Abstract:

Information retrieval systems should seek to match resources with the reading ability of the individual user. Traditional readability formulas, including the popular Flesch-Kincaid Reading Age, rely on numerical representations of text characteristics, including syllable counts and sentence lengths, to suggest the audience level of the resource. However, as HTML pages contain tables, hyperlinks, and menus to present information versus text in printed books, the values used as inputs by readability formulas are usually distorted.

Rather than manually editing resources before calculating numerical values for word and sentence characteristics, machine learning methods, including cosine 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.

This preliminary study has compared the performance of readability formulas and machine learning programs in their ability to correctly predict audience levels. 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 have predicted the specific audience level with F-measures of 0.10 and 0.05, respectively. On the other hand, cosine, Naïve Bayes, the Collins-Thompson and Callan model, and SVMAUD have improved these F-measures to 0.57, 0.61, 0.68, and 0.78, respectively. When titles, keywords, and abstracts were used for training, the F-measures for cosine, Naïve Bayes, the Collins-Thompson and Callan model, and SVMAUD have improved to 0.61, 0.68, 0.75, and 0.86, respectively. In addition, when cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD have been trained and tested using resources from a single subject category, the specific audience level prediction performance has further improved to F-measures of 0.63, 0.70, 0.76, and 0.87, respectively. This preliminary study has clearly shown that SVMAUD has been able to predict the audience level of digital library resources with much higher performance than readability formulas.
However, since the NSDL has mainly held STEM topics, the audience level prediction tools that have been developed in this study would need to be extended to other collections and subject categories found in digital libraries in the final dissertation. In addition, the performance of SVMAUD would be compared with the prediction performance of readability formulas and other machine learning methods when predicting the audience level for books, pamphlets, articles, and other written materials. SVMAUD has been found to outperform readability formulas and other machine learning systems in this preliminary study and could also suggest the audience level for any written work.

Bio:
Todd C. Will is a PhD candidate in the Department of Information Systems at NJIT. He also holds a BS in Management Information Systems awarded by NJIT and an MBA in Management of Technology also awarded by NJIT. His research interests include knowledge management, digital libraries, automatic metadata generation (AMG), database management systems, and information retrieval. He is the co-author of academic papers published in the proceedings of the Americas Conference on Information Systems and the Joint Conference on Digital Libraries.

Dissertation Proposal Committee:
Dr. Yi-Fang Brook Wu, Associate Professor of Information Systems, NJIT (Advisor)
Dr. Vincent Oria, Associate Professor of Computer Science, NJIT (Committee Member)
Dr. Lian Duan, Assistant Professor of Information Systems, NJIT (Committee Member)
Dr. Julian M. Scher, Associate Professor of Information Systems, NJIT (Committee Member)
Dr. Il Im, Associate Professor, School of Business, Yonsei University, Korea (External Committee Member)