2011 International Conference on Document Analysis and Recognition

Performance Evaluation of Algorithms for Newspaper Article Identification

Roberto Beretta
Telpress, S.p.A.
Rieti, Italy
roberto.beretta@telpress.it

Luigi Laura
Dept. of Computer and System Sciences
Sapienza University of Rome
Roma, Italy
laura@dis.uniroma1.it

***Abstract*—A typical modern newspaper recognition system operates in distinct phases: i) page segmentation (also called page decomposition or zoning), that is the process of decomposing a page into its structural and logical units (called regions or zones); ii) region (or zone) labeling, where the previously identified units are labeled according to their types (title, text, images, and lines); iii) article identification (or tracking or clustering), in which all the units that belong to a single article are clustered together; and iv) read order identification, in which each item in an article is assigned its reading order inside the article.**

So far, in the literature, several works appeared describing algorithms and metrics for the first two phases, i.e. page segmentation and region labeling, that indeed play a crucial role in the whole process; however, few results focused on article identification, that is a difficult task mainly due to the rich and complex variety of newspapers layouts.

In this paper we propose a methodology to evaluate newspapers article identification algorithms; our approach is based on well-established tools from graph theory: in particular, we reduce the newspaper article clustering problem to a specific graph clustering problem, that is therefore evaluated using the appropriate *coverage* and *performance* measures.

The advantages of our approach are twofold: on one side, the proposed measures correctly detects that not all the errors are equals, i.e. some errors are worse than others, and the scores are assigned properly. On the other side, we show how to *reverse* the reduction, in order to exploit the large number of graph clustering algorithm available: indeed, given a graph clustering algorithm, to obtain a full working newspaper article identification algorithm we only need to define a similarity measure between units in the article. We provide some examples, using a specifically designed dataset.

Finally, we would like to point out that both our dataset, together with its ground-truth base, and the software tool, that implements the proposed approach, are freely available.

***Keywords*-newspaper article identification, performance evaluation, graph clustering**

I. Introduction

The Layout Analysis is a complex task for several distinct types of documents, and this is particularly true when we focus on newspapers, that present a rich and complex variety of layouts.

In this paper we focus on the problem of the evaluation of algorithms for the article identification (or tracking or clustering) problem: here, we assume that the input, for the algorithms to be evaluated, is the output of the previous phases of the Layout Analysis, i.e. Page Segmentation and Region Labeling. Therefore, we assume that the input is a series of (labeled) blocks, and the requested output is a clustering of these blocks into the logical units they belong: the articles of the newspaper page.

We propose a methodology based on well-established tools from graph theory: in particular, we reduce the newspaper article identification problem to a specific graph clustering problem. More precisely, we transform a newspaper page into a graph, where each block is a node and all the nodes that belong to the same article are connected together, i.e. they form a clique. The algorithms are therefore evaluated using the appropriate *coverage* and *performance* measures [4], that focus on intra-cluster density and extra-cluster sparsity, i.e. the evaluation favors the algorithms that form cluster in which most of the edges of the graph are internal to the clusters, and few edges are from cluster to cluster. Note that, with the above modeling, the algorithms that identify correctly all the articles achieve the maximum score in both the metrics.

Our approach allows to easily distinguish between different errors: the measures reward the correct detection of bigger articles, i.e. the one with several blocks, that the intuition suggests to be more important, at least from a layout based point of view.

Furthermore, we *reverse* our reduction, and show how to turn a graph clustering algorithm into a a full working newspaper article identification algorithm. This way, we can exploit the large number of graph clustering algorithm available: we only need to define a similarity measure between units in the article.

We implemented our approach in a software tool that is freely available[1], together with the dataset we used and its ground-truth base.

This paper is organized as follows: next section presents a brief overview of related work, whilst the graph clus-

[1]It can be downloaded from the web address www.dis.uniroma1.it/~laura/PEANAI/

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.87

IEEE computer society

tering metrics that we use, *coverage* and *performance*, are discussed in Section III. Then, in Section IV we present our framework for the evaluation of article identification algorithms; in Section V we show how to use graph clustering algorithms to develop article identification ones, and concluding remarks are addressed in Section VI.

