Recognition of Table of Contents for Electronic Library Consulting

A. Belaïd

LORIA-CNRS Campus Scientifique B.P. 239
54506 Vandoeuvre-les-Nancy Cedex France
Phone: (33) 3 83 59 20 82, Fax: (33) 3 83 41 30 79, e-mail: Abdell.Belaïd@loria.fr

Received April 5, 2000 / Revised February 19, 2001

Abstract. A labeling approach for automatic recognition of Tables of Contents (ToC) is described in this paper. A prototype is used for electronic consulting of scientific papers in a digital library system named Caliope. This method operates on a roughly structured ASCII file, produced by OCR. The recognition approach operates by text labeling without using any a priori model. Labeling is based on a Part of Speech Tagging (PoS) which is initiated by a primary labeling of text component using some specific dictionaries. Significant tags are first grouped in homogeneous classes according to their grammar categories and then reduced in canonical forms corresponding to article fields: "title" and "authors". Non labeled tokens are integrated in one or another field by either applying PoS correction rules or using a structure model generated from well detected articles. The designed prototype operates with a great satisfaction on different ToC layouts and character recognition qualities. Without manual intervention, 96.3% rate of correct segmentation was obtained on 38 journals including 2020 articles and 93.0% rate of correct field extraction.

Key words: Caliope project – Digital Library – Table of Contents Recognition – Part-of-Speech Tagging – OCR Combination

1 Introduction

This work was elaborated in the framework of the Caliope project in collaboration between the Xerox Research Centre in Grenoble (XRCE) and the French Academic Research Institute INRIA. Caliope is a digital library system that allows simultaneous consultation of scientific periodicals. Caliope is based on the concept of "scanning on the request" of scientific papers and their remote print. Paper selection is possible according to Tables of contents (ToCs) given in an XML format on a central server.

Even though nowadays ToCs database is sizeable, many periodicals must be integrated into Caliope re- quire an intensive manual keyboarding. The solution retained was to dispose of tools adapted to the automatic recognition of those ToCs that allow an automatic feeding of the server.

ToC recognition leads to identify their articles. Articles have a relatively simple structure which is composed of three essential fields: "title", "authors" and "page number". In spite of the simplicity of this structure, many factors may complicate their straightforward extraction: order of changing of the fields, bad separation of the two textual fields ("title" and "authors"), etc. These two factors can be accentuated by OCR errors, deleting thus the separation between the two fields and introducing errors in the word field which may complicate their identification.

Few ToC recognizers have been proposed in the literature. On one hand, Takasu et al. [1,2] proposed a system named CyberMagazine, based on image segmentation into blocks and syntactic analysis of their contents. The article recognition combines the use of decision tree classification and a syntactic analysis using a matrix grammar. On the other hand, Story and O’Gorman [3,4] proposed a method that combines OCR techniques and image processing. Blocks are first located by the image processing “docstrum”. Then, the ToC layout and relationships between the different article references are found according to an a priori model given manually for each kind of journal. These relationships are used, for example, to determine automatically the page number when the user clicks on one article title, or to give specific information concerning one article.

In this work, we propose a method based on text coding which in turn is based on Part of Speech tagging (PoS) [5,6]. The idea of this method, employed in language processing and text indexing, is to reassemble nouns in nominal syntaxes representing the same information. The nouns are given by a specific morphological tagging. This method can be applied in ToC recognition for article field identification by reassembling in the same syntagm "title" or "authors", words having similar tags. The process of tagging consists of three stages: tokenization, morphological analysis, and syn-
tactical grouping and disambiguation. The tokenizer isolates each textual term and separates numerical chains from alphabetic terms. The morphological analyzer contains a transducer lexicon. It produces all the legitimate tags for words that appear in the lexicon. If a word is not in the lexicon, a guesser is consulted. The guesser employs another finite-state transducer which examines the context and decides to assign the token to “title” or to “authors” depending on prefixes, inflectional information and productive endings that it finds.

