**PREDICTION OF LEVEL OF INFLUENCE OF A RESEARCH PAPER**

**Avijit Sengupta, Kimia Keshanian**, **Arpita Nayak**, **Shaily Saigal**

The primary objective of the study is to explore the possibility of making a successful prediction about the level of influence a research paper has on the paper which has cited the research paper.

To address our research questions, we have developed a unique dataset. Our dataset consists of 10 MISQ papers with 1129 number of instances and 40 human engineered features to correctly identify the level of influence a research paper on another paper in a particular citation context

Before deploying any machine learning algorithm, we used the least absolute shrinkage and selection operator (LASSO) regression method to performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability. Upon selection of the important attributes or variables, we deployed a machine learning approach to automatically identify the level of influence a research paper has on a paper which has cited it in a particular context. We deployed eight different kinds of model to successfully classify the level of influence (high or medium or low). These eight machine learning models are logistic regression, support vector machine (RBF), support vector machine (linear), decision tree, Ensemble (hard voting), bagging, random forest and Adaboost. Multiple evaluation matrix (accuracy, precision, recall, f1 score and RMSE) were also calculated for all eight different models.

Among eight different models, SVM (RBF) classifier gave the best results.

Our current results were obtained based on a very small dataset. The accuracy of classification may vary on a much larger dataset. One of the primary challenges was the creation of dataset as the whole process of creating the dataset is quite a labour intensive in nature. In future, we would like to generate new data to develop a much larger dataset and try to deploy a method for optimizing the weights of our predictor variables. We also would like to incorporate new features regarding the characteristics of the publication outlet and characteristics of coauthors of the paper cited.

Our current study shows the possibility of automatic detection of the level of influence imparted by a research paper on a paper which has cited it. It also shows specific features are important for the successful identification of the level of influence.

**Problem Statement**

Our research focused on the following questions:

RQ1: What is feasible and required for the development of a system which will automatically detect the level of influence of a research paper referred (cited) by another research paper?

RQ2: What are the features based on which such an automatic level of influence identification system can be developed?

**Introduction**

The use of different citation matrixes (h-index, number of current citations of paper, etc.) and impact factor-based indicators (Bordons et al 2002) for different policy purposes has increased over the last two decades (Bordons et al 2002). These citation matrixes are often used by various entities like academic publishers, public universities, national science foundation (NSF), funding council, etc. for various purposes. These are often used to provide tangible evidence of benefit to weigh against the costs of research by public universities. Often, they are used for undertaking strategic decision-making, benchmarking as well as the allocation of limited resources. They also provide different entities with the required tools for comparing peer research programs across disciplines. For an independent researcher, these indicators can help him to monitor and manage his performance and also demonstrate the impact of his research work to the government, stakeholders and funding bodies.

Though very powerful these indicator needs to be deployed carefully before making any decision as all these indicators come with their significant inherent limitations. Metrics in themselves can’t convey the full impact, however, they are often viewed as powerful and unequivocal forms of evidence (Millar and Kelly 2013). Although metrics can provide evidence of quantitative changes or impact from research, they are unable to adequately provide evidence of the qualitative impacts that take place and hence are not suitable for all the impact we will encounter. The full influence will not be realized if we only focus on easily quantifiable indicators (like h-index, i10 index, etc.).

These inherent limitations are often related to the assumptions that their proponents make before developing the actual matrix or indicator. One of such assumptions is that all the cited papers have equal influence and often counted as 1 for each of them. This particular assumption paints an incomplete picture of the influence because it does not consider important cues regarding the type of knowledge utilization and level of influence of a research paper (Holsapple, 2008; Serenko and Bontis, 2013). In our current study, we have changed this particular assumption and assume that all the research papers cited impart a different level of influence on the paper. In other words, all the papers listed under the reference section of a paper have a different level of influence on the current paper. Current bibliometric or citation-based methods do not provide any scope of differentiating the different level of influence. We have introduced a higher granularity of knowledge utilization by enumerating the purpose of citation.

