情 報 学 基 礎 63－7
（2001．7．24）

特許コーパスを用いた検索タスクの提案

岩山真† 藤井敦‡ 高野明彦†† 神門典子††
iwayama@pi.titech.ac.jp fujii@ulis.ac.jp takano@acm.org kando@nii.ac.jp

† 東京工業大学 / 日立製作所
‡ 図書館情報大学 / 科学技術振興事業団 CREST
†† 国立情報学研究所

NTCIR-3 において特許コーパスを用いた検索タスクを提案する。提案するタスクでは、「公開特許公報」「JAPIO 抄録」「PAJ(Patent Abstracts of Japan)」などの特許コーパスを用い、「基本検索タスク」「自由タスク」の二種類のタスクを実施する。基本検索タスクは、ある製品に関する一般的な記述からそれを支える特許を検索するタスクであり、具体的には新聞記事から関連特許を検索する。自由タスクは、特許コーパスを対象に、参加者が自由に課題を設定し評価する実験的なタスクである。

Patent Retrieval Challenge in NTCIR-3

Makoto Iwayama† Atsushi Fujii‡ Akihiko Takano†† Noriko Kando††
iwayama@pi.titech.ac.jp fujii@ulis.ac.jp takano@acm.org kando@nii.ac.jp

† Tokyo Institute of Technology / Hitachi Ltd.
‡ University of Library and Information Science / Japan Science and Technology Corporation
†† National Institute of Informatics

In NTCIR-3 workshop, we propose a new task of "Patent Retrieval Challenge" using patent corpora. The main task is "cross DB search" whose purpose is to find a set of patent documents relevant to a news paper article described on some products. We will also try an experimental task, in which participants themselves can propose, execute and evaluate their own tasks. This free-styled task intends to explore future directions of patent information processing.

1 はじめに

近年多くの国がプロパテント (特許重視) 政策を推進するようになり、「創造→権利化→権利活用 (資源回収) →新たな創造」といった知的創造サイクルの強化、加速化が叫ばれている。長年にわたって蓄積されてきた特許情報は、新技術創出、製品開発および経営戦略に欠かすことができないという認識も得ている。特に昨今は、ビジネスモデル特許、遺伝子配列特許など新しいタイプの特許も現われてきて、特許出願量に加え、内容の多様性も増してきた。このような状況で、膨大かつ多様な特許データベースを効率良く利用するための技術が強く望まれており、その実用化は重要な課題となっている。

更に産業のグローバル化に伴い、世界特許 (Universal Patent) の必要性も増してきた。日米欧の三極間では、特許情報の共有化や同時審査に向けての検討も進んでいる。ここでは言語横断検索など言語の壁を越えて特許を扱う技術が必要不可欠になるであろう。

上記の背景の下で、特許に関する様々な検索技術や言語処理技術を比較検討することを目指して、NTCIR-3において特許検索タスク (Patent Retrieval Challenge) を提案する。今回は一回目ということもあり、特許コーパスに慣れ親しむことも一つのねらいである。よって、後述する「基本検索タスク」と「自由タスク」の二つのタスクを設けるが、参加者はいずれのタスクで参加してもよい。自由タスクとは文字通り自由な課題であり、配布する特許コーパスを使えば何を行ってもよい。

以下、まず 2 節で、本タスク提案の背景について振り返る。3 節で、特許文書に関する特徴を復習したあと、4 節で、配布する予定のコーパスについて説明する。次に、5 節と 6 節で、今回提案する基本検索タスク、自由タスクを説明する。最後に、配布コーパスのデータ例を付録に示す。

なお、本論文での記述は 2001 年 6 月現在での状況をまとめたものであり、実際のタスクでは変更になる可能性がある点を了承して頂きたい。

2 背景

特許検索については、各社共に研究開発が行われており有料サービスも多い。ところが、情報検索の基礎研究において特許情報が扱われることはあまりなかった。従来の情報検索は、どちらかというとジャンルに依存しない一般的な枠組を目指していたからである。よって、広く研究者間で特許検索特有の現象について議論したり情報を交換したりする機会は少なかった。それでも最近は、知的財産権の重要性が増してきたことから、特許検索や特許分類に関する研究発表が増えてきた。米国の TREC においても、特許を扱う独立したタスクは行われていないものの、検索対象の一部として特許が含まれている。

