一般口演
# 一般口演19
# 医療データ分析７（ DWH・臨床研究データベース）
2018年11月25日(日) 09:00 〜 10:30 F会場 (5F 502+503)

[4-F-1-3] PDF形式で保存される検査レポートから特定の結果値を取得するプログラムの開発

○張 冬堯, 和田 聖哉, 中川 彰人, 真鍋 史朗, 武田 理宏, 松村 泰志 （大阪大学大学院医学系研究科　医療情報学）

背景：電子カルテデータの二次利用に際して、後ろ向き研究では、非構造化データの取得が課題となる。検査レポートは、検査種ごとに特徴的なレイアウトとなっていることが多い。目的： PDF形式で保存される検査レポートから、臨床研究で必要な結果値を取得するプログラムを作成し、その精度を検証すること。方法：プログラミング言語は pythonを使用した。 PDF形式は、文字列とその座標で定義される。最初に複数ページから目的のページを同定するため、目的ページのみに存在する文字列をマーカとして設定した。唯一のマーカがない場合は複数のマーカあるいはマーカとその座標から同定を行った。次に目的データの取得を行った。目的データが文字列である場合、同じ列のマーカから目的データを取得した。目的データが表形式の値である場合、表の横と縦の第一列をマーカとして座標を標記し、目的データを取得した。スキャン PDF文書については、 OCRソフトウェアの精度影響があるため、座標の容認範囲を設定した。最後に、目的データが数値である場合は、取得値があらかじめ設定した範囲に入らない場合はエラーを返すことで精度管理を行った。結果：本プログラムを用い、スキャン PDF文書である X線骨密度測定検査から腰椎と大腿骨の骨密度値を取得した。目的ページの同定マーカとして「 f Left Hip」、「 f Right Hip」、「 f Lumbar Spine」と「 Total」を用いた。患者識別情報は、「 Patient ID」、「 Scan Data」をマーカとし取得した。骨密度値は表の横「 BMD」、「 PR」と縦「 Neck」、「 Total」をマーカとして取得した。延べ598患者、2,735ページのレポートを処理した。目的のデータは1,057ページで記載され、精度チェックでエラーは165件であった。エラーなく取得できたデータ892件は、目視による確認の結果、すべて正しい値が取得できていた。結語：電子カルテに蓄積される PDF文書から特定の検査値を取得することが可能であった。

PDF 形式で保存される検査レポートから特定の結果値を取得するプログラムの開発

張冬堯[*1]、和田聖哉[*1]、中川彰人[*1]、
真鍋史朗[*1]、武田理宏[*1]、松村泰志[*1]

*1 大阪大学大学院医学系研究科医療情報学

Program development to acquire specific result values from inspection report saved in PDF format

DONGYAO ZHANG[*1], Syo-ya Wada[*1], Akito Nakagawa[*1],
Shirou Manabe[*1], Toshihiro Takeda[*1], Yasushi Matsumura[*1]

*1 Osaka University Graduate School of Medicine Department, Medical Informatics

In the secondary use of electronic medical record data for retrospective studies, acquisition of unstructured data becomes an issue. Inspection reports are often characteristic layouts for each type of examination. Our study was to develop a program to obtain the result values necessary for clinical research from the inspection reports saved in PDF format and to verify its accuracy. We used the programming language python to develop the program. In the program, we read contents of PDF files into strings by utilizing PDF layout characters and coordinate information. And then, we search for the target data by keyword and coordinate. As for scanned PDF documents, we used OCR software ABBYY FineReader 14 to convert images of printed text into machine-encoded text. Due to the accuracy influence of OCR software, we set the acceptance range of coordinates and adjusted the length of keywords. Using this program, we conducted an experiment to acquire the bone density values of the lumber spine and femur from the X-ray bone density measurement reports which are the scanned PDF documents. As a result, it is possible to acquire specific result values from PDF documents stored in the electronic medical records.

