Scientometrics (2016) 109:615–628
DOI 10.1007/s11192-016-2050-y

Classifying modeling and simulation as a scientific discipline

Ross Gore[1] · **Saikou Diallo**[1] · **Jose Padilla**[1]

Received: 23 March 2015 / Published online: 9 July 2016

Abstract The body of knowledge related to modeling and simulation (M&S) comes from a variety of constituents: (1) practitioners and users, (2) tool developers and (3) theorists and methodologists. Previous work has shown that categorizing M&S as a concentration in an existing, broader disciple is inadequate because it does not provide a uniform basis for research and education across all institutions. This article presents an approach for the classification of M&S as a scientific discipline and a framework for ensuing analysis. The novelty of the approach lies in its application of machine learning classification to documents containing unstructured text (e.g. publications, funding solicitations) from a variety of established and emerging disciplines related to modeling and simulation. We demonstrate that machine learning classification models can be trained to accurately separate M&S from related disciplines using the abstracts of well-index research publication repositories. We evaluate the accuracy of our trained classifiers using cross-fold validation. Then, we demonstrate that our trained classifiers can effectively identify a set of previously unseen M&S funding solicitations and grant proposals. Finally, we use our approach to uncover new funding trends in M&S and support a uniform basis for education and research.

Keywords Simulation · Research · History of OR · Machine learning

✉ Ross Gore
rgore@odu.edu

Saikou Diallo
sdiallo@odu.edu

Jose Padilla
jpadilla@odu.edu

[1] Virginia Modeling, Analysis and Simulation Center, Old Dominion University, Norfolk, VA, USA

Introduction

There has been significant work in identifying scientific disciplines and the changes within them (Vinkler 1988; Mayr 2004; Glenisson et al. 2005; Herrera et al. 2010; Bourke and Butler 1998; Vessey et al. 2005; Katz and Hicks 1995; Börner et al. 2012; Wallace et al. 2012; Searls 2010; Kaur et al. 2012; Ioannidis 2006). While these efforts have been fruitful, they have generally focused on: (1) structured data and (2) the relationships among research publication authorship and citations.

Furthermore, the study of emerging disciplines, such as modeling and simulation (M&S), have not focused on identification. Instead, researchers have focused on analyzing the content of the emerging discipline to elucidate the contributions of concepts, relationships and information in established disciplines (Hinze 1994). In addition, there are specific difficulties associated with identifying content that reflects M&S as a discipline versus content that uses modeling and/or simulation as a methodological approach (i.e. engineering, physics, astronomy, atmospheric sciences).

Here, we refer to an *emerging discipline* as one which integrates some subset of the following from two or more bodies of knowledge: (1) perspectives, concepts and theories, (2) tools and techniques and (3) information and data (Salter and Hearn 1997; Aboelela et al. 2007). In contrast, an *established discipline* is an area of specialization that focuses on one narrow topic (Salter and Hearn 1997; Aboelela et al. 2007). For example, computational biology is not an established discipline because it is not a specialization of a single topic. Instead, it is an emerging discipline because it combines concepts of biology and computer science but is separate and identifiable from both.

We address all of these issues with a flexible framework to classify M&S from other emerging and established disciplines. The framework leverages machine learning to construct classification models. The classification models are trained on research publications to identify M&S as well as ten other scientific disciplines. Once trained, the classification models are deployed on new data sets to classify content in each discipline. The performance of the models is evaluated using established measures to demonstrate their effectiveness. Finally, newly identified M&S content is analyzed to gather new insight into the funding of the discipline by the National Science Foundation (NSF) and National Institute of Health (NIH).

A methodology to classify and analyze M&S is needed. Previous work has shown that categorizing M&S as a concentration in an existing, broader disciple is inadequate. This labeling does not provide a uniform basis for research and education across all educational institutions (Balci 2001; Sarjoughian and Zeigler 2001; Crookall 2010). The analysis based on the M&S content our approach classifies supports the development of such a basis.

Ultimately, our work has four main contributions:

1. We present a formal definition of the *discipline classification problem* so that the effectiveness of our framework can be evaluated. This problem definition is adapted from problem of text classification in the field of machine learning (Sebastiani 2002).
2. We demonstrate how to use machine learning techniques to build and train a classifier that can classify documents related to M&S from other established disciplines (i.e. medicine) as well as emerging ones (i.e. computational biology).
3. We apply our trained classifiers to data sets of funding solicitations and grants from NSF and NIH. We show that our algorithm can effectively identify the M&S funding solicitations and grant abstracts.

4. We demonstrate the utility of our framework by applying it to gather new insight into the funding of M&S by the National Science Foundation (NSF) and National Institute of Health (NIH).

Data and problem definition

A number of different data sets are used in our study. First, we describe each data set. Then we formally define the discipline classification problem.

