

# Categorization of Computing Education Resources into the ACM Computing Classification System

Yinlin Chen  
Department of Computer Science  
Virginia Tech  
Blacksburg, VA  
+1 540 808 9053  
ylchen@vt.edu

Edward A. Fox  
Department of Computer Science  
Virginia Tech  
Blacksburg, VA  
+1 540 552 8667  
fox@vt.edu

Paul Logasa Bogen II  
Intelligent Computing Research Team  
Computational Data Analytics Group  
Oak Ridge National Laboratory  
Oak Ridge, TN 37931  
+1 865 241 0337

bogenpl@ornl.gov

Haowei Hsieh  
School of Library and Information  
Science  
University of Iowa  
Iowa City, IA 52242  
+1 319 335 5713  
haowei-hsieh@uiowa.edu

Lillian N. Cassel  
Department of Computing Sciences  
Villanova University  
Villanova PA 19085  
+1 610 519 7341  
lillian.cassel@villanova.edu

## ABSTRACT

The Ensemble Portal harvests resources from multiple heterogenous federated collections. Managing these dynamically increasing collections requires an automatic mechanism to categorize records in to corresponding topics. We propose an approach to use existing ACM DL metadata to build classifiers for harvested resources in the Ensemble project. We also present our experience on utilizing the Amazon Mechanical Turk platform to build ground truth training data sets from Ensemble collections.

## Categories and Subject Descriptors

H.3.7 [Digital Libraries]: *Collection*. I.5.2 [Design Methodology]: *Classifier design and evaluation*.

## General Terms

Design, Experimentation.

## Keywords

Digital libraries, machine learning, classification, Amazon Mechanical Turk.

## 1. INTRODUCTION

The Ensemble Project is a multi-university project funded by NSF to add a computing education portal to the NSDL family of STEM Pathways. Currently the Ensemble portal has 4432 metadata records harvested from 15 different content providers. All of these resources are harvested in Dublin Core format.

In order to help managers and users efficiently gather accurate information, resources need to be correctly classified. Correctly categorized resources help digital librarians to manage their

collections and improve faceted searches.

We attempt to use the ACM Computing Classification System (CCS) as a source of categories for Ensemble resources. By using the CCS, we have a potential source of training data through the ACM DL for building classifiers that we believe will be able to classify Ensemble resources correctly. In this paper, we describe the building of our ground-truth test collection, our preliminary experiments with the ACML DL and our current results.

Section 2 describes related work about classification and other research using Amazon's MTurk service. Section 3 describes our procedure for training data preparation and ground truth selection. Section 4 describes our preliminary experiments. Section 5 describes our conclusions and future work.

## 2. RELATED WORK

Documents can be categorized manually, via an automatic technique (such as clustering), or semi-automatically using a classifier [1]. Manual categorization yields good results, but carries additional costs in terms of personnel. Automatic techniques are useful for finding patterns in data, but do not guarantee the classes found will correspond to pre-defined categories [2]. In many cases the difficulty in working with semi-automatic techniques is building adequate training data. Corpus analysis aims to mitigate this problem by using available document collections that have been pre-categorized as a source of training material. Large digital are ideal candidates for corpus-based training. Already this kind of technique has been used with Reuters's news archive to classify medical documents [3], and with Wikipedia articles to classify educational resources [4].

Amazon Mechanical Turk (MTurk) is a crowdsourcing service in which researchers can post tasks to be completed and give rewards to the workers who complete them [5][7]. These tasks are called Human Intelligence Tasks (HIT), researchers design HITs to conduct behavioral research [7], or do annotation tasks for NLP [8]. In this paper, we designed and conducted two types of HITs in the MTurk and collected our ground truth records from the Ensemble project for evaluating classifiers we have built.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

JCDL '12, June 10-14, 2012, Washington, DC., USA.

Copyright 2010 ACM 1-58113-000-0/00/0010...\$10.00.

### 3. DATA SETS AND GROUND TRUTH

This section describes procedures to build training data sets from ACM metadata and a ground truth set from Ensemble collection.

#### 3.1 ACM Metadata Set

The ACM metadata set describes conference papers and journal articles from 1954 to 2011. There are 168,639 metadata records, classified according to the ACM computing classification system (CCS). The CCS is a hierarchical classification scheme with varying levels of specificity. To balance between specificity and ample numbers of documents for our test collection, we looked at CCS at the second-level of the hierarchy. There are 61 second-level categories. Because the Ensemble collection is relatively small compared to the ACM metadata set and we wanted at least 4 Ensemble records for each selected category as testing data set, we selected 14 second-level categories from CCS for which we have a good number of entries in Ensemble collection.

