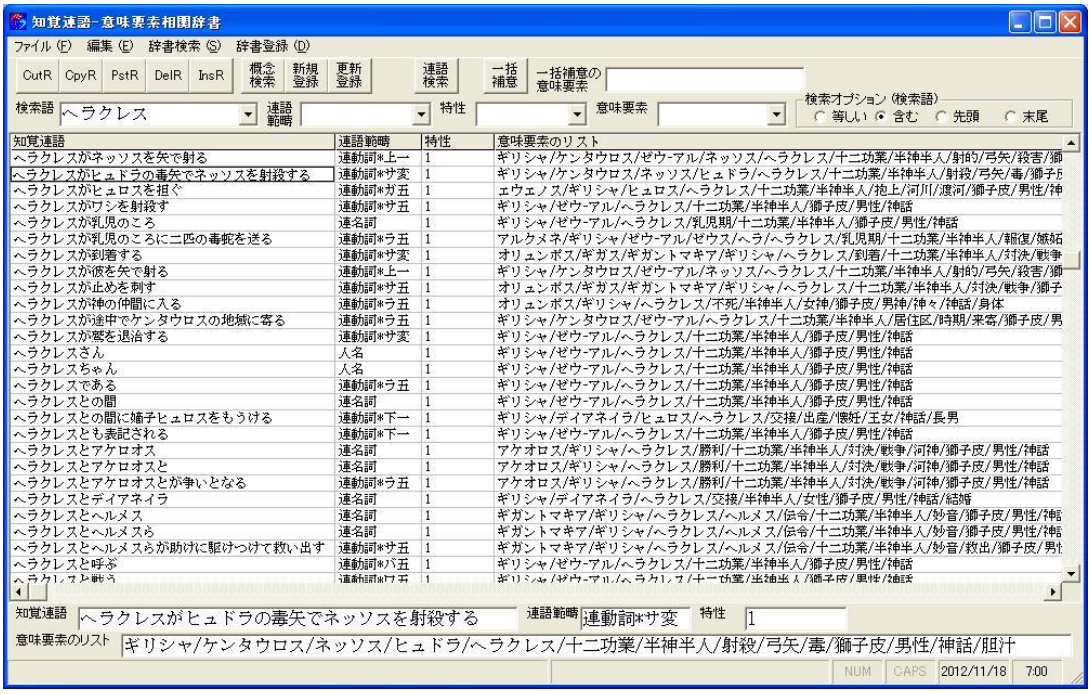


図 9 知覚連語-意味要素相関辞書（概念辞書）

(3) 専門用語辞書

特定の専門分野を指定して，ある用語がその分野の専門用語かどうかを判断するには，専門用語辞書を用意して，当該の用語が専門用語辞書に記載されているかどうかで判定するのが有効である．我々の専門用語辞書は**図 10** に示されるようなインターフェイスを持っている．

表示例では，分野として“神話”を選択し，用語として“アフロディテ”をテキストとして記入し，[検索] ボタンを押すことで大域変数 ^NWTTDIC（分野，用語，品詞，番号を添字とする 4 階層データ；用語の解説，その他の情報をデータ値とする）の内容を表示している．

グリッドの編集機能は，知覚連語辞書や概念辞書のものと同様である．グリッドの各情報を編集して [新規登録] もしくは [更新登録] のボタンを押すことで大域変数 ^NWTTDIC が登録・更新・削除される．

知覚連語に専門用語が文字列として含まれることと，専門用語が構成要素として知覚連語に含まれることとは必ずしも同じことではないが，我々が「専門用語が知覚連語に含まれる」あるいは「専門用語が知覚連語と相関をもつ」と表現する場合は，後者の意味での判断である．この意味で専門用語が使用されている知覚連語の集合が**第 3 節**でのべた共通語同値類である．

共通語同値類を高速に検索するには知覚連語と専門用語を一対一に対応させる大域変数と，一対一に逆対応させる大域変数が存在するのが機能的であるが，我々はこれらの大域変数を，^NWCWAMP と ^NWCWIAMP という名前で定義している．

専門用語辞書に新しい用語が追加された結果を知覚連語全般に反映させる手続きは [T-PC 相関] ボタンを押すことで実行されるが，この作業は知覚連語全体をたどる必要があるので，相当の時間が必要である．

専門用語辞書に新規追加が無くて新しい知覚連語が定義されたときに専門用語辞書を参照して，専門用語が構成要素として含まれるかどうかの判断はほぼ瞬時に行われるので，この手続きは (1) で述べた知覚連語辞書の [新規登録], [更新登録] の 2 ボタンに含められた．

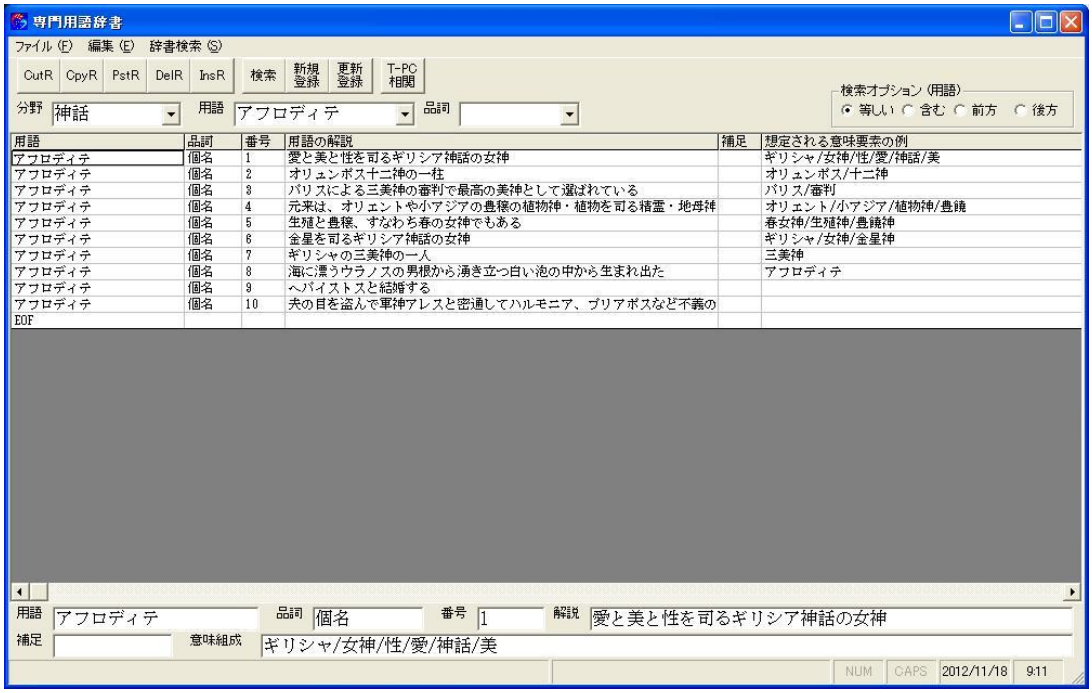


図 10 専門用語辞書

知覚連語辞書の用語の解説には専門用語を含まないで専門用語を解説する文が多いので，これをテキストにして知覚連語学習を行わせれば同義性同値類の要素の一つを与えることになる．しかし，これはあくまでも同義表現の一つの例を与えるのであって，同義性同値類検索の真の対象は，多くの文献で機に応じて叙述された様々な表現であることを意にとめておかなければならない．我々のシステムにとって専門用語辞書の解説記述そのものはあくまでも補助的なものでしかない．

(4) 共通の用語を含む知覚連語の検索

これまでの 3 インターフェイスが知覚連語の共通語同値類，同義性同値類をデータベースとして定義するための基礎的なものであるのに対し，この項と次の項は二つの同値類を利用し活用するためのインターフェイスである．

まずは，専門用語が如何に使用されているのかという専門用語の使用例を検索する共通語同値類検索のインターフェイスである（**図 11**）．**図 11** では，専門用語で検索対象の言葉“アポロン”を検索している．

“アポロン”というギリシャ神話の神の名前を用いた表現は，日本語解析システム「ささゆり」によって数多く機械学習されているが，知覚連語として登録されるときに知覚連語の構成要素として“アポロン”という用語を含むものは **(3)** でのべた ^NWCWIAMP によって用語“アポロン”と関連づけられているので，[検索]ボタンを押す操作に応じて，ほぼ一瞬で検索される．条件を満たす知覚連語をさらに限定するのが［補助キー］であり，**図 11** では補助キーとして“太陽”が入力されているので“アポロン”が含まれ，尚かつ“太陽”も含まれる知覚連語に限定されて表示されている．

図 11 共通の用語を含む知覚連語の検索

さらに各知覚連語に対応する意味要素のリストが概念辞書から検索されてグリッド第 4 列に表示されている．このグリッドは概念辞書と同様の編集機能を持ち，意味要素のリストに追加記入を行ってから［追加登録］ボタンを押すことで追加された意味要素が概念辞書のデータに追加登録される．したがって，このインターフェイスは，専門用語がどのような使われ方をしているのかを検索するのみならず，専門用語と補助キーで知覚連語の意味の同一性を絞り込んだ上で意味要素を一括して補足する目的で使用することが可能である．

このインターフェイスの本来の目的に立ち返ってみれば，用語“アポロン”の使われ方の検索内容からアポロンが“太陽の神”，“太陽神”であることや“黄金の弓”を持っていること，“月の女神アルテミス”と双子の兄妹であること，息子に“オルフェウス”を持つことなどの知識が取得できることが観察される．

(5) 同じ意味を持つ言い換え表現の検索

いよいよ難解な日本語の言い換えを検索するインターフェイスである．**図 12** に示すのは同義性同値類検索を行うためのものである．

このインターフェイスでは，専門分野を選択し，課題文を書き込み，課題文から意味的距離がどの程度の範囲ものまで検索するかという境界距離（境界半径と言っても良い；意味的距離がこの値より小さい知覚連語を検索するという目安）を設定して，［同義検索］ボタンを押すと課題文について，含まれる専門用語，課題文の構文，課題文が保持する意味要素が判断され，それぞれ［用語］欄，［構文］欄，［意味要素］欄に表示される．その上で，課題文と意味的距離が境界距離内にある知覚連語が［同義文］欄に表示される．［同義文］欄に表示される内容は知覚連語，連語範疇，課題文との意味的距離の 3 点セットのリストである．

図 12 では，課題文として，専門用語が含まれない，“海の泡から生まれた”という単文が記入されているが，その意味要素として｛アフロディテ/オリュンポス/ギリシャ/ビーナス/女神/性/愛/泡/神話/美/美神/誕生/豊饒/金星｝が検出され，［同義文］欄に，これまで日本語解析システム「ささゆり」が機械学習してきた知覚連語がリストされている．

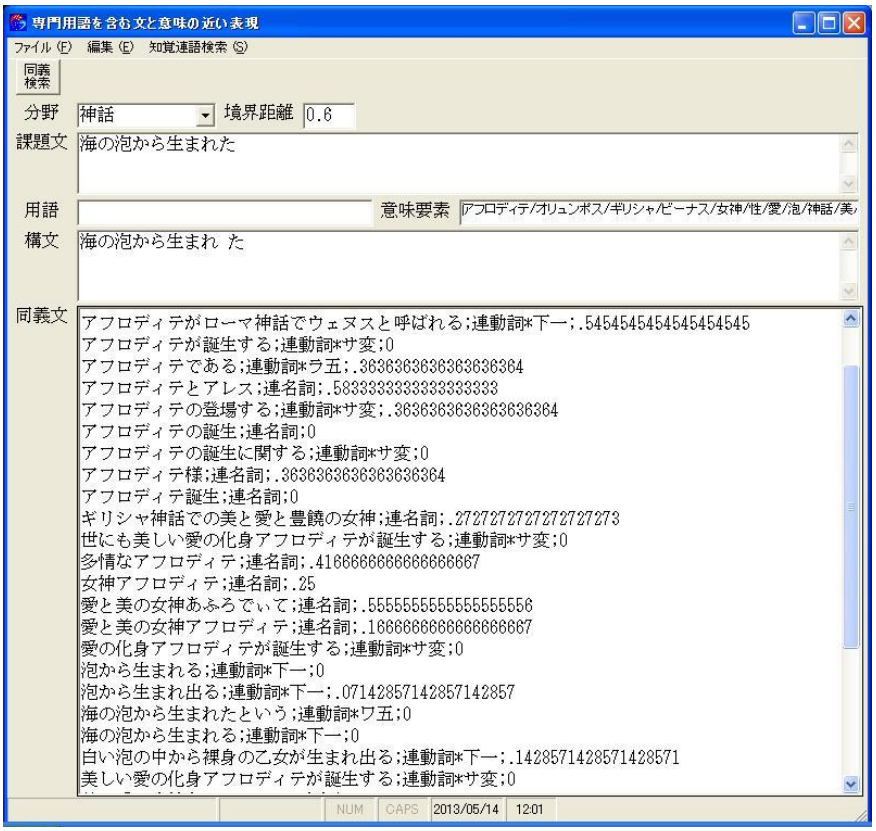


図 12 意味の近い表現の検索（単文の場合）

検索結果の中には，キーワードとして含まれていなかった“アフロディテ”を含む知覚連語が数多く見られ，“アフロディテがローマ神話でウェヌスと呼ばれる”，“ギリシャ神話での美と愛と豊饒の女神”といった，ギリシャ神話からローマ神話への伝承の知識や，“愛と美の女神アフロディテ”のような“アフロディテ”の本質を表現するもの，また少し文学的もしくは芸術的表現“白い泡の中から裸身の乙女が生まれ出る”，“美しい愛の化身アフロディテが誕生する”，“アフロディテ”の交際相手を示唆する“アフロディテとアレス”のようなものまで含まれている．これらの情報は“アフロディテ”の多くを解説している．

こうした広汎な情報は日本語解析システム「ささゆり」がテキストとして学習する文献が多岐にわたるにつれて豊富になり，学問の本質的考察の上からも文学的な表現のモデルとしても有益なヒントを提供してくれることが期待される．このことは，こうした知識収集のあり方の本質的な性格を提示していると思われる．

図 13 は，**図 12** と同じインターフェイスで，課題文として“父神ゼウスを心から愛したために純潔の化身として崇められた”という複文が記入されたものである．このインターフェイスで，日本語解析システムの構文解析機能が発揮される．[構文]欄には“ため”という形式名詞を接合名詞とする構文解析結果

(1) 父神ゼウスを心から愛し た ⇔ ため [原因/根拠/理由;根拠]

<ため>(1)に純潔の化身として崇められた

が表示され，“ため”の意味として“根拠”という内容語が想定されている．課題文には“ゼウス”という専門用語が含まれているが，これが[用語]欄に検出され，[同義文]欄には課題文の意味内容が示唆する女神“アテナ”についての日本語解析システム「ささゆり」がこれまで知覚連語として学習してきた叙述がリストされている．“オリュンポスの主神ゼウスとティ

タン族の娘メティスの子”“処女神のアテナ”，“戦いと芸術の女神アテナ”，“知恵と戦いの神”，“知恵と戦争の女神アテナ”など，直接言及のなかった知識内容を検索していることに注目されたい．

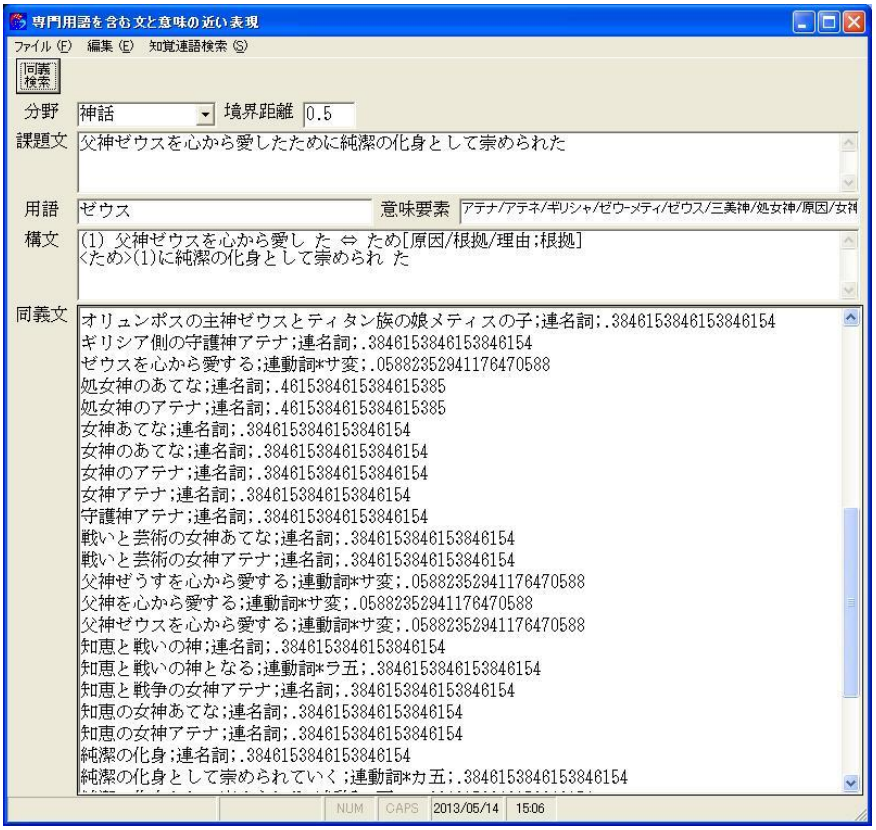


図 13 意味の近い表現の検索（複文の場合）

ここに示されたように，我々のシステムでは，指定された分野の用語が含まれるか含まれないかにかかわらず，また単文であるか複文であるかにかかわらず，与えられた文の構文解析・意味解析を行って，当該の文に意味的に近い知覚連語を検索する．こうした機能は，これまでにも物理の分野の例 [8] や医学の分野の例 [10] が示されてきており，この報告にも第 2 節，第 3 節で，オノマトペや物理に関する解析例がいくつも示されてきた．今，この項でギリシャ神話の分野での適用例が紹介されたのである．例に挙げられたものは一部であるが，現時点でギリシャ神話に登場する大半の人物について，同様の検索が可能になったことを報告しておきたい．