Data

ACM data set

The ACM data set is obtained by collecting the abstracts from conference proceedings and print periodicals from 1960 to 2011 (White 2001). The publications in the ACM data set span several disciplines: modeling and simulation, computer science, computer engineering, electrical engineering and systems engineering.

The discipline of a particular publication in the ACM data set is determined by the title of the conference or periodical in which it appears. For example, a publication in the *Winter Simulation Conference* pertains to the discipline of modeling and simulation. Similarly, a publication in the *Transactions on Programming Languages* pertains to the discipline computer science. In most cases the mapping between the periodical and discipline is straightforward. However, in some cases it is possible that reasonable people could disagree. In ambiguous cases we chose to classify periodicals and their publications as contributors to the discipline of computer science. We made this choice because computer science is seen as the overarching discipline of the ACM.

The ACM data set consists of 213,725 abstracts which are mapped to the disciplines as follows: computer science (160,293), modeling and simulation (12,823), computer engineering (4247), electrical engineering (14,961) and systems engineering (19,235).

PLoS data set

The PLoS data set is formed by collecting the all abstracts of the six domain specific Public Library of Science periodicals (Jahn et al. 2013). The domain specific PLoS periodicals are: *PLoS Biology*, *PLoS Medicine*, *PLoS Computational Biology*, *PLoS Genetics*, *PLoS Pathogens*, *PLoS Neglected Tropical Diseases*.

The inception date of these journals varies from 2003 to 2005, but each has continued to be published through 2013. The six disciplines covered in these periodicals match their respective titles: biology, medicine, computational biology, genetics, pathogens and disease. The disciple of a particular publication is determined by the journal in which it is published.

The PLoS data set consists of 13,095 abstracts which are mapped to the disciplines as follows: biology (1334), medicine (832), computational biology (2976), genetics (3041), pathogens (3348) and disease (1564).

National science foundation data set

The NSF data set is formed by collecting the abstracts from solicitations and funded grants during 1990–2003 from the following divisions: (1) Atmospheric Sciences, (2) Computing and Communication Foundations, (3) Civil and Mechanical Systems, (4) Chemical and Transport Systems, (5) Design and Manufacturing Innovation, (6) Materials Research, (7) Mathematical Sciences and (8) Undergraduate Education (Pazzani and Meyers 2003). The discipline these abstracts are classified by is discussed in the Methodology Section. The NSF data set consists of 961 abstracts.

National Institute of Health data set

The NIH data set is formed by collecting the abstracts from solicitations and funded grants during 1990–2012 from the following divisional institutions: (1) National Heart, Lung and Blood Institute, (2) National Library of Medicine, (3) National Institute of Allergy & Infectious Diseases, (4) National Institute of Diabetes, Digestive & Kidney Diseases, (5) National Eye Institute, (6) National Institute of General Medical Sciences, (7) National Cancer Institute, (8) National Institute of Dental & Craniofacial Research (NIH 2003). The discipline these abstracts are classified by is discussed in the Methodology Section. The NIH data set consists of 6837 abstracts.

Discipline classification problem definition

Discipline categorization focuses on assigning documents to predefined disciplines. It is an application of text classification—the study of classifying any text-based document into a set of predefined categories.

Applying text classification to solve categorization problems is now new. Text classification techniques have been applied to: (1) spam filtering to discern e-mail spam messages from legitimate emails (Wang and PAN 2005), (2) email routing which sends an email sent to a general address to a specific address based on the topic (Argamon et al. 1998), (3) language identification to automatically determining the language of a text (Rajman and Besançon 1998), (4) story genre classification to automatically determining the genre of a text (Lin et al. 2009; Xiao et al. 2009), (5) readability assessment to quantify the degree of readability of a text based on age groups or reader type (Miltsakaki and Troutt 2008), (6) determining the sentiment of a speaker or a writer with respect to a given topic (Liu and Zhang 2012), (7) article triage which selects articles that are relevant to a specified topic or annotation (Wei et al. 2012).

Furthermore, there have been several studies related to the automated categorization of bibliometric data. Using graphical navigation tools researchers have created bibliometric maps of science to answer policy-related question (Noyons 2001). In addition mapping and citation analysis have been combined to evaluate the research scope and performance of the academic institutions and research centers (Nederhof and Noyons 1992; Noyons et al. 1999). The success of these efforts have motivated the development of bibliometric standards to improve the reliability of bibliometric results and guarantee the validity of bibliometric methods (Glänzel 1996). Standards have facilitated the analysis of journal impact measures (Glänzel and Moed 2002) and scientific collaboration networks (Glänzel and Schubert 2005).

