

Fall 1-31-2014

## **SVMAUD: Using textual information to predict the audience level of written works using support vector machines**

Todd Will  
*New Jersey Institute of Technology*

Follow this and additional works at: <https://digitalcommons.njit.edu/dissertations>



Part of the [Databases and Information Systems Commons](#), and the [Management Information Systems Commons](#)

---

### **Recommended Citation**

Will, Todd, "SVMAUD: Using textual information to predict the audience level of written works using support vector machines" (2014). *Dissertations*. 153.  
<https://digitalcommons.njit.edu/dissertations/153>

This Dissertation is brought to you for free and open access by the Electronic Theses and Dissertations at Digital Commons @ NJIT. It has been accepted for inclusion in Dissertations by an authorized administrator of Digital Commons @ NJIT. For more information, please contact [digitalcommons@njit.edu](mailto:digitalcommons@njit.edu).

## **Copyright Warning & Restrictions**

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be “used for any purpose other than private study, scholarship, or research.” If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of “fair use” that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

**Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation**

Printing note: If you do not wish to print this page, then select “Pages from: first page # to: last page #” on the print dialog screen

The Van Houten library has removed some of the personal information and all signatures from the approval page and biographical sketches of theses and dissertations in order to protect the identity of NJIT graduates and faculty.

## **ABSTRACT**

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

**by  
Todd Will**

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 level 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 is 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, the 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.

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

**by  
Todd Will**

**A Dissertation  
Submitted to the Faculty of the  
New Jersey Institute of Technology  
in Partial Fulfillment of the Requirements for the Degree of  
Doctor of Philosophy in Information Systems**

**Department of Information Systems**

**January 2014**

Copyright © 2014 by Todd Will

ALL RIGHTS RESERVED.

**APPROVAL PAGE**

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

**Todd Will**

---

Dr. Yi-Fang Brook Wu, Dissertation Advisor Date  
Associate Professor and Chair of Information Systems, NJIT

---

Dr. Il Im, Committee Member Date  
Associate Professor, School of Business, Yonsei University, Korea

---

Dr. Vincent Oria, Committee Member Date  
Associate Professor of Computer Science, NJIT

---

Dr. Lian Duan, Committee Member Date  
Assistant Professor of Information Systems, NJIT

---

Dr. Julian M. Scher, Committee Member Date  
Associate Professor Emeritus of Information Systems, NJIT



## BIOGRAPHICAL SKETCH

**Author:** Todd Will  
**Degree:** Doctor of Philosophy  
**Date:** January 2014

### **Undergraduate and Graduate Education:**

- Doctor of Philosophy in Information Systems,  
New Jersey Institute of Technology, Newark, NJ, 2014
- Master of Business Administration in Management of Technology,  
New Jersey Institute of Technology, Newark, NJ, 2003
- Bachelor of Science in Management Information Systems,  
New Jersey Institute of Technology, Newark, NJ, 2002

**Major:** Information Systems

### **Presentations and Publications:**

- Will, T., & Wu, Y.-F. (2012). Improving Access to Digital Library Resources by Automatically Generating Complete Reading Level Metadata. *Proceedings of the Americas Conference on Information Systems (AMCIS)*, Seattle, Washington.
- Will, T., Srinivasan, A., Im, I., & Wu, Y.-F. (2009). Search Personalization: Knowledge Based Recommendation in Digital Libraries. *Proceedings of the Americas Conference on Information Systems (AMCIS)*, San Francisco, California.
- Will, T., Srinivasan, A., Bieber, M., Im, I., Oria, V., & Wu, Y.-F. (2009). GRE: Hybrid Recommendations for NSDL Collections. *Proceedings of the Joint Conference on Digital Libraries (JCDL)*, Austin, Texas.

*This dissertation is dedicated to my beloved family*

To my parents, Karen and Thomas, who have supported me through my entire education,  
To my brother, Mark, and my sister, Christy,  
With whom I shared many wonderful moments of my life.

## ACKNOWLEDGMENTS

I would like to express my deep and everlasting gratitude to my dissertation advisor, Dr. Yi-Fang Brook Wu, for her constant help and guidance while conducting my dissertation research. I would also like to especially thank one of my committee members, Dr. Il Im, my initial student advisor, for his continued support in my pursuit of this degree. My other committee members, Dr. Vincent Oria, Dr. Lian Duan, and Dr. Julian M. Scher, have provided excellent comments in the development of the research study and the analysis of results. It has been a great honor to know the members of my committee and I cannot thank them enough for their hard work and dedication to make this dissertation possible.

I would also like to thank Quanzhi Li, Nkechi Nnadi, Umar Qasim, and Anand Srinivasan for their help in developing the experimental systems used in this research. The participants in the case study portion of this research should also be recognized since they took time from their busy schedules to provide useful data for analysis. Finally, I would like to thank my fellow PhD students and faculty in the Information Systems Department for their encouragement and guidance as I completed this research.

This work would not have been possible without the generous financial support of the National Science Foundation and the Institute for Museum and Library Services to allow me to focus on this research. I would also like to acknowledge the collaboration provided by the Digital Library for Earth System Education (DLESE), Econport, and Teacher's Domain to develop integrated interfaces and providing a portion of documents for the collection used in the study.

## TABLE OF CONTENTS

Chapter	Page
1 INTRODUCTION.....	1
1.1 Objective.....	4
1.2 Background.....	6
1.3 Research Overview.....	8
1.4 Contributions.....	9
1.5 Dissertation Organization.....	10
2 LITERATURE REVIEW.....	11
2.1 Readability Formulas.....	12
2.1.1 Flesch-Kincaid Reading Age.....	12
2.1.2 Dale-Chall Reading Ease Score.....	15
2.1.3 Gunning-Fog Index.....	16
2.1.4 Simple Measure of Gobbledygook (SMOG).....	17
2.1.5 Spache Readability Formula.....	18
2.1.6 Advantage Open Standard for Readability (ATOS).....	20
2.1.7 Lexile Framework for Reading.....	25
2.1.8 Readability Formulas Summary.....	28
2.2 Machine Learning Methods.....	30
2.2.1 Cosine.....	33
2.2.2 Naïve Bayes.....	37
2.2.3 Clustering.....	41

**TABLE OF CONTENTS**  
**(Continued)**

<b>Chapter</b>	<b>Page</b>
2.2.4 Support Vector Machines.....	50
2.2.5 Latent Semantic Indexing.....	55
2.2.6 Machine Learning Methods Summary.....	61
2.3 Conclusion.....	63
<b>3 SVMAUD SYSTEM DESIGN AND IMPLEMENTATION.....</b>	<b>64</b>
3.1 SVMAUD System Design.....	64
3.1.1 SVMAUD Document Language Model.....	64
3.1.2 SVMAUD System Architecture.....	66
3.2 SVMAUD System Implementation.....	69
3.2.1 SVMAUD Training Program.....	69
3.2.2 SVMAUD System Interface.....	72
3.2.3 SVMAUD System Output.....	74
3.3 SVMAUD Summary.....	75
<b>4 AUDIENCE LEVEL PREDICTION.....</b>	<b>76</b>
4.1 Research Questions.....	78
4.2 Creating the Digital Library Resource Collection.....	81
4.3 Digital Library Collection Overview.....	86
4.4 Digital Library Audience Level Prediction Evaluation.....	90
4.4.1 Readability Formulas versus SVMAUD.....	91
4.4.2 Cosine and Naïve Bayes versus SVMAUD.....	96

**TABLE OF CONTENTS**  
**(Continued)**

<b>Chapter</b>	<b>Page</b>
4.4.3 Collins-Thompson and Callan Method versus SVMAUD.....	100
4.4.4 Machine Learning Using Readability Formula Inputs.....	104
4.5 Digital Library Audience Level Prediction Discussion.....	110
4.5.1 Readability Formulas vs. SVMAUD.....	111
4.5.2 Cosine, Naïve Bayes, and Collins-Thompson and Callan vs. SVMAUD.	113
4.5.3 SVMAUD Performance Using Readability Formula Inputs.....	114
4.6 Digital Library Audience Level Prediction Evaluation.....	116
5 ADJUSTING TERM WEIGHT BASED ON HTML TAGS.....	118
5.1 Previous Studies.....	118
5.2 Document Processing.....	128
5.3 Evaluation.....	130
5.3.1 General Audience Levels.....	132
5.3.2 Specific Audience Levels.....	135
5.4 Performance Improvement.....	139
5.5 Summary and Conclusion.....	142
6 REDUCING NOISE IN THE TRAINING DATASET.....	144
6.1 General Audience Level Noise-Reduced Classification Performance.....	144
6.2 Specific Audience Level Noise-Reduced Classification Performance.....	147
6.3 Effect of Resource Length on SVMAUD Classification Performance.....	151
6.4 SVMAUD Performance by Digital Library Collection.....	154

**TABLE OF CONTENTS**  
**(Continued)**

<b>Chapter</b>	<b>Page</b>
6.5 SVMAUD Performance Improvement.....	157
6.6 Performance Summary and Conclusion.....	159
7 SUBJECT-SPECIFIC CLASSIFICATION.....	162
7.1 Digital Library Collection Overview by Subject.....	162
7.2 Home School Resource Collection Overview.....	164
7.3 Health Sciences Subject-Specific Classifier Performance.....	171
7.4 History and Geography Subject-Specific Classifier Performance.....	176
7.5 Mathematics Subject-Specific Classifier Performance.....	181
7.6 Reading and Writing Subject-Specific Classifier Performance.....	186
7.7 Science Subject-Specific Classifier Performance.....	191
7.8 Technology and Engineering Subject-Specific Classifier Performance.....	196
7.9 Overall Subject-Specific Classifier Performance.....	201
7.10 SVMAUD Subject-Specific Classifier Improvement.....	208
7.11 Machine Learning Using Subject-Specific Classifiers Discussion.....	211
7.12 Machine Learning Using Subject-Specific Classifiers Conclusion.....	213
8 CONCLUSION.....	214
8.1 Completion of Study Objectives.....	214
8.2 Answers to All Research Questions.....	216
8.3 Theoretical and Practical Implications.....	221
8.3.1 Theoretical Implications.....	222

**TABLE OF CONTENTS**  
**(Continued)**

<b>Chapter</b>	<b>Page</b>
8.3.2 Practical Implications.....	225
8.4 Contributions.....	229
8.5 Future Work.....	231
8.6 Conclusion.....	234
APPENDIX A DALE COMMON WORD LIST.....	236
APPENDIX B NSDL METADATA GUIDELINES.....	251
REFERENCES.....	257



## LIST OF TABLES

<b>Table</b>	<b>Page</b>
2.1 ATOS Readability Predictors.....	21
2.2 ATOS and Book Length Grade Level Weights for Long Books.....	23
2.3 Lexile to Grade Level Correspondence.....	27
2.4 Readability Formulas Summary.....	29
2.5 Machine Learning Methods Summary.....	62
4.1 Test Collection Digital Library Sources.....	87
4.2 Unique Feature Words Summary.....	89
4.3 Readability Formulas vs. SVMAUD – General Audience Levels.....	92
4.4 Readability Formulas vs. SVMAUD – Specific Audience Levels.....	94
4.5 Cosine vs. Naïve Bayes vs. SVMAUD – General Audience Levels.....	98
4.6 Cosine vs. Naïve Bayes vs. SVMAUD – Specific Audience Levels.....	99
4.7 Collins-Thompson and Callan vs. SVMAUD – General Audience Levels.....	102
4.8 Collins-Thompson and Callan vs. SVMAUD – Specific Audience Levels.....	103
4.9 SVMAUD using Readability Formula Inputs – General Audience Levels.....	105
4.10 SVMAUD using Readability Formula Inputs – Specific Audience Levels.....	107
4.11 Average Readability Formula Inputs - General Audience Levels.....	108
4.12 Average Readability Formula Inputs - Specific Audience Levels.....	110
5.1 Six Categories and Associated HTML Tags.....	119
5.2 Six Categories and Associated HTML Tags.....	121
5.3 HTML Tag Classes.....	123

**LIST OF TABLES**  
(Continued)

<b>Table</b>	<b>Page</b>
5.4 HTML Tag Classes with PDF Equivalent.....	129
5.5 Cosine vs. Naïve Bayes vs. SVMAUD – General Audience Levels.....	133
5.6 Collins-Thompson and Callan vs. SVMAUD – General Audience Levels.....	134
5.7 Cosine vs. Naïve Bayes vs. SVMAUD – Specific Audience Levels.....	137
5.8 Collins-Thompson and Callan vs. SVMAUD – Specific Audience Levels.....	138
6.1 Noise-Reduced Classification Performance – General Audience Levels.....	145
6.2 Collins-Thompson and Callan vs. SVMAUD – General Audience Levels.....	146
6.3 Noise Reduced Classification Performance – Specific Audience Levels.....	148
6.4 Collins-Thompson and Callan vs. SVMAUD - Specific Audience Levels.....	150
7.1 Digital Library Subject Category Coverage.....	163
7.2 Home School Resource Collection Summary.....	170
7.3 Subject-Specific Classifier Document Collection Summary.....	170
7.4 Health Sciences Specific Audience Level Prediction Results.....	172
7.5 Health Sciences Specific Audience Level Prediction–Thompson & Callan.....	173
7.6 Health Sciences General Audience Level Prediction Results.....	174
7.7 Health Sciences General Audience Level Prediction – Thompson & Callan.....	175
7.8 History & Geography Specific Audience Level Prediction Results.....	177
7.9 History & Geography Specific Audience Prediction–Thompson&Callan.....	178
7.10 History & Geography General Audience Level Prediction Results.....	179
7.11 History & Geography General Audience Prediction – Thompson&Callan.....	180

**LIST OF TABLES**  
**(Continued)**

<b>Table</b>	<b>Page</b>
7.12 Mathematics Specific Audience Level Prediction Results.....	182
7.13 Mathematics Specific Audience Level Prediction Results–Thompson&Callan....	183
7.14 Mathematics General Audience Level Prediction Results.....	184
7.15 Mathematics General Audience Level Prediction Results–Thompson&Callan....	185
7.16 Reading & Writing Specific Audience Level Prediction Results.....	186
7.17 Reading & Writing Specific Audience Level Results-Thompson&Callan.....	188
7.18 Reading & Writing General Audience Level Prediction Results.....	189
7.19 Reading & Writing General Audience Level Prediction – Thompson&Callan.....	190
7.20 Science Specific Audience Level Prediction Results.....	192
7.21 Science Specific Audience Level Prediction Results – Thompson&Callan.....	193
7.22 Science General Audience Level Prediction Results.....	194
7.23 Science General Audience Level Prediction Results – Thompson&Callan.....	195
7.24 Technology & Engineering Specific Audience Level Prediction Results.....	197
7.25 Tech & Eng. Specific Audience Level Prediction Results–Thompson&Callan....	198
7.26 Technology & Engineering General Audience Level Prediction Results.....	199
7.27 Tech & Eng. General Audience Level Prediction Results-Thompson&Callan.....	200
7.28 Overall Specific Audience Level Prediction Results.....	202
7.29 Overall Specific Audience Level Prediction Results – Thompson&Callan.....	204
7.30 Overall General Audience Level Prediction Results.....	205
7.31 Overall General Audience Level Prediction Results – Thompson&Callan.....	206

## LIST OF FIGURES

Figure	Page
2.1 Graphical representation of the Vector Space Model.....	33
2.2 Intrusion detection system sample cluster.....	47
2.3 Sample output of MClust program.....	49
2.4 SVM graphical representation of hyperplanes for a two-class classifier.....	52
3.1 SVMAUD learning function and classification engine.....	67
3.2 Input directory structure for training data.....	70
3.3 SVMAUD classification result.....	73
4.1 Screen capture of a water cycle activity.....	83
4.2 A water cycle activity HTML source code excerpt.....	84
4.3 Screen capture of comparative advantage resource.....	85
4.4 HTML source code excerpt for comparative advantage resource.....	85
5.1 Weight distribution among <i>ctags</i> for different values of $\beta$ .....	126
5.2 $\beta$ versus F-measure for general audience levels.....	132
5.3 $\beta$ versus F-measure for specific audience levels.....	136
5.4 SVMAUD general audience level performance improvement.....	140
5.5 SVMAUD specific audience level performance improvement.....	141
6.1 Effect of resource length on performance – general audience levels.....	152
6.2 Effect of resource length on performance – specific audience levels.....	153
6.3 SVMAUD performance by collection – general audience levels.....	155
6.4 SVMAUD performance by collection – specific audience levels.....	156

## LIST OF FIGURES

<b>Figure</b>	<b>Page</b>
6.5 SVMAUD general audience level performance improvement.....	158
6.6 SVMAUD specific audience level performance improvement.....	159
7.1 Screen capture of Cuneiform program.....	165
7.2 Page 18 image from The Complete Book of Math, Grades 5-6.....	166
7.3 Page 18 text extracted from The Complete Book of Math, Grades 5-6.....	167
7.4 Page 8 image from The Complete Book of Reading.....	168
7.5 Page 8 text extracted from The Complete Book of Reading.....	169
7.6 SVMAUD specific audience level performance by subject category.....	207
7.7 SVMAUD general audience level performance by subject category.....	208
7.8 SVMAUD specific audience level performance improvement.....	209
7.9 SVMAUD general audience level performance improvement.....	210

# CHAPTER 1

## INTRODUCTION

One of the largest problems facing mass communication today is the challenge of appealing to a wide audience with a variety of reading abilities while maintaining the individual reader's interest. The reading ability of a person is influenced by a number of factors, including years of education, interests, prior knowledge, and life experiences. While everyone can understand simple words, a smaller audience with more advanced reading abilities is able to comprehend more complex words. Readers must be challenged to maintain their interest in the article or book but not to the point of becoming frustrated due to an inability to understand the resource content.

Targeting vocabulary to the reading ability of individual users is an ever-present challenge faced by authors. For example, readers of school textbooks include authors, editors, publishers, teachers, and students; however, the vocabulary should be targeted toward the student, or the audience, rather than other potential readers. In a study comparing the usage of the National Science Digital Library (NSDL) against the web search engine Google, the authors suggest that retrieving resources appropriate for the audience level is an important component of digital library search systems (McCown, Johan, & Michael, 2005). The General Pediatrics digital library provides resources to health care providers, patients, and families covering a variety of pediatric topics; the resources in this collection, while providing resources toward families and children, have not been found to be understandable by the average adult, much less a child (D'Alessandro, Kingsley, & Johnson-West, 2001). Recognizing that resources should be

targeted toward the reading abilities of the current user, the Learning Resource Metadata Initiative (LRMI) standard requires the age range of the reader, such as twelve to fifteen years old, to be entered for each resource in the collection (Learning Resource Metadata Initiative, 2013). While digital libraries are most noticeably faced with the problem of labeling all resources with the most appropriate audience level, these problems also extend to other areas of daily life. As President Obama seeks to appeal to the working class, his speeches are targeted toward people who complete eighth grade versus those of former President Bush, who has appealed to constituents with a higher level of education (Ostermeier, 2010). Similarly, the *Wall Street Journal* targets business executives and affluent customers who understand financial markets (Adage.com, 2010); the text must be written using advanced vocabulary and longer sentences than mass market newspapers. Textbook authors, Presidents, newspaper editors, and librarians must be able to verify that the vocabulary is appropriate for their audience.

As manual identification of audience level for resources is both time consuming and labor intensive, computer-based readability tests can predict the most appropriate audience level of literature. One of the most popular readability tests is the Flesch Reading Ease that relies on word and sentence characteristics to predict the reading ability required by the audience (Flesch, 1948). On the other hand, the Dale-Chall Reading Ease Score considers the sentence length and vocabulary chosen by the author by comparing the words in the resource with a list of words appropriate for a fourth grade student (Chall & Dale, 1995). These readability formulas rely on structural and semantic characteristics of text, such as word length, syllables per word, and sentence length, to predict the difficulty of understanding text. However, these methods cannot account for

variability in authors' sentence structures or consider the terms appropriate for each audience level. For example, the word "television" contains four syllables but a pre-kindergarten reader would understand this term while "rhinitis," only containing three syllables, requires a medical professional to explain. Web-based documents contain bullet points and headers and footers that do not follow the conventions of traditional written English and contain far fewer terms than books or articles. In fact, Collins-Thompson & Callan (2005) report that readability formulas tend to perform extremely poorly when identifying the audience level of web-based resources, with a correlation between human and Flesch-Kincaid assigned grades of 0.25 for grades 1-6 and 0.47 for grades 1-12. This problem is shared by scanned books that require Optical Character Recognition (OCR) to extract text; while OCR is able to identify typed words in clear fonts, words written in script are more difficult to detect and end-of-sentence tokens, such as periods, may be missed. To overcome these problems, other methods rely on the holding pattern of libraries to identify the audience level of the resource (Bernstein, 2006). However, this method requires that all resources are held by at least one library and can only predict the general audience level. For these reasons, readability formulas are not expected to predict the audience level of web-based digital library resources with high performance.

Rather than relying on simplistic syllable and sentence length calculations, essay grading models consider the vocabulary chosen by the author to determine quality. The essay grading and audience level prediction problems share the same inputs (text of the resource), processes (match the terms in a new resource with vocabulary appropriate for each rating), and outputs (numerical score representing the quality of the written work).



These methods typically rely on supervised machine-learning algorithms to match the vocabulary in an ungraded resource with the vocabulary appropriate for each essay grade. As these methods do not rely on the simplistic syntactic and structural characteristics of text, the audience level prediction performance of machine learning methods should exceed the performance of traditional readability formulas for these resources.

### **1.1 Objective**

If a document is targeted toward a population with high audience levels, lower level readers typically become frustrated as they cannot understand the content without frequent trips to a dictionary. To ensure the document text is targeted toward the appropriate reader, readability formulas consider textual attributes, such as syllable counts and sentence length; these attributes are entered into a formula including constants derived from regression analysis to return audience level. While these formulas can be applied to a large number of domains, they suffer from serious limitations due to variability with authors' chosen vocabulary and sentence structures. Web-based resources pose a new set of challenges since they contain headers, images, and tables that do not follow these grammatical rules and fewer words than traditional printed books. In order to overcome these issues, machine learning methods borrowed from the essay grading domain are proposed to predict the audience level of digital library resources with higher performance than readability formulas; these techniques generally rely on matching terms in a resource with a predefined vocabulary appropriate for each grade.

The first objective seeks to improve audience level prediction performance for digital library resources by employing classification methods drawn from the essay

grading domain. The prediction performance of common readability formulas are compared with classification algorithms, including the proposed SVM-based program SVMAUD, with respect to digital library resources.

The second objective seeks to improve the performance of the machine learning methods by adjusting term weight based on the HTML tags in which it occurs. The terms that appear in the title and header text should more succinctly describe the content than text that appears as plain text or captions on the web page. By assigning additional weight to terms appearing in certain tags, the prediction performance by all machine learning methods should improve over assigning all terms the same weight independently of the tags in which they appear.

In addition to the content of the resource, web pages also contain menus, headers, footers, and scripts that appear on every page in the collection regardless of the audience level of the resource. By removing this extraneous information and only using the resource content appropriate for an audience level, the prediction performance should improve since a lower percentage of terms will overlap between adjacent audience levels.

Finally, taking advantage of subject category metadata information stored with each resource can be used to improve prediction performance. By using math resources to predict the audience level for other math resources, the topics covered in each audience level for that subject can be extracted, leading to an increased ability by the different machine learning methods to make fine-grained distinctions between adjacent audience levels.

## 1.2 Background

Previous linguistic research suggests the audience level for documents with readability formulas containing constants derived from regression analysis. Some of the semantic and structural characteristics employed by traditional readability formulas include the length or number of words, phonetic syllables, polysyllables, and number of sentences in the document to measure semantic and syntactic difficulty (Flesch, 1948; McLaughlin, 1969). These methods rely on linear regression models to suggest the relationship between audience levels (output parameter) by using numerical representations of word and sentence characteristics in documents (input parameters). Computer-based systems can predict the appropriate audience level using these readability formulas (McCallum & Peterson, 1982; McLaughlin, 1969), but these systems suffer from serious limitations such as an inability to analyze concepts or terms (George, 2000).

