A Roadmap for Composing Automatic Literature Reviews: A Text Mining Approach

Eugênio Monteiro da Silva Júnior(✉) and Moisés Lima Dutra

PGCIN, Federal University of Santa Catarina, Florianópolis, SC, Brazil
eugenio.monteiro@posgrad.ufsc.br, moises.dutra@ufsc.br

Abstract. Due to accelerated growth in the number of scientific papers, writing literature reviews has become an increasingly costly activity. Therefore, the search for computational tools to assist in this process has been gaining ground in recent years. This work presents an overview of the current scenario of development of artificial intelligence tools aimed to assist in the production of systematic literature reviews. The process of creating a literature review is both creative and technical. The technical part of this process is liable to automation. For the purpose of organization, we divide this technical part into four steps: searching, screening, extraction, and synthesis. For each of these steps, we present artificial intelligence techniques that can be useful to its realization. In addition, we also present the obstacles encountered for the application of each technique. Finally, we propose a pipeline for the automatic creation of systematic literature reviews, by combining and placing existing techniques in stages where they possess the greatest potential to be useful.

Keywords: Systematic review · Text mining · Automation

1 Introduction

It is remarkable that the scientific production keeps growing at an accelerated rate. According to [9], at August 2018 there were 33,100 active English-language peer-reviewed journals, which published together 3 million papers per year, resulting in an annual growth of approximately 5%. The large number of publications on certain topics means that writing literature reviews consumes many hours of human work, since it requires the analysis of several texts. Although information technology tools have facilitated the access to a myriad of journals around the world and have made the search process more streamlined, the human effort to find potentially useful information when a large number of documents is retrieved is still too high. According to [22], an experienced reviewer can evaluate on average two abstracts per minute and, in the case of more complex topics, each abstract may require several minutes to be evaluated. This time multiplied by hundreds or even thousands papers results in a total of many hours

E. Bisset Álvarez (Ed.): DIONE 2021, LNICST 378, pp. 229–239, 2021.
https://doi.org/10.1007/978-3-030-77417-2_17

of work, when considering only the initial stage of selecting the relevant papers. The evolution of Artificial Intelligence (AI) techniques observed in recent years, especially in the subarea known as Natural Language Processing (NLP), allows us to envisage scenarios in which these modern techniques and their associated tools can be used to enhance the process of creating literature reviews, from an automatic composition approach.

This paper aims to present an overview of the current scenario of the application of AI techniques for the automatic creation of literature reviews. Furthermore, we propose a general pipeline resulting from the combination of these techniques, in order to highlight the challenges and possibilities currently existing in this area of research. The main contribution of this work is to present the current possibilities for automating systematic reviews of literature and how they can be put to work together to facilitate the reduction of the operational workload of researchers during the conduct of a literature review.

2 Literature Review

Before thinking about automating a literature review process, it is necessary to know how it is traditionally conducted. Therefore, this section aims to conceptualize and briefly describe how to manually create a literature review.

There are several types of literature review, each one with its own objectives [7]. Among these types, the state-of-the-art and the systematic reviews (SR) stand out, as they are better known. A state-of-the-art review considers mainly the most current research in a given area or on a given topic. It often summarizes current and emerging trends, research priorities and standards in a particular field of interest. This review aims to provide a critical survey of the extensive literature produced in recent years, along with a synthesis of current thinking in the area. It may offer new perspectives on an issue or point out an area that needs more research [5]. Systematic reviews are a widely used method to gather the results of multiple studies in a reliable manner. According to [6], as a research method, systematic reviews are undertaken according to explicit procedures. The term "systematic" distinguishes them from reviews undertaken without clear and accountable procedures. According to [7], a SR seek to gather all available knowledge on a given topic with the guarantee of being transparent in reporting their methods to facilitate other researchers to replicate that process. Another function of a systematic review is to identify research gaps, in order to develop new ideas [11].

