

Finding the Boundaries of Information Resources on the Web

Pavel Dmitriev
Cornell University
304 Upson Hall
Ithaca, NY 14853
+1 607 255-5431

dmitriev@cs.cornell.edu

Carl Lagoze
Cornell University
301 College Ave
Ithaca, NY 14853
+1 607 255-6046

lagoze@cs.cornell.edu

Boris Suchkov
Cornell University
303 Upson Hall
Ithaca, NY 14853
+1 718 986-3911

bvs2@cornell.edu

ABSTRACT

In recent years, many algorithms for the Web have been developed that work with information units distinct from individual web pages. These include segments of web pages or aggregation of web pages into web communities. Using these logical information units has been shown to improve the performance of many web algorithms. In this paper, we focus on a type of logical information units called *compound documents*. We argue that the ability to identify compound documents can improve information retrieval, automatic metadata generation, and navigation on the Web. We propose a unified framework for identifying the boundaries of compound documents, which combines both structural and content features of constituent web pages. The framework is based on a combination of machine learning and clustering algorithms, with the former algorithm supervising the latter one. Experiments on a collection of educational web sites show that our approach can reliably identify most of the compound documents on these sites.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *clustering, retrieval models*. I.2.6 [Artificial Intelligence]: Learning – *parameter learning*.

General Terms

Algorithms, Human Factors.

Keywords

WWW, Clustering, Compound Documents.

1. INTRODUCTION

Recent research has demonstrated the value of treating the web as a set of logical information units that don't necessarily correspond to individual web pages; i.e., the information accessible via a single URL. For example, algorithms that segment web pages into semantic blocks have been shown to improve the performance of web information retrieval [1]. At the other end of the spectrum, algorithms that cluster individual web pages into web communities, have proven effective in numerous applications, such as finding collections of high quality pages on a given topic [3], finding related pages and identifying relationships among topics [5], and modeling evolution of the Web [4].

In this paper, we focus on a distinguished type of logical information units called a "compound document" (cDoc) [2] and

describe a framework for automatically generating them. Intuitively, a cDoc is a set of pages that in aggregate correspond to an individual's perception of a larger information entity. This is analogous to membership principles in the physical world, where books are composed of constituent pages, photo albums of individual pictures, etc. A typical example of such a cDoc is a web news article, consisting of several html pages, or set of web pages describing a subject such as the biology of monarch butterflies, with different pages dedicated to anatomy, sensory systems, life cycle, etc.

The ability to identify cDocs would be useful in a number of important applications. It can improve both recall and precision of the information retrieval on the Web [2]; it can help link analysis algorithms, such as PageRank, to identify links serving solely navigational purpose; it would allow better automatic metadata generation and association for Digital Libraries, and, finally, it can be used as a basis for improved user interfaces.

We present in this paper a technique that combines machine learning and clustering to automatically identify cDocs.

2. DESCRIPTION OF THE FRAMEWORK

Intuitively, it seems rare that a cDoc spans across multiple web sites (by a "web site" we mean all pages under the same domain name). Thus, the work presented in this paper looks at one web site at a time. However, it can be easily generalized to deal with an arbitrary set of pages. We represent a web site as a graph with nodes corresponding to the pages, and edges corresponding to the hyperlinks. Our framework consists of two stages.

The first (training) stage takes as input several web sites with the cDocs on them labeled by a human. Given these web sites, we train our Machine Learning algorithm (a weighted variant of Naïve Bayes) to learn the right way to compute the weights on the edges of the graph as follows. First, for each pair of pages connected by an edge, a set of useful features is extracted. Second, the value of every feature is transformed into a real number and attached to an edge between the two pages. Thus, for each edge, there is a vector of real numbers attached to it. We also know from the labeling whether the two pages connected by this edge belong to the same cDoc. The Machine Learning algorithm learns a mapping from the space of these vectors to the space of real values representing weights on the graph edges, trying to make the weight higher if the two pages connected by a hyperlink corresponding to the edge are more likely to belong to the same cDoc.

On the second stage, given a new web site, we wish to group the pages on the web site into a set of clusters, each cluster

corresponding to a cDoc. We repeat the process of extracting the features from web pages, and transforming them into vectors of real values (steps 1 and 2 on the figure 1). Then, we use the Machine Learning algorithm we trained to compute the weights on the edges (figure 1, step 3). Finally, the Graph Clustering algorithm (a variant of a Shortest Link clustering) is applied to the weighted graph (figure 1 step 4). The clusters are the cDocs we are looking for. This whole process is depicted on figure 1.