II. Related Work

The performance evaluation of the page segmentation phase is a well studied problem; indeed, there is also a series of competitions within ICDAR conference, i.e. *ICDAR Page Segmentation Competition*, since the 2001 edition; see the work of Antonacopoulos *et al.*, [1] for a discussion on the performance evaluation of the 2009 edition.

However, when we focus on the article identification phase, there is no standard technique to evaluate the results; thus, we witnessed several distinct approaches: we cite the work of Gatos *et al.*, [2], in which the evaluation is simply done by computing the percentage of correctly identified articles; note that, since every time that *one* block is incorrectly assigned to an article, there are *two* article not correctly identified: the one the block belongs, and the one the block was assigned. The authors, therefore, report also the percentage of blocks correctly identified, that is, in their experiments, a higher value.

A different approach, and more close to our one, has been proposed by Aiello and Pegoretti in [3]: here the authors present and evaluate three distinct algorithms for article identification. These algorithms' common approach is to build a graph, called *connection graph*, in which each node is a block, and each connected component is an article. All the algorithms start with an empty graph, i.e. a graph with all the nodes and no edges, and add edges between nodes if a computed similarity value exceeds a given threshold. The ground-truth for each page is the graph in which each article is a clique. The authors evaluate the algorithms by building a Weighted Harmonic Mean of three distinct function (precision, recall, and distribution) that are computed comparing the ground-truth graph with the output graph (i.e. the connection graph produced by their algorithms).

We will discuss the differences between our approach and the one of Aiello and Pegoretti in Section IV; here we only mention the main issue: that it does not adapt to a non graph-based algorithm, i.e. an algorithm that do not explicitly compute a graph structure between blocks.

III. Graph Clustering Metrics

As mentioned before, we follow the approach of Brandes *et al.*, [4], where a good clustering of a graph separates dense subgraphs from each other. Different indices have been proposed and, as in [4], we will use the *coverage*, originally introduced in [5], and *performance*, proposed for the first time by van Dongen [6]. Given a clustering $\mathcal{C}$, in the following we denote with $m(\mathcal{C})$ the number of intra-cluster edges, and with $m(\overline{\mathcal{C}})$ the number of inter-cluster edges; obviously $m = m(\mathcal{C}) + m(\overline{\mathcal{C}})$.

Coverage of a graph clustering $\mathcal{C}$ is the fraction of *intra-cluster* edges (denoted with $m(\mathcal{C})$) within the complete set of edges. Intuitively, high values of coverage correspond to high quality clustering, but note that the trivial clustering, in which all the nodes are in only one cluster, achieves the maximum possible value (i.e., 1).

$$coverage(\mathcal{C}) := \frac{m(\mathcal{C})}{m} = \frac{m(\mathcal{C})}{m(\mathcal{C}) + m(\overline{\mathcal{C}})}$$

Performance of a graph clustering counts the number of "correctly interpreted pairs of nodes" in a graph

$$performance(\mathcal{C}) := \frac{m(\mathcal{C}) + \sum_{\{v,w\} \notin E, v \in C_i, w \in C_j} 1}{(1/2)\ n\ (n-1)}$$

As we can see from the above formula, a pair of nodes is correctly interpreted by a clustering if both nodes belong to the same cluster and are adjacent or if they are not adjacent and belong to different clusters. Analogously to coverage, performance counts the number of correctly interpreted pairs and is then normalized by the total number of node pairs. It was originally introduced in [6] and is closely related to the editing distance of a clustering, i. e., the minimum number of edges that have to be deleted or inserted in order to transform the graph into a set of disjoint complete subgraphs. Although it is similar to coverage, it additionally evaluates the sparsity between clusters and can thus favor fairly different clusterings than coverage. Generally speaking, for sparse graphs performance often prefers fine clusterings.

In this case a trivial clustering does not achieve a good value of performance (unless the graph is clique-like); coverage and performance together provide a good picture of the quality of the clustering. It is important to note that, as pointed out in [4], under some assumptions trying to maximize coverage or performance is NP-hard.