Here, we apply text classification to categorize any document within a data set into a set of predefined disciplines. Each document is composed of unstructured text and a data set is composed of many documents. In the discipline classification problem, a classifier, *CLASS* must assign a Boolean value to each pair $\langle d_j, c_i \rangle \in D \times C$, where D is the data set of documents and $C = \{c_1, \ldots, c_{|C|}\}$ is a set of predefined disciplines. A value of *true* assigned to $\langle d_j, c_i \rangle$ indicates that document d_j is part of discipline c_i, while a value of *false* indicates it is not. Membership within a discipline is mutually exclusive. This means that a document, d_j can only belong to one discipline, c_i, at most. The set of disciplines C are symbolic labels, no additional knowledge of their meaning is given during classification. Similarly, only data within the documents can be used during classification.

Methodology

Given the definition of the data sets and the discipline classification problem, we propose a solution centered around machine learning. Machine learning is a general inductive process that trains classification models by learning, from a set of pre-classified documents, the characteristics of a particular discipline. From these characteristics, the inductive process gleans the characteristics that a new unseen document should have in order to be classified into a particular discipline (Sebastiani 2002; Alpaydin 2004).

Our study consists of three phases: *training* the classifiers, *classifying* new data sets and *evaluating* effectiveness. Once these are complete we employ our framework to provide insight into M&S. In the remainder of this section we review each phase of our methodology.

Training

Pre-classified documents are a key resource in machine learning (Sebastiani 2002; Alpaydin 2004). Our framework leverages pre-classified documents in large, well-indexed repositories of research publications: The Association of Computing Machinery (ACM) and Public Library of Science (PLoS). For each abstract in the ACM and PLoS data sets we use a negative dictionary to filter out words that obfuscate the identity of a discipline. We employ the negative dictionary developed by Fox et al. for general texts because it has been shown to be maximally efficient and effective in filtering semantically neutral words in the English language (Fox 1989; Yu 2008). Next, we describe the machine learning algorithms we apply to the filtered abstracts to train the discipline classification models.

Machine learning algorithms

We employ two different algorithms for classification: Naive Bayes (NB) and Stochastic Gradient Descent (SGD) (Lewis 1998; Jordan 2002). While these algorithms do not span the spectrum of machine learning algorithms they represent two different approaches to classification—one based on bayesian statistics (NB) and the other based on logistic regression (SGD). Here we summarize both algorithms and discuss the parameterizations we use to train our classification models for each of the eleven disciplines in the ACM and PLoS data sets.

The Naive Bayes approach to classification is based on applying Bayes' theorem with independence assumptions. Given a particular classification, it assumes that the presence or

absence of a particular feature is unrelated to the presence or absence of any other feature. For example, consider the description of an apple as a red, round, fruit that is 3″ in diameter. Naive Bayes approach to classification considers each of these four features (red, round, fruit, 3″ diameter) to contribute independently to the probability that a given object is an apple, regardless of the presence or absence of the other features (Lewis 1998; McCallum et al. 1998; Eyheramendy et al. 2003; Kim et al. 2006).

Stochastic Gradient Descent (SGD) applies a different approach to classification. It randomly shuffles the documents in the training set and then iteratively employs a logistic regression cost function to determine how well the current shuffle classifies the documents into categories. The next shuffle is determined by a stochastic function and a cost function that calculates how well (or poorly) the current shuffle separated the documents in comparison to the previous shuffle. This process repeats until further improvement does not seem to be possible. Typically, SGD is most successful when applied to large, text-based

A ACM

True \ Predicted	CS	M&S	SE	CE	EE
CS	**72934**	441	189	3095	3482
M&S	307	**5803**	12	127	164
SE	61	105	**1889**	35	34
CE	230	161	17	**6807**	261
EE	351	68	13	582	**8858**

B PLOS

True \ Predicted	Bio.	Med.	Comp Bio.	Gen.	Path.	Dis.
Bio.	**618**	4	33	5	4	5
Comp Bio.	6	**388**	4	7	4	5
Gen.	73	9	**1371**	12	16	8
Path.	11	23	14	**1414**	31	38
Dis.	12	24	29	31	**1543**	34
Med.	17	11	3	18	16	**717**

Fig. 1 Cross validation results for discipline classification using Stochastic Gradient Descent (SGD's) on ACM (**a**) and PLoS (**b**) datasets

document repositories with a moderate number of ($\sim$20) predefined categories (Zhang 2004; Bottou 2010; Baird and Moore 1999).

Accuracy

We applied the NB and SGD algorithms to both the ACM and PLoS data sets and evaluated the algorithms' accuracy for each data set using a ten-fold cross validation. This process consisted of 10 rounds where the documents in each data set were partitioned into 10 equal portions. In each round 9 of the portions of the abstracts (90 %) were used to train the models, while the remaining portion of abstracts (10 %) were used for testing the accuracy of the classification models (Efron and Gong 1983; Kohavi et al. 1995).

