# Enhancing Table of Contents Extraction by System Aggregation

Thi-Tuyet-Hai Nguyen, Antoine Doucet, Mickaël Coustaty

# Enhancing Table of Contents Extraction by System Aggregation

Thi Tuyet Hai NGUYEN, Antoine DOUCET, Mickael COUSTATY
L3i Laboratory, University of La Rochelle, France
Email: {hai.nguyen, antoine.doucet, mickael.coustaty}@univ-lr.fr

***Abstract*—The OCR-ed books usually lack logical structure information, such as chapters, sections. To enrich the navigation experience of users, several approaches have been proposed to extract table of contents (ToC) from digitised books. In this paper, we introduce an aggregation-based method to enhance ToC extraction using system submissions from the ICDAR Book structure extraction competitions (2009, 2011, and 2013). Our experimental results show that the union of two best approaches outperforms the existing approaches using both the title-based and link-based evaluation measures on a dataset of more than 2000 books. By efficiently combining the results of existing systems in an unsupervised way, we consistently beat the state-of-the-art in book structure extraction, with performance improvements that are statistically significant.**

## I. Introduction

Nowadays, many libraries make an effort to digitise the printed books, especially historical ones. The conversion of each scanned book into searchable text is implemented by optical character recognition (OCR). However, the existing OCR technologies only provide the full text of books with some structural information such as paragraphs and pages. More complex structures, including chapters, sections, etc., are not systematically detected. The extraction of ToC is suggested as the supplement of current OCR technologies to give more structural information. Such technique provides a convenient way to browse inside books and increases the access and usability of digitised books.

The task of ToC extraction confronts several challenges due to the limitations of OCR technologies, as well as the diversity of ToCs' layouts. In fact, most approaches rely on keywords such as "chapter", "section" to identify ToC entries. Therefore, their performances will fall down if the OCR process makes mistakes when recognising such keywords. In addition, each book can have different layouts: some books have "flat" ToC pages, some have "ordered" or "divided" ToC page(s) [1], and others simply do not have a single ToC page.

There are several approaches trying to reconstruct the book's ToC. They can be divided into three main types. Some approaches, including the state-of-the-art approach, rely on the detection of ToC pages, then analyse them for ToC entries. Instead of depending on ToC pages, others take the whole document into consideration and utilise some strong indicators to extract ToC entries (larger font, vertical spaces, etc.). The hybrid approaches consider whether the input book has ToC pages or not, then apply the suitable method either find and analyse or build the book's ToC.

In this paper, we present an approach based on the aggregation of the existing approaches. We utilise the combination of two set operators (the union and the intersection) and two properties (title and page number) to aggregate submissions of the ICDAR book structure extraction competitions in 2009 [2], 2011 [3], and 2013 [4]. Our method is evaluated by the title-based and link-based measures over three book structure extraction competitions' datasets.

The paper is organised as follows. In Section II, we describe previous work related to ToC extraction. Section III is the detailed description of our approach. Our experimental results are displayed in Section IV. Finally, we present our conclusions in Section V.

## II. Related Work

Several approaches are meant to address the extraction of books' ToCs. They can be classified into 3 types, including approaches based on the detection of ToC pages, on the whole book content, and hybrid ones [4].

First, there are approaches relying on the detection of ToC pages. Typically, such techniques include three main steps. The first one concentrates on detecting the book's ToC pages, before extracting ToC entries from such pages. Finally, the remaining book's content is processed for identifying links between titles and pages.

The state-of-the-art approach belongs to this type and is developed by Dresevic et al. [5](MDCS). It also recognises TOC pages and assign each physical page with a logical page number. After that, each ToC page is analysed for ToC sections whose important parts will be processed to detect titles and corresponding page numbers. In the next step, a fuzzy search method is applied to identify links between titles and page numbers. All parts of this ToC extraction engine are based on pattern occurrences obtained from their training dataset.