The paper is organized as follows. Section 2 briefly reviews the Calliope architecture. Section 3 describes the ToC classes, their structure and OCR drawbacks that affect their recognition performance. Section 4 outlines the proposed scheme for ToC recognition by detailing the operation of labeling, article segmentation and field extraction. Based on OCR recognition, we demonstrate how labeling method and contextual extension can be used to localize articles in a noisy text. Finally, Section 5 contains a summary and conclusions of the study.

2 System architecture

2.1 Calliope architecture

As shown in Figure 1, Calliope has a distributed architecture. It contains a ToC central server, two storing document servers (it can be more) and several scanning sites, all connected to the Web. The ToC server is the main entry point for interacting with the libraries. Article images already scanned are on the storing servers. These images have been created on request by scanning sites, located in documentation centres associated to the project. Images are converted from TIFF (standard for the scanning format) to GIF (in a degraded version) allowing users to display them and to print them at will. When a document is selected, a clearer version of the document (Postscript format) is then transmitted for printing.

2.2 XmlOCR based ToC server

The ToC recognition is based on the use of OCR and on XML for the format exchange with Calliope. First, the scanned image of the ToC is recognized by combining OCRs; the result is formatted by using XML (XmlNodeLayout). Second, this file is analyzed and article references are extracted and organized in another XmlNodeLayout format. XmToC. The result is verified manually by the operator and the final output is sent to Calliope server in XmlCal, a format compatible with Calliope format.

3 ToC structure

We have limited our application to textual ToCs as they appear in books, periodicals and conference proceedings (see Fig. 2). Although some of them are already produced today in an electronic format, a fair proportion of ToCs still remain on paper medium. It is then required to convert and structure these ToCs in an electronic form.

A textual ToC supported by our application is a document composed of at least one page, one header, a list of sections, article references and footnotes. An article reference is made of an article title, a list of authors and a page number, written in one, two or three separated columns. Article references may have more or less complicated structure but must be separated from each other. Sections corresponding to session names are given to group papers. These section titles will not be taken into account during the recognition phase because they are not considered as index for the ToC server.

4 ToC Analysis system

The ToC analysis system is decomposed into two main steps:

- The first step consists on text labeling in order to highlight the text components representing articles embedded in the text. The proposed method analyzes each line and marks each component dealing with an article reference, such as numerical strings (for page numbers), initials, proper nouns and connecters (for authors), common words (for titles), long spaces (for tabulations), etc.
- The second step is devoted to article location and article reference identification. The proposed method (1) selects the more regular indices reflecting the article reference structure, (2) tries to enlarge this labeling to the article references, and (3) generates a structure model of reliable references for correcting unstructured ones. Once the article reference extraction is completed, an XML file is generated containing all these references which are linked to the corresponding articles.

4.1 Document capture

ToCs are digitized with 300 dpi scanning resolution using an HP ScanJet 6100C scanner. Three commercial OCRs are applied and their result are combined in order to overcome the individual drawbacks. In spite of this combination, several residual defects remain in the final text : (1) All characters are not properly recognized due to the poor quality of the original document. Therefore, this results in a bad recognition rate for punctuation and non-abbreviations which are important for the recognition. (2) The font styles are not homogeneously restored, leading to a lack of information for the reference separation. (3) The space lines which are function of the proportional fonts are poorly interpreted and, as a result, may disappear implying a non clear separation between article references.

Because all of the OCR drawbacks, our recognition method operates in a bottom-up manner, labeling progressively the safe text components and tries to enlarge this labeling to the corrupted components by using the context.
4.2 ToC analysis

ToC analysis consists of article extraction and field (title and authors) delineation within each article. Intuitively we have thought of a top-down approach by proposing an a priori model for article structure, but at our great surprise, we have noticed that almost each journal has its specific structure. As shown in figure 3, 28 different models have been registered for the 38 ToC classes studied. These models differ by:

- the relative position of their main components: Title (Tit) and its continuation (STit), Authors (Auts) and its continuation (SAuts), and page number (Num).
- Every field can appear on the left, on the right, or in the middle;
- the existence or not of a tabulation (text indentation);
- the existence or not of a dotted line, either between the title and the authors, or between the authors and the page number;
- the existence or not of an item mark in the beginning indicating the article starting;
- etc.