The tenfold cross validation results are shown in Figs. 1 and 2. The SGD classifier was able to achieve an accuracy greater than 90 % in each data set and the NB classifier was able to achieve an accuracy greater than 80 %. Given the nature of emerging disciplines like M&S, the ambiguity of natural language and the similar domains encompassed in each data set the performance of both classifiers, especially SGD, is significant. Furthermore, our accuracy results show that the approach is not dependent on a single machine learning classification technique. SGD classification may yield superior accuracy but a bayesian algorithm still performs well. Given the superior accuracy of the SGD classifier we employ it in the remainder of this paper. However, our approach is capable of supporting any existing machine learning algorithm for classification.

Classification

Next, we propose an algorithm which applies our trained classification models to identify M&S funding solicitations and accepted grant proposal abstracts from NSF and NIH. This data is truly unstructured. It contains multiple types of documents from different time periods. Given the lack of structure in the data sets it is not obvious how existing discipline identification techniques would be able to be applied to identify separate corpora.

Our algorithm applies the trained classification models to each of the document (i.e. funding solicitation and/or accepted grant proposal) in the data sets. For each document, we use a negative dictionary to filter out words that obfuscate the identity of a discipline. Then, each of the eleven trained models return a probability measure p, which reflects the probability the document is part of the discipline. The maximum probability measure of the eleven trained models is p_{max}. If p_{max} is above the probability threshold for the discipline t then the algorithm categorizes the document as a part of the discipline. Otherwise it moves it into the *OTHER* category. The *OTHER* reflects documents that are not part of any

Fig. 2 Classification accuracy for discipline classification using tenfold cross validation

Data Set \ Classifier	SGD	Naive Bayes
ACM	**90.8**	82.6
PLOS	**92.4**	86.1

of the eleven disciplines. The tendency of the algorithm to classify documents in the *OTHER* category is dependent on the probability threshold t for each discipline. Next, we evaluate how effective the algorithm is in classifying the NSF and NIH abstracts using different values for t.

Evaluation

Our evaluation tests the effectiveness of our trained classifiers on 1000 randomly selected abstracts that were manually classified into disciplines by research groups at two different institutions. The other researchers were not restricted to classifying the abstracts into the eleven disciplines previously discussed. For any abstract, that did not fit the eleven disciplines they could classify the document as *OTHER*.

In general the two research groups agreed on the classification of the abstracts. The interrater reliability is $\kappa = 0.704$ with $p < 0.001$. This reflects a level of agreement most experts consider substantive and is a statistically significant result (Landis and Koch 1977). For those abstracts where the two research groups expressed disagreement we examined the abstract and chose one of the two classifications chosen by the groups.

We measure effectiveness using two metrics: (1) accuracy and (2) coverage. Accuracy reflects the ratio of: (1) documents the algorithm correctly classified for a given discipline compared to (2) the total number of documents the algorithm classified for a given discipline ($\text{discipline}_{\text{correct}} + \text{discipline}_{\text{incorrect}}$). Accuracy is computed using Equation 1.

$$\text{accuracy} = \frac{\text{discipline}_{\text{correct}}}{\text{discipline}_{\text{correct}} + \text{discipline}_{\text{incorrect}}} \tag{1}$$

Coverage reflects the ratio of documents the algorithm correctly classified for a given discipline compared to the total number of documents cataloged in that discipline by the other researchers ($\text{discipline}_{\text{total}}$). Coverage is computed using Eq. 2.

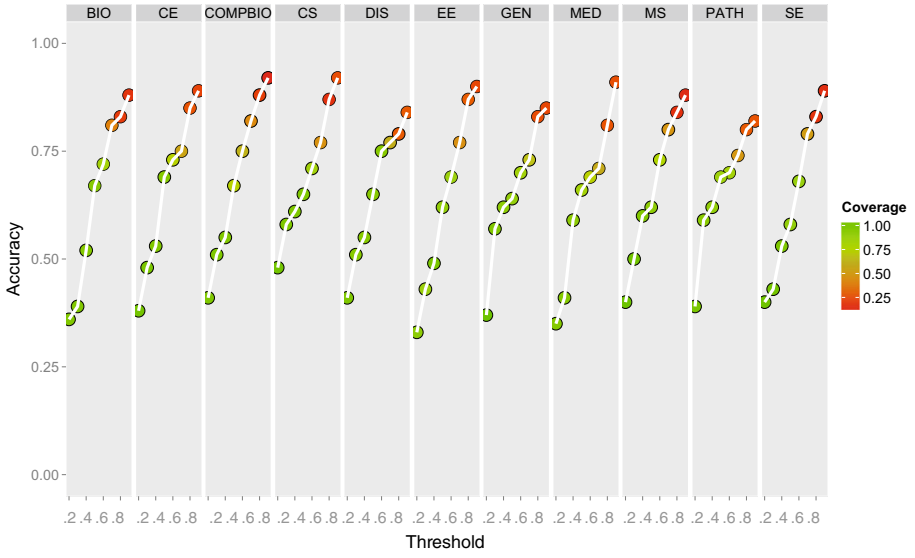

Fig. 3 Accuracy and Coverage for disciplines using different probability thresholds

$$\text{coverage} = \frac{\text{discipline}_{\text{correct}}}{\text{discipline}_{\text{total}}} \quad (2)$$

