Type of proposal: Long paper

Mode of presentation: Oral

Title: Linking Article Parts for the Creation of a Newspaper Digital Library

Topic: Document Processing- Digital Libraries, Information Retrieval - Document Classification

Keywords: document layout analysis, document understanding, newspaper article tracking, text matching

Authors:

S. L. Mantzaris, Lambrakis Press S.A,

Address: 8 Heyden Str, 104 34 Athens, Greece

e-mail :slm@dolnet.gr, FAX: +30-1-8250040, phone: +30-1-8252680,

B. Gatos, Lambrakis Press S.A

N. Gouraros, Lambrakis Press S.A

S. J. Perantonis, Institute of Informatics and Telecommunications, National Research Center “Demokritos

Required facilities: data videoprojector

Linking Article Parts for the Creation of a Newspaper Digital Library

S. L. Mantzaris [1], B. Gatos [1], N. Gouraros[1], S. J. Perantonis [2]

[1] Lambrakis Press S.A., 8 Heyden Str,

104 34 Athens, Greece

slm@dolnet.gr

[2] Institute of Informatics and Telecommunications,

National Research Center “Demokritos”,

153 10 Athens, Greece

sper@iit.demokritos.gr

Abstract

An important issue pertaining to the retro-conversion of newspapers, i.e. the conversion of newspaper issues into digital resources, is the identification and appropriate digital representation of an article. To complete this task, a number of steps have to be followed, from segmentation of the newspaper image to optical character recognition and linking of different items belonging to the same article. In this paper, an evaluation of different information retrieval techniques is presented that aim at linking textual parts of an article that can be found on different pages of a newspaper issue. Three document matching techniques are evaluated, namely title-to-title, title-to-text and text-to-text matching. In addition, the effect on the matching accuracy of using a stemmer and of employing appropriate conflict resolution techniques is studied for each of the above approaches. Experimental results involving a number of issues of a Greek newspaper show that the best technique, namely text-to-text matching augmented with a stemmer and conflict resolution, can reach a high linking accuracy rate of 96%.

1. Introduction

The problem of document image understanding has attracted many researchers. The mainstreams of the techniques that are used in order to face this problem are based on the geometric layout of the page. Here we are focusing our attention to newspaper pages. Newspapers have a very complex and diverse layout. Combining geometric layout algorithms with a rule or knowledge based approach, as in [1,2], it is possible to obtain an adequately efficient method for understanding a newspaper page. However, in this case it is very common that a single page contains many article units, some of

which are continued on another page. The identification of all article units in different pages that comprise an article of a newspaper issue would be very helpful for the creation of a Digital Library based on newspaper material. On the other hand the manual linking of article parts is a time consuming and expensive procedure. Clearly the layout information is not useful in automatically linking the parts of an article that spans two or more pages. For this reason, three different approaches based on text matching, stemming and full text retrieval techniques are presented and compared.

In order to have the necessary textual data, as well as an indication that an article is continued on another page, a number of steps should be followed:

a) segmentation of image pages into various items (titles, text, images etc.),
b) identification of special symbols or patterns indicating the article's continuation to another page,
c) article unit identification and reconstruction,
d) recognition of the textual components.

In order to avoid the exaggerated cost of a manual implementation of the above four steps, we have proceeded to an automatic scheme that minimizes human intervention. Below we briefly describe our automatic approach to carry out all these steps.

Algorithms for page segmentation referred to in the literature are mainly based on the smearing and labeling of regions [3], on the image profiling in various directions [4] and on texture information [5]. The fully successful application of the above-mentioned algorithms is prevented from the haphazard layout of newspaper articles and the close contact of different segments. We use a new technique for newspaper page segmentation based on smearing and labeling of regions and on gradual extraction of

image components in the following order: Lines, images and drawings, background lines, special symbols, text and title blocks ([1], [2]). Special symbols correspond to specific regions on the newspaper page such as references, which in our test case were text regions lying on the right side of arrows indicating that an article continues on another page (Fig. 1). Connected Component Analysis helps us to locate special symbols according to their geometric features (such as the ratio of height to width). An extra verification for the existence of special symbols is carried out using a pattern matching technique [6].

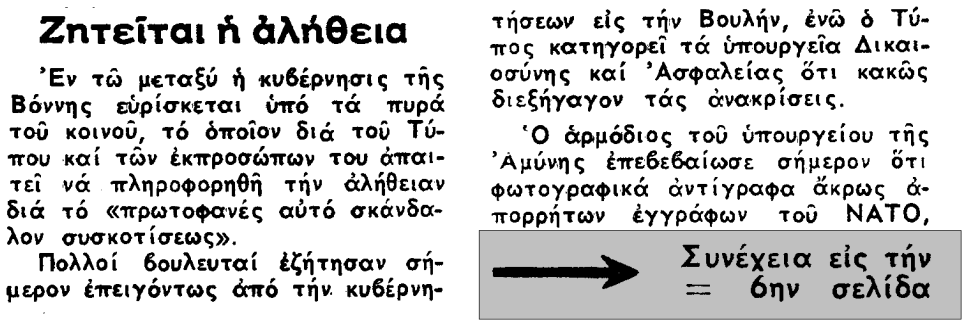

Figure 1. An example of a special symbol that is a reference indicator

After the segmentation phase, the article units of a newspaper page are identified. An article unit is considered to consist of various elements such as headline, title, text, picture, caption and article continuation indicators (references). Text areas are labeled as headlines or sub-headlines after sufficient evidence for the starting point of an article unit is found. Our approach for article tracking exploits the segment relationships

existing in the page layout of a newspaper. These relations are formulated as a set of rules [2]. An example of newspaper article identification is shown at Fig. 2. Before proceeding to the last step, the results produced so far were checked and corrected as necessary.