There is no consensus regarding how many steps the production of a systematic review can be divided into. Several different proposals in this regard can be found in the literature. While some authors propose only 3 steps, others like [21] suggest 15 steps. In this work, we consider the 4 steps shown below. According to [1], those steps are usually part of a review process:

– **Searching:** extensive searches are carried out to locate as much relevant research as possible according to a query. These searches include scrutinizing electronic databases, scanning reference lists, and searching for published literature.

- **Screening:** it narrows the scope of search by reducing the collection to only the documents that are relevant to a specific review. The aim is to highlight key evidence and results that may impact on the policy.
- **Mapping:** the Evidence for Policy and Practice Information and Co-ordinating Centre (EPPI-Centre)[1] has pioneered the use of "maps" of research as a method to both understand research activity in a given area and as a way of engaging stakeholders and to identify priorities for the focus of the review.
- **Synthesizing:** it correlates evidence from a plethora of resources and summarizes the results.

The process of preparing a systematic review is both creative and technical. It is worth mentioning that there is a natural dichotomy of tasks: creative tasks are performed during the development of the core question to be answered and the protocol to be applied, while technical activities can be performed automatically following exactly the applied protocol [21].

There are some standards for the development of systematic reviews in a traditional way that can serve as guides for the automation process, such as the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) statement and the PICO (Patient, Intervention, Comparison, and Outcome) framework. PRISMA consists of a checklist with 27 items and a four-phase flowchart to help authors improve the reporting of systematic reviews and undertake meta-analysis. It is focused on randomized trials, but it can also be used as a basis for reporting SR from other types of research, particularly evaluations of interventions. PRISMA may also be useful for critical appraisal of published SR [16]. Regarding the PICO framework, according to [4], it can be used to develop a well-formulated research question with a clear statement of objectives. Some other standard models that are worth mentioning are PEO (Patient, Exposure, and Outcome) and PIO (Patient, Intervention, and Outcome). They are used to formulate the inclusion and exclusion criteria defined to select relevant studies, in order to answer the research question.

3 Text Mining

Text data mining or text mining is a derivation of data mining that, instead of working with numerical and structured data, works with textual data. The main difference between regular data mining and text mining is that in text mining the patterns are extracted from natural language text rather than from structured databases of facts. Databases are designed for software applications to process them automatically; text is written for people to read. There are no programs that can "read" text as humans do and there is no evidence that they will exist in the near future. Many researchers think it will require a full simulation of how the mind works before we can write programs that read the way people do [8].

[1] A specialist center for: (i) developing methods for systematic review and synthesis of research evidence; and (ii) developing methods for the study of the use research. https://eppi.ioe.ac.uk.

However, there is a research field called computational linguistics (also known as Natural Language Processing - NLP) that is making great progress in carrying out small sub-tasks in text analysis. For example, it is relatively easy to write a program to extract sentences from a paper or book that, when shown to a human reader, appear to summarize its content [8]. The main methods of NLP used in systematic reviews are text classification and data extraction. The classification methods look for models that can automatically associate documents (abstracts, full texts or parts of these texts) with previously defined categories. The data extraction methods try to identify parts of the text or individual words/numbers that correspond to a variable of interest [15]. Since scientific production is mostly presented in textual form, AI techniques specifically aimed at processing textual data have a wide field of application to aid in the production of literature reviews.

4 Automating the Creation of Systematic Reviews

Since the production of a SR are both creative and technical, it is expected that all stages considered technical are subject to automation. Indeed, the idea of automating the steps of a systematic review is not exactly new. According to [10], the first paper to propose the use of Machine Learning (ML) to this purpose was published in 2005. From that year on, several works were published regarding the application of computational techniques in each of the SR stages. One good way to observe the evolution of this idea is by reading systematic reviews published on this subject. In 2015, while [11] published a review that exclusively covers works of data extraction in SR, [18] dedicated to review papers related to automatic identification of relevant studies.

There are several methods for implementing text mining and related tasks. The methods currently considered the most relevant to support systematic reviews are: automatic term recognition (ATR), text clustering, text classification, and text summarization [20]. There is also a large amount of software applications specifically developed to assist in the production of systematic reviews. The SR Toolbox[2] website provides a list of various available tools for supporting systematic reviews of literature.