Rather than relying on the syntactic and structural characteristics of text, other methods do not require any textual input from the resources. Recognizing that different types of libraries typically serve a small segment of a population, the audience level can be inferred from the libraries holding the material, with the holding symbols weighted by a numeric code for the library audience type. This method defines “difficulty-level” of comprehending a resource based on the number and audience level of libraries holding the title; these weights are averaged over all bibliographic records for the title in the Online Computer Library Center (OCLC) WorldCat database (Bernstein, 2006; White, 2008). However, this method suffers from serious limitations, including the absolute scales method employed to assign the threshold levels for various audience levels.

Therefore, the OCLC WorldCat method cannot be used to identify the audience level of a random webpage that is not held by any library and is not suitable for this application.

Resources in digital library collections are typically HTML pages that contain headers, footers, menu items, bullet points, tables, and comprise one or two pages; these characteristics can distort the true audience level if readability formulas are used due to the low amount and unconventional structure of text. Rather than employing simplistic textual characteristics, essay grading methods rely on pre-labeled essays to identify vocabulary appropriate for each score and then compare terms in a new essay with the predefined vocabulary to assign a grade. In an overview paper comparing different automated essay grading techniques, three major approaches are identified: statistical methods, natural language processing, and hybrid (combinations of these two) methods, with hybrid methods generally experiencing higher performance (Valenti, Neri, & Cucchiarelli, 2003).

Two hybrid essay grading systems created by the Educational Testing Service (ETS), ETS-I and E-Rater, experience the highest performance among all essay grading systems reviewed (Valenti, Neri, & Cucchiarelli, 2003). ETS-I matches a specific lexicon combined with grammatical rules to an essay grade. After manually creating the training dataset by entering all possible synonyms and metonyms for all key words, classification techniques can predict the human-assigned grade for new essays with an accuracy of 93% (Whittington & Hunt, 1999). E-Rater considers the syntactic variety, organization of ideas, vocabulary chosen by the author, and selected predictive features; by automatically scoring 750,000 GMAT essays, this model achieves a 97% agreement of plus or minus one grade level between human assigned and computer labeled scores

(Burstein, Kukich, Wolff, Chi, & Chodorow, 1998; Burstein, Leacock, & Swartz, 2001; Larkey, 1998). Since the inputs (vocabulary chosen by the author) and outputs (a numerical score representing essay quality) are the same for essay grading and the audience level prediction problem, essay grading methods should improve the audience level prediction performance over readability formulas in the digital library domain.

Recent research areas in vocabulary-based classification methods borrowed from the essay grading domain include sentiment identification, identification of information sources (author, publisher, etc.), biomedical text categorization, and hierarchical categorization of web pages. These methods require positive or negative samples based on the average semantic orientation of the phrases (“bag-of-words”) in the resource. Semantic orientation is represented by frequency of occurrence of the words in the document or as part-of-speech tag occurrences (Bo, Lillian, & Shivakumar, 2002; Turney, 2001). Highly-dimensional text classification for authorship identification is implemented using SVM classifiers, resulting in high precision for low percentages of true positive samples (Diederich, Kindermann, Leopold, & Paass, 2003). For these reasons, machine learning methods borrowed from the essay grading domain are proposed to improve the audience level prediction performance over readability formulas.

### **1.3 Research Overview**

Digital librarians and publishers must be able to verify that the author’s chosen vocabulary challenges but does not frustrate the reader. As manual audience level labeling by experts is time consuming and labor intensive, librarians and publishers can employ computer-based systems that accept resource text as input and output the

audience level for that resource. Even if the resource contains needed information, the individual may elect to continue searching for other resources with simpler vocabulary.

This research seeks to improve the state of the art in audience level identification. First, a number of different readability formulas and classification algorithms drawn from the essay grading area consider textual information to suggest the most appropriate audience level for resources held in a digital library collection. Then, the performance is tuned by adjusting the weight assigned to terms appearing in various HTML tags. Other methods to improve the prediction performance consist of reducing the level of noise in the training data and developing a series of subject-specific classifiers. The labeling of resources with complete and consistent audience level is not only useful for librarians as they catalog resources, but also ensures that written works, as in the case of books and newspapers, maintain reader interest by using appropriate vocabulary.

#### **1.4 Contributions**

This dissertation seeks to provide several contributions in the area of automatic audience level identification. As manual judgments of audience level vary among different human experts, automated methods should be able to suggest audience level more consistently with less effort.

SVMAUD should be able to predict the audience level for digital library resources with high performance. SVMAUD performance is compared with the prediction performance of two readability formulas and three other machine learning methods.

A number of performance tuning methods are then used to improve performance. First, since a digital library mainly holds web pages, term weights are adjusted based on the HTML tags in which they appear. Second, digital library resources contain menus, headings, footings, and other elements common to all pages in a collection regardless of audience level; these common elements should be removed when identifying the audience level for digital library resources. Finally, a series of subject-specific classifiers are developed, so that resources from one subject are used to predict the audience level for resources discussing the same subject. The performance of all four machine learning methods is compared for each of these performance tuning methods.

## **1.5 Dissertation Organization**

This dissertation is composed of eight chapters. Chapter 1 provides an overview of the work that is to be conducted and background of audience level prediction. Chapter 2 reviews readability formulas and classification models. Next, Chapter 3 describes the algorithm used by SVMAUD and the system implementation. Chapter 4 evaluates the performance of different readability formulas and classification methods using a digital library test collection. Chapter 5 describes the prediction performance improvement by adjusting term weight based on the HTML tags in which the term occurs. In Chapter 6, the training data is composed of metadata information instead of the full text of the resource. In Chapter 7, the resources from a home school collection are used to augment the digital library collection to increase subject coverage; a series of six subject-specific classifiers are developed to improve prediction performance. Finally, Chapter 8 concludes the dissertation, providing the summary of results and implications.

## **CHAPTER 2**

### **LITERATURE REVIEW**

Due to the wide range of possible audience level values that can be entered for resources held in library collections, a standard coding scheme for audience level must be followed among all collections. This coding scheme can identify the age of the reader, grade level of the reader, the audience level, or the number of years of formal education required to comprehend the resource. After the standard coding scheme is developed, computer-based audience level identification methods can use textual information to suggest the most appropriate audience level for all resources in the collection with missing or incompatible audience level metadata.

This chapter first reviews traditional readability formulas and then describes classification methods borrowed from the essay grading domain. Readability formulas rely on a variety of different textual characteristics, ranging from average syllables or characters per word to sentence structures. These methods require simple calculations of numerical values that represent text difficulty, and then these numbers are entered into the formula to obtain the audience level of the resource. The next part of this chapter discusses more complex classification methods borrowed from the essay grading domain that can be implemented to improve the audience level prediction performance by comparing the terms in an unlabeled resource with a predefined set of terms appropriate for each audience level. The performance of each of these methods with respect to a number of different applications is also presented.



## 2.1 Readability Formulas

Some common readability formulas consider the length and number of words, phonetic syllables, polysyllables, vocabulary chosen by the author, and sentence length to calculate a score for the semantic and syntactic difficulty of understanding the text. These methods use some of these text characteristics to suggest the most appropriate audience level in combination with constants derived from regression analysis. All of these methods suggest the ease or difficulty of reading a text and some methods can even suggest the most appropriate specific audience level of the reader that should be able to understand yet not be frustrated by the content of the resource. The seven most common readability formulas are the Flesch-Kincaid Reading Age, Dale-Chall Reading Ease Score, Gunning-Fog Index, Simple Measure of Gobbledygook (SMOG), the Spache Readability Formula, Advantage Open Standard for Readability (ATOS), and Lexile Framework for Reading. Each of these methods is described in the following seven sections.

### 2.1.1 Flesch-Kincaid Reading Age

The Flesch-Kincaid Reading Age seeks to determine the most appropriate audience level for a particular document based on the relationship between the syntactic structure and word choice of the document and the relative ease of understanding the content. As humans generally follow grammatical rules in their writing, the predominant focus is on word and sentence difficulties. This method employs a linear regression model to identify the relationship between the textual information in the document, namely syllables, word count, and sentence length, and the difficulty of comprehending the document content.

The Flesch Reading Ease formula suggests the ease or difficulty of understanding a particular text based on average syllables per word and average sentence length. This method calculates an index ranging from 0 to 100, with higher scores indicating that the material is easier to understand (Flesch, 1948). The Flesch Reading Ease is calculated as follows:

$$FRE = 206.876 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right) \quad (2.1)$$

This formula is able to calculate the ease of reading a document's content but is not able to identify the specific grade level or age of the reader.

To correlate the ease of reading a resource with grade level appropriateness, this formula is further updated to calculate the most appropriate reading age of the person that is challenged by the document content while not becoming frustrated. The Flesch Kincaid Reading Age (the audience level for the document) is calculated by the following formula:

$$FKRA = (0.39 \times ASL) + (11.8 \times ASW) - 15.59 \quad (2.2)$$

where ASL is the average sentence length (words in document divided by the number of sentences) and ASW is the average number of syllables per word (total syllables in document divided by the number of words) (Kincaid et al., 1975). A good approximation for syllable counts can be found by counting all of the vowels in the word and counting all vowels that commonly appear to each other, such as e and a in the word eat, as one syllable; eat is counted as only one syllable. On the other hand, the sentence lengths are calculated by taking the sum of all spaces that do not appear next to each other divided by the number of periods, exclamation points, hard returns, semicolons, and question marks. While these methods are not exact, they can give a good approximation without requiring

a dictionary that contains the number of syllables in every word in the English language or requiring a human expert to manually count the words per sentence. In this formula, a score of 5.4 indicates that the content is appropriate for a fifth grade student in the fourth month of the school year. With the advent of computer based systems, the full text from a resource can be plugged into a computer algorithm that suggests the appropriate audience level rather than requiring a person to manually count these text features (McCallum & Peterson, 1982).

The Flesch Reading Ease formula is probably the easiest to understand and most widely used readability formula, since it only requires the inputs of average sentence length and average syllables per word to suggest the most appropriate audience level. For this reason, this formula is used to suggest the audience level for many government documents that need to be understood by a majority of the population. For example, insurance declarations and related paperwork must be found to have a Flesch Reading Ease score of at least 45 before presenting them to the customer (Onecle, 2010).

While this formula is able to easily calculate the appropriate reading age for the resource, it suffers from a variety of drawbacks. This method cannot suggest the audience level for ideas or text without any structure and the cutoff for each reading age is based on previous documents that may not have commonality with the current document. Another serious drawback arises as this method does not consider the vocabulary appropriate for each audience level but, rather, depends on the number of syllables, as more syllables typically indicate harder words that require a higher reading ability to understand. For example, some words, such as “television” contain four syllables yet can be understood by a pre-kindergarten student while other words, such as

“rhinitis,” only contain three syllables yet require a medical professional to define. The average number of syllables per word is not necessarily indicative of the ease or difficulty of understanding the document.

### 2.1.2 Dale-Chall Reading Ease Score

The Dale-Chall Reading Ease Score takes a different approach from the Flesch-Kincaid Reading Age by not only considering the structural and semantic characteristics of the text but also considering the vocabulary chosen by the author. The Dale Common Word List consists of approximately 3,000 words that are typically understood by a fourth grade student; as the proportion of words in this list increases in a document, the resource is considered easier to understand. All of the words in the document are extracted and compared with this list, shown in Appendix A. The end-of-sentence tokens, namely periods, exclamation points, semicolons, hard returns, and question marks, are also counted and then divided by the number of words to determine the average sentence length of the resource. After these values have been calculated, then the following formula can be used to arrive at the Dale-Chall Reading Ease Score:

$$R = 0.1579 * (\textit{Proportion of Words not in Dale Common Word List} * 100) + 0.0496 * \left( \frac{\# \textit{Words}}{\# \textit{Sentences}} \right) (+3.6365 \textit{ if } > 5\% \textit{ of words not in Dale Common Word List}) \quad (2.3)$$

After this portion of the score is calculated, the following heuristic is used to identify the audience level of the resource. For all values under 5, the audience level is less than or equal to Grade 4; for all values between 5 and 9, the audience level is Grade 5-12; all scores between 9 and 9.9 are considered to be college level; and all scores over 10 are graduate and above (Chall & Dale, 1995). By converting the result of this formula to different audience levels, the specific audience level can be generalized from the

numerical result provided by this score. While this method incorporates the vocabulary chosen by the author as part of the formula, the true audience level calculation can still be distorted by the average sentence length parameter. In addition, this formula can be distorted if words are misspelled in the original resource, particularly in the case of scanned resources that must be converted to text using Optical Character Recognition (OCR) due to an exact match being required between the word in the resource and the words appearing on the Dale Common Word List.

### 2.1.3 Gunning-Fog Index

The Gunning-Fog index is another readability formula that can determine the ease or difficulty of comprehending works written in the English language. This index, like the FKRA, suggests the number of years of formal education that the reader must complete in order to understand the text. The readability score is calculated by using the following formula, where complex words are words with three or more syllables (Gunning, 1952):

$$\text{Gunning - Fog Index} = 0.4 * \left( \left( \frac{\# \text{ words}}{\# \text{ sentences}} \right) + 100 * \left( \frac{\# \text{ complex words}}{\# \text{ words}} \right) \right) \quad (2.4)$$

In this formula, a calculated score of 12 indicates that a high school senior should be able to understand the text, while a score of 8 indicates that the text is appropriate for an eighth grade student. The original formula counts clauses as well since most people see a clause as a complete thought; however, in later versions of the formula, the clause calculation is dropped as it must be done manually. This formula does not consider the vocabulary appropriate for a particular grade level but, rather, only the characteristics of words and sentence structures in the document. This method can result in higher grade levels required to understand words that a kindergartner can comprehend, such as broccoli or television, while some shorter words may be more difficult to understand.

The original formula requires only samples of text since computers have not been available to automate the calculation; however, the full text of digital resources available today can be used as input to determine the most appropriate audience level.

#### **2.1.4 Simple Measure of Gobbledygook (SMOG)**

The Simple Measure of Gobbledygook (SMOG) is another readability formula that suggests the most appropriate audience level for any textual work. This formula is easier to calculate than the Gunning-Fog index while increasing accuracy over the Flesch-Kincaid Reading Age. To make the calculation even easier, representative samples of text can be extracted from the resource rather than using the entire resource as input to the formula. The calculation requires the extraction of a number of sentences from the document, with at least ten from the beginning third, ten from the middle third, and another ten sentences from the remaining third of the document. Within each set of representative sentences, the number of polysyllabic words, or words with three or more syllables, is counted. The grade level is calculated by entering these values into the following formula:

$$Grade\ Level = 1.043 * \sqrt{30 * \frac{\text{number of polysyllabic words}}{\text{number of words}}} + 3.1291 \quad (2.5)$$

This formula calculates the grade level of the student that can comprehend the text. This formula correlates with readers with complete comprehension of test materials within 1.5159 grade levels at the 0.985 level (McLaughlin, 1969). With the introduction of computer based readability methods, especially in the case of digital libraries, where resources are stored in digital format, this formula can use the entire body of text as the representative sample and then easily count the number of polysyllabic words to predict

the most appropriate audience level of the resource. However, this formula still suffers from the same problems as other readability formulas, namely that the resource must be well edited and the actual words chosen by the author are not considered.

### 2.1.5 Spache Readability Formula

The Spache Readability Formula relies on a different set of input parameters than other readability formulas. Rather than requiring the calculation of syllables contained in each word, this formula relies on a previously determined set of 769 words appropriate for everyday reading; this list is compared with the words in the document chosen by the author. This formula does not rely on syllables per word to measure difficulty but, rather, relies on the presence or absence of unfamiliar words for each grade level. There are two different formulas, the original and then a revised one with different coefficients but the same required inputs. The audience level for a resource can be calculated using one of the following formulas:

Original Spache Reading Level Formula:

$$\begin{aligned} & \textit{Original Spache Reading Level} = \\ & (0.141 * \textit{Average Sentence Length}) + (0.086 * \textit{Unique Unfamiliar Words}) + 0.839 \end{aligned} \quad (2.6)$$

Revised Spache Reading Level Formula:

$$\begin{aligned} & \textit{Revised Spache Reading Level} = \\ & (0.121 * \textit{Average Sentence Length}) + (0.082 * \textit{Unique Unfamiliar Words}) + 0.659 \end{aligned} \quad (2.7)$$

In these formulas, the average sentence length is the number of words divided by the number of end-of-sentence tokens and unique unfamiliar words is the count of the words in the document that are not contained in the everyday word list, with each word counted only once independent of the number of times it appears in the resource (Spache,

1953). This formula results in the number of years of formal education that the student should complete in order to comprehend the text in a given resource. Two major issues contribute to the inaccuracy of grade level calculations when using this formula, namely the generation of the list of common words and the average sentence length calculations. In a later article, the list of 769 common words is criticized as not being complete and containing many difficult words, such as quarter and reason, while not including many other common words that are present in everyday life (Stone, 1956). The choice of words on the familiar list can severely impact the score depending on the word choice by the author, as in the case of synonyms; for example, if the word car appears on the common words list but the author uses auto instead, auto is counted as an unfamiliar word and serves to increase the audience level of the resource while, in actual fact, the difficulty should be lower. Similar to other formulas that rely on counts of end-of-sentence punctuation marks, poorly edited resources may be missing periods and question marks that serve to unduly increase the audience level necessary to understand the resource. This formula is simpler to implement in a computer based system when compared to other formulas but suffers from serious drawbacks, especially in the choice of words that appear on the familiar words list.



### **2.1.6 Advantage Open Standard for Readability (ATOS)**

The Advantage Open Standard for Readability (ATOS) is a readability formula that can be applied to different written works to match students with appropriate resources. This readability formula is used by Renaissance Learning to suggest resources to students based on reading ability when using two other Renaissance Learning products, namely the Standardized Test for Assessment of Reading (STAR) and Accelerated Reader (AR). STAR attempts to identify the most appropriate audience level for each student, while AR measures the ability of students to comprehend different passages of text. These two tests are designed to enable students to advance their reading skills as well as match students to books that both challenge and inform the reader. These two different tests are computer based and typically require less than 10 minutes to complete. STAR requires the student to complete a number of sentences by choosing the most appropriate word from a list of provided words; after the student completes the test, he or she is presented with a report measuring such areas as reading ability and grade equivalency. AR, on the other hand, provides a list of titles geared toward the individual student's reading ability that the student can check out from the library; after the student finishes reading the book, he or she then completes a series of reading comprehension questions based on the book and is given points based on the difficulty of the book as well as the number of correct answers (Milone, 2010).

To combat the problem of unique students attaining unique audience levels, the Advantage Open Standard for Readability seeks to match students with the most appropriate reading materials independent of grade level. This formula is designed to easily identify the readability of new books, be understandable to teachers, and appear

instructionally sound to educators. Three different readability formulas are used to measure the reading difficulty of different written works: the ATOS for Text Readability Formula to suggest the readability of short passages; the ATOS Grade Level to convert ATOS for Text score into grade level equivalents; and two alternate formulas of ATOS for Books to convert the grade level scale to either a 100 point scale (similar to Flesch Reading Ease) or a 2000 point scale (similar to Lexiles). The following table summarizes the readability predictors under consideration for possible inclusion in the audience level prediction model.

**Table 2.1** ATOS Readability Predictors

Predictor Abbreviation	Predictor Description
AvgChar	Average number of characters per word
SDChar	Standard deviation of the number of characters per word in complete sentences
AvgWords	Average words per sentence
FamWords	Relative frequency of familiar words to total words found in complete sentences. Familiar words are easy words that are commonly found in written works
AvgGrade	Average grade level for words found on a previously graded vocabulary list. These words are categorized by the audience level of the person that should be able to understand the word.
AvgGrad100	Average grade level for words found on graded vocabulary list excluding the top 100 most frequent words in the Advantage Learning Systems corpus.
SDGrade	Standard deviation for the average grade level for words found on the previously categorized vocabulary list.
AvgSyll	Average number of syllables per word referenced in a dictionary of 69,794 words
SDSyll	Standard deviation of the number of syllables per word based on the same dictionary of 69,794 words
Mono	Count of monosyllabic words divided by the total number of words in the resource
Poly	Count of polysyllabic words divided by the total number of words in the resource

Various combinations of these different factors are combined with coefficients derived from regression analysis to determine the best predictors of audience level for resources. This work provides three different audience level formulas as detailed in the following paragraphs.

The ATOS for Text Readability Formula is based on the number of words per sentence, the average grade level of words, and the average number of characters per word. This formula labels the audience level of the resource based on the scaling method developed by Rasch (1980). The ATOS Rasch Difficulty Formula (ATOSRD) is as follows:

$$ATOSRD = -8.54 + 1.95 * \ln(AvgWords) + 0.46 * AvgGrade100 + 1.74 * \ln(AvgChar) \quad (2.8)$$

In this formula, AvgWords is the average number of words per sentence, AvgGrade100 is the average grade level of words found on a previously graded category listing excluding the most popular 100 words found in the corpus, and AvgChar represents the average number of characters per word.

Now that the reading difficulty is calculated for all resources, the next step is to convert this value to a grade level equivalent. This formula is based on a study with a database containing over 950,000 Accelerated Reader (AR) records from more than 30,000 students reading and testing on different books. This dataset is then used to plot the average ATOSRD values against the average audience level and a quadratic function is fit to the data points. This study results in a formula to convert the ATOSRD into a grade level equivalent.

$$ATOS \text{ Grade Level} = 5.86 + 2.86 * ATOSRD + 0.32 * ATOSRD^2 \quad (2.9)$$

Another study finds that book length is an important predictor for the reading difficulty of a resource. This study samples three million AR quiz records captured during Fall 2008 to determine the relationship between book length and difficulty of comprehending the content of the book. Students are able to correctly answer 87% of questions on books containing less than 500 words and 84% for books containing 501-5,000 words. This number decreases until students are able to correctly answer only 65% of the reading comprehension questions for books containing 250,001 to 500,000 words. As books become longer, the reading comprehension generally decreases. Therefore, an additional formula suggests the relationship between the length of the book and its difficulty. The first formula calculates the Book Length Grade Level (BLGL) for Books with over 500 words as follows:

$$BLGL \text{ for Books containing over 500 words} = 0.68 * \ln(\text{Book Length}) - 1.87 \quad (2.10)$$

This formula can now be used to calculate the ATOS for Books Readability Formula as follows:

$$ATOS \text{ for Books Readability Formula} =$$

$$ATOS \text{ Wght} * ATOS + BLGL \text{ Wght} * BLGL \text{ for Books containing over 500 words} \quad (2.11)$$

In this formula, ATOS Wght and BLGL Wght are chosen based on the number of words in the book according to the following table:

**Table 2.2** ATOS and Book Length Grade Level Weights for Long Books

Number of Words	ATOS for Text Weight	Book Length Grade Level Weight
500-4,999	0.50	0.50
5,000-49,999	0.60	0.40
50,000-99,999	0.80	0.20
100,000-249,999	0.85	0.15
250,000 and up	0.90	0.10

This method attempts to adjust the grade level formula relative to the number of words contained in the book as longer books are found to result in a lower level of reading comprehension (Milone, 2010).

Out of these three different formulas, the ATOS Grade Level is most appropriate to determine the audience level for a resource for digital library collections. While this formula can be applied to determine the grade level for books that contain a large amount of text, this formula cannot predict the grade level of web pages that contain little text and do not follow normal grammatical and sentence conventions. For example, one part of the formula relies on the average number of words per sentence and is found by dividing the number of end-of-sentence tokens, such as periods, exclamation points, and question marks, by the number of words in the text; web pages that contain sentence fragments, bullet points, tables, and figures typically contain fewer punctuation marks relative to word count, artificially increasing the reading difficulty. This formula is based on regression analysis performed on a test collection with a limited number of resources; a new resource may not share complete similarity with the resources chosen to suggest the constants used in this formula. In the case of web pages that contain less text and may not follow normal sentence and grammatical conventions of written English, the formula is believed to perform poorly. Lastly, the vocabulary chosen by the author is not considered in this formula. This formula relies on word characteristics to suggest the most appropriate audience level; longer words or words with higher syllable counts are not necessarily indicative of more difficult words found in higher audience levels.