The interested reader can refer to the work of Gaertler [7] for a survey on graph clustering metrics and algorithms.

IV. Performance Evaluation of Article Identification Algorithms

In this section we describe our approach to evaluate the performance of article identification algorithms, that are algorithms whose input is a set of blocks, eventually classified (i.e. text, image, title, etc.), and the output is a clustering of this blocks into articles; as we said in the introduction, the input to this kind of algorithms is the output of a Page Segmentation algorithm, and we assume that all the blocks are correct, i.e. there has been no error in the Page Segmentation phase.

Now, let us assume we want to analyze the relative performance of an algorithm. We have a ground-truth base,

(a) (b) (c)

Figure 1. An example of the construction of the ground-truth graph: (a) the original newspaper page; (b) the distinguished article blocks, where each block is assigned a (numeric) label; (c) the ground-truth graph: each block corresponds to a node connected to the ones whose blocks are in the same article.

made of some newspaper pages whose articles have been correctly grouped together (from correctly identified blocks). The whole approach can be summarized in the following two steps:

- We build the ground-truth graph G in the following way: for each newspaper page, we build a graph whose nodes are the blocks of the page, and two nodes are linked if their blocks belong to the same article. In this way, each article in a page is represented by a clique (see Figure 1).
- We identify the articles in the page, by using the algorithm to be evaluated; let us call $\mathcal{C}$ the clustering of blocks produced by the algorithm.
- We compute the *coverage* and *performance* of the clustering $\mathcal{C}$ of the graph G.

The process of building the ground-truth graph is depicted in Figure 1: as we mentioned above, we transform each article in a clique whose nodes are the blocks that belong to the article. Then, we evaluate the performance of an article identification algorithm by computing the *coverage* and *performance* on the clustering it produced.

It is important to emphasize that the article identification algorithms can be of any type: they can vary from layout to semantic based; we do not need them to "guess" the graph structure of the ground-truth; we only need them to produce a clustering of the blocks into articles.

How do we evaluate, and weight, errors? In Figure 2 we can see a paradigmatic example: here we see the correct clustering, and two incorrect ones; which one is better? How do we evaluate them?

Before discussing the metrics, let us take a closer look to the (artificial) newspaper page of Figure 2: it is made of two "major" articles and a "minor" one. We have two distinct incorrect clustering: the first ($\mathcal{C}_A$) merges the picture and a column of the second article with the first article, whilst the third article is correctly identified; the second clustering ($\mathcal{C}_B$), instead, assigns to the two "major" articles one column each of the third article (the "minor" one).

Before discussing the metrics, the natural question is: between (the incorrect) clusterings $\mathcal{C}_A$ and $\mathcal{C}_B$, which one we do prefer? We believe that clustering $\mathcal{C}_B$ is better: the two "major" articles are complete, whilst each of them has a half of the third article. On the other side, clustering $\mathcal{C}_A$ splits one of the major articles, assigning two blocks from it to the other major article, and correctly identifies the third article. Summing up, we prefer clustering $\mathcal{C}_B$ because, if we look at the newspaper page, the first thing that we see are the two major articles, and we would appreciate them to be preserved.

In Table I we can see some metrics relative to the two clusterings. If we count the number of correctly identified articles, than $\mathcal{C}_A$ is the only that scores one. If we focus on the number of incorrectly assigned blocks, it is a tie: two each. When we consider the number of complete articles, where an article is considered complete if all its blocks are clustered together, eventually with blocks from other articles, also in this case there is a tie.

Now, let us focus on *coverage* and *performance* of the clusterings $\mathcal{C}_A$ and $\mathcal{C}_B$ against the ground-truth graph. In Figure 3 we can see the graphical representation of the graph clusterings. Here, it seems natural to say that $\mathcal{C}_B$ is a better clustering: a smaller number of edges crosses the boundaries between clusters. The reported measures of *coverage* and *performance* support this intuition: $\mathcal{C}_B$ is a better cluster when using these measures.