ここで，さらに我々の同義検索の有効性を模式的に示すために**第 3 節**で述べた連鎖的同義検索の，ギリシャ神話についての実例を示しておきたい．例としてよく知られている“ヘラクレス”というギリシャ神話の英雄について検索を実行してみる．“ヘラクレス”から意味的半径 1.4 以内に入る知覚連語をリストしてみると次のようなものが挙がってくる．(但し，スペースを節約するために，下記リストの知覚連語に内包されるという意味で独立的でないものはリストから省略した)

12 の功業を行う
「ヘラの栄光」を意味する
そこにヘラクレスが到着する
のちにオリュンポスの神に連なる
アケロオスとヘラクレス
エウエノスの急流にさしかかる
エリュマントスの猪を生け捕りにする
オイタの山に燃える
オイテ山の頂に火葬の薪を組ませる
カリュドン王オイネウスの親戚の一人を殺してしまう
ケンタウロスと戦って討伐する
ゼウスとアルクメネの子
ゼウスとアルクメーネーの子
ゼウスの子ヘラクレスである
ティリュンスの英雄ヘラクレスが現身の部分を捨てる
デイアネイラと息子ヒュロスとともトラキアへと旅立つ
ネッソスの「毒薬」がしだいに効いてくる
ネッソスらと戦って討伐する
ネメアの大獅子を締め殺す
ネメアの獅子の皮をひろげる
ヒュドラーを退治する
ヘスペリデスの園からリンゴをとってくる
ヘラクレスがこれを着る
ヘラクレスがネッソスを矢で射る
ヘラクレスがヒュロスを担ぐ
ヘラクレスがワシを射殺す
ヘラクレスが乳児のころに二匹の毒蛇を送る
ヘラクレスが彼を矢で射る
ヘラクレスが止めを刺す
ヘラクレスが途中でケンタウロスの地域に寄る
ヘラクレスが鷲を退治する
ヘラクレスである
ヘラクレスとの間
ヘラクレスとも表記される
ヘラクレスとアケロオスとが争いとなる
ヘラクレスと呼ぶ
ヘラクレスと戦う
ヘラクレスと死の衣
ヘラクレスにその嫌疑がかけられる
ヘラクレスの一家惨殺
ヘラクレスの下に嫁ぐ
ヘラクレスの冒険
ヘラクレスの助けが必要とする
ヘラクレスの寝所
ヘラクレスの柱
ヘラクレスの気を狂わせる
ヘラクレスの継父
ヘラクレスの誕生
ヘラクレスの選択
ヘラクレスを天上界へ連れ去る
ヘラクレスを捜しにいかせる
ペルセウスの子孫である
ポロスの酒甕を開けてしまう
ユピテルを父とする
多くの半神半人の英雄の中で最大の存在である
幼名をアルケイデスという
強靭な肉体のへーラクレース達を見る
彼の妹デイアネイラを妻とする;連動詞*サ変
怪力のヘラクレス
息子で英雄の誉れ高いヘラクレス
新妻デイアネイラをともなう
森深いオイタの山じゅうに叫び声をひびかせる
死の衣をはぎとろうとする
水蛇ヒュドラを退治する
河神アケロオスとの対決で勝利する
河神アケロオスと対決して勝利する
父オイネウスの意向で婿候補になる
父親の市ティリュンスへ帰る
獅子の皮を着る
獅子の皮を頭からかぶる
獅子座にしたという
祖父の名のままアルカイオスとも呼ばれている

禁断の酒甕を開けてしまう
英雄ヘラクレスが神の仲間に入る
酒甕を無理矢理開けてしまう
雪の中で網を投げる

大半は“ヘラクレス”の生涯の事件（出生，功業，末路に関するもの）を象徴するであるが，この検索結果からは各事件の詳細は見えてこない．

そこで，例えば，出生にまつわる情報の表現である知覚連語“ゼウスの子ヘラクレスである”から意味的半径 1.4 以内にあって，“ヘラクレス”から意味的半径 1.4 の外にあるものをリストしてみると，次のようなものが挙げられる．

もう一人がゼウスの子ヘラクレスである
アルクメネに接近する
アルクメネに生ませようと考える
アルクメネに目をつける
アルクメネの子供
アルクメネをその母に選ぶ
アルクメーネーの子
アルクメーネーを見初める
アンピトリュオンに変身して近づく
アンピトリュオンの子イピクレスとゼウスの子ヘラクレスを双子で出産する
ゼウスがアンピトリュオンに姿を変えて近づく
ゼウスが人間の女性と浮気をする
ゼウスと人間の女性アルクメネの子供
息子で英雄の誉れ高いヘラクレスの助けを必要とする

リストされたものは“ヘラクレス”の出生についての情報，つまり「ヘラクレスは，ゼウスがアンピュトリオンの妻である人間のアルクメネに，夫に変身して接近し妊娠させた子供である」ということを如実に示すものである．

もう一つ“ヘラクレス”の末路についての情報であろうと推測される知覚連語“死の衣をはぎとろうとする”から意味的半径 1.4 以内にあって，“ヘラクレス”から意味的半径 1.4 の外側にあるものをリストしてみると，次のようなものが挙げられる．

デイアネイラがネッソスの媚薬を密かに袋に入れて携帯する
ネッソスの「媚薬」を肌衣に塗って持たせる
ネッソスの「毒薬」がしだいに効いてくる
ネッソスの媚薬
ネッソスの甘言を信じる
ネッソスの秘薬
ネッソスの血がしみこむ
ネッソスの血に浸す
ネッソスの血の毒による
ネッソスの血を塗る
ネッソスの血を採っておく
臨終際のネッソスの甘言
臨終際のネッソスの甘言を信じる

これらの知覚連語は，「ヘラクレスの死の原因が，妻デイアネイラがネッソス（ヘラクレスが矢で射たケンタウルス）の甘言にだまされて，媚薬と信じてヘラクレスの肌着に塗って届けた毒薬にある」という情報を認識させるものである．(“ヘラクレス”の検索結果に既にあるように「ヘラクレスがオリュンポスの神に引き入れられた」ことは補足しておかなければならない)

以上のように連鎖的同義検索は，様々に表現された知覚連語の中にある系統性や関連性を観察することを容易にし，神話に登場する神々や人物たちの理解を大きく促進することが期待される．このような用語理解の基礎になっているものは多くの著作やインターネットの記事をテキストにして機械学習された知覚連語の集合と，この節で述べられた 5 つのインターフェイスを通して知覚連語に割りふられた意味要素の集合である．言い換え技術の基礎を為すこのようなデータの完全性を問うことはあまり意味を持たないと考えられる．なぜなら知識のソースと考えられるもとのテキストに際限はなく，たとえジャンルを限定したとしても，今後も増え続ける性質のものだからである．重要なことは，このようなシステムで供給される知識は，テキストの収集とテキストからの知覚連語の機械学習，知覚連語に意味要素を割りふっていく地道

な努力によって，次第に増強され，対応可能な知識の範囲と供給される知識の正確さが際限なく進展していくことである．

以上で，難解な日本語を同義異表現の知覚連語を検索し，直接言及のなかった別の表現を提示することによって知識内容を補完する技術の近年の進展の例証が完了したわけであるが，この節を締め括るに当たって，同義検索技術が進展するにつれ，日本語の表記の混乱が引きおこす自然言語としての日本語のあり方について，懸念される一つの現象について述べておきたい．それは外来語の表記法の問題である．例えば“アフロディテ”という女神の名前についても長母音を強調した“アフロディーテー”や発音のモデルを採った国や地方，時代の差違を強調した“アプロディテ”，“アプロディーテー”，“アプロディタ”等々，同一の概念を指し示す表記に広汎な多様性が表れている．我々の同義検索ではこれらは同じ概念を表現しているものとして検索されるが，知覚連語を蓄積するデータベースには，この多様性を収納するだけで大きな負担容量を要求することになる．その意味では日本語解析システムにこうした表現の差違を同一視して標準的な表現をデータベースに保存し，表記法の差違にはインターフェイスレベルで対応するような機能が必要なのかも知れない．こうした機能のプログラムも M 言語の文字列処理にとってはそれほど困難なことではないと思われる．

5. 日本語解析システム「ささゆり」の基礎となる階層型データの即時的修正の方法

第 4 節で，知覚連語辞書の更新にともなって人工知能の基礎を与えるいくつかの大域変数(知能データ) を更新する方法が，旧来と異なって，即時的に対応する方式に変更されたことについて触れた．この節では，長期にわたって即時的対応がなされなかったことに対する弁明と，今般即時的対応が可能になった着想とについて述べてみたい．これは， M 言語というプログラミング言語に固有のものであって，通常のプログラミング言語には見られないプログラミング法についての紹介でもある．

まず，最初に，弁明から述べると，知覚連語辞書の更新情報を知能データへ即時的対応させることに筆者が躊躇していた最大の理由は，知能データの中に知覚連語の構成文字階層構造を保持する大域変数 ^NWTREE が存在するということにあった．知覚連語の構成文字階層構造というのは**図 14** に示すような知覚連語を構成する各文字を階層構造のノードとするデータの構造であり，これが日本語文を知覚連語に切断していく基本的なアルゴリズム [1] を与えている．

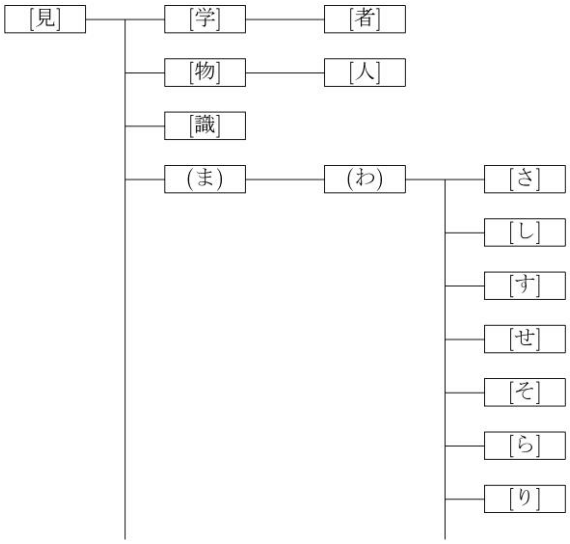


図 14 知覚連語の構成文字階層構造 [1]
見で始まる知覚連語の場合; [ ] のノードにはデータがあり，( ) のノードにはデータが無い．

知覚連語辞書のデータ削除に対する知能データへの即時的対応は，複雑な木構造を保持する大域変数 ^NWTREE の中間ノードの削除を要求するわけである．よく知られているように，Kill コマンドによる大域変数の中間ノードの削除は，そのノードの下位の枝葉ノード総ての削除を意味するのであるから，この枝葉ノードのバックアップを如何にするのか？そしてその実行にどのくらいの時間がかかるのか？バック

アップのルーチンが大変複雑なものになるのではないか？こうした懸念要素のため筆者は大域変数の中間ノードの削除を避けて，知覚連語辞書の削除を伴わない修正もしくは追加データに対してのみ即時的に対応する機能を［知能暫定］ボタンに託していたわけである．知覚連語辞書のデータ削除に伴う知能データの修正には，知能データを知覚連語辞書に従って全く新しく定義し直す方式を取っていたわけで，［知能更新］ボタンがこれに対応していたのである．しかし，知能更新には知覚連語辞書の 91 万をこえる大域変数を総なめするだけの時間が必要であり，この時間が知覚連語辞書の肥大化に伴って増大する傾向にあったことも否めない．

このようなまどろっこしい状況を克服するために，未確認数の枝葉を持つ大域変数の枝葉ノードのバックアップに取り組むことにした．未確認数の添字列を含むデータの存在確認と保護保存を一般的に表現するプログラミングの記述方式を M 言語は内包していたのである．つまりプログラムの叙述を文字列として変数に記憶させ，これを Xecute コマンドで実行させることが M 言語では可能である．したがって，削除するノードのノード指定の情報を ""S1","S2",･･･,"SN"" のような文字列として与え，これをもとに，下位のノードを手繰っていくコマンドを文字列として構成すれば，Xecute コマンドによってそれを実行に移すことが可能である．このような思案のもとに枝葉ノードの保護保存のプログラムの一例が構成された．懸念していたものとは異なり，構成されたバックアップ・ルーチンは非常に簡単なものであった．バックアップに必要な時間についても，ルーチンを日本語解析システムに組み込んで実行した結果，通常の知覚連語辞書の編集に支障をきたすような遅延は観察されなかった．

しかしながら、こうした努力の後で M 言語の拡張型コマンドの存在が指摘された．“Z”の文字ではじまる拡張型コマンド（“Z-命令”）に Kill コマンドに対応した，ZKill コマンドが存在し，これを使用すれば枝葉ノードを保持したまま当該ノードの削除が可能で，名前間接演算の形式（@ を冠する変数名）を用いる方式が Xecute コマンドを用いる方式より機能的であることが示唆された。

結果として，これまで Kill コマンドや Xecute コマンドを用いて記述してきたルーチンを ZKill コマンドと名前間接演算の形式を用いることで，知覚連語辞書更新に対する知覚連語の構成文字階層構造への即時対応性は保証され，［知能更新］ボタンや［知能暫定］ボタンは実質的に不要になった．

6 まとめと展望

M 言語の階層型データをアルゴリズムに組み込んだ日本語解析システム「ささゆり」による同義異表現検索機能と難解な日本語の言い換え技術の近年の進展が述べられた．

第 2 節では，日本語解析システム「ささゆり」の同義検索の基礎を与える機能のうちで特に，複文を単文化して意味解析する機能がレビューされた．

我々の日本語解析システムでは意味解析の基礎に知覚連語を置く．知覚連語は文法的繋がりを基礎に機械的にテキストから抽出される．知覚連語によって日本語文を切断する結果として，システムは複文を単文に分解する機能を保持する．システムは，複文を修飾子（修飾文）と接合名詞（被修飾名詞）の対応関係と骨格文（文の修飾文を除去した残余部分）とに分けて接合名詞の意味推定を行う．接合名詞が形式名詞の場合も同様である．形式名詞の意味を推定し，内容名詞に置きかえることは難解語の言い換え技術の第一の手法である．

第 2 節の後半では，物理学の解説・翻訳で実際に使われた複文を例にして，我々の日本語解析システムが多様な複文に対応できることが示された．骨格文に含まれる複数の名詞のそれぞれに修飾子がつく例や修飾子に含まれる名詞にさらに修飾子がつく例などに対して日本語解析システム「ささゆり」が正確に構文判断できることが示される一方で，修飾子に先行する句が

修飾子に後続する句の主語になっているような複文では我々のシステムでも判断を誤ることがあることについても触れられた．しかし，日本語解析システムにとって判断が困難な文は人間にとっても判断が難しいものであること，このような，いわゆる分かりにくい構文を避けるためには，修飾子と直接主語述語の関係にない先行詞と修飾子の間に読点を打つことが人間にもコンピュータにも分かりやすい文を作成する方法であることも事実である．

第 3 節では，同じく同義検索の基礎を与える機能のうちで特に，同義異表現の知覚連語を検索する機能についてレビューされた．

日本語解析システム「ささゆり」では知覚連語を単位に意味要素を割り振っていく．したがって知覚連語が意味空間の基礎を張る．難解語を，分野を分けて考えると専門用語の集合と考えられる．知覚連語は専門用語を含むものと専門用語を含まないものに分類されるが，特定の用語を含む集合を，共通の用語を含むという意味で，共通語同値類と呼ぶ．特定の用語が含まれていたとしても用語の使い方によって知覚連語の意味は異なることが多い．

その一方で，特定の用語が使われていてもいなくても叙述が同じ意味内容を指している知覚連語の集合も考えられる．この知覚連語の集合を知覚連語の同義性同値類と呼ぶ．共通語同値類の知覚連語は同義性同値類の知覚連語で類別されるのである．難解語があるときにその同義性同値類を検索すれば，難解語を用いないで同等の知識内容を表現しているものを見つけることが出来る．これが難解語の言い換え技術の第二の手法である．これは，キーワード検索とは異なる新しい検索技術でもある．

こうした技術の動機を与えたオノマトペをふり返っての適用例として“カンカン”という言葉が例に出され，宮沢賢治などの文学作品に対しても有効な言い換え文が検索されることが示された．

第 3 節ではさらに，同義性同値類の連鎖的検索の有効性についても言及された．日本語の一つの文があったときに，この文から意味的距離が一定の半径内にあって，もとの文とは少し意味の違う知覚連語を見つける．そしてふたたび，この知覚連語を起点として，ある半径の内側にある知覚連語を検索する．このように興味のつながる経路に沿って知覚連語の意味空間をたどることによって，一つの系統だった知識が検索される．このことが可能なのは，知覚連語の意味空間が連続性を保持しているからに他ならないが，我々の検索システムの展開性を物語っていると考えられる．例として取り上げられたのは素粒子の“ニュートリノ”という言葉で，この言葉を起点にして“弱い相互作用”，“弱い相互作用と電磁相互作用の統一”，“大統一への示唆”などの一連の概念が検索された．

第 4 節では分野を神話，ギリシャ神話に特定して近年の進展が例証された．日本語解析システム「ささゆり」は，知覚連語の文法的側面や意味的側面などの諸側面を制御するために 5 つのインターフェイスを持っているが，ギリシャ神話について 5 つのインターフェイスがどのように機能するのかを示す形式で例証が展開された．5 つのインターフェイスとは，(1) 知覚連語辞書，(2) 概念辞書，(3) 専門用語辞書，(4) 共通の用語を含む知覚連語の検索のインターフェイス，(5) 同じ意味を持つ言い換え表現の検索のインターフェイスの 5 つであるが，(1) と (2) は知覚連語の文法的側面や意味的側面などの基礎的性質を構築，編集，管理するためのものである．(3) は専門用語の定義を与えるという意味での役割を持つものである．この論文の主要なテーマである同義異表現検索機能と難解な日本語の言い換え技術の例証が最も端的に表現されるのは (4) と (5) である．

(4) では，“アポロン”というギリシャ神話の神の名前を例にとって，専門用語の使用例が用語を含んだ叙述の前後関係が多くの“アポロン”の特性を含んだ情報を提供することが示された．そして，このインターフェイスでは用語の他に補助キーを指定して，情報をさらに絞り

込むことで概念辞書の意味要素を効率的に増強する機能が備わっていることが例示された.