Our 40 features are humanly engineered and developed on the basis of contemporary literature available in the domain of Scientometrics. While a significant number of these 40 features represent the author’s specific purpose for citing the paper there are features which specify different sections in which the paper was cited. Yet, there are a set of features which deals with the reputation of the coauthors and the reputation of the paper.

We have identified the purpose of the author for citing each of the paper listed in the reference section. Apart from the author’s purpose, we also have identified in which section the paper was cited, the orientation of the citation (positive or negative), the specificity of the citation, etc. Finally, they have assigned each of the instances of citation to a high, medium or low level of influence.

**Method and Materials**

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1** | | | |
| **Number** | **Attribute/Feature** | **Description** | **Scale** |
| 1 | A1. Support for Argument (General) | The author has cited the paper to derive support for making an argument or a chain of arguments | Binary (1/0) |
| 2 | A2. Support for Model Development and Empirical Framework | The author has cited the paper to derive support for developing the model that he proposes and/or to create the empirical framework | Binary (1/0) |
| 3 | A3. Support for Individual Hypothesis Development | The author has cited the paper to derive support for creating hypothesis. | Binary (1/0) |
| 4 | A4. Support for Theoretical Explanation | The author has cited the paper to derive support for explaining results or to provide theoretical explanation for certain outcomes | Binary (1/0) |
| 5 | A5. Support for the Methodology Used | The author has cited the paper to derive support for following a particular methodology (both qualitative and quantitative: statistical, analytical, empirical, etc.) | Binary (1/ 0) |
| 6 | A6. Support for the Scale and Variables Used | The author has cited the paper to derive support for selecting and using specific measurement scale, dependent variable, independent variable, control variable, etc. | Binary (1/0) |
| 7 | A7. Support for the Example Provided | The author has cited the paper to derive support for providing an example or a metaphor or making an analogy. | Binary (1/0) |
| 8 | A8. Support for the Focus of the Study | The author has cited the paper to derive support for putting forward the focus of the study or revealing the focus of the current study | Binary (1/0) |
| 9 | A9. To derive Support from a Particular Theory/Conceptual Model | The author has cited the paper to derive support from a particular theory or conceptual model | Binary (1/0) |
| 10 | A10. To derive Support for Making Assumption | The author has cited the paper to derive support for making important assumptions | Binary (1/0) |
| 11 | A11. To Provide Crucial Information | The author has cited the paper to describe some important figures or information crucial to the context of the study | Binary (1/0) |
| 12 | A12. To Describe an Important Research Trend | The author has cited the paper to describe a specific trend in the current context of the research | Binary (1/0) |
| 13 | A13. To Describe a Concept | The author has cited the paper to describe a specific construct/principle/rule/idea/concept/topology/terminology/taxonomy | Binary (1/0) |
| 14 | A14. To Describe Other Seminal Work in the Area | The author has cited the paper to suggest some important findings of the cited paper or other studies closest to the current study | Binary (1/0) |
| 15 | A15.To Describe Important Statistical Method | The author has cited the paper to describe a particular statistical method | Binary (1/0) |