Besides the typical steps of the ToC-recognition-based approaches, Wu et al. [1] take the diversity of ToC's layout into consideration. They introduce three basic layout styles of book's ToCs, namely "flat", "ordered", and "divided". They further design three corresponding rule-based techniques for processing each of these styles. The ToC layout style classification is used as a complement step before the extraction of ToC entries.

The main disadvantage of this type of approaches is that it mainly relies on ToC pages to extract ToC entries, therefore its performance can be significantly decreased in

case of books without ToC pages, or whenever the physical or digitised version of the ToC page is damaged or altered.

To overcome this, the second type of approaches focuses on the analysis of the entire book content instead of focusing on its ToC pages. The representative approach of this type is presented by Giguet and Lucas [6] (GREYC). They use a four-page window to find the large whitespace which is considered as a strong indicator of the ending of a chapter and the beginning of a new one. That approach concentrates on the entry title which is extracted from the third page of the sliding window.

In our point of view, the main benefit of such an approach is that it is totally unsupervised and language-independent. However, it will require large memory for processing the whole document even in the case of a book with clear and exhaustive ToC pages.

In out-of-copyriqht books, it has been observed that as many as 20% books do not contain a ToC [7]. It thus seems necessary to use hybrid approaches, so as to be able to deal with books with and without a ToC.

Liu et al. [8] (NANKAI) proposed such a hybrid method. They consider whether a book has ToC pages or not, then apply the appropriate method. A rule-based method is designed for books with ToC pages while machine learning is used to deal with books without ToC pages.

Instead of using traditional rule-based method with classical boolean logic, the approach of Gander et al. [9] (INNSBRUCK) utilises the power of rule-based technique with the flexibility of the fuzzy logic, with the aim to better handle several OCR flaws as well as variations in the books' styles. Additionally, results are carefully refined by a grammar-based method in the final step.

Another hybrid approach which combines a rule-based technique, a supervised method and similar strong indicators of Giguet and Lucas [6] to extract the ToC entries is the approach of Djean and Meunier [10], [11] (XRCE). There are four methods in their suggestion. The first and second methods use a rule-based approach to parse ToC pages and index pages. The supervised method relying on five generic properties (contiguity, textual similarity, ordering, optional elements, no self-reference) and on some document layout specificities is the core of the third method. The last one relies on trailing page whitespace.

Hybrid approaches are promising in that they shall properly handle all books, with or without ToC pages. However, hybrid approaches still underperform the top of the first type of approaches (MDCS) in all three ICDAR book structure competitions.

In general, no approach has fully combined the ToC pages' features and the book content. According to our analysis of the submissions to the 3 rounds of the ICDAR book structure extraction competition, MDCS always obtains the best performance on 1653 books with ToC pages. The hybrid approach XRCE achieves the highest performance on the 187 books without ToC pages of the competitions' datasets in ICDAR 2009 and ICDAR 2011. The GREYC approach is the best on the 167 books without ToC pages of ICDAR 2013 competition dataset. The following tables (Table I, II and III) illustrate the detailed performance scores observed over the three ICDAR competitions' datasets.

TABLE I
DETAIL PERFORMANCE SCORES OVER THE ICDAR 2009 COMPETITION DATASET

| Method | F-measure | | | |
|---|---|---|---|---|
| | Books with ToC | | Books without ToC | |
| | Title-based measure | Link-based measure | Title-based measure | Link-based measure |
| GREYC | 0.07 | 1.5 | 0.13 | 0.5 |
| NOOPSIS | 10 | 47.7 | 0.87 | 2.8 |
| **XRCE** | 33.17 | 72.4 | **7.81** | **22.6** |
| **MDCS** | **50.84** | **78.8** | 0.13 | 7.4 |