Faced to the structure changing of ToC articles avoiding the use of a generic model of each periodical, we proposed a bottom up approach by labeling the article terms and trying to climb up some term groupings related to the article fields. We will first introduce the PoS tagging approach and then show its use in ToC recognition.

4.2.1 PoS tagging Knowing the complexity of the text structure representation and the information extraction from this structure, several text analysis methods have been proposed based on information tagging of keyword extraction [5-17]. A linguistic model is then built from
these keywords in order to highlight the important content of textual fields and to search for linguistic components where the reference to the reality is stable. For example, the first hypothesis of the model SYDO [7] is that the parts of speech constructed around the noun (or nominal syntagm) are those containing references to the speech field objects and are those we have to identify. The proposed linguistic model reflects the mechanism allowing the moving from word predicate to nominal syntagm.

There are two classes of automatic tagging methods in PoS: rule based methods [6,16] and stochastic methods [9,13,14], functioning in supervised and non supervised mode.

The former typically uses a contextual information to assign tags to unknown or ambiguous words. These rules are often known under the name: contextual frame rules. In addition to this contextual information, several taggers use the morphological information to resolve the ambiguity caused by unknown words. Some systems go beyond the contextual and morphological information by including rules considering some indices such as the punctuation or the capitals.

The latter incorporate the frequency or the probability in the validation process. The easiest use of the probability that a word occurs with a particular tag for its identification, but these approaches have an uncertain behavior knowing their local vision which can leads to inadmissible sequences of tags. An alternative to these approaches is to perform the probability of the given sequence of frequent tags. This approach is based on the n-gram method considering that the best tag for a given word is determined by the probability that it occurs with the n previous tags. The method generally retained in a stochastic tagger combines the two previous approaches, using the tag sequence probabilities and the word measurement frequency. This is known under the name of HMM [Hidden Markov Model][10,14].

4.2.2 PoS application on ToC analysis. The application technique of PoS tagging on ToC recognition consists in morphological labeling of the different ToC terms and grouping similar consecutive labels in order to climb up the interesting fields. These groupings are called : parts of speech. Figure 4 shows such a grouping of an article where the labels used are common nouns, proper nouns and numerals. Common nouns allow to climb up the title, while proper nouns climb up the authors and the numerals climbs up the page number.

By refining this morphological labeling, we can observe some situations in better way as illustrated in figure 5. Knowing that a title cannot finish by a preposition, we have succeed in moving back the separation frontier between the title and the author and to climb up the prepositional separator.

5 ToC Structure analysis

The structure analysis follows three major steps. In first step, a primary labeling of lines and text components is performed. In second step, based on these labels more syntactic forms are constructed to represent each article reference, and in third step, the final structure of ToCs is built.

5.1 Primary labeling

The text file is examined line by line and each line or each component (string or space) in a line is labeled according to its content as shown in table 1. By the end of this
phase, the text file is modeled using a line grammar. For this basic grammar, a token is used as a primary label.

Figure 6.b shows the primary labeling of the ToC given in figure 6.a.

<table>
<thead>
<tr>
<th>Table 1. Primary labels</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical Labeling</strong></td>
</tr>
<tr>
<td>SL</td>
</tr>
<tr>
<td>LS</td>
</tr>
<tr>
<td>CL</td>
</tr>
<tr>
<td>NS</td>
</tr>
<tr>
<td>DL</td>
</tr>
<tr>
<td>TB</td>
</tr>
<tr>
<td><strong>Morphological Labeling</strong></td>
</tr>
<tr>
<td>CN</td>
</tr>
<tr>
<td>PN</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>PR</td>
</tr>
</tbody>
</table>

5.1.1 Article location. A ToC is rarely written in a distinctive manner in the page. Textual zones accompanies the ToC as headers on the top, footnotes at the bottom, or sometimes as editorial zones at different places. So, in order to make the method location independent, we have made the search for the ToC location in the detection of page numbers.

