Chapter 14
Logical Structure Recovery in Scholarly Articles with Rich Document Features

Minh-Thang Luong
National University of Singapore, Singapore

Thuy Dung Nguyen
National University of Singapore, Singapore

Min-Yen Kan
National University of Singapore, Singapore

ABSTRACT
Scholarly digital libraries increasingly provide analytics to information within documents themselves. This includes information about the logical document structure of use to downstream components, such as search, navigation, and summarization. In this paper, the authors describe SectLabel, a module that further develops existing software to detect the logical structure of a document from existing PDF files, using the formalism of conditional random fields. While previous work has assumed access only to the raw text representation of the document, a key aspect of this work is to integrate the use of a richer representation of the document that includes features from optical character recognition (OCR), such as font size and text position. Experiments reveal that using such rich features improves logical structure detection by a significant 9 F1 points, over a suitable baseline, motivating the use of richer document representations in other digital library applications.

INTRODUCTION
The pace of scholarly exploration, publication and dissemination grows faster every year, reaching unprecedented levels. To support this level of innovation, scholars increasingly rely on open-access mechanisms and digital libraries, portals and aggregators to disseminate their findings (Brown, 2009). While there is controversy over which of the trends of search engines, open access, preprint and self-archiving have most influenced the growth of scientific discovery, the consensus
Logical Structure Recovery in Scholarly Articles with Rich Document Features

is that these batteries of methods have bettered the dissemination of scholarly materials. Now, an arguable bottleneck in the scientific process is in the processing, sensemaking and utilization of scholarly discoveries for the next iteration. Scholars are still largely confined to printing, reading and annotating the papers of their interest offline, without the help or guidance of a digital library to organize and collect their thoughts.

We believe a key component of a strategy to address this gap is in building applications that take advantage of the logical structure and semantic information within the documents themselves. Even within the limited domain of computer science, searching for competing methodologies to solve a problem, analyzing empirical results in tables, finding example figures to use in a presentation, or determining which datasets have been used to evaluate an approach, are all comparative tasks that researchers do on a regular basis. Unfortunately, currently these can only be done manually, without aid from any computing infrastructure.

To support such analytics is not trivial and requires groundwork. One important subtask that is common to all of the above problems is to obtain the logical structure of the scholarly document. We paraphrase Mao, Rosenfeld, and Kanungo’s (2003) earlier definition and define a document’s logical structure as “a hierarchy of logical components, such as (for example) titles, authors, affiliations, abstracts, sections, etc.” Note that the logical structure we seek is more comprehensive than what most other published systems identify. Namely, we identify not only metadata such as title, authors, abstract and parsing references, but also the logical structure of the internals of the document – sections, subsections, figures, tables, equations, footnotes and captions.

In this paper, we present SectLabel, an open source system to solve two related subtasks in logical structure discovery: 1) logical structure classification, and 2) generic section classification. In the first task, we consider a scholarly document as an ordered collection of text lines, and need to label each text line in a document with a semantic category, representing its logical role. In the second task, we take the headers of each section of text in a paper as evidence to deduce a generic logical purpose of the section.

We accomplish our implementation by extending an existing, freely available platform for reference string parsing, ParsCit\(^1\). ParsCit uses the machine learning methodology of conditional random fields (CRF), a model that blends sequential labeling techniques with pointwise entropy-based classification. We extend the use of CRFs in ParsCit to also provide logical structure discovery through the addition of the SectLabel module.

A further reality of document processing is that inputs come in many forms of markup: from richly annotated XML representations of OCR output to noisy, raw text dumps provided by copy and paste operations. Robustness is thus highly desirable, where the tool does not fail but where output quality may gracefully degrade as the input quality becomes poorer.

We summarize our contributions as follows:

1. We present the first logical structure discovery tool that expressly caters for scientific documents. Unlike previous scholarly document processing systems, it attempts logical structure discovery over the span of the entire paper at a fine-grained (per line) level.
2. We infer the generic logical purpose of each major section of text, mapping specific sections names to generic ones (i.e., “5. Text Features” → Methodology). This promotes comparative viewing of sections with identical purpose across articles.
3. Our implemented system handles both rich formatted input from an optical character recognition system, as well as from plain text dumps of scholarly articles. We evaluate both modes of input and conclusively show that rich input enhances classification performance, especially for certain logical structure classes.