OCR 認識誤りを含む書誌情報の確率的パターン解析手法

2M−1

早川 公泉　　高須 淳宏　　安達 淳
東京大学 工学系研究科　　学術情報センター 研究開発部

1 はじめに

既存の出版物を電子図書館で扱う場合、一般にスキャナを用いて紙から文書画像化されることになる。全文検索を行うために画像データからOCRを用いてテキストデータを得ることは可能であるが、認識誤りが避けられないという問題がある。そのため、あくまで画像データを基本とし、OCRを通して得られたテキストも補助的に使用する場合がある。

筆者らは文書画像間にハイパーリンクを作ることにより、利便性を向上させることを試みている。例えば、学術論文の参考文献と当該文献の間、目次と記事の間などのリンクが挙げられる。本稿では、論文から参考文献へのリンクの自動作成を行うための方法に焦点を絞る。このためには、

1. スキャナを用いての画像化
2. 画像処理等による参考文献領域の特定
3. OCRによるテキストデータの生成
4. テキストからの参考文献項目の解析
5. データベースとの照合による参考文献の特定

というプロセスを経て実現されることになる。書誌的事項を確認し、これからデータベース中のリンク先文書画像を特定し、ハイパーリンクを形成する。筆者らはすでに [1] において 4 以外の処理について検討を行い、文書画像から矩形領域を切り出してOCR処理を行うこと及び誤りを含むテキストのマッチングが可能となっている。

本稿では、OCRを通して得られたテキストデータから参考文献項目の解析を実現する手段として、確率文脈自由文法 [2] を拡張した確率的パターン解析法を提案する。

2 参考文献の書式

学会によって参考文献の書式は異なる。情報系の学会誌の参考文献の書式例は次のようになる。

- 著者名:タイトル, 書誌情報.
- 著者名,"タイトル", 書誌情報.

書誌情報はさらに雑誌名、出版社、巻号、頁などに細分される。

学会によって各要素の区切りが異なっている上、OCRによる認識誤りが生じるため、デリミタによらずに参考文献の情報を認識するのが望ましい。また、著者によって参考文献の表記は微妙に異なっており、その曖昧さを吸収しなければならない。

そこで、参考文献の項目の書式の解析を行って文法規則を抽出し、その文法を用いて参考文献の項目の各要素を特定する。各要素の特定が行えれば、OCR誤りに対応したマッチング手法を用いて既存のデータベースとの照合が可能となり、参考文献の項目に記述されている文献の特定が行える。

3 確率的パターン解析

3.1 確率文脈自由文法の適用

参考文献の項目に対して、確率文脈自由文法を適用して各要素の特定を行う。確率文脈自由文法は文脈自由文法に対して確率を付与し、複数生じる構文木の順序付けを行うものである。

確率文脈自由文法では、各生成規則 $A \rightarrow \alpha$に対して、

$$\sum_{\alpha} P(A \rightarrow \alpha) = 1 \qquad (1)$$

となる確率 $P(A \rightarrow \alpha)$ を付与する。ある特定の構文木 t を得る文 wの生成確率は

$$P(w,t) = \prod_{A \rightarrow \alpha} P(A \rightarrow \alpha) \qquad (2)$$

と、各生成規則の確率の和で表され、文 w全体の生成確率は

$$P(w) = \sum_{t} P(w,t) \qquad (3)$$

と表せる。確率文脈自由文法では、全ての文の生成確率の和に対して

$$\sum_{w} P(w) = 1 \qquad (4)$$

が成り立っている。

A Probabilistic Analysis of Bibliography including OCR Misrecognition
Kimimoto Hayakawa[1], Atsuhiro Takasu[2], Jun Adachi[2]
[1]Graduate School of Engineering, The University of Tokyo
[2]Research & Development Department, National Center for Science Information Systems

3.2 確率文脈自由文法の拡張

参考文献の項目に確率文脈自由文法を適用する際の障害は、参考文献の項目の各要素に対してあらかじめ品詞分類が不可能である点である。そこで、確率文脈自由文法を適用する範囲を前終端記号までとし、前終端記号と終端記号間にはあらたな確率を付与することにする。