It is important to highlight that all the technical steps of the review process can be automated through some computational technique with the main objective of reducing the human workload.

4.1 Challenges and Opportunities

After defining the theme of the research and the inclusion criteria, the first technical stage of a SR is the search for correlated studies. Ideally, 100% of the existing studies on the topic should be retrieved. Text mining can help by suggesting possible query terms. Even if the researcher already has found some documents that meet his/her inclusion criteria, he/she can always use a term

[2] http://www.systematicreviewtools.com/.

recognition service that suggest new terms and concepts to be used in a new query [20]. According to [1], term extraction improves the search strategy by creating additional metadata that can increase accuracy by automatically identifying key phrases, concepts or technical terms, within the documents. The improvement in the set of search terms has the potential to expand the coverage of the results, which may be sufficient in cases where the object of study is very specific. Thus, the number of papers retrieved is relatively small, but the entire literature is covered. In some other cases, the number of papers retrieved is very high, which makes it even more challenging to find works that are really relevant to the subject to be searched. Consequently, it is possible to think about methods that can be applied to the set of retrieved papers in order to find among them those that are really relevant.

In a systematic review, the term *screening* refers to the manual process of sifting through, at times, thousands of titles and abstracts that are retrieved from database searches. In order to improve reliability, the titles and abstracts are often screened by two people. This is a very labour-intensive task and adds considerably to the review's cost and time [20]. This is a stage where machine learning techniques can be very useful, by means of filtering not only titles and abstracts, but also full texts. According to [12], the first study to consider this possibility was [3]. According to [20], there are two ways to use text mining to automate this step: the first aims to prioritize the list of items for manual screening so that the studies at the top of the list are those most likely to be relevant; the second one uses the studies manually labeled (included/excluded) as a training dataset, so that the system can "learn" to automatically classify the other works. Apparently, conducting this step in a semi-automatic manner can bring many benefits to the researcher. However, [15] highlights that the main limitation of the automatic screening of abstracts is the fact that it is not clear at what point it is "safe" for the reviewer to the interrupt the manual screening. Even systems that, instead of providing a definitive and dichotomous classification, provide classifications based on probabilities are not free from the risk of loss. For example, a paper that has received a low probability may be relevant, and if a researcher chooses to stop screening in a certain threshold of probability, this paper may not be included in the results.

Another way to find relevant studies to a literature review is by citation mining. As an example, the study of [2] proposes a method to systematically mine the various types of citation relations between papers to retrieve documents that may be related to the topic searched by a specific systematic review. The author's proposal is conceptual and was conducted manually. This method, according to this author, despite having potential for automation, had some limitations related to the available databases API's that made it impossible to create a computational algorithm at that time. Currently, the existing databases API's provide more and more information about the indexed papers, which makes it possible to write algorithms that can automatically retrieve related papers through the citation mining technique. Moreover, [19] stresses it is possible to think about the integration of the aforementioned content-based methods with

the citation-based methods, in order to create a more efficient model for retrieving relevant papers.

Once the set of relevant studies are identified and retrieved, the next step is to extract the useful information present in each one of them. According to [21], extracting data from texts is one of the most time consuming tasks in a systematic review. Therefore, there are already several works whose objective is to automatically extract data from texts. According to [15], when considered specifically reviews of randomized controlled trial (RCT), there are only few prototypes of platforms that make these technologies available, such as ExaCT[3] [13] and RobotReviewer[4]. For basic science reviews, the NaCTeM (the United Kingdom National Center for Text Mining) has developed several systems that use structured models to extract concepts such as genes and proteins from texts. Since the desired information can be present in several sections of the paper, extracting it can become a complex cognitive task. Consequently, even partial automation can reduce the time required to complete this task, as well as reduce errors and save time [21].

One obstacle for achieving better data mining models is the lack of training data. ML systems need a dataset with manually assigned labels in order to adjust model parameters. Associating labels with individual terms in documents to enable the training of data-extraction models is an expensive task. EXaCT, for example, was trained on a small set (132 in total) of full-text papers. RobotReviewer was trained by using a much larger dataset, but the 'labels' were semi-automatically induced, using a strategy known as 'distant supervision'. This means the annotations used for training were imperfect, thus introducing noise to the model [15]. Recently, [17] released the EBM-NLP dataset, which comprises about 5000 abstracts of RCT reports manually annotated in detail. This may provide training data helpful for the development of data extraction models [15].