TABLE II
DETAIL PERFORMANCE SCORES OVER THE ICDAR 2011 COMPETITION DATASET

| Method | F-measure | | | |
|---|---|---|---|---|
| | Books with ToC | | Books without ToC | |
| | Title-based measure | Link-based measure | Title-based measure | Link-based measure |
| GREYC | 9.47 | 52.5 | 6.9 | 42.5 |
| **XRCE** | 19.02 | 58.4 | **26.32** | **53.8** |
| NANKAI | 38.85 | 71.6 | 7.93 | 26.9 |
| **MDCS** | **48.96** | **78.2** | 5.12 | 8.6 |

TABLE III
DETAIL PERFORMANCE SCORES OVER THE ICDAR 2013 COMPETITION DATASET

| Method | F-measure | | | |
|---|---|---|---|---|
| | Books with ToC | | Books without ToC | |
| | Title-based measure | Link-based measure | Title-based measure | Link-based measure |
| **GREYC** | 8.74 | 47 | **9.18** | **35.4** |
| EPITA | 18.06 | 41.9 | 0.07 | 1.8 |
| WURZBURG | 22.13 | 48.6 | 7.53 | 26.5 |
| INNSBRUCK | 36.17 | 74.2 | 8.2 | 33.6 |
| NANKAI | 42.65 | 73.8 | 0.7 | 7.5 |
| **MDCS** | **52.67** | **80.1** | 0.2 | 1.9 |

As a consequence, we propose in this paper a new method which relies on the combination of existing methods. This aggregation step is based on two set operators (the union and the intersection) applied on two of the properties of ToC entries (title and page number). Full details are given in the following section.

## III. OUR APPROACH

We aggregate pairs of submissions based on two set operators and the properties of their ToC entries.

Our purpose is to evaluate the performance of an aggregation submission which only contains the common entities of two submissions or all the entities extracted by these submissions. We apply two fundamental set operators to combine two submissions, based on the intersection and the union operators.

It is simple to apply set operators on primitive sets such as integers, floats or strings. However, a ToC submission is a derived set which consists of books' ToCs. A book's ToC has a unique book id and a list of ToC entries, that each include three properties: title, page number and depth level [2]. We suppose a two-step combination. Firstly, we combine two submissions based on the book id (bookid). Secondly, the ToC entries will be aggregated by applying each set operator to each property of a ToC entry (which belongs to a primitive set).

The difficulty in choosing appropriate properties is supported by users' reading-behaviours. When reading a book, most users first pay attention on the ToC to find contents and identify the corresponding pages. Hence, we consider two properties of ToC entries', title and page, for combining submissions.

In the two following sub-sections, the set operators as well as the ToC entries' properties are introduced.

### A. The properties

It is simple to identify whether two page numbers match or not. Then to deal with minor title variations, we apply a measure used in three book structure extraction competitions [3], which considers that two strings A and B are similar if the following distance between string A and B is lower than 20% and the distance between their first and last five characters (or less if the string is shorter) is lower than 60%.

$$D = \frac{LevenshteinDist * 10}{Min(length(A), length(B))} \tag{1}$$

where LevenshteinDist is a measure based on a modified version of the Levenshtein algorithm. In the modified Levenshtein algorithm, if the character is alphanumeric then the cost of edit operations (substitution, deletion and insertion) is 10 , otherwise their cost remains 1.

### B. The operators

In set theory, the intersection (AND) of set A and set B is the set which contains all elements of A that also belong to B.

$$A \cap B = \{x : x \in A \wedge x \in B\} \tag{2}$$

As to the union (OR), the union of set A and set B is the set of all elements in two sets, which are in A, in B, or in both A and B. In other words, this set contains all elements of set A and set B's elements which are different from set A.

$$A \cup B = \{x : x \in A \vee x \in B\} \tag{3}$$

We propose 8 combinations, namely *AND pages, OR pages, AND titles, OR titles, AND pages AND titles, OR pages OR titles, OR pages AND titles, OR titles AND pages*. Each of them will be carefully discussed in the following two subsections, organising the combinations in two types, based on whether they use a single or a double operator.