特許検索に特化した学術会議としては、SIGIR2000 で開催された「特許検索に関するワークショップ [1]」が初めての会議である。このワークショップでは、9 件の発表とパネルディスカッションが行われ、様々な角度から特許情報処理の現状、将来について議論された。国内においては、非公開の会議などで一部の研究者や関係者が集まることはあったものの、特許検索に関する公開の会議はまだ開かれていない。

今回、特許コーパスの整備が進んだことを契機に、NTCIR において特許検索タスクを提案する。共通のコーパス上で様々な技術の比較検討をすることが主目的であるが、特許検索またはその周辺分野に関わる研究者間で情報を共有することも重要な目的である。

3 特許文書の特徴

特許制度は、発明などの知的財産を保護するために制定された。新しい発明は、特許公報という形で公開され、その代償として一定期間、独占的に利用することが許されている。発明者は、その発明を明細書という形で表現して特許庁に出願する。出願された特許は一定期間 (18ヶ月) 経つと全て公開される (公開特許公報)。その後に、必要であれば審査請求をして、審査をパスすれば特許として登録される (特許公報)。

ここで、明細書 (ほぼ公報と同じ) の特徴を挙げると、

- 構造を持った文書である。構成要素は「請求項」「発明の詳細な説明」「発明が解決しようとする課題」「発明の実施の形態」「実施例」「発明の効果」など。また、公知例(既存の特許など)への参照もある。
- 発明の範囲を定める上で重要なのが「請求項」であり、そのためか請求項は独特のスタイルで記述される。請求項は基本的には一文で書かれるため、文が長く、係り受け関係も複雑になりやすい。
- 特に「請求項」においては、発明の適用範囲を広げるために、一般的な用語を用いることが多い。
- 発明であるため新語や専門用語が多い。
- 全体の長さにばらつきがある。長いものは非常に長い。

一次情報である明細書(公報)の他にも、一次情報を加工した文書が目的により作成される。その一部については今回のタスクでも配布する予定である(4節参照)。

4 特許コーパス

特許検索タスクを実施するにあたって、2001年6月現在、以下のコーパス群を配布する予定である。全てのコーパスは、(株)パトリスからの提供により、NTCIR事務局で配布する。また、これらのコーパスはGSK(言語資源コンソーシアム)経由でも後日入手可能になる予定である。

- 公開特許公報全文データ(98,99)。以降「公開特許公報」で参照。
- JAPIO出願抄録データ(98,99)。以降「JAPIO抄録」で参照。
- 日本国英語特許出願抄録データ(98,99)。以降「PAJ」で参照。
- パトリス標準検索課題。以降「パトリス検索課題」で参照。

簡単に各コーパスを説明する。まず「公開特許公報」は、出願から18ヶ月たって公開される特許全文情報で、基本的には特許庁より公開されているものと同じである。特許査定を経て公開される「特許公報」とは異なり、出願した特許全てが含まれている。98年99年公開の約34万件をテキスト形式で配布する。ただし図表情報は含まない。

「JAPIO抄録」は、JAPIO((財)日本特許情報機構)により作成された特許抄録コーパスである。公開公報には、もともと出願人により要約が付与されているが、JAPIO抄録は、これらを専門家が(必要であれば)修正したものである。専門家は、以下の基準により修正が必要な要約を選別して書き換える。

- 適切な長さか(400字程度を目安)
- 該当特許にふさわしいものか
  - 請求項の内容を含んでいるか
  - 使用語句について発明の詳細からの引用がおこなわれているか
  - etc.

JAPIO抄録も、98年99年公開の約34万件を配布する。

「PAJ」は、JAPIO抄録を英語に翻訳したコーパスである。同じく、98年99年公開の約34万件を配布する。

「パトリス検索課題」は、JAPIOにより作成され、(株)パトリスに引き継がれたデータセットであり、検索課題とその正解集合から成る。検索課題は、3分野34課題である。正解集合の作成手順は以下の通りである。まず各課題につき、専門家が論理検索式を作る。この検索式も公開する。次に全文検索を行い正解