At the last step, we automatically recognize all textual components. We integrated an OCR module developed in earlier work ([7]) which involves the stage of character segmentation, the stage of extracting reliable features for every character and finally, the stage of recognition using a fast and effective classifier.

(a) (b)

Figure 2.(a) original image, (b) segmented image in article units

We selected as character features appropriately weighted averages originating from the application of overlapping Gaussian masks to the character body. Using those features as an input to a k-Nearest-neighbor classifier we achieve high recognition rates. At the worst case of deformed and noisy newspaper segments, the recognition rate is approximately 95%. For a normal page, the recognition rates are in the range of 99.7%.

All the aforementioned steps are fundamental for the digital retro-conversion of the printed material. Thus, we consider that no additional effort is mandatory in order to have the reference and textual data, which are necessary for the automated linking of article parts.

Here, for simplicity, will assume that an article has at most two parts, which is the most common case in practice. On the other hand, if an article is divided in more than two parts, we can apply our methods to link the first and the second parts to form a new first part. Then, we can link the new first part to the third part in order to form yet another first part and so on. Another assumption is that in the first part of an article there is a reference indication of the page that holds the continuation of this article.

2. Linking article parts

Three different approaches were applied in order to find the continuation of an article part. The first was to match the title of the first part of an article to the titles of the possible continuations. The second was to use the full texts of the candidate continuations in addition to their titles. In the third approach the full text of both the first part of an article and the candidate continuations were employed. Apart from the amount of text used, the aforementioned approaches have an additional parameter that

differentiates them. This is the correctness of the output from the OCR preprocess step. Only in the first approach the output of the OCR is manually corrected and so is considered to be error free.

To achieve the best possible results, instead of using the words in the form they appeared in the newspaper page, a stemmer is applied to all textual data. In this way the stems of the words are used during the matching phase. Since our texts are in Greek, a Greek stemmer that handles the various forms of a noun, an adjective and a verb was used. The stemmer was developed during the creation of an electronic edition of the literature magazine EPOCHES (published by Lambrakis Press SA from 1963 to 1967) and was used in our experiments without any modifications. Even though lately there is a skepticism for the value of stemming in information retrieval, the Greek morphology is very rich (a single noun has at least four forms and a single verb has at least twenty forms) making the use of a stemmer imperative. Our experiments showed that stemming offered improvements up to 35% in effectiveness.

In all cases a score value is computed in order to rank the results of the matching between a first part of an article and a set of candidate second parts. A set of candidate second parts consists of all article units that lay on the reference page, which is mentioned in the first part of an article. The top ranked candidate article unit in the results is considered to be the expected continuation. Since it is possible for two or more different first parts to have sets of candidate continuations containing common elements, possible conflicts may arise. In order to resolve such conflicts, a technique that increases the effectiveness of the matching phase is proposed. Consider two articles, i.e. A_1 and A_2, which have their second parts on the same page. Furthermore,

assume that among their possible continuations the top ranked matching second part for both of them is the same, i.e. P, with scores S_1 and S_2 respectively. Without loss of generality, if S_1 is greater than S_2 then P is the continuation of A_1 and P is eliminated from the set of possible continuations for A_2. Thus, A_2 has now a new top ranked matching second part. In case new conflicts appear, the same rule is applied again. This technique can be easily extended to cover more than two conflicting first parts.

a) Title to title matching

At a first glance it is expected that the two parts of an article have the same title. However, this is often not the case and for that reason this approach gave the poorest results. An additional drawback is that in many articles the continuation does not have a title. Since a title contains few words, OCR errors can create serious problems to the matching procedure. For this reason, the OCR output was manually corrected. In this approach the scoring function used was the number of common stems between the titles (stopwords excluded).

b) Title to full text matching

In this approach each candidate set of continuations is considered as a document collection and the title is considered as a query to that collection. Except for a list of stopwords all the stems contained in the document collection are considered as index terms. An implementation of the extended boolean model [8] is used to estimate the similarity between the query and each document in the collection. In this case the similarity value is used as the score value. The stems contained in the title of a first part

are joined using a *p-norm* conjunction operator. The value of *p* is dependant on the number of stems (stopwords excluded) that a title has. If a title has more than 5 stems then *p* was set to 2, otherwise *p* was set to 5. These values of *p* resulted from experiments conducted using the EPOCHES text corpus. The weight of an index term appearing in a query or a document is estimated as given in [8]. Only titles of the first parts of articles were manually corrected.