The majority of the Ensemble records were published within the past ten years, thus we selected ACM DL metadata that was published after 2001 as our training data set.

For our selected categories, we extracted records from the ACM metadata set into a SQLite database to do feature extraction. We considered several factors when processing this metadata set. Since our purpose is to classify educational materials available through the Ensemble Portal, we focused on analogous metadata fields in both the ACM and Ensemble collections. This proved to be the title and abstract in ACM or description in Ensemble. Additionally, the documents in the ACM collection can be in multiple categories. Since no category is privileged over any other, we treat a document with multiple categories as multiple documents – each in a single category. For each document we store the title, abstract, and a category. After the database was populated, we started building our ground truth set from the Ensemble collection. We asked two different user groups, information science students at University of Iowa and MTurk, to accomplish the ground truth selection tasks. The detailed procedures are described in the following two sections.

#### 3.2 Iowa

The ground truth selection task was planned as an in-class group exercise after a lecture about HCI research where Domain Experts and Human Intelligence Tasks (HIT) can be used.

##### 3.2.1 Participants

The 25 participants are students taking the required Computing Foundation class at SLIS (School of Library and Information Science). In the class, they learn basic knowledge, terminology, and skills of computing and programming. Most of the students are also taking another SLIS required course, Conceptual Foundation, which covers “Theory, principles, and standards in organization of information; function of catalogs, indexes, bibliographic networks; introduction to meta-data descriptions, name and title access, subject analysis, controlled vocabularies, classification systems.” The foundation classes should provide students with adequate basics for the exercise.

##### 3.2.2 Methodology

The 25 students in the computing foundation class were divided into 10 groups (5 2-person groups and 5 3-person groups). We chose 5 computing topics (Programming Techniques, Data Structures, Formal Languages, Discrete Mathematics, and

Database Management) where each topic is assigned to two groups. Each group was asked to spend a little time to learn what that topic is about, a link to Wikipedia about that topic was also provided. We expect that students should be able to learn enough to categorize related resources and be able to identify records in the Ensemble collection that can be part of their given topic. Each group was asked to find 20 items that belong to the given topic. For each record they found they were asked to include the URL to that record and a short explanation to justify their choice.

Table 1 - Comparison of collection techniques

Experiment	Resources	Participants	Cost (\$)	Time
Iowa	1180	90	\$11.8	8.61 hrs
MTurk	485	47	\$4.85	2.58 hrs

The exercise started in-class and continued two more days for the students to finish typing the results. The reports were collected online through the course system. At the end, 9 student reports were collected and 97 documents were categorized (one team did not finish the exercise in time).

#### 3.3 Amazon Mechanical Turk (MTurk)

Additionally, we used MTurk to also build our ground truth set from the Ensemble portal to covers all 14 selected categories. We first employed *Single assignment* strategy to gather records from Ensemble portal, and then we used *Plurality* strategy to evaluate these records identified by both Iowa students and MTurk workers. Based on the majority vote result of these records, we selected our final ground truth records for testing our classifiers.

##### 3.3.1 HIT design – Ground Truth

We used the same approach as in the Iowa class to collect these 14 selected categories records from Ensemble portal. We created a HIT for each category. In each HIT, we provided a short description of the category and a link to Wikipedia about that category. Workers were asked to use the Ensemble portal search interface to find 14 records that matched their assigned category. They were also instructed to not simply use the category as their search terms to find records. To verify this, they must document what search terms they used in this HIT and optionally provide reasons why they think this record is matched to this category. We estimated that each HIT required at least one hour to finish so we set the time allotted per HIT to 6 hours.

In order to get high quality results from workers, we set a high worker qualification on our HIT. Our qualification is a worker who has a HIT approval rate greater than or equal to 95%, has had greater than 50 HITs approved, and whose location is in the United States. We paid \$1 for each approved assignment; this reward is pretty high in the MTurk HIT pool.

We created another 9 HITs with the same rewards for Iowa students to have records covering all 14 categories. We announced these HITs to Iowa students twice during the experiment. However, no Iowa student accepted these HITs. One possible reason is that this experiment was hosted during the winter break and that was a bad time for students to do experiments. A second possible reason is that even though a \$1 reward is high in MTurk, since the average hourly rate among campus works is at least \$5, the HIT reward is relatively low.

After three weeks, a total of 14 HITs were completed by at least one MTurk worker. We were able to get 263 records classified

across 14 categories. The average completion time per HIT is one hour and 22 minutes. The effective hourly rate is \$0.79. Workers had carefully answered each HIT. The workers not only just gave the Ensemble record URL and search terms, but also detailed the reason why they chose this record for this category.