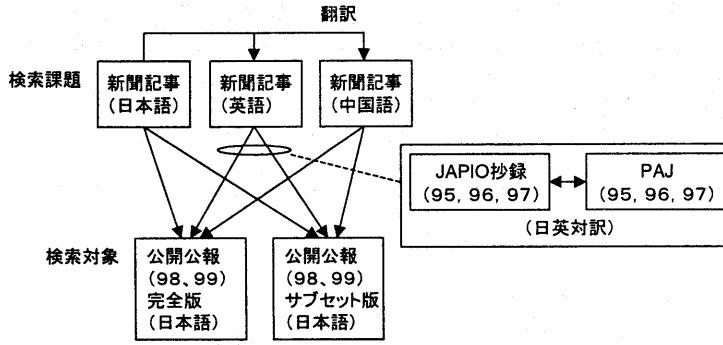

図 1: 基本検索タスク案の概要

候補集合を得る。最後に専門家が判定を行い、最終の正解集合を得る。正解集合の特許数は総計 1500 程度である。

なお、今回提案するタスクでは、上記コーパスを全て使うとは限らない。「公開特許公報」「PAJ」「JAPIO抄録」の例を付録に示す。

5 基本検索タスク

配布コーパスを使って、「基本検索タスク」と「自由タスク」の二つのタスクを提案する。

基本検索タスクでは、ある新聞記事に関連する特許を検索する。従来のように検索トピックの形で検索課題を与えるのではなく、例 (新聞記事) の形で検索課題を与える点が特徴である。これには幾つかの理由がある。

- 従来型の検索に関しては、部分的ながらテストコレクションがある (「パトリス検索課題」参照)。
- 専門家 (サーチャなど) というより一般ユーザ (企業の管理職など) を意識した検索状況である。新聞、雑誌など、ある技術の一般的な記述から、それを支えている特許を探すという検索の需要は今後増してくるであろう。
- クロスデータベース検索としての要素技術をみきわめたい。新聞コーパスと特許コーパスでは、使用する用語の傾向が異なっている。例えば特許 (特に請求項) では「ゴム」と書くべき所を一般化して「弾性体」のように書く傾向がある。新語や専門用語が使われる割合も特許のほうが高いであろう。このような問題を解決し異なるデータベースを連携して利用する技術は、新聞/特許間に限らず様々なデータベースの組み合わせにおいて有効であろう。

基本検索タスク案の概要および使用するコーパスを図 1 に示す。以下、詳細について説明する。

5.1 検索課題

既に述べたように、検索課題には新聞記事をそのまま与える。各検索課題求につき単一の記事を与える予定である。また、記事の種類としては、「製品発表記事」「技術や製品のトレンドを解説した記事」の二種類を考えている。

「製品発表記事」からの検索は、「製品」に関する具体的な特許の検索である。一方、「トレンド解説記事」からの検索は、その分野の技術情報を、特許という観点から概観することが目的である。特許マップ作成 (6.1 節参照) の前処理としても重要な検索である。

また、5.3 節で後述するように、多言語 (英語、中国語) での検索課題記事も用意する予定である。

5.2 検索対象および検索結果の提出

検索対象は、「公開特許公報」全文 (2 年分) とする。ただし、2 年分の公開特許公報は規模が大きいため、これとは別にサブセット版を検索対象として用意することも考えている。また、ベースラインとなる検索システムおよびツール群も提供する予定である。

検索結果は、通常の NTCIR にならい、各検索要求につき上位 1000 件程提出してもらう。また、公開特許公報は長いため、可能ならば各公報が検索された根拠も提出してもらう。根拠は公報の部分パッセージ (集合) とする。根拠の長さについては自由である。よって、パッセージ検索をサポートしていない参加団体については、全文を根拠として提出してもかまわない。根拠は、何らかの形で評価結果に反映させる予定である (5.4 節参照)。

5.3 言語横断検索

検索課題として日本語オリジナルの新聞記事だけでなく、対応する英語、中国語の記事も用意する予定である。対応記事は人手による翻訳で作成する。

よって、参加団体は、検索課題として日本語、英語、中国語が選べる。いずれの場合も検索対象は日本語の公開特許公報である。

言語横断検索 (特に英日検索) の場合、提供コーパスの PAJ, JAPIO 抄録を対訳訓練データとして用いてもよい。ただし、検索対象と重複する年のデータは使用しないことが望ましい。そのため、95,96,97 年の PAJ, JAPIO 抄録を訓練データとして配布することを検討している。この他にも、NTCIR-1 および 2 で提供された多言語コーパスなど、特許以外のコーパスや既存の辞書等を用いてもかまわない。