(5) のインターフェイスが言い換え技術の本領と言えるものであるが, 専門用語が含まれるか含まれないかにかかわらず, 同義検索をによって, 同じ意味内容を保持する多くの情報・表現を抽出することに成功している. このことは, 課題文が単文, 複文にかかわらず実行され, 特に複文の場合には課題文が構文解析され, 修飾文によって修飾されている名詞の意味限定についてもその効用が実証された. 実例として取り上げられたのは, キーワード的に表現して“アフロディテ”と“アテナ”という非常にポピュラーなものであるが, 2 人の女神の多面的な属性が検索された.

ギリシャ神話というジャンルについての意味的空間の連続性が, “ヘラクレス”という英雄についての連鎖的検索を実行してみるという方式で, 確認された. 特に“ヘラクレスの誕生”にまつわる父神“ゼウス”と母“アルクメネ”との経緯や, “ヘラクレスの死”にまつわるケンタウルスの“ネッソス”と妻“デイアネイラ”との経緯などが, 系統的に検索された.

第 5 節では, 複雑な木構造を持つ大域変数の中間ノードのデータ削除に伴う, Kill コマンドに対応する枝葉ノードのデータ保護保存について, M 言語に固有のプログラミング法によるシンプルかつ一般的なルーチンが記述出来ることが検討されたが, 拡張型コマンドの ZKill コマンドを用いれば自然に枝葉ノードを保護しつつ当該ノードの削除が可能であることが述べられた. 結果として知覚連語辞書の更新に関連する高階層データ群への即時的反映は容易に実現されることが述べられた.

今後の問題として, **第 4 節**の最後に述べた外来語のカタカナ表記などに見られる表現法のばらつきや混乱に対応する機能をどのように設計するべきかという思案があるが, こうした問題についても近々に対応したいと考えている. この論文で紹介されたことは, 日夜進展しつつある技術であり, 今後さらに広汎な分野への適用とコミュニケーション支援への還元が期待される.

参考文献

[1] 高橋 亘, 『コミュニケーション支援の情報科学』, 現代図書 (2007, 4 月, 相模原).

[2] 高橋　亘, “M 言語による日本語解析システム「ささゆり」の意味解析---連体修飾のある日本語文の意味解析---”, 『Mumps』, Vol. 24 (2008) 27～33.

[3] 宮地絵美, 高橋　亘, “M 言語による聾者のための日本語簡易化機能---連体修飾のある日本語文の単文化と形式名詞の意味推定---”, 『Mumps』, Vol. 24 (2008) 35～40.

[4] 高橋　亘, “日本語解析システム「ささゆり」における連体修飾のある日本語文の意味解析”, 『関西福祉科学大学紀要』, Vol. 12, 21～30 (2009).

[5] 高橋　亘, 宮地絵美, “聾者のための日本語簡易化法---連体修飾のある日本語文の単文化と形式名詞の意味推定---”, 『関西福祉科学大学紀要』, Vol. 12, 31～39 (2009).

[6] 高橋 亘, “日本語解析システム「ささゆり」における日本語文簡易化の方法と知覚連語間の意味的距離”『総合福祉科学研究』, Vol. 1 (2010) 91～100.

[7] 高橋 亘, 津村雅稔, “オノマトペを含む日本語文の代替表現機能---聾者のための情報保障の技術---”, 『総合福祉科学研究』, Vol. 1 (2010) 115～122.

[8] 高橋 亘, “知覚連語の同値性と日本語文簡易化の方法---M 言語による日本語解析システム「ささゆり」の意味解析---”, 『Mumps』, Vol. 25 (2010) 9～21.

[9] 津村雅稔, 高橋 亘, “オノマトペを含む日本語文の M 言語による代替表現機能---聾者のための情報保障の技術---”, 『Mumps』, Vol. 25 (2010) 23～33.

[10] 高橋 亘, “日本語解析システム「ささゆり」における知覚連語の同義性同値類を用いた検索技術と日本語の言い換え技術”,『Mumps』, Vol. 26 (2011) 3～26.
[11] ミチオ・カク, ジェニファー・トレイナー共著,久志本克己訳, 広瀬立成監修, 『アインシュタインを超える---超弦理論が語る宇宙の姿---』, ブルーバックス, 講談社 (1988, 東京).
[12] 宮沢賢治, 『風の又三郎』, 青空文庫
[13] オウィディウス著, 中村善也訳, 『変身物語（下)』, ワイド版岩波文庫 (2009, 東京) .

統計情報提供サービスが可能な Electronic Health Record アプリケーション構築支援

Electronic Health Record Application Development Framework of Clinical Information Statistics

徳永達也[1)], 粂直人[2)], 岡本和也[2)], 竹村匡正[3)], 黒田知宏[2)], 吉原博幸[2)]
Tatsuya Tokunaga[1)], Naoto Kume[2)], Kazuya Okamoto[2)],
Tadamasa Takemura[3)], Tomohiro Kuroda[2)], Hiroyuki Yoshihara[2)]

1) 京都大学大学院情報学研究科, 2) 京都大学医学部附属病院医療情報企画部,
3) 兵庫県立大学応用情報科学研究科
1) Graduate School of Informatics, Kyoto University
2) Division of Medical Information Technology and Administration Planning, Kyoto University
3) Graduate School of Applied Informatics, University of Hyogo

京都府京都市左京区聖護院川原町 54
54 Kawahara-choShogoin Sakyo-ku Kyoto Japan
TEL : 075-751-3165
FAX : 075-751-3077
e-mail : tokunaga@kuhp.kyoto-u.ac.jp

要旨

Electronic Health Record (EHR)は診療情報の地域間での共有にとどまらず、 患者個人を軸として診療情報を集約する概念である。診療情報を患者に提示する場合、患者が内容をそのまま理解し難いため、付加情報を与え解説することが望ましい。特に Web 上の医学用語説明や患者間のデータ比較等が有用であると期待される。著者らは EHR アプリケーションの開発を効率化するために web サービス連携を用いたフレームワークを構築してきた。本論文ではフレームワークに対してデータ統合処理モジュールを導入し統計的データを提供する方法を提案する。横断的なデータ処理は EHR のポリシーで制限されているため、 提案モジュールはアクセス制御が必要となる。また応答速度を確保するためにデータの再利用を行う。実験では提案モジュールを用い、年齢別薬剤総費用の患者個人データと平均値の比較提示アプリケーションを構築した。また、外部 web サービスを利用し統計結果を可視化した。

キーワード : 電子健康記録、 Web サービス、 統計情報提供

平成 24 年 11 月 19 日受付　　平成 25 年 1 月 7 日受理

Abstract

Electronic health record (EHR) is regarded to provide patient-centralized information service as well as regional clinical information sharing between facilities. EHR provides patient's clinical information for the patient. However, clinical information is not suitable for patients to understand. Therefore, EHR should assist the understanding by additional information that explains the meaning of the clinical information. Especially,

the additional information would be given by the comparison between patients as well as the description of medical terms on the internet. The authors had been developed an EHR application framework that provides web-based service collaboration to make application development efficiently.

In this paper, a design of a data integration module on the application framework is proposed. The module realizes to generate the additional information of the comparison between statistics and user's data. Because traversal data handling is limited by EHR policy, the module is required to have access control properly. Also, the statistics should be reused to accelerate the response of the EHR application. A sample application, which provides the average of the total amount expense of medicine in a year, was implemented by using the module. Also, the statistic result was handed to an external web service to visualize the result by a line chart. The result indicated that the EHR application framework with the proposed module is effective to accelerate the development of the additional information integrated EHR applications.

keyword：EHR, Web service, Statistical Information

1.　背景

情報技術の発展に伴って Electronic Medical Records (EMR) が普及し医療情報が電子化されつつある。地域医療連携のためにEMRを施設間で参照し、二重検査を防止するなど、診療の効率化の動きが広まりつつある。また、診療機関単位の記録ではなく、患者個人を軸として医療情報を集約するための概念として Electronic Health Record (EHR)が提示され、徐々に普及しつつある [1 - 3]。EHRは、患者個人を軸とした記録であることから患者が自身の記録を閲覧するためのサービスとしての役割をもっている[1]。 従来、EHRは診療機関から提供された診療記録をそのまま患者に提示するものがほとんどである。一方で、診療情報は医療従事者であれば内容を理解できるのに対し、患者が直接内容を理解することは困難な場合が多い。そこで、インターネット上に展開されている様々な医療関連の解説をもとに診療情報に付加的な情報を追加することで、患者の理解を促進することができると期待できる。また、診療情報の理解に際しては、疾患の平均的な推移と自身の推移を比較することで、より現在の患者本人の健康状態の理解につながるものと考えられる。例えば、検査の結果値が平均や通常に比べた場合に高いか低いかを知るだけでも、単に正常値かどうかという判定以上に疾患の推移を把握できる。ところが、他との比較を行う場合、そもそも他の全患者情報をもとに比較情報を生成する必要があるのに対し、地域をまたがる医療情報の取り扱いに関する統一的な枠組みは皆無である。そこで、EHRが提供する患者を軸とした医療情報の集合を、相互に参照するためのデータソースとして扱うことで、相対的な評価の基準となる統計的データを生成できることは容易に想像できる。

EHRをデータソースとした付加価値を提供するアプリケーションを開発するために、インターネット上のリソースやEHRの診療情報を一元的に扱い統計などのデータ統合処理を行う仕組みが必要である。EHRのデータに対するアクセス権を制御し、情報漏えい等のセキュリティのリスクを回避しつつ、患者横断的なデータ処理を行う必要がある。また、統計等の全データを対象とした処理を実施する場合、明らかに処理時間が膨大になることから、患者に対するサービスとしては成立させるための応答速度を確保するための工夫が必要になる。さらに、上記二点を個別のアプリケーションごとに実装することはセキュリティ上も、開発効率の上でも現実的ではない。EHRのデータに対するアクセス権制御と高速な患者横断的処理を容易に実装できる統一的な開発環境フレームワークにより、様々なサービスを提供するEHRアプリケーションを展開できる。英国では1987年から臨床研究を目的として General Practice Research Database (GPRD)と呼ばれる657万人超の記録を格納した診療情報大規模データベースを構築している [4]。日本では、病院

経営指標の提供や臨床研究を目的として複数施設の医事データや検査データを取り扱う医療統計情報プラットフォーム（CISA）が構築され、診療行為分析に基づく施設間比較などに用いられている [5-7]。これらは患者を対象としたデータ提供を目的としておらず、またEHRアプリケーションを開発するフレームワークとしても提供されていない。

2. システム構成

背景で述べた通り、病院経営や臨床研究に対して診療データを再利用し統計情報として利用する取り組みが存在する。しかしながら、患者への情報提供サービスにおける診療データの再利用はあまり行われておらず、付加価値のあるサービスが提供されていない。したがって、本研究は、診療データを統合することで得られる付加価値を提供可能なEHRアプリケーション開発環境の実現を目的とする。本開発環境の提供により、Webサービス連携を用いることで開発に要する労力を最小限に抑えたまま統計結果などの付加価値をもつEHRアプリケーションの開発を容易にする [8]。診療データの統合に際しては、EHRで管理される診療文書のデータフォーマット間の突合や統合が課題となる。また、EHRアプリケーションからの問い合わせに対して、処理結果を現実的な時間で応答する必要がある。以上より、次の三要件を満たすEHRアプリケーション開発環境を構築する。

- ・診療情報の取得と統合結果の管理におけるアクセス権制御
- ・標準フォーマットを対象としたデータの取得と処理
- ・EHRデータの統合処理が可能なWeb連携APIの構築

本研究でいうデータ統合処理とは、主に統計処理や項目間のデータ比較をさす。数値として加算可能な値に関しては統計処理を行うことが可能である。一方、自然文などの非構造化データに関しては、例えば対象項目のみを取り出して並べて時系列で比較することが可能である。個々のEHRアプリケーションでこれらの機能を実装することは可能であるが、Webサービスリクエストを投げることで、これらの半構造化データの解釈と興味対象のデータの抽出が行えるようなデータ統合処理モジュールを提供することで、アプリケーション開発の効率を向上させることができる。診療情報データを患者横断的に統合して扱う場合、どの項目を比較、統合するかをどのように管理するかが重要となる。特に、統計データを出力する際には、検査データなどは数値の加算による単純な統計を得られると期待しうるが、実際にはEHR間で横断的にデータを統合して統計情報を得る場合、検査方法や検査機器の違いによって、そもそも加算対象の数値の意味が異なるため事前に項目の突合を行わないと統計データとして扱うことは困難である。また、統計結果を現実的な応答速度で返すための高速化が重要となる。データ取得に要する時間を短縮するために統計対象のデータセットをデータ統合処理モジュール内に確保することが有効であると考えられるが、キャッシュされたデータや計算結果に対するアクセス権の管理はデータソースへのアクセス権の管理とは別途制御される必要がある。以上のことから、データ統合処理にあたり、データ保持の運用規約、統合対象となるデータのスクリーニング、項目間のマッピング、統計データの格納方法、応答速度の高速化手法を決定する必要がある。

複数のEHRから集約された診療情報データを統合して扱う場合、個々の診療情報が全く異なるフォーマットを用いて記述されているとデータの意味を解釈する必要があるため突合が困難である。したがって、ある程度標準化されたデータフォーマットをもつデータを対象とすることで突合にかかる処理を低減できる。既存の医療情報交換規約には、Health Level 7（HL7）や Medical Markup Language（MML）などがある[9、 10]。EHRデータソースがこれらの交換規約に則ったデータを出力できることを前提として利用するものとする。その上で、交換規約のもつデータの各項目が自然文等で記述されている場合、自然言語処理を用いて統合対象となる項目を抽出するなどの処理が要求される。

EHRアプリケーションを構築する上で、個々のアプリケーションを開発するコストを低減するためにも既存の機能を再利用することが望ましい。特に、

データ取得やデータ可視化などのアプリケーションの基本的な機能に関しては、既存のアプリケーションで実装された機能を呼び出して利用することで開発コストが低減されうる。EHRはネットワークを介したサービスを基本とすることからも、個々の機能を Webサービス連携により組み合わせてEHRアプリケーションを構築することが望ましい。Webサービスの連携によって EHRアプリケーションを構築する既存の仕組みとして EHR App Formがある[11]。EHRをWebリソースの一種としてとらえ、データソースとして扱うことで、複数のEHRのデータを横断的に取得することが可能である。データソース、データ処理、データ提示がそれぞれApplication Program Interface (API)を通して呼び出すことで、APIコールの組み合わせでEHRアプリケーションを実装することができる。本研究はEHR App Formを採用し、再利用可能な形でパッケージ化された機能やデータセットを利用する。EHRアプリケーションがデータ統合処理を呼び出す場合、ユーザに対する応答を維持するために、統合処理が一定時間で完了することを保証する必要がある。統合処理を高速化するために、EHRデータソースからのデータ取得に要する時間をできるだけ短縮したい。そこで、単純なWeb連携によるオンタイムのデータ取得に加え、フレームワーク内に利用履歴のあるオリジナルデータを格納するキャッシュデータベース(DB)を設ける。また、データの処理結果を再利用するために、結果格納DBを用意し、格納場所を番号指定などで呼び出すことで、他のWebサービスに対するデータの授受を簡素化して高速な処理を実現できる。さらに、利用頻度の高い一連の統計処理に関しては、常時最新のデータに基づき最新の結果を準備したい。そこで、事前に定義された定形処理を実施し処理結果を定期的に更新する。以上の三要件を満たすことにより、EHRデータソースからWebアクセスでデータを取得し、取得したデータを統合処理した上で、一時的にキャッシュしてEHRアプリケーションから統合結果を任意に利用できる環境を構築できる。ここで、EHRデータソースのもつポリシーを踏襲したデータ取得や、データの保存に関しては、EHR App Formの基本機能で提供される。図1に、EHRが扱う診療情報の連携と再利用の概要を示す。病院からある標準フォーマットに則って出力された診療情報が、EHRのリポジトリに蓄積し、医療機関向け、患者向け、研究向けに利用される。文書を作成した病院は当然、EHRに提示している全データを閲覧できる。連携病院間であれば、個々の契約に則って、EHRの診療情報リポジトリへのアクセスの可否がルール化される。例えば診療情報に加え医事情報もEHRで扱うことができれば、病院間の経営分析結果の比較等にも用いることができる。医療機関向けに対しては、医師の権限で担当患者のデータを閲覧できる必要がある。一方、患者向けに対しては、個人の診療情報に限って閲覧できる必要がある。研究向けには、個々の個人が特定されない形での匿名化データとして扱う必要があり、またデータの傾向を知るための臨床統計結果などでデータを提示する必要がある。

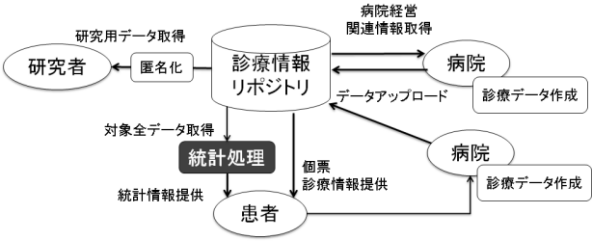


図1 EHRが扱うデータの流れと再利用の形態

2．1　診療情報へのアクセス制御

EHRで扱われる診療情報データのアクセス権にはおおまかに、診療情報作成施設からのデータ利用、患者個人に対するデータ開示、診療情報作成施設の連携機関に対するデータ開示・利用の三種類がある。全ユーザのデータを利用する場合、EHRアプリケーションを実行可能なユーザは、基本的には診療情報作成施設である必要がある。一方、患者の個人向けEHRアプリケーションの場合は、患者個人に対するデータ開示が許可されている診療情報データに関してのみ利用可能となる必要がある。連携機関が診療情報作成施設とデータ利用の契約をしている場合は、EHRの文書を利用することが可能であるが、特定の患者に関してのみ連携機関と共有を許可する場合も多いと考えられる。

