Journal Decision Using Machine Learning Techniques

Shuangping Liu
Northwestern University
shuangping-liu@u.northwestern.edu

Li Zeng
Northwestern University
lizeng2016@u.northwestern.edu

Yu Zhou
Northwestern University
yuzhou2016@u.northwestern.edu

I. Introduction

One of the toughest choices that many researchers have to face during their academic career is to find an appropriate journal or conference to publish their works. Besides the impact of the journals, many other factors are also seriously considered by researchers before submitting their manuscripts, such as the correlation between their research fields and that of the journal, the researcher?s publication history on this journal, the popularity of the research topics in this journal, etc. It would make this choice much easier if the rationality of submitting to a certain journal could be automatically estimated and predicted by machine based on the former publications in this journal. More importantly, this system can also help reduce the workload of journal editors if a paper can be automatically graded and recommended beforehand.

Herein we propose to build up a model to characterize the rationality to make decision to selecting jounrnal. In this final project, we first established that the machine can distinguish research subjects between materials science and chemisty. The original dataset will be acquired by directly crawling the information of the papers from the American Chemistry Society (ACS) for Journal of Chemistry Society (JACS) and Wiley Publication Group for Advanced Materials (A&M) website in the past 10 years (2005 2015). These information such as titles, authors, contents of abstract, affiation, publication year etc, will be treated as input attributes for machine to learn and test. The idea of this project will be similar to typical bag-of-words approach, we will process the title and abstract to study the importance of representative word with respect to the journals. The machine learning methods that we are going to employ are Naive Bayes, Support vector machines, and decision trees.

II. Data

In the project we will work on two research areas: chemistry and material science. Chemistry is a a very foundamental research subject as physics or mathematics, while material science is a newly research area merged since roughly 50 years ago. Chemistry focuses on synthesis of new molecular compounds or explain new phenomena at molecular scales but material science emphasizes more on applications and enginneeing at interfaces of novel nano-scale materials, such as thin film devices, nano-dots. These two areas certainly have mutually interested topics but they study different aspects of these topics. Hence, this makses this project very interesting and useful to see whether the machine can learn the features of these two closely related but different research subjects.

TABLE I
The sample of the dataset

Data	Content
info	Article first published online: 23 DEC 2014 — DOI: 10.1002/adma.201404849
doi	10.1002/adma.201404849
author	Peter Tseng, Jonathan Lin, Keegan Owsley, Janay Kong, Anja Kunze, Coleman Murray and Dino Di Carlo
title	Flexible and Stretchable Micromagnet Arrays for Tunable Biointerfacing (pages 1083?1089)
year	2015
type	Communications
abstract	A process to surface pattern polydimethylsiloxane (PDMS) with ferromagnetic structures of varying sizes (micrometer to millimeter) and thicknesses (¿70 ?m) is developed. Their flexibility and magnetic reach are utilized to confer dynamic, additive properties to a variety of substrates, such as coverslips and Eppendorf tubes. It is found that these substrates can generate additional modes of magnetic droplet manipulation, and can tunably steer magnetic-cell organization.
affiliation	Department of Bioengineering, University of California, Los Angeles, Los Angeles, California, USA

We particularly chose Journal of Chemistry Society (JACS) and Advanced Materials(A&M) because both are top and prestigious journals with high impact. Therefore, the articles from these two journals will be the most indecative ones to show the research trends of the past decades. An example were chosen from A&M and shown in Table I.

The journal information usually comes with several indexes and information for reader to quickly locate and learn the topics of a publication. The most informative attributes are tilte and abstract. Therefore, by counting the frequency of the ”key words” used in titles and abstracts, we can teach machine to learn what are the most related topics to either chemisty and materials science. Angewandate Chemie is also one of the top journals in chemstry area, and one of the competing journals to JACS. We pulled the data from its website as well. And they will be used as test samples for machine to make a prediction.