5.4 評価

各団体からの検索結果をプーリングし、各特許につき 3 段階 (A:完全に適合、B:部分的に適合、C:不適合) で適合度を付与する。システムの評価は、`trec_eval` を用いた recall/precision 評価を基本とするが、これとは別に新たな評価基準を模索することも考えている。

特に、特許は長いためユーザが適合度を判定するコストも評価要因に加えることが望ましい。ユーザに適合度を判定させるための手段として、文書の一部 (パッセージ) を提示する手段があり、その有効性は Google に代表される検索エンジンによって実証されている。そこで、人手による適合度判定でも、文書全体に加え、その根拠となったパッセージも判定要因に加えることを考えている。評価方法の詳細については、現在検討中である。

またこの場合、参加システムは検索結果一覧に加えパッセージ集合を出力することが期待されているが必ずしも必須ではない。パッセージ集合を出力しないシステムの場合、検索された特許公報全文を根拠パッセージとみなすことを考えている。

6 自由タスク

自由タスクは、参加団体が自由に定義し評価するタスクである。配布する特許コーパスを使えば、どのような問題を設定しても構わない。自由タスクは、次回以降の方向性を探る課題でもある。評価も含め詳細はまだ未定であるが、できるだけ制限を設けない形にしていきたい。以下、幾つかの例を挙げる。

6.1 特許マップの作成

特許マップとは、ある特定の技術や分野における特許を様々な視点でまとめたものである。公開された技術情報を調査することで重複した開発を防いだり、最先端の技術開発動向を把握することが主な利用目的である。特許という観点からある業界全体の権利関係を調査したり、他社注目特許の発明者が過去に出願した特許を分析して傾向を探るといった個別の用途にまで用いられている。特許庁が作成した特許マップの例も `http://www.jpo.go.jp/ryutu/tokumap.htm` で公開されている。

ところが、特許マップの作成には多くの人的資源、経済的資源を必要とする。よって、即時性が要求される状況では特許マップの半自動作成や、作成支援が望まれている。ここでは、情報の整理分類加工の技術が必要になる。近年脚光をあびているテキストマイニングなども深く関連してくるだろう。

6.2 請求項の書き換え、実施例との対応付け

請求項は、特許の適用範囲を定める重要な項目であるが、通常の文章とはかけ離れたスタイルで記述されることが多い。複数のトピックが一文に詰め込まれるため、長い文が多く係り受け関係も複雑で、内容を一読して把握することは困難である。そのため請求項を解析して読みやすく書き換える技術は、非常に有用であり、近年研究も進んでいる。

また、公報には実施例など、請求項を説明する項目があり、そこでは請求項と同じトピックが、より平易な文章でかつ具体的な用語を使って説明されている。これらの間の対応付けを用いて請求項の読解を支援するのも一つの方法である。

6.3 抄録の自動作成

配布する「JAPIO 抄録」はいわば正規化された抄録であり、その作成にかかる時間的、経済的なコストは高い。特許は出願数が多いために、JAPIO 抄録に相当する抄録を自動的に作成したり作成を支援したりするシステムが必要とされている。

また、JAPIO 抄録および公開公報は原文/抄録の対応がとれた大規模パラレルコーパスであるため、特許に限らず一般的な抄録作成における基礎的なデータとしてこれらのコーパスを用いるのも興味深い。

6.4 特許の自動分類

特許には国際特許分類 (IPC) と呼ばれる分類が付与している。たいていの検索システムは国際特許分類、もしくは FI, F タームといった更に詳細な分類を用いて検索結果にふるいをかけるため、これらの分類コードを正確かつ漏れなく付与することは重要である。近年、分類コードの付与を支援するための様々な手法が提案されている。また、この分類コード付与は基本的にはテキスト分類と同じである。従来のテキスト分類は新聞記事を対象にしたものが多いが、特許を対象にどの程度有効なのかを調べることも興味深い。

7 おわりに

NTCIR-3 において特許検索タスクを提案する。「公開特許公報」「JAPIO 抄録」「PAJ」「パトリス検索課題」の 4 つのデータセットを配布し、「基本検索タスク」と「自由タスク」の二つのタスクを実施する。参加者はいずれのタスクで参加しても良い。

「基本検索タスク」では新聞記事を検索課題とし、それに関連する公開特許公報を検索する。多言語の新聞記事を与えることで、言語横断検索も同時に実施する。一方「自由タスク」は配布コーパスを使う限り、何を行っても自由である。

参考文献

[1] ACM-SIGIR workshop on patent retrieval, 2000.