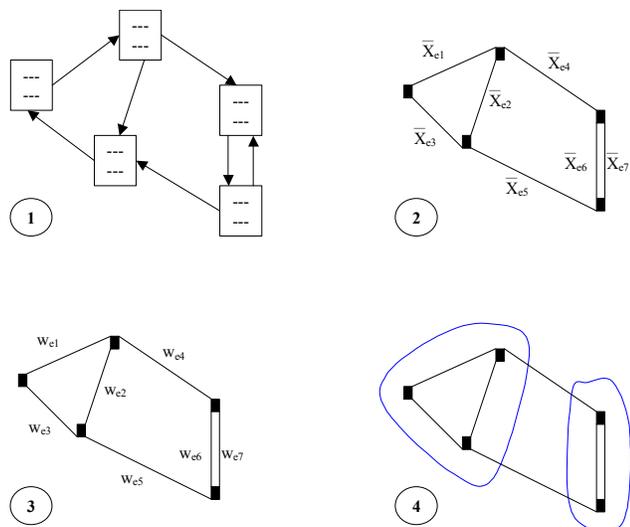


Figure 1. The process of detecting compound documents on a new web site.

3. EXPERIMENTAL SETUP

To evaluate our approach we used a dataset of 50 web sites on various educational topics. The web sites were downloaded and then all processed by a human to manually identify cDocs on them. The size of a web site in the dataset varied from 16 to 1011 web pages, averaging at 258.9 pages. The number of cDocs on a web site varied from 1 to 169, averaging at 26.8 documents. Finally, the number of web pages in a cDoc varied from 1 to 274, averaging at 9.6 pages.

In addition to evaluating the ability of our approach to identify compound documents correctly by comparing the set of documents identified by our approach to the one identified by human, we conducted several sets of experiments testing different properties of our algorithms, such as the sensitivity to the selection of training examples, and the influence of inclusion/exclusion of a particular feature.

4. RESULTS AND CONCLUSION

The main results of our experiments are the following:

- The vast majority of clusters produced by our approach were evaluated as *good*, or *ok* by a human rater. There were few really bad clusters, most of which were produced due to one of the three reasons discussed below.
- Our approach can identify correctly most discriminative features from a relatively small number of training examples (6-10 web sites with at least 100 cDocs on them).
- The approach is insensitive to the choice of training examples. This suggests that there is a high degree of redundancy present

in the data, which allows detecting the correct discriminative features from any (large enough) set of training examples.

- Dropping one feature at a time did not lead to a significant decrease in performance, which means that no single feature is vital to the success of our approach.

There were three main reasons for mistakes made by the algorithm, which we plan to address in our future work. The first problem was due to *singletons*, cDocs consisting of only one page. There was a considerable number of such cDocs in our dataset, but the current clustering algorithm is unable to detect them. The second problem is due to so-called *contents pages*. These are the pages resembling a contents page of a journal. A contents page links to many different cDocs, and ideally should either be a part of every cDoc it links to, or should be declared a singleton. Finally, the third problem is due to so-called *vocabulary pages*. These are pages resembling a vocabulary. Such pages typically have a large number of inlinks coming from other pages of the site. Clearly, we would like all pages in such a vocabulary to be a single cDoc, however, our current algorithm mistakenly assigns every page in such a vocabulary to one of the cDocs linking to it.

5. CONCLUSION

In this paper we consider a problem of automatically identifying compound documents on the Web. Our approach to solve this problem is based on the combination of machine learning and clustering algorithms, with the former providing the necessary information for the latter one. Using machine learning to identify the features that define a compound document makes our approach applicable in a variety of different contexts and adaptable to a number of different target audiences. The experiments on a dataset consisting of 50 educational web sites showed that our approach could be used to reliably find most of the compound documents in the dataset.

6. ACKNOWLEDGEMENTS

This work was funded by the National Science Foundation under grant 0227648. The ideas in this paper are those of the authors and not of the National Science Foundation.

7. REFERENCES

- [1] Cai, D., Yu, S., Wen, J.-R., Ma, W.-Y. Block-based Web Search. 27th ACM Conference on Research and Development in Information Retrieval, 2004.
- [2] Eiron, N., McCurley, K. S. Untangling compound documents on the web. 14th ACM Conference on Hypertext and Hypermedia, 2003.
- [3] Kleinberg, J. Authoritative sources in a hyperlinked environment. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998.
- [4] Kumar, P., Raghavan, P., Rajagopalan, S., Tomkins, A. Extracting large-scale knowledge bases from the Web. 25th Conference on Very Large Data Bases, 1999.
- [5] Toyoda, M., Kitsuregawa, M. Creating a Web community chart for navigating related communities. 12th ACM Conference on Hypertext and Hypermedia, 2001.