科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１０　日現在

機関番号：３４４１７
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５９０６２９

研究課題名（和文）診療チーム意思決定のバスケット分析とＡＰＭＬを用いた道筋解析と行動予測

研究課題名（英文）Analysis on decision-making process and prediction of subsequent behavior of medical care team using medical records with the basket analysis and attention profiling mark-up language

研究代表者
渡辺　淳（WATANABE, Jun）

関西医科大学・医学部・准教授

研究者番号：４０１４８５５７

交付決定額（研究期間全体）：（直接経費）　　1,200,000円、（間接経費）　　360,000円

研究成果の概要（和文）：診療チームにおける意思決定の道筋解析と展開予測法を確立するために、非構造化データである日本語自然文で記載された診療記録のアセスメント記述の解析にAPMLとアソシエーション分析を適用した。診療記録のアセスメント記述から予測した診療計画と、実際の処置・処方との比較から、主にアソシエーション分析を用いた隠蔽情報・暗黙知の顕在化によって的中率の向上(40%前後->60～70%)が示された。他方、自然文処理精度の重要性が判明したため、異なったアルゴリズムを組み合わせた構文解析法を開発して日本語自然文の処理過程に導入し、予測精度を向上させ（的中率約80%）、意思決定の過程と展開の予測を概ね可能とした。

研究成果の概要（英文）：To establish a suitable method for determining intendment formed between a medical care team staff and predict the subsequent behavior, we examined the suitability of the association (basket) analysis and the attention profiling mark-up language for this purpose by exposing hidden or implicit relationship from medical records. Comparison with medical plan written in assessment and the actual records of treatments, predictive value was improved by the exposure of implicit knowledge and information using the association analysis (from around 40% to 60 to 70%). Since the importance of the accuracy of natural language processing has been found, we thus developed and introduced a method for syntactic analysis by a combination of the dependency parsing and the syntactic tree analysis. This procedure allowed us to figure out the decision-making process and the subsequent behavior of the team at a certain level (about 80%).

研究分野：医歯薬学

科研費の分科・細目：境界医学・医療社会学

キーワード：医療情報学　診療記録　意思決定　展開予測　バスケット分析　APML　自然言語処理　非構造化データ

１．研究開始当初の背景

電子カルテ使用による情報の共有化によってチームベースの診療情報の共有が容易となった。また、セマンティック web の導入によってディジタル化された記載情報からの意味・意図を含めた情報抽出が可能となりつつある。そこで、チームに属する個人の意志・意図を診療記録の記載とメンバー相互の関係とを組み合わせ、その後の診療の展開を推定することが可能になれば、より良いチーム医療の実現に資するための多くの知見が得られることが期待できると考えられる。しかしながら隠蔽・暗示的な関係検出およびそれらから事後の展開を予測するための好適な手法がなく、それらの手法の確立が待たれていた

種々のオブジェクトの相互関係は、基本的にノードとエッジ（リンク）からなる相関図として描出する事が可能である。普遍的にノードとエッジを記述可能な方法に、XML の書式を応用した Attention Profiling Mark-up Language (APML) があり、APML では隠蔽された関係や行為者の意図を表現すること、および以降の展開を予測する事などが可能とされている(http://apml.pbworks.com/)。

他方、申請者は、多数の送信サーバ・ドメインから特定のアカウントに対し、送信サーバ・ドメインを頻繁に変更しながら迷惑メールを送信する標的型 snowshoe 攻撃集団の関係解析と送信ドメイン予測に APML を適用し、APML に高い隠蔽・暗黙関係顕在化能力および展開予測能力を有することを明らかにした（渡辺ら、医療情報学 29S:859-864, 2009；第 14 回日本医療情報学会春季学術大会-シンポジウム 2010 in 高松)。これらの結果は、隠蔽・暗黙関係の顕在化およびその後の展開予測に APML が有用である可能性を強く示唆する。また、上述の研究を遂行する過程で、隠蔽情報の顕在化にアソシエーション（バスケット）分析を用いたデータマイニングが有用であることも判明した。そこで、我々は APML とアソシエーション分析を併用した展開予測アルゴリズムを開発し、構造化データを対象とした比較的単純な事象の高精度予測を可能とした。