The above example, whilst artificial, is to clarify the crucial point: not all the errors are equals, i.e. we cannot account the same when different blocks are incorrectly classified. This is well known in the evaluation of Page Segmentation algorithms: see, e.g., the distinction between *allowable* and *non-allowable* errors in the evaluation of ICDAR 2009 Page Segmentation Competition [1].

Our proposed framework has the advantage that it does not require users to specify the relative importance of the errors: in some sense, we can say that they are deduced from the graph structure of the articles; intuitively, more blocks are in an article, more important it is, and therefore it should be weighted more.

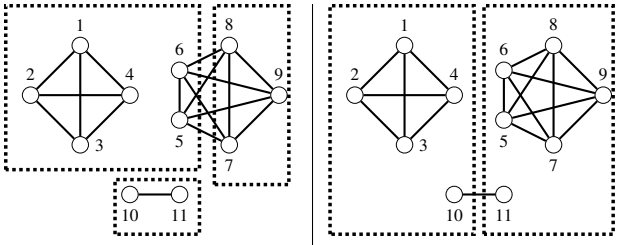
Furthermore, we note that, if we want to offer users some degree of customization, we could switch to weighted graph: we can put weights on the edges, expressing how strong we believe that the corresponding nodes belong to the same

(a) (b) (c)

Figure 2. The evaluation of article identification algorithms: the ground-truth base, i.e. the correct decomposition of the blocks into articles (a); the articles identified by algorithms ALG_A (b) and ALG_B (c).

	ALG_A	ALG_B
# of correctly identified articles	1	0
# of incorrectly assigned blocks	2	2
# of unsplit identified articles	2	2

Table I
SOME METRICS RELATIVE TO THE EXAMPLE SHOWN IN FIGURE 2.



cov = 0.647 perf = 0.745 (a)

cov = 0.941 perf = 0.818 (b)

Figure 3. Relatively to the example shown in Figure 2, here we see a graph(ical) view of clustering $\mathcal{C}_A$ (a) and $\mathcal{C}_B$ (b); also the corresponding values of *coverage* (cov) and *performance* (perf) are shown.

article, and then use the weighted version of *coverage* and *performance* [7].

We end this section with a comparison with the approach of Aiello and Pegoretti in [3]: as we said in Section II, they build a graph, called *connection graph*, using three distinct algorithms. As in our approach, the nodes of the graph are the blocks, and also their ground-truth base is made of one clique for each article. Then, their algorithms identifies the articles as the connected components of the connection graph.

The major difference between our approach and their one is that they evaluate their algorithm with respect to how closely their *connection graph* matches the ground-truth graph. Assume two algorithms identifies correctly each article in a page, but one of them builds a tree (that is the smallest connected subgraph possible) for each article, whilst the second one builds a clique, like in the ground-truth graph. The algorithms, therefore, would have been evaluated differently, whilst both of them correctly identified the articles. Going along the same lines it is possible to build

Figure 4. A screenshot of the software tool, in the evaluation phase of an article identification algorithm. Here, blocks with the same color belong to the same article.

examples in which algorithms that output the perfect article clustering are evaluated less than algorithms that fail in it, but build a connection graph closer to the ground-truth base graph. This is because the object of the evaluation is not the article clustering, but the graph structure. Furthermore, we also note that the approach of Aiello and Pegoretti can be used only to evaluate algorithms that build a graph structure, whilst our approach does not suffer of such a limitation.

A. Software tool and ground-based Dataset

We implemented the evaluation metric in a small tool, that allows to perform the following operations:

- Load a set of PDF files with their associated ground-truth base (i.e., the correct clustering of the articles within the newspaper).
- Define (and save) a ground-truth base for a PDF file.
- Evaluate the performance of algorithms: in the tool we implemented two article identification algorithms, and it is easy to add other algorithms, following a template interface.

We tested the effectiveness of our approach, by using the mentioned two simple article identification algorithms against a dataset made of 40 pages (10 covers and 30 internal pages) from italian newspapers.

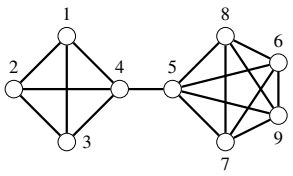


Figure 5. If an edge express the likelihood of two nodes of belonging to the same article, with the graph shown would you bet on one article (connected component) or two (clustering)?