Keywords: PDF, OCR, Acquired, Data Accuracy

1. 緒論

電子カルテの普及に伴い、多施設の電子カルテデータの臨床研究等への二次利用が取り組まれている。MID-NET（Medical Information Database Network）は国の医療情報データベース基盤整備事業で電子カルテやレセプトデータを匿名化して収集しているが、その対象は処方・注射オーダや検体検査情報など、電子カルテに構造化して蓄積されている情報に限定されている [1)]。一方、臨床研究での二次利用を考えると、電子カルテに記載する経過記録等のデータを構造化して取得する必要がある。診療録直結型全国糖尿病データベース事業（J-DREAMS）では、処方・注射オーダ等に加え、医師が記載する経過記録を、テンプレートを使用することで構造化して、データ収集している [2)]。我々は、電子カルテに電子症例報告書システムを組み込み、テンプレートを用いて診察記事を記載しながら症例データを蓄積するシステムを構築し、電子カルテメーカー4 社の電子カルテを導入する 15 病院に展開している [3)]。これらの構造化データ取得の取り組みは前向き研究で可能となる。一方、ビッグデータを用いた臨床研究を考えると、電子カルテに蓄積されたデータを用いた後ろ向き研究を考慮する必要がある。このためには、電子カルテに蓄積された非構造化データの利活用が課題となる。

心臓超音波検査や呼吸機能検査、X 線骨密度測定検査などの検査レポートは、この中に記載される数値データが臨床研究に用いられることが想定される。これらのレポートはフリーテキストで記載されることが多く、そのデータは部門システムのデータベースへの保管、システム連携やスキャンにて PDF 化して電子カルテデータベースへの保管することが想定されるため、多施設から構造化データを収集して臨床研究に二次利用することは容易ではない。一方、これらのレポートは施設ごとに異なるフォーマットではあるものの、施設内では統一のフォーマットで記載され、臨床研究で必要となる数値データは表形式で決まったところに記載されていることが多い。このため、PDF 形式でレポートを収集し、その PDF ファイルから目的の数値データを収集することができれば、臨床研究での二次利用が可能となる。

2. 目的

電子カルテに蓄積された PDF 形式の検査レポートから、臨床研究で必要な結果値を取得するプログラムを作成し、その精度を検証することを目的とする。

3. システム概要

本システムはプログラミング言語 python を使用した。PDF コンテンツに対する操作に、pdfminer という python ライブラリを使用した。指定のフォルダに保存した PDF ファイルを読み込むために、python の glob モジュールを利用した。最後の目的データを書き込みすることは python の csv モジュールを利用した。

読み込んだ PDF コンテンツは、文字列とその4つの頂点の座標がページの左隅を原点座標（0,0）として定義される。文字列は先に Y 軸座標値の降順、次に X 軸座標値の昇順で並べられる。

本システムの流れは、図1の通り、PDF コンテンツの読み込み、目的ページの同定、目的データの取得、取得データの精度管理、取得データの書き込み、五つの部分で構成された。

3.1 目的ページの同定

最初に、複数のページで構成されるレポートから、目的データの記載されるページを同定するために、目的ページのみに存在する文字列をマーカーとして設定した。目的ページの

みに存在する唯一のマーカーがない場合は、複数のマーカーあるいはマーカーとその座標値からページの特定を行った。

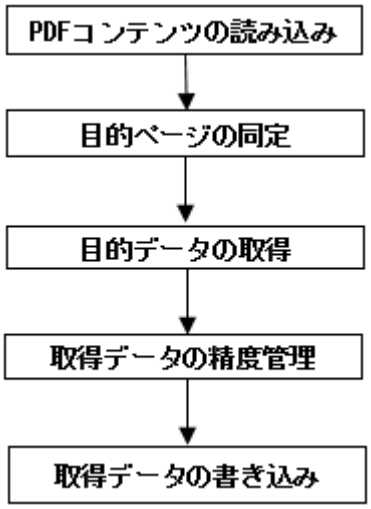

図1 システムの流れ

3.2 目的データの取得

次に、目的データの取得を行った。目的データが文字列の形式である場合、同じ列の記載が不変の文字列をマーカーとして利用した。例えば、「患者 ID:」をマーカーとして、ページの中で目的の文字列を探し、その後に続く数値を患者IDとして取得することとした。