２．研究の目的

本研究は、患者の病態に応じて刻々変遷する診療チームにおける治療方針の意志決定の道筋解析とその後の展開を予測する手法の開発を目的とする。この目的達成に向け、本研究では APML とアソシエーション分析を併用し、構造化データを用いた試行系（SMTP ヘッダ解析を用いた spam 送信組織の関係解析・意思推定と展開予測）から得られたアルゴリズムを改善しながら診療記録の記述解析に適用・検証することで、非構造化データを用いた意思決定の道筋解析と展開予測を試みた。

３．研究の方法

(1) 概要

本研究では、道筋解析・展開予測アルゴリズムを、まず、単純な構造化データ（SMTP ヘッダ）を用いて検証し、その結果に応じて非構造化データ（日本語自然文で記載された診療記録）の解析に適用する方略を採った。

解析対象としたデータは、診療記録のうち「アセスメント」項に日本語自然文で記載された記述とした。

(2)研究 デザイン

検証に際しては、まず、蓄積されたデータを用いて、後ろ向き（レトロスペクティブ）アプローチを用いた診療チーム意志決定推測アルゴリズムの改良と検証を実施した。具体的には、まず、蓄積データを形態素分析によって単語に分割した後、キーワードを任意に抽出してデータを定型化（構造化）して検証した。次に、意味解析を行って抽出した構造化データを用いて検証を行った。

次に、上述の結果にもとづいて、構造化データおよび非構造化データを用いた展開予測の前向き（プロスペクティブ）アプローチによるリアルタイム検証、を行った。

なお、研究遂行の過程で、日本語自然文の処理が精度を大きく左右することを示唆する結果が得られた。そこで、個人情報を除去した後に解析用サーバに移出した日本語自然文記述について形態素解析を実施し、次に得られた品詞情報付き単語リストを、オープンソースソフトウエアを用いて構文木解析および係り受け解析に供した。

(3) 解析方法

解析用データファイルは電子カルテの SOAP（主観的所見、客観的所見、評価、治療指針と指示）記載から個人情報を除いて転記（複写）した。形態素解析で単語に分かち書きした後、当初は用手法で抽出または意味解析を経てキーワードを抽出し、データを記載項目ごとにタグ付けして XML 形式で記載し解析用データファイルを APML のリンク先として指定した。

APML 解析は XML:: APML モジュールを組み込んだ Perl 記載のスクリプト（LyoKato, http://search.cpan.org/~lyokato/XML-APML-0.04/lib/XML/APML.pm. 2007）と、APML 仕様に準じて記載した APML ファイル(渡辺ら、医療情報学 29S:859-864, 2009）を用いた。事象（ケース）ごとに事後の結果から項目ごとに感度と特異度を算出し、得られた尤度比を APML の Profile に、明示的関係については<ExplicitData>、隠蔽・暗黙的関係については<ImplicitData>の項に、それぞれ 0－1.0 の重みに換算して登録した。

アソシエーション（バスケット）分析は、第一段階では、抽出した単語について、それらの特徴を「前提部（条件）(antecedent)」、その結果がアセスメントの結果を受けて記

載される Plan の項目において、処置・処方事項または他のチームメンバーへの指示に反映されていることを帰結部（consequent)」とし、アプリオリ（Apriori）のアルゴリズムに準じて、信頼度（confidence） と支持度（support)を 算出した。また、クリニカルパス記載事項や該当する疾患または投与する薬剤の効果や副作用に関する情報が記載に欠落している可能性をバスケット分析を用いて推定し、必要事項を顕在化させて解析対象のデータに随時追加した。