In order to evaluate these records and compare the quality between Iowa students and MTurk workers, we used *Plurality* strategy – majority vote to assess the records gathered from both user groups and described in the next section.

### 3.3.2 HIT design – Plurality

Our second MTurk experiment has two purposes: to find the ground truth records for testing our classifiers and to investigate whether MTurk workers can do the same work as students who have background knowledge in computing education materials.

To assess work quality between these groups, we used a plurality review strategy to design our second MTurk experiment. First, we reviewed the Iowa students’ reports and got 97 resources for evaluation. Second, we got 236 resources for evaluation from our first MTurk experiment. For each resource we created a corresponding HIT. Inside each HIT, we listed the resource’s title, description and category description. We asked workers to rate the decision from 1 to 5 based on the provided information. We assigned 5 workers to each HIT and accept the categorization of that resource if the average score of the HIT is above 3.

We created a total of 485 assignments from the Iowa resources and 1180 assignments from the MTurk resources. The worker qualification was set to an overall hit approval rate of 70%. Since each HIT was relatively easy, and our estimated time requirement was 5 to 10 minutes for each HIT, we set our HIT reward to be 1 cent.

We used the time difference to detect suspicious cheating workers. We monitored worker’s response time on each HIT. If a worker keeps having the lowest completion time to answer HITs compared to other worker’s completion time in the same HIT, we suspect that worker was cheating and rejected all the HITs answered by that worker. We found two workers who kept using the minimal time to answer HITs. Most of HITs they answered used only 3 seconds, compared to other workers’ completion time average 15 seconds. We rejected these two workers and republished HITs to the MTurk pool for other workers.

We got all the HITs completed by workers in three days. There were a total of 137 workers involved in our second experiment. Table 1 lists the detailed information about our second experiment. We calculated each record’s average score and selected records with score over 3 as our ground truth for that category.

## 3.4 Analysis

We compared the Iowa students’ work with MTurk workers’ work using the data from our second experiment. Our hypothesis was that MTurk workers could perform the same quality work as Iowa students. We used a t-test to verify whether our hypothesis is accepted or rejected. The t-test result shows that there is no

significant difference between these two user groups at a confidence level of 0.95. Thus we can conclude that workers can do the same categorization work as students.

We inspected the entries identified by Iowa students and MTurk workers. We found the resources selected by the two groups differed. Table 2 shows that the number of common and total records identified by these two groups.

**Table 2 - Results of ground truth set building.**

<i>ACM second-level category</i>	<i>Records (Iowa)</i>	<i>Records (MTurk)</i>	<i>Records (Both)</i>
Programming Techniques	26	15	1
Data Structures	33	15	1
Formal Languages	9	15	3
Discrete Mathematics	10	15	3
Database Management	19	15	2

We can clearly see that the number of records for each category from Iowa varied, but they are consistent from MTurk. This is because not every document identified by Iowa was valid and some are duplicates. These invalid records were removed for our second experiment.

While both groups used the same interface, MTurk workers carefully followed instructions and used complex and different search terms to find documents. However, students in class often used one-search term and listed the top 20 records returned. There were only four Iowa students who gave us the detailed reason why they chose a document for a category. Others groups only gave us URLs without any explanation. Workers need to get HITs approved by the requester in order to get paid. A rejected HIT not only is a waste of the works’ time and effort, but can get them negative feedback, affecting their overall approval rate and reputation.

## 4. PRELIMINARY EXPERIMENTS

To move beyond manual-classification of documents in Ensemble Collections, we experiment with classifiers trained on metadata from the ACM Digital Library. We extracted three distinct term vectors – one over titles, one over abstracts, and one over the combination of title and abstract. We remove stopwords using the Stanford stopword list, stem the remaining terms using Porter’s Snowball algorithm, and trim out terms that are present in less than 5% and over 95% of documents. Classifiers were validated first via 10-fold cross-validation on ACM metadata entries.

We found three techniques that performed well on training data – Nearest Neighbor, Random Trees, and Random Forest each of which obtained an 85.697% accuracy for validation. However, when applied to hand-classified\*\* Ensemble documents, performance dropped considerably. The best results obtained overall were by a locally weighted Multinomial Naïve Bayes classifier with 25.525% accuracy and the bagging extension of Multinomial Naïve Bayes with 24.625% accuracy.