Given these measures we evaluate the effectiveness of our classification algorithm on the NSF and NIH datasets using different values for the probability threshold parameter *t*. The results of this evaluation are shown in visualized in Fig. 3. In Fig. 3 the x-axis reflects the value of *t*, the y-axis reflects the *accuracy* and the color of each point reflects the *coverage*.

Figure 3 shows a very clear tradeoff between the accuracy and coverage for different probability thresholds. As the probability threshold decreases the accuracy decreases and the coverage increases across all of the 11 disciplines. On the whole this trend is not surprising. One would expect that as the algorithm classifies more abstracts into disciplines with less certainty the algorithm's accuracy will decrease. However, it is unexpected that this pattern is visible and relatively uniform across all eleven disciplines. In particular, the coverage in each discipline begins to drastically improve once *t* reaches 0.70. Furthermore, the accuracy of each model does not significantly decay until *t* falls below 0.60.

Ultimately, the overall performance of the algorithm for the NSF and NIH could be optimized by identifying the value for *t* within each discipline that provided the best balance of accuracy and coverage. However, using a *t* value of 0.65 for all disciplines provides a balance of accuracy and coverage.

Classification algorithm application

Thus far we have reviewed how our approach: (1) develops a trained classifier for M&S using publications from large repositories and (2) effectively classifies previously unseen funding solicitation and grant abstracts from NSF and NIH. Here, we demonstrate its utility

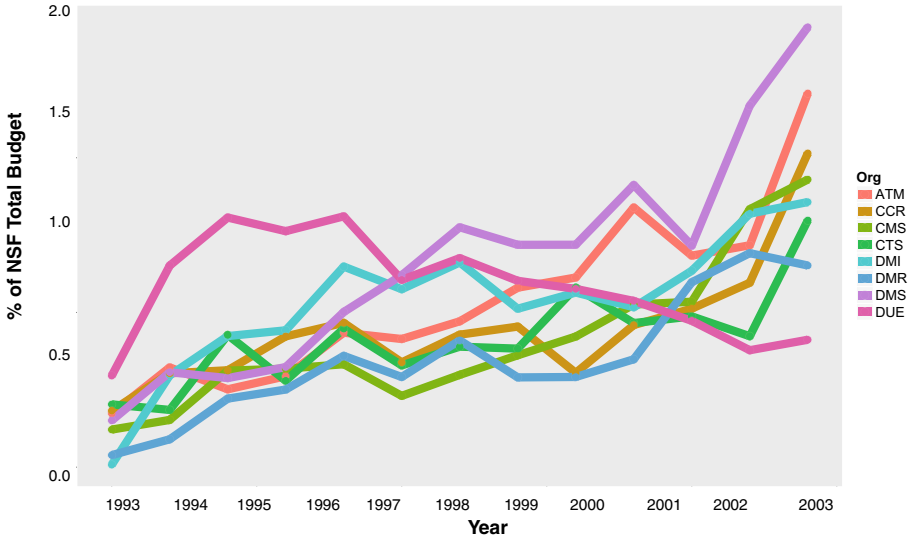

Fig. 4 Exploration of the percentage of NSF's budget devoted to modeling and simulation from 1990 to 2003

by providing examples of the types of analysis about M&S that can be performed on the classified content to support a uniform basis for M&S research and education.

Our example explores the funding of M&S by the NSF from 1990 to 2003 (Pazzani and Meyers 2003) and the NIH from1990 to 2012 (NIH 2003). While this analysis is straightforward, it relies on our approach to effectively classify the modeling and simulation corpus from the NSF and NIH data sets. Recall, the NIH and NSF funding solicitations and grant abstracts do not match the publication-citation structure required by existing discipline classification techniques and there is not a modeling and simulation division within either NSF or NIH. The classification step is the only automated technique we are aware of that is capable of isolating those solicitations and grants related to modeling and simulation. We use the set of 371 abstracts classified as M&S by our approach with a probability threshold value of $t = 0.65$, which yielded and accuracy measure of 0.741 and a coverage measure of 0.729. Unlike our previous evaluation this classification is performed on all the funding solicitation and grants in both data sets ($\sim$7000) as a result all of these abstracts were not manually inspected to verify classification accuracy or coverage.