### C. Single operator

In this type of combination, we only use one set operator with one property. With the page property, we use the intersection (*AND pages*) and the union of two submissions (*OR pages*).

The *AND pages* set only contains ToC entries which contain the same book pages of two submissions, while the *OR pages* set contains the union of all ToC1s and ToC2s pages.

For example, let us assume we have two submissions: submission 1 of participant ID1 and submission 2 of participant ID2, as illustrated in Fig. 1 and Fig. 2. For book id 1, let us name the ToC entry of submission 1 as ToC1 and the ToC entry of submission 2 as ToC2.

```
<?xml version="1.0" encoding="UTF-8"?>
<bs-submission participant-id="ID1">
    <book>
        <bookid>1</bookid>
        <toc-entry title="CLARISINE THE COUNTESS" page="32"/>
        <toc-entry title="THE BALLAD OF BLOODY ROCK" page="39"/>
        <toc-entry title="WHEN FIRST LOVE COMES"  page="46"/>
        <toc-entry title="THE SWALLOW" page="50"/>
    </book>
</bs-submission>
```

Fig. 1. Example submission of participant ID1

```
<?xml version="1.0" encoding="UTF-8"?>
<bs-submission participant-id="ID2">
    <book>
        <bookid>1</bookid>
        <toc-entry title="CLARISINE COUNTESS;" page="32"/>
        <toc-entry title="1.THE BALLAD BLOODY ROC" page="40"/>
        <toc-entry title="THE POPLARS" page="46"/>
        <toc-entry title="THE FINISHED TASK" page="53"/>
    </book>
</bs-submission>
```

Fig. 2. Example submission of participant ID2

Before describing each case, it is worth clarifying that according to the distance measure mentioned in Eq. 1, the first title of ToC1 "CLARISINE THE COUNTESS" is similar to that of ToC2 "CLARISINE COUNTESS;" and the second title of ToC 1 "THE BALLAD OF BLOODY ROCK" is similar to that of ToC 2 "1.THE BALLAD BLOODY ROC". We consider four cases which can happen between book titles and book pages in our examples, including "similar title and same page" (the first ToC entries), "similar title and different page" (the second ToC entries), "different title and same page" (the third ToC entries), and "different title and different page" (the fourth ToC entries).

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ANDpages">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="46" title="WHEN FIRST LOVE COMES" />
    </book>
</bs-submission>
```

Fig. 3. AND pages set

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ORpages">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="39" title="THE BALLAD OF BLOODY ROCK" />
        <toc-entry page="46" title="WHEN FIRST LOVE COMES" />
        <toc-entry page="50" title="THE SWALLOW" />
        <toc-entry page="40" title="1.THE BALLAD BLOODY ROC" />
        <toc-entry page="53" title="THE FINISHED TASK" />
    </book>
</bs-submission>
```

Fig. 4. OR pages set

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ANDtitles">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="39" title="THE BALLAD OF BLOODY ROCK" />
    </book>
</bs-submission>
```

Fig. 5. AND titles set

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ORtitles">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="39" title="THE BALLAD OF BLOODY ROCK" />
        <toc-entry page="46" title="WHEN FIRST LOVE COMES" />
        <toc-entry page="50" title="THE SWALLOW" />
        <toc-entry page="46" title="THE POPLARS" />
        <toc-entry page="53" title="THE FINISHED TASK" />
    </book>
</bs-submission>
```

Fig. 6. OR titles set

Given our examples, the *AND pages* set contains the first and the third ToC entries which have the same page numbers (32, 46). Similarly, the *OR pages* set consists of six ToC entries, including four from ToC1 and two from ToC2 which have different page numbers (40, 53) than in ToC1. These combination sets of our examples are illustrated in Fig. 3 and Fig. 4.

As regards the title property, the intersection set of two submissions (*AND titles*) contains ToC entries which have the similar book's titles of two submissions; and the union of two submissions (*OR titles*) consists of all ToC1s and ToC2s whose titles are different from ToC1s. We have *AND titles* set and *OR titles* set of our examples in Fig. 5 and Fig. 6.