| 16 | A16. To Describe Characteristics of a Theory/Conceptual Model | The author has cited the paper to describe or define characteristics of a particular theory or a conceptual model | Binary (1/0) |
| 17 | A17. To Describe the Creation of a Measure/Variable | The author has cited the paper to describe the creation of a measure, scale, variable (IV, DV, Control) | Binary (1/0) |
| 18 | A18. To Describe the Origin of a Model | The author has cited the paper to describe the origin of a model or focus of a model or to describe the body of research that the study followed | Binary (1/0) |
| 19 | A19. To Showcase the Contribution | The author has cited the paper to showcase the contribution of the study (theoretical and practical) | Binary (1/0) |
| 20 | A20. To Showcase the Results | The author has cited the paper to showcase the results obtained by the study | Binary (1/0) |
| 21 | A21. Other Reasons | The author has cited the paper for a reason other than the above 20 reason | Binary (1/0) |
| 22 | Specificity | The citation includes quotes or page number from the cited paper, or the author has specifically mentioned about the purpose of the citation | Binary (1/0) |
| 23 | Valence/Orientation | The author has cited the paper | Binary (1/0) |
| 24 | Cited in the Introduction Section | The author has cited the reference paper in the introduction section of the paper | Binary (1/0) |
| 25 | Cited in the Literature Review Section | The author has cited the reference paper in the literature review section of the paper | Binary (1/0) |
| 26 | Cited in the Theoretical Development Section | The author has cited the reference paper in the theoretical development section review section of the paper | Binary (1/0) |
| 27 | Cited in the Methodology/Method/ Empirical Model Development Section | The author has cited the reference paper in the methodology or method or the empirical model development section of the paper | Binary (1/0) |
| 28 | Cited in the Result Section | The author has cited the reference paper in the result section of the paper | Binary (1/0) |
| 29 | Cited in the Discussion Section | The author has cited the reference paper in the discussion section of the paper | Binary (1/0) |
| 30 | Cited in the Implication or Contribution Section | The author has cited the reference paper in the implication or contribution section of the paper | Binary (1/0) |
| 31 | Cited in the Limitation and/or Future Work section | The author has cited the reference paper in the limitation and /or the future work section of the paper | Binary (1/0) |
| 32 | Cited in the Conclusion Section | The author has cited the reference paper in the conclusion section of the paper | Binary (1/0) |
| 33 | Cited Together with Other Papers | The reference paper was cited along with other reference papers and part of a cluster of citation | Binary (1/0) |
| 34 | Cited in Multiple Sections | The reference paper was cited in multiple sections of the paper | Binary (1/0) |
| 35 | Current Citation Count of the Reference Paper | The current number of citations the reference paper has | Continuous |
| 36 | The Reference Paper Published in the Same Journal or Conference | The reference paper was published in the same journal | Binary (1/0) |
| 37 | Number of Coauthors Present in the Reference Paper | The total number of coauthors the reference paper has | Continuous |
| 38 | Average h-Index of all the Coauthors of the Reference Paper | The average of h-index of all the coauthors of the reference paper | Continuous |
| 39 | Published in a Journal Which is Under Bucket of Eight AIS Journals | The reference paper was published in one of the journals which is currently under AIS Journal bucket | Binary (1/0) |
| 40 | Number of Years Passed After the Publication of the Reference Paper | The number of years passed after the first publication of the reference paper | Continuous |