### 2.1.7 Lexile Framework for Reading

The Lexile Framework for Reading takes a different approach than other readability formulas by matching readers to resources independent of grade level. As different states require different curriculum standards necessary to pass each grade, this formula does not pinpoint a specific grade level for each resource but, rather, calculates the Lexile for each person and then that person should find resources labeled with approximately the same Lexile. The Lexile is an indicator of the reading ability of the student rather than the grade level necessary to understand the written work. The scale generally ranges between 0L and 1700L, but it is possible to have scores outside this range. In these cases, scores below 0L indicate beginning readers while scores over 1700L indicate advanced readers.

The Lexile Measure is a measure of the individual's reading ability or the ease of understanding text; this number is then followed by an "L" for Lexile. For example, an individual with a Lexile level of 500 would be given 500L as the reading ability. This Lexile measure for the individual reader must first be obtained by completing a reading comprehension test offered by a number of different companies, including McGraw Hill and Scholastic, rather than the student being enrolled in a particular grade level in the American educational system. After the test is completed by a large number of students, the Lexile score can then be calculated by using the following formula:

$$\text{Lexile Score} = H + \left( \frac{180 + S^2}{1040} \right) * \log \left( \frac{R}{L-R} \right) \quad (2.12)$$

In this formula,  $L$  is the number of questions on the test,  $H$  is the average slice Lexile,  $S$  is the Lexile Standard Deviation, and  $R$  is the number of the reader's correct answers on the test. This Lexile measure for a reader indicates the reading difficulty of text where the

reader will succeed on 75% of the slices. The 75% measure is the point at which a person's reading ability peaks, or the point at which he or she is challenged to understand the text but does not become frustrated (Stenner, 1992).

The Lexile measure must also be calculated for every document in the corpus to identify the reader that is able to understand yet be challenged by that resource. In order to Lexile a book or other written work, the book is divided into slices of 125 to 140 words where each slice contains complete sentences or paragraphs. The slices are then calibrated to the Lexile scale by using the following formula:

$$\text{Lexile Score} = 582 + 1768 * SL_i + 386 * WF_i \text{ Lexiles} \quad (2.13)$$

In this readability formula,  $SL_i$  is the log of the mean sentence length, or the average number of words per sentence.  $WF_i$  is the average number of times that the word appears in a work containing five million words as found in the Word Frequency Book (Carroll, Davies, & Richmond, 1971). In this formula, more common words are more likely to appear in a work; these words indicate a high level of familiarity for all readers, implying that they are understood by readers with lower reading abilities. This measure indicates the point at which the reader can comprehend 75% of the book's slices.

Now that the Lexile measure is calculated for both the reader and the documents in the corpus, the reader can now find resources that are plus or minus 100L but still within his or her comfort level, where the reader is both challenged yet not frustrated by the resource content (Wright & Stenner, 1998). The important distinction between Lexile measures and other readability formulas is the Lexile measure does not directly correspond to the grade level of a reader that can understand the resource but, rather, levels the playing field for all students that complete the test. However, there can be a

rough correspondence to grade levels as higher Lexile measures indicate a higher degree of reading difficulty. The following table shows the rough correspondence between grade level and Lexile score.

**Table 2.3** Lexile to Grade Level Correspondence

Grade Level	Reader Measure, mid year
1	Up to 300L
2	140L to 500L
3	330L to 700L
4	445L to 810L
5	565L to 910L
6	665L to 1000L
7	735L to 1065L
8	805L to 1100L
9	855L to 1165L
10	905L to 1195L
11 – 12	940L to 1210L

In common practice, the grade level correspondence is not used to ensure that all readers are on a level playing field and readers do not feel disparaged for having lower reading abilities than their peers. There is a lot of overlap between different grade levels as the audience level appropriateness of different books is not typically limited to a single grade level (Lexile.com, 2010). The Lexile measure does not consider the quality or content of the book but, rather, the reading difficulty and is a good indicator of the student most likely to comprehend the book's content.

While this method may be able to match users to appropriate reading resources, the Lexile measure suffers from a number of drawbacks. Teachers and other educators in the United States think in terms of first grade and second grade and not in terms of Lexile measures, leading to a difficulty for teachers to find grade level appropriate resources for



use in their classrooms. Like other readability formulas that rely on the syntactic structure of text, web pages that do not follow the conventional grammatical and sentence structures of books would probably be given a higher Lexile measure than warranted. The slices or samples of text that are chosen for input to the Lexile formula can influence the calculated score upwards of 450L simply due to variability in sampling sentences; similarly, the Lexile score can be influenced 100L by eliminating some end of sentence tokens (MediaMetrics, 2007). While this measure is successful at Lexiling books that contain text organized into sentences, the performance with respect to Lexiling web pages is expected to suffer as these files contain figures, tables, bullet points, URLs, and missing end-of-sentence tokens that can impact the Lexile score calculation.

### **2.1.8 Readability Formulas Summary**

Readability formulas are based on the idea that the grade level or reading ease of a text can be calculated with a reasonable degree of accuracy based on the syntactic and semantic structure of the text. For example, sentences that contain a large number of words are typically harder to understand, requiring a higher reading ability to comprehend, than text that contains a smaller number of words per sentence. Similarly, words that contain a larger number of syllables are considered more difficult to understand than words containing fewer syllables. The following table on the next page summarizes the calculation of these readability formulas:

**Table 2.4** Readability Formulas Summary

<b>Readability Formula</b>	<b>Formula</b>	<b>Meaning of Formula Result</b>
Flesch Reading Ease	$\text{Reading Ease} = 206.876 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)$	Score 0 to 100; higher numbers are easier to read
Flesch-Kincaid Reading Age	$\text{Grade Level} = (0.39 \times \text{ASL}) + (11.8 \times \text{ASW}) - 15.59$	Grade level of student required to understand text
Dale-Chall Reading Ease Score	$R = 0.1579 \left( \frac{\text{Proportion of Words Not in Dale Common Word List} * 100}{\text{Sentences}} \right) + 0.0496 \left( \frac{\text{Total Words}}{\text{Sentences}} \right)$ (Add 3.6365 if greater than 5% of words not in Dale List)	Grade level of student required to understand text
Gunning-Fog Index	$\text{Grade Level} = 0.4 * ((\# \text{ words}) / \# \text{ sentences}) + 100 * ((\text{complex words}) / \text{words})$	Grade level of student required to understand text
Simple Measure of Gobbledygook (SMOG)	$\text{Grade Level} = 1.043 * \sqrt{30 * \frac{\# \text{ of polysyllabic words}}{\# \text{ of words}}} + 3.1291$	Grade level of student required to understand text
Revised Spache Readability Formula	$\text{Grade Level} = (0.121 * \text{Average Sentence Length}) + (0.082 * \text{Unique Unfamiliar Words}) + 0.659$	Grade level of student required to understand text
ATOS	$\text{ATOSRD} = -8.54 + 1.95 * \ln(\text{AvgWords}) + 0.46 * \text{AvgGrade100} + 1.74 * \ln(\text{AvgChar})$ $\text{ATOS Grade Level} = 5.86 + 2.86 * \text{ATOSRD} + 0.32 * \text{ATOSRD}^2$	Grade level of student required to understand text
Lexile Framework for Reading	$\text{Lexile Measure} = 582 + 1768 * SL_i + 386 * WF_i \text{ Lexiles}$	Ease or difficulty of understanding text; requires test to determine reader's Lexile

In the case of web pages that do not follow the syntactic and semantic conventions necessary for usage by these formulas, the prediction performance of these formulas is expected to suffer. These methods generally do not consider the vocabulary chosen by the author, only the characteristics of text. By matching the vocabulary chosen by the author with a pre-defined vocabulary appropriate for each audience level, the

audience level prediction performance should improve over traditional readability formulas.

## 2.2 Machine Learning Methods

This section reviews different machine learning techniques that can be employed to automatically predict the audience level of a resource based on the text contained in that resource. These methods do not rely on the formatting of the text within the document or even word characteristics to make audience level predictions. Rather, they attempt to identify the most appropriate audience level based on the vocabulary contained in the document or resource. In the case of web pages that are not required to follow the same structure or layout, these methods can be employed to suggest the audience level without depending on word characteristics or sentence structures required by readability formulas. These methods generally fall into one of two categories – supervised, used in classification, and unsupervised, used in clustering. Supervised machine learning methods require a dataset with pre-labeled training samples; each unlabeled resource is then labeled with one of these predefined categories. On the other hand, unsupervised machine learning seeks to place documents that share some similarity close to each other; these methods can provide a visualization of complex data but cannot label documents with predefined categories such as audience levels.

Collins-Thompson and Callan (2005) propose a language modeling approach that relies on a previously defined lexicon for each grade level that can then be used to predict the audience level of web documents. The results show that deriving individual text-based grade level models to predict the appropriate audience level perform much higher

than readability formulas; in fact, the correlation between human-expert assigned labels and Naïve Bayes machine-learning labels for web-based resources is found to be 0.69 for grades one through six and 0.79 for grades one through twelve (Collins-Thompson and Callan, 2005). In addition, the overlap of “bag-of-words” across grade levels is considered in the model, as the complexity of certain words can contribute to difficulty in reading for more than one grade level. However, this approach suffers from a serious drawback whereby resources may not contain many of the words in the pre-defined lexicon, especially in the case of rare or highly specialized words. These types of methods do not rely on predefined grammatical structures to determine the most appropriate audience level but, rather, on the terms that appear in the full text of the resource.

Advances in computational language technologies attempt to understand audience level identification as a function of text coherence or cohesion (McNamara et al., 2004). Cohesion is defined as the explicit characteristic features such as words, phrases, and sentences that help readers to understand and connect the ideas present in the text. Coherence, on the other hand, describes the characteristics of the reader’s mental representation of the text in which ideas, concepts, or subjects are linked together. This model does not perform well when the text is poorly written or unstructured. The Coh-Metrix system maintained at University of Memphis lists sixty cohesion parameters by combining the readability methods and other computational linguistic methods to measure the text cohesion using Latent Semantic Analysis. The Coh-Metrix user studies indicate that readability formulas perform well for a low-coherence population while

classifiers perform well for expert-level resources with high cohesion (McNamara et al., 2004).

The identification of audience level for an unlabeled resource can be recast as a machine learning problem, whereby each audience level is a class and all documents with human-expert entered audience level are used for training samples. These methods are successfully applied in other areas, such as identifying the quality of an essay based on the vocabulary and sentence structure chosen by the author. By borrowing methods from the essay grading domain, the audience level prediction performance of web-based resources should be improved. The Si & Callan method incorporates the document content to measure the readability metric and approaches the audience level prediction problem as a traditional text classification system (2001). Combining 91% of the unigram language model to represent the term-document linear relationships and 9% of the sentence length distribution model, web documents are labeled with audience level with an accuracy of 75%. Another study proposes an automated audience level detection system for search engine user queries employing SVM to incorporate both syntactic features and frequency of  $n$ -word sequences ( $n$  consecutive words); expert judged datasets are evaluated and an overall accuracy of 83% is achieved using kernel based SVM classifiers (Liu et al., 2004).

While these programs demonstrate that algorithms borrowed from other areas can be successfully applied to predict the audience level for unlabeled resources, these methods have not been applied to label digital library resources. This section reviews five different potential machine learning algorithms and evaluates the ability of each to

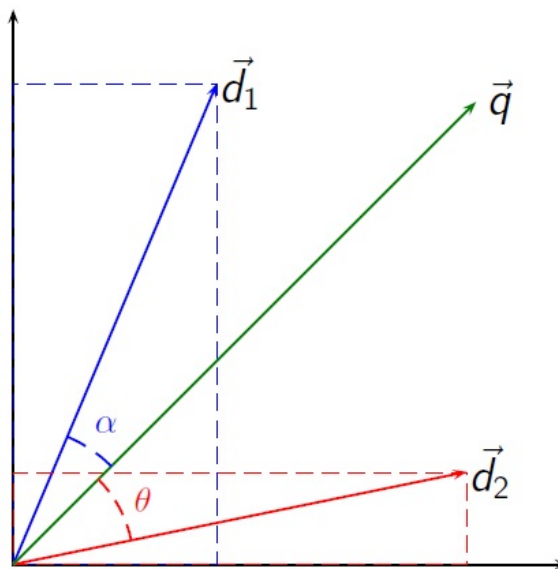
correctly predict the audience level based on the textual information contained in digital library resources.

### 2.2.1 Cosine

Perhaps the simplest text classification algorithm in use today is cosine. Cosine is based on the classic vector space model that represents documents, queries, or other textual information in a vector space (Salton, Wong, & Yang, 1975). In this model, documents and queries can be represented by the following vector:

$$d_j = (w_{1,j}, w_{2,j}, \dots, w_{i,j}) \quad (2.14)$$

where  $w_{i,j}$  refers to the weight of term  $i$  in document  $j$ . These terms can be represented by keywords, individual terms, or key phrases depending on the application of the model. A pair of documents with a smaller cosine separating them is considered to be more similar than those with a larger cosine, as represented in the following diagram:



**Figure 2.1** Graphical representation of the Vector Space Model.

In this diagram, the cosine can be measured between the query and documents in the collection and the document with a smaller cosine when compared to the original query is considered to be more relevant to the user's query. The cosine between the vectors is calculated by using the following formula:

$$\cos\theta = \frac{a_2 \cdot q}{\|a_2\| \|q\|} \quad (2.15)$$

In order to calculate the term weight, several weighting schemes may be used. The weighting scheme proposed in the Vector Space Model is term frequency – inverse document frequency (TF-IDF), as described by Salton, Wong, & Yang (1975), and probably the most popular weighting scheme currently in use. The weight of each term in each document can be calculated by using the following formula:

$$w_{t,d} = tf_{t,d} * \log \frac{|D|}{|\{d \in D \mid t \in d\}|} \quad (2.16)$$

The parameter  $tf_{t,d}$  is simply the count of each term in each document. For example, if a document contains five instances of the term “test,” then the term frequency is five for that term for the document. The inverse document frequency measures the relative importance of each term for each document in the collection. The  $|D|$  refers to the number of documents in the collection while  $|\{d \in D \mid t \in d\}|$  refers to the number of documents in the collection that contain the term. As more documents contain the term, the discrimination value of the term decreases; that term is consequently given lower weight when calculating the similarity between documents and queries or between documents and classes. After the term weights for each document under consideration are calculated, the cosine separating two document vectors is calculated using the following formula.

$$sim(d_j, q) = \frac{d_j \cdot q}{\|d_j\| \|q\|} = \frac{\sum_{i=1}^N w_{i,j} * w_{i,q}}{\sqrt{\sum_{i=1}^N (w_{i,j})^2} * \sqrt{\sum_{i=1}^N (w_{i,q})^2}} \quad (2.17)$$

The similarity can be calculated using this formula by comparing all pairs of documents in the collection, the similarity of each document to a query, or the cosine between the pre-labeled class documents and a new document. The cosine between document and query or document and document can be calculated and, then, the pair with the lowest cosine is considered most similar.

One of the most common applications of cosine is in information retrieval, by comparing the terms entered in a query with textual document content to identify relevant resources in the collection. The bag-of-words model does not consider the relationships between terms in a document but, rather, the number of occurrences of different terms in the entire document. One method suggests the extraction of concepts from WordNet that are related to the query and then combining these terms in different ways to form a new representative vector for the query that does not require exact term matching between documents and queries. This study raises the macro-averaged F1, or F-Measure where precision and recall are weighted equally, from 0.649 to 0.714 for the Reuters collection and 0.667 to 0.719 for twenty newsgroups on a variety of topics (Elberrichi, Rahmoun, & Bentaalah, 2008). Cosine can be used to effectively match keywords entered by users with documents in a collection.

In a related application, Will et al. propose a recommendation system for digital libraries (2009). One component of this system is a content based recommendation implementation that relies on the cosine model to identify resources that are similar to the one being viewed. In this model, the cosine between each resource and all other resources in the collection is calculated and stored offline. Then, as the user browses



around the site, documents with similar content to the one being viewed are presented to the user for consideration. This cosine-based recommendation system presents relevant resources to the user 72% of the time.

Cosine can also be used in virus detection in order to separate files that are modified to contain a virus from legitimate files without any modifications. As virus creators continually improve their techniques by changing the virus signature as each infected file is created, traditional signature detection based techniques may fail. Rather than requiring static matches in the infected file with known virus signatures, the cosine similarity measure can be used to compare two files based on analysis of the portable executable (PE) format of files that contain the virus with ones that are not infected. By comparing the code using the cosine similarity measure, similarities within the two files can be identified even if the signature is changed. In the test dataset, the changed code is identified in five out of ten code samples, defined as a cosine similarity threshold less than 0.97; the lowest similarity value is over 0.85, indicating that this measure can be used to identify variants of existing viruses (Karnik, Goswami, & Guha, 2007). By comparing the original file instructions with the modified instructions that possibly contains a virus, potential viruses can be identified. The cosine similarity measure can be used to classify a new code sample as a possible virus or safe for installation and usage.

This model is easy to understand and simple to calculate the similarities between pairs of documents, documents and queries, or documents and pre-defined classes. This method allows for partial matching of queries and documents but requires that the term, and not its synonym, appear in the document. The similarity between all document pairs can be calculated offline and stored for later retrieval while a query requires calculation

on the fly. However, a major problem exists as the cosine separating a pair of documents may be small but the documents may actually be far apart in the vector space. For this reason, Euclidean distance, or the straight-line distance between two points, can be used to calculate the similarity by using the following formula:

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \quad (2.18)$$

The distance between the two points is simply the square root of the sum of the squares of the difference between point  $p$  ( $p_1, p_2$ ) and point  $q$  ( $q_1, q_2$ ). In this way, if two documents have a small cosine separating them but are actually far apart in the vector space, the distance can be calculated and the two documents are not considered as being similar in content. However, both calculations do not consider the relationship between terms contained within the same document and may perform poorly under certain conditions where this relationship exists.

### 2.2.2 Naïve Bayes

In a different vein, the Naïve Bayes classification model is based on Bayes theorem from statistics, which states that the presence or absence of a feature is unaffected by the presence or absence of any other feature. In classification, this means that the presence or absence of a word is unaffected by the presence or absence of any other word, and words are assumed to appear randomly throughout the document. In most cases, the Bayes decision rule is based on the maximum likelihood that a document belongs to a particular class. After all training documents are placed into their respective class or classes, the set of terms that compose the documents in each class is used as the training dataset. Assuming that all terms in the document are independent of each other and appear randomly throughout the document and class, the probability of the term

appearing in the document is simply the sum of all occurrences of the term divided by the total terms in the document. However, Naïve Bayes classifiers tend to perform poorly in certain situations where terms are not independent of each other (Zhang, 2004). The probability that a document  $d$  appears in class  $c$  can be written by:

$$P(d \text{ in } c) = P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (2.19)$$

where  $P(t_k|c)$  is the conditional probability of term  $t_k$  occurring in class  $c$ . The  $P(t_k|c)$  is a measure of the individual term contribution toward the document belonging in the correct class. In text classification, the class to which the document is most likely to belong, or the maximum a posteriori (MAP), is calculated using the following formula:

$$c_{map} = \operatorname{argmax}_{c \in C} \hat{P}(c|d) = \operatorname{argmax}_{c \in C} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c) \quad (2.20)$$

As the true values of  $\hat{P}(t_k|c)$  and  $\hat{P}(c)$  are unknown, these values are estimated from the training dataset. To simplify the problem based on the formula  $\log(xy) = \log(x) + \log(y)$ , the following formula is obtained:

$$c_{map} = \operatorname{argmax}_{c \in C} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)] \quad (2.21)$$

Classes that contain more terms that match the terms in the unlabeled document are more likely to be the correct class than those classes that do not contain many of the same terms. The probabilities of  $\hat{P}(c)$  and  $\hat{P}(t_k|c)$  must now be estimated to solve the formula; the document is assigned to the class with the highest probability. The approximate  $\hat{P}(c)$  can be estimated using the following formula:

$$\hat{P}(c) = \frac{N_c}{N'} \quad (2.22)$$

In this formula,  $N_c$  is the number of documents in the class and  $N'$  is the total number of documents in the training dataset. The  $\hat{P}(t_k|c)$  can now be estimated by using the following formula:

$$\hat{P}(t_k|c) = \frac{T_{ct}}{\sum_{t' \in v} T_{ct'}} \quad (2.23)$$

$T_{ct}$  is the number of occurrences of the term  $t$  in class  $c$  in the training dataset while  $t'$  is the count of terms that are not  $t$  in the class. To eliminate zeroes in the above equation, one is added to each count, yielding:

$$\hat{P}(t_k|c) = \frac{T_{ct}+1}{\sum_{t' \in v} (T_{ct'}+1)} = \frac{T_{ct}+1}{(\sum_{t' \in v} T_{ct'})+B} \quad (2.24)$$

where  $B = |V|$ , the number of terms in the vocabulary (Chessman & Stutz, 1996).

The Naïve Bayes classification model is used in a large variety of systems with varying success rates. One application seeks to measure the readability of health related information to determine whether the reading difficulty is appropriate for the audience. Rather than relying on word and sentence characteristics as required by readability formulas such as the Flesch Kincaid Reading Age, the Naïve Bayes classifier labels documents with one of three reading difficulty levels. This method results in 98% accuracy for 250 health documents; by using this method, 70%-90% of resources held in the test collection are appropriate for intermediate readers (Miller et al., 2007; Leroy et al., 2008).

In another study, the concept of code readability is explored. When writing code, the programmer should not only use comments to explain each part of the program but also use descriptive variables and spacing. A group of human raters identify a set of code features that contribute to code readability and then, based on the presence or absence of these features, the readability of the code is assessed. The automated measure seeks to determine whether the code is more readable or less readable, resulting in a binary Naïve Bayes classification model. In fact, the authors report that the usage of comments to

explain code is less important than placing blank lines between different code segments (Buse & Weimer, 2008).

The Naïve Bayes machine learning method can also be effectively used to detect unwanted commercial email, otherwise known as spam. This research measures the performance of the Naïve Bayes method when the inputs, including lemmatization, training corpus size, and stop word lists, are modified. When using both lemmatization and stop word lists to modify the input to the Naïve Bayes model, the performance accuracy is reported at 99.99%; however, the ability to find spam is only 63% accurate in a test collection containing 3,000 messages (Androutsopoulos, et al., 2000).

Another study considers the ability of the Naïve Bayesian approach to match users with audience-level appropriate documents. As manually obtaining the audience level for every document in a large collection is not feasible, this study seeks to apply a Naïve Bayes classifier incorporating additional language modeling to suggest the audience level of web-based documents. This study places training documents into twelve different categories for grades first through twelfth and six different categories for grades first through sixth. By creating a model that represents the terms appropriate for each grade level in the training dataset, an unlabeled document can be labeled with one of these predefined grade levels. Rather than relying on a simplistic Naïve Bayes model whereby all terms are weighted based on their word probabilities in the training dataset, additional tuning functions are performed on the training dataset to reduce the effect of words occurring with high frequencies. For example, stop words tend to occur more frequently at lower audience levels but, on the other hand, also reduce the importance of less-frequently occurring words (Collins-Thompson & Callan, 2004). For this reason, the

simple Good-Turing method is used to smooth the word frequency data in each class and reduce the importance of frequently occurring words that may distort the audience level prediction; similarly, words with low frequency in the training dataset are given more weight (Gale, 1995). The findings show a root mean squared error of between one and two grades for nine out of twelve grades and a correlation between human-expert identified and machine-suggested audience levels of 0.69 for grades one through six and 0.79 for grades one through twelve (Collins-Thompson & Callan, 2005).