目的データが表形式の値である場合、表の縦の第一列（タイトル列）の文字列、横の第一行（タイトル行）の文字列をマーカーとして、その文字列が記載される座標を取得した。次に、縦の第一列のマーカーの Y 座標と、横の第一行のマーカーの X 座標に記載される文字列を目的データとして取得した。

表１ 検査レポートの検査結果値表

Region	Area	BMC	BMD	T-Score	PR	AM
	(cm^2)	(g)	(g/cm^2)		(%)	(%)
L2	10.88	10.52	0.967	-0.5	95	160
L3	9.92	11.50	1.159	1.0	111	170
L4	9.03	12.02	1.331	2.4	126	180
Total	29.84	34.04	1.141	1.2	113	163

表1のような検査結果値表が例として、具体的なデータ取得手法を説明する。表1の中で、「1.141」と「113」二つの検査値は目的データである。まず、目的データを対応する縦と横の第一行の文字列「BMD」、「PR」、「Total」をマーカーとして、別々の座標値を取得する。次に、「BMD」の Y 軸座標値と「Total」の X 軸座標値に記載される文字列「1.141」を目的データとして取得した。同じ手法で、「PR」の Y 軸座標値と X 軸座標値に記載される文字列「113」を取得する。

3.3 スキャン PDF 文書の取り扱い

3.3.1 OCR による文字情報の取得

紙の文書をスキャンして生成するPDF文書は、画像の形式で表現するため、前述の方法で読み込んで、文字列の形で表現することができない。そこで、スキャン PDF 文書については、OCR ソフトウェアを利用して、スキャンした画像から OCR エンジンによって文字情報を読み取り、テキストデータに変換した。本研究では、ABBYY Fine Reader 14 という OCR ソフトウェアの ABBYY Hot Folder 機能を利用して、スキャン PDF 文書の文字情報を識別した。

3.3.2 OCR ソフトウェアの精度対策

スキャン PDF 文書については、OCR ソフトウェアの精度影響があるため、マーカーの工夫、座標値の容認範囲の設定、取得データの精度管理が必要となった。

OCR ソフトウェアが画像から文字情報を認識する際に、「r」と「i」、「r」と「a」、「f」と「i」、「:」と「;」等の認識間違いが多い。このため、マーカーとして使用する文字は、できるだけこれらの文字が含まれないように設定を行った。例えば、「Scan Type:」をマーカーとして使用したい場合、「:」の判読エラーが生じやすいため、マーカーが見つからないケースが多くなる。このため、我々はマーカーの長さを調整し、「Scan Type」をマーカーとして設定した。

次に、座標値については、OCR 精度問題のために、表の中の座標値が完全に一致しない。そこで、我々は座標値の容認範囲を設定した。表の中で取得対象となる結果値は、マーカーの座標に基づいて設定した範囲に入れば、座標値が一致とみなし、その値を取得した。

最後に、OCR ソフトウェアの認識精度の影響より、目的データが取得できない、あるいは取得したデータが正しくない状況が発生する可能性がある。目的データは、固定の形式があるので、取得したデータが標準の形式であるかどうかの検証を行った。例えば、当院では、患者 ID は八桁の数字で表現される。取得したデータが全て数字ではなく、あるいは八桁ではない場合はこのデータがエラーを返すことで精度管理を行った。ページごとに取得した目的データ列のうち、一つのデータがエラーになった場合、このページの目的データは csv ファイルに書き込まない仕様とした。

3.4 対象

本研究では、X 線骨密度測定検査報告書を対象とし、腰椎と大腿骨の骨密度値を構造化して取得することを目的とした。解析対象は、2010 年 1 月から 2017 年 7 月に大阪大学医学部附属病院で施行された X 線骨密度測定検査報告書とした。本報告書はスキャンにて電子カルテに取り込まれ、PDF ファイルで保管されている。本システムを用い、腰椎と大腿骨の骨密度値を構造化して取得し、その検出精度を評価した。

3.5 システムの設定

本研究対象の例として、腰椎骨密度測定検査報告書を図2に示す。目的データは、「レポート種類名」、「患者 ID」、「検査測定日」、「BMD 値」、「PR 値」の 5 つである。目的データのうち「レポート種類名」、「患者ID」、「検査測定日」は文字列形式からデータの取得であり、図中に黒枠で記載した。「BMD 値」と「PR 値」は表形式からデータ取得であり、図中に赤枠で記載した。