形態素解析には MeCab および IPA 辞書と ComeJisho を用い，構文木解析には Cocke-Younger-Kasami (CYK 法)、係り受け解析には CaboCha)を用いた。なお、この過程で日本語自然文を高精度で解析可能な２段階アルゴリズムの前段部を構築した（後述)。

因子分析、多重対応分析、アソシエーション（バスケット）分析およびベイズ推定には GNU R を用い、有意水準を 5%として検定した。

４．研究成果

(1) 単純構造化データ（SMTP ヘッダ）を用いた予測アルゴリズムの検証

① アルゴリズムの事前検証 「意思決定の道筋・展開予測アルゴリズム」の診療データ解析への応用に際し、このアルゴリズムが持つ問題点の洗い出しを,構造化データ(SMTP ヘッダ)を用いて試みた。アルゴリズムの概要は、まず、既知のデータから特徴を抽出し、APML とアソシエーション分析を用いてそれらの特徴の信頼度、支持度、相互関係から次に用いられる可能性の高い送信サーバ（ドメイン)、送信アドレス等を推定するものである。推定された事項をフィルタに実装して推定精度を検証した。この時点での本アルゴリズムを用いたフィルタによる snowshoe spam の検出精度は 98%以上で、推定結果にもとづいた予測精度（それまで受信歴のなかった送信サーバからの snowshoe spam の捕捉率）は 95%以上であり、それ以外の spam/scam（迷惑メール/詐欺・情報詐取メール）を含めた検出・予測精度も 95%以上を示していた。

問題点の洗い出しによって、この高検出率・高予測精度が、それらの spam/scam の送信特性の検出に都合良く合致している可能性が示唆された。Snowshoe spam は送信者が IP アドレス領域を次々と取得してそこに多数のメールサーバを置いて送信する方式である。そこで、APML・アソシエーション分析によって Whois 情報を顕在化し、特定の送信者・送信組織がいつ、どこの送信ドメイン（IP アドレス範囲とドメイン名）を取得したかを随時フィルタに追加してゆくだけで、結果的に好結果得られている可能性が判明した。また、この時点での spam の大半は、不正プログラムが感染した個人ユーザの PC (bot net) から送信されており、APML、アソシエーション分析を用いた予測に際しては、ISP 等がユーザ用にアサインした IP アドレス範囲を特定し、その情報と頻用される送信アドレスとを組み合わせるだけで、高効率な検出・補足が実現できていることが判明した。これらのことから、このアルゴリズムをより複雑なデータに適用する際には、複雑なデータに対応したアルゴリズム検証のためのモデルが必要であることが示唆された。

一方、ここまでの検討の過程で、このアルゴリズムでは検出が容易でない迷惑メールが存在することが判明した。それは 419 scam と称される詐欺メールで、一部は bot net から送信されているものの、多くは ISP や企業の正規メールサーバおよび大手フリーメール業者の正規メールサーバから送信されており、この時点での検出効率は 25～30%に留まっていた。

② アルゴリズムの改善

上述の事項に加えて、APML では陳古化したデータの処理が容易でなく、予測時に古いデータ(out of date)の特性が混入することで予測精度を低下させる可能性があることが判明した。そこで、従来は APML が主、アソシエーション分析が従であった方式を、アソシエーション分析を主とし APML を補足手段に用いる方式に変更し、419 scam の出現予測とリアルタイム検出に適用し、随時検証を実施しながら、成果を診療記録の解析にフィードバックした。アルゴリズムの改良によって、偽装情報が含まれた SMTP 程度のデータの種類・量でも 60-70%の的中率が得られるようになった。次に、送信者の送信方略に関する意図を顕在化させるステップをアルゴリズムに組み込んだうえでエンベロープ情報をバスケット分析とベイズ推定を用いて解析する手法に改良することで、419scam の高効率の検出と阻止（感度 95%以上、特異度 99.8%以上)が可能となった。