### D. Double operators

This sub-section describes approaches where we rely on two operators. Firstly, there are two combinations utilising the same set operator on book properties separately: *AND pages AND titles, OR pages OR titles*. The result of the combination (*AND pages AND titles*) is the set that only contains ToC entries which have the same book page number and similar book titles. The output set of the combination (*OR pages OR titles*) includes all entries of ToC1 and ToC2.

With our examples, we the *AND pages AND titles* set only contains the first entry. The *OR pages OR titles* set includes all ToC1 entries whose page numbers are 32, 39, 46 or 50, as well as the ToC2 entries whose page numbers and titles are different from those in ToC1 (only the entry of page 53 is added for that reason, while the entry of page 40 will be ignored because its title is similar to that of the entry of page 39). Fig. 7 and Fig. 8 demonstrate these combinations.

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ANDpagesANDtitles">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
    </book>
</bs-submission>
```

Fig. 7. AND pages AND titles set

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ORpagesORtitles">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="39" title="THE BALLAD OF BLOODY ROCK" />
        <toc-entry page="46" title="WHEN FIRST LOVE COMES" />
        <toc-entry page="50" title="THE SWALLOW" />
        <toc-entry page="53" title="THE FINISHED TASK" />
    </book>
</bs-submission>
```

Fig. 8. OR pages OR titles set

Secondly, two combinations are utilising different set operators on two properties: *OR pages AND titles*, and *OR titles AND pages*. The *OR pages AND titles* set is a subset of the *OR pages* set, since we can obtain it by removing the ToC entries which have different titles from the ones in the *OR pages* set. Similarly, the *OR titles AND pages* set is a subset of the *OR titles* set, obtained by deleting the ToC entries with page numbers different of those in the *OR titles* set. These combinations are illustrated in Fig. 9 and Fig. 10.

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ANDtitlesORpages">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="39" title="THE BALLAD OF BLOODY ROCK" />
        <toc-entry page="40" title="1.THE BALLAD BLOODY ROC" />
    </book>
</bs-submission>
```

Fig. 9. OR pages AND titles set