V. Graph-based algorithms for Article Identification

In Figure 6 we show the pseudo-code of a generic article identification algorithm. The idea here is to compute a graph structure over the blocks, where an edge between two nodes denotes a certain degree of confidence that the two corresponding blocks belong to the same article. Then, we compute a clustering of the graph. In some sense, this is an approach similar to the one proposed by Aiello and Pegoretti [3], with the exception that they compute the connected components as articles, while we propose to compute clustering.

In Figure 5 we can see an example of way we believe that clustering might be better than connected components (that, by the way, are indeed a form of clustering): here there are two *natural* clusters, connected by an edge. Which are the articles behind this graph structure? Of course, we cannot be sure by simply looking at the graph structure, without seeing the actual newspaper page; but consider the following argument: when something *really important* happens, like a war or a tsunami, it usually occupies more than one page on a newspaper, and there are several related articles on each of these pages. If we use a semantic approach (i.e. not a layout based one) that makes use of the text in each block to estimate the similarity with other blocks, it is more than possible to *connect* blocks from different articles, simply because all the articles in the page deal with the same topic.

Therefore, we propose to use an approach that might drop some edge, if the whole structure suggests it (as in Figure 5). We are currently experimenting this approach, starting from the algorithms proposed in [3], and combining them with some layout based rule. A convenient feature of this approach is, indeed, the opportunity to use several engines to decide whether two blocks belong to the same article, adding edges accordingly to each of the engine, and then use a graph clustering algorithm to identify the articles.

VI. Conclusions

In this paper we presented a framework to evaluate article identification algorithms, based on graph clustering tools. Our approach allows to distinguish different degrees of errors, depending on the number of blocks of the articles incorrectly identified. We implemented our evaluation framework in a software tool that allows to evaluate article identification algorithm; the tool, together with the dataset we used for our experiments, is freely available.

```
INPUT: a set of blocks b_i
OUTPUT: a clustering of the blocks into
articles

1. For every block b_i add node i to graph G
2. For every pair of blocks b_i and b_j
    compute Similarity(b_i,b_j);
3. If Similarity(b_i,b_j) > threshold then
    add edge (i,j) to graph G
4. Cluster graph G
5. Output the clusters of G as articles
```

Figure 6. Pseudo-code of a generic article identification algorithm based on graph clustering

We also showed how to *reverse* this idea to produce article identification algorithms, based on graph clustering algorithms together with some similarity function between the blocks. We are currently experimenting some algorithms that exploit this idea.

References

[1] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos, "Icdar 2009 page segmentation competition," in *ICDAR '09: Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1370–1374.

[2] B. Gatos, S. L. Mantzaris, K. V. Chandrinos, A. Tsigris, and S. J. Perantonis, "Integrated algorithms for newspaper page decomposition and article tracking," in *ICDAR '99: Proceedings of the Fifth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 1999, p. 559.

[3] M. Aiello and A. Pegoretti, "Textual article clustering in newspaper pages," *Applied Artificial Intelligence*, vol. 20, no. 9, pp. 767–796, 2006. [Online]. Available: http://dx.doi.org/10.1080/08839510600903858

[4] U. Brandes, M. Gaertler, and D. Wagner, "Engineering Graph Clustering: Models and Experimental Evaluation," *ACM Journal of Experimental Algorithmics*, vol. 12, no. 1.1, pp. 1–26, 2007. [Online]. Available: http://portal.acm.org/citation.cfm?id=1227161.1227162

[5] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad, and spectral," in *Proceedings of the 41st Annual Symposium on the Foundation of Computer Science*. IEEE Computer Society, Nov. 2000, pp. 367–380.

[6] S. van Dongen, "Graph clustering by flow simulation," Ph.D. dissertation, University of Utrecht, 2000.

[7] M. Gaertler, "Clustering," in *Network Analysis*, ser. Lecture Notes in Computer Science, U. Brandes and T. Erlebach, Eds., vol. 3418. Springer, 2004, pp. 178–215.