While the Naïve Bayes probability model is simple to understand and implement, the base assumption, that all terms and documents are independent of each other, is inherently flawed as words appear in mostly the same order or certain words appear only in certain parts of the document. Its effectiveness at labeling new documents with the correct class varies widely as it depends on the exact term appearing in the document (it does not consider synonyms) and all words in a document share some relationship with each other.

### **2.2.3 Clustering**

Clustering takes a different approach than other machine learning methods by not labeling documents with a pre-determined set of class names but, rather, by grouping documents or items with similar content together. Clustering methods typically require a set number of clusters to be determined before the process can begin but that optimal number is not always known in advance and can be difficult to estimate. However, if the main concepts or similarities between documents are unknown, clustering can be a good place to begin. These algorithms typically take one of many different forms. Perhaps the most common clustering algorithm is that of hierarchical clustering, whereby all

documents are placed into one cluster and then split into smaller clusters (“top down”), or each document is placed into an individual cluster and then the clusters are combined to form larger clusters (“bottom up”). Partitional algorithms are able to determine all clusters in one pass rather than joining or dividing clusters and, therefore, are more efficient than hierarchical ones. Density based algorithms draw clusters in irregular shapes based on a certain threshold of items that must exist in each cluster. While these algorithms function in different ways, almost all clustering is based on distances, namely the distance between a cluster and a document. Documents that are placed closest together share the highest degree of similarity with each other. There are two main categories of clustering algorithms – hierarchical and partitional methods.

Hierarchical methods either start with one cluster which is split into successively smaller clusters or with all documents in individual clusters that are then joined together to form a predetermined number of clusters. After each document is placed within the clustering space based on some comparison metric, the distance between all pairs of clusters can be calculated using one of the following methods. The Euclidean distance is the straight-line distance between any two points and can be calculated by the following formula:

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \quad (2.25)$$

where  $a$  and  $b$  are two different points. Another formula that can be used to calculate the distance is the squared Euclidean distance, which is shown on the next page:

$$\|a - b\|_2^2 = \sum_i (a_i - b_i)^2 \quad (2.26)$$

Another common distance measure is the Manhattan distance, which is calculated by the following formula:

$$\|a - b\|_1 = \sum_i |a_i - b_i| \quad (2.27)$$

Finally, cosine similarity can also be used to calculate the cosine between two different points.

$$\cos\theta = \cos^{-1} \frac{a \cdot b}{\|a\| \|b\|} \quad (2.28)$$

However, cosine does not consider the distance between different points but, rather, the angle between two points so the points could be far apart in the clustering space yet have a cosine between them of zero. The advantage of the hierarchical clustering algorithm is that any distance measure can be used and the clustering algorithm can be stopped either when a certain number of clusters is reached or the distance between new clusters is sufficiently large. After the first set of clusters are created, the larger clusters are linked together or split apart to form new clusters. This process can be carried out using one of the following formulas, where  $A$  and  $B$  are two different clusters. The maximum distance between two clusters can be calculated by using the following formula:

$$\text{Max distance} = \max \{d(x, y): x \in A, y \in B\} \quad (2.29)$$

Another distance calculation uses the minimum distance between two clusters:

$$\text{Min distance} = \min \{d(x, y): x \in A, y \in B\} \quad (2.30)$$

The final distance calculation uses the mean distance between elements of two different clusters:

$$\text{Final distance} = \frac{1}{|A| \cdot |B|} \sum_{x \in A} \sum_{y \in B} d(x, y) \quad (2.31)$$

However, while hierarchical clustering seems easy to implement and performs well, it can still perform poorly when compared to partitioning algorithms. The distinguishing elements of different classes are the frequency of keywords contained in each document; each document contains only a small subset of the total terms contained in the entire



document set. A pair of documents can share many of the same words and be placed together in the early stages of the clustering process but this process, once complete, does not repeat so the documents are fixed in specific clusters even though they may share a higher degree of similarity with documents placed in other clusters (Steinbach, Karypis, & Kumar, 2000).

As hierarchical clustering methods share many weaknesses, namely that documents once joined cannot be split and the high level of computational resources required to solve the quadratic problem, partitional methods can be used, whereby documents are placed next to the center of the group that shares the highest degree of similarity. The three main partitioning methods are  $k$ -means clustering, fuzzy  $c$ -means clustering, and partitioning around medoids (PAM). The goal of  $k$ -means clustering is to partition a set of  $n$  documents into  $k$  different clusters whereby each document belongs to the cluster that has the smallest distance. The  $k$ -means algorithm attempts to minimize the Within Cluster Sum of Squares (WCSS) of  $x$  documents in  $S$  clusters:

$$\text{WCSS} = \underset{S}{\text{argmin}} \sum_{i=1}^k \sum_{x_j \in S_i} \|x_j - \mu_i\|^2 \quad (2.32)$$

Given an initial set of means either determined randomly or by some heuristic, the  $k$  clusters are then associated with the nearest mean value, or center of the cluster, by using the following formula:

$$S_i^{(t)} = \left\{ x_j : \|x_j - m_i^{(t)}\| \leq \|x_j - m_{i^*}^{(t)}\| \text{ for all } i^* = 1, \dots, k \right\} \quad (2.33)$$

In this formula,  $x$  is each document and  $m$  is the mean of the cluster. The centroid of each of the means becomes the mean of the new cluster, which is calculated by using the following formula:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j \quad (2.34)$$

After the means for each cluster are calculated, these become the new means and then the assignment and centroids are recalculated. This process continues until convergence is reached when the association of each resource with each cluster does not change (Lloyd, 1982). *K*-means is advantageous to use over hierarchical clustering methods because the calculation of clusters is not quadratic but, rather, linear as additional documents are added to the clusters. Documents are assigned to each cluster based on individual merit, not their relationship to other clusters so it experiences higher performance than other clustering methods.

Partitioning around Medoids (PAM) is a variation of *k*-means clustering where, rather than partitioning around randomly chosen points, the data is partitioned around actual data points drawn from the collection of  $n$  data points (medoids) based on the predefined number of clusters. To begin, the set of  $k$  data points are randomly chosen from the initial set  $n$  of input documents. Each additional data point is then associated with the nearest medoid based on similarity measures such as Euclidean distance or cosine. For each medoid  $m$ , each non-medoid data point is then swapped with  $m$  and the WCSS is calculated to minimize the WCSS and the data point with the lowest WCSS is then selected as the new medoid. This process continues until there are no longer any changes in the medoid, at which point, the clusters are identified (Theodoridis & Koutroumbas, 2006).

Fuzzy *c*-means clustering takes a different approach by allowing a document to belong to one or more clusters and is based on work done by Dunn (1973) and further improved by Bezdek (1981). The idea is based on the minimization of the objective function:

$$J_m = \sum_{i=1}^N \sum_{j=1}^C u_{ij}^m \|x_i - c_j\|^2, \text{ where } 1 \leq m \leq \infty \quad (2.35)$$

where  $m$  is any real number over one,  $u_{ij}$  is the degree of membership of  $x_i$  in the cluster  $j$ ,  $x_i$  is the  $i$ th of  $d$ -dimensional measured data,  $c_j$  is the  $d$ -dimension center of the cluster, and  $\|\cdot\|$  is any norm expressing the similarity between any measured data and the center. The partitioning function is carried out by reiterating through a process whereby membership in a cluster  $u_{ij}$  and the cluster centers  $c_j$  are updated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^C \left( \frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}} \text{ where } c_j = \frac{\sum_{i=1}^N u_{ij}^m \cdot x_i}{\sum_{i=1}^N u_{ij}^m} \quad (2.36)$$

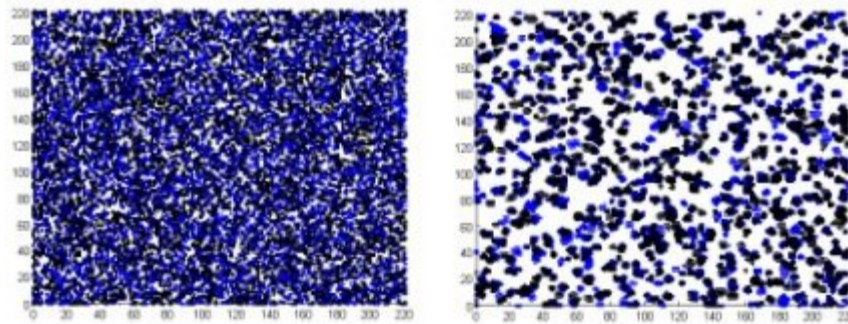
The iteration stops at the point when the following function meets a predetermined threshold value  $\delta$  between zero and one:

$$\delta > \max_{ij} \left\{ \left| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right| \right\} \quad (2.37)$$

This iteration eventually converges on a local minimum or a saddle point  $J_m$ .

Clustering can be used to find similar strands of Deoxyribonucleic Acid (DNA) in different animals and people, whereas a human expert may not know these relationships at the outset. DNA contains the genetic material that controls the functioning, appearance, and growth of organisms. One study employs a self-organizing map approach to place similar DNA sequences close together. For example, the DNA of all people experiencing a common disease can be sequenced and then different sequences can be compared to identify similarities or differences between people that contracted this disease. This method allows for much simpler analysis of DNA sequences rather than requiring a human to manually read through all sequences to identify possible causes for the disease (Elhadi & Abbas, 2010). In fact, the only input required by this model, beyond the dataset, is the number of clusters that should be created by the model.

Clustering can also be used for intrusion detection by identifying the most likely characteristics of unauthorized users. The following figure demonstrates the result of the clustering algorithm (Ramos & Abraham, 2005).



**Figure 2.2** Intrusion detection system sample cluster.

While this model may not appear to provide a lot of meaningful information, the clustering represents the similarities between different attributes, such as the number of failed logins, the time between logins, or the number of bytes transmitted. By grouping these results together, the differentiation between valid entries and invalid entries to the system can be discerned. After the characteristics of network intrusions are identified by using this model, then a formula or relationship between these different characteristics can be developed.

Clustering can also be used in strategic group analysis in order to determine whether different firms have similar strategic positions within a particular industry. As different firms seek to develop individual competitive advantage on such attributes as price and quality, cluster analysis can be used to identify firms with similar attributes. By developing a framework around the Turkish construction industry, three clusters are

identified with significant performance differences between the firms in each of the three clusters. In this study, the firms that perform well compete on the basis of increasing quality, gaining access to necessary resources, employing a systematic approach, and encouraging a collaborative environment for decision making (Dikmen, Birgonul, & Budayan, 2009). By identifying the characteristics of firms in the best performing cluster, the other firms can modify their strategic plan to incorporate these characteristics to move towards the optimal cluster containing high-performing firms.

All of these clustering methods share two major disadvantages over supervised machine learning methods, namely that the optimal number of clusters is unknown and the relationship between documents in the same cluster is similarly unknown. Documents may need to be labeled based on the most appropriate audience level but the grouping is based on terms contained in each document, possibly causing documents to be clustered based on subjects or authors. MClust is a computer program that analyzes the documents in the dataset and suggests the optimal configuration for clustering (Fraley & Raftery, 2009). It provides for parameter estimation for normal mixture models with a variety of covariance structures and can also provide options for simulation using these models. Other included functions support hierarchical model-based clustering, Expectation-Maximum for mixture estimation, and Bayes Information Criterion (BIC) for suggesting comprehensive strategies for clustering, density estimation, and discriminate analysis. The following figure displays sample output from the MClust program after inputting a document set:

identifier	Model	HC	EM	Distribution	Volume	Shape	Orientation
E		•	•	(univariate)	equal		
V		•	•	(univariate)	variable		
EII	$\lambda I$	•	•	Spherical	equal	equal	NA
VII	$\lambda_k I$	•	•	Spherical	variable	equal	NA
EEI	$\lambda A$		•	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k A$		•	Diagonal	variable	equal	coordinate axes
EVI	$\lambda A_k$		•	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k A_k$		•	Diagonal	variable	variable	coordinate axes
EEE	$\lambda D A D^T$	•	•	Ellipsoidal	equal	equal	equal
EEV	$\lambda D_k A D_k^T$		•	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k D_k A D_k^T$		•	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k D_k A_k D_k^T$	•	•	Ellipsoidal	variable	variable	variable

**Figure 2.3** Sample output of MClust program.

The circles in the above figure indicate that hierarchical clustering can be used; the program can also indicate the optimal number of clusters for a particular dataset (Fraley & Raftery, 2009). Even if the optimal number of clusters is known, the relationship between documents in different clusters cannot be determined beforehand. This method is good for visualizing large datasets or attempting to find similarities between a number of different documents when the similarities between documents are not initially known.

Clustering is extremely useful to group similar items together when the relationship between those items is not initially known. In the case of audience level prediction, the label for each cluster is known, namely first grade, second grade, etc. For example, if a librarian tries to catalog resources in the collection based on audience level, clustering cannot be used to identify the audience level of each new resource and, therefore, is not appropriate to solve the audience level prediction problem.

### 2.2.4 Support Vector Machines

The Support Vector Machines (SVM) method relies on mathematical formulas to learn the best-separation hyperplane from a set of positive and negative training samples and then splits classified entities into two subsets according to certain independent parameters that represent the properties of the data to be classified. As the separation between two sample sets increases, the probability of correctly labeling the document similarly increases (Joachims, 1998; Joachims 1999). This method is used successfully in many different classification tasks ranging from computer grading of student essays (Page, 1994) to other text categorization tasks (Yang & Liu, 1999). In the text classification application, SVM uses a vocabulary-based method that considers each word in the text of the entire training set as a unit word vector and then normalizes the frequency weights of all words in the documents that are measurement units along the word vector. After the model is created, an  $N$ -dimensional vector space model represents documents in which  $N$  represents the number of feature words in the document. SVM classification is well suited for sparse document vectors that contain a high proportion of unique terms for each class.

SVM classifiers maximize the distance of positive and negative data, also called the margin, from the hyperplane. Suppose the training data for a two-class classifier is represented by the following formula:

$$\text{Training data} = (x_1^+, y_1), \dots, (x_k^+, y_k), (x_{k+1}^-, y_{k+1}), \dots, (x_l^-, y_l) \quad (2.38)$$

In this formula,  $k$  is the number of positive samples,  $(l-k)$  is the number of negative samples,  $y \in \{+1, -1\}$  and each  $x_i$  is an  $N$ -dimensional vector in  $X \subseteq \mathfrak{R}^N$  real-valued

space. The decision boundary hyperplane  $H_d$  is an  $\mathfrak{R}^N$  dimensional plane containing no training data as represented by a linear function:

$$H_d: (w \cdot x) + b = 0 \quad \text{OR} \quad H_d: w^T x + b = 0 \quad (\text{matrix form}) \quad (2.39)$$

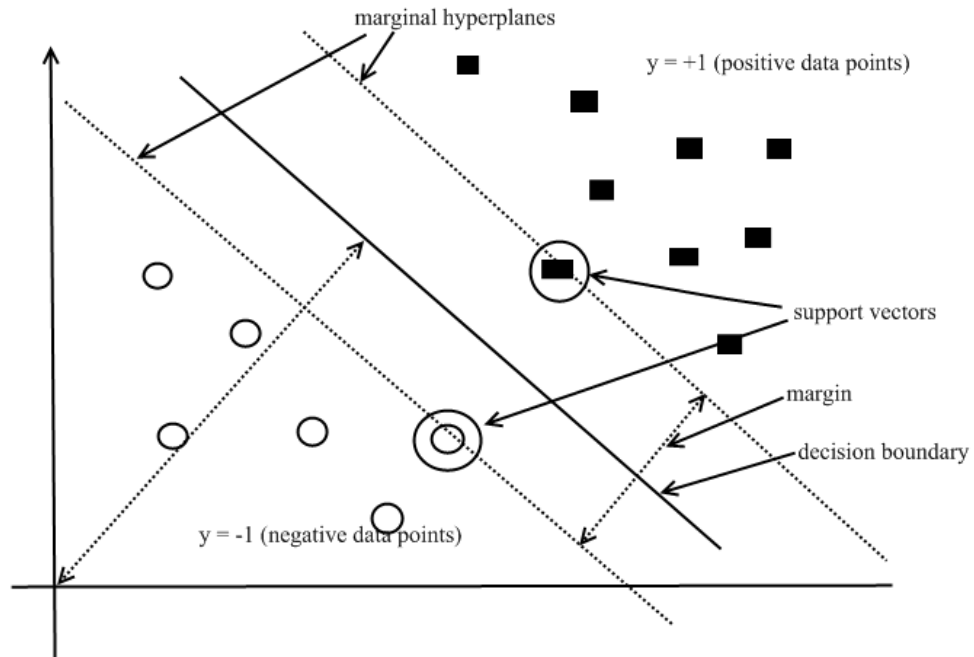
where  $\cdot$  represents the vector dot product. The SVM model then maximizes the margin that separates these two data classes; as the margin increases, the possibility of a classification error is reduced. Given two support vectors, the closest positive sample and negative sample to the decision boundary,  $(x_i^+, +1)$  and  $(x_j^-, -1)$  respectively, are identified. Then, two marginal hyperplanes,  $H_m^+$  and  $H_m^-$ , can be defined that pass through the positive and negative support vector and are also parallel to the decision boundary hyperplane  $H_d$ . When a normal vector of positive and negative samples with respect to marginal hyperplanes is measured, the positive and negative directions indicate the classification boundary between the two hyperplanes, as shown in the following equations:

$$H_m^+: (w \cdot x_i^+) + b = +1 \quad (2.40)$$

$$H_m^-: (w \cdot x_j^-) + b = -1 \quad (2.41)$$

Figure 2.4 on the next page provides a graphical representation of optimal hyperplanes separating two datasets in a two dimensional vector space model. As the positive and negative samples are on opposite sides of the optimal hyperplane, the direction of the normal vector  $w$  with respect to the data points is the decision rule for classification.





**Figure 2.4** SVM graphical representation of hyperplanes for a two-class classifier.

As SVM, by its very nature, is a binary classification problem, whereby a document falls into one class or the other, a problem arises when more than two classes exist. To solve this problem, one of two methods is generally chosen, namely One versus All or One versus One. The One versus One method considers all pairs of classes when performing the calculation, whereby one class in the pair is considered to be the negative class while the other is considered to be the positive class; the document is then placed into one of the two classes based on the decision hyperplane and the winning class is incremented by one point. After the comparison between all pairs of classes is completed, the votes for each class are counted and then the document is labeled with the class with the highest number of votes. The One versus All method, on the other hand, considers a single class to be the positive class while documents in all other classes are considered to be negative; the  $p$  value, or the highest separation between the decision

hyperplane and the document vector, is calculated for each iteration. After all classes are the positive class, then the document is labeled with the class that has the highest  $p$  value separation.

In one of the most common applications for machine learning, spam detection, email messages are identified as either spam or not spam. Since a large number of mass emails are sent out every day, this system can help users filter relevant messages from unwanted ones. In addition to email, spam is also prevalent on social networking sites possibly to boost rankings in search engines or encourage more people to buy a particular product. This SVM-based detection system attempts to separate forum spam messages from valid postings based on a number of characteristics, including post counts and post tags. By incorporating both URL and tag information into the classification model, the ability to correctly predict spam is found to be 94.54% (Kyriakopoulou & Kalamboukis, 2008).

SVM is also used in financial time series forecasting. This study compares the performance of a multi-layer back-propagation neural network with SVM. By using five real futures contracts compiled from the Chicago Mercantile Market, SVM outperforms the neural network based algorithm with respect to normalized mean square error (NMSE), mean absolute error (MAE), directional symmetry (DS), and weighted directional symmetry. These results show that SVM can be advantageous in financial time series forecasting (Tay & Cao, 2001).

SVM is also prevalent in the field of bioinformatics. One study introduces a sequence-similarity kernel in combination with support vector machines to solve the protein classification problem. With experiments using the SCOP database, this method

performs well in homology detection by using linear time classification of test sequences. By using the ROC50 curve, SVM outperforms all other methods under consideration (Leslie, Eskin, & Noble, 2002).

SVM can also be used in the facial recognition domain by verifying that the features in a picture represent a face versus some other object. After identifying possible faces by looking for skin color pixels, the eyes are found by identifying the white around the iris. In this model, 300 facial images and 300 non-face images are used as the training data. After selecting the attributes to identify possible facial images, the model is trained on these attributes to separate facial images from non-facial images. By using an SVM-based system over a neural networking approach, the face detection rate is found to improve from 88% of samples to 96% while the false detection rate decreases from 6% to 4% (Lin, Yen, Yeh, & Lin, 2008). SVM experiences high performance with respect to identifying faces, outperforming other machine learning methods.

The complexity of SVM modeling is independent of the features encountered in the training dataset and the number of support vectors that must be computed to develop the model. These classifiers also have the major advantage that they are minimally affected by outliers and, after the training model is computed, the complexity of the model does not increase as the number of unlabeled documents increases. SVM runtime computation is faster during the training and model-building phases using linear optimization techniques over quadratic computations that require the calculation of the vector dot product for every document.

### 2.2.5 Latent Semantic Indexing

One of the problems experienced by most search engines is the inability to match words if they are misspelled or misunderstanding the context in which words are used. To try to minimize this impact on search performance, Latent Semantic Indexing (LSI) is much smarter than other classification methods in that, rather than identifying similar terms between documents, LSI considers similar concepts (Furnas et al., 1987). This method is based on the premise that words used in similar contexts have similar meanings. In this way, even if the term itself does not appear in the document, a query can still identify the document as relevant to the search keywords; this method also handles the problem of synonyms, where two different words have similar meanings, and polysemy, where a single word has several different meanings (Deerwester, Dumais, & Harshman, 1988). For example, the term notebook may refer to a movie (*The Notebook*), a laptop computer, or even a pad of paper. This method uncovers the latent semantic structure of text and is able to identify similarities even if words are misspelled or do not even exist in the document. LSI can also be applied to many other areas, including spam detection (Gee, 2003) and even summarizing a body of text (Gong & Liu, 2001).

LSI can be applied to the problem of document classification, where an unlabeled document is assigned to one or more predefined categories based on the similarity of concepts contained in the document when compared with the concepts for each class contained in the training data set. In this way, a document may be mapped to a particular class even if the document does not contain any terms that are identified as belonging to that class. During the training process, example documents for each category are used to identify the key concepts contained within each category. Then, each unlabeled

document is assigned to a category based on the highest similarity of concepts contained in the document with concepts in the unlabeled resource (Dumais et al., 1998). LSI is not strictly limited to the exact spelling of words in the document and, therefore, is forgiving of misspellings or character strings. LSI can also be applied to match documents across languages as long as the languages are structured similarly. In these ways, LSI is shown to be very effective at matching documents to pre-defined classes (Ding, 1999).