Figures 4 and 5 show the eight divisions which were the most frequent funders of modeling and simulation within NSF and NIH from 1990 to 2003 and 1990 to 2012 respectively. Tables 1 and 2 provide a key for each acronym in the legends of Figs. 4 and 5. The most notable trend in Fig. 4 is that starting in the year 2000, funding for modeling and simulation within NSF drastically increased. It experienced significant growth from 2000 to 2003 in seven of the eight divisions and the only division where it declined (DUE) had been the leading funder of modeling and simulation from 1990 to 2000. In contrast, it appears that funding for modeling and simulation grew steadily from 1990 to 2012. The number of funded NIH grants related to modeling and simulation doubled approximately every 5 years culminating in almost 5000 funded modeling and simulation grants in 2012.

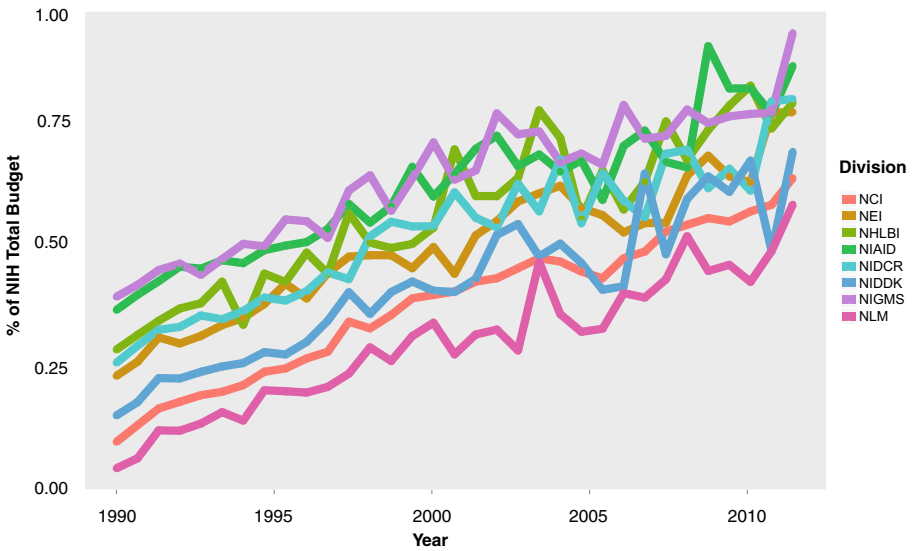

Fig. 5 Exploration of the percentage of NIH's budget devoted to modeling and simulation from 1990 to 2012

Table 1 NSF acronyms

NSF Division acronym	Full name of NSF division
ATM	Atmospheric Sciences
CCR	Computing and Communication Foundations
CMS	Civil and Mechanical Systems
CTS	Chemical and Transport Systems
DMI	Design and Manufacturing Innovation
DMR	Division of Materials Research
DMS	Division of Mathematical Sciences
DUE	Division of Undergraduate Education

Legend of NSF acronyms shown in Fig. 4

Table 2 NIH acronyms

NIH Institute acronym	Full name of NIH Institute
NHLBI	National Heart, Lung and Blood Institute
NLM	National Library of Medicine
NIAID	National Institute of Allergy & Infectious Diseases
NIDDK	National Institute of Diabetes, Digestive & Kidney Diseases
NEI	National Eye Institutue
NIGMS	National Institute of General Medical Sciences
NCI	National Cancer Institute
NIDCR	National Institute of Dental & Craniofacial Research

Legend of NIH acronyms shown in Fig. 5

We also employ our classification algorithms to determine the prevalence of NSF funding for each of eleven identified disciplines. This analysis is shown in Fig. 6. One of the visible trends in Fig. 6 is that more grants related to the established disciplines of computer science, biology and genetics have been funded compared to the less established disciplines of modeling and simulation, systems engineering and computational biology. Further analysis is needed to determine if this is true for all emerging disciplines compared to their established counterparts. However, our approach has elucidated this research question.

Conclusion and future work

The vast majority of research on the classification of established and emerging scientific disciplines has been restricted to research publications and citations. Given the increased availability of unstructured text this restricted model of discipline identification is becoming less applicable. Furthermore, difficulties associated with identifying content that reflects M&S as a discipline versus content that uses modeling and/or simulation as a methodological approach has limited the study of the M&S body of knowledge.