A　公開特許公報の例(一部)

実際の配布フォーマットとは異なる.

<SDO BIJ><DP M=0001><RTI ID=000001 HE=150 WI=170 LX=0200 LY=0300>(19)【発行国】日本国特許庁（ＪＰ）
(12)【公報種別】公開特許公報（Ａ）
(11)【公開番号】特開平１０－５５
(43)【公開日】平成１０年（１９９８）１月６日
(54)【発明の名称】コーヒーバック
(51)【国際特許分類第６版】
　A23F　5/36
　　5/46
　B65D 77/08
　　81/20
【ＦＩ】
　A23F　5/36
　　5/46
　B65D 77/08　　G
　　81/20　　Z
【審査請求】未請求
【請求項の数】１４
【出願形態】ＯＬ
【全頁数】８
(21)【出願番号】特願平９－４８７３０
(22)【出願日】平成９年（１９９７）３月４日
(31)【優先権主張番号】６１０７７１
(32)【優先日】１９９６年３月４日
(33)【優先権主張国】米国（ＵＳ）
(71)【出願人】
【識別番号】５９７０３０２６８
【氏名又は名称】クラフト・ジェイコブズ・サッチャード・リミテッド
【氏名又は名称原語表記】ＫＲＡＦＴ　ＪＡＣＯＢＳ　ＳＵＣＨＡＲＤ　ＬＩＭＩＴＥＤ
【住所又は居所】イギリス国グロスターシャー　ジーエル５０・３エイイー，チェルトナム，ベイズヒル・ロード（番地なし），セント・ジョージズ・ハウス
【住所又は居所原語表記】Ｓｔ．Ｇｅｏｒｇｅ'ｓ　Ｈｏｕｓｅ，Ｂａｙｓｈｉｌｌ　Ｒｏａｄ，Ｃｈｅｌｔｅｎｈａｍ，Ｇｌｏｕｃｅｓｔｅｒｓｈｉｒｅ　ＧＬ５０　３ＡＥ，Ｕｎｉｔｅｄ　Ｋｉｎｇｄｏｍ
(72)【発明者】
【氏名】ニール・サンダース
【住所又は居所】イギリス国バンベリー　オーエックス１６・９エイユー，グランジ・ロード　４４
(74)【代理人】
【弁理士】
【氏名又は名称】杜本　一夫　（外５名）
</RTI></SDO><SDO ABJ><TXF FR=0001 HE=055 WI=080 LX=0200 LY=1800>(57)【要約】
【解決課題】　改良された粉末コーヒー製品の小包を提
供することを目的とする。
【解決手段】　コーヒー成分区画室（１４）と、アロマ
成分区画室（１６）と、両区画室の間を連通可能とする
開口（２６）と、を形成するシェル（１２）を備えるコ
ーヒーバック（１０）が提供される。上記開口（２６）
は、揮発性アロマ成分のみ通過可能であり、コーヒーア
ロマ液体（２２）自身は通過不能とし、粉末コーヒー
（２０）とコーヒーアロマ液体（２２）とが少なくとも
実質的に混合しないように構成されている。
<EMI ID=000002 HE=100 WI=080 LX=1100 LY=1800></SDO><SDO CLJ><DP M=0002><TXF FR=0001 HE=250 WI=080 LX=0200 LY=0300>【特許請求の範囲】
【請求項１】　コーヒー成分区画室とアロマ成分区画室
とを形成するシェルを備えるコーヒーバックであって、
上記コーヒー成分区画室には一定量の粉末コーヒーが含
有され、上記アロマ成分区画室には一定量のコーヒーア
ロマ液体が含有されており、上記コーヒー成分区画室と上記アロマ成分区画室とは離
隔されていて、上記シェルは両区画室の間を連通させ且
つ両区画室の間に延びる開口を形成する連通部分を含
み、
上記開口を介して上記アロマ成分区画室から上記コーヒ
ー成分区画室まで、上記コーヒーアロマ液体の揮発性成
分は通過するが、上記コーヒーアロマ液体自身は通過し
ないことを特徴とするコーヒーバック。
【請求項２】　請求項１のコーヒーバックであって、さ
らに、前記アロマ成分区画室内に吸水性材料を具備し、
該吸水性材料によって前記コーヒーアロマ液体が吸着さ
れることを特徴とするコーヒーバック。
【請求項３】　請求項１のコーヒーバックであって、前
記一定量のコーヒーアロマ液体はコーヒー製品の１重量
％以下であることを特徴とするコーヒーバック。
【請求項４】　請求項３のコーヒーバックであって、前
記一定量のコーヒーアロマ液体はコーヒー製品の０．１
～１．０重量％の間にあることを特徴とするコーヒー
バック。
【請求項５】　請求項１のコーヒーバックであって、前
記コーヒー成分区画室と前記アロマ成分区画室とには大
気圧を超える圧力が負荷されていることを特徴とするコ
ーヒーバック。
【請求項６】　請求項１のコーヒーバックであって、前
記コーヒー成分区画室は、噴出部分を形成する下部を有
することを特徴とするコーヒーバック。
【請求項７】　請求項６のコーヒーバックであって、前
記シェルは、前記コーヒー成分区画室から前記コーヒー
製品を出すために、前記噴出部分を横切り該シェルを引
き裂きやすくする手段を含むことを特徴とするコーヒー
バック。