The LASSO is one of the popular methods for variable selection and shrinkage estimation which is introduced by Tibshirani (1996). LASSO assigns zero weights to most irrelevant or redundant features (Li et al., 2006). For this purpose, a regularization term is added to the cost function to keep the model weight as small as possible. Therefore, LASSO automatically performs feature selection and output of a model with few nonzero feature weights (G´eron, 2017). In this project, we apply LASSO to find the most relevant features in our dataset and run our prediction models based on LASSO results.

**Results**

We have taken 10 papers to analyze what factors influence or doesn’t influence a particular paper. Our Y variable is rank of the paper that is classified into high, medium and low that has been encoded as 3,2 and 1 respectively. We have around 761 references from all the 10 papers and 1129 instances that were used to run the model.

Lasso regularization method has been used to do the feature selection to weed out the unimportant features. We randomly assigned alpha (Lasso) values from which the model selected it be 0.01 and the coefficients of less important features were found to be 0. Taking 0.01 as alpha, we reran the model by dropping the less important features.

Features which are not considered for the final model building are presented in Table 2

|  |  |
| --- | --- |
| Cited in the Discussion Section | Binary (1/0) |
| Cited Together with Other Papers | Binary (1/0) |
| The Reference Paper Published in the Same Journal or Conference | Binary (1/0) |
| Number of Coauthors Present in the Reference Paper | Continuous |
| Number of Years Passed After the Publication of the Reference Paper | Continuous |
| Support for the Methodology Used | Binary (1/0) |
| To Describe Characteristics of a Theory/Conceptual Model | Binary (1/0) |
| To Describe the Creation of a Measure/Variable | Binary (1/0) |
| To Showcase the Results | Binary (1/0) |

We got rid of above features and reran the model using several machine learning algorithms and the results are as follows:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 3** | | | | | | | |
|  | **Accuracy** | **Precision** | **Recall** | **F1 Score** | **Support** | **RMSE** | **Hyperparameter Values** |
| Logistic Regression | 65.48% | 69% | 73% | 70% | 86% | 0.584 | N/A |
| SVM-RBF | 74% | 74% | 78% | 76% | 86% | 0.469 | C=4 |
| Decision Tree | 60.61% | 72% | 78% | 75% | 90% | 0.646 | Max\_Depth=5 |
| Ensemble (Voting) | 61.94% | 69% | 72% | 71% | 90% | 0.566 | N\_estimator=6 |
| Bagging | 65.04% | 64% | 78% | 70% | 86% | 0.654 | Max\_leaf\_nodes=16 |
| Random Forest | 65.04% | 64% | 78% | 70% | 86% | 0.654 | N-Estimator=3 |
| Adaboost | 58.40% | 70% | 72% | 71% | 86% | 0.575 | Learning\_Rate=0.3 |
| SVM-Linear | 66.8% | 68% | 74% | 71% | 86% | 0.61 | C=0.5 |

Comparing the above algorithms, we found SVM-RBF to be the best model with F1 score to be around 76% with the least RMSE value 46.9%.

Confusion matrix of the model (SVM-RBF) is provided below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4** | | | |
|  | **Low** | **Medium** | **High** |
| **Low** | 52 | 10 | 6 |
| **Medium** | 6 | 49 | 17 |
| **High** | 10 | 9 | 67 |

The confusion matrix suggests comparatively higher performance for predicting the higher level of influence which for a medium level of influence the performance is not that accurate. We believe the lack of data point is currently responsible for such a result.

**Limitations**

Our study has some limitations. One such limitation is the amount of data taken to do the analysis. While this may be a concern for the results, we tried to compensate this by increasing the number of features up to 40.

The second limitation of this study may be that the results are more along the lines of predictive analysis of the dataset. Since the dataset was less, it was difficult to go beyond this analysis.

The third limitation one might say is one could have automated some part of data collection like collecting keywords and the citations which resulted in a bit more of time consumption that eventually led to problems mentioned above.

Finally, considering the gold standard (human prediction accuracy: 88-90%) our best model still underperforming. In other words, our current models are grossly underfitting and we need to consider more relevant features to improve model prediction accuracy.

**Future Work**

We are planning to do more research in the future of this project. We are planning to add more data that will bolster our results. We will be adding more important features required as we go along with the project. We are also planning to undertake unsupervised methodology and do a bit of text mining with our dataset. Optimization of weights according to their ranks is also an idea as of now. Finally, we will try to publish our research studies.

**Reference**

1. Holsapple, C. W.: “A publication power approach for identifying premier information systems journals.” Journal of the American Society for Information Science and Technology. vol. 59, 2, pp.166--185 (2008).
2. Serenko, A., Bontis, N.: “The intellectual core and impact of the knowledge management academic discipline.” Journal of Knowledge Management. vol. 17, 1, pp.137--155 (2013).
3. Bordons, Maria, M. Fernández, and Isabel Gómez. "Advantages and limitations in the use of impact factor measures for the assessment of research performance." Scientometrics 53.2 (2002): 195-206
4. Tibsharani, R.: “Regression shrinkage and selection via the Lasso.” Journal of the Royal Statistical Society, 1996.
5. Millar, Ross, and Kelly Hall. "Social return on investment (SROI) and performance measurement: The opportunities and barriers for social enterprises in health and social care." Public Management Review 15.6 (2013): 923-941