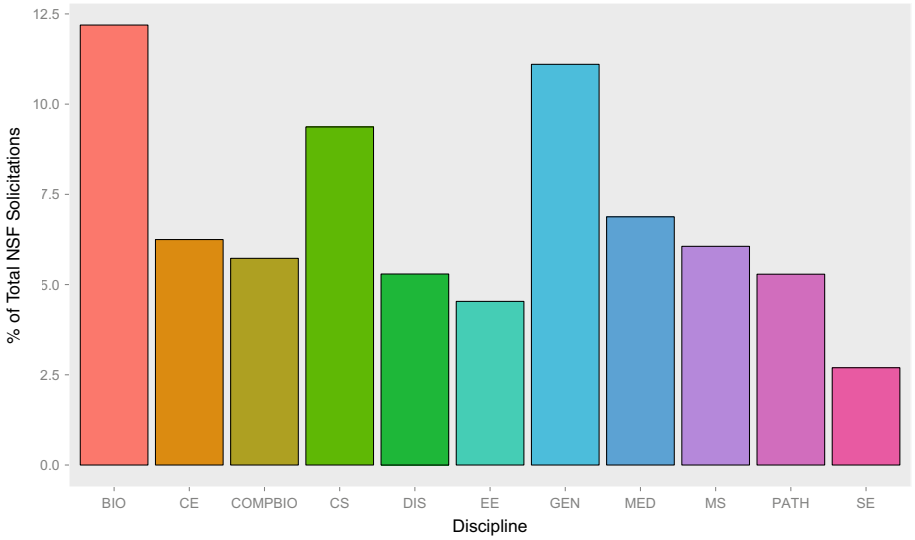

Fig. 6 Exploration of the funding of the different disciplines identified in this paper within NSF from 1990 to 2003

We address these issues with a framework to classify M&S from other emerging and established disciplines. The framework leverages machine learning to construct classification models. The classification models are trained on research publications to identify M&S as well as ten other scientific disciplines. Once trained, the classification models are deployed to identify M&S funding solicitations and grants from NSF and NIH. The performance of the models is evaluated using established measures to demonstrate their effectiveness. Finally, the proposed approach analyzes content representing M&S to uncover new fundings trends in support of a uniform basis for research and education. In future work, we will look to apply our approach for additional types of analyses which build the M&S body of knowledge.

Acknowledgments We gratefully acknowledge the support of our colleagues at the Virginia Modeling, Analysis and Simulation Center (VMASC), University of Virginia (UVA) and Gettysburg College in manually classifying the 1000 NSF and NIH Grants used in the evaluation.

References

Aboelela, S. W., Larson, E., Bakken, S., Carrasquillo, O., Formicola, A., Glied, S. A., et al. (2007). Defining interdisciplinary research: Conclusions from a critical review of the literature. *Health Services Research*, *42*(1p1), 329–346.

Alpaydin, E. (2004). *Introduction to machine learning*. Cambridge: The MIT Press.

Argamon, S., Koppel, M., & Avneri, G. (1998). Routing documents according to style. In *First international workshop on innovative information systems*, pp. 85–92. Citeseer.

Baird, L., & Moore, A. W. (1999). Gradient descent for general reinforcement learning. *Advances in Neural Information Processing Systems, 20*, 968–974.

Balci, O. (2001). A methodology for certification of modeling and simulation applications. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, *11*(4), 352–377.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The ucsd map of science. *PLoS One*, *7*(7), e39464.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer.

Bourke, P., & Butler, L. (1998). Institutions and the map of science: Matching university departments and fields of research1. *Research Policy*, *26*(6), 711–718.

Crookall, D. (2010). Serious games, debriefing, and simulation/gaming as a discipline. *Simulation and Gaming*, *41*(6), 898–920.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, *37*(1), 36–48.

Eyheramendy, S., Lewis, D., & Madigan, D. (2003). On the naive bayes model for text categorization. In *Proceedings of the ninth international workshop on artificial intelligence and statistics*, pp. 705–722.

Fox, C. (1989). A stop list for general text. In *ACM SIGIR forum* (Vol. 24, pp. 19–21). ACM.

Glänzel, W., & Schubert, A. (2005). Analysing scientific networks through co-authorship. In H. F. Moed, W. Glänzel, K. U. Leuven & U. Schmoch (Eds.), *Handbook of quantitative science and technology research* (pp. 257–276). New York, NY: Springer.

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics*, *35*(2), 167–176.

Glänzel, W., & Moed, H. F. (2002). Journal impact measures in bibliometric research. *Scientometrics*, *53*(2), 171–193.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing and Management*, *41*(6), 1548–1572.

Herrera, M., Roberts, D. C., & Gulbahce, N. (2010). Mapping the evolution of scientific fields. *PloS One*, *5*(5), e10355.

Hinze, S. (1994). Bibliographical cartography of an emerging interdisciplinary discipline: The case of bioelectronics. *Scientometrics*, *29*(3), 353–376.

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. *American Music*, *183*(5,049), 2–209.

Ioannidis, J. P. A. (2006). Concentration of the most-cited papers in the scientific literature: Analysis of journal ecosystems. *PLoS One*, *1*(1), e5.

Jahn, N., Fenner, M., & Schirrwagen, J. (2013). PlosopenR–exploring FP7 funded PLOS plosopenR–exploring FP7 funded PLOS. *Information Services & Use, 33*(2), 93–101.