（以下略）

B　JAPIO抄録の例

実際の配布フォーマットとは異なる.

```
<PATDOC>
<B210>1997048730</B210>
<B220>19970304</B220>
<B110>1998000055</B110>
<B140>19980106</B140>
<B310>19960304US96     6107</B310>
<B711>クラフト　ジエイコブス　サチヤード　ＬＴＤ</B711>
<B721>ニール　サンダ゛ース</B721>
<B511>A23F  5/36       </B511>
<B511>A23F  5/46       </B511>
<B511>B65D 77/08      G</B511>
<B511>B65D 81/20      Z</B511>
<B542>コーヒーパック</B542>
<SDOAB LA="J">
<P>
改良された粉末コーヒー製品の小包を提供する。£チャンネル、溝、フランジ、芳香、吸水性パッド
</P>
<P>
コーヒーパック１０は、コーヒー成分区画室１４と、アロマ成分区画室１６と、両区画室の間を連通
可能とする開口２６とを形成するシェル１２を備えるよう構成されている。その場合、開口２６は、
揮発性アロマ成分のみ通過可能であり、コーヒーアロマ液体自身は通過不能とし、粉末コーヒーと
コーヒーアロマ液体とが少なくとも実質的に混合しないように構成されている。
</P>
</SDOAB>
</PATDOC>
```

C　PAJの例

実際の配布フォーマットとは異なる.

```
<PATDOC>
<JPPAT>
<SDOBI LA="E">
<B110>10000055</B110>
<B121>PATENT ABSTRACTS OF JAPAN</B121>
<B130>A</B130>
<B140>19980106</B140>
<B190>JP</B190>
<B210>09048730</B210>
<B220>19970304</B220>
<B310>96    610771</B310>
<B320>19960304</B320>
<B330>US</B330>
<B511>   A23F  5/36    </B511>
<B512>   A23F  5/46    </B512>
<B512>   B65D 77/08    </B512>
<B512>   B65D 81/20    </B512>
<B541>EN</B541>
<B542>PACKED COFFEE</B542>
<B711>KRAFT JACOBS SUCHARD LTD</B711>
<B721>SANDERS NEIL</B721>
</SDOBI>
<SDOAB LA="E">
<SEC>
<P>
PROBLEM TO BE SOLVED: To obtain a small pack of an improved powder
coffee product.
</P>
<P>
SOLUTION: This packed coffee 10 is equipped with a shell 12 forming a divided room 14 for
a coffee powder component, a divided room 16 for acoffee aroma component and an opening 26
enabling the communication between both the rooms, and in this case the opening is
constituted so as to pass only a vaporizable aroma component, reject the coffee aroma
liquid, and at least not to substantially mix the powder! coffee with the coffee aroma
liquid.
</P>
<P>
COPYRIGHT: (C)1998,JPO
</P>
</SEC>
</SDOAB>
<SDODR LA="E">
<EMI ID="00000001" HE="089" WI="066" TI="AD"IMF="TIFF"></EMI>
</SDODR>
</JPPAT>
</PATDOC>
```