**Table 3 - Best techniques per category by F Score**

Category	Technique	F-Score	Precision	Recall
B.3	J48 Graft Tree	0.190	0.333	0.133
B.7	Complement Naïve Bayes	0.200	0.400	0.133
C.1	Linear	0.316	0.750	0.200
C.2	LogitBoost	0.684	0.684	0.684
D.1	REP Tree	0.160	0.444	0.098
D.2	REP Tree	0.233	0.136	0.800
D.3	Decision Table	0.262	0.170	0.571
D.4	LogitBoost	0.491	0.464	0.520
E.1	Naïve Bayes Multinomial	0.222	1.000	0.125
E.5	J48 Graft Tree	0.000	0.000	0.000
F.4	Discriminative Multinomial	0.424	0.778	0.292
G.2	Naïve Bayes Random Committee of Random Trees	0.263	0.385	0.200
H.2	Discriminative Multinomial Naïve Bayes	0.286	0.364	0.235
K.4	Naïve Bayes Multinomial	0.300	0.185	0.800

When results are examined per class, some classes were better classified than others. Hyper Pipe classifiers obtained perfect recall on the B.3 category, albeit at the cost of very low precision (0.045). The Complement Naïve Bayes classifier obtained perfect precision at the cost of low recall (0.267) for C.1, high recall (0.789) and moderate precision (0.441) for C.2., and decent recall (0.643) and low precision (0.107) for D.3. The one rule classifier had perfect recall on D.2 again at the cost of low precision (0.048). The voting feature intervals classifier was perfect precision classifier for E.1 and F.4 but obtained poor recall on both (0.021 and 0.041 respectively). These results lead us to believe that our classifiers were overfit on training data and, thus, techniques need to be adopted to mitigate this effect. The best techniques for each category by F Score can be seen in table 3.

## 5. CONCLUSIONS & FUTURE WORK

We gained valuable experiences about conducting research with MTurk from this research:

- Experiments can be conducted on MTurk anytime and results obtained in a short period. It is difficult to engage students to participate in experiment during breaks even with participation rewards.
- Workers can quickly complete HITs with simple tasks, thus we need to have a validation technique to identify possible low quality results. An automatic approval and rejection workflow is needed.
- Workers cannot easily pretend they did a good job for HITs with complex tasks. It is easy to identify low quality work through completion times and result reports.
- Workers with higher approved rate are willing to spend time to participate in research even the average hour rate is low.
- Worker skills and accuracy vary widely, thus it is important to find a group of trusted workers.
- We are working on an approach to automatically evaluate results from workers, approve qualifying HITs, reject low quality HITs and resubmit to the HIT pool.

Our preliminary experiments also provided us with valuable insight for moving forward. While our validation results were

very good and test results in individual categories show promise, the poor test results in other categories continue to hold back the overall performance of our techniques. In order to better understand this issue, we intend to perform a more in depth examination of the Ensemble documents, expand the features sets used for analysis, and apply other ensemble learning techniques.

## 6. ACKNOWLEDGMENTS

This research is supported by NSF Grants DUE-0840713, 0840715, 0840719, 0840721, 0840668, 0840597, 0836940, and 0937863.

This document was prepared by Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, Tennessee 37831-6285; managed by UT-Battelle, LLC, for the US Department of Energy under contract number DE-AC05-00OR22725.

This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Our thanks to ACM for providing us the ACM DL Metadata to use in this research, to the Academy for Advanced Telecommunications and Learning Technologies at Texas A&M University for time on Brazos and to NICS/XSEDE/TeraGrid for the time on Nautilus.

## 7. REFERENCES

- [1] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.* 34, 1, 1-47.
- [2] Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data clustering: a review. *ACM Comput. Surv.* 31, 3, 264-323.
- [3] Chen, G., Warren, J., and Riddle, P. (2010). Semantic Space models for classification of consumer webpages on metadata attributes. *J. of Biomedical Informatics* 43, 5, 725-735.
- [4] Meyer, M., Rensing, C., and Steinmetz, R. (2008). Using community-generated contents as a substitute corpus for metadata generation. *Int. J. Adv. Media Comm.* 2, 1, 59-72.
- [5] Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *In Proc. of CHI 08.*
- [6] Mason, W., & Suri, S. (2010). Conducting Behavioral Research on Amazon's Mechanical Turk. *Behavior Research Methods*, 5(5), 1-23.
- [7] Chen, J. J., Menezes, N. J., Bradley, A. D., & North, T. A. (2011). Opportunities for Crowdsourcing Research on Amazon Mechanical Turk. *Human Factors*, 5, 3.
- [8] Yetisgen-yildiz, M., Solti, I., Xia, F., & Halgrim, S. R. (2010). Preliminary Experience with Amazon's Mechanical Turk for Annotating Medical Named Entities. *Computational Linguistics*, 180-183