本研究で提案するデータ統合処理APIが扱うデータ集合は、フレームワークが提供するデータソース系APIから取得されるものとする。本研究ではEHR App Form　を採用しているため、EHRが設定している診療データ公開のポリシーに従っているデータのみを取得可能であり、ポリシーを踏襲する範囲で複数の機関の診療データを扱うことができる。具体的には、EHR App Formは、運用規定でEHRデータソースから患者横断的にデータを取得する場合、事前にEHRからアクセス許可を得る必要がある。例えば、EHR App FormがEHRとの間の契約で、データの匿名化を条件に全データを利用する許可を得る、といった運用が求められる。 データ取得に関する取り決めは、EHR App Formの運用ポリシーを踏襲することで、フレームワークとEHR間に任せるものとし、本研究ではデータ取得の取り決めに関して議論しない。

本来、EHRが保有しているすべてのデータに対して統計などの処理を実施できることが望ましいが、事実上困難である。実際にはポリシーの制約の範囲内で得られたデータで処理を行うことになるため、統計の結果の精度評価は困難になる。一方で、本研究では、処理の元になったデータ集合の属性が明確であれば、それをユーザに提示することで、ユーザ自身が統計処理の精度を判断できる仕組みとなる。以上より、データ統合処理APIが用いたデータ集合の属性が統合処理結果のヘッダ情報として返される機能を構築する。

統計処理の結果、得られたデータ数が少ないことで個人が特定されることを回避するための統計情報利用時のルールを策定して運用する必要がある。具体的には、通常の臨床疫学研究などで採用されている、統計結果が5件以下の場合、結果を開示しないといった処理をおこなう必要がある。反対に、EHRアプリケーションが例えば少人数でお互いにデータを開示し比較し合うグループに対するものであれば、統計結果数が少なくともユーザの許可を得て結果を出力する場合を想定しうる。したがって、データ統合処理APIにデータソースアクセスの際にユーザの開示リクエストを受け付ける仕組みを実装する。

2. 2　標準フォーマットに基づくデータの取得と処理

多くの医療機関は独自のEMRで運用されており、EMR内では独自のデータ構造で診療情報を保持している。EHRから得られる診療情報データの多くは、各医療機関からEHRにデータが集約される際に特定のデータ交換規約に則った標準フォーマットとなっている。一方で、データ交換規約として広く普及している標準フォーマットはHL7、 CDAr2、MMLなど多種あり、統一されていない。EMRから得られる診療情報を標準フォーマットで出力する際には、すべての情報を構造化することは原理的に困難である。結果として、診療情報データを構造化できる部分と、非構造のままあつかう部分に分けられることが多い。例えば MMLを採用した場合、検査や処方といった文書種別と文書内の項目に関しては構造化されるが、項目の中身に関しては非構造化データのまま扱われる。図2に半構造化データの例を示す。自然文で記述されているデータを扱う際に、自然言語処理を用いて項目を抽出することで、構造化データのみならず、非構造化データも統合対象となる項目として利用できるようになる。図2では、*<mmlPc:rxOrder>*タグがタグ内のテキストデータが処方オーダであることを指している。処方オーダの中身が自然言語で記述されている場合、自然言語処理によって薬剤名と個数、単位、分量、日数などを抽出することは可能である。

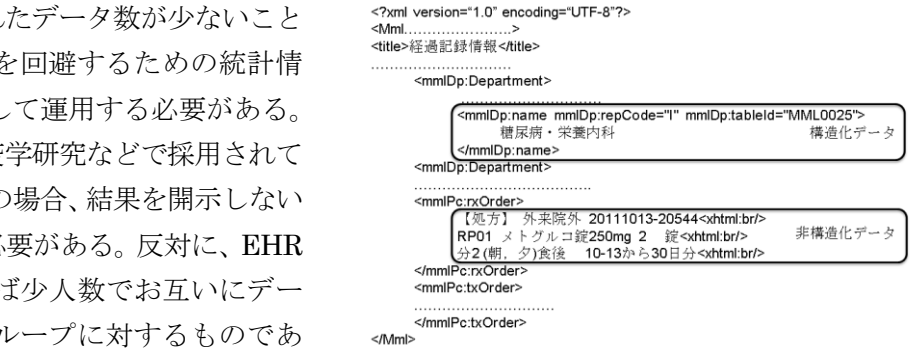


図2 データ交換規約の標準フォーマットに格納された半構造化データ（MMLの例）

ゆえに、データ統合処理モジュールは、非構造化データを定型データに変換して利用可能にする前

処理モジュールを持つものとする。ここで、前処理モジュールは自然言語処理エンジンを必ずしも独自に実装する必要はなく、可能であればWebサービス連携により外部の自然言語処理エンジンを利用する。

2.3　データ統合処理モジュールの構築と API 定義

本研究は、Web連携により呼び出せる、統計やデータ比較などの集合演算が可能なデータ統合処理モジュールを提案する。EHR App Form は、API呼び出しで連携する外部Webリソースを三種類に分けて利用可能としている。EHRからデータを取得するデータソース系API、データの加工を行うデータ変換系API、およびデータのビューを提供するデータ提示系APIからなる。それぞれのAPIは、個々のWebサービスにアクセスするためのモジュールとしてドライバをもつ。図3に、EHR App Form に提案するデータ統合処理をモジュールとして組み込んだ際の構成図を示す。

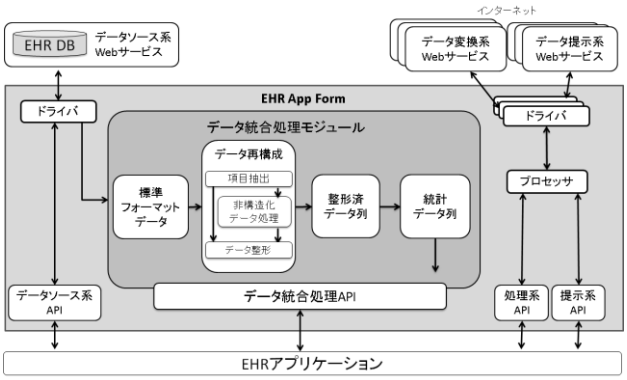


図3 EHR App Formを基盤としたデータ統合処理モジュールによるAPI拡張

2. 3. 1　処理結果の再利用

一般的に統計、比較などに必要なリソースは増加する傾向にある。したがって、一度計算した結果はできるかぎり再利用することで、システム全体の負荷を下げると同時に、アプリケーションの負荷を下げることが望ましい。再利用の対象としては、統計処理前のデータ、統計処理後のデータ、統計処理方法の三つが対象となる。統計処理前のデータに関してはEHRデータソースに対するデータ取得リクエストは必ずWeb経由となり、当然処理対象となる文書数が増加するたびに負荷として大きくなる。そこで、新たに利用するデータに関してはEHRデータソースに問い合わせて取得せざるを得ないとしても、一度でも利用されたデータに関しては極力手元にキャッシュしておくことが望ましい。また、統計処理後のデータに関しても、EHRアプリケーションからのリクエストがあるたびに同じ計算を実行することは負荷が高い。そこで、統計処理後のデータを保存し、同じデータを利用する場合には過去の計算結果に対して差分のみの計算で応答を返すことが望ましい。そのため、すでに構築された計算ロジックを用いる統計結果を常に最新の状態で準備しておくために、定期的に当該計算ロジックを呼び出して実行する仕組みが必要である。さらに、一度組まれた統計等の計算ロジックに関しても、同じ計算ロジックを構築する手間を省くために、Stored procedureとして保存し、再利用することで、開発を簡略化できると考える。

データ、ロジックの再利用に際し、もう一つの側面として、統計処理前のデータ、統計処理後のデータ、統計処理方法それぞれに対するアクセス権の制御が必要である。一方、EHR App Formが実装しているアクセス制御モジュールに認証の判断を委ねることで、データ統合処理APIを利用する際に、アプリケーション開発者がアクセス権管理を実装する必要は無くなる。具体的には、アクセス制御モジュールは、開発者IDとそれに紐付くアクセスが可能なデータソースのリストを管理している。あるアプリケーション開発者のIDで作成されたEHRアプリケーションが実行された場合、アクセス制御モジュールにセッション変数が設定される。以降の通信においては、EHRアプリケーション側でのAPI呼び出しの際に、セッションIDをリクエストに追加することで、各モジュールが独自にアクセス制御モジュールにリクエストの正当性を確認することでアクセス制御を実現する。計算ロジック、処理結果へのアクセスに関しても、その計算ロジックを実行する対象になったEHRデータソース、処理結果のもとになったEHRデータソースをもとにアクセス権の確認

を行うこととする。以上より、すべての格納されるデータには元になったEHRデータソースの情報が付加されることで、あとはEHR App Formのアクセス制御モジュールに処理を委ねることで、統計データへのアクセス権を管理できる。

2. 3. 2. モジュール設計

本研究では、簡便にEHRアプリケーションを実装するために、既存のフレームワークを採用しデータ統合機能を追加する。図4に統合処理モジュールの構造を示す。Webサービスを経由してEHRデータソースから毎度全データを取得し統計計算を行うことはデータの取得と統計計算共に非常に時間がかかり、通信量も膨大となる。そこで、本手法では事前に統計計算の結果を一時的に保持する。EHRアプリケーションに対する応答速度を確保するために、データ圧縮、処理の高速化、応答の高速化の三つの観点でモジュールを構成する。そもそも計算対象となるデータ量を削減するために、過去に計算した結果を再利用できる機構が必要である。また、計算に要するデータの取得時間を短縮するために過去に実施したデータ整形処理等を保存し、定期的に新しいデータを加えた再計算を実施する。処理リクエストがきた時点ですでに再計算済みの結果を返すことで、常に最新のデータでの処理結果を迅速に応答することができる。

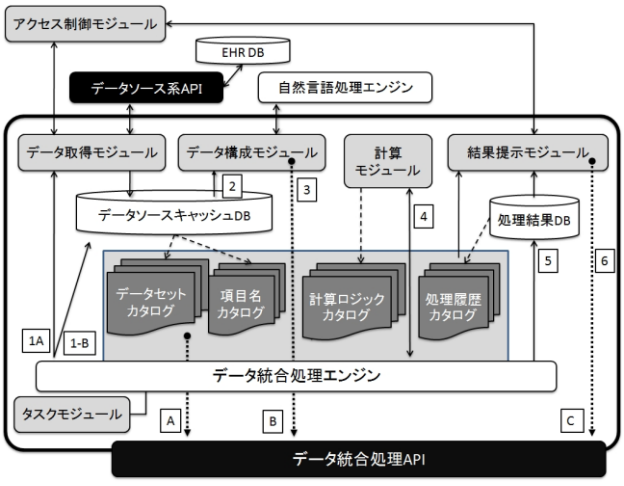


図4 データ統合処理モジュールの構造

提案するデータ統合処理モジュールは、準備するデータソースキャッシュDB、 計算ロジック、処理結果DBのそれぞれをカタログとして保持し、データ統合処理エンジンはカタログにある項目番号をもとにそれぞれのリクエストをハンドリングする設計とする。具体的には、データセットカタログには採用するEHRデータセンターのネットワークアドレス、 API仕様、 開示項目名、 データ加工履歴等が列として定義される。項目名カタログには、薬剤名、検査結果項目名、病名、薬価、医事点数等のEHRを経由して得られるマスター情報を登録する。計算ロジックカタログには、平均、分散といった基本的な演算や、投薬から数日後に検査値の上昇が見られた患者の抽出、といったデータベースクエリを格納する。ただし具体的な演算内容は計算モジュールに格納されているものとする。処理履歴カタログの各項目は、過去の計算ロジックカタログの処理と実際に利用されたデータセットカタログのインデックスを組み合わせとして持つ。たとえば、特定薬剤の薬価情報と処方量というデータセットに対して、計算ロジックカタログから平均値と年齢毎の処方量の分散を適用した、という処理内容を保存する。

すべてのリクエストはWeb連携をベースにすることで、モジュール内、モジュール外への通信が同様の手続きで実行されるため開発が容易になる。また、先の節でのべたアクセス権管理の際に、カタログに対して各項目が関連するEHRデータソースを記録することで、当該項目にあるアプリケーションがアクセスを許可されるかどうかを判断することができる。データソースとしてキャッシュした内容は、データセットカタログ、項目名カタログに展開しておくことで、EHRアプリケーション開発の最初の工程で最新のカタログを取得して新たに事前に準備されている統計処理結果やデータソース項目を検索できるようにする。また、複数データソース間の項目名マッピングなども一度実行された内容は履歴として項目名カタログに登録しておくことで、以降は突合済みの項目名を用いて統計処理を実施することができる。データ統合処理APIは、すでに処理された結果が存在する場合、結果格納DBから迅速に結果を返すことが可能である。さらに、一

時的に統計結果を保持しておくDBをフレームワーク内に設置する必要がある。もし統計出力に際してなんらかの個人を特定しうるデータが残っている場合は、統計データのマスキングや出力件数の最低件数設定などにより、結果を隠匿できるようにする。以上の手続きは、データ統合処理エンジンがカタログにある項目番号をもとにしたメッセージングにより実施する。一方、事前にデータ取得して統計情報を準備する場合、定期的にデータソースにアクセスしてデータを取得し統計情報を更新するモジュールが必要になる。タスクモジュールは、定期的にデータソース系APIにアクセスし、取得したデータと事前に与えられた処理リクエストを元にデータ統合処理APIを呼び出して処理を行ったのち、結果を結果格納DBに格納する。

2．3．3　API 定義

データ統合処理APIの定義を表1に示す。データ統合処理APIは、EHRアプリケーションからの要請に応じて統計処理を行った後、結果を呼び出し元アプリケーションに返す。表1にデータ統合処理APIの仕様と実装例を示す。

表1 データ統合処理APIの仕様と実装例

0. 認証	
入力	<アクセス制御モジュール>?<開発者コード>&<パスワード> http://~/auth?developer-id=123
出力	<セッション番号> session=123
1. カタログリクエスト	
入力	<カタログモジュール>?<セッションID>&<カタログ指定>&<検索キー指定> http://~/catalog?session=123&store=1&key="薬剤" http://~/catalog?session=123&store=2&key="患者ID"
出力	<カタログリスト> Source[1]=<薬価データベース:key1=薬剤名,key2=金額,source="EHR1"> Source[2]=<患者処方オーダ:key1=患者ID, key2=薬剤名,source="EHR1">
2. カタログ内のStored Procedure 番号に基づく統合処理リクエスト	
入力	<処理リクエスト>?<セッションID>&<データソース指定>&<計算ロジック番号>&<突合主キー番号> http://~/request?session=123&source=1(EHR1)&order="1(集計)"&key="薬剤名"
出力	<データ格納先Box番号> box=4
3. データ取り出し要求	
入力	<データ出力系API>? <セッション番号>&<データ格納Box番号>&<出力モード> http://~/view?session=123&target={box:4}&mode={1:chart(graph)}
出力	<折れ線グラフデータ> image.jpg

はじめに、EHRアプリケーションの開発者IDを受け取り、アクセス制御モジュールに渡して開発者IDがアクセスを許可されているEHRデータソースを確認し、セッション番号を取得する。次に、セッション番号と閲覧したいカタログと検索キーを指定し、カタログのなかから使いたいデータ項目を探索する。ここで、データソースとしてEHRデータだけではなく、薬価マスターデータベースなども含めて利用可能なデータとして列挙される。具体的な処理要求に関しては、セッション番号、利用するデータソース、どの計算方法を用いるか、主キーになる項目はどれかを指定すると、処理が実行され、結果は結果格納DBに保存される。EHRアプリケーションに対して、格納場所を示すBox番号が返される。最後に、セッション番号とBox番号を指定することで、結果をデータ列として取得することができる。ここで、データ統合処理モジュールはフレームワークにある提示系API対して、結果格納DBのBox番号を引き渡すことが可能である。結果格納DBの内容をデータソースとして提示系APIに出力処理を依頼できる。

3．実証実験

提案したデータ統合処理モジュールをEHR App Formに導入し、EHRアプリケーションが付加価値サービスをWeb連携によって提供できるか確認する。はじめに、データ統合処理モジュールによりEHRの全処方データを取得し年齢別に集計するアプリケーションの構築を試みる。集計結果を特定のユーザの処方データと重畳して比較するEHRアプリケーションの構築を試みる。また、比較は統計データを EHR App Formのもつデータ提示系APIを用いてグラフ表示する。次に、得られた統計結果を再利用し、データ統合処理モジュールとデータ提示系APIを利用して、統計結果に付加的情報をホップアップできるEHRアプリケーションの構築を試みる。

3．1　個人の処方費用と年齢別統計を比較するEHR アプリケーションの構築

提案するデータ統合処理モジュールは、統計処理を行う際に必要な項目を簡便に選択するために、eXtensible Markup Language (XML) で記述された定型データフォーマットを採用しているEHRデータソースを利用することを前提とする。さらに、XML を Document Object Model (DOM) や Simple API for XML (SAX) をもちいて解釈する。

以上より、EHRデータソースとしては、データフォーマットに MMLを採用している『まいこネット』を指定して利用する[12]。まいこネットは、処方データを文書単位では経過記録情報モジュールで定義して格納しており、MMLを解釈することで処方情報がどこからどこまでかを抽出することができる。しかしながら、MMLタグの中身の処方オーダは構造化されておらず、自然文のまま記述されている。そこで、MMLタグの中身に関しては、外部Webサービスの自然言語処理エンジンを用いて薬剤名と個数、単位、分量、日数を抽出する。処方費用を算出するに際し、処方された薬剤から当該の薬剤の薬価を知る必要がある。薬価マスターとして、JAPICが提供する医療用医薬品添付文書情報にある各薬剤の一処方あたりの単価を利用する。