```xml
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<bs-submission participant-id="ORtitlesANDpages">
    <book>
        <bookid>1</bookid>
        <toc-entry page="32" title="CLARISINE THE COUNTESS" />
        <toc-entry page="46" title="WHEN FIRST LOVE COMES" />
        <toc-entry page="46" title="THE POPLARS" />
    </book>
</bs-submission>
```

Fig. 10. OR titles AND pages set

## IV. Experimental Results

### A. Datasets

We used the Book Structure Extraction competitions' datasets and their participants' submissions to experiment our fusion algorithm. A different dataset was used for each of the 3 rounds of the competition, each of them being composed of books selected from the INEX book search track, containing 50,239 digitized books [12]. In each of those subsets, the specificities and the diversity of this large book collection were preserved, both in terms of book genre, and in terms of the observed ratio of books with and without a physical table of content pages (80:20). Details of the three competitions' datasets and participants are given below:

- ICDAR 2009 dataset - 527 books (436 with ToC, 91 without ToC): MDCS, XRCE, NOOPSIS, GREYC
- ICDAR 2011 dataset - 513 books (417 with ToC, 96 without ToC): MDCS, NANKAI, XRCE, GREYC
- ICDAR 2013 dataset - 967 books (800 with ToC, 167 without ToC): MDCS, NANKAI, INNSBRUCK, WURZBURG, EPITA, GREYC

### B. Evaluation

As stated before, in order to assess the quality of the results and to be able to compare our results to the methods proposed in those competitions, two main metrics have been used: a title-based measure [7] and a link-based measure [13]. In the title-based measure, ToC entries are firstly assessed on whether their title is similar to the ground truth according to a distance measure mentioned in Eq. 1, then the links and the depth levels are considered.

Concerning the link-based measure, first of all, ToC entries are first assessed based on whether they "link" to a page number that truly matches an existing ToC entry. After that the similarity of the titles is computed using the following INEX weighted Levenshtein distance, and the depth levels are tested:

$$simil(s1, s2) = 1 - \frac{weightedLevenshtein(s1, s2)}{max(weight(s1), weight(s2))} \quad (4)$$

where weightedLevenshtein is similar to LevenshteinDist mentioned in the Eq. 1, and weight(s) is the sum of each character's weight in the string s (if a character is a letter or a number then its weight is 10, otherwise, its weight is 1).

The global performances of our systems, computed on these 3 datasets, are presented in tables IV, V, and VI. Each table is horizontally split into 3 blocks of information. The first block evaluates to two best approaches (best appr.) from the competition. Two next blocks correspond to the results obtained with the single and the double operators presented in III-C and III-D. In addition, as we got the same results on both the *AND bookid* set and the *OR bookid* set with the AND OPERATOR, only one row is presented.

Our results show that the union operator applied to one property outperforms the sole state-of-the-art approach (MDCS) on both title-based and link-based evaluation measure. In terms of the link-based measure, the aggregation of two best competition approaches using *OR pages* always gets higher performance than the MDCS approach, with 4.0%, 9.9% and 6.6% improvements respectively over the ICDAR competitions' datasets from 2009, 2011, and 2013, respectively.

As to the title-based measure, the *OR pages* aggregation is 3.75% higher than MDCS for the 2011 competition. With the 2009 and 2013 datasets, the *OR pages OR titles* aggregation achieves better results than MDCS, by 0.6% and 0.92% respectively.

The union operator outperforms other set operators because it combines the best of two worlds, by integrating results from a) methods that are good at extracting ToC entries from books with ToC pages, and b) methods that are good over books without ToC pages. This confirms our initial hypothesis that both types of approaches are complementary. Indeed, the main F-measure improvement is due to strong recall improvement while precision remains stable.