LSI first requires the construction of a term-document matrix, then performing a Singular Value Decomposition on that matrix, and, finally, by using the new set of matrices to identify similar concepts within the collection. LSI creates a term-document matrix, where each term is a row, each document is a column, and each cell within that matrix identifies the number of occurrences of the term in the document. These matrices tend to be very large and sparse as few documents contain mostly the same terms. After the matrix is constructed, then local and global weighting functions can be applied to determine the importance of term weight for each document. Some common local weighting schemes include binary (value of 1 if term exists in the document or 0 if the term does not exist), term frequency (simply the number of occurrences of the term in the document, 0 if the term does not exist), and log (log of the term frequencies + 1). Global weighting schemes commonly fall into one of several categories, namely binary ( $g_i = 1$ ), normal ( $g_i = 1 / (\text{square root of the sum of the term frequencies squared})$ ),  $g_i = gf_i / df_i$ , where  $gf_i$  is the number of occurrences of term  $i$  in the entire collection and  $df_i$  is the number of documents in which the term occurs, or the inverse document frequency ( $g_i = 1 + \log_2 (n / df_i)$ ). Another log entropy weighting function is also proposed as shown in the following equation (Berry & Browne, 2005):

$$g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n}, \text{ where } p_{ij} = \frac{tf_{ij}}{gf_i} \quad (2.42)$$

After each of the local and global weighting schemes are defined, the value in each cell can then be calculated by the following sample formula; in this formula, the log based weighting scheme is used:

$$a_{ij} = g_i \log(tf_{ij} + 1), \text{ where } g_i = 1 + \sum_j \frac{p_{ij} \log p_{ij}}{\log n} \quad (2.43)$$

Now that the value of each cell is adjusted, Singular Value Decomposition is performed on the matrix to determine the relationship and patterns between concepts within the different documents in the training collection (Berry, Dumais, & O'Brien, 1995). During this process, three different matrices are computed out of the original adjusted term-frequency matrix, a term-concept vector matrix  $\mathbf{T}$ , a singular values matrix  $\mathbf{S}$ , and finally a concept-document matrix  $\mathbf{D}$  subject to the following conditions  $\mathbf{A} = \mathbf{TSD}^T$ ,  $\mathbf{T}^T \mathbf{T} = \mathbf{D}^T \mathbf{D} = \mathbf{I}_r$ ,  $\mathbf{TT}^T = \mathbf{I}_m$ ,  $\mathbf{DD}^T = \mathbf{I}_n$ , and  $S_{1,1} \geq S_{2,2} \geq \dots \geq S_{r,r} > 0$ ,  $S_{ij} = 0$  where  $i \neq j$ .  $\mathbf{A}$  is the original supplied term-document matrix where  $m$  is the number of unique terms and  $n$  is the number of documents in the collection,  $\mathbf{T}$  is the computed  $m$  by  $r$  matrix of term vectors where  $r$  is the rank of  $\mathbf{A}$  (the measure of unique dimensions),  $\mathbf{S}$  is a computed  $r$  by  $r$  diagonal matrix of decreasing singular values, and  $\mathbf{D}$  is a computed  $n$  by  $r$  matrix of document vectors. After all of these additional matrices are populated, LSI reduces the concept matrix  $\mathbf{S}$  to a much smaller size  $k$ , typically between 100 and 300 dimensions. This reduction eliminates much of the noise that is generated as a result of the sparse matrix while preserving the most important concepts in the work. Rather than calculating the entire matrix  $\mathbf{S}$  and then truncating it to a much smaller dimension, more efficient LSI algorithms prefer to only calculate the first  $k$  dimensions and then ending this process.

After these three matrices are created, the LSI algorithm can now query the matrix to retrieve documents from the collection. The similarity between documents and the query can be calculated as a function of the angle between the corresponding vectors. These three matrices store the conceptual information that has been gathered from the collection. By transforming the original formula of  $\mathbf{A} = \mathbf{TSD}^T$  into  $\mathbf{D} = \mathbf{A}^T \mathbf{T} \mathbf{S}^{-1}$ , the LSI matrix can now be queried or additional documents may be added to the LSI space. The query can be added to a new column in  $\mathbf{A}$  and the original global and term weights, drawn from the original dataset, are multiplied by the term count in each row in the new column; the new column in  $\mathbf{A}$ , representing the query, is multiplied by  $\mathbf{T} \mathbf{S}^{-1}$ . The similarity between the query and the concepts representing each document can be calculated using the vector space model; if the document discusses all of the concepts found in the query, it is considered highly relevant. New terms present in the additional documents or queries are ignored during this process. In text classification, the term-document matrix can be a term-class matrix, where all of the class terms are considered to be a document. Then, after the additional matrices are calculated, the concepts contained in an unlabeled document are compared with the existing classes and the document is assigned to the class with the highest level of similarity.

In a study employing the LSI concept, new relationships can be identified between different research papers. Scientific ideas can be compared in literature and possible new connections can be found that may have been previously undiscovered. One study seeks to identify nearby literature which may make incremental improvements and, also, to uncover far reaching relationships that may introduce new hypotheses that can be tested. For example, the term blood viscosity is closely related to both Reynauds

and fish oil but no documents may contain both Reynauds and fish oil; by employing LSI in this application, the underlying concept of fish oil may be linked to Reynauds through the additional term blood viscosity. Even though the system does not discover this association with a high confidence level, two other treatments are identified as possible cures for Reynauds, namely calcium dobesilate and Niceritrol; these two drugs are related to treating Reynauds. By using LSI, these latent relationships can be uncovered and suggest a new hypothesis that medical researchers can test (Gordan & Dumais, 1998).

Filtering unsolicited email, otherwise known as spam, from valid emails that the user would like to read is a continual problem. Many different methods are proposed to filter spam from non-spam messages, particularly using Naïve Bayesian methods. LSI tends to be more effective at filtering spam from non-spam than these other methods, as evidenced by both high precision and high recall. One study applies LSI to the problem of spam filtering, resulting in precision and recall well over 98% for identification of both legitimate and spam documents (Gee, 2003)

Since LSI performs extremely well at identifying important concepts in documents, LSI can extract important sentences from a document in order to create a summary. The main goal of this study is the selection of sentences that both describe the important content in the document as well as find sentences that contain different information from one another. The results from the LSI document summarization model are compared with the manual creation of summaries by three human evaluators. This study compares the performance of this LSI model (extracting the sentences with the highest singular vectors) with the performance of traditional summarization techniques that select sentences with the smallest cosine between the document text and each



sentence. By using a test collection consisting of 549 closed-caption news stories containing between 3 and 105 sentences, the recall is around 53% and the precision around 60%, with LSI slightly outperforming traditional models (Gong & Liu, 2001).

One of the advantages of using LSI is the ability to identify documents that are structured similarly even if the documents are written in different languages. Most other machine learning methods require exact matches between terms rather than matches between concepts as in LSI. Organizations typically seek to create knowledge repositories of best practices that can be used to increase competitive advantage. The creation of these repositories in multinational organizations is a problem since the documents are often written in the language of the country where the organization is located. In order to navigate through documents stored in different languages, documents can be clustered according to concepts rather than terms. One study seeks to create an LSI-based document clustering technique to organize the knowledge in the repository on a navigational map. While this study results in a proof-of-concept system, the cross-lingual document clustering map is comparable to cluster precision and recall of single-language repositories (Wei, Yang, & Lin, 2008).

LSI is more accurate than other classification methods, since it works by identifying concepts in documents instead of merely considering the existing terms. It is forgiving of misspelled words and is even able to find a relevant document even if none of the terms in the query exist in that document. It faces many challenges that need to be overcome, most noticeably in the processing time required to calculate the different matrices. It also experiences a limit on the number of concepts that can be considered and, if the document contains only the truncated concepts from the matrix, then the

document is unable to be classified. There is also a serious problem in the ability to determine the optimal number of dimensions, or the number of important concepts represented in the collection. While Bradford suggests that the optimal number of dimensions can range from 300 for smaller collections to 400 for larger document collections containing millions of documents (2008), the optimal number of dimensions cannot be determined in advance, possibly causing relevant concepts to be inadvertently removed from the matrix. While this method is highly effective with respect to information retrieval, the large amount of computer processing power required as well as the large number of unknowns makes LSI a poor choice for identifying the audience level of documents held in digital library collections.

#### **2.2.6 Machine Learning Methods Summary**

This section summarizes five different classification methods that can be applied to automatically determine the most appropriate audience level for textual resources. The table on the next page summarizes the algorithms and performance of each method in different applications.

**Table 2.5** Machine Learning Methods Summary

<b>Learning Method</b>	<b>Basic Operation</b>	<b>Application</b>	<b>Performance</b>
Cosine	Measures cosine between documents placed in the vector space model	Information Retrieval	Increase from 0.649 to 0.714 for the Reuters collection and 0.667 to 0.719 for 20 newsgroups on a variety of topics by using WordNet + Cosine
		Recommendation	72% of documents provided to the user are relevant
		Virus Detection	Successfully detected changed code in five out of ten code samples
Naïve Bayes	Probability that class terms are used to create the document	Health Information	98% accuracy at identifying the correct audience level from three classes
		Software Code Readability	80% effective at making readability judgments
		Spam Detection	99.99% accuracy but 63% of total spam detected
		Audience Level	Root mean squared error between 1 and 2 grades for 9 out of 12 grades; correlation of 0.79 between human and computer suggested grade levels
Clustering	Places similar documents together based on their content	Genetics	N/A
		Intrusion Detection	N/A
		Corporate Strategy	N/A
Support Vector Machines	Maximize margin between two classes	Social Networking Spam Detection	94.54% accuracy
		Financial Time Series Forecasting	Outperforms multi-layer back-propagation neural network
		Protein Classification	Outperformed all other methods under consideration using ROC50 curve
		Facial Recognition	Correctly identifies face 96% of time with false detection rate of 4%
Latent Semantic Indexing	Uses dimensionality reduction to uncover latent semantic structures in text	Literature Based Discovery	Sodium Dobesilate and Niceritrol may be effective at treating Reynauds
		Spam Detection	Over 98% precision and recall to detect both spam and legitimate messages
		Document Summarization	53% precision, 60% recall at identifying the most important sentences in the document
		Document Clustering	Clustering performance across multiple languages similar to single-language performance

\*\* N/A = Not Applicable (Performance is not available)

Clustering suffers the serious drawback of not being able to associate an unlabeled document with a class based upon a pre-defined set of classes. Latent Semantic Indexing is highly accurate but is also expensive both in time and computational resources required to develop the model; each time a new document is added, the entire model must be recomputed to retain all important concepts in the collection. Naïve Bayesian methods report lower performance than other machine learning methods, including SVM. The most appropriate classification methods to use for automatic audience level prediction are SVM, Naïve Bayes, and cosine, whose performance is highest in a variety of applications and required computational resources do not increase based on the number of documents to be labeled.

### **2.3 Conclusion**

This chapter first reviewed several readability formulas that have relied on syntactic and semantic characteristics of text to suggest the most appropriate audience level of a resource. These methods should have experienced poor performance when identifying the appropriate audience level for web-based digital library resources that have not followed the conventions of written English. Then, a number of different machine learning methods borrowed from the essay grading domain were reviewed; these methods could have been used to automatically suggest the most appropriate audience level for an unlabeled resource. Among these machine learning methods, cosine, Naïve Bayes, and SVM were found to be the most appropriate methods to use in this application, with SVM generally experiencing the highest performance.

## **CHAPTER 3**

### **SVMAUD SYSTEM DESIGN AND IMPLEMENTATION**

The algorithm for the proposed SVM-based audience level prediction program, or SVMAUD, is first described in this chapter. Then, the second part of this chapter demonstrates the system operation. SVMAUD is a Windows-based classification program, with initial training conducted using a Java-based program.

#### **3.1 SVMAUD System Design**

Since SVM performs well in a variety of vocabulary based classification applications, this algorithm is proposed to automatically suggest the audience level for digital library resources and should outperform other machine learning methods and readability formulas. This system suggests the most appropriate audience level for a set of unlabeled written resources. The first section describes the document language model used by SVMAUD, while the second section provides the system architecture.

##### **3.1.1 SVMAUD Document Language Model**

The document language model relies on the “bag-of-words” vector space model approach, where each dimension of the document model represents a weighted term drawn from the individual document vocabulary. The weights of these individual terms are used to create a vector space model based on the importance of the term relative to both the document and the collection. In other words, the input document text is transformed into a feature vector that contains a number of words that describe the document content.

To construct the document language model based on selective features, several preprocessing tasks are performed to reduce the feature space by eliminating stop words, or words with high frequency of occurrence, such as “a,” “the,” “of,” etc. Even though the stop words are removed from the input documents, the removal of these words does not materially impact audience level labeling performance as they appear in many different categories. The remaining term features are still adequate to predict the audience level accurately; if the features are too numerous, the dimensionality for computation increases with little affect on performance. Spelling errors are eliminated during pre-processing after tokenization of the text only documents. As the unique terms appear sparsely in the collection, considering these terms in the model helps with fine-grained classification of documents between adjacent audience levels. After parsing and tokenization of the documents, for each audience level ( $G_i$ ), the vocabulary is constructed with the tokenized term ( $w_k$ ), the number of occurrences of the term in all documents within the audience level  $G_i$  ( $N(w_k \in G_i)$ ), number of documents across all audience levels containing the term ( $|D(w_k)|$ ), number of occurrences of each term in every category ( $N(w_k \in G_*)$ ), and number of classes that contain the term ( $|G_*(w_k)|$ ). The final model-building step involves the calculation of the weights for each feature word to reflect its relative importance within a particular audience level.

A simple classical model such as Bayes theorem calculates the value of each word within the audience level using the apriori and conditional probabilities based on word frequencies. Previous studies show text classification systems using term frequency - inverse document frequency (TF-IDF) to assign feature weights to be highly effective (Salton, Wong, & Yang, 1975). The TF-IDF weight is proposed as a measure used to

evaluate the importance of a word with respect to a document in a collection or corpus; the importance of a term increases proportionally to its frequency in the document and decreased based on the number of documents containing the term in the entire corpus as shown in the following formula:

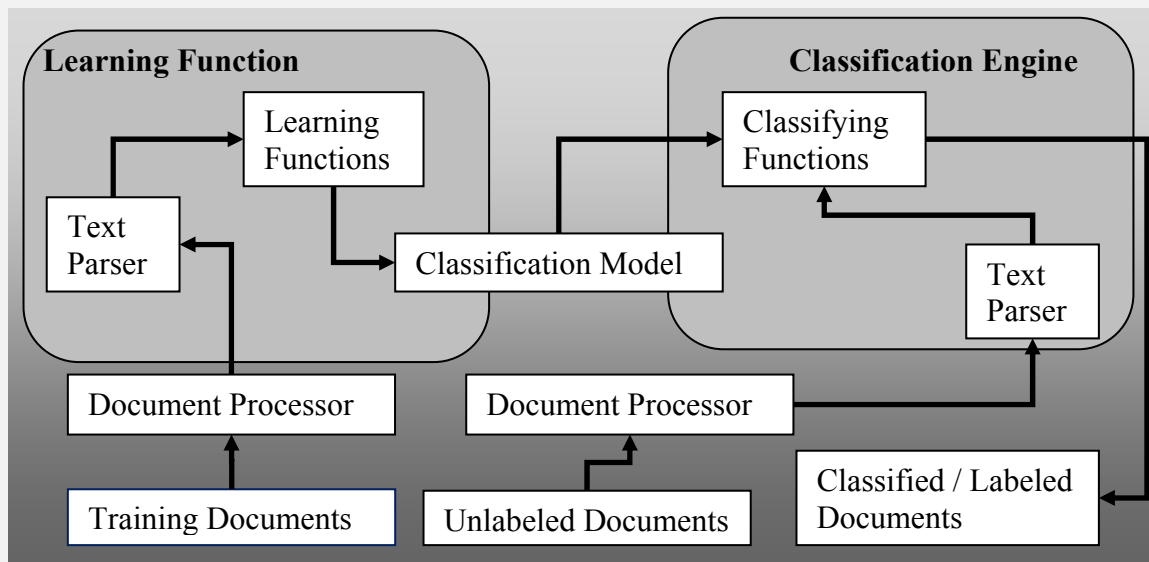
$$f(w_k \in G_i) = \sum_{D_j \in D(w_k); D_i(w_k) \in D_i} \left( \frac{N(w_k \in D_j)}{\sum N(w_* \in D_j)} \right) \times \log \left( \frac{|D_i \in G_i|}{1 + |D(w_k)|} \right) \quad (3.1)$$

$$f(w_k \in G_i) \approx \left( \frac{N(w_k \in G_i)}{\sum N(w_k \in G_*)} \right) \times \log \left( \frac{|D_i \in G_i|}{1 + |D(w_k)|} \right)$$

In SVMAUD, the TF-IDF weighing formula is modified to incorporate the term frequency of overlapping terms across multiple audience levels or classes. Similar to the TF-IDF model, where the term weight is reduced proportionally as additional documents in the collection contain the term, the term weight is dependent on the number of audience levels that contain the term. If a term appears in a number of classes or audience levels, then its importance as a discriminator between classes is reduced. These feature weights are computed after constructing the vocabulary for each audience level.

### 3.1.2 SVMAUD System Architecture

The SVMAUD machine learning algorithm is composed of two different phases – a learning phase and a classification phase. The learning phase uses documents with human-expert labeled audience level values to train the model, while the classification phase predicts the most appropriate audience level for unlabeled documents. The following figure demonstrates the different steps involved in classifying documents using SVMAUD.



**Figure 3.1** SVMAUD learning function and classification engine.

In the learning phase, the training dataset containing pre-labeled documents is provided to the text parser to perform text parsing and tokenization. The document processor extracts information from documents, e.g., in HTML documents, HTML tags and stop words are removed and the information located in the body and header is extracted. The learning function module trains SVMAUD using the document language models and generates the classification model in terms of support vector parameters; the classification engine then uses this information to classify unlabeled documents. The process to generate SVMAUD classification models for each audience level is described in the following steps:

1. The training dataset is ranked descending according to the number of documents in each class.
2. The class with the highest number of documents is identified and its documents are labeled as positive samples. Documents in all other classes are labeled as negative samples.



3. Using the positive and negative labeled documents, SVMAUD is trained to generate model  $M$  for the specific class.
4. The next largest training category sample is then selected and steps 2 to 4 are repeated until all classification models are generated.

After the learning component of the system generates the feature model for each audience level, the decision function classifies the unlabeled documents by transforming these documents into word vectors using the class information. Each test document is classified individually within these different models and the process ends when the document is positively labeled with any one of the models. In this way, the possibility of incorrectly labeling a document is reduced as it is more likely to be labeled with a class that contains a higher number of training documents. The classification procedure for a test dataset is described in the following steps:

1. The training data is converted to document vectors based on feature weights in the class with the highest number of documents.
2. Document vectors are provided to SVMAUD along with the top ranked classification model.
3. If SVMAUD labels the document positively, the document is labeled with the top ranked class label.
4. If the document is negatively labeled, the next ranked class model is used to classify the document and steps 1 to 4 are repeated. If the process does not stop with step 3 by placing a document in the positive samples, then the document is labeled with the lowest ranked class label provided all other models label the document negatively.

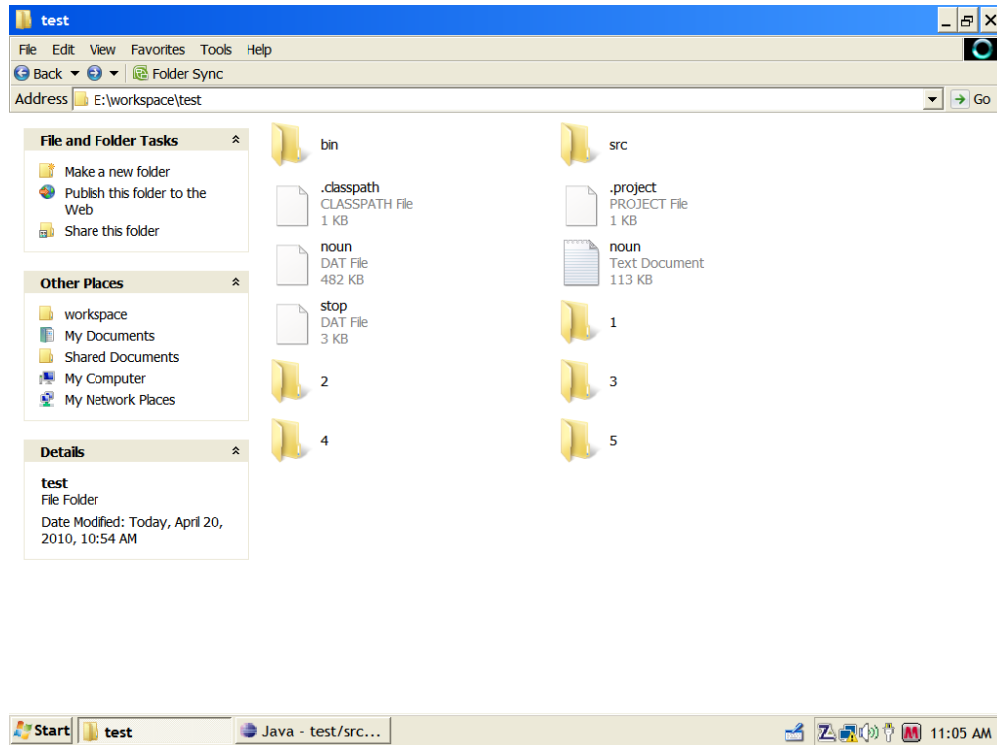
After this process is completed, all documents are labeled with the most appropriate audience level by SVMAUD. This complete and consistent audience level information can then be stored in a database for later usage by a retrieval system.

## **3.2 SVMAUD System Implementation**

This section describes the usage of SVMAUD to generate consistent and complete audience level metadata for all resources in the test collection. The first section discusses the training portion of the program, the second section describes the user interface, and the third section describes the interpretation of the output.

### **3.2.1 SVMAUD Training Program**

This section describes the training portion of SVMAUD; after the training is completed, SVMAUD can then label resources with the most appropriate audience level. The program requires a set of human-expert labeled text files that are associated with each class. For example, if there are five predefined classes (class 1, class 2, class 3, class 4, and class 5) with documents in each class, each document is placed into its associated class as in the screen capture on the next page.



**Figure 3.2** Input directory structure for training data.

After these directories are created and the resources placed into the proper classes, the training program can be run. This java based program is run through the command line to create the model for each class by issuing the following commands:

- `java CreateClassifier 1 "2,3,4,5"`
- `java CreateClassifier 2 "1,3,4,5"`
- `java CreateClassifier 3 "1,2,4,5"`
- `java CreateClassifier 4 "1,2,3,5"`
- `java CreateClassifier 5 "1,2,3,4"`

The training program requires that no class names contain spaces or special characters, such as asterisks and exclamation points. Rather than using the class names of 1, 2, 3, 4, and 5, more descriptive labels can be used, such as First, Second, Third, Fourth, or Fifth, corresponding to different elementary school grades. With respect to general audience levels, the classes can be labeled as Elementary, Middle, High, or College.

After the training program is completed, a number of text files are generated that need to be copied to the SVMAUD directory to label resources with the most appropriate audience level. A text file is created for each class; if the classification problem consists of five different classes, then five text files are created, representing the terms and their associated weights for the class. A sample text file consists of the term followed by a comma, the number of occurrences of the term in the class (term frequency) followed by a comma, the total term frequency in the entire training dataset followed by a comma, and, finally, the total number of classes in which the term occurs. A sample excerpt from one of these text files appears as follows:

```
aa,5,350,5
aachen,1,1,1
aad,2,99,5
aalen,1,1,1
aamir,1,1,1
aaps,3,2,2
aarhus,2,1,1
aaron,4,18,5
aarts,1,2,2
aas,4,39,5
aasen,1,1,1
ab,9,313,5
ab-dependent,1,1,1
aba,1,6,5
abaeva,1,1,1
```

In addition to these files, one additional text file is named category.txt and contains the class label, followed by a comma, and then the number of documents in the class. This information is entered by the user in the category.txt file, as follows.

```
1,900
2,1000
3,700
4,900
5,500
```

After all of the text files are created, they are copied into the root directory of SVMAUD's file location. The text files 1.txt, 2.txt, 3.txt, 4.txt, and 5.txt contain the training data that is inputted to SVMAUD; category.txt contains information regarding the number of documents in each category. These text files are then imported into the SVMAUD program to classify unlabeled documents.

### **3.2.2 SVMAUD System Interface**

After the model is created for each audience level, the classification program is now run to suggest the most appropriate audience level for unlabeled documents. After the classification algorithm is chosen, either cosine, Naïve Bayes, or SVM, then the text documents generated from the training portion are used to initialize the program. The Windows based portion of the program is able to identify the most appropriate audience level for resources in the collection by labeling each document with its respective class.

SVMAUD supports audience level prediction by using cosine similarity, Naïve Bayes, or Support Vector Machines (SVM). The first Directory field represents the home directory of the program; after this directory is entered, then the Initialize button is clicked. The second Directory field and Browse button represent the directory containing the text files that need to be labeled with audience level. After inputting all of this information, the program is run. In this case, SVMAUD predicts the specific audience level of resources. On the next page is a sample screen capture of SVMAUD after it predicts the audience level for each resource.