Out datasets include 32268 papers in JACS, 9008 papers in A&M and 21887 papers in Angewandte Chemie ranging from

TABLE II
THE ACCURACY OF THE THREE MACHINE LEARNING METHODS WITH ONLY TITLES INCLUDED IN DATASETS

Methods	Training Set	Test Set	AC Test
Naive Bayes	0.9666	0.9484	0.8479
Linear SVC	0.9011	0.9004	0.7806
Decision Tree	0.9967	0.9964	0.1010

TABLE III
THE ACCURACY OF THE THREE MACHINE LEARNING METHODS WITH TITLES AND ABSTRACTS INCLUDED IN DATASETS

Methods	Training Set	Test Set	AC Test
Naive Bayes	0.8955	0.8838	0.9430
Linear SVC	0.9589	0.9529	0.8718
Decision Tree	0.9966	0.9962	0.1007

2005 to 2015.

III. EXPERIMENTS

The python machine learning package scikit-learn was employed to classify the documents with a bag-of-words approach. We started with estimating how frequently the words show in the titles and abstracts of a certain journal. Term frequency?inverse document frequency (TF-IDF) was therefore used to measure the importance of the words by considering the appearance of a certain word in both the given papers and the general corpus. In bag-of-wods approach, each title and abstract are vectorized to a sparse vector filled with the values given by TF-IDF transform as an instance of training set. The target value of each instance is the journal type (e.g., JACS corresponds to 0 and A&M is 1 in our case).

80% of our data were randomly chosen as the training set and the rest 20% were used as test set. 10 fold cross validation was performed for the training set to check how well the machine learning model can be generalized to independent dataset. The papers in the journal Angewandte Chemie were used as a independent dataset to test our model, since it is as prestigious as JACS in the chemistry field. Herein we use Naive Bayes, Supporting Vector Machine and Decision Tree as the machine learning techniques in order to obtain the best results.

IV. RESULTS

We apply the three machine learning methods to two different datasets:

- Including only titles information
- Including both titles and abstracts information

A. Dataset with only titles

The results are shown in Table II.

1) Naive Bayes: Naive Bayes gives the most reasonable results among the three methods. The model derived from the training set not only fit the test set well, but also gives good prediction for the Angewandte Chemie test. Around 85% of the papers of Angewandte Chemie are classified to chemistry.

2) SVM: Specifically we are using SVM with linear kernel and minimization of the squared hinge loss (LinearSVC in scikit-learn). LinearSVC is very sensitive to the regularization parameter C. Large C will result in overfitting of the training set and very poor accuracy for Angewandte Chemie test. In our experiment, we use C = 0.004. Around 78% of the papers in Angewandte Chemie are identified as acceptable to JACS.

TABLE IV
THE ELAPSED LEARNING TIME OF THE THREE MACHINE LEARNING METHODS WITH TITLES AND ABSTRACTS INCLUDED IN DATASETS

Methods	Time (Second)
Naive Bayes	0.036
Linear SVC	0.197
Decision Tree	

3) Decision Tree: Decision Tree: We attempt to avoid overfitting problem by limiting the maximum depth of the decision tree. However, it does not improve much of our results: Decision Tree can fit perfectly with training set and test set, while it only put 10% of the Angwandte Chemie papers to the correct class.

B. Dataset with titles and abstracts

The results are shown in Table III.

With both titles and abstracts included in the dataset, the prediction accuracies for the Angewandte Chemie set of NB and LinearSVC are improved (NB: 94%, LinearSVC: 87%). For NB, the accuracy of the training set and test set is lower than before, since more information is taken into consideration for the classification. It is likely that some words in the abstract can be both included in JACS and A&M. For LinearSVC, one can adjust the parameter C to obtain better accuracy for all the datasets. For DT, the accuracy of all the datasets are almost the same as before.

C. Performance

The performance of the three machine learning methods are shown in Table IV. Apparently, Naive Bayes uses the least amount of time while giving relatively good results compared with other methods.

V. CONCLUSION

In conclusion, NB and Linear SVC perform well in predicting the appropriate journal to submit, while DT is not suitable for this type of problem.

The future work of this project could be including more attributes. For example, Keywords can be included and treated with similar methods. Year is another important attribute to study the trending of each research area, and the correlations between two journals. More practical test datasets can also be introduced to verify our machine learning model. Additionally, we can certainly include more dataset by extracting data since a few decades ago, and expending the research subjects to physics or biology, and some sub-areas such as organic chemistry vs. inorganic chemistry.

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.