In fact, in our method, only references accompanied by their page numbers are considered. So, as the numerical strings are easy to extract, the first step for the reference location is to find NS location. Knowing that these NSs are regular, they are first extracted and then only those presenting some location regularities are considered. The regularities correspond to 1) starting and end position within the references, 2) vertical alignment, 3) position after a specific punctuation like dotted line, 4) position after or before a blank line. A weight between 6 and 100 is assigned to each kind of regularity. Then, for each NS, a total weight is performed by adding regularity weights. The NSs presenting an amount greater than a given threshold are kept for the following steps.

5.1.2 Column extraction. Tabulations are very important to identify columns for better separation between reference components. As for the page numbers, columns are detected by using the tabulation regularities based on the same position in the lines.

5.1.3 Article delimitation. Some logical rules are used to mark up the obvious starting or ending lines of ToC article references. Rules used for article reference location are given in table 2.

Then, knowing the starting or the ending of an article, the system tries to delimit the missing limit by deducing it from the limits of the other articles. This is done either at this level if the context is rich enough or later by using more information from labeling result and article model.

Figure 6.c shows the delimitation result of the articles of the ToC given in figure 6.a. The sign "=" is used for lines outside the articles and "=" for lines inside the articles.

5.2 Authors and title extraction

In this step, we start the real morpho-syntactical analysis of the text. It consists first in identifying the more representative tokens of the two fields “title” and “authors”: common words for “title” (Tit) and proper nouns plus initials for “authors” (Avt), and then in their progressive and syntactical extention until the entire delimitation of fields. This process is realized separately for each article, i.e. for each list of lines surrounded by AS and AE.

The extraction procedure is applied recursively on labeling rules until any labeling changing is possible. In each extraction step, the priority is given for “authors” extraction rules because of the low representativity of proper nouns and essentially initials in the “authors” field forming. The main extraction phases are: initial field forming from basic tokens by combination of linguistic aspects, and field correction by syntax re-establishment for the authors and syntagn re-establishment for the title. The latter is based on contextual grouping and association of unknown tokens (badly recognized by OCR).

5.2.1 Linguistic aspects. Linguistic aspects are related to the morphological tagging of words and to the grouping of some words having similar sense, commonly called parts of speech. This can be done progressively in successive different steps: forming, grouping, context enlarging, etc. These steps will be detailed in the following.

5.2.1.1 Forming rules. The first step uses the “morphology proximity” for forming the "authors" and "title" field, based on the gathering of some obvious morphologically close consecutive labels. Table 3 outlines these gathering rules. The sign “*” means “followed by”.

5.2.1.2 Continuity rules. Since the fields can be written in many consecutive lines and words shared between lines, it is needed to reassemble them by noticing the field continuity beyond the end of the current lines. This is possible when the current line terminates by a non-terminal word (such as a preposition, an article or a connector for a title, or by an initial or a first name

<table>
<thead>
<tr>
<th>Table 2. Article delimitation rules</th>
<th>where AS represents the article starting and AE the article ending</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL = AS if</td>
<td>CL = AE if</td>
</tr>
<tr>
<td>CL begins by PU or NS</td>
<td>CL ⊃ NS and NL = SL</td>
</tr>
<tr>
<td>CL ⊃ NS and PL = SL</td>
<td>CL ⊃ NS and NL = BL</td>
</tr>
<tr>
<td>CL ⊃ NS and PL = EL</td>
<td></td>
</tr>
</tbody>
</table>
for an author]. Table 4 gives some continuity rules and shows an example for author and title written in adjacent columns. For the title, the continuity is maintained by the presence of the connector “and”, while for the author, the continuity is maintained thanks to the last name completing the beginning of the author on the first line.

Table 4. Continuity rules

<table>
<thead>
<tr>
<th>Continuity Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>TIT</td>
</tr>
<tr>
<td>Tit + CC</td>
</tr>
</tbody>
</table>

5.2.1.3 Reduction rules

At the end of this first labeling step, reduction and filtering rules are applied initially to put together, a list of identical labels, and then to “assimilate” some intermediate tokens representing punctuation, spaces or connectors. Table 5 exhibits some of these rules.