他方、展開予測アルゴリズムの妥当性検証および、アソシエーション分析と APML の予測能力特性の評価を目的とし、某フリーメール業者から当方の受信中継サーバ群に送信されている漏洩クレデンシャルを利用した詐欺メール（419 scam と phishing メール）の SMTP ヘッダ情報を材料とし、それらの検出効率を指標として予測と検出をリアルタイムで（プログレッシブに）解析した。
検証期間中に捕捉した詐欺メールのエンベロープ送信者アドレスの 20%弱が漏洩・流出したもので、アソシエーション分析、APML とも、流出ログイン情報は詐欺メールのフィルタリング回避を企図したものと推定した。アソシエーション分析の結果に基づいたフィルタでの検出精度は 98%以上だったか APML では 90%に達しなかった。

これらの、少量データを用いた意思決定の道筋解析と展開予測に際しては、a．アソシエーション分析を主とするのが好適であること、b．暗黙知を顕在化して形式知とすることが重要であること、c．適切な構造化デ

ータを解析対象とした場合には 95%以上の予測精度が期待できること等が示され、これらの結果を、随時、診療記録の解析にフィードバックした。

(2)日本語自然文の解析による意思決定道筋解析と展開予測のためのレトロスペクティブ解析

①抽出キーワードを用いた解析

転帰・事後の結果が判明している 76 事例のカルテのアセスメント記載について形態素分析を用いて単語分かち書きの後、上述（当時は改善中）のアルゴリズムを適用し、アルゴリズムの問題点を洗い出した。対象とした診療記録は１日毎に主観的情報（S，問診結果等)、客観的情報（O,検査結果等)、評価（A，アセスメント；得られた情報の評価と診療方針の策定等)、処方・処置の指示（P，プラン;オーダ発行記録や診療チームメンバーへの指示等）が SOAP の順に記載された問題指向型診療記録で、アセスメント項の記載事項がプランに反映されているかどうかの的中率によってアルゴリズムの精度判定を試みた。

記載量が多い場合には、ある程度期待どおりの結果が得られる傾向があった。しかしながら、記載が短いフレーズや単語の羅列であった場合の予測能力は低かった。また、経過に問題が生じたケースでは記載が多く、順調だった場合には記載が少ない傾向があることから、評価に際して記載量の多寡によるバイアスが生ずることが判明した。

RDF/RDF Schema に基づいて記載事項を定型化変換して上述の事例を再解析した。定型化によってデータ量の多寡によるバイアスはやや軽減したが、予測能力の低下は顕著となった。この段階におけるアセスメントの全記載事項に関するプラン記載の的中率は、40%弱に留まった。

これらの過程で、APML 解析前のバスケット分析による特徴抽出精度が結果に影響を与えること、および予測精度を向上させる動詞・助動詞の使用頻度が診療記録の記載では一般文（新聞等）に較べて低いことが判明した。

②自然言語処理を用いた解析

同病名が付与された患者の電子カルテから匿名化して抽出したアセスメント（日本語、自然文）およびプランの記載事項について、分かち書きの後、上述の抽出法の代わりに用手法または既知の意味・文脈解析法を用いて推定し、推定結果をプラン記載のオーダ発行歴と比較して推定精度を調べるとともに、結果を因子分析、多重対応分析、アソシエーション（バスケット）分析およびベイズ推定に供して記載者の治療指針・アウトカムの推定を試みた。併せて、推定結果に影響を及ぼす要因の洗い出しを試みた。

用手法または既知の意味・文脈解析法を用いた際の予測精度も上述の方法と同様、40%前後と低く、上述の方法と用手法を組み合わせた場合のみ 60%弱の値に到達した。この原因を調べたところ、要因として a．上述医療従事者・診療チームの暗黙知が診療記録に記載されない、b.クリニカルパス適用症例ではパスの内容を補充する必要がある、c．記載量の多寡によって推定精度が大きく影響されること等に加え、d．自然文処理の方法が精度を左右すること、および e．同義語・同義フレーズ記載対応のためのシソーラス整備の必要性が示唆された。

