Automatic classification of scientific papers in PDF for populating ontologies

Juan C. Rendón-Miranda, Julia Y. Arana-Llanes, Juan G. González-Serna and Nimrod González-Franco
Department of Computer Science
National Center for Research and Technological Development, CENIDET
Cuernavaca, México
{juancarlos, juliaarana, gabriel, nimrod}@cenidet.edu.mx

***Abstract*— Classification of scientific papers is a task performed by specialized libraries. In a research institution, classification on this type of papers is realized in a very superficial manner and the criteria for classifying documents depends on several people. This paper describes the work related to identify different sections of the paper, automatically classify and instantiate them in an ontology in order to perform inferences about them.**

Keywords—Document classification; Information extraction; Ontology population, Classifiers

I. INTRODUCTION

Document classification is a task made to order information and to give a faster access to it. The quality of classification relies on the algorithm used and the training of the classifier. Ontology population is made to enrich the information storage and it can be used to perform inferences with the store knowledge. There are tools able to make each task (classification and ontology population) separately. But, the combination of both will help to have a faster and order access to the documents.

Scientific papers are usually published in portable document format (PDF). This format is very useful because it makes possible to have a correct visualization of the papers in a platform-independent way. However, this format was not created to automatically process text. This represents a problem when processing the text, especially when the paper is written in more than one column. When working with classification the most common thing is to create a classifier or to adapt an existing algorithm. There are open source libraries that can be used to do this, we only need to figure out how they work and adapt them to our needs.

This paper is centered in the classification of scientific papers in PDF format according to the first level of the ACM Classification System (CCS) [1]. Once the classification is done, the result is instantiated in a document ontology in order to be able to perform inferences about the papers.

II. METHODOLOGY

This section describes the methodology created to extract the required information from the PDF files, classify and instantiate them in the document ontology.

A. Text extraction from PDF files

There are several open source tools to extract text from PDF files but very few focus on identifying the text sequence with the goal of getting relevant and comprehensible result or human lecture. To extract text from PDF files we use LA-PDFText [2]. This tool is created to extract information from scientific papers and it is able to automatically process this type of documents. We use this tool to solve problems such as the heterogeneity of document formats, the identification of sections and the sequence of the paper.

LA-PDFText provides a command line interface to extract text just by providing the path of the PDF document. Once the file is analyzed, it generates a plain text document.

B. Paper characteristics identification

In order to identify elements from the paper such as: title, authors, affiliation, author's email, keywords and abstract we use LA-PDFText and GROBID [3]. LA-PDFText identifies the elements of the document through a set of rules by conceptualizing a PDF file as a XML file. GROBID is a learning machine capable of identify the mentioned elements once is trained. As a result of the analysis process with GROBID, we obtained an XML file with the extracted information.

C. Preprocessing

It is necessary that the text obtained in the extraction phase gets a cascade preprocessing before the classification of the document. This consists on normalizing the characters to lower-case, remove punctuation signs and stop words, stemming using the Porter algorithm [4] and selecting the keywords that will be useful in the classification phase.

D. Training

To train the classifier we use 13 bag of words, each one corresponding to one of the first level classes in the ACM CCS taxonomy. Every bag of word have a set of text lines, where every line is an abstract from a scientific paper taken from the ACM Digital Library, under that class. The papers used to form the bag of words also received the treatment of the preprocessing phase.

978-1-4799-3010-4/14 $31.00 © 2014 IEEE
DOI 10.1109/CSCI.2014.153

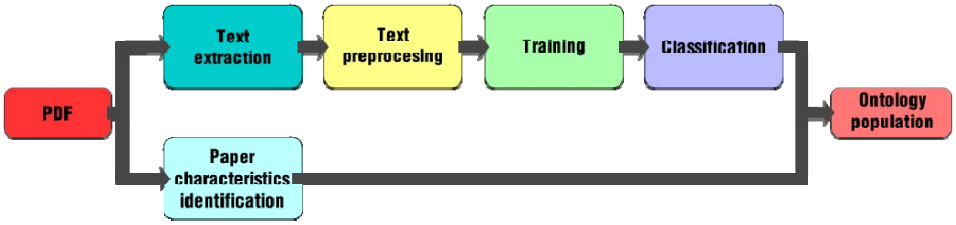

Fig. 1. Methodology for the classification of scientific papers in PDF

E. *Classification*

Naïve Bayes classification module is used to classify the document. This module is included in the Natural Language Toolkit (NLTK) [5]. To correctly classify the documents, this module takes the result of the training and preprocessing phase. Once it analyses the document, it returns the class to where the document belongs.

F. *Ontology population*

The ontology population phase is performed to store the result of the identification and classification phases. Every paper is modeled as a class with the set of characteristics extracted from the document, such as: title, authors, affiliation, abstract, keywords, etc. With this, we are able to make some inferences and to exploit the stored information in the ontology.

III. CONCLUSIONS

Within the results of the current work it is expected to create a web application where the user will be able to select a scientific paper in PDF and classify it automatically. This paper will be under every phase of the methodology in order to be classified and instantiated in an ontology that models knowledge objects (papers included). Once the ontology is populated it can be used to performed inferences and obtained implicit knowledge from the papers such as: to know how many papers an author has, to know which authors have worked with a specific one, diversity of topics of an author and the scientific production of one author compare with his peers.

REFERENCES

[1] Association for Computing Machinery, Inc, "The 2012 ACM Computing Classification System", New York, NY, 2012.

[2] Ramakrishnan, C., A. Patnia, E. Hovy and G. Burns (2012). "Layout-Aware Text Extraction from Full-text PDF of Scientific Articles."Source Code for Biology and Medicine 7(1): 7.

[3] P. Lopez. "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications". Proceedings of the 13th European Conference on Digital Library (ECDL), Corfu, Greece, 2009.

[4] M.F. Porter, (1980) "An algorithm for suffix stripping", Program: electronic library and information systems, Vol. 14 Iss: 3, pp.130 – 137

[5] E. Loper, S. Bird, "NLTK: the Natural Language Toolkit". Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics, p.63-70, July 07-07, Philadelphia, Pennsylvania, 2002.