Figure 7b shows the effects of the continuity and reduction rules.

5.3 Syntax correction

Due to the OCR errors and to the presence of unknown proper nouns in the “authors” fields, a great number of tokens remain non labeled (labeled by UN). The use of PoS can help to find contextual information to correct these labels. This is made by the re-establishment of the syntax for the author fields, and of the syntagm for the title fields.

5.3.1 Syntax re-establishment. The author syntax correction leads to its syntax re-establishment by correcting the misslabeled components representing either the first or the last name. List of the rules corresponding to these situations is shown in Table 6. In the first example, “and”, image of “and” is corrected thanks to the presence of authors surrounding it. In the second example, the correct second line is “B. Zavidovique”. This is re-established thanks to the presence of a tabulation on the left, indicating the field changing, and the existence of a connector followed by an author on the right.

5.3.2 Syntagm re-establishment. The title syntagm correction leads to obtain homogenous term groupings. The correction methodology tends to associate unknown words to the surrounding groupings using some continuation words such as connectors, prepositions or articles.
Table 6. Syntax re-establishment where PN+ means a list of proper nouns, and Aut+ means the beginning of an author field. Some rules include the ending line label (EL) in order to precise the favourable context for the correction.

<table>
<thead>
<tr>
<th>Agglutination Rules</th>
<th>Extension Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aut + UN + Aut ⇒ Aut</td>
<td></td>
</tr>
<tr>
<td>Tit + EL + TB + UN + ⇒ Tit + EL + TB + Aut+</td>
<td></td>
</tr>
<tr>
<td>Aut+ + CC + FN</td>
<td>IT + UN ⇒ Aut</td>
</tr>
</tbody>
</table>

Examples

J. D. Bailey, J. Hall, 75
P. A. S. Reed and C. J. Colbourn

Generic Functions for On-Clip Vision.
BY. Zavidavique and T.M. Bernard

These connectors cannot end the title and should connect the parts before and after them. This is possible by the use of some agglutination and extension rules, such as mentioned in the table 7 (here, Tit⇒ means extended title on the right, Tit⇐ means extended title on the left, Tit+ simply represents a grouping set of common nouns. The first example shows the agglutination of a bad recognized word “evolution” due to the existence of surrounding prepositions and valid common nouns). The second example is also corrected because the error is located on the term group "K artificial neural networks", surrounded by several valid words. Furthermore, the left sentence cannot end by the word “feedforward” (the direct object is missing), and the right sentence begins by a preposition “for” indicating a connection with the left part. The third example is difficult to correct because errors are dominant. This can be resolved in a post-processing step of word segmentation and lexicon verification methods.

Table 7. Syntagm re-establishment

<table>
<thead>
<tr>
<th>Agglutination Rules</th>
<th>Extension &amp; Anticipation Rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tit+ + UN + Tit+ ⇒ Tit+</td>
<td>Tit + PR</td>
</tr>
<tr>
<td>Tit+</td>
<td>Tit+ + Prep + Aut + Tit+ ⇒ Tit+</td>
</tr>
<tr>
<td>Tit+ + Prep + Tit+ ⇒ Tit+</td>
<td></td>
</tr>
</tbody>
</table>

Examples

bernoulli error measure approach to train feedforward K artificial neural networks for classification problems
aworkshopapproach using spreadsheets for the teaching of statistics K and probability

Figure 7.c illustrates the syntax and the syntagm re-establishments rules.

5.4 Article model generation

The objective of this step is to restore the incompletely structured article references in order to guaranty the extraction of all the references in a ToC. We generate a structure model based on the recognized reference structure in order to complete the labeling of the incomplete article references.

A generic structure model describes the content of each column by the tags {Aut, Tit, Pag, TB} associated with generic constructors such as “+” for repetitive (at least one), “-” for repetitive (can be zero), and “?” for optional.