The last step of a SR that can be assisted by text mining techniques is the synthesis of information. According to [15], although the software tools to support the synthesis of revision data have been around for a long time (especially for performing meta-analyzes), the methods for automating it are beyond the capabilities of ML and NLP.

Furthermore, it is also possible to think about ways to automatically summarize the texts that were selected for review, by extracting information from the full texts of the papers and not just from their abstracts. For this, there is the technique that creates automatic text summaries. According to [14], this technique either generates a summary for a single document at once or for multiple documents together (MDS - multi-document summarization), by extracting the most relevant information found within the texts. Automatic summarization is quite important in systematic review processes, as it condenses the information that was discovered and classified and thus provides a solution to the information overload problem [20].

[3] https://exact.cluster.gctools.nrc.ca/ExactDemo/.
[4] https://www.robotreviewer.net/.

The use of MDS methods offers the benefits of reducing the overwork on the reviewer, as well as enabling an overview of a body of research. However, the proper place and use of such summarization must be established for it to offer the greatest benefit, regarding the current state of the art. This is partly an issue for a system designer but also partly an issue of training and experience for the reviewer. Thus, running a MDS on a large collection of texts from many domains, on many subjects, would probably not be a useful exercise and would indicate a lack of understanding about getting the most from summarization. However, if the reviewer have previously produced a cluster or a classification of documents, then it makes sense to apply the MDS, since documents in a cluster or class can reasonably be expected to have something in common, consequently, the results would be meaningful [20].

Finally, there are the Natural Language Generation (NLG) technologies, which can be used to automatically write specific paragraphs of the review, such as a description of the types of documents retrieved, results of the evaluation, and summary of the conclusions [21]. Currently existing techniques are not able to produce perfect texts like those written by humans. However, the automatically generated text can serve as basis for the text to be written manually by reviewers, e.g. avoiding errors in data transfers from multiple sources. Importantly, this kind of technology still has a lot to be improved. Thereby, in search for more integrated tools to automate systematic reviews, researchers should be aware of the new text generation methods that are emerging.

4.2 A Pipeline for Creating Systematic Literature Reviews

The intention of this paper is not to present a strict rule of how an SR should be automated, nor to indicate specific tools or technologies for that purpose. The objective here is just to highlight, based on what has already been presented in the scientific literature on this topic, the stages of the literature review pipeline with the greatest potential for automation. We try to indicate what should be the focus of researchers on AI, when they go to work within this theme. Disregarding the steps that naturally involve a creative process and, consequently, must be performed by humans, the next paragraphs focus on the operational tasks that are part of the reviewing process. In Fig. 1, we propose a pipeline that combines several techniques used by different projects to automatically generate a literature review. This pipeline shows not only a sequence of technical steps required for the creation of an automatic literature review, but also the respective AI techniques that can be useful in each phase.

In the searching phase, computational techniques can help in suggesting terms to maximize the amount of documents retrieved, however, the human operator remains essential to carry out the process. Thus, this phase is considered to have a medium automation potential. In spite of that, the works found in the literature propose techniques with great potential for increasing the degree of automation in this stage. Machines usually perform this task better than humans do. Besides, scientific databases are increasingly providing structured data on references, which facilitates the automation process. In addition, it is