Information: 25 classified, 21 correct (precision: 0.8400) SVM Pairwise Cl  Limit to existing categories only

Directory: ments and Settings\Todd\My Documents\HomeSchoolingReadingLevelStudy\TextClassifier Initialize

Document: c:\temp\google.htm Browse Classify

Query: Num: 10 Classify

Directory: C:\Downloads\TestDocs Browse Classify

Document	Actual	Assigned
C:\Downloads\TestDocs\3_31.txt	3	5
C:\Downloads\TestDocs\3_32.txt	3	5
C:\Downloads\TestDocs\3_33.txt	3	5
C:\Downloads\TestDocs\3_34.txt	3	5
C:\Downloads\TestDocs\5_22.txt	5	5
C:\Downloads\TestDocs\5_23.txt	5	5
C:\Downloads\TestDocs\5_24.txt	5	5
C:\Downloads\TestDocs\5_25.txt	5	5
C:\Downloads\TestDocs\5_26.txt	5	5
C:\Downloads\TestDocs\5_27.txt	5	5
C:\Downloads\TestDocs\5_28.txt	5	5
C:\Downloads\TestDocs\5_29.txt	5	5
C:\Downloads\TestDocs\5_30.txt	5	5
C:\Downloads\TestDocs\5_31.txt	5	5
C:\Downloads\TestDocs\5_32.txt	5	5
C:\Downloads\TestDocs\5_33.txt	5	5
C:\Downloads\TestDocs\5_34.txt	5	5
C:\Downloads\TestDocs\5_35.txt	5	5
C:\Downloads\TestDocs\5_36.txt	5	5
C:\Downloads\TestDocs\5_37.txt	5	5
C:\Downloads\TestDocs\5_38.txt	5	5
C:\Downloads\TestDocs\5_39.txt	5	5
C:\Downloads\TestDocs\5_40.txt	5	5
C:\Downloads\TestDocs\5_41.txt	5	5
C:\Downloads\TestDocs\5_42.txt	5	5

**Figure 3.3** SVMAUD classification result.

The first column in the output screen represents the document filename that is classified. The second column shows the human-expert label for the resource, if this information is previously known. The final column to the right displays the predicted audience level of the resource. If the two right most columns display the same information, then the document is correctly classified. In this case, SVMAUD correctly predicts the human-expert entered audience level with a precision of 0.84.

### 3.2.3 SVMAUD System Output

After SVMAUD is finished running, an output text file is generated that identifies the most appropriate audience level for each document. In the following file, the first column displays the file location, the second column shows the original expert-identified class, and the third column presents the class label assigned by the program. If the second and third columns are the same, then SVMAUD correctly labels the document.

C:\Downloads\TestDocs\3_31.txt	3	5
C:\Downloads\TestDocs\3_32.txt	3	5
C:\Downloads\TestDocs\3_33.txt	3	5
C:\Downloads\TestDocs\3_34.txt	3	5
C:\Downloads\TestDocs\5_22.txt	5	5
C:\Downloads\TestDocs\5_23.txt	5	5
C:\Downloads\TestDocs\5_24.txt	5	5
C:\Downloads\TestDocs\5_25.txt	5	5
C:\Downloads\TestDocs\5_26.txt	5	5
C:\Downloads\TestDocs\5_27.txt	5	5
C:\Downloads\TestDocs\5_28.txt	5	5
C:\Downloads\TestDocs\5_29.txt	5	5
C:\Downloads\TestDocs\5_30.txt	5	5
C:\Downloads\TestDocs\5_31.txt	5	5
C:\Downloads\TestDocs\5_32.txt	5	5
C:\Downloads\TestDocs\5_33.txt	5	5
C:\Downloads\TestDocs\5_34.txt	5	5
C:\Downloads\TestDocs\5_35.txt	5	5
C:\Downloads\TestDocs\5_36.txt	5	5
C:\Downloads\TestDocs\5_37.txt	5	5
C:\Downloads\TestDocs\5_38.txt	5	5
C:\Downloads\TestDocs\5_39.txt	5	5
C:\Downloads\TestDocs\5_40.txt	5	5
C:\Downloads\TestDocs\5_41.txt	5	5
C:\Downloads\TestDocs\5_42.txt	5	5

After the output text file is generated, this file can be imported into a database to populate the missing or inconsistent audience level metadata values. Using SVMAUD is much faster than manually labeling each document with the correct class information.

After the audience level is automatically predicted for each document by SVMAUD and the resulting output text file imported into a database, then users of the retrieval system can draw upon this complete and consistent audience level information to reduce the effort required to find relevant documents in the collection.

### **3.3 SVMAUD Summary**

This chapter demonstrated the system design and operation of SVMAUD that could have automatically suggested the audience level for all documents in the digital library collection with missing or incompatible audience level information. This section first described the classification algorithm to automatically generate audience level metadata by SVMAUD. In the first section, the document language model was described, and then the algorithm used by SVMAUD to label documents with audience level was provided. The second part of this chapter described the training portion of SVMAUD, the system interface, and the interpretation of the output of the program. SVMAUD could have also automatically identified the appropriate class label for any metadata element that could have been limited to a set of known values, including subject category and learning object type.



## CHAPTER 4

### AUDIENCE LEVEL PREDICTION

As the audience level metadata can be incomplete or inconsistent for all resources in a digital library collection, automatic classification methods can be employed to label all resources in the collection with complete and consistent audience level metadata. This study measures the performance of two popular readability formulas, Flesch-Kincaid Reading Age and Dale-Chall Reading Ease Score; cosine, Naïve Bayes, and Collins-Thompson and Callan machine learning methods; and SVM/AUD in their ability to correctly identify the human-expert provided audience level for a collection of digital library resources.

Different machine learning methods and readability formulas are utilized to label web-based resources with the most appropriate audience level. The text from documents with expert-provided audience level information is extracted and used to train the four classifiers under evaluation. The performance of the two readability formulas and four machine learning methods are measured using the standard classification performance measures of precision, recall, and F-measure. Precision (P) is defined as the proportion of resources labeled with an audience level by the automated method that matches the human-expert identified level. Recall (R) is the proportion of resources associated with a human-expert identified audience level that the automated method correctly identifies. The F-measure (F) is defined as the harmonic mean between precision and recall, calculated by  $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$ . In addition, the correlation between human-expert entered and the predicted audience level is also reported, since an

incorrect prediction of plus or minus one audience level is a smaller error than an incorrect prediction by five audience levels; correlation is calculated by the following formula, where  $X$  and  $Y$  are the values of two variables (in this case,  $X$  is the human-expert entered audience level for the resource and  $Y$  is the predicted audience level):

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (4.1)$$

In this formula,  $\mu_X$  and  $\mu_Y$  represent the expected values, and  $\sigma_X$  and  $\sigma_Y$  represent the standard deviation. In addition, the t-test level of significance whereby SVMAUD exceeds the performance of the method under consideration is also provided. Then, the method with the highest performance can be used to suggest the audience level for all documents with missing or incompatible audience level metadata.

In order to ensure a fair comparison among all machine learning methods and readability formulas, each method is asked to predict both the general and specific audience level for each resource in the test collection. While each resource should be associated with the single most specific audience level, resources with audience levels of plus or minus one audience level should also be appropriate for the single target audience level. Since a resource that is appropriate for first grade can probably also be understood by a second grade student, the general audience levels encompass a few adjacent audience levels. For purposes of this research, early elementary refers to students in kindergarten or below to second grade students, late elementary refers to students in grades three through five, middle school refers to students in grades six through eight, high school refers to students in grades nine through twelve, and college refers to all students taking undergraduate and graduate courses. In all studies, the general and

specific audience level prediction performance is compared with the human-expert assigned audience levels.

First, this chapter describes the research questions that the study seeks to answer. The next two sections describe the composition of the collection used for evaluation purposes. Then, the next section describes the results of different readability formulas and machine learning systems when predicting the audience level of digital library resources by using full text for training and testing. Finally, the results from these studies are summarized and the chapter is concluded.

#### **4.1 Research Questions**

This section identifies the five research questions that this research seeks to answer. The research questions mainly revolve around comparing the performance of the proposed audience level prediction system SVMAUD to readability formulas and other machine learning systems, and then attempting to improve SVMAUD's prediction performance through a variety of performance tuning methods.

RQ1: Could SVMAUD be used to predict the audience level for digital library resources with performance exceeding readability formulas?

This question determines if the audience level metadata could be identified using the computer based classification method SVMAUD with higher performance than traditional readability formulas. Since these machine learning methods are successfully used in the essay grading domain to suggest the grade or score of an essay based on the vocabulary chosen by the author, they should also perform well to solve the audience

level prediction problem. Even though these methods are not applied to predict the audience level of resources held by digital library collections, these methods should still outperform traditional readability formulas that rely on word and sentence characteristics.

RQ2: Which machine learning method, among cosine, SVM, Naïve Bayes, and the Collins-Thompson and Callan method, would result in the highest performance when suggesting the audience level for all documents in a collection?

This question seeks to determine the method that would have the highest level of performance when identifying the most appropriate audience level for previously unlabeled resources. Each method would be called upon to identify the audience level for all documents in the test collection and then their predictions would be compared with the audience level information provided by the human-expert collection managers.

RQ3: Since digital library resources have been predominantly web pages, could the machine learning audience level prediction performance be improved if term weights have been adjusted according to the HTML tags in which they have appeared?

Most digital library collections contain HTML pages that are hosted on a web server to be accessible to all users at all locations and all times. The title and header information should hold important clues describing the major ideas in the resource, while the table data could solely consist of numbers and should warrant smaller weight. By assigning different weights to terms appearing in different HTML tags, the prediction performance should improve over assigning all terms the same weight to all terms independent of the HTML tag in which the term appears.

RQ4: By training the machine learning methods using metadata associated with each resource in a digital library collection, could the audience level classification performance be improved?

As documents held in digital library collections mainly consist of online resources, these documents would likely contain headers and footers common to every page in the collection; by including this noisy data, the performance of the classification methods should be reduced since this information would be common to all resources in the collection independent of audience level. On the other hand, metadata elements, such as title, keywords, and abstract, should be unique for every resource in the collection. By reducing the level of noise in the training dataset, the prediction performance by all machine learning methods should be improved over using full text for training. Since not all documents would be cataloged with complete metadata information, with only title and URL being required, the full text of unlabeled resources would need to be used to ensure a sufficient number of words would be available for comparison to the terms found in the training data.

RQ5: Could the audience level classification performance be improved if the machine learning methods have been trained and tested using resources discussing the same subject?

SVMAUD could be used to predict the audience level for a wide range of subjects, where one classifier could predict the audience level for all resources covering all subjects contained in a digital library collection. This study should improve SVMAUD performance by developing a series of subject-specific classifiers, where

resources discussing a single subject, such as mathematics, would be used to predict the audience level of other mathematics resources.

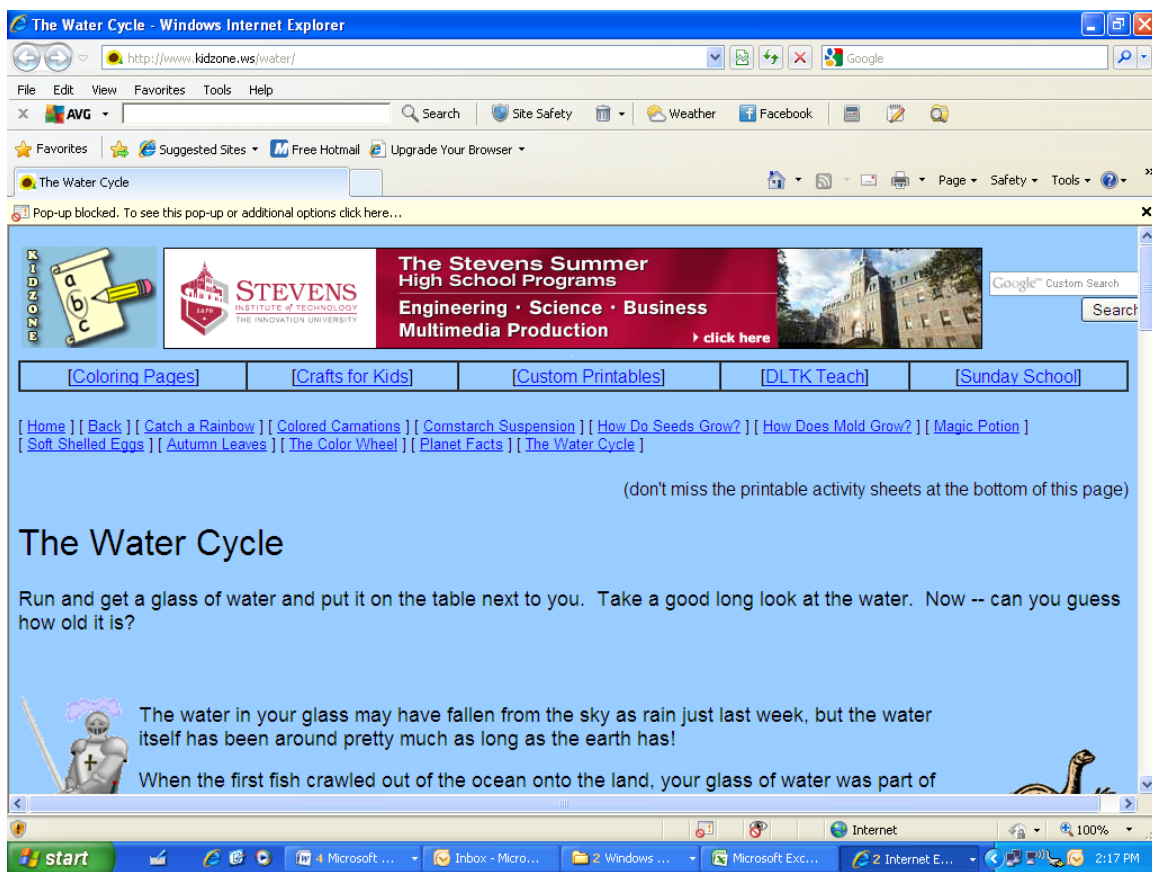
#### **4.2 Creating the Digital Library Resource Collection**

The NSDL library required all collections to enter metadata values according to the NSDL\_DC metadata standard; these elements and their descriptions are listed in Appendix B. Even though NSDL member collections were required to follow this standard, only title and URL were required with all other elements being optional and no controlled vocabulary existed to restrict the values entered for these metadata elements, leading to incomplete and inconsistent entries. SVMAUD should have been used to complete the education level, or audience level, metadata for all resources held by the digital library collections. All of the resources and associated metadata, such as title, author, keywords, URL, and abstract, were downloaded from the digital libraries by using the jOAI program. This software was an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) harvester program that ran under Tomcat and allowed resources and associated metadata to be harvested and stored in the file system (Weatherley, 2012). After this program downloaded all of the resource metadata to the file system, then the resource metadata was imported into a database to be used for training and testing the machine learning methods and readability formulas.

After compiling the list of URLs for all resources in the collection, the full text of the resource could have been downloaded for input to the readability formulas and machine learning techniques. The freeware program, URL2File, was called upon to retrieve the full text of the HTML page. This program could have retrieved any file

available on the World Wide Web and been run in batch mode by inputting a list of URLs (Chami.com, 2002). This program was used to download the text of the HTML pages, excluding any pictures or video files that have been embedded on the page.

The source code of HTML pages consisted of a plain text file that could have been directly inputted to the readability formulas and machine learning methods in order to suggest the most appropriate audience level of resources. However, these pages also contained headers, footers, scripts, tables, ordered lists, metadata tags, and other features not available in books and magazines; these features could have distorted the calculations required for readability formulas. The following figures show two sample digital library resources and their associated HTML source code. The first sample page is hosted by Kidzone, which catalogs elementary school activities covering all subjects, from math to science and geography; this sample resource describes an activity for a student to learn about the water cycle (Kidzone.ws, 2012).



**Figure 4.1** Screen capture of a water cycle activity.

Figure 4.2 on the next page shows an excerpt from the source code for the water cycle activity screen capture shown in Figure 4.1 (Kidzone.ws, 2012).



```

<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/html4/loose.dtd">
<html>

<!-- #BeginTemplate "../_kidzone_tb.dwt" -->

<head>
<!-- #BeginEditable "metadesc" -->
<meta http-equiv="Content-Type" content="text/html; charset=windows-1252">
<meta name="description" content="Fun facts for kids about the Water Cycle. Includes
photos, activity suggestions and some printable worksheets.">
<meta name="keywords" content="homework help, water cycle, precipitation,
evaporation, condensation, learning, printable worksheets, printable coloring pages">
<!-- #EndEditable -->
<!-- #BeginEditable "doctitle" -->
<title>The Water Cycle</title>
<!-- #EndEditable -->
<link href="../kidzonestyles/main.css" rel="stylesheet" type="text/css">
<!-- #BeginEditable "cssjs" -->
<link href="../kidzonestyles/watercycle.css" rel="stylesheet" type="text/css">
<!-- #EndEditable -->
<!-- Casale Media: Pop Under -->
<script type="text/javascript"><!--
var casaleD=new Date();var
casaleR=(casaleD.getTime()%8673806982)+Math.random();
var casaleU=escape(window.location.href);
var casaleHost=' type="text/javascript" src="http://as.casalemedia.com/s?s=';
document.write('<scr'+ipt'+casaleHost+'58882&amp;u=');
document.write(casaleU+'&amp;f=1&amp;id='+casaleR+'"><\scr'+ipt>');
//--></script>

```

**Figure 4.2** A water cycle activity HTML source code excerpt.

Another digital library collection, Econport, hosts economic resources appropriate for high school and college level students, including games and instructional texts. This web page describes David Ricardo's theory of Comparative Advantage, where nations should have specialized in products that could have been created using readily available resources and traded with other nations that have created other products; in this way, the total output of all nations would have increased. The following two figures show the displayed page as well as a sample of the HTML source code for this resource (NetMBA, 2012).

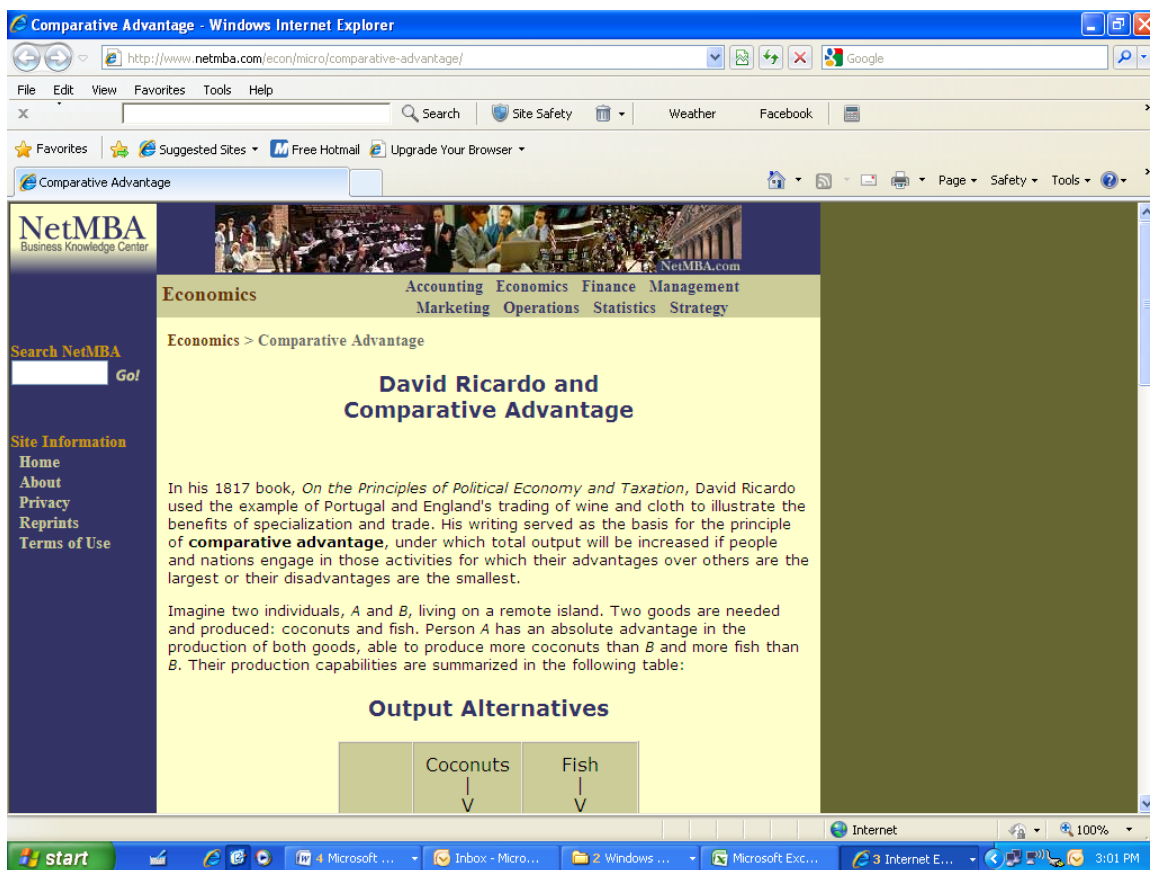


Figure 4.3 Screen capture of comparative advantage resource.

```

<body onselectstart="return false">
<div class="content">
<!--begincontent-->
<p
class="currentpath"><a
class="currentpath"
href="http://www.NetMBA.com/econ/">Economics</a> &gt; Comparative
Advantage</p>
<h3 class="title">David Ricardo and<br />Comparative Advantage</h3>
<br />
<p>In his 1817 book, <i>On the Principles of Political Economy and
Taxation</i>, David Ricardo used the example of Portugal and England's trading of wine
and cloth to illustrate the benefits of specialization and trade. His writing served as the
basis for the principle of <b>comparative advantage</b>, under which total output will
be increased if people and nations engage in those activities for which their advantages
over others are the largest or their disadvantages are the smallest.</p>
<p>Imagine two individuals, <i>A</i> and <i>B</i>, living on a remote island. Two
goods are needed and produced: coconuts and fish. Person <i>A</i> had an absolute
advantage in the production of both goods, able to produce more coconuts than <i>B</i>

```

Figure 4.4 HTML source code excerpt for comparative advantage resource.

If the text contained in the HTML source code, even after removing all tags and script information, would have been to be used to suggest the audience level of the resource, the audience level would have been much higher than actually warranted due to bullets, tables, menus, and other text common to all resources in the collection. However, the pages have been cataloged by a variety of collections and each collection structured the page differently, leading to an inability to remove all of this extraneous information and leaving only the text representing the resource content. Therefore, a broadly applicable audience level prediction system was needed to account for the common information, formatting tags, and structures present in the resource.

### **4.3 Digital Library Collection Overview**

Since not all resources in the digital library collections were associated with specific audience levels, only those resources that contained an expert-identified specific audience level were used in this evaluation. The resources used in this experiment were provided by a number of National Science Digital Library (NSDL) collections, normally targeting students and educators in grades kindergarten through twelfth grade, with additional resources provided by Springerlink to represent college level. Since the Springerlink collection contained many tens of thousands of college-level resources with audience level metadata, this collection was sampled to more evenly distribute resources among all audience levels. These libraries have typically held science, technology, engineering, and math (STEM) resources. The following table summarizes the digital libraries that have provided resources used in this evaluation.