**Significance of our results.** To determine whether our results are statistically conclusive, we computed the student's t-test to compare the distributions of our best combinations to the best-performing methods over each of the three competition datasets.

TABLE IV
PERFORMANCE SCORES OVER THE ICDAR 2009 COMPETITION DATASET

| | Method | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Title-based | Link-based | Title-based | Link-based | Title-based | Link-based |
| **Best appr.** | Books with ToC pages (MDCS) | 41.33 | 65.90 | 42.83 | 70.30 | 41.51 | 66.40 |
| | Books without ToC pages (XRCE) | 30.28 | 69.20 | 28.36 | 64.80 | 28.47 | 63.80 |
| **Single Operators** | AND pages (MDCS-XRCE) | 42.94 | 66.80 | 34.68 | 52.60 | 36.90 | 56.20 |
| | AND titles (MDCS-XRCE) | 38.51 | 54.20 | 24.38 | 33.70 | 27.40 | 38.00 |
| | AND bookid OR pages (MDCS-XRCE) | 41.01 | 70.10 | 44.63 | 77.30 | **41.70** | 70.30 |
| | AND bookid OR titles (MDCS-XRCE) | 36.12 | 59.60 | 46.08 | 76.50 | 39.05 | 63.20 |
| | OR bookid OR pages (MDCS-XRCE) | 41.01 | 70.20 | 44.63 | 77.50 | **41.70** | **70.40** |
| | OR bookid OR titles (MDCS-XRCE) | 36.12 | 59.70 | 46.08 | 76.70 | 39.05 | 63.40 |
| **Double Operators** | OR pages AND titles (MDCS-XRCE) | 37.27 | 55.60 | 24.45 | 34.80 | 27.11 | 38.80 |
| | OR titles AND pages (MDCS-XRCE) | 35.53 | 56.10 | 36.17 | 54.30 | 34.11 | 51.40 |
| | AND pages AND titles (MDCS-XRCE) | 38.44 | 54.30 | 23.24 | 31.70 | 26.54 | 36.50 |
| | AND bookid OR pages OR titles (MDCS-XRCE) | 41.75 | 70.60 | 44.59 | 76.00 | **42.11** | 69.90 |
| | OR bookid OR pages OR titles (MDCS-XRCE) | 41.75 | 70.70 | 44.59 | 76.20 | **42.11** | **70.10** |

TABLE V
PERFORMANCE SCORES OVER THE ICDAR 2011 COMPETITION DATASET

| | Method | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Title-based | Link-based | Title-based | Link-based | Title-based | Link-based |
| Best appr. | Books with ToC pages (MDCS) | 40.40 | 64.50 | 43.17 | 70.20 | 40.75 | 65.10 |
| | Books without ToC pages (XRCE) | 27.39 | 79.30 | 18.69 | 52.50 | 20.38 | 57.60 |
| Single Operator | AND pages (MDCS-NANKAI) | 39.72 | 64.10 | 34.14 | 54.40 | 34.96 | 55.60 |
| | AND titles (MDCS-NANKAI) | 38.48 | 58.90 | 27.60 | 39.80 | 30.00 | 43.80 |
| | AND bookid OR pages (MDCS-XRCE) | 43.52 | 75.00 | 48.82 | 83.20 | **44.50** | **75.00** |
| | AND bookid OR titles (MDCS-XRCE) | 39.55 | 63.50 | 51.86 | 79.60 | 42.25 | 65.00 |
| | OR bookid OR pages (MDCS-XRCE) | 43.52 | 75.00 | 48.82 | 83.20 | **44.50** | **75.00** |
| | OR bookid OR titles (MDCS-XRCE) | 39.55 | 63.50 | 51.86 | 79.60 | 42.25 | 65.00 |
| Double Operators | OR pages AND titles (MDCS-NANKAI) | 36.14 | 56.60 | 27.88 | 41.30 | 29.44 | 44.00 |
| | OR titles AND pages (MDCS-NANKAI) | 35.59 | 56.30 | 35.37 | 55.60 | 33.79 | 52.30 |
| | AND pages AND titles (MDCS-NANKAI) | 37.86 | 58.10 | 25.76 | 36.80 | 28.53 | 41.40 |
| | AND bookid OR pages OR titles (MDCS-XRCE) | 43.96 | 74.90 | 47.37 | 79.60 | **43.64** | **72.50** |
| | OR bookid OR pages OR titles (MDCS-XRCE) | 43.96 | 74.90 | 47.37 | 79.60 | **43.64** | **72.50** |

TABLE VI
PERFORMANCE SCORES OVER THE ICDAR 2013 COMPETITION DATASET

| | Method | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Title-based | Link-based | Title-based | Link-based | Title-based | Link-based |
| Best appr. | Book with ToC pages (MDCS) | 42.77 | 64.90 | 45.92 | 71.50 | 43.61 | 66.60 |
| | Book without ToC pages (INNSBRUCK) | 33.63 | 75.70 | 32.14 | 68.90 | 31.34 | 67.20 |
| Single Operator | AND pages (MDCS-NANKAI) | 43.87 | 65.50 | 37.49 | 54.80 | 38.85 | 56.50 |
| | AND titles (MDCS-NANKAI) | 42.12 | 60.50 | 30.12 | 40.30 | 32.94 | 44.50 |
| | AND bookid OR pages (MDCS-INNSBRUCK) | 41.28 | 68.20 | 48.97 | 84.10 | 43.41 | 72.00 |
| | AND bookid OR titles (MDCS-INNSBRUCK) | 37.66 | 60.20 | 49.72 | 82.00 | 41.38 | 66.20 |
