Dissertation Defense

SVMAUD: USING TEXTUAL INFORMATION TO PREDICT THE AUDIENCE LEVEL OF WRITTEN WORKS USING SUPPORT VECTOR MACHINES

Presented By: Todd Will
PhD Candidate, Department of Information Systems, NJIT
todd.will@njit.edu

DATE: Wednesday, November 13, 2013
COFFEE: 1:15-1:30
TIME: 1:30-4:30
LOCATION: GITC Room 1403

ABSTRACT

Information retrieval systems should seek to match resources with the reading ability of the individual user; similarly, an author must choose vocabulary and sentence structures appropriate for his or her audience. Traditional readability formulas, including the popular Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score, rely on numerical representations of text characteristics, including syllable counts and sentence lengths, to suggest audience level of resources. However, the author’s chosen vocabulary, sentence structure, and even the page formatting can alter the predicted audience level by several levels,
especially in the case of digital library resources. For these reasons, the performance of readability formulas when predicting the audience level of digital library resources is very low.

Rather than relying on these inputs, machine learning methods, including cosine, Naïve Bayes, and Support Vector Machines (SVM), can suggest the grade of an essay based on the vocabulary chosen by the author. The audience level prediction and essay grading problems share the same inputs, expert-labeled documents, and outputs, a numerical score representing quality or audience level. After a human expert labels a representative sample of resources with audience level, the proposed SVM-based audience level prediction program, SVMAUD, constructs a vocabulary for each audience level; then, the text in an unlabeled resource is compared with this predefined vocabulary to suggest the most appropriate audience level.

Two readability formulas and four machine learning programs are evaluated with respect to predicting human-expert entered audience levels based on the text contained in an unlabeled resource. In a collection containing 10,238 expert-labeled HTML-based digital library resources, the Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score predict the specific audience level with F-measures of 0.10 and 0.05, respectively. Conversely, cosine, Naïve Bayes, the Collins-Thompson and Callan model, and SVMAUD improve these F-measures to 0.57, 0.61, 0.68, and 0.78, respectively. When a term’s weight is adjusted based on the HTML tag in which it occurs, the specific audience level prediction performance of cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD improves to 0.68, 0.70, 0.75, and 0.84, respectively. When title, keyword, and abstract metadata are used for training, cosine, Naïve Bayes, the Collins-Thompson and Callan model, and SVMAUD specific audience level prediction F-measures are found to be 0.61, 0.68, 0.75, and 0.86, respectively. When cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD are trained and tested using resources from a single subject category, specific audience level prediction F-measure performance improves to 0.63, 0.70, 0.77, and 0.87, respectively. SVMAUD experiences the highest audience level prediction performance among all methods under evaluation in this study. After SVMAUD is properly trained, it can be used to predict the audience level of any written work.

**BIO**
Todd is a PhD candidate in the Information Systems department at the New Jersey Institute of Technology. His research interests including applying machine
learning methods to solve classification problems, information retrieval, improving access to digital resources, and text mining. He is currently a teaching assistant in the information systems department.

**Dissertation Committee:**

Dr. Yi-Fang Brook Wu, Associate Professor and Chair, IS, NJIT (committee chair)

Dr. Vincent Oria, Associate Professor, CS, NJIT

Dr. Lian Duan, Assistant Professor, IS, NJIT

Dr. Julian M. Scher, Associate Professor Emeritus, IS, NJIT

Dr. Il Im, Associate Professor, Yonsei University, Korea (external committee member)

[WebEx Login](https://njit.webex.com/njit/j.php?ED=253232182&UID=1688560272&RT=MiMxMQ%3D%3D)

**Topic:** Todd Will Dissertation Defense  
**Date:** Wednesday, November 13, 2013  
**Time:** 1:00 pm, Eastern Standard Time (New York, GMT-05:00)  
**Meeting Number:** 641 204 763  
**Meeting Password:** (This meeting does not require a password.)

To join the online meeting (Now from mobile devices!)

2. If requested, enter your name and email address.
3. If a password is required, enter the meeting password: (This meeting does not require a password.)
4. Click "Join".