最初に、目的ページの同定は検査報告書の種類情報「f Lumbar Spine」、「f Left Hip」、「f Right Hip」と目的データがあるページの特徴文字列「Total」とその座標値を利用した。そのうち、検査報告書の種類情報「f Lumbar Spine」、「f Left Hip」、「f Right Hip」は認識間違いやすい英語アルファベットが多いため、「f Lu」、「f Le」、「f R」をマーカーとして設定した。

文字列形式の目的データである「レポート種類名」、「患者。

ID」、「検査測定日」は同じ列の記載不変の文字列「Scan Type」、「Patient ID」、「Scan Da」をマーカーとして設定した。「レポート種類名」の取得データは検査報告書で記載された英語から日本語に変更して、結果ファイルに書きこんだ。「検査測定日」には検査報告書で記載された「January 01, 2010」の形式から「20100101」のような形式に変更し、結果ファイルに書きこんだ。

表形式の目的データである「BMD 値」と「PR 値」は表の縦と横のタイトル行の記載が不変の文字列をマーカーとして取得した。「BMD 値」は縦タイトル列の「Total」と横タイトル行の「BMD」をマーカーに利用した。「PR 値」は縦タイトル列の「Total」と横タイトル行の「PR」をマーカーに利用した。

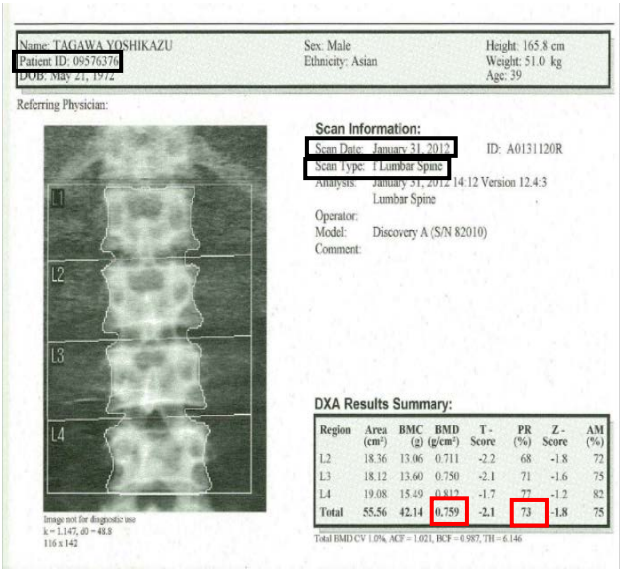

図 2 検査報告書の実例

3.6 取得データ

本研究の対象に対して、取得したデータには、「ファイル No.」、「レポート種類名」、「患者 ID」、「検査測定日」、「BMD 値」、「PR 値」の 6 つである。その内、「ファイル No.」は、学習データを検証する際に必要な情報で、PDF ファイル名から取得した。ほかの 5 つのデータは PDF コンテンツから取得した。取得した 6 つのデータは、表 2 の形式で、csvファイルで保管した。なお、患者 ID は患者情報保護の観点から本論文中ではアスタリスクで表示している。

表 2 取得した目的データの例

ファイル No.	レポート種類名	患者 ID	検査測定日	BMD 値	PR 値
0000	椎体	********	20120131	0.759	73
0001	椎体	********	20120131	0.506	50
0001	大腿骨頸部（左）	********	20120131	0.339	43
0002	椎体	********	20120131	0.590	58
0003	椎体	********	20120131	0.783	77
0003	大腿骨頸部（左）	********	20120131	0.545	69

4. システム評価

本研究は学習データとして、延べ 598 患者、2,735 ページのレポートを処理した。目的データは 1,057 ページで記載され、精度チェックでエラーとなったデータは 165 件であった。エラーなく取得できたデータ 892 件は、目視による確認の結果、すべて正しい値が取得できていた。

