How near-duplicate detection improves editors' and authors' publishing experience

Yury Kashnitsky
Elsevier, The Netherlands
`y.kashnitskiy@elsevier.com`

Vaishnavi Kandala
Elsevier, The Netherlands
`v.kandala@elsevier.com`

Egbert van Wezenbeek
Elsevier, The Netherlands
`e.wezenbeek@elsevier.com`

IJsbrand Jan Aalbersberg
Elsevier, The Netherlands
`ij.j.aalbersberg@elsevier.com`

Catriona Fennell
Elsevier, The Netherlands
`c.fennell@elsevier.com`

Georgios Tsatsaronis
Elsevier, The Netherlands
`g.tsatsaronis@elsevier.com`

Abstract

We describe a system that helps identify manuscripts submitted to multiple journals at the same time. Also, we discuss potential applications of the near-duplicate detection technology when run with manuscript text content, including identification of simultaneous submissions, prevention of duplicate published articles, and improving article transfer service.

Keywords: scholar ethics, near-duplicate detection, peer-review process, editorial speed, article transfers

1 Introduction

Slow editorial processes and lack of communication around manuscript statuses can cause authors to submit to multiple journals in parallel. Publishers want to prevent processing such simultaneous submissions for various reasons including reputational concerns, legal ramifications, and to prevent overloading peer reviewers (Torgerson et al., 2005; Wager, 2009). A simple approach to the problem is to use manuscript content only (titles, authors, keywords, abstracts) to run near-duplicate detection, e.g. with Locality Sensitive Hashing (Leskovec et al., 2014).

2 Analysis of near-duplicate manuscripts

In this section, we describe the outcomes of a near-duplicate analysis that we did with manuscripts submitted to Elsevier journals.

Technical and methodological details:

- Following (Leskovec et al., 2014), we say that two pieces of text are near-duplicates if the Jaccard similarity of their shingle sets exceeds some threshold, 0.8 in our case
- We collected 4.6 mln. manuscripts submitted to >2K Elsevier journals between January 2018 and October 2020.
- We used titles and abstracts to run MinHash LSH for near-duplicate detection
- For analysis, we implemented MinHash LSH with PySpark, the MLlib MinHash LSH implementation [1] can serve as a reference
- The technique is limited to the analysis of lexical text similarity only, i.e. semantic similarity (paraphrasing, synonym replacement, etc.) can not be captured with LSH.

Insights and outcomes:

- **25%** of manuscripts were found to have a near-duplicate submitted earlier
- Over 90% of near-duplicate pairs of manuscript titles are those where one manuscript is a *resubmission* of another, meaning that it is submitted following a rejection of a similar manuscript
- **2.5%** of manuscripts were found to be under consideration simultaneously at more than one journal.

---

[1] `http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html`

3 Applications of manuscript near-duplicate detection

Having developed a technology for large-scale detection of near-duplicate manuscripts, we see several applications of this technology for various use cases across Elsevier: catching simultaneous submissions, prevention of duplicate published article, and improving article transfer service.

3.1 Catching simultaneous submissions

With over 2 mln. manuscripts submitted to Elsevier journals annually, and 2.5% of them being simultaneous submissions, flagging such cases to editors makes sense for the sake of saved reviewers' time. Having productionized the solution, we estimate we would be able to prevent reviewers from working on 50K redundant article reviews.

Preventing simultaneous submissions also ensures that authors get fair credit for their work which contributes to trust in the academic record (Errami et al., 2008). It can also help mitigate the problem of biased meta-analyses where duplicate studies are inadvertently included more than once (Fairfield et al., 2017).

3.2 Prevention of duplicate published articles

Article retraction is a very undesired publication outcome for any publisher. One of the reasons for retractions is a simultaneous submission not caught on time. Although such cases are rare (we analyzed 300 of them), they often lead to reputational ramifications not only for authors but for publishers as well. With near-duplicate detection technique in place, we can prevent some of these retractions.

3.3 Improving article transfer service

One possible optimization of the editorial process would be the improvement of the article transfer service which can be done if we avoid transfers to a journal that had already processed the given manuscript (or its near-duplicates) in the past. With near-duplicate detection technique, we can build such "manuscript journeys" that track journals that had processed a given manuscript (or its near-duplicates), see Fig. 1. A red arrow corresponds to such a "bad" transfer. With a preliminary analysis, we found that 1% of transfers can be improved this way.

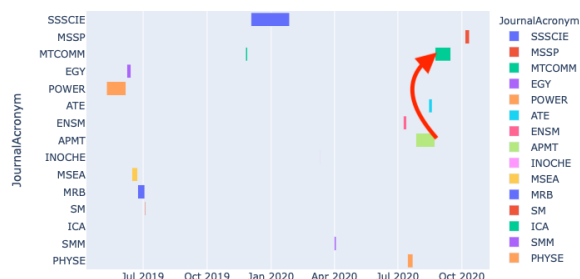Also, we can collect statistics from past resubmissions and use historic data on final successful

Figure 1: A manuscript journey and a bad article transfer. Axis X stands for time, axis Y stands for journals.

publication to make better (higher chance of acceptance) recommendations for transfers and thus improve author experience.

4 Conclusions

We see that a simple near-duplicate text detection technique, if implemented efficiently and scaled, has attached to it many important use cases leading to better and safer editorial processes, an increased author experience and a larger degree of integrity in the publication process. The technique can further be improved to account for semantic difference of the textual content like paraphrasing or changing word order.

References

Mounir Errami, Zhaohui Sun, Tara C. Long, Angela C. George, and Harold R. Garner. 2008. Déjà vu: a database of highly similar citations in the scientific literature. *Nucleic Acids Research*, 37(1):D921–D924.

Cameron J. Fairfield, Ewen M. Harrison, and Stephen J. Wigmore. 2017. Duplicate publication bias weakens the validity of meta-analysis of immunosuppression after transplantation. *World journal of gastroenterology*, 23(39):7198–7200.

Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of Massive Datasets*, 2nd edition. Cambridge University Press, USA.

David Torgerson, Joy Adamson, Sarah Cockayne, Jo Dumville, and Emily Petherick. 2005. Submission to multiple journals: A method of reducing time to publication? *BMJ (Clinical research ed.)*, 330:305–7.

Elizabeth Wager. 2009. Why you should not submit your work to more than one journal at a time. *African journal of traditional, complementary, and alternative medicines : AJTCAM*, 7(2):160–161.