A possible model can be:

\[
\begin{align*}
\text{Aut+} \\
\text{TB ? Tit*} \\
\text{TB ? Stit+ Pag}
\end{align*}
\]

This model means that the article references are structured in two columns. The right column contains the page number and the left column contains a stack of lines composed of several author lines, followed by at least one title line. The page number is in the last line.

Thus, if in that article reference, some components remain labeled with UN, then these components will be labeled as Tit.

Figure 8 shows the model correction of a ToC. The model allows the separation of two successive articles and the correction of the page number of one of them. This model is obtained by the synthesis of all the specific structures of each valid article reference. The repetition of similar components at the same position for this component in the final model.

Generally, the model depends on the number of columns determined. We will give some correction rules based on this number in the following. Formally we consider that each article reference is made of a stack of lines SoLi containing either Tit, Aut, Pag or UN otherwise.

5.4.1 Case of three columns In this case, SoLi is composed of three columns C1, C2 and C3. We assume that each column contains only one label Tit, Aut, Pag or UN. Knowing the page number column, the identification may concern the other two columns.

Let \( C_1 = \{x_i\}_{i=1,n} \) be the column of titles, i.e. \( x_i \in \{\text{UN}, \text{Tit}\} \). \( C_2 = \{y_j\}_{j=1,m} \) is the column of authors, i.e. \( y_j \in \{\text{UN}, \text{Aut}\} \). Many configurations are possible as shown by each line in the left part of Equation (1).

\[
\begin{align*}
x \mid y & \mid TB y \mid x \mid TB Pag \\
\text{Pag} & \text{TB} x \mid y & \mid TB y \mid x \\
x \mid y & \mid TB Pag \text{ TB} y \mid x
\end{align*}
\]

(1)
5.4.2 Case of two columns. There are only two possible configurations allowing the identification of reference components. The first one corresponds to equation (2).

Let \( z = \{z_i\}_{i=1, n} \) be the column containing authors and titles.

\[
\text{Pag TB } z \Rightarrow \{ \begin{array}{c}
\text{if } z_1 = \text{Aut} \Rightarrow z_n = \text{Tit}
\end{array} \}
\]

(2)

The second case corresponds to (3) where authors and titles are separated by either DL or PN.

\[
\text{Pag TB } x^1 | y^1 \Rightarrow \{ \begin{array}{c}
\text{if } \exists k / x_k = \text{Tit} \Rightarrow x_i = \text{Tit} \forall i, y_j^2 = \text{Aut} \forall j
\end{array} \}
\]

\[
\text{Pag TB } x^1 | y^1 \Rightarrow \{ \begin{array}{c}
\text{if } \exists l / y_l = \text{Aut} \Rightarrow y_j = \text{Aut} \forall j, x_i^2 = \text{Tit} \forall i
\end{array} \}
\]

(3)

5.4.3 Case of one column. This case is somewhat similar to the previous case and two reduction possibilities are considered:

\[
\text{Pag } z \Rightarrow \{ \begin{array}{c}
\text{if } z_1 = \text{Aut} \Rightarrow z_n = \text{Tit}
\end{array} \}
\]

\[
\text{Pag } x^1 | y^1 \Rightarrow \{ \begin{array}{c}
\text{if } \exists k / x_k = \text{Tit} \Rightarrow x_i = \text{Tit} \forall i, y_j^2 = \text{Aut} \forall j
\end{array} \}
\]

\[
\text{Pag } x^1 | y^1 \Rightarrow \{ \begin{array}{c}
\text{if } \exists l / y_l = \text{Aut} \Rightarrow y_j = \text{Aut} \forall j, x_i^2 = \text{Tit} \forall i
\end{array} \}
\]

(4)

6 Experiments and discussion

Experiments were performed on 32 journals and 9 proceedings, containing 2703 articles and 2020 author fields. The most frequent ToC format is the one column format. The article location rate is 96.5 %, Articles were totally located in 81 % of ToCs. Only one ToC has less than the half of the articles localised; this ToC contains some structure errors. The recognition rate for fields is 96.5 % for page numbers and 93.0 % for the separation between titles and authors. Authors have been identified completely in 40% of the cases, and 4.5% have been identified in less than 50%. Figure 9 summarises the results given for the three ToCs formats, while figure 10 shows the results for the different journals.