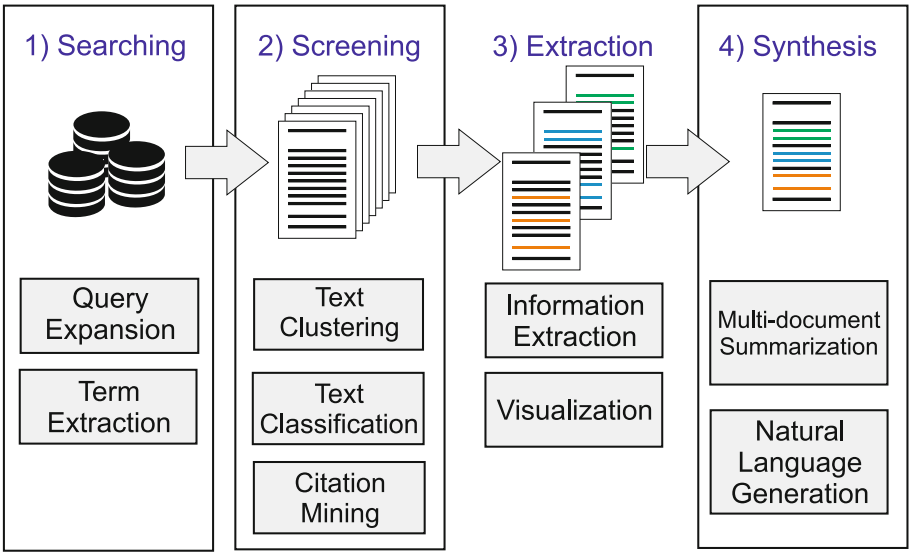

Fig. 1. A pipeline for automatically creating systematic literature reviews

possible to work on the direct extraction of references from papers (i.e. PDF files). Therefore, this is a step that deserves more attention and investment from researchers.

As for the screening stage, we consider that it still has a low potential for automation, currently. That statement comes from the fact that the conversion of texts into vectors while preserving semantic relations is still incipient, among other causes, due to the small number of positive training examples available. Text classification methods are extremely dependent on a good conversion of texts into vectors, since their accuracy in classification is highly impacted by the extraction of text characteristics. Usually, ML-based classification methods depend on training data to 'learn' patterns. In systematic reviews, the number of papers labeled as 'included' is less than the size of the set of 'excluded' papers, which makes it difficult to properly adjust ML-based models due to this imbalance of sets. As previously mentioned, given the risk of losing potentially relevant papers during the screening stage, researchers may not feel secure in delegating the exclusion of much of the retrieved papers to an automatic classification process. Thus, we believe it is still necessary to develop new methods for extracting more precise text characteristics, so that it is possible to consider that automatic sorting as a secure time saver for researchers.

The extraction and synthesis steps present a great potential to be automated. In these stages, computational techniques operate by extracting and organizing important information from texts. Various techniques for extracting certain data from texts are being developed and can be applied at this stage of automatic creation of literature reviews. Especially for medical reviews, which are already

more standardized, there is great potential for applying these techniques. Nevertheless, as the natural language processing keeps evolving, it is possible to imagine for a near future the extraction of texts for automatically creating literature reviews to be applied to some other areas of knowledge, such as the social sciences.

5 Conclusions and Next Steps

Automating literature reviews is a promising research field because the number of published papers grows every year. The large amount of available texts makes human work difficult in writing scientific literature reviews. For this reason, the development of computational tools to assist researchers in this purpose continues to arouse interest in the scientific community. It is important to highlight that, due to the many limitations of the existing computational techniques, there are still no definitive/standardized tools to help in the automatic creation of systematic reviews.

In this way, the papers found in the literature only present specific/partial solutions for certain stages of the construction of a systematic review. The supervised methods, despite being very useful in some of these phases, face the problem of lack of data for training. Consequently, these techniques present less potential for development in SR. As for the unsupervised methods, there are greater possibilities. Summarization, visualization and document clustering are examples of tasks that can help researchers deal with the large number of publications available, without relying on previously-labeled databases for training. For this reason, the development of computational models that will contribute to the reduction of human workload, especially during the operational stages of SR, can provide more agility to the process of generating scientific knowledge. This article brings together some existing initiatives aimed at this purpose. In the stages of search, screening, extraction and synthesis, some computational techniques have already been used in order to facilitate the reviewer's work.