| | OR bookid OR pages (MDCS-INNSBRUCK) | 41.97 | 69.50 | 49.69 | 85.40 | **44.07** | **73.20** |
| | OR bookid OR titles (MDCS-INNSBRUCK) | 38.36 | 61.50 | 50.45 | 83.40 | 42.04 | 67.40 |
| Double Operators | OR pages AND titles (MDCS-NANKAI) | 39.67 | 59.60 | 30.21 | 41.80 | 32.07 | 44.70 |
| | OR titles AND pages (MDCS-NANKAI) | 38.80 | 58.40 | 38.95 | 56.70 | 37.29 | 53.80 |
| | AND pages AND titles (MDCS-NANKAI) | 42.49 | 60.00 | 28.27 | 37.00 | 31.69 | 41.80 |
| | AND bookid OR pages OR titles (MDCS-INNSBRUCK) | 42.16 | 68.30 | 48.70 | 81.10 | 43.87 | 71.00 |
| | OR bookid OR pages OR titles (MDCS-INNSBRUCK) | 42.86 | 69.70 | 49.42 | 82.40 | **44.53** | **72.20** |

In terms of linked-based measure, this showed clear significance over all competition subsets ($p<0.001$), demonstrating the added-value of our approach over the state-of-the-art. In terms of the title-based measure, statistical significance was obtained for the 2011 and 2013 datasets ($p<0.001$), but not for the 2009 dataset.

## V. CONCLUSION

This paper presents an aggregation approach using two set operators on two properties of ToC entries in order to combine the output of top-performing methods in book structure extraction. Our experimental results demonstrate that the union operator applied on ToC entries' properties performs better than the top-performing methods of the state-of-the-art for both title-based and link-based evaluation.

## REFERENCES


[1] Z. Wu, P. Mitra, and C. L. Giles, "Table of contents recognition and extraction for heterogeneous book documents," in *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013.

[2] A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "ICDAR 2009 Book Structure Extraction Competition," in *Proceedings of the Tenth International Conference on Document Analysis and Recognition*, Barcelona, Spain, july 2009.

[3] A. Doucet, G. Kazai, and J.-L. Meunier, "ICDAR 2011 book structure extraction competition," in *Proceedings of the Eleventh International Conference on Document Analysis and Recognition*. IEEE, 2011.

[4] A. Doucet, G. Kazai, S. Colutto, and G. Mühlberger, "ICDAR 2013 Competition on Book Structure Extraction," in *Proceedings of the Twelfth International Conference on Document Analysis and Recognition*. IEEE, 2013.

[5] B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "Book layout analysis: Toc structure extraction engine," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2008.

[6] E. Giguet and N. Lucas, "The book structure extraction competition with the resurgence software at caen university," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2009.

[7] A. Doucet, G. Kazai, B. Dresevic, A. Uzelac, B. Radakovic, and N. Todic, "Setting up a competition framework for the evaluation of structure extraction from ocr-ed books," *International Journal on Document Analysis and Recognition*, vol. 14, no. 1, 2011.

[8] C. Liu, J. Chen, X. Zhang, J. Liu, and Y. Huang, "Toc structure extraction from ocr-ed books," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2011.

[9] L. Gander, C. Lezuo, and R. Unterweger, "Rule based document understanding of historical books using a hybrid fuzzy classification system," in *Proceedings of the 2011 Workshop on Historical Document Imaging and Processing*. ACM, 2011.

[10] H. Déjean and J.-L. Meunier, "Xrce participation to the 2009 book structure task," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, 2009.

[11] H. Déjean and J.-L. Meunier, "On tables of contents and how to recognize them," *International Journal of Document Analysis and Recognition (IJDAR)*, vol. 12, no. 1, 2009.

[12] G. Kazai, A. Doucet, M. Koolen, and M. Landoni, "Overview of the inex 2009 book track," in *International Workshop of the Initiative for the Evaluation of XML Retrieval*. Springer, pp. 145–159.

[13] H. Déjean and J.-L. Meunier, "Reflections on the inex structure extraction competition," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*. ACM, 2010.