**Table 4.1** Test Collection Digital Library Sources

<b>Collection</b>	<b>Web URL</b>	<b>Docs</b>
American Museum of Natural History	<a href="http://www.amnh.org/">http://www.amnh.org/</a>	200
BioMed Central	<a href="http://www.biomedcentral.com/">http://www.biomedcentral.com/</a>	27
Digital Library for Earth Systems Education (DLESE)	<a href="http://www.dlese.org">http://www.dlese.org</a>	391
Digital Library Network for Engineering and Technology (DLNET)	<a href="http://www.dlnet.vt.edu/">http://www.dlnet.vt.edu/</a>	137
Digital Library of Indigenous Science Resources (DLISR)	<a href="http://www.dliser.org/">http://www.dliser.org/</a>	17
Digital Library of Information Science and Technology (DLIST)	<a href="http://dlist.sir.arizona.edu/">http://dlist.sir.arizona.edu/</a>	3
EconPort	<a href="http://www.econport.org">http://www.econport.org</a>	267
Educational Benchmarks	<a href="http://strandmaps.nsdler.org/AAAS-Collection/NSDLbenchmarksContent.jsp">http://strandmaps.nsdler.org/AAAS-Collection/NSDLbenchmarksContent.jsp</a>	323
iCPalms	<a href="http://www.floridastandards.org">http://www.floridastandards.org</a>	268
Math Common Core	<a href="http://mixinginmath.terc.edu">http://mixinginmath.terc.edu</a>	222
Math Forum	<a href="http://www.mathforum.org">http://www.mathforum.org</a>	2,668
Math Landing	<a href="http://www.mpt.org/">http://www.mpt.org/</a>	198
Mathematics Gateway	<a href="http://mathgateway.maa.org/">http://mathgateway.maa.org/</a>	103
Middle School Portal: Math and Science Pathways	<a href="http://www.msteacher2.org">http://www.msteacher2.org</a>	623
My NASA Data	<a href="http://mynasadata.larc.nasa.gov/about_page.php">http://mynasadata.larc.nasa.gov/about_page.php</a>	29
Pacific Resources for Education and Learning	<a href="http://www.prel.org">http://www.prel.org</a>	357
SMILE Pathway	<a href="http://www.howtosmile.org/">http://www.howtosmile.org/</a>	94
Springerlink	<a href="http://www.springerlink.com/">http://www.springerlink.com/</a>	1,743
STEM Education Gateway	<a href="http://www.nsdler.org/collection/stem-education/">http://www.nsdler.org/collection/stem-education/</a>	117
Teach Engineering	<a href="http://www.teachengineering.com/">http://www.teachengineering.com/</a>	267
Teachers Domain	<a href="http://www.teachersdomain.org/">http://www.teachersdomain.org/</a>	2,094
The Teaching Company - Science and Mathematics Courses	<a href="http://www.teach12.com/storex/courses.aspx?t=&amp;sl=&amp;s=910&amp;sbj=Science%20and%20Mathematics&amp;fMode=s">http://www.teach12.com/storex/courses.aspx?t=&amp;sl=&amp;s=910&amp;sbj=Science%20and%20Mathematics&amp;fMode=s</a>	10
Tool Factory	<a href="http://www.toolfactory.com">http://www.toolfactory.com</a>	42
Trinity Remembered	<a href="http://www.trinityremembered.com">http://www.trinityremembered.com</a>	7
Visual Materials from the Tissandier Collection	<a href="http://lcweb2.loc.gov/pp/tischtml/tiscabt.html">http://lcweb2.loc.gov/pp/tischtml/tiscabt.html</a>	6
Web Adventures	<a href="http://webadventures.rice.edu/">http://webadventures.rice.edu/</a>	25
<b>Total Documents</b>		<b>10,238</b>

These resources spanned a wide range of audience levels, from kindergarten through college, including graduate audience levels. This wide variety of audience levels should have challenged SVMAUD to correctly predict the audience level for all resources in the collection. Since Springerlink and Project Euclid have held resources appropriate for undergraduate and graduate students, complete audience level information was not critical. However, with respect to digital libraries, such as Teacher's Domain and DLESE, a wide range of audience levels were covered, ranging from kindergarten through twelfth grade; complete and consistent audience level metadata was required to filter resources appropriate for the current user. Since 10,238 resources in the digital library test collection were associated with expert-labeled audience-level metadata, SVMAUD could have used the vocabulary in these resources to complete the missing or inconsistent audience level information for the remaining unlabeled resources in the collection.

As SVMAUD requires a large number of documents associated with each audience level in order to be effective, the distribution of resources across all audience levels are summarized in the next table. Since all classifiers generally perform well when each class contains a high number of unique terms, the proportion of words that appear only in a single class is also provided as well.

**Table 4.2** Unique Feature Words Summary

<b>Specific Audience Level</b>	<b>General Audience Level</b>	<b>Docs</b>	<b>Words That Appear in One Class</b>	<b>Total Words</b>	<b>Percent Unique</b>
Kindergarten	Early Elementary	698	17,539	58,733	29.86%
First	Early Elementary	719	7,894	32,295	24.44%
Second	Early Elementary	606	6,705	29,709	22.57%
Third	Late Elementary	418	7,348	31,777	23.12%
Fourth	Late Elementary	528	13,697	45,511	30.10%
Fifth	Late Elementary	532	11,789	40,751	28.93%
Sixth	Middle School	664	20,508	61,342	33.43%
Seventh	Middle School	663	30,020	78,247	38.37%
Eighth	Middle School	631	19,622	53,678	36.56%
Ninth	High School	693	33,869	76,958	44.01%
Tenth	High School	640	15,746	47,915	32.86%
Eleventh	High School	552	15,224	52,927	28.76%
Twelfth	High School	644	31,867	79,930	39.87%
UG Lower (Sampled)	College (Sampled)	750	26,574	64,327	41.31%
UG Upper (Sampled)	College (Sampled)	750	23,334	56,197	41.52%
Graduate (Sampled)	College (Sampled)	750	21,641	53,728	40.28%
<b>Total Documents</b>		<b>10,238</b>	<b>303,377</b>	<b>864,025</b>	<b>35.11%</b>

\*\* UG = Undergraduate

Lower audience levels typically contained a higher number of overlapping words between adjacent audience levels, since these students generally have a smaller and simpler vocabulary than a person taking a college level class. For this reason, the proportion of highly specialized words that only appeared in one class was much higher for resources at higher audience levels. Out of the entire test collection, 35% of the words were considered unique by only appearing in a single audience level. Therefore, SVMAUD and the other machine learning methods should have been used to suggest the audience level with high performance.

#### 4.4 Digital Library Audience Level Prediction Evaluation

For evaluation purposes, the documents in each class were divided into five different folds, with each fold containing an approximately equal number of resources. As an example, 640 resources were labeled with the tenth grade audience level so, for this audience level, 512 resources were used for training while the remaining fold was used for testing. Then, this process was repeated five different times until each fold of resources was used once for the testing part of the evaluation.

Four different evaluations are conducted in this experiment. First, the performance of two readability formulas, the Flesch-Kincaid Reading Age and Dale-Chall Reading Ease Score, are compared to the results provided by the SVM-based classifier SVMAUD. Second, the performance between two classification methods, cosine and Naïve Bayes, is compared with SVMAUD performance. Next, the Collins-Thompson and Callan method audience level prediction performance is compared with the performance of SVMAUD. Finally, SVMAUD is trained and tested using the inputs to the readability formulas to determine whether textual characteristics, such as average syllables per word and average sentence length, are reliable indicators of the difficulty of a resource. All of these evaluations use the standard classification performance measurements of precision, recall, and F-measure. In addition, the correlation between human-entered and suggested values by the machine learning or readability formula method is also provided, since an incorrect prediction of plus or minus one audience level is less of an error than an incorrect prediction of plus or minus five audience levels. The t-test level of significance at which SVMAUD outperforms the other computer-based methods or readability formulas under evaluation is also provided for each evaluation.

#### 4.4.1 Readability Formulas versus SVMAUD

This evaluation considered the audience level performance of two common readability formulas against the prediction performance of SVMAUD. Probably the most popular formula was the Flesch-Kincaid Reading Age that relied upon the number of words per sentence and the average syllables per word to suggest the most appropriate audience level for the resource. This formula used these parameters, combined with constants derived from regression analysis, to predict the number of years of formal education required to understand the resource:

$$FKRA = (0.39 \times ASL) + (11.8 \times ASW) - 15.59 \quad (4.2)$$

The Dale-Chall Reading Ease score took a different approach by comparing the vocabulary chosen by the author against a list of 3,000 words that should have been learned by the average fourth grade student; in addition, this formula considered the average sentence length in the document. The Dale Chall Reading Ease formula is given as follows:

$$R = 0.1579 (\text{Proportion of Words Not in Dale Common Word List} * 100) + 0.0496 \left( \frac{\# \text{ of Words}}{\# \text{ of Sentences}} \right) \quad (\text{Add } 3.6365 \text{ if } > 5\% \text{ of words not in Dale List}) \quad (4.3)$$

The Dale-Chall Reading Ease Score was the most similar readability formula to SVMAUD among those previously reviewed, since it also considered the vocabulary chosen by the author rather than only relying on word and sentence characteristics. Two different evaluations were conducted to measure the performance of these readability formulas against SVMAUD. The first evaluation considered the performance when predicting general audience levels, while the second evaluation considered specific audience levels.



The precision, recall, and F-measure were measured with respect to the Dale-Chall Reading Ease Score, the Flesch-Kincaid Reading Age, and SVMAUD in their ability to correctly predict the human-expert entered general audience levels. For all of these evaluations, P referred to Precision, R referred to Recall, and F referred to F-Measure; these methods were the standard classification prediction performance measures. The next table summarizes the results from the first part of the study.

**Table 4.3** Readability Formulas vs. SVMAUD – General Audience Levels

General Audience Level	Docs	Flesch-Kincaid			Dale-Chall			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	2,023	0.16	0.03	0.05	0.00	0.00	0.00	0.86	0.89	0.88
Late Elementary	1,478	0.05	0.00	0.00	0.10	0.02	0.04	0.92	0.75	0.83
Middle School	1,958	0.18	0.02	0.04	0.18	0.01	0.02	0.81	0.85	0.83
High School	2,529	0.29	0.43	0.34	0.02	0.01	0.01	0.82	0.89	0.85
College (Sampled)	2,250	0.30	0.77	0.43	0.19	0.67	0.30	0.98	0.94	0.96
<b>Overall</b>	<b>10,238</b>	<b>0.28</b>	<b>0.28</b>	<b>0.28</b>	<b>0.16</b>	<b>0.16</b>	<b>0.16</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The correlation between human-expert entered and readability-formula suggested audience level values was found to be 0.05 for Flesch-Kincaid and 0.10 for the Dale-Chall Reading Ease Score. In this evaluation, the Flesch-Kincaid Reading Age experienced the poorest performance by predicting the human expert identified audience level with an F-measure of only 0.28. The Dale-Chall Reading Ease Score also experienced extremely poor performance with an overall F-measure of 0.16. However, SVMAUD far outperformed these readability formulas by correctly identifying the general audience level with an F-measure of 0.87. The correlation between human-expert

entered and readability-formula suggested values was found to be extremely low, indicating that the incorrect predictions were far away from the human-expert entered values. In fact, the audience level predictions mainly fell into the college level, since web pages contained a number of tables, menus, and bullet points that distorted the average sentence length far upward. Since SVMAUD relied on the vocabulary chosen by the author rather than the sentence structures in the text, its performance was found to be much higher than the readability formulas under evaluation; in fact, SVMAUD outperformed both readability formulas at the 0.0004 level of significance.

In a digital library setting, the resources should have been associated with the most specific audience level to best match resources to users. If the resource was stored with the specific audience level, such as first grade, second grade, etc., and the user would have desired elementary school resources, the retrieval system could have presented all resources in grades one through five. However, if each resource was stored with its general audience level and the user required early elementary resources in grades kindergarten through second, the retrieval system could not have used the audience level metadata to further refine the retrieved resources. The next evaluation considers the ability of the readability formulas and SVMAUD to correctly predict the human-expert provided specific audience levels; the results from this study are shown in table 4.4 on the next page.

**Table 4.4** Readability Formulas vs. SVMAUD – Specific Audience Levels

Specific Audience Level	Docs	Flesch-Kincaid			Dale-Chall			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	698	0.08	0.01	0.02	0.00	0.00	N/A	0.61	0.78	0.68
First	719	0.00	0.00	N/A	0.00	0.00	N/A	0.91	0.80	0.85
Second	606	0.05	0.00	0.01	0.00	0.00	N/A	0.56	0.92	0.70
Third	418	0.06	0.03	0.04	0.01	0.00	0.00	0.99	0.66	0.79
Fourth	528	0.00	0.00	N/A	0.03	0.00	0.01	0.81	0.71	0.76
Fifth	532	0.03	0.00	0.00	0.08	0.02	0.03	0.99	0.63	0.77
Sixth	664	0.03	0.00	0.00	0.12	0.01	0.02	0.99	0.68	0.81
Seventh	663	0.05	0.01	0.01	0.00	0.00	N/A	0.89	0.67	0.77
Eighth	631	0.06	0.01	0.02	0.00	0.00	N/A	0.46	0.85	0.60
Ninth	693	0.03	0.01	0.02	0.01	0.00	0.00	0.79	0.82	0.81
Tenth	640	0.09	0.08	0.08	0.01	0.01	0.01	0.96	0.80	0.87
Eleventh	552	0.07	0.17	0.10	0.00	0.00	0.00	0.90	0.77	0.83
Twelfth	644	0.06	0.14	0.08	0.00	0.00	0.00	0.77	0.81	0.79
UG Lower (Sampled)	750	0.07	0.24	0.11	0.16	0.19	0.17	0.75	0.93	0.83
UG Upper (Sampled)	750	0.07	0.13	0.09	0.14	0.12	0.13	1.00	0.75	0.85
Graduate (Sampled)	750	0.21	0.57	0.31	0.04	0.32	0.07	0.94	0.81	0.87
<b>Overall</b>	<b>10,238</b>	<b>0.10</b>	<b>0.10</b>	<b>0.10</b>	<b>0.05</b>	<b>0.05</b>	<b>0.05</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>

\*\* UG = Undergraduate; N/A = Not Applicable (F-Measure could not be calculated due to precision and recall of zero)

Both the Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score performed extremely poorly in this evaluation, correctly predicting the human-expert identified audience level with F-measures under 0.10. In addition, the correlation between human-expert entered and readability formula suggested values was extremely poor, with a correlation of 0.13 for Flesch-Kincaid and 0.07 for Dale-Chall. SVMAUD far outperformed both of these methods once again with an overall F-measure of 0.78; in fact, SVMAUD performance exceeded the performance of both readability formulas at the 0.0001 level of significance. The prediction performance decreased when compared

with general audience levels, since the classifier had a larger number of classes, or audience levels, with fewer documents in each class for training samples. The Flesch-Kincaid Reading Age labeled the majority of the documents with the college audience level due to the parameter average sentence length distorting the true audience level far upward; HTML pages generally contained tables, figures, lists, and other attributes that have not required end-of-sentence tokens, resulting in a much longer average sentence length. As the Dale-Chall Reading Ease Score used the set of words understood by an average fourth grade student, the prediction performance was highest with respect to the fourth through sixth grade levels; in addition, since this formula relied on the sentence length calculation that was much higher than warranted, a large number of resources were labeled with college audience levels. Since SVMAUD did not rely on numerical representations of text characteristics, but rather on the author's chosen vocabulary, its performance was much higher than readability formulas at a high level of significance.

As the documents used in this study were originally HTML documents that contained headers, footers, tables, and figures in addition to the full text commonly found in books, the performance of the readability formulas was severely impacted due to the sentence length parameter. The number of sentences, as calculated by counting the number of end-of-sentence tokens and line breaks, was much lower than actually present in the document. Some digital library resources used tables, figures, and lists in the body; this text format did not follow the grammatical and sentence conventions of traditional written English. SVMAUD had not suffered from these same limitations since, rather than relying on the structural characteristics of the text, this program relied on the vocabulary chosen by the author.

In all of the experiments conducted as part of this evaluation, SVMAUD outperformed both readability formulas and other machine learning methods at high levels of significance. Both the Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score experienced extremely poor performance due to their reliance on the number of words per sentence as part of the formula, with F-measures under 0.30 for general audience levels and correlations less than 0.10. With respect to digital libraries, in addition to the full text that typically followed grammatical and sentence conventions, resources also contained headers, footers, bullet points, sentence fragments, and other elements that distorted the average words per sentence calculation. These formulas have not matched the vocabulary chosen by the author with the vocabulary appropriate for each audience level. These readability formulas were able to correctly predict the audience level for resources with F-measures less than 0.30 for general, and 0.10 for specific audience levels. SVMAUD was found to outperform the two readability formulas under evaluation at the 0.0004 level for general audience levels and 0.0001 for specific audience levels. Even though the Dale-Chall Reading Ease Score incorporated the vocabulary appropriate for each audience level, the calculation was distorted upward by the sentence length parameter.

#### **4.4.2 Cosine and Naïve Bayes versus SVMAUD**

The baseline method in this study is cosine. This classification algorithm measures the cosine between an unlabeled document and all terms associated with each class extracted from the training samples. This formula describes the calculation of the cosine between an unlabeled document and the class by using the formula:

$$\begin{aligned} \text{Cosine}(d, c) &= \frac{\sum_K d_K c_K}{\left(\sqrt{\sum_K d_K^2}\right)\left(\sqrt{\sum_K c_K^2}\right)} \\ d_K &= \sum_K c_K * \log\left(1 + \frac{16}{N_C}\right) \end{aligned} \quad (4.4)$$

In this formula,  $d_k$  represents the frequency of feature  $k$  in document vector  $d$ ,  $c_k$  represents the frequency of the feature  $k$  in category vector  $c$ , and  $N_C$  represents the number of classes in which the feature term  $c_k$  occurs. The document is assigned to the class where the cosine is the smallest between the document and class, indicating that the terms in the document are most similar to the specific class.

The Naïve Bayes machine learning method used term frequency to suggest the probability that a document belonged to a particular class. Stop words were removed from documents before training the method, as they occurred in many different classes and reduced the importance of more important feature words. Stemming was completed as the results were found to improve slightly across all audience levels. This model predicted the difficulty of understanding a particular text  $T$  relative to the grade level  $G_i$  by calculating the probability that the language model of the particular grade represented the words contained in the unlabeled text.

$$P(T | G_i) = P(|T|) |T|! \prod_{w \in V} \frac{P(w | G_i)^{C(w)}}{C(w)!} \quad (4.5)$$

In this formula,  $V$  represents the vocabulary for grade  $G_i$ ,  $w$  represents one of the key terms in  $V$ , and  $C(w)$  represents the entire tokens in the text  $T$  containing words similar to  $w$ . The resource is labeled with the audience level whose terms have the highest probability of generating the written work.

Similar to the readability formula study previously described, this study compares the performance of these three classifiers, namely cosine, Naïve Bayes, and SVMAUD,

based on their prediction performance with respect to precision, recall, F-measure, and correlation between human-expert and machine-learning suggested audience levels. The first part of this study compares the prediction performance for general audience levels, while the second part considers specific audience levels. The results from the general audience level prediction performance study are shown in the following table:

**Table 4.5** Cosine vs. Naïve Bayes vs. SVMAUD – General Audience Levels

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	2,023	0.60	0.65	0.62	0.65	0.72	0.68	0.86	0.89	0.88
Late Elementary	1,478	0.58	0.26	0.36	0.67	0.55	0.60	0.92	0.75	0.83
Middle School	1,958	0.47	0.55	0.50	0.51	0.69	0.59	0.81	0.85	0.83
High School	2,529	0.55	0.66	0.60	0.84	0.62	0.71	0.82	0.89	0.85
College (Sampled)	2,250	0.93	0.84	0.88	0.90	0.87	0.89	0.98	0.94	0.96
<b>Overall</b>	<b>10,238</b>	<b>0.62</b>	<b>0.62</b>	<b>0.62</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

SVMAUD was found to outperform both cosine and Naïve Bayes classification methods in this part of the study. The cosine-based classifier performed worst with an F-measure of 0.62. The Naïve Bayes classifier performed slightly better with an overall F-measure of 0.70. However, once again, SVMAUD outperformed both of these methods with an overall F-measure of 0.87. In addition, the correlation between human-expert entered and machine-learning suggested values also greatly increased over the readability formulas, with a correlation of 0.74 for cosine and 0.77 for Naïve Bayes; however, once again, the SVMAUD correlation measure outperformed these other two methods with a

correlation of 0.91. SVMAUD was found to far outperform both cosine and Naïve Bayes machine learning methods at the 0.0150 level of significance.

The next part of this evaluation considers the performance of these three machine learning methods when predicting the specific audience level for resources in the test collection. The results from this part of the study are summarized in the following table.

**Table 4.6** Cosine vs. Naïve Bayes vs. SVMAUD – Specific Audience Levels

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	698	0.38	0.60	0.46	0.43	0.63	0.51	0.61	0.78	0.68
First	719	0.81	0.63	0.71	0.80	0.64	0.71	0.91	0.80	0.85
Second	606	0.42	0.85	0.56	0.43	0.87	0.58	0.56	0.92	0.70
Third	418	0.97	0.38	0.54	0.98	0.41	0.58	0.99	0.66	0.79
Fourth	528	0.57	0.49	0.53	0.64	0.52	0.57	0.81	0.71	0.76
Fifth	532	0.98	0.34	0.50	0.99	0.38	0.55	0.99	0.63	0.77
Sixth	664	0.97	0.36	0.52	0.98	0.39	0.56	0.99	0.68	0.81
Seventh	663	0.69	0.39	0.50	0.73	0.44	0.55	0.89	0.67	0.77
Eighth	631	0.24	0.65	0.36	0.27	0.69	0.39	0.46	0.85	0.60
Ninth	693	0.60	0.67	0.63	0.64	0.71	0.67	0.79	0.82	0.81
Tenth	640	0.89	0.58	0.71	0.90	0.63	0.74	0.96	0.80	0.87
Eleventh	552	0.77	0.57	0.65	0.77	0.61	0.68	0.90	0.77	0.83
Twelfth	644	0.55	0.57	0.56	0.62	0.65	0.63	0.77	0.81	0.79
UG Lower (Sampled)	750	0.54	0.83	0.66	0.58	0.87	0.70	0.75	0.93	0.83
UG Upper (Sampled)	750	0.99	0.47	0.63	0.99	0.52	0.68	1.00	0.75	0.85
Graduate (Sampled)	750	0.86	0.56	0.68	0.88	0.60	0.71	0.94	0.81	0.87
<b>Overall</b>	<b>10,238</b>	<b>0.57</b>	<b>0.57</b>	<b>0.57</b>	<b>0.61</b>	<b>0.61</b>	<b>0.61</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>

\*\* UG = Undergraduate

In this part of the study, the cosine and Naïve Bayes prediction performance decreased slightly to F-measures of 0.57 and 0.61 for cosine and Naïve Bayes,



respectively. However, SVMAUD again outperformed these other methods with an overall F-measure of 0.78. The correlation between human-expert entered and machine-learning suggested values slightly increased over general audience levels, from 0.74 to 0.77 for cosine and 0.77 to 0.79 for Naïve Bayes. However, once again, SVMAUD was found to outperform both the cosine and Naïve Bayes methods under evaluation at the 0.0001 level of significance.

All of these machine learning methods outperformed the readability formulas presented in the first part of this evaluation. Rather than relying on the structure of text to reason the difficulty or ease of understanding the document content, these methods relied on the vocabulary chosen by the author. The cosine and Naïve Bayes classifiers correctly predicted the specific audience level with F-measures of 0.57 and 0.61, respectively. However, SVMAUD again outperformed these machine learning methods by correctly predicting the general audience level with F-measures of 0.87 for general and 0.78 for specific audience levels. All of these machine learning methods predicted the audience level of the resources with far higher performance than readability formulas.