エラーとなった 165 ページに対して分析した結果、エラーを発生する原因は 3 つに分類された。一つ目は、ページの同定に使用したマーカーの OCR 認識間違いである。このようにページの同定に失敗した場合、目的データの取得を行うことができない。この原因でエラーとなったページは 62 ページあった。二つ目は、目的データの欠損である。例えば、6 つの目的データが全て取得できず、5 つのみ取得した場合は目的データ列の欠損があると認め、エラーを返して最後の結果ファイルで書き込まない。取得データが欠損する理由は主に設定したマーカーの認識間違いと目的データ文字列全体の認識失敗の 2 つがある。この原因でエラーとなったページは 73 ページあった。三つ目は、精度管理を行った際、取得したデータがエラーになる場合である。目的データは、数値データや文字数など、それぞれの固定形式があるので、取得したデータが OCR 認識の原因で指定した固定形式ではない場合、この取得したデータ列は間違いデータがあると判断し、エラーを返して最後の結果ファイルで書き込まない。この原因でエラーとなったページは 30 ページあった。

5. 考察

電子カルテに蓄積されたデータを臨床研究等に二次利用することを考えると、構造化されて保存されていないデータの利活用が特に後ろ向きの臨床研究で課題となる。

部門システムで作成されるレポートは PDF 化して電子カルテに保存することで、臨床目的でのレポート閲覧には問題がない。もし、これらのファイルに記載される数値データを構造化して取得することができれば、様々なレポートでの活用が期待される。

本研究ではスキャンされ電子カルテに PDF 形式で取り込まれたレポートを OCR 解析し、マーカーとなる文字列から目的の数値を取り出すことを行った。スキャン PDF 文書に対する認識には、複数の OCR ソフトウェアを試したが、認識の精度が 100％となるソフトウェアはなかった。本研究が利用した ABBYY Fine Reader 14 ソフトウェアは認識精度が一番良いと判断し、利用した。しかしながら、主に OCR の精度の問題で目的ページの同定エラー、目的データの同定エラー、取得したデータの認識エラーにより 15.6%のデータが取得できなかった。一方、本システムの精度チェックで正しいと判断されたデータは、目視によりすべて正しいデータであることが確認できた。取得データを臨床研究に使用することを考えると、間違ったデータが解析に用いられることは避ける必要があり、本システムによる精度チェックは正しく機能していた。エラーとなったデータは目視で確認しデータ登録を行うなど、対応をとることは可能となる。

本システムの精度向上に向けては、OCR の精度向上が最も効果的となる。具体的には、OCR ソフトウェアの中の予備処理オプションで、「反転色の画像を修正」にチェックを入れれば、画像の背景と字体の色が反転され、OCR の精度が「反転色の画像を修正」オプションにチェックを入れないより向上した。我々は OCR の調整を行った結果エラーページは 142 ページ（目的ページの同定エラー：50 ページ、目的データの同定エラー：62 ページ、目的データの認識エラー：30 ページ）に

減らすことに成功しており、エラーなく取得できたデータ 913 件は目視の結果、引き続きすべて正しいデータであった。OCR の認識ミスをゼロにすることは難しいと思われるが、マーカーの工夫などでエラーを減らすことができる可能性はある。一方、目的データの認識エラーは OCR ソフト自体の機能に依存しており、精度を上がる手法を別に考える必要がある。

取得したい対象レポートのフォーマットは医療機関ごとに異なることが想定されるため、本手法では医療機関ごとにマーカーを設定する必要が出てくる。しかし、一旦マーカーを設定すればデータ収集を行うことは可能であるため、多施設での臨床研究でも本プログラムを用いることは可能である。一方、目的のデータを指定したうえで機械学習を実施し、マーカーを認識し、目的データを取得することが可能となれば、データ取得をさらに一般化することが可能となる。

6. 結論

電子カルテに蓄積される PDF 文書から特定の検査値を取得することが可能であった。

参考文献

1) 大江 和彦. MID-NET:医薬品安全対策のための医療情報データベース. 生体医工学 2017; 55(4) :159-164.
2) 診療録直結型全国糖尿病データベース事業 http://jdreams.jp/
3) Matsumura Y, Hattori A, Manabe S, Tsuda T, Takeda T, Okada K, Murata T, Mihara N. A Strategy for Reusing the Data of Electronic Medical Record Systems for Clinical Research. Stud Health Technol Inform. 2016; 228: 297-301.