文wのi番目のトークンw_iが品詞aを表している確率を$P(w_i \leftarrow a)$と表すことにする。ここでは

$$\sum_a P(w_i \leftarrow a) = 1 \qquad (5)$$

が成り立つように$P(w_i \leftarrow a)$の値を付与する。いま構文木t_jにおいて、トークンw_iの品詞をa_jとすると、文wに対する構文木t_jの生成確率は、

$$P(t_j, w_i|w) = \frac{P(w_i \leftarrow a_j)P(t_j)}{\sum_j (P(w_i \leftarrow a_j)P(t_j))} \qquad (6)$$

と表せる。従って全てのトークンの確率を考慮すると、

$$P(t_j|w) = \frac{\prod_i (P(w_i \leftarrow a_j))P(t_j)}{\sum_j (\prod_i (P(w_i \leftarrow a_j))P(t_j))} \qquad (7)$$

と表すことが出来る。

3.3 文法規則への確率の付与

確率文脈自由文法を用いる場合、あらかじめ文法を用意しておかなければならず、各生成規則に対しても確率を求めておく必要がある。表１は実際に情報処理学会論文誌 Vol36,No.1 の参考文献の項目の中から、文法が判断出来た 275 件について品詞分類を行い、頻度から確率を求めた結果である。

頻繁に出現する文法規則があるのは当然ながら、そうでない文法規則についても、無視できるほど稀ではなく規則の種類も多様である。そのため、出現頻度の低い文法規則を認識するためには、単語の表す品詞の確率と組み合わせることが重要となってくる。

3.4 品詞確率

単語が特定の品詞を表している確率については、文法規則の場合とは別の手段を用いて付与する。確率文脈自由文法の文法規則にはすでに単語間の位置や前後関係などの統計的情報が含まれているため、それぞれの語の品詞の確率は他の語とは独立に与えられるべきである。そのため、もし品詞を全く特定できない語があった場合、その語には全ての品詞の確率を一様に付与するのが妥当となる。

実際には単語に含まれている文字種やその並びによってある程度品詞を判断していくことになる。具体的には、あらかじめ各品詞毎にその品詞に含まれる語の集合を用意しておき、その集合内の語との類似度を求めて重み付けをおこなうことによって、単語の表す品詞の確率を求めることができる。

参考文献	→	著者情報:タイトル部 , 書誌情報 (発行年月).	:269/275
	→	著者情報:タイトル部 (発行年月).	:4/275
	→	書誌情報 (発行年月).	:2/275
書誌情報	→	頁, 出版元情報	:8/271
	→	出版元情報, 版	:2/271
	→	出版元情報	:53/271
	→	雑誌情報, 頁, 出版元情報	:2/271
	→	雑誌情報, 頁	:69/271
	→	雑誌情報	:1/271
	→	研究報告情報, 出版元情報, 頁	:11/271
	→	研究報告情報, 出版元情報	:12/271
	→	研究報告情報, 頁	:82/271
	→	研究報告情報	:28/271
	→	版	:1/271
出版元情報	→	出版元, 地名	:15/88
	→	出版元	:60/88
	→	地名	:13/88
雑誌情報	→	雑誌名, 巻, 号	:65/74
	→	雑誌名, 巻号	:3/74
	→	雑誌名	:4/74
研究報告情報	→	研究報告, 出版元情報	:23/133
	→	研究報告, 巻号	:32/133
	→	研究報告	:78/133
タイトル部	→	タイトル, サブタイトル	:6/271
	→	タイトル	:265/271

表 1:参考文献の項目の文法例

4 今後の展開

本稿では、参考文献項目の情報の認識を行う手段として、確率文脈自由文法を発展させて参考文献に適用できるように拡張した、確率的パターン解析手法を提案した。

今後の課題として、文法規則の確率を得るためのサンプルを増やして出現頻度の低い文法規則の確率の精度を向上させると共に、適切な品詞確率を求めるためのアルゴリズムの検討を進めていくことが挙げられる。

さらに、本手法を用いた参考文献項目の認識システムの実装を進めていき、文書画像間のリンクを張ることを試み、有効性の検証を行っていく予定である。

参考文献

[1] 高須 淳宏, 文書画像データからの書誌情報の抽出とマッチング, 情報学基礎, 45-6, 1997

[2] E.Charniak, Statical Language Learning, The MIT Press, 1993