3．1．1　方法

データ統合処理モジュールを用いて、ある患者の処方の総費用を当該患者と同年代の平均と比較してグラフにプロットするEHRアプリケーションを構築する。診療データを集約するために、まいこネットは標準規約としてMMLを採用している。参加する医療機関が日次でEMRから当日作成された診療文書をMMLで出力し、まいこネットデータセンターに送信している。MMLは文書単位のモジュールで構成され、検歴情報や処方、臨床サマリ等からなる。患者基本情報、生年月日、文書種別、文書作成日等の文書のヘッダ情報をPostgreSQLに格納している。また同時に、診療文書の文書本体を高速なXML処理が可能なCache’（InterSystems社）に格納している。PostgresSQLに格納されたカタログ情報を参照して文書リストを取得し、文書IDをもとに文書本体をCache’から取得する。Cache' は他のデータベースに比べ、非常に高速なBitmap検索が可能である。したがって、相対時間検索のように、あるイベントを軸にした前後数日のイベントを取得するような検索範囲の指定に対して高速な応答が可能である。たとえば、『ある薬剤Aを投与した前後n日間に検査項目Bの結果がCの患者』といったクエリに対して高速な応答が可能であることから、EHRフレームワークの実装に関しては特に、単にデータソースとしてのデータベースではなく、計算ロジックカタログのプラットフォームとしても Cache’をベースに実装することが望ましい。

提案するEHRアプリケーションでは、まいこネットに格納されている患者基本情報から生年月日等を取得し、そこから当該年度の当該まいこネットユーザの年齢を算出する。データ統合処理Webサービスを通じて、処方文書リストと、薬価データベースを参照する。特定のまいこネットユーザの処方情報を薬価マスターと突合して年度別処方総費用を算出した。さらに、データ統合処理Webサービスを用いて2011年度の1年分の全まいこネットユーザの処方総費用の平均をもとめ、当該まいこネットユーザの情報と比較した。最後に、当該集計結果を結果格納DBに保存して、処理履歴カタログに保存し、外部Webサービスから処理履歴カタログを参照して集計結果のグラフ表示をおこなった。

3．1．2　結果

用いた2011年度の全まいこネットユーザの文書数は 1168875件、文書取得対象となった患者IDは42519件、薬価計算の対象となった薬剤の総数は1952227件であった。図5に、処方費用の比較結果を折れ線グラフで示すEHRアプリケーションの出力結果を示す。横軸は年齢、縦軸は処方の年間総費用を指す。各年齢における処方の年間総費用の平均値を0歳から100歳辺りまで示している。また、58歳過ぎから62歳までの山なりの線は、特定のまいこネットユーザ一名の処方履歴から得られた診療を受けた年齢別の処方総費用を示している。当該ユーザは、61歳の時に平均値を大きく上回っていることが確認できる。また、統計精度に関してユーザの判断に委ねるために、統計出力に用いられたデータソースのサマライズ情報を右下に提示している。以上より、特定患者の診療データと統計情報を同時に提示して比較することができたといえる。

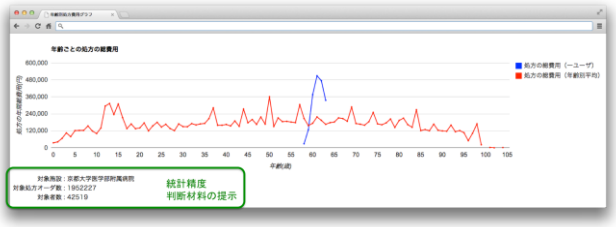


図5 個人の処方費用と年齢別統計の比較アプリケーションの実装例

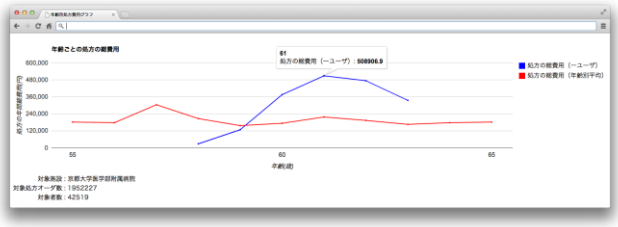


図6 統計結果と外部Webサービスの連携によるEHRアプリケーションの拡張

3. 2.　統計結果と外部 Web サービスの連携による EHR アプリケーション構築

統計情報を用いた新たなEHRアプリケーションを開発する場合、すでにある統計情報を再利用することで開発コストを下げ効率の良い開発ができる。本実証実験では、先の実験で求めた統計データを統計結果DBに保存し、処理履歴カタログを参照しながら、データを再利用してデータ統合処理モジュール外にあるWebサービスを呼び出して連携することで、新たなアプリケーションを開発することを試みる。すでに用意されている統計情報を利用し、Web連携によってアプリケーションを構築できるか確認する。

3. 2. 1　方法

先の統計データと個人の処方総費用データのグラフに対して、統計結果の一部をアノテーションして提示するアプリケーションを、アノテーションポップアップWebサービスを用いて実現する [11]。ここで、アノテーションポップアップWebサービスは、与えられたデータをグラフ等の特定の場所にポップアップして表示するものである。具体的には、グラフのあるポイントを選択すると、その場所の時点での対象ユーザの年齢と、処方総費用が出力されるものを構築する。

3. 2. 2　結果

図6に、統計結果と外部Webサービスの連携によるEHRアプリケーションの例を示す。グラフ上にプロットされた点がオンマウス時に年齢と総費用の数値をポップアップで提示できた。

EHR App Form のWeb連携の枠組みを利用することで、データ統合処理モジュールの出力結果を他のWebサービスに引き渡してグラフ化し、さらにアノテーション情報を付加することができた。図6では、データの参照範囲を当該個人ユーザのデータの近傍だけに限定してすでに結果が得られている統計データを取り出し、グラフ上で拡大表示することができた。以上より、EHR App Formに則ったWeb連携で、データ統合処理モジュールの処理結果を再利用することで、多様なビューを提供するEHRアプリケーションを簡便に構築できたといえる。

3. 3　統計結果比較 EHR アプリケーションの応答速度評価

先の統計結果比較EHRアプリケーションを実行する際の実行時間を評価する。特定ユーザの集計を行う際の処理時間を表2に示す。

表2 統計結果比較EHRアプリケーションにおける特定ユーザの処方総費用算出時間

	ID数	文書数	薬剤数	実行時間 [s]
文書リスト取得	1			1.7
文書本体取得		125		89.03
項目抽出		125		0.32
前処理			189	0.29
薬価計算			189	0.06
年齢別計算				0.01
DBデータ呼出				0.01
Webサービス呼出				0.01
合計 (秒)				91.43
合計 (分)				1.52

表2の特定ユーザの処方総費用算出にかかる各工程の実行時間をみると、文書本体取得に要する時間が約90 [s]であり、通常のWebサービスアプリケーションの応答待ち時間としては許容できない。ここで、図4のデータソースキャッシュDBに、一度でも

利用したユーザ情報は簡略化した形で保存し、保存している情報をカタログとして、データセットカタログ、項目名カタログに記録する。この時、文書本体取得で直接EHRデータソースに取得リクエストを投げる必要がなくなるため、全体の算出時間から90 [s] を引いた、2.39 [s] が応答時間として計算できる。2.39 [s] は十分に統計結果の表示待ちに耐える時間であると考える。

次に、個人の処方総費用算出時間にもとづく、全ユーザの処方総費用算出時間の推定結果を表3に示す。

表3 統計結果比較EHRアプリケーションにおける特定ユーザの処方総費用算出時間

	実行時間 [s] (個人)	実行時間[s] (全データ)
文書リスト取得	1.7	72282.3
文書本体取得	89.03	832472.78
項目抽出	0.32	3004.01
前処理	0.29	3006.43
薬価計算	0.06	585.67
年齢別計算	0.01	0.01
データ呼出	0.01	0.01
Webサービス呼出	0.01	0.01
合計 (秒)	91.43	911351.22
合計 (分)	1.52	15189.19

2011年度の全処方データを算出する際の延べID数は 42519、文書数は 1168875、薬剤総出現数は1952227であった。ここから、全処方の統計を算出するための処理時間を推定すると、約 253時間、およそ 10.55日かかることになる。ここでも、文書取得リクエストにかかる時間が約 10.47日であり、統計対象となる文書数に比例して増加するため、単純に新規統計処理を実行した場合にはEHRアプリケーションとして実用的ではない。したがって、データソースキャッシュDBに全ユーザの処方情報を格納しておくことで、応答速度を向上させることができるといえる。この場合、全ユーザの処方情報の集計時間だけであれば約 1.83時間で結果を準備できることになる。結果を処理結果DBに保存しておくことで、ある特定のユーザが統計と自身のデータの比較を実行した場合、自身のデータの集計に要する時間程度で応答を得られることが期待できる。以上より、データ統合処理モジュールにデータソースキャッシュDBと処理結果DBがあることで、EHRアプリケーションの応答速度を確保できるといえる。

4. 考察

応答速度の高速化に関しては、データの再利用、計算方法の改良の二点が考えられる。毎回対象とする全ての個票データを収集して計算を行う方法に比べて、差分データを用いて統計処理することによる高速化ができる。また、統計精度を犠牲にして良い場合、近似計算によって応答時間の短縮を図ることは可能である。一方、統計結果の精度を重視する場合は最新のデータを取得して計算をする必要が出てくるため、応答時間は遅くなる。応答時間と統計結果の精度のトレードオフに応じて、高速化の手段を動的に変更することでEHRアプリケーションの動作品質に影響すると考えられる。例として構築した患者処方総費用のグラフでは、ある患者のデータが58 歳から62 歳の期間のデータしか存在しない。今後、EHRが安定して長期間運用され、患者の生涯に渡ってデータが記録されることで、比較できる年齢の幅も当然広がると期待できる。

実験では、処方の総費用をまいこネットに提供されている診療データの集合から統計結果を算出しており、まいこネットのデータの大半は京都大学医学部附属病院のデータである。したがって、まいこネットにデータ提供している施設の傾向が大きく現れていると考えられることから、今後、急性期病院のみならず、亜急性期病院や慢性期病院の診療データが集約されることで同様の統計をおこなっても当然結果は異なる傾向を持つと考えられる。

本研究ではデータ交換規約としてMMLのタグから文書種別を特定し、処方内容をある程度整形された自然文から自然言語処理によって取り出した。実験では、年齢ごとの薬剤総費用の算出の際に、処方オーダから得られた薬剤名とJAPICの薬価情報に記載された薬剤名とのマッチングが年度間で正確に実施できない場合があった。また、ミリグラムやグラム等記載されている単位が統一されていないものもあり比較計算が困難な薬剤が存在した。自然文で記述された内容の解釈は自然言語処理の精度に依存しているといえる。また、そもそも処方デー

タに記載された薬剤名が利用した薬価情報の中に存在しない例も確認された。公開されている薬価情報と診療機関で入力される情報の標準化がさらに必要である。

統計結果が正確な比較情報として提供されるためには、ある患者の薬剤費を同一疾患の患者群の平均と比較して提示することが望ましい。一方で、複合疾患をかかえる患者の場合、薬剤費に最も相関の高い疾患が主病名の疾患とは限らず、平均との比較は困難である。したがって、より詳細な比較を行うためには、個々の処方情報の中から比較対象とする疾患に相関の高い薬剤を抽出した上で、薬価を積算する必要があると考えられる。この場合、処方から疾患と薬剤の相関を決定的に算出することは困難なため、全体の統計結果から相対的に相関の重みを各疾患と薬剤の間に重みを印加するなどの手法が求められる。以上より、各パラメータの重み付けのための統計処理と、算出された重みを用いた統計処理による結果と個々の症例との比較、という二段階の統計処理により比較情報の精度を向上させることができると期待できる。

5.　まとめ

本研究では、蓄積されたEHRデータを有効活用するために、患者に対する情報提供において統計情報を提供することを目的として、EHR App Form を経由して診療データ取得し、得られた統計結果を患者に提供するWebサービスモジュールを構築した。さらに、構築した機能を利用した統計情報を提供するEHRアプリケーションの構築を行った。統計情報の提供によって患者が自身の健康状態を比較しながら把握するための情報の提供が可能になった。以上より、Web連携によるEHRアプリケーションの開発フレームワークに、本研究で提案したアクセス権を制御しながら容易にデータ統合処理が可能なモジュールを追加することで、EHRアプリケーション開発を促進し、患者に対する付加価値のあるデータ提供が可能になったといえる。

文献リスト

[1]　吉原博幸：Dolphin Project 地域医療連携システムの現状,治療,90 (2),pp.359-364, 2008.

[2]　柴田真吾：長崎県県央地区を中心とした広域地域医療連携ネットワークシステムの構築～長崎地域医療連携ネットワークシステム（あじさいネット）構築の経緯～, 全国自治体病院協議会雑誌, 44 (5), pp.711-714, 2005.

[3]　原量宏: かがわ遠隔医療ネットワーク K-MIX, 電子情報通信学会誌　94 (10), pp.863-865, 2011.

[4]　Hershel Jick, James A.Kaye , Susan S. Jick：Antidepressants and the Risk of Suicidal Behaviors, 292(3), pp.338-343, 2004.

[5]　荒木賢二：病院経営分析への CISA データ活用～とくに薬剤費について～, CISA 研究会「CISA Study：医療統計情報の有効な活用事例」発表資料, 2010.

[6]　寺島健史 :CISA によるパーキンソン病内服治療についての大学間ベンチマーク分析, CISA 研究会「CISA Study：医療統計情報の有効な活用事例」発表資料, 2010.

[7]　粂直人, 川田康友, 真鍋史朗：分散型臨床情報データベース構築による四大学病院間の医事・検査データ横断検索, 医療情報学連合大会論文集, 31, pp.750-751.

[8]　電子情報通信学会(青木利晴　監修)：Web サービスコンピューティング ,電子情報通信学会,Chapter2, 2005.

[9]　日本 HL7 協会, http://www.hl7.jp/. (最終アクセス日 2013.11.01)

[10] MedXML コンソーシアム, http://www.medxml.net/XOOPS/html/modules/news/ (最終アクセス日 2012.11.01)

[11] 徳永達也, 竹村匡正, 粂直人, 岡本和也, 黒田知宏, 吉原博幸：地域医療連携システムにおけるインフラとしてのウェブサービス構築の試み, 第 38 回日本 M テクノロジー学会大会論文集, 3-1, 2011.

[12] まいこネット – 京都地域連携医療推進協議会, http://www.e-maiko.net.（最終アクセス日 2013.11.01）

Caché 各バージョンのパフォーマンス比較

Comparison of performance for each Caché version

木村一元[1)]
Kazumoto Kimura[1)]

1) 獨協医科大学病院医療情報センター
1) Center for Medical Informatics, Dokkyo Medical University Hospital

要旨

Intersystems の Caché には、発売以来多くのバージョンがある。その中から 5 つのバージョンの評価版を用いて、Caché Object Script (M 言語)によるパフォーマンスを調べた。計算速度、文字列検索速度、グローバル書き込み速度について 6 種の M 言語プログラムを作成して測定した。各バージョンのパフォーマンスに大きな違いは無かった。

Abstract

There are many versions in Caché. of Intersystems. Using the evaluation version of the five versions, I examined the performance by Caché Object Script Language (M). Calculation speeds, speed of string search, the global write speed were measured using six programs created in M language. There was no significant difference in the performance of each version.

キーワード：Caché Object Script (M 言語), バージョン, パフォーマンス
Keyword：Caché Object Script, version., performance, mumps,

１．はじめに

Mをベースとした Intersystems の Caché は、発売以来多くの機能の改良が成されて来た。このバージョンアップ（図１)に伴い、どの程度のパフォーマンスの向上があったのか知る目的で、５つの Caché バーションについて、パフォーマンスを調べた。

２．対象と方法

調査した Caché は、5.1, 2007.1, 2009.1, 2010.1, 2010.1 for Mac の５つのバーションで、その Caché Object Script (M 言語) について、そのパフォーマンスを評価版を用いて調べた。

平成 25 年 1 月 3 日受付　　平成 26 年 1 月 6 日受理

- Caché 2.0 1997, Caché 2.1 1997/9
  Caché 3.0 1998/1, Caché 3.1 1999/1
  Caché 3.2 2000/1, Caché 4.0 2000/9
  Caché 4.1 2001/9, Caché 5.0 2002/12
  5.0.21
- Caché 5.1 2005/11, Caché 5.2 2006/2
  Caché 2007.1 '07/11,
  Caché 2008.1 '08/3, Caché 2008.2 '08/10
- Caché 2009.1, Caché 2010.1

図１　Caché の各バージョン

調査項目詳細

調査は以下の 6 つの項目を測定する Caché Object Script (M 言語)ルーチンを作成して行った。
１）主メモリ上での計算速度：スピアマン順位相関係数の有意確率の直接法による計算

(Spearman04.int) (図２)。

２） １部グローバルを用いた計算速度：

１）と同様の計算であるが、データの一部をグローバルデータとして保持して計算する方式 (Spearman14.int) (図３)。

３） 文字列検索速度：シーケンシャルファイル（41,880 問題、840,781 行、21.02MB）中の４つのキーワードの文字列検索(KKS01.int) (図４)。

４) グローバル(１階層)への書き込み速度: シーケンシャルファイル(３)と同一のデータ)の文字列データのグローバル(１階層)への書き込み速度 (KKSGS.int) (図５)。

５） ４）のグローバルデータからの文字列検索速度(KKS10.int) (図６)。

６）カナ文字の同姓同名のチェックのためのグローバルデータ(２階層)への書き込み速度。ファイルサイズ：656,296 件、47.687MB (patidchk01.int) (図７)。

1）主メモリ上での計算速度:

```
genperm ;
 set cnt=0 for i=0:1:(n-1) { set p(i)=i+1,c(i+1)=p(i) }
 set k=1
 while k<n {
     set cnt=cnt+1
     if k#2=1 set i=c(k)
     else set i=0
     set tt=p(k),p(k)=p(i),p(i)=tt
     set d2=0
     for i2=0:1:(n-1) { set d2=(i2+1-p(i2))*(i2+1-p(i2))+d2 }
     set t(d2)=t(d2)+1,k=1
     while c(k)=0 { set c(k)=k,k=k+1 }
     set c(k)=c(k)-1
 }
 q
```

図２　(Spearman04.int)

2）1部グローバルを用いた計算速度:

```
genperm ;
 set cnt=0 for i=0:1:(n-1) { set p(i)=i+1,c(i+1)=p(i) }
 set k=1
 while k<n {
     set cnt=cnt+1
     if k#2=1 set i=c(k)
     else set i=0
     set tt=p(k),p(k)=p(i),p(i)=tt
     set d2=0
     for i2=0:1:(n-1) { set d2=(i2+1-p(i2))*(i2+1-p(i2))+d2 }
     set ^t(d2)=^t(d2)+1,k=1
     while c(k)=0 { set c(k)=k,k=k+1 }
     set c(k)=c(k)-1
 }
 q
```