For article delimitation, problems arise in two cases:

- the suppression by the OCR of empty lines due to the presence in the neighboring of specific font styles,
- the proximity between rubric titles and articles. This leads to some ambiguities on the article starting essentially in the case where it doesn’t exist enough completely separated articles allowing the construction of an article model.

For field delimitation, the problem is essentially due to the separator between the title and the author, which is either absent, erroneous (because of OCR), or confused with a punctuation sign. The field identification runs well in the case where the separation is clear between authors and article titles; fields in separated columns, authors on a separated line. Figure 9 gives the global results obtained for each document class, showing the author extraction rates and the article location rates. It is easier to observe that the segmentation rates are high for three column ToCs and low for one column ToCs.

![Table](image)

**Fig. 9. Global results**
Figure 8. Model correction of ToCs. From right to left: generic model for all the ToC, specific models generated for each article reference, original ToC text.

Figure 10 gives the detailed results obtained for each kind of journal and proceeding. The different rates obtained cannot allow us to separate the journals in terms of their difficulty not only by the number of columns, but also by the complexity of their structure.

6.1 Error analysis

The main errors encountered can be:

- the bad restitution by OCR of tabulations or space lines between two successive references, because a space line or a tabulation is either too narrow or filled by noise;
- the heterogeneity of the reference format;
- the ambiguity between the section title and the first line title of a reference. This ambiguity is caused by a narrow space between the two lines and the absence of indication on the reference starting.

Figure 11 shows the various occurrences corresponding to the several errors incurred by the system.

6.2 OCR improvement

In order to improve the recognition rates, we have attempted to improve the input data quality by combining different OCRs. Four commercial OCRs have been used in this sense. Figure 12 shows the different results obtained from the four OCRs, in an increasing order, and for the manual correction of the data. It is observed that the quality is higher in journals than proceedings, and magazines. Furthermore, the recognition rate is maximum on a corrected data which is a good test for the approach performance measurement of the technique.

We observed that the results obtained by OCR combination are better than those obtained from individual
OCRs as seen in figure 13. For the combined approach, we have used the GNU diff program based on the Miller and Myers algorithm [18], for the file synchronisation, and a weighted majority vote for the error correction.

6.3 User interface

A user interface was designed to help user to perform specific job. It shows "stacks" containing images of journal front page that one can view on library shelves (see Figure 14.a). To see the contents, one selects a journal and the system opens a set of ToC images for all the journal issues (see Figure 14.b). The user can select one ToC. The system displays the corresponding ToC image. A menu of functions is available for image processing and recognition on the selected ToC. Once the ToC references are identified, these references are displayed in HTML (see Figure 14.c) allowing a link of each reference to its corresponding article in GIF format (see Figure 14.d).

7 Conclusion and perspectives

In this paper, we have presented an efficient labelling method for ToC analysis. ToCs are scanned and converted to ASCII format by OCR. This ASCII file is roughly structured and contains many corrupted characters. The segmentation method uses a syntactical reduction approach. It is based on a linguistic tagging of the text words and on a canonical reduction of the tags. The method operates field by field and line by line, by separating the articles and then by separating fields within the articles. The system doesn’t use any a priori model. It adapts the extraction process on each new ToC by only taking into account a general logical knowledge on the article structure. In this prototype, only one OCR has been used without any specific parameterisation concerning the document quality. In the future, we plan to extend this prototype to 1) combine many OCRs in order to improve the data quality, 2) reinforce the tagging rules of PoS in order to refine the use of linguistic rules, and 3) extend the application to more complicated ToCs like those of magazines.

Acknowledgements. We are very grateful to the Xerox Research Centre in Grenoble for authorizing the publication of this paper.

References

Fig. 14. a-d: User interface