Jordan, A. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems*, *14*, 841.

Katz, J. S., & Hicks, D. (1995). The classification of interdisciplinary journals: A new approach. In *Proceeding of the fifth biennial conference of the international society for scientometrics and informatics*, pp. 7–10.

Kaur, J., Hoang, D. T., Sun, X., Possamai, L., JafariAsbagh, M., Patil, S., et al. (2012). Scholarometer: A social framework for analyzing impact across disciplines. *PloS One*, *7*(9), e43235.

Kim, S.-B., Han, K.-S., Rim, H.-C., & Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. *IEEE Transactions on Knowledge and Data Engineering*, *18*(11), 1457–1466.

Kohavi, R., et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International joint conference on artificial intelligence* (Vol. 14, pp. 1137–1145). Lawrence Erlbaum Associates Ltd.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In D. E. Chemnitz (Ed.), *Machine learning: ECML-98* (pp. 4–15). New York, NY: Springer.

Lin, F.-R., Hsieh, L.-S., & Chuang, F.-T. (2009). Discovering genres of online discussion threads via text mining. *Computers and Education*, *52*(2), 481–495.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. Aggarwal & C. Zhai (Eds.), *Mining text data* (pp. 415–463). New York, NY: Springer.

Mayr, E. (2004). *What makes biology unique? Considerations on the autonomy of a scientific discipline*. Cambridge: Cambridge University Press.

McCallum, A., Nigam, K., et al. (1998). A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (Vol. 752, pp. 41–48). Citeseer.

Miltsakaki, E., & Troutt, A. (2008). Real-time web text classification and analysis of reading difficulty. In *Proceedings of the third workshop on innovative use of NLP for building educational applications*, pp. 89–97. Association for Computational Linguistics.

Nederhof, A. J., & Noyons, E. C. M. (1992). Assessment of the international standing of university departments' research: A comparison of bibliometric methods. *Scientometrics*, *24*(3), 393–404.
NIH. (2003). National Institute of Health Research Awards 1990–2012 via Exporter. http://exporter.nih.gov/. Accessed June 19, 2013.
Noyons, E. (2001). Bibliometric mapping of science in a policy context. *Scientometrics*, *50*(1), 83–98.
Noyons, E. C. M., Moed, H. F., & Luwel, M. (1999). Combining mapping and citation analysis for evaluative bibliometric purposes: A bibliometric study. *Journal of the Association for Information Science and Technology*, *50*(2), 115.
Pazzani, M., & Meyers, A. (2003). NSF Research Award Abstracts 1990–2003 Data Set. http://archive.ics.uci.edu/ml/datasets/NSF+Research+Award+Abstracts+1990-2003. Accessed June 19, 2013.
Rajman, M., & Besançon, R. (1998). Text mining: Natural language techniques and text mining applications. In S. Spaccapietra & F. Maryanski (Eds.), *Data mining and reverse engineering* (pp. 50–64). New York, NY: Springer.
Salter, L., & Hearn, A. (1997). *Outside the lines: Issues in interdisciplinary research*. Montreal: McGill-Queen's Press-MQUP.
Sarjoughian, H. S., & Zeigler, B. P. (2001). Towards making modeling & simulation into a discipline. *Simulation Series*, *33*(2), 130–135.
Searls, D. B. (2010). The roots of bioinformatics. *PLoS Computational Biology*, *6*(6), e1000809.
Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys (CSUR)*, *34*(1), 1–47.
Vessey, I., Ramesh, V., & Glass, R. L. (2005). A unified classification system for research in the computing disciplines. *Information and Software Technology*, *47*(4), 245–255.
Vinkler, P. E. T. E. R. (1988). An attempt of surveying and classifying bibliometric indicators for scientometric purposes. *Scientometrics*, *13*(5–6), 239–259.
Wallace, M. L., Larivière, V., & Gingras, Y. (2012). A small world of citations? The influence of collaboration networks on citation practices. *PloS One*, *7*(3), e33339.
Wang, B., & PAN, W. (2005). A survey of content-based anti-spam email filtering [j]. *Journal of Chinese Information Processing*, *5*, 000.
Wei, C.-H., Harris, B. R., Li, D., Berardini, T. Z., Huala, E., Kao, H.-Y., et al. (2012). Accelerating literature curation with text mining tools: A case study of using PubTator to curate genes in PubMed abstracts. *Database, 2012*, bas041. doi:10.1093/database/bas041.
White, J. (2001). Open portal for digital library. *Communications of the ACM*, *44*(7), 14–44.
Yu, B. (2008). An evaluation of text classification methods for literary study. *Literary and Linguistic Computing*, *23*(3), 327–343.
Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *Proceedings of the twenty-first international conference on machine learning*, p. 116. ACM.