図３　(Spearman14.int)

3）文字列検索速度：

```
st   kill   set swc=0
     read "Key Word ? kw1 kw2 kw3 kw4 ",key,!!
     do keysp
     set fn="C:¥MPS¥Kokushi95-102x10.txt",kw=0,swt=0
     open fn
     for i=1:1 use fn read:kw=0 da q:da=":E:"  do job10
     close fn
     quit
job10  kill d,obj,wrk,wrk2
     set d(1)=da
     for j=2:1 read da q:da="問題"!(da=":E:")  set d(j)=da,je=j
     set kw=0 if $E(da,1,3)="問題"!(da=":E:") set kw=1
     do search
     q
```

図４　(KKS01.int)

4）グローバル(１階層)への書き込み速度：

```
IN   set stm=$H
     set fn="C:¥MPS¥Kokushi95-102x10.txt",kw=0,swt=0,n=0
     open fn
     for i=1:1 use fn read:kw=0 da q:da=":E:"  do job10
     close fn
     set etm=$H
     use 0 write "Ex time = ",$p(etm,",",2)-$p(stm,",",2),!
     q
     ;
job10  set n=n+1
     set ^kdat(n)=da
     q
```

図５　(KKSGS.int)

5）4)のグローバルデータからの文字列検索速度：

```
st kill  set swc=0
   read "Key Word ? kw1 kw2 kw3 kw4 ",key,!!
   do keysp
   set kw=0,swt=0,n=0,L=0,ksw=0
   for i=1:1 q:ksw=1  set:kw=0 L=L+1,da=^kdat(L) do job10
   q
   ;
job10 set d(1)=da
   for j=2:1 set L=L+1 q:'$d(^kdat(L))  set da=^kdat(L)
q:da="問題"  set d(j)=da,je=j
   set ksw=0 if '$d(^kdat(L)) set ksw=1 q
   set kw=0 if $E(da,1,3)="問題" set kw=1
   do search
   q
```

図６　(KKS10.int)

6）カナ文字の同姓同名のチェックのためのグローバルデータ(2階層)への書き込み速度：

```
st   kill ^patid set fn="C:¥HOSP¥patidtest.txt"
     open fn for i=1:1 use fn read dat do job10
eoj  close fn
     quit
job10 ;id,kana,kanji,sex,birthday
     set d1=$tr($p(dat,",",1)," "),d2=$tr($p(dat,",",2)," ")
     set d3=$tr($p(dat,",",3)," "), ….,d5=$tr($p(dat,",",5)," ")
     set d1=$tr(d1,"  "),d2=$tr(d2,"  "), ….,d5=$tr(d5,"  ")
     set da=d1_","_d2_","_d3_","_d4_","_d5
     if '$d(^patid(d2)) set ^patid(d2)=1,^patid(d2,1)=da
     else  set sq=^patid(d2)+1,^patid(d2)=sq,^patid(d2,sq)=da
     q
```

図７　(patidchk01.int)

使用機器

Windows 版の Caché である Ver.5.1, 2007.1, 2009.1, 2010.1 は、以下の PC を使用して測定した。

HP Compaq dc5100 (P4 640 3.2GHz, 1GB RAM 533MHz, WinsowsXP Pro SP3)

Mac 版 Caché の 2010.1Mac は、以下の PC を使用して測定した。

MacBook (Intel Core 2 Duo 2GHz,4GB DDR2 SDRAM 667MHz, OS 10.6.4)

Disk キャッシュサイズは 306MB 固定とした。

３．結 果

５回の測定結果の平均値を表１にまとめた。

	5.1	2007.1	2009.1	2010.1	2010.1 Mac
1) N=10 N=11	9.0 99.0	9.0 109.4	9.2 107.4	9.0 107.8	6.4 76.0
2) N=10 N=11	17.4 201.0	16.0 190.4	17.4 202.8	16.6 190.2	12.0 134.0
3) 検索1	1.8	2.0	2.0	2.0	1.4
4) 1階層	7.2	7.6	8.6	8.2	8.6
5) 検索2	2.2	1.6	2.0	2.0	1.2
6) 2階層	19.0	18.8	21.2	20.0	15.6

5回の測定の平均値(秒)

表１ 測定結果表

１)および２)の項目による計算速度はN!を反映しており、グローバルを用いた場合の計算速度は主メモリの 1.85 倍となった。

３）のシーケンシャルファイルからの読み込みと、５）のグローバルからの読み込みの速度はほぼ同じであった。

４）の１階層グローバル書き込み速度は以下のようになった。

１行当たりの書き込み ７．５μsec

１行当たりの文字数 ２５Bytes

６）の２階層グローバル書き込み速度は以下のようになった。

１件当たりの書き込み ３０μsec

１件当たりの文字数 ７３Bytes

計算速度、検索速度は、Caché のバージョンよりもハードウェア性能への依存が強い。

グローバルアクセスは、初期 CACHE.DAT の状態の違いがみられた。またバージョンによる改良がみられた。

４．考 案

主メモリ上でのアクセス速度は、どのバージョンもマシンの性能を反映している。

グローバルアクセス（書き込み時）性能において、１階層グローバルの作成時間は８秒であり、同じものをシーケンシャルファイルでの作成時間２０秒に比べて２．５倍速かった。しかし CACHE.DAT の拡張時にはグローバルへの書き込みに時間が必要なことがある。キャッシュバッファのサイズに依存した。

１階層グローバルの書き込みに比べて２階層グローバルの書き込み時間が遅かった*。両者の１データ当りの文字数は大幅に異なっている。

グローバルアクセス（読み込み時）性能においては、１階層グローバルの場合、シーケンシャルファイルからの読み込みとの差異は僅かであったが、バージョンによってはグローバルからの読み出しが速い場合もあった。キャッシュバッファの違いも考えられる。

５．まとめ

評価版を用い各バージョンの Caché のメモリアクセス速度とグローバルへの R/W アクセス速度を Caché Object Script の６つの項目について調べた。

Windows 版でのパフォーマンスは大きくは違いが無い様であった。

Caché には多くの機能があり、この調査だけではバージョンの違いを一概に議論できないが、１つの判断材料となろう。

謝　辞

今回の Mac 版の操作方法に関してインターシステムズジャパンのカスタマーサポートセンターの田中氏には多大なサポートを頂いた。
ここに謝辞を申し上げる。

参考文献

1)木村一元, 矢口裕子: 図書館所蔵雑誌検索システム，第 28 回日本 MTA 大会論文集，25-26, 2001, 北海道
2)木村一元: WebLink アプリケーションから CSP アプリケーションへ, 第 33 回日本 MTA 大会論文集，55-60, 2006, 東京
3)医事試験制度研究会: 平成 17 年版 医師国家資格試験出題基準, pp148, まほろば, 2004
4)http://en.wikipedia.org/wiki/JPEG_File_Interchange_Format
5)http://epi.fm.senshu-u.ac.jp/~ph150187/distribution/050722_jpeg_etc.doc
6)http://hp.vector.co.jp/authors/VA032610/contents.htm
7)まつもと ゆきひろ, 石塚 圭樹：オブジェクト指向スクリプト言語 Ruby，pp576，アスキー出版，1999
8)木村一元: WebLink を使った医学部学生向け教育システム，第 27 回日本 MTA 大会論文集, 41-44, 2000, 名古屋

＊編者注　本文中では「１階層グローバルの書き込みに比べて２階層グローバルの書き込み時間が遅かった」となっておりますが、実際のデータを計算しますと大きな違いはありませんでした。

故木村日本 M テクノロジー学会長より預かりました書きかけの原稿を、西山強と土屋喬義が編集を行い作成いたしました。

院内に蓄積された低解像度文書画像を対象とした文書検索システムの提案とその評価

Proposal of Medical Document Retrieval System for Low-resolution Document Images in Hospital Information Systems

中村峻太 [1)], 川中普晴 [1)], 土井俊祐 [2)], 鈴木隆弘 [2)], 高林克日己 [2)], 山本晧二 [3)], 高瀬治彦 [1)], 鶴岡信治 [1)]

Shunta Nakamura [1)], Hiroharu Kawanaka [1)], Shunsuke Doi [2)], Takahiro Suzuki [2)], Katsuhiko Takabayashi [2)], Koji Yamamoto [3)], Haruhiko Takase [1)], Shinji Tsuruoka [1]

1) 三重大学, 2) 千葉大学医学部附属病院, 3) 鈴鹿医療科学大学

1) Mie University, 2) Chiba University Hospital, 3) Suzuka University of Medical Science

〒514-8507 津市栗町屋町 1577 番地

1577 Kurima-machya, Tsu, Mie 514-8507, JAPAN

Tel & Fax.: 059-231-9737

E-mail: kawanaka@elec.mie-u.ac.jp

要旨

近年，病院情報システムの普及にともなって多くの診療文書が電子化されるとともに，これまでの紙の医療文書のスキャニングに関する試みも数多く行われている．しかしながら，一部のシステムでは記憶容量の制約等により十分な解像度で文書がスキャンされていないのが現状である．このような文書は，病院情報システム内に多量に蓄積されているものの，解像度が不足しているために汎用 OCR 等では電子化できない状況にある．本研究では，これら蓄積されている低解像度医療文書から文書タイトルを抽出して検索タグを付与する方法，およびそれを利用した紙文書画像の検索システムについて検討する．本論文では，研究の基礎的検討として，スキャンされた文書の解像度・フォントと文字認識精度に関する調査を行うとともに，その結果に基づいた文書タギング法を提案した．また，試作したシステムを用いて実際の医療文書による評価実験を行った．

キーワード：スキャニングシステム，低解像度画像，文字認識，医療文書検索システム

平成 25 年 2 月 22 日受付　　平成 25 年 3 月 22 日受理

Abstract

Recently, many paper-based documents used in hospitals have been computerized because of diffusion of Hospital Information Systems (HIS). However, some of previously scanned documents do not have enough resolution for document image processing, e.g. OCR, due to storage limitation of the systems. Currently, these documents are archived in HIS, but not used effectively now. These documents should be converted to electrical data for medical

document retrieval. This study discusses document image processing methods to search low-resolution document images. As the first step of this study, we propose the tagging method for low-resolution images archived in HIS and develop a prototype system for evaluation experiments using M language. This paper shows the detail of the proposed method and experimental results, and also discusses the effectiveness of the developed system. We also describe problems and future works of this study in the end of paper.

Keyword : Scanning System, Low-resolution Image, Character Recognition, Medical Document Retrieval System

1. はじめに

近年，病院情報システム（Hospital Information System: HIS）の普及にともない，様々な診療データが電子化されつつある．しかしその一方で，HIS の導入以前に作成された診療文書は，主に紙文書として保管されている．これらの紙文書は臨床研究においては貴重な知識データベースである．しかしながら，これらの文書は上記の理由や電子データに変換する際，手作業による入力では時間と費用がかかるといった理由により，有効に活用されていないのが現状である．

一方，昨今では多くの医療機関において紙文書のスキャニングや診療情報の利活用に関する研究や試みが数多く報告されている[1-5]．しかしながら，一部の HIS ではシステムの持つ記憶容量の制約等といった理由により，十分な解像度でスキャンされていない文書も存在する．このような文書は，システム内に文書画像データとして多量に蓄積されているものの，解像度が不足していることから，市販の OCR ソフトウェア等を用いて文書中の文字を認識する，あるいは文書構造を認識することは容易ではない．また，蓄積されている文書の中には，他院からの診療情報提供書（紹介状）や各種検査の報告書，診断書等も含まれている．そのため，診療情報の二次活用という観点からは HIS との連携，すなわち各文書に何らかのタグを付けて検索や参照が可能な状態にすることが望ましい．しかしながら，これらの文書に検索タグを付与するには手作業に頼らざるを得ない状況にある．そのため現在，検索タグが付与されている文書は極めて少なく，これらの文書は有効に活用されていない．

そこで本研究では，これら HIS 内に蓄積された低解像度の文書画像に対して文書種（文書タイトル）をタグとして付与（タギング）するための方法について検討する．本論文では，研究の第一段階として，解像度と文字認識エンジンの認識精度について調査・検討する．次に，得られた調査結果を活用し，低解像度でスキャンされた医療文書に対して文書種タグを自動的に付与する（タギング）ための文書タイトル認識方法を提案する．評価実験のために，千葉大学医学部附属病院の病院情報システムから M 言語（Caché）を用いて画像を抽出する環境を構築するとともに，抽出された文書に対してタイトルをタグ付けする評価実験を行った．

2. 実験材料

本研究では，三重大学医学部附属病院ならびに千葉大学医学部附属病院にて使用・蓄積されている文書，さらにそれらのフォーマットを参考にした帳票形式の文書を複数種類作成し，評価実験の材料として使用した．通常，文書をスキャンする際には十分な見読性を確保するために，300dpi 程度の解像度が要求される[6]．しかしながら，e - 文書法（あるいはそれに先立つ厚生省健康政策局からの通知）施行前にスキャンされた文書や，参照用として一時的にスキャ

ンしたものについては，解像度が極端に低いものも少なくない．実際，千葉大学医学部附属病院の病院情報システムに蓄積されている文書画像を精査したところ，スキャン時の解像度が80dpi 程度の文書画像も 3 万通ほど存在した．本論文では，これら文書の中から

A. 排尿に関する質問票（10 枚）
B. 診療情報提供書（7 枚）
C. 肺がん検診問診票（10 枚）
D. 検査報告書（3 枚）
E. NICU 入院総括（10 枚）
F. 入院総括（10 枚）

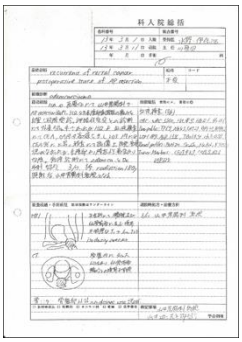
の 6 種類の文書（計 50 枚）を実験材料として用いることとした（図 1）．なお，診療情報提供書や検査報告書については，文書作成元の医療機関や検査によってその書式が少しずつ異なるが，同種の文書として扱えるものについては，同じカテゴリとして扱うこととした．

（a）文書 A

（b）文書 B

（c）文書 C

（d）文書 D



（e）文書 E

（f）文書 F

図 1　実験材料（一例）

3. 解像度と認識精度の関係

本研究ではまず，市販されている OCR エンジンの持つ認識精度について確認するため，入力画像の解像度，フォントの種類と文字認識精度に関する予備実験を行った．実験では，平仮名および図 2 に示すような，院内文書によく用いられている文字（計 269 文字）を印字した文書を作成し，フラットベッドスキャナを用いて文書画像化する．次に，スキャンされた画像の解像度を 75～300dpi まで変化させ，各画像に対して汎用の OCR エンジンを用いて文字認識を行う．なお，図 2 中の文字列は複数の院内文書を参考に決定した．また，文書に印刷する文字のサイズは 12pt，フォントは院内文書にてよく使用されている明朝，ゴシック，丸ゴシックの 3 種類とした．OCR エンジンはパナソニックカラーOCR ライブラリーを用いた．

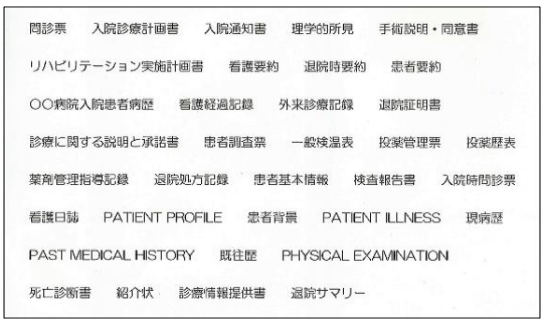


図 2　文字認識実検用文書（丸ゴシックの例）

図 3 に予備実験の結果を示す．図において横軸は入力文書画像の解像度を，縦軸は誤認識率

を示している．図からもわかるように，市販の OCR ソフトを用いた場合，入力文書画像の解像度が 150dpi を下回るとともに誤認識が発生し始め，100dpi 以下となると急激に誤認識率が上昇することが明らかとなった．また，入力画像の解像度が 150dpi より高い場合においては，フォントタイプと文字認識精度の間には有意な関係性は認められなかった．

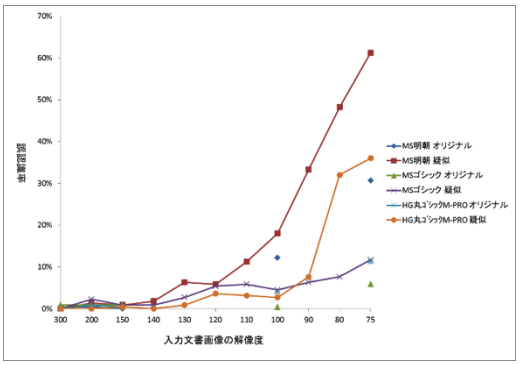


図 3　入力画像の解像度と誤認識率の関係

一方，解像度が 100dpi 以下となった場合，フォントの種類によって認識精度に大きな差が生じた．一般的に，OCR エンジンでは，文字の形状（主に輪郭）を認識のための特徴量として使用している．そのため，低解像度化により文字の一部が消失してしまい，結果として特徴量が大きく変化してしまったものと考えられる．これらの結果より，文字の輪郭を特徴量として用いる汎用の OCR エンジンを用いて低解像度の文書画像に検索タグを付与することは容易ではなく，新たなアプローチによるタグ付け方法が必要であると考えられる．

4.　方法

(1)　辞書データベースの作成

図 4 に提案法の概要図を示す．提案法ではまず，「診療情報提供書」や「退院時要約」などといったタグ付けの対象となる文書種名の画像を作成する．作成された文書タイトル画像は，グレースケール化や 2 値化といった前処理の後，垂直方向の射影ヒストグラムを用いて文字間のスペース等の空白部分が除去される．提案法では，文字間のスペースを除去した画像を辞書データとして用いることにより，例えば「診□療□情□報□提□供□書」のように各文字間にスペースが入っているケースについても対応することが可能となる．次に，作成された文書タイトル画像は，その外接矩形が切り出された後，フォントサイズによる依存性を排除するためにアスペクト比を保持した拡大・縮小処理により画像サイズを規格化する．なお，80dpi でスキャンされたタイトル画像にサイズを合わせた場合では，対象となる画素数が少ないため各辞書画像間の差異が顕著に表れない可能性が高い．そこで提案法では，タイトル部分に使用されている 12 ポイント相当の活字を法令等で規定されている解像度（300dpi）でスキャンした時の文字サイズを考慮し，辞書に登録する文字画像の高さを 50 ピクセルとした．また，予備実験の結果から，使用するフォントによって文字の