c) Full text to full text matching

In this approach each candidate set of continuations is considered as a document collection and the full text of the first part is considered as a query to that collection. An implementation of the vector space model [9] is used to compute the similarity between the query and each document in the collection. In this case the similarity value, which was calculated using a cosine measure, is used as the score value. The weight of an index term i for a query Q equals to the Inverse Document Frequency (IDF) of the term in the document collection. IDF was calculated according to the formula log(N/df), where N is the total number of articles in the collection, and df is the number of articles that contain the term under consideration. The weight w_{Di} of an index term i for a document D is $w_{Di}=0.5+0.5freq_i/maxfreq_{D,}$ where $freq_i$ is the number of occurrences of term i in document D and $maxfreq_D$ is max ($freq_j$) for all $j \in D$. Although more sophisticated weighting schemes are well known today ([10]), these simple weighting schemes proved to be sufficient for our task. The OCR output was used without manual interventions.

3. Experimental results

The test set contains 50 first parts of articles having 247 candidate second parts. All test data was taken from the Greek newspaper «TO VIMA» with publication dates from 1965 to 1974. The articles were chosen randomly with only criterion that both the first and the second parts of an article had a title. The experimental results are summarized in table 1.

	Title to Title matching	**Title to Full text matching**	**Full Text to Full Text matching**
Without stemming or conflicts handling	66%	74%	80%
Stemming only	86%	88%	92%
Conflicts handling only	66%	78%	84%
Stemming and conflicts handling	86%	92%	96%

Table 1. The percentages represent successful matching.

As we can observe from the table:

- A fully automatic method, that is Full Text to Full Text matching, gives the best results compared to the other two approaches.
- Stemming had a strong positive impact on the effectiveness of the matching procedure for Greek texts
- The conflicts resolution technique helps to improve the results
- Title to Title matching technique was the weakest matcher, taking into consideration that its input was 100% correct. The main problem with this

approach was that in some cases the titles of the first and the second part of an article did not have any common words or stems excluding the stopwords.

4. Conclusion

This paper aimed at studying an important component in our effort of identifying and representing in electronic form a newspaper article, namely the process of automatically linking different parts of an article residing on different pages of a newspaper issue. To this end, software capable of segmenting a newspaper page image and identifying article parts on a single page has been augmented with methods capable of comparing article parts residing on different pages of the same issue. Three such methods were studied, namely title to title, title to full text and full text to full text matching, and their linking accuracy was evaluated using a testbed consisting of several issues of the Greek newspaper “TO VIMA”. Moreover, the effectiveness of using auxiliary methodologies such as stemming and conflict resolution criteria was studied and evaluated. It was shown that a high linking accuracy (up to 96%) can be achieved by using the appropriate combination of techniques (text to text matching with word stemming and conflict resolution criteria).

We consider that the above results are encouraging and can be used as a springboard for tackling harder cases when for example the segmentation has errors or the page where the continuation lays is incorrectly recognized by the OCR.

Acknowledgements

The authors wish to thank Dr. Ion Androutsopoulos for his helpful comments.

References

1. Gatos, B., Mantzaris, S. L., Perantonis, S. J. and Tsigris, A.: Automatic Page Analysis for the Creation of a Digital Library from Newspaper Archives. Accepted for publication to the special issue "In the Tradition of Alexandrian Scholars" of IJODL
2. Gatos, B., Mantzaris, S. L., Chandrinos, K. V., Tsigris A., and Perantonis, S. J.: Integrated Algorithms for Newspaper Page Decomposition and Article Tracking, Proccedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99), IEEE Computer Society, 559-562 (1999)
3. Fan, K., Liu, C., Wang, Y.: Segmentation and Classification of Mixed Text/Graphics/Image Documents. Pattern Recognition Letters, Vol. 15, 1201-1209 (1994)
4. Verikas, A., Bachauskene, M., Vilunas, S., Skaisgiris, D.: Adaptive Character Recognition System. Pattern Recognition Letters, Vol. 13, 207-212 (1992)
5. Strouthopoulos, C., Papamarkos, N., Chamzas, C.:Identification of text-only areas in mixed type documents. Engineering Applications of Artificial Intelligence, Vol. 10, No. 4, 387-401 (1997)
6. Jain, A. K.: Fundamentals of Digital Image Processing. Prentice-Hall, Englewood Cliffs, NJ (1989)

7. Gatos, B., Karras, D., Perantonis, S.:Optical Character Recognition Using Novel Feature Extraction & Neural Network Classification Techniques. Proc. of the Workshop on Neural Network Application and Tools, 65-72 (1993)
8. Salton, G., Fox, E. A. and Wu H.: Extended Boolean Information Retrieval, Communications of the ACM, Vol. 26, No. 12, 1022-1036 (1983)
9. Salton, G., Ed: The SMART Retrieval System – Experiments in Automatic Document Retrieval, Prentice Hall Inc., Englewood Cliffs, New Jersey, 1971
10. Singhal, A. K: Term Weighting Revisited, Ph.D. dissertation, Cornell University, 1997, available as TR97-1626 technical report