Automated Extraction of Results from Full Text Journal Articles

R. Wes Crues
University of Illinois
Dept. of Educational Psychology
1310 South Sixth Street
Champaign, Illinois
crues2@illinois.edu

ABSTRACT

Recent mandates by federal funding agencies and universities to create open access repositories of published research allow researchers a wealth of texts to analyze. Furthermore, some publishers of academic texts have begun creating policies to permit non-commercial text mining of journal articles. This project follows the approach of [7], which automatically extracts result sentences from full-text biomedical journal articles by using support vector machines and naive Bäyes classifiers. I also experiment with using the least absolute shrinkage and selection operator (LASSO) [6, 18] as a method to select features for the classifiers. I compare this new approach with other feature selection strategies used in previous studies.

Keywords

Information extraction, text classification, feature selection

1. INTRODUCTION

Information overload is hardly a new concept, with even the Ancient Roman scholar Seneca the Elder claiming in 1 AD, "the abundance of books is distraction" [8]. Similarly, the automatic summarization of text has been researched since at least the 1950's, with Luhn's work on creating abstracts automatically [11]. In concert, United States (US) federal funding agencies, such as the National Institutes of Health (NIH) [13], the National Science Foundation (NSF) [14], and the Institute for Educational Sciences (IES) [9], and university systems such as the University of California (UC) [1] have adopted open access policies for funded and published research. Publishers of academic journals, such as Elsevier [4] and Springer [15], have adopted policies for non-commercial research of texts. Finally, some national governments (e.g., the United Kingdom (UK) [10]) have adopted changes to copyright law allowing for non-commercial research of copyright protected works.

Given these open-access and legal policy changes, a wide swath of researchers now have access to a wealth of texts to automatically analyze. Specifically, the shifts in policies and laws allows for text mining to extract result sentences from full-text journal articles. Further, publishers have created APIs which allow for access to texts. It is unlikely that future researchers will be able to carefully read and analyze all of the texts in order to extract pertinent results. However, open-access policies in the US by the NIH have enabled automated extraction since the late 2000s in some fields.

My research seeks to first expand the work done in the biomedical sciences, particularly in [7] to the educational sciences, but also to explore an additional feature selection technique. This experiment is to complement the work in [20] by using the LASSO as a feature selection technique.

2. BACKGROUND

Text mining has been recognized as a tool to reduce the time required to complete a systematic literature review [17]. There are several tasks text mining can simplify when creating a systematic review. Current text mining approaches allow relevant studies to be identified, by identifying relevant search terms, and describing the characteristics of prior investigations can be accomplished by automatic summarization [17]. This proposal is inspired by the systematic search of literature using targeted queries by the information scientist, Don Swanson, who revealed a link between magnesium and migraines in the late 1980s [16]. This finding is novel because it linked medical literature with chemistry literature. Thus, I want to uncover previously unrealized links, contradictions, and confirmations in the current literature on on how students utilize computers to enhance or hinder their educational experience.

Supervised learning using text has been heavily researched in the biomedical sciences. For example, [12] proposed to use a modified naïve Bayes classifier which can determine whether an abstract is relevant for a given topic, based on the words in previously seen abstracts. They also propose a unique weighting scheme which allows for high recall and reasonable precision. In their work, they show their proposed process can significantly reduce the time required to conduct a systematic literature review. Given the amount of publications available following from the aforementioned changes, these results could help educational researchers significantly reduce time to determine which previously published work is most relevant.

More broadly, this work addresses the need to have a "living systematic literature review" where the most up-to-date published findings can be included for practitioners and researchers to implement and be informed of these findings [3]. One study found the average time between a published finding and inclusion in a systematic literature review to average between 2.5 and 6.5 years [3]. This relates directly to an initiative by the US's Institute of Educational Sciences to use evidence based practices [19]; that is, connecting the knowledge from research to practicing the knowledge.

3. APPROACH

This project will extract sentences containing results from full-text journal articles in peer-reviewed journals. Given that journals have dozens of volumes and issues, it is likely not feasible to read and find all relevant articles needed to understand prior research. This process will create a systematic review of literature from educational journals in a targeted area: student interaction and behavior in computing environments. The systematic review will inform researchers on previous findings and update practitioners on the most current research.

3.1 Extracting Results

To extract result sentences, I will parse full-text journal articles into sentences, using a tokenizer, for example, Python's NLTK [2]. Next, I label the sentences as either containing a result or not, as well as indicate the section of the article where the sentence lies, and whether the sentence is the first or last in the respective paragraph, following from [7]. In [7], result sentences were distributed throughout the journal articles and were most common in the first or last sentence of the paragraph. Then, I will experiment with various classifiers, such as support vector machines, naïve Bayes classifiers, decision trees, and various ensemble models. The output of the classifiers will be the sentences containing results, which can then be used to form a thorough systematic review.

To train these models, I will select features using traditional metrics, such as information gain, mutual information, and the χ^2 statistic [20], which are the ones used by [7]. Interestingly, using these three feature selection strategies, not one term was selected by all three methods; however, there was overlap with terms for the χ^2 statistic and information gain, and information gain and mutual information. Because of this finding, I propose to use a different feature selection technique to select words or surface level knowledge (e.g., sentence position, section of paper) to train these classifiers.