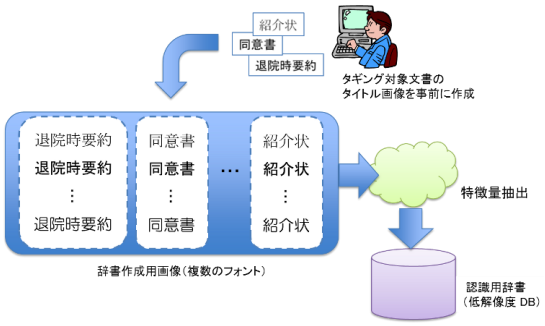


図 4　辞書データベースの作成プロセス

特徴が大きく変化することが予想される．そのためここでは，院内文書にてよく用いられている，明朝，ゴシック，丸ゴシックの 3 種類のフォントを用いて文書タイトル画像を作成し，辞書に登録することとした．

図 6 に，作成された辞書用の文書タイトル画像の一例を示す．作成された画像は横方向の空白部分を全て除去するため，文字や使用しているフォントによっては，「偏」と「旁」の間にある空白部分も削除されてしまう可能性がる．しかし，これらの空白は使用しているフォントの

種類とその大きさに強く依存するため，複数のフォントを用いた辞書を作成するケースにおいては大きな問題とならない．

(2)　文書タイトルの識別

図 7 に文書タイトルの識別方法の概要図を示す．ここではまず，スキャンされた文書画像（入力画像）に対して 2 値化や傾き補正といった前処理が行われる． 次に，入力された文書画像から文書種を表す部分，すなわち文書タイトルに相当する部分が抽出される．ここでは射影ヒストグラムを用いてタイトル部分を抽出した．なお，タイトル部分を抽出する方法には種々の方法が考えられる，その方法については，本論文では限定しない．文書タイトルは

（1） 文書の最上段に書かれていること，
（2） タイトルの文字サイズは他のそれと比較して大きいことが多い

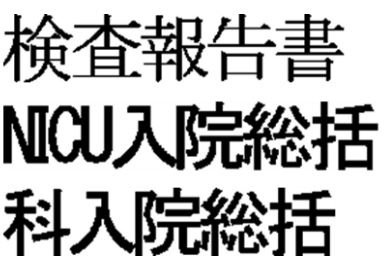


図 5　作成された辞書登録用画像の一例

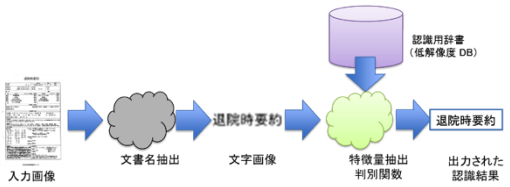


図 6　文書タイトルの識別

という特徴を利用すれば，その領域を特定することができるからである．抽出されたタイトル部分は，辞書に登録されている文字画像に合わせて拡大・縮小処理が行われる．ここでは，入力画像の高さが 50 ピクセルとなるように縦横比を保ったまま画像を拡大・縮小する．

次に，抽出されたタイトル画像を用いて文書タイトル（文書種）を識別する．前章にて述べたとおり，これらの文字画像はその解像度の低さゆえ，市販の OCR エンジンや文字認識アルゴリズムを使用することは難しい．また，同様の理由により，通常の文字認識の処理過程において用いられる輪郭情報等の特徴量を利用することも現実的な方法ではない．当然のことながら，OCR の認識率が著しく低下する低解像度の場合，その認識結果について信頼性が担保できないため，追加辞書による補正も困難である．一方，タギング対象となる低解像度文書は「診療情報提供書」,「紹介状」,「検査報告書」や「退院時要約」といった特定の文書である．言い換えれば，画像中の文字を個々に認識する必要はなく，文字（あるいは文字列）の塊として認識できればよい．そこで提案法では，個々の文字を認識せず，辞書に登録された文字画像とのマッチングを取ることにより，入力文字画像の識別を行うこととした．

入力された文字画像はまず，その画像の横幅に基づいてその文書種の候補を絞り込む．本研究では，図 1 の文書のようにタイトル部分が活字印刷された文書を対象としているため，使用フォントが決まっていれば，その文字幅は大まかに決まる．また，処理対象となる文字画像は規格化処理により文字列の高さが揃えられているため，画像の幅を用いて識別結果候補の絞り込みが可能となる．ここでは，入力画像の横幅に対して±10%の許容量を設け，その範囲を超える画像については識別結果の候補から除外することとした．

次に，絞り込まれた候補に対して辞書画像との相違度を計算する．ここでは，最も単純な方法として，比較対象となる画素数のうち画素値が異なっていた画素数を対象画素数で除したものを用いた．この相違度を全ての辞書画像について計算し，その値が最小となったものを識別結果として出力した．

5.　結果と考察

(1)　試作システムの概要

本論文では，提案法の有効性について検討するため，評価実験を行った．図 7 に千葉大学医学附属病院における医療文書画像の抽出プロセスを示す．ここでは，千葉大学医学部附属病院における病院情報システム（東芝情報システム製，データベースは Caché を使用）及びファイルサーバの文書管理用ストレージから M 言語を用いて文書を抽出する．

千葉大学医学部附属病院では，画像ファイルのアドレスや属性などの情報を病院情報システムに，画像ファイルそのものをファイルサーバに保存している．このうちアドレスや属性については病院情報システムを構成している Caché データベースのグローバルに保存されており，管理には M 言語を利用している．

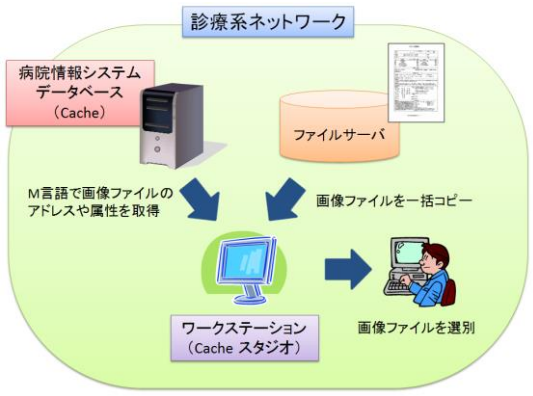


図 7　医療文書画像の抽出プロセス

まず，千葉大学医学部附属病院の診療系ネットワークに接続されたワークステーション端末を用いて抽出用のプログラムを M 言語で作成する．これには千葉大学医学部附属病院企画情報部の端末にあらかじめインストールされている Caché スタジオを利用した．また，属性情報にはファイルの種別，診療科，登録時間，新規・更新，保存先アドレス，コメントなどがタグづけられているが，1 章で述べた通り，取り込み文書の画像ファイルについては十分なタグづけがされていない状況にある．そのため，本研究で利用するファイルをプログラムで自動的に選別することは不可能であった．そこで，病院情報システムから画像ファイルの保存先アドレスを取得した後，該当する可能性があるファイルを研究用の外付けハードディスクに一括ダウンロードした．対象とする画像ファイルの選別については，千葉大学医学部附属病院企画情報部の情報管理責任者の指導の下，目視にて抽出を行った．画像ファイルの中には回転しているもの，2 枚の文書を結合させて保存しているもの，著しい損失があるもの等が混在しており，完全にイレギュラーであると判断したファイルについては対象から除外した．

また，一連の作業は全て情報管理区域である企画情報部内で行い，本研究に関係のないタイトル部分以外の情報については，塗りつぶし等を行うことで個人情報の保護に努めた．

(2)　実験結果

2 章にて述べた 6 種類（計 50 枚）の文書画像に対して提案法を適用し，その認識率について調査した．表 1 に評価実験の結果を示す．表中における文書種 A～F は，2 章にて示した文書種と対応している．表からもわかるように，文書種 A，C，E については，スキャンした全ての文書を間違いなく識別できている．また，文書 F についても 90%と高い識別率が得られており，実験に用いた全ての文書に関する識別率は 94%であった．一方，文書 B, D, F については，識別に失敗しているケースが見受けられた．特に，文書 D のケースでは，スキャン時に発生した紙のズレにより画像が大きく傾いていたため，文字部分の外接矩形サイズが辞書画像と大きく異なってしまったことが原因となっていた（図 8）．なお，この画像に対して傾き補正処理を再度適用した結果，文書 D として正しく識別された．また，文書 B のケースでは，入力文字画像の横幅が辞書画像の横幅と近い値であったこと，本手法で用いているマッチング方法が画素単位でのマッチング方法であることが原因となっていた．

表 1　実験結果

識別結果

入力文書種	A	B	C	D	E	F	認識率(%)
A(10)	10	0	0	0	0	0	100
B(7)	0	6	0	0	1	0	85.7
C(10)	0	0	10	0	0	0	100
D(3)	1	0	0	2	0	0	66.6
E(10)	0	0	0	0	10	0	100
F(10)	0	0	0	1	0	9	90

図 8 識別に失敗したタイトル画像（文書 D）

(3)　考察

実験結果から，本問題のようにタギング対象となる文書種が限定的である場合，提案法のような比較的単純なマッチング方法をもちいることで，低解像度文書に対しても検索タグを付与することができると考えられる．

一方，識別に失敗した原因としては，傾き補正が不十分であったこと，各画像の画素値の差異のみを用いて相違度を算出しているため，入力されるタイトル画像によっては相違度の値が高くなってしまい，結果として誤認識されてしまったと考えられる．また，提案法では識別ミスを防ぐために予め画像の横幅を用いて候補画像を絞り込んでいる．しかし，スキャン条件によってはデバイスに起因する「にじみ」や「欠け」等によって文字の形状や横幅が変化してしまい，その結果，誤認識を引き起こす可能性もある．今後は，単純な相違度を用いるだけでなく，文字列の横幅やおおまかな文字形状（黒画素の分布),各文字の局所的な特徴量などを相違度計算に取り入れていく必要があると思われる．

(4)文書検索システムとしての有用性

先に述べた方法により，90%以上の文書については，その文書種を識別することが可能である．また，本章（1）にて示したシステムに提案法を組み込むことにより，付与されたタグを用いて低解像度文書を検索することも可能となる．また，本システムは現行の病院情報システムとも連携しているため，システム内に蓄積されている診療データとも容易に紐付け可能である．すなわち，現在の病院情報システムに蓄積された診療データと過去の紙文書を横断的に検索・閲覧することができるため，これまで活用されなかった診療データの利活用が可能になると考えられる．

6. まとめ

本論文では，病院情報システムに蓄積されている低解像度文書画像を対象としたタギング法の提案のための予備実験を行った．また，予備実験の結果をもとに，マッチング法を用いた文書タイトル認識とタギング法の概要を紹介した．文書検索のために，Caché をベースとしたシステムの提案と試作を行った．評価実験の結果，提案法を用いることにより 94%の識別率が得られた．実験結果から，HIS に蓄積された低解像度の文書画像についても，タギング対象となる文書画像が限定的であれば, OCR を使用することなく文書種を認識・タギングができることが明らかとなった．

現在も筆者らは，千葉大学医学部附属病院と協働で識別エンジンの開発とシステムへの実装，評価試験を進めており，医療現場での利用を目的としたシステムを開発している段階である．院内には本論文で取り上げたような文書に加えて，多種多様の文書が低解像度で保存されているのが現状である．今後は，これらの文書に対する評価実験を進めるとともに，提案法のさらなる精度向上や問題点の改善にも取り組んでいく予定である．

参考文献

[1] 桑田成規，稲田拓，大越厚，浜本政志，近藤博史：e-文書法対応スキャンシステムの構築および稼働実績評価診療録管理，日本診療録管理学会誌，　20(2)，26，2008

[2] 稲岡則子，紀ノ定保臣，宇都由美子，石原謙：伏見清秀データウェアハウスとデータ利活用医療情報学，27(3)，261-268，2007

[3] S. Doi, T. Suzuki, G. Shimada, M. Takasaki, S. Fujita, T. Tamura and K. Takabayashi: Auto-Selection of DPC Codes from Discharge Summaries by Text Mining in Several Hospitals and Analysis of Differences in Discharge Summaries, Journal of Computational Intelligence and Intelligent Informatics 16(1), 48-54, 2012

[4] 竹村匡正：A knowledge extraction from medical practical documents on medical terminology and ontology 日本放射線技術学会雑誌，65(7), 962-3 (20), 2009

[5] H. Kawanaka, K. Yamamoto, H. Takase and S. Tsuruoka, Document Image Processing for Hospital Information Systems, Modern Information Systems (C. Kalloniatis *ed.*), 65 - 86, 2012

[6] 首相官邸，e-文書法の施行について，http://www.kantei.go.jp/jp/singi/it2/others/e-bunsyou.html

「日本エム・テクノロジー学会」ご入会のご案内

日本エム・テクノロジー学会（日本ＭＴＡ）は、M言語（ＭＵＭＰＳ）の利用・改良・普及を目的とした団体で、個人や法人が加入して活発な活動を行っております。M言語はＡＮＳＩに FORTRAN 及び COBOL に続いて３番目の標準コンピュータ言語として制定され、米国連邦情報処理標準言語にも採用されました。さらに１９９２年５月にはＩＳＯ標準言語として制定されるに至っております。一方、近年のコンピュータのダウンサイジングの流れにあって、ユーザも着実に増えつつあります。

日本ＭＴＡは先に述べたような目的に向けて種々の活動を続けておりますが、貴方にも、是非とも日本ＭＴＡに参加し活動を盛り上げて頂きたくご案内申し上げる次第です。

Ａ．日本ＭＴＡの活動

１）年次学術大会、研究会や講習会の開催
２）M言語に関する技術情報の提供
３）学術雑誌　「Mumps」の出版
４）M言語改良仕様の検討･･･米国M　Development Committee と連携
５）国際ＭＴＡ、各国ＭＴＡ（ＭＵＧ）との交流
６）M言語のＪＩＳ化推進
７）ソフトウエアの公開流通

Ｂ．会員の特典

会員になることにより次のような特典が考えられ、充分満足頂けるものと考えられます。

＊個人会員の特典

１）日本ＭＴＡ年次大会、M言語関係学術集会、研究会、講習会のお知らせ
２）日本ＭＴＡ主催の学術集会、研究会、講習会などの参加費用の割引
３）M言語に関する各種情報・資料の提供
４）流通、ソフトウエア（ＭＴＡＰＡＬ）の低額頒布
５）M言語ベンダーの折々のプロダクツ紹介・パンフレット・カタログ類の頒布
６）雑誌「Mumps」の無料配布

・上記の各種活動を通じて、M言語に関する全世界の最新の技術情報が得られます。

＊法人会員の特典

法人会員は「日本ＭＴＡの目的に賛同する法人で、日本ＭＴＡの目的を遂行するために積極的に事業を後援する事を表明した者とし、正副各１名の代表者を登録し、正副代表者とも個人会員と同等の資格を持つ」ことになります。尚、正副代表者には正会員と同様の日本ＭＴＡの役員としての道があります。

１）日本ＭＴＡ主催の集会には５名迄、会場費、講習会費などを会員割引
２）日本ＭＴＡ主催の医療人、企業人を対象とする講習会へ法人会員から優先的に出講

３）日本ＭＴＡ主催の集会への出品、展示に関する料金の割引
４）日本ＭＴＡ学術大会論文集、ＭＴＡニュース等への広告費の割引
５）法人会員のプロダクトのパンフレット、カタログ類の会員への頒布
６）ユーザ法人にはM言語ベンダーないしシステムエンジニアの紹介
７）日本ＭＴＡの流通パッケージ（ＭＴＡＰＡＬ）を割引料金で利用
８）ＭＴＡニュースを単なる広告ではなく、新しいプロダクツの紹介等の質の高いＰＲのために利用可能

注意）法人会員は、国際ＭＴＡが設けている施設会員と企業会員に相当するものですが、学校法人・国立施設など税法上非営利団体扱いの法人を非営利法人とし、国際慣例よりも 40%低い基本会費を申し受けます。その他は企業法人ないしベンダー法人としての会費を申し受けます。ご入会の手続きは下記事務局までご連絡をお願い申し上げます。

・上記の各種活動を通じて、M言語に関する全世界の最新の技術情報が得られます。
・M言語ユーザ間、M言語を取り扱うベンダー・メーカー間とのコミュニケーションが充実します。

Ｃ．会費

ア）個人会員　入会費　　￥４，０００（学生会員：￥１，０００）
　　年会費　　￥６，０００（学生会員：￥２，０００）
イ）法人会員　入会費　￥１０，０００（営利・非営利法人共通）
　　年会費　￥５０，０００（営利法人：１口）
　　￥３０，０００（非営利法人：１口）

注意）会計年度は、毎年４月１日から翌年３月３１日までです。

Ｄ．ご入会手続き

１）入会資料請求＜電話・ＦＡＸ・郵便＞
２）事務局から送付された「会員登録票」（法人会員の場合は正・副代表者の「会員登録票」及び「法人会員申込書」）に必要事項を記入の上、事務局までお送り下さい。
３）郵便払込で入会金、年会費を事務局に納金して下さい。

＜学会事務局＞

〒264-0026　千葉県千葉市中央区亥鼻 1-8-1
千葉大学医学部附属病院企画情報部内
事務担当：土井　俊祐
TEL：043-226-2346　FAX：043-226-2373
Email：mta-office@mta.gr.jp