(3) 隠蔽情報の顕在化と構文解析の導入

(2)項と同一の記載事項にクリニカルパス記載情報を顕在化して補い、併せて類語処理を行ったデータついて、CYK 法を用いた構文木解析および係り受け解析を行って単語の相互関係を明確化するとともに、因子分析、多重対応分析、アソシエーション（バスケット）分析およびベイズ推定を行って記載者の治療指針・アウトカムの推定を試みた(図１)。

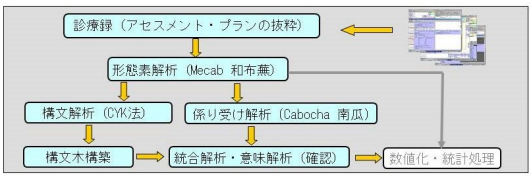

図１　構文解析を導入した処理フロー

まず、異なった原理に基づく２つの構文解析手法（構文木解析および係り受け解析）を組み合わせた構文解析プロセスを導入した（図２)。

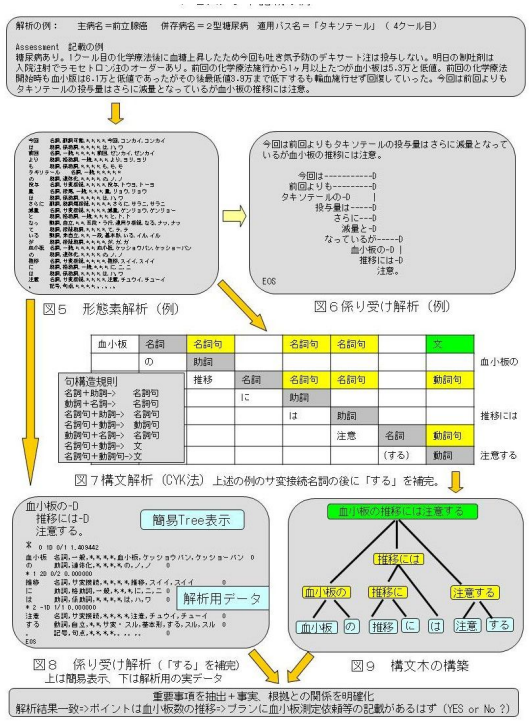

図２　構文解析の例

同時に、隠蔽情報・暗黙知の顕在化、およ

び同義語（類語）対応のためのシソーラスの整備（形態素解析による分かち書き時に用いる辞書に実装）を実施した。これらのことによって単語の相互関係が明らかとなり、隠蔽情報・暗黙知を補完した場合の的中度（予測精度）は 80%に達し、研究開始当初に設定した最低限の目標値には到達した（図３）。

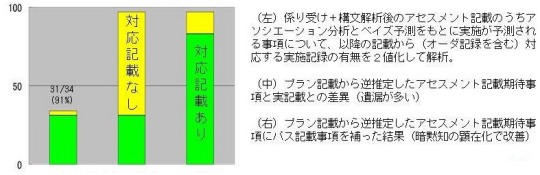

図３　隠蔽情報顕在化と構文解析導入による的中度（予測精度）の向上

また、これまでの解析結果で得られた低値が、顕在化情報の不足（例えば、クリニカルパス適用症例では、パスと差異のあった事項しか記載されない傾向がある）に起因する部分が大きいことも確認できた。しかしながら、(1)項で示したアルゴリズムの予測精度の期待値が 95%を超えるのに対し、同アルゴリズムを用いた診療記録解析についての予測精度は 80%強で足踏み状態となった。

(4) 構文解析とパラレルコーパス・機械学習を組み合わせた自然言語処理法の提案

上述の「80%の壁」を乗り越えるために、解析過程を段階ごとに再調査した。その結果、上述の精度向上を阻む要因が、日本語自然文記述における単語の相互関係把握の正確性不足にあることが判明した。このことから、解析対象となる文は、機械的（自動）処理が可能な正しい構文で単語相互の関係が明確な記述であることが求められる。しかしながら、診療記録への自然文記載に際して、記載者が構文および単語の相互関係を意識して記述することは、現実には容易でないと考えられる。

