Information Sciences 422 (2018) 542–557

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Network-based approach to detect novelty of scholarly literature

Reinald Kim Amplayo, SuLyn Hong, Min Song*

Department of Library and Information Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul, South Korea

ARTICLE INFO

Article history:
Received 4 October 2016
Revised 9 September 2017
Accepted 13 September 2017
Available online 15 September 2017

Keywords:
Novelty detection
Scholarly literature
Autoencoder neural network
Network-based feature extraction

ABSTRACT

We present a method to detect the novelty of a research paper. Because novelty in scholarly literature also examines the larger research community, a network-based approach for extracting features is proposed. Two graphs are introduced, a macro-level graph, where authors and documents are used as nodes, and a micro-level graph, where keywords, topics, and words are used as nodes. After constructing the seed graph, papers are incrementally added while changes in the graph are recorded as the feature set of a paper. An autoencoder neural network is then used as the novelty detection model. The experimental results show that the commonly used text feature representations, TF-IDF and one-class SVM, are not suitable for detecting the novelty of a research paper. Among the constructed graphs, keyword-level graph features exhibit the best performance using regression analysis as the metric. We also combine the macro-level graph, micro-level graph, and all features and find that the combination of keywords, topics, and word features perform the best using regression and citation count analysis. Other factors that could affect the citation counts, impact, and audience, are also discussed.

1. Introduction

Regardless of their specific domains, academic scholars publish new findings at pertinent conferences or in journals, and it has become profoundly important to identify papers that convey novel ideas and findings. However, identifying a novel research paper is a very difficult task. A major problem with such a task is the sudden rapid increase in the number of research papers published each year. As a reference, the number of scientific papers published from 2000 to 2015 in arXiv [16] in the field of computer science increased by approximately 3765% in 15 years. Another problem is the peer review process to which papers are subjected for publication in journals and at conferences, which takes a substantial amount of time and the judgment may be biased by the subjectivity of the reviewer.

Although there are multiple ways of addressing these problems, one potential method is to infer the novelty of a paper automatically. If the novelty of a paper could be inferred, one could filter the number of papers published in a journal in an automatic and objective manner. We define the novelty of a research paper according to Kaufer and Geisler's definition [16], which divided the definition of novelty in academic writing into three groups: 1) novelty represents the relationship between ideas and communities rather than a property or individual trait, 2) novelty can be found in the complex process

* Corresponding author.
E-mail addresses: rktamplayo@yonsei.ac.kr (R.K. Amplayo), lynn.hong@yonsei.ac.kr (S. Hong), min.song@yonsei.ac.kr (M. Song).

http://dx.doi.org/10.1016/j.ins.2017.09.037
0020-0255/

of the exchange of ideas between authors and communities, and 3) novelty can be referred to as a set of standards that must be followed to contribute to the growth of the community.

There have been a substantial number of past reports regarding the novelty detection of natural language texts using statistical [9,31,2,17] and neural network methods [20,32,19]. However, most of these methods used general text data in the form of documents, which are very different from research papers. Thus, evaluating the novelty of a research paper would need to be different from general texts. As defined earlier, measuring the novelty of a research paper requires considering the community structure to which a research paper belongs, as well as the extent to which the research paper contributes to changing the community structure.

The purpose of this study is to propose a novelty detection model for research papers that takes cognizance of the properties according to which the novelty of a research paper is measured. We propose a network-based approach that reflects the change in the structure of a graph as the research paper is added to the graph. In addition, we utilize the following three representations for detecting the novelty of a research paper that correspond to Kaufer and Geisler's [16] definitions of research paper novelty:

- The *novelty* of a paper can be measured using the change in the graph when the paper is added.
- The *ideas* and *techniques* presented in the paper can be measured using the tokens, keywords, or topic discussed in the paper.
- The *quantitative and qualitative difference* from the current literature can be measured using the degree of change of the graph when the paper is added.

The remainder of the paper is organized as follows: the section on related work discusses previous research on measuring the novelty of research papers, past techniques used in the novelty detection of texts, and network-based representation and its usage in novelty detection. The methodology section describes our datasets and the preprocessing stage, the construction of multiple graphs and the extraction of features from them, and the novelty detection model and its evaluation metrics. The results section reports the experimental results and the evaluation of the model compared to other methods and baselines. The discussion section discusses the issues pertaining to the proposed novelty detection model. Finally, the conclusion section summarizes the contributions of our work and presents future directions.

2. Related work

The aim of this study is to create a novelty detection model for research papers using a network-based approach. Therefore, we conduct literature reviews on novelty detection using texts and network-based approaches.

2.1. Measuring the novelty of research papers

Research papers are reviewed based on novelty as one of the primary criteria. Surprisingly, there have been no past efforts to automate the reviewing of research papers by measuring the novelty of a paper. However, since novelty is always assumed to be the future impact of the article on the field [27], it can be the inference of the scholarly impact of an upcoming research paper.

The most common measurement for scholarly impact is Bibliometrics, which provides a powerful set of methods to measure and evaluate the structure and process of scholarly communication [5]. There are several ways to evaluate these attributes of documents: 1) using a simple citation count, 2) counting only the recent citations, and 3) identifying papers that have won an award [6]. In recent years, many researchers have suggested that there is more to measuring scholarly impact than Bibliometrics. Piwowar [25] argued that alternative metrics, or altmetrics, should also be considered when measuring scholarly impact. This includes download counts, the number of shares to social networking sites, likes, and subscriptions. Ding et al. [7] argued that looking deeper into the text, extracting entities from that text, and using them as the main operands for the measurements have become the new frontier for measuring scholarly impact. They refer to this technique as "Entitymetrics," which is a term coined from entity and Bibliometrics.

2.2. Novelty detection of texts

Novelty detection of texts has existed for several years and was most active after the introduction of TREC 2002 Novelty Track [11], in which a task is provided to find the relevant and novel sentences pertaining to a topic and which then generates an ordered list of relevant documents. Tsinghua University [33] emerged as the leader for this task and created a technique specific to this objective, in which they performed two-step research, first to find relevant sentences and second, to eliminate repetitive information using a query, term expansion, and topic classification.

Techniques for novelty detection can be divided into two main groups: 1) statistical techniques [22], and 2) neural network-based techniques [23]. Both techniques have been used to detect the novelty of texts. Earlier methods used statistical approaches to detect the novelty of texts. Ertoz et al. [9] treated the novelty detection problem as a task of identifying topics in collections of documents. This type of treatment is understandable if one topic contains only one novel document