「日本Mテクノロジー学会」規約

第一章　総　　則

第１条　本会は日本Mテクノロジー学会（M　Technology Association of Japan)という。

第２条　本会の事務所は幹事会の承認を経て、学会長が指定するところに置く。

第二章　目的および事業

第３条　本会は「M言語」並びにこれに関する情報システムの利用、応用、改良、並びに普及を行うことを目的とする。

第４条　本会は前条の目的を達成するため次の事業を行う。

1）学会大会、フェア、研究会、講習会などの開催
2）学会誌、ニュースなどの刊行物の発行
3）M言語の日本語装備の標準化
4）M言語の標準装備の監視
5）海外のＭＴＡ（ＭＵＧ）などとの連携活動
6）内外の関連諸学会との連絡ならびに協力活動
7）M言語利用技術の相互交換の促進、本会に提供された資源の整備、管理ならびに会員への還元
8）日本Mテクノロジー学会出版会に関する事業
9）その他目的達成のために必要な事業

第三章　会　員

第５条　本会会員は個人会員と法人会員からなる。

1）個人会員は本会の目的に賛同し、本会の対象とする領域、又はそれと関連する領域において活動する個人とする。
2）法人会員は本会の目的に賛同する法人で、本会の目的を遂行する為に積極的に事業を後援する事を表明したものとする。法人会員においては正副各１名の代表者を登録するものとする。正副代表者は個人会員と同等の資格を有する。

第６条　本会に入会を希望する者は所定の申込書に入会金及び会費を添えて本会事務所に申し込まねばならない。

第７条　本会会員は、毎年所定の会費を前納しなければならない。

第８条　本会会員で住所変更のあったものは速やかに住所変更届を、また退会しようとするものは退会届を本会事務所に提出しなければならない。本会会員で、住所不明となるか催促にも拘らず２か年を越えて会費納入遅滞のあったものは退会の扱いを受ける。物故会員は退会の扱いを受ける。

第９条　本会の規約に背く行為のあった会員は、幹事会の議決を経てこれを除名することができる。

第四章　役員その他

第10条　本会に次の役員を置く

１）学会長	１名
２）日本Mテクノロジー学会大会長(以下「大会長」という)	１名
３）日本Mテクノロジーフェア実行委員長(以下「フェア実行委員長」という)	１名
４）幹事　庶務財務担当	１名
国際担当	１名
流通担当	１名
広報担当	１名
雑誌担当	１名
ネットワーク担当	１名
M言語標準化担当	１名
ＪＩＳ・ＩＳＯ担当	１名
５）会計監事	１名
６）評議員	若干名
７）日本Mテクノロジー学会出版会理事長	１名
８）日本Mテクノロジー学会出版会理事	若干名

第11条　各役員の選出または構成を次のように定める。

１）評議員に欠員が生じた場合、学会長は評議員会の推薦者を総会に諮り、その承認を得て決定する。評議員の定数は学会長が定める。但し、各評議員の構成割合は会員の職域構成割合に近いものとする。
２）学会長及び会計監事は、評議員会の推薦者を総会に諮り、その承認を経て決定する。
３）幹事は学会長が推薦し、総会の承認を経て決定する。学会長と幹事は併任できない。
４）大会長は学会長が幹事会の推薦者を総会に諮り、その承認を経て決定する。
５）フェア実行委員長は学会長が幹事会の推薦者を総会に諮り、その承認を経て決定する。
６）出版会理事長並びに理事は学会長が推薦し、総会の承認を経て決定する。

第12条　各役員の任務は次のように定める。

1）学会長は会を代表し、総会、幹事会、評議員会の議長となる。
2）大会長は、年次日本Mテクノロジー学会大会を総括する。
3）フェア実行委員長は、年次日本Mテクノロジーフェアを総括する。
4）庶務財務担当幹事は、本会に関する庶務及び全ての資金及び財産の管理を行う。また、最新の名簿の管理、総会その他の議事録の管理を行う。
5）国際担当幹事は、海外のＭＴＡ（ＭＵＧ）組織との連携並びにM言語開発委員との協力を司り、その他の国際的協力を行う。
6）流通担当幹事は、M言語応用プログラムのユーザー間相互交換の促進、ＭＵＧプロトタイプ・アプケーション・ライブラリー（ＭＵＧＰＡＬ）など M言語資源の整備、管理、維持、会員に対する資料提供等のサービスを行う。
7）広報担当幹事は,Mテクノロジーニュース等を通じ広報活動を行う。
8）雑誌担当幹事は、学会誌 「Mumps」 の編集を兼ね、出版の進行を司る。
9）ネットワーク担当幹事は、ネットワークを活用した会員間のコミュニケーションの向上を図る。
10）M言語標準化担当幹事は、M言語の標準化を図る。
11）ＩＳＯ・ＪＩＳ担当幹事は,M言語のＩＳＯとＪＩＳ標準制定に関することを司る。
12）会計監事は、年次会計の監査を行い総会に報告する。

第13条　各役員の任期を次のように定める。

1）学会長、幹事、会計監事の任期は、４月１日より翌々年３月31日までの２年間とし再任を妨げない。
2）大会長の任期は、前学会終了時に始まり学会の残務処理の終了までの期間とする。
2）フェア実行委員長の任期は、前Mテクノロジーフェア終了時に始まりMテクノロジーフェアの残務処理の終了までの期間とする。
3）評議員の任期は特に定めないが、４年間続けて評議員会に出席しなければ評議員資格を失う。

第五章　会議および委員会

第14条　（総会）

1）総会は本会の最高の議決機関である。
2）総会は学会長が毎年１回召集する。但し、幹事会の議決による場合または会員の５分の１以上から請求された場合、学会長は臨時総会を召集しなければならない。
3）総会の議長は学会長とする。
4）次の事項は総会に提出してその承認を受けなければならない。
　ａ．事業報告および収支決算
　ｂ．事業計画および収支予算
　ｃ．その他幹事会が必要と認めた事項
5）総会の成立に必要な出席者数は会員のうち50名または10％の少ない方を上回る数とする。

6）総会の議決は本規約に別に定めるものの他、出席会員の過半数による。

第15条　（幹事会）

1）学会長が必要に応じて召集する。但し、幹事の過半数から請求があった時は、学会長は幹事会を召集しなければならない。
2）幹事会の議長は学会長とする。
3）幹事会は学会長、大会長、フェア実行委員長、幹事、会計監事により構成される。
4）学会長は必要に応じて各種委員会の委員長を出席させることができる。
5）幹事会の議決は構成員の過半数による。

第16条　（評議員会）

1）学会長が毎年１回召集する。但し、学会長は必要に応じて臨時評議委員会を召集する。
2）評議員会は学会長の諮問に答え本会の重要案件を審議する。議長は学会長とする。
3）評議員会は学会長、会計監事、Mumps 編集委員、新評議員を総会に推薦する。

第17条　（学会誌 Mumps 編集委員会）

1）雑誌担当幹事は必要に応じて学会誌 Mumps 編集委員会を召集する。
2）学会誌 Mumps 編集委員会の議長は雑誌担当幹事とする。
3）学会誌 Mumps 編集委員は編集委員会が任命する。任期は３年とし、再任を妨げない。

第18条　（各種委員会）

1）学会長は必要に応じて幹事会の議を経て各種委員会を設置、統合、分化、改廃することができる。

第19条　（日本Mテクノロジー学会大会）

1）本会は年１回以上の日本Mテクノロジー学会大会を開催する。

第20条　（日本Mテクノロジーフェア）

1）本会は年１回以上の日本Mテクノロジーフェアを開催する。

第21条　（日本Mテクノロジー学会出版会）

1）日本Mテクノロジー学会出版会の規約は別途定める。

第六章　資産および会計

第22条　本会の資産は次の通りとする。

1）本会の設立当初からの財産
2）入会金および会費
3）事業に伴う収入
4）資産から生ずる利子など
5）寄付金品
6）負担金
7）その他

第23条　本会の資産は、学会長及び庶務財務担当幹事が管理する。

第24条　本会の重要な財産（基本財産）に関しては、これを消費し、または担保にしてはならない。但し、本会の事業遂行上止むを得ない理由があるときは、幹事会の出席者の２／３以上の議決と総会の出席者の３／４以上の議決を経てその一部に限り処分し、または担保に供することができる。

第25条　本会の事業計画およびこれに伴う収支予算は、年度毎に学会長および庶務財務担当幹事が編集し、幹事会の議決を経て総会の承認を得なければならない。

第26条　本会の事業報告書および収支決算は、年度毎に学会長および庶務財務担当幹事が作成し、会計監事が監査し、幹事会の議決を経て総会の承認を得なければならない。

第27条　本会支援のため各種団体よりの負担金、寄付、研究費などの交付があった場合、幹事会の承認により本会の資産として受け入れる。

第七章　規約の変更ならびに解散

第28条　本規約の改正は幹事会および総会において各々出席会員の２／３以上の議決を経なければならない。

第29条　会を解散するには総会において出席会員の３／４以上の同意を必要とする。

第30条　会の解散に伴う残余財産は、法律による制限のあるものの他は世界保健機構（WHO）に寄付するものとする。

第八章　付　則

第31条　本会の略称を日本ＭＴＡ、英文略称をＭＴＡ－ＪＰという。

第32条　本会の入会費、年会費は別に定めるものとする。

第33条　学会長は本会の発展に功績のあった特定個人に対し名誉会長、名誉会員の称号を与えることができる。

第34条

１）本規約は１９７７年１０月２９日より発効するものとする。
２）本規約は１９７９年　９月１４日より改訂し発効するものとする。
３）本規約は１９８７年　７月２９日より改訂し発効するものとする。
４）本規約は１９９１年１０月３１日より改訂し発効するものとする。
５）本規約は１９９２年　８月　１日より改訂し発効するものとする。
６）本規約は１９９２年１０月２９日より改訂し発効するものとする。
７）本規約は１９９３年　４月　１日より改訂し発効するものとする。
８）本規約は１９９４年　８月　６日より改訂し発効するものとする。
９）本規約は１９９５年　９月３０日より改訂し発効するものとする。
１０）本規約は１９９６年　９月１５日より改訂し発効するものとする。

※　学会規約につきましては、現在改訂作業を行っております。
掲載内容は2013年8月時点のものです。

「Mumps」投稿規程

（1991 年 7 月 10 日制定）
（1994 年 12 月 1 日改正）
（2008 年 11 月 24 日改正）
（2012 年 9 月 12 日改正）

本規定は日本 M テクノロジー学会誌「Mumps」に、会員が自発的に寄稿する論文（以下投稿論文という）に関する必要事項を定めたものです。学会誌「Mumps」には、編集委員会が依頼する原稿（依頼原稿）も掲載しますが、それについての必要事項はそのつど定めます。

１．論文の主題

投稿を受ける論文の主題は、コンピュータシステム／言語である MUMPS に直接、間接に関係するものとします。

例えば、MUMPS の利用技術についての考案や開発、MUMPS の言語についての直接仕様や提言、MUMPS システム装備、MUMPS と他の世界とのインターフェース、MUMPS の教育など、MUMPS に関係するあるいは関係しそうなテーマについて広く受け入れます。ただし、他の雑誌に掲載された、あるいは投稿中の論文はお断りします。

２．投稿論文の種類

投稿論文は次の 6 種類に限ります。

１）原著論文

未投稿で、論文の主要部分に独創性、独自性のある論文。既に発表した問題について別の視点からまとめた論文も未投稿原著論文であり得ます。また、応用開発、調査等であっても、その過程で創意工夫や独自性があれば原著論文の対象とします。

２）総説

ある主題について、過去の研究業績を詳細にまとめ文献を伴って記述し、その主題に関する現状と将来展望を明らかにした論文。

３）研究速報

新しい研究成果が原著になるほどにはまとまっていないが発表に価値があると考えられるもの。

４）技術ノート

作成したプログラムや新しいシステムの紹介など、MUMPS 技術に関する論文で、会員の相互の利益になると思われるもの。

５）フォーラム

意見、提案、提言、感想、著書や学術集会の紹介など、上記以外で会員の利益になると思われるもの。

６）Letter to the editor

原著論文に対する質問やコメント、日本 MTA の活動に関係のあるコメントなど。

３．投稿論文の長さ

原則として下記の表の通りの長さとします。Ａ４用紙（横 21 字×縦 41 行×2 段組＝1722 文字）で刷り上がりページ１枚となります。ただし、これを越える場合でも、編集委員会が必要と認めた場合には別に定める超過料金を支払って掲載することができます。

論文の種類	論文のページ数（刷り上がり）
原著	１０ページ（以内）
総説	３０ページ
研究速報	６ページ
技術ノート	６ページ
フォーラム	４ページ
Letter to the Editor	１ページ

４．投稿者の条件

１）筆頭著者は日本 M テクノロジー学会会員であること。

２）共著者も原則として会員であることとします。

５．原稿の送付

原稿（2 段組の印刷形式原稿でも可）を下記編集委員会宛てに送って下さい。原稿到着日を投稿の受付け日としその日付を誌上に明記致します。

原稿送付先・連絡先

〒260-8677

千葉市中央区亥鼻 1-8-1

千葉大学医学部附属病院　企画情報部内

日本Mテクノロジー学会事務局　担当：土井俊祐

TEL：043-226-2346　FAX：043-226-2373

e-mail：　mta-office@mta.gr.jp

６．掲載の採否

投稿された原稿は、編集委員会が依頼する２名の査読者が査読します。そしてその査読の意見を考慮して編集委員会がその原稿の採否を決定します。査読の結果によっては、原稿の内容や論文の種類を修正変更することを投稿者にお願いすることもあります。

７．原稿作成要領

１）原稿の構成

投稿原稿はおよそ次の構成に従って作成して下さい。

a）論文の題名

b）著者名、所属、所在地

a）と b）は日本語と英語の両方を記入して下さい。

c）キーワード・・・８語以内（日・英）

d）和文要旨・・・・２００字から４００字

e）英文要旨・・・・２００words から３００words

f）本文

g）謝辞・・・・・・必要に応じて

h）文献リスト

文献の引用は本文中の引用箇所に出現順に通し番号[１]，[３－５]等を記し、本文の末尾に一括して引用番号順に並べて下さい。雑誌の文献は引用番号、著者名、論文題名、雑誌名、巻号、最初と最後の頁数、西暦年号の順です。

単行本の文献は引用番号、著者名、題名、版数、引用頁、発行社、発行地、西暦年号の順です。

（例）

１．福井太郎：糖尿病患者管理システムの開発，医学情報学，10(2):30-35(1990).

i）図表

図や表はそれぞれを本文中に入れて下さい。図や表の大きさは基本的に著者の意向に沿いますが、大き過ぎると判断された場合は、縮小されることがあります。

j）特殊文字

特殊文字は原則として禁止しますが使用される場合は使用位置を別紙にて明示して下さい。

２）投稿原稿（FD, CD 等の記憶媒体または電子メールで提出）

原稿は標準的なワープロ（一太郎、MS−Word）で、Ａ４用紙に横４２文字×縦４１行を１頁として作成して下さい。印刷原稿の形式でも受付けます。

また、原稿には表紙を付け、表紙にはつぎの事項を記入して下さい。

表紙・・・題名

連絡先（氏名・住所・電話・FAX）

原稿の種類

原稿の枚数（本文・図・表別に）

別冊希望部数（５０部の倍数）

その他・・特殊文字等を使用されている場合は明記して下さい。

3）印刷原稿（FD, CD 等の記憶媒体または電子メールで提出）

採用が決定した印刷原稿は、標準的なワープロにて A4 用紙 2 段組のカメラレディの原稿にて提出して下さい。

（印刷原稿、カメラレディ原稿作成時の注意事項）

＊基本的に横 21 字×縦 41 行×2 段組が 1 頁になって印刷されます。

＊原稿は題名（日本語）、題名（英語）、著者名（日本語）、著者名（英語）、著者所属・住所（日本語）、著者所属・住所（英語）、和文抄録、キーワード（日本語）、英文抄録、キーワード（英語）、本文の順で同一ファイル名に保存して下さい。

＊印刷は、モノクロで行われます。原稿にてカラーを使われる場合には、この点に留意して原稿を作成して下さい。

＊なお、編集側にてタイトル、著者名、所属、要旨の形式の統一を行います。また、タイトルページの左下に原稿受付の月日、査読後の受理月日を入れますので提出して戴いた原稿と異なることがあります。

８．版権

掲載論文の版権は日本 M テクノロジー学会に帰属するものとする。

９．別刷

別刷は 30 部まで無料とし、それ以上は実費とする。別刷の部数は投稿時または校正原稿提出時に申し出て戴ければ、10 部単位で増刷する。

編集後記

皆様のご協力を得まして学会誌「Mumps」の第 27 巻を発行することができました。今回は編集委員長をされていた木村一元前学会長の急逝により、春木が引き継ぐ形となりました。前回の刊行から 2 年半と長い期間を要しましたが、査読者の協力を得て、何とか発行の運びとなりました。

今回の雑誌は 5 編の掲載論文のうち 3 編が学生からの投稿と、若返りを象徴するものとなりました。また、内容としても M の基礎技術をはじめとして、蓄積されたデータの応用に M を利用するものまで、非常に多彩なものとなりました。

今後、種々の分野において M の柔軟な機能を生かした新たな利用法や、それに基づく知見が報告され、各分野の活動が一層活発になることに期待します。

最後になりますが、当会の発展に多大なるご尽力をいただいた木村前学会長のご功績をたたえますとともに、謹んでご冥福をお祈り致します。

Mumps (The Official Journal of M-Technology Association-Japan)

第２７巻　　２０１４年2月１７日発行

発行者　日本Mテクノロジー学会
会長　　土屋　喬義
〒346-0003　埼玉県久喜市久喜中央 3-1-10
土屋小児病院
TEL：0480-21-0766
FAX：0480-21-2230

編集者　日本Mテクノロジー学会　編集委員会
委員長　　春木　康男
〒251-1193　神奈川県伊勢原市下糟屋 143
東海大学医学部　基礎医学系　医学教育・情報学
TEL：0463-93-1121　Ex.2140，2143
FAX：0463-93-5418

事務局　日本Mテクノロジー学会　事務局
庶務財務担当幹事　　鈴木　隆弘
〒260-8677　千葉県千葉市中央区亥鼻 1-8-1
千葉大学医学部附属病院　企画情報部内
TEL：043-226-2346
FAX：043-226-2373

印刷　　三陽メディア株式会社
〒260-0824　千葉市中央区浜野町 1397
TEL：043-266-8437
FAX：043-266-1498