そこで、本研究で開発した「異なったアルゴリズムを組み合わせた構文解析法（構文木解析・係り受け解析）」を前段とし、この手法で前処理した文を後段で用手的または機械学習法を用いて作成したパラレルコーパス（診療録記載に頻用される文とそれと同意味の解析容易な文との対比文例集）を用いて機械的処理が可能な可読性の高い構文に変換する手法を考案するに至った。現在、この手法で解析した診療記録記載日本語自然文を対象とした予測精度は、対象とする疾患によって差異はあるが、高いものでは 90%以上に達しつつある。この手法はパラレルコーパスの作成に時間を要する点などの欠点はあるものの、日本語自然文で記載された診療記録の２次利用を容易にする一手法として有用性が期待できると考えられる。

５．主な発表論文等

〔雑誌論文〕（計５件）

① 渡辺　淳、仲野俊成、松本掲典、新貝欣久、高木真平、西野典宏、某大手フリーメール業者から送信される詐欺メールにおける流出ログイン情報の利用、査読無、医療情報学 33 (S)1006-1009 2013

② 渡辺　淳、仲野俊成、日本語自然文で記載された診療記録の構文解析の試み、査読無、医療情報学 33 (S)832-835 2013

③ 渡辺　淳、仲野俊成、北村　臣、電子カルテに記載された少量情報を用いた意思決定の道筋解析と展開予測に及ぼす暗黙知の影響、査読無、医療情報学 32 (S)1458-1460 2012

④ 渡辺　淳、仲野俊成、松本掲典、新貝欣久、高木真平、SMTP ヘッダにおける限局された情報を用いた詐欺メール送信者の意志決定の道筋解析と展開予測に及ぼす暗黙知の影響、査読無、医療情報学、32 (S)642-645 2012

⑤ 渡辺　淳、仲野俊成、松本掲典、新貝欣久、SMTP エンベロープの特徴解析による 419 scam メッセージの検出、査読無、医療情報学 31 (S)699-702 2011

〔学会発表〕（計５件）

①渡辺　淳、仲野俊成、松本掲典、新貝欣久、高木真平、西野典宏、某大手フリーメール業者から送信される詐欺メールにおける流出ログイン情報の利用、第 33 回医療情報学連合大会 2013 年 11 月 21 日～2013 年 11 月 23 日 神戸ファッションマート

② 渡辺　淳、仲野俊成、日本語自然文で記載された診療記録の構文解析の試み、第 33 回医療情報学連合大会 2013 年 11 月 21 日～2013 年 11 月 23 日 神戸ファッションマート

③ 渡辺　淳、仲野俊成、北村　臣、電子カルテに記載された少量情報を用いた意思決定の道筋解析と展開予測に及ぼす暗黙知の影響、第 32 回医療情報学連合大会 2012 年 11 月 14 日～2012 年 11 月 17 日 新潟朱鷺メッセ

④ 渡辺　淳、仲野俊成、松本掲典、新貝欣久、高木真平、SMTP ヘッダにおける限局された情報を用いた詐欺メール送信者の意志決定の道筋解析と展開予測に及ぼす暗黙知の影響、第 32 回医療情報学連合大会 2012 年 11 月 14 日～2012 年 11 月 17 日 新潟朱鷺メッセ

⑤ 渡辺　淳、仲野俊成、松本掲典、新貝欣久、 SMTP エンベロープの特徴解析による 419 scam メッセージの検出、第 31 回医療情報学連合大会 2011 年 11 月 21 日～2011 年 11 月 23 日 鹿児島市民文化ホール

６．研究組織

(1)研究代表者

渡辺　淳（WATANABE, Jun）
関西医科大学・医学部・准教授
研究者番号：４０１４８５５７