3.2 Feature Selection

Another experiment I plan to conduct to extract words from the corpus of sentences from the journal articles is to utilize the LASSO to select words to use to train classifiers to discern sentences containing results from those that do not. Given that the LASSO is used for high dimensional data sets as a variable selection technique, in fields such as gene-expression analysis [5], this approach seems reasonable given the high dimensionality and sparseness of text data. I will experiment with various parameters of the LASSO to ensure reasonable feature selection; that is, a feature set which is not prohibitively small to provide high recall and reasonable precision, but one which is not too big to prohibit generalizablity.

The specific binomial logistic LASSO model I will use to select terms is

$$\log \frac{P(result = 1|\mathbf{x})}{P(result = 0|\mathbf{x})} = \beta_0 + \mathbf{x}^T \beta, \qquad (1)$$

where $result$ equals one if the sentence x_i contains a result, and zero otherwise. Note that $\mathbf{x}$ is a matrix, where each row is a sentence, one column is $result$, and the other columns are words and surface-level features about the sentence. In the estimation phase, the model's likelihood function is penalized by a shrinkage parameter λ. This shrinkage parameter shrinks unimportant βs towards zero, thus leaving only the most important terms with nonzero βs. These terms will then be used to train the classifiers to extract result sentences to be used in systematic literature reviews. Further, the magnitude of each β can be beneficial in determining relative importance of a term.

For this portion of the project, I will experiment with various λs to determine which give the best performance when training the models to extract result sentences. A comparison of the feature selection strategies in [7, 20] will be conducted to determine any relationship between these feature selection strategies and the LASSO.

4. CURRENT STATUS

My current tasks are to complete a literature review of text classification. In this literature review, I address traditional classifiers from multivariate statistics and machine learning, but also accompany background on generating systematic literature reviews. The literature review also includes a discussion of evidence based practices and speculates on how a living systematic literature review might impact education research.

A concurrent stage is procuring and processing texts for analysis. In [7], seventeen full-text articles were analyzed, with around 2550 total sentences being considered. Thus, once all texts have been selected, I will begin labeling the sentences as containing a result or not containing a result. Efforts are underway to procure a small research fund to pay a research assistant to also label sentences as a measure of inter-rater reliability.

5. PROPOSED CONTRIBUTIONS

This work provides contributions to the fields of information science and educational data mining. One contribution is an alternative feature selection strategy which could improve performance of supervised learning methods. Because feature selection is arguably the most important analysis phase in text classification, using the LASSO in addition to strategies already used might help better performance in text classification.

Another contribution of the work is introducing the concept of a living systematic literature review to educational research. Due to the explosion of the amount of published research in education, and the interest in evidence based practice to be utilized in education, this work can address those desires.

6. ADVICE SOUGHT

I would like advice on any or all of these concerns:

1. Are there other approaches, besides classifiers such as support vector machines, naïve Bayes, discriminant analysis, neural networks, and decision tree classifiers that would be useful for this approach?

2. What suggestions do you have for analyzing the result sentences once they have been discovered by the classification algorithms?

3. Do you have any suggestions for experiments with the shrinkage parameter, λ, for selecting terms when using the LASSO?

4. Are there any specific metrics you would suggest to use for analyzing the results of either result extraction or selecting terms?

7. REFERENCES

[1] Academic Senate of the University of California. UC systemwide academic senate open access policy, 2013.

[2] S. Bird. NLTK: The natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics, 2006.

[3] J. H. Elliott, T. Turner, O. Clavisi, J. Thomas, J. P. Higgins, C. Mavergames, and R. L. Gruen. Living systematic reviews: an emerging opportunity to narrow the evidence-practice gap. *PLoS Med*, 11(2):e1001603, 2014.

[4] Elsevier, Inc. Text and data mining policy, 2014.

[5] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*, volume 1. Springer, 2001.

[6] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[7] H. A. Gabb, A. Lucic, and C. Blake. A method to automatically identify the results from journal articles. *iConference 2015 Proceedings*, 2015.

[8] Hewlett Packard. Dizzying volumes of data is nothing new.

[9] Institute of Educational Sciences. IES policy regarding public access to research, 2016.

[10] Intellectual Property Office. Exceptions to copyright: Research, 2014.

[11] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165, 1958.

[12] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Blenis. A new algorithm for reducing the workload of experts in performing systematic reviews. *Journal of the American Medical Informatics Association*, 17(4):446–453, 2010.

[13] National Institutes of Health. Revised policy on enhancing public access to archived publications resulting from NIH-funded research, 2008.

[14] National Science Foundation. NSF's public access plan: Today's data, tomorrow's discoveries (NSF 15-22), 2015.

[15] Springer. Springer's text- and data-mining policy, 2016.

[16] D. R. Swanson. Migraine and magnesium: Eleven neglected connections. *Perspectives in Biology and Medicine*, 31(4):526–557, 1988.

[17] J. Thomas, J. McNaught, and S. Ananiadou. Applications of text mining within systematic reviews. *Research Synthesis Methods*, 2(1):1–14, 2011.

[18] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[19] US Department of Education: Institute of Educational Sciences. Identifying and implementing educational practices supported by rigrous evidence: A user friendly guide, 2003.

[20] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.