As for future work, the computational implementation of the proposed pipeline will be carried out. Ideally, this implementation will use mainly unsupervised methods to avoid relying on training data, which is still very scarce. We intend to use existing algorithms for grouping, extracting and synthesizing information, available in the literature, that best adapt to the scenario that is being worked on. Our ultimate goal is to achieve a complete solution to automate the operational steps of a systematic literature review. As a subsequent step to the development of the prototype, we intend to test it in application scenarios from different areas of knowledge, and make it available for specialized researchers in these areas to qualitatively evaluate the results obtained.

References

1. Ananiadou, S., et al.: Supporting systematic reviews using text mining. Soc. Sci. Comput. Rev. **27**(4), (2009). https://doi.org/10.1177/0894439309332293

2. Belter, C.W.: Citation analysis as a literature search method for systematic reviews. J. Assoc. Inf. Sci. Technol. (2015). https://doi.org/10.1002/asi.23605
3. Cohen, A.M., et al.: Reducing workload in systematic review preparation using automated citation classification. J. Am. Med. Inform. Assoc. **13**(2), (2006). https://doi.org/10.1197/jamia.M1929
4. Davis, D.: A practical overview of how to conduct a systematic review. Nurs. Stand. **31**(12), (2016). https://doi.org/10.7748/ns.2016.e10316
5. Dochy, F.: A guide for writing scholarly articles or reviews for the Educational Research Review (2006)
6. Gough, D., Thomas, J., Oliver, S.: Clarifying differences between review designs and methods. Syst. Rev.**1**, 28 (2012). https://doi.org/10.1186/2046-4053-1-28
7. Grant, M.J., Booth, A.: A topology of reviews: an analysis of 14 review types an associated methodologies. Health Inf. Libr. J. (2009). https://doi.org/10.1111/j.1471-1842.2009.00848.x
8. Hearst, M.A.: What is text mining? https://people.ischool.berkeley.edu/~hearst/text-mining.html. Accessed 12 Oct 2020
9. Johnson, R., Watkinson, A., Mabe, M.: The STM Report: An Overview of Scientific and Scholarly Publishing. 5ª edição. STM: International Association of Scientific, Technical and Medical Publishers, The Hague (2018)
10. Jonnalagadda, S., Petitti, D.: A new iterative method to reduce workload in systematic review process. Int. J. Comput. Biol. Drug Des. **6**(1–2), 5–17 (2013). https://doi.org/10.1504/IJCBDD.2013.052198
11. Jonnalagadda, S.R., Goyal, P., Huffman, M.D.: Automating data extraction in systematic reviews: a systematic review. Syst. Rev. (2015). https://doi.org/10.1186/s13643-015-0066-7
12. Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., Ouzzani, M.: Learning to identify relevant studies for systematic reviews using random forest and external information. Mach. Learn. **102**(3), 465–482 (2015). https://doi.org/10.1007/s10994-015-5535-7
13. Kiritchenko, S., et al.: ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med. Inform. Decis. Mak. **10**, 56 (2010). https://doi.org/10.1186/1472-6947-10-56
14. Mani, I.: Automatic Summarization, John Benjamins, Amsterdam (2001)
15. Marshall, I.J., Wallace, B.C.: Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst. Rev. (2019). https://doi.org/10.1186/s13643-019-1074-9
16. Moher, D., et al.: Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. (2009). https://doi.org/10.1371/journal.pmed.1000097
17. Nye, B. et al.: A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistic (2018)
18. O'Mara-Eves, A., et al.: Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst. Rev. (2015). https://doi.org/10.1186/2046-4053-4-5
19. Sarol, M.J., Liu, L., Schneider, J.: Testing a citation and text-based framework for retrieving publications for literature reviews. In: BIR 2018 Workshop on Bibliometric-Enhanced Information Retrieval (2018)
20. Thomas, J., McNaught, J., Ananiadou, S.: Applications of text mining within systematic reviews. Res. Synth. Methods (2011). https://doi.org/10.1002/jrsm.27

21. Tsafnat, G., et al.: Syst. Rev. Automat. Technol. Syst. Rev. (2014). https://doi.org/10.1186/2046-4053-3-74
22. Wallace, B.C., et al.: Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinform. (2010). https://doi.org/10.1186/1471-2105-11-55