#### **4.4.3 Collins-Thompson and Callan Method versus SVMAUD**

In their 2005 paper, Collins-Thompson and Callan modified the Naïve Bayes machine learning method to improve its performance at predicting the audience level for web-based resources. Stop words were not removed as they tended to occur more frequently at lower audience levels; stemming was completed to reduce the number of unique terms. In addition, all words that occurred only in one class or all words that appeared only once in the training dataset were removed. In the standard Naïve Bayes model, the probability for all terms was based on the number of occurrences of each term in the training dataset;

the Collins-Thompson and Callan method took a different approach by reducing the importance of highly occurring terms, such as stop words, and implementing the simple Good-Turing method to smooth the word frequency data in each class by adjusting term frequencies (Gale, 1995). Modifying the base Naïve Bayes equation using a mixture model of nearby classes on the logarithmic scale resulted in the following formula; the mixture model considered word frequencies across different audience levels as words typically appeared in more than one class:

$$\log P(G_i | T) = \sum_{w \in V} C(w) \log P(w | G_i) - \sum_{w \in V} \log C(w)! + \log \left( \frac{1}{N_G} \right) + \log S \quad (4.6)$$

$$L(T | G_i) \underset{G}{\overset{G_i}{>}} \sum C(w) \log(P | G_i)$$

In this modified formula,  $N_G$  represented the number of grade levels (twelve in the paper) and  $S$  represented the contribution of the passage length.

After the training process was completed, this method could now have been used to predict the audience level for unlabeled resources in the collection. As the difficulty of understanding the resource content varied across different sections, the unlabeled resource was split into chunks of one hundred words in length, and the probability that the training class terms were used to create the one hundred word chunk was calculated and stored; then, the two highest probabilities for the word chunks found in each audience level were averaged to represent the probability that the document terms were drawn from the training class terms. The probabilities that each unlabeled document belonged to a particular class were sorted in ascending order (Collins-Thompson & Callan, 2005). Since a reader's comprehension typically peaked when he or she

understood 75% of the words (Stenner, 1992), the most likely audience level was chosen to be the one that had occurred at the 75<sup>th</sup> percentile of this distribution. The findings showed a root mean squared error of between one and two grades for nine out of twelve grades and a correlation between human-expert identified and machine-suggested audience levels of 0.69 for grades one through six and 0.79 for grades one through twelve (Collins-Thompson & Callan, 2005).

The performance of the Collins-Thompson and Callan method was compared with the performance of SVMAUD based on the precision, recall, and F-measure for each audience level. The following table summarizes the results from this study.

**Table 4.7** Collins-Thompson and Callan vs. SVMAUD – General Audience Levels

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	2,023	0.80	0.83	0.81	0.86	0.89	0.88
Late Elementary	1,478	0.87	0.64	0.74	0.92	0.75	0.83
Middle School	1,958	0.71	0.77	0.74	0.81	0.85	0.83
High School	2,529	0.75	0.82	0.79	0.82	0.89	0.85
College (Sampled)	2,250	0.96	0.92	0.94	0.98	0.94	0.96
<b>Overall</b>	<b>10,238</b>	<b>0.81</b>	<b>0.81</b>	<b>0.81</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The additional language modeling performed by the Collins-Thompson and Callan method improved the F-measure prediction performance from 0.70 using simple Naïve Bayes to 0.81 in this study. The correlation between human-expert entered and machine-suggested values also improved from 0.77 for Naïve Bayes to 0.87 for the Collins-Thompson and Callan method. Even though the performance improved,

SVMAUD still outperformed this language modeling method, with an overall F-measure of 0.87 and a correlation of 0.91 between human-expert entered and SVMAUD-suggested audience levels. SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0931 level of significance using this test collection.

The next study compared the ability of the Collins-Thompson and Callan method and SVMAUD to correctly predict the human-expert entered specific audience level. The results from this study are shown in the following table.

**Table 4.8** Collins-Thompson and Callan vs. SVMAUD – Specific Audience Levels

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	698	0.49	0.69	0.57	0.61	0.78	0.68
First	719	0.86	0.72	0.78	0.91	0.80	0.85
Second	606	0.50	0.88	0.64	0.56	0.92	0.70
Third	418	0.99	0.57	0.72	0.99	0.66	0.79
Fourth	528	0.69	0.61	0.65	0.81	0.71	0.76
Fifth	532	1.00	0.50	0.66	0.99	0.63	0.77
Sixth	664	0.99	0.53	0.69	0.99	0.68	0.81
Seventh	663	0.81	0.55	0.65	0.89	0.67	0.77
Eighth	631	0.34	0.75	0.47	0.46	0.85	0.60
Ninth	693	0.71	0.76	0.74	0.79	0.82	0.81
Tenth	640	0.93	0.71	0.80	0.96	0.80	0.87
Eleventh	552	0.84	0.67	0.74	0.90	0.77	0.83
Twelfth	644	0.66	0.70	0.68	0.77	0.81	0.79
UG Lower (Sampled)	750	0.63	0.88	0.74	0.75	0.93	0.83
UG Upper (Sampled)	750	0.99	0.61	0.75	1.00	0.75	0.85
Graduate (Sampled)	750	0.91	0.68	0.78	0.94	0.81	0.87
<b>Overall</b>	<b>10,238</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>

\*\* UG = Undergraduate

Once again, SVMAUD outperformed the 0.68 F-measure prediction performance of the Collins-Thompson and Callan method by correctly predicting the human-expert entered audience level with an overall F-measure of 0.78. The Collins-Thompson and Callan method also improved the correlation between human-expert entered and machine-learning suggested from 0.79 for simple Naïve Bayes to 0.83; however, SVMAUD, with a correlation of 0.88, still outperformed the Collins-Thompson and Callan method with respect to the correlation between human-expert entered and machine-learning predicted values. SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0013 level of significance. In all of these performance comparisons, SVMAUD far outperformed the Collins-Thompson and Callan method.

#### **4.4.4 Machine Learning Using Readability Formula Inputs**

Since the readability formulas required simple textual characteristics, such as syllables per word and sentence length, to be plugged into a formula to suggest the audience level for a resource, this study trained SVMAUD by using the inputs to the readability formulas to predict the audience level for unlabeled resources. The Flesch-Kincaid Reading Age required the average syllables per word and the average sentence length as inputs, while the Dale-Chall Reading Ease Score required the proportion of difficult words and the average sentence length as inputs. Since SVMAUD required the formation of a term-document matrix, where all occurrences of each term in each document were counted and stored in each cell, the two inputs to each formula were calculated and stored for each resource in the training and testing collection. With respect to the two readability formulas, SVMAUD was used to predict the audience level for each resource using the inputs from the readability formulas; for comparison purposes, the prediction

performance of SVMAUD when trained and tested using the full text was also measured.

The following table displays the general audience level prediction performance.

**Table 4.9** SVMAUD using Readability Formula Inputs – General Audience Levels

General Audience Level	Docs	Flesch-Kinkaid			Dale-Chall			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	2,023	0.65	0.11	0.19	1.00	0.13	0.23	0.86	0.89	0.88
Late Elementary	1,478	0.46	0.11	0.18	0.45	0.14	0.21	0.92	0.75	0.83
Middle School	1,958	0.59	0.12	0.19	0.80	0.14	0.24	0.81	0.85	0.83
High School	2,529	0.33	0.48	0.39	0.16	0.14	0.15	0.82	0.89	0.85
College (Sampled)	2,250	0.32	0.79	0.46	0.23	0.71	0.35	0.98	0.94	0.96
<b>Overall</b>	<b>10,238</b>	<b>0.35</b>	<b>0.35</b>	<b>0.35</b>	<b>0.26</b>	<b>0.26</b>	<b>0.26</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The correlation between human-expert and method-suggested audience levels was found to be 0.18 for SVMAUD using Flesch-Kinkaid Reading Age inputs, 0.26 for SVMAUD when using Dale-Chall Reading Ease Score inputs, and 0.88 for SVMAUD when using the full text. SVMAUD, when trained and tested using full text, was found to outperform SVMAUD when using Flesch-Kinkaid Reading Age inputs at the 0.17 level of significance and the Dale-Chall Reading Ease Score at the 0.07 level of significance. While SVMAUD, when using full text as input, far outperformed both readability formulas, the prediction performance of the two readability formulas when using SVMAUD to suggest audience level, rather than using constants derived from regression analysis, also improved. When plugging the average syllables per word and the average sentence length into the Flesch-Kinkaid Reading Age formula, the overall F-measure

prediction performance was found to be 0.28, versus 0.35 when SVMAUD was used to predict the audience level using the same inputs. Similarly, the Dale-Chall Reading Ease Score overall F-measure prediction performance increased from 0.16, when using the original formula, to 0.26 when using SVMAUD to predict the audience level.

The next part of the study considered the specific audience level prediction performance of the Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score when SVMAUD was trained and tested using the inputs to the readability formulas. Table 4.10 on the following page summarized the results for the specific audience level prediction performance. SVMAUD, using full text for training and testing, far outperformed the prediction performance of SVMAUD when trained and tested using the inputs to the two readability formulas. The correlation between human-expert and SVMAUD suggested audience levels was 0.20 for Flesch-Kincaid Reading Age inputs, 0.15 for Dale-Chall Reading Ease Score inputs, and 0.91 for SVMAUD when using full text for training and testing. SVMAUD was found to outperform SVMAUD when using Flesch-Kincaid Reading Age inputs at the 0.0035 level of significance, and SVMAUD when using Dale-Chall Reading Ease Score inputs at the 0.0018 level of significance. However, the performance of both readability formulas improved when using SVMAUD to predict audience level rather than using the constants found by using regression analysis in the original formula.

**Table 4.10** SVMAUD using Readability Formula Inputs – Specific Audience Levels

Specific Audience Level	Docs	Flesch-Kinkaid			Dale-Chall			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	698	0.30	0.06	0.10	1.00	0.05	0.10	0.61	0.78	0.68
First	719	0.96	0.06	0.12	1.00	0.06	0.12	0.91	0.80	0.85
Second	606	0.53	0.06	0.11	1.00	0.05	0.10	0.56	0.92	0.70
Third	418	0.17	0.09	0.12	0.18	0.08	0.11	0.99	0.66	0.79
Fourth	528	0.80	0.07	0.12	0.31	0.06	0.10	0.81	0.71	0.76
Fifth	532	0.57	0.07	0.13	0.32	0.08	0.13	0.99	0.63	0.77
Sixth	664	0.64	0.07	0.13	0.51	0.07	0.13	0.99	0.68	0.81
Seventh	663	0.44	0.08	0.13	0.61	0.06	0.10	0.89	0.67	0.77
Eighth	631	0.36	0.09	0.14	0.74	0.06	0.10	0.46	0.85	0.60
Ninth	693	0.20	0.08	0.11	0.32	0.06	0.11	0.79	0.82	0.81
Tenth	640	0.17	0.13	0.15	0.12	0.07	0.09	0.96	0.80	0.87
Eleventh	552	0.09	0.24	0.13	0.04	0.05	0.04	0.90	0.77	0.83
Twelfth	644	0.08	0.18	0.11	0.04	0.06	0.05	0.77	0.81	0.79
UG Lower (Sampled)	750	0.10	0.30	0.15	0.20	0.24	0.22	0.75	0.93	0.83
UG Upper (Sampled)	750	0.11	0.19	0.14	0.19	0.18	0.18	1.00	0.75	0.85
Graduate (Sampled)	750	0.24	0.60	0.34	0.05	0.37	0.08	0.94	0.81	0.87
<b>Overall</b>	<b>10,238</b>	<b>0.16</b>	<b>0.16</b>	<b>0.16</b>	<b>0.11</b>	<b>0.11</b>	<b>0.11</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>

\*\* UG = Undergraduate

The Flesch-Kinkaid Reading Age audience level prediction performance improved from an overall F-measure of 0.10 when using the original formula, to 0.16 when using SVMAUD to suggest audience level; similarly, the Dale-Chall Reading Ease Score overall F-measure improved from 0.05 when using the original formula, to 0.11 when using SVMAUD to predict audience level.

Even though the readability formulas were tested on a large number of resources in order to obtain the constants to be multiplied by each variable, the constants still needed to be adjusted for the digital library collection used in this evaluation; SVMAUD



did not require this adjustment to be made. In addition, the readability formulas required all inputs to either increase or decrease at approximately the same rate between adjacent audience levels; SVMAUD did not require this relationship to be followed. The average sentence length, average syllable counts, and proportion of difficult words were not found to be reliable indicators of the audience level of the resources in the test digital library collection.

Since both readability formulas required the inputs to increase or decrease at the same rate among adjacent audience levels, this part of the study compared the Flesch-Kinkaid Reading Age and the Dale-Chall Reading Ease Score inputs versus each audience level. The following table shows the results from this study.

**Table 4.11** Average Readability Formula Inputs - General Audience Levels

General Audience Level	Docs	Word Count		Syllables/ Word		% Difficult Words		Sentence Length	
		Avg	SD	Avg	SD	Avg	SD	Avg	SD
Early Elementary	2,023	1,143	1,859	2.11	0.43	40%	13%	11	6
Late Elementary	1,478	945	1,259	2.10	0.52	44%	15%	12	15
Middle School	1,958	1,606	2,107	2.02	0.57	49%	15%	13	10
High School	2,529	1,666	2,364	2.04	0.52	49%	15%	15	21
College (Sampled)	2,250	841	769	2.07	0.40	33%	11%	20	12

\*\* Avg=Average; SD=Standard Deviation

The average among all resources in the human-expert audience level of the particular readability formulas input were shown. For example, as the audience level increased, the average syllables per word should have also increased. The Flesch-

Kinkaid Reading Age required the average syllables per word and the average sentence length as inputs; on the other hand, the Dale-Chall Reading Ease Score required the proportion of difficult words and the average sentence length as inputs. For the readability formulas to perform well, the averages should have increased or decreased at approximately the same rate as the audience level increased, and the standard deviation should have been small. However, the syllables per word and the proportion of difficult words did not increase at the same rate between adjacent audience levels; in fact, the lowest proportion of difficult words was found in the college level, where the difficult words proportion should have been highest according to the readability formulas. On the other hand, the average sentence length increased as the audience level increased, but the standard deviation was very large, indicating that the average sentence length could have varied by well over 21 words in the high school audience level.

The next part of the study considered the inputs to the readability formulas when averaged for each specific audience level. The results from this study are shown in Table 4.12 on the next page. Due to an assumption of a linear relationship between document attributes and audience level, the syllables per word, proportion of difficult words, and average sentence length should have increased at the same rate from one audience level to the following one. However, even though the average syllable count remained roughly two per word, the average syllable count increased and decreased as the audience level increased. The proportion of difficult words also increased and decreased as the audience level had increased. The lowest proportion of difficult words occurred at the graduate level with one-third difficult words; the proportion of difficult words should have been highest at this audience level.

**Table 4.12** Average Readability Formula Inputs - Specific Audience Levels

Specific Audience Level	Docs	Word Count		Syllables/Word		% Difficult Words		Sentence Length	
		Avg	SD	Avg	SD	Avg	SD	Avg	SD
Kindergarten	698	1,841	2,535	2.05	0.48	44%	15%	12	6
First	719	758	1,170	2.13	0.43	39%	13%	11	7
Second	606	796	1,289	2.15	0.37	38%	11%	10	3
Third	418	953	1,310	2.10	0.47	44%	12%	11	12
Fourth	528	1,034	1,304	2.10	0.61	45%	15%	14	23
Fifth	532	849	1,165	2.10	0.47	43%	15%	12	6
Sixth	664	1,382	1,965	2.04	0.51	48%	14%	13	12
Seventh	663	2,159	2,602	1.99	0.57	50%	16%	12	6
Eighth	631	1,261	1,460	2.03	0.61	50%	14%	13	10
Ninth	693	2,195	2,384	2.04	0.50	47%	14%	17	37
Tenth	640	973	1,905	2.00	0.51	51%	14%	14	7
Eleventh	552	1,359	2,186	2.05	0.52	52%	14%	13	8
Twelfth	644	2,047	2,673	2.06	0.54	46%	16%	16	9
UG Lower (Sampled)	750	994	1,101	2.08	0.36	34%	10%	18	10
UG Upper (Sampled)	750	763	428	2.10	0.41	34%	11%	19	11
Graduate (Sampled)	750	767	590	2.03	0.42	33%	13%	24	14

\*\* UG = Undergraduate; Avg=Average; SD=Standard Deviation

Similarly, the sentence length had not followed a consistent pattern as the audience level had increased. Since the inputs to the two readability formulas have not followed a consistent pattern, the formulas could not have predicted the audience level with high performance.

#### 4.5 Digital Library Audience Level Prediction Discussion

This section summarizes and discusses the results from the digital library audience level prediction study. The first part of this section considers the readability formulas versus

SVMAUD study, while the second part discusses the results from the machine learning methods evaluation.

#### **4.5.1 Readability Formulas vs. SVMAUD**

In general, readability formulas considered the semantic and syntactic features present in the text to predict the difficulty of understanding the text. Rather than considering the vocabulary chosen by the author, other aspects, such as sentence length, syllables per word, and characters per word, were used to predict the audience level for written works. While these formulas were able to predict the audience level of books and other textual works that followed grammatical and sentence conventions, web-based digital library resources posed a new set of challenges that could not have been easily solved. These documents were typically shorter than books and, in addition to the full text, contained headers, footers, and even scripts that distorted the true audience level. In particular, the average sentence length was calculated to be much higher than had been warranted, since the proportion of end-of-sentence tokens was much lower in web-based documents when compared with published books. In addition, readability formulas have not considered the vocabulary chosen by the author; simple words such as “television” contained more syllables than more complex words such as “rhinitis,” but “television” should have been associated with a much lower audience level than “rhinitis.”

The Flesch-Kincaid Reading Age relied on the average number of words per sentence, and the average number of syllables per word, in addition to coefficients derived from regression analysis to predict the audience level. The Dale-Chall Reading Ease Score, on the other hand, considered the vocabulary chosen by the author as part of the formula by comparing the words in the document, with a list of 3,000 words

commonly known to a fourth grade student. As the proportion of words on this list decreased, the audience level should have increased. However, since this formula also considered the average sentence length, the reading difficulty was much higher than warranted, since many traditional end-of-sentence tokens have not appeared in web documents at the same rate as written English. The performance of these methods should have improved if the abstract was used for suggesting the audience level of the resource; however, the abstract had not been completed for most resources in the collection, and was much shorter than the full text of the resource. In addition, this study sought to develop a broadly applicable program that could have predicted the audience level for any web-based digital library resource; if the abstract or keywords were missing for these resources, they could not have been used to suggest the most appropriate audience level.

Whereas readability formulas focused on the structure of the text, SVMAUD took a different approach by considering the vocabulary chosen by the author, which was compared to the terms present in a set of predefined classes. Since the end-of-sentence tokens were ignored, the audience level prediction performance was found to have increased over that of readability formulas. However, this method required a set of resources that have been pre-labeled with the most appropriate audience level, whereas the readability formulas only required the user to perform some simple calculations to arrive at the audience level. The readability formulas were able to calculate the most appropriate audience level with much less user input. Even though the readability formulas were much easier to use and are domain-independent, once SVMAUD was trained using a number of resources appropriate for each audience level, its performance was found to be much higher by matching the vocabulary contained in the unlabeled

resource with the predefined vocabulary drawn from human-expert labeled resources appropriate for each audience level. The initial cost of creating a training dataset of documents appropriate for each audience level would have been balanced by the increased performance available by using SVMAUD to label all resources with missing or incompatible audience level information.

#### **4.5.2 Cosine, Naïve Bayes, and Collins-Thompson and Callan vs. SVMAUD**

This evaluation compared the performance of SVMAUD against two other machine learning methods, cosine and Naïve Bayes. Cosine experienced decent performance, with F-measures of 0.62 for general and 0.57 for specific audience levels. The Naïve Bayes method experienced higher performance with F-measures of 0.70 for general and 0.61 for specific audience levels. The Collins-Thompson and Callan method experienced still higher performance with F-measures of 0.81 for general and 0.68 for specific audience levels. However, SVMAUD was able to outperform all of these methods with F-measures of 0.87 for general and 0.78 for specific audience levels.

All machine learning methods under evaluation were able to predict the audience level of digital library resources with much higher performance than readability formulas. These machine learning methods required a human expert to identify the audience level for a set of resources that would have been used to train these models; readability formulas have not required this initial training step. Digital library web pages, unlike books and magazines, were structured differently since they contained headers, footers, hyperlinks, and other textual information, in addition to the full text of the resource. All resources held by a collection typically shared a number of common attributes, such as headers and footers, independent of the actual audience level of the resource. SVM

classifiers, in general, required a high proportion of unique terms in each class in order to make fine-grained distinctions between adjacent audience levels. SVMAUD prediction performance exceeded the other machine learning methods and readability formulas under evaluation, more than balancing the required initial effort required of a human to identify representative training samples appropriate for each audience level.

Finally, as the audience level for a resource could have likely spanned many grade levels, such as a resource being appropriate for grades six through eight, the training documents were placed into all applicable classes, increasing the number of terms that were common to more than one audience level. As the resources became more specialized at higher audience levels, the proportion of unique terms also increased; in fact, when predicting the specific audience level for college level resources, the SVMAUD prediction performance resulted in F-measures around 0.85. SVMAUD relied on a large proportion of unique terms for each audience level in order to better make fine-grained distinctions between adjacent audience levels.

#### **4.5.3 SVMAUD Performance Using Readability Formula Inputs**

This study took the inputs to the two readability formulas, Flesch-Kincaid Reading Age and the Dale-Chall Reading Ease Score, and used the inputs to these readability formulas as the training and testing data for each resource in the collection. The performance when using SVMAUD to predict the audience level using these inputs was found to improve as compared to using the constants in the original formulas. SVMAUD did not require that the inputs follow a consistent pattern of increasing or decreasing as the audience level increased.

The readability formulas relied on well-edited resources that followed the convention that the inputs, such as syllables per word and average sentence length, should have increased or decreased at a constant rate as the audience level increased. The results from this study showed that there was not a consistent pattern among resources held in the digital library collections. In fact, the highest average sentence length was found to be 242.3 for a ninth grade document titled "Microbes : Too Smart for Antibiotics?" held by the Middle School Portal: Math and Science Pathways collection; this resource was a lesson plan converted to text using OCR, and missed nearly all the end of sentence tokens. The highest average syllable count per word was 10.6 for a fourth grade resource titled "Cross-Cultural Studies in Cognition and Mathematics" held by the Pacific Resources for Education and Learning collection; words ran together due to a lack of spaces when converted from text using OCR. The resource with the highest proportion of difficult words was found to be 90% for an eleventh grade resource titled "Falling Football" held by the Math Forum collection; this document contained many words not on the Dale Common Word List, even though the synonyms appeared on this list. The readability formulas required well-formatted resources that followed the conventions of traditional written English and a consistent pattern of the inputs increasing or decreasing as the audience level increased.

If the inputs to the readability formulas truly represented the ease or difficulty of understanding a resource, then simply adjusting the constants that were calculated using regression analysis should have resulted in much higher performance. However, the inputs have not followed a consistent pattern, and the results would still have been extremely poor, as evidenced by the predictions of SVM-AUD when trained and tested



using these inputs. Even though the overall F-measure audience level prediction performance increased by approximately 0.08 for both readability formulas with respect to general audience level, and 0.06 for specific audience level, SVMAUD, relying on the full text of the resource to suggest audience level, far outperformed both readability formulas. Rather than relying on simplistic word and sentence characteristics, the vocabulary in a pre-labeled set of resources should have been used as training data to predict the audience level for unlabeled resources.

#### **4.6 Digital Library Audience Level Prediction Evaluation**

This part of the study compared the performance of a number of different machine learning methods and readability formulas when asked to predict the audience level for resources held in digital library collections. The readability formulas experienced extremely poor performance, due to the nature of web pages that, in addition to the full text, also contained headers, footers, scripts, tables, and figures that distorted the average sentence length parameter. Even though the Dale-Chall Reading Ease Score considered the vocabulary appropriate for each audience level, the audience level prediction performance was reduced due to the average sentence length parameter. These readability formulas experienced extremely poor performance with F-measures under 0.30 for general, and under 0.10 for specific audience levels. These readability formulas failed, due to the inconsistent pattern of increasing or decreasing values for inputs to these formulas as the audience level increased. By training SVMAUD to use these same inputs to predict the audience level, the performance marginally increased, indicating that these text characteristics were not indicative of reading difficulty.

To overcome the limitations imposed by readability formulas, SVMAUD was proposed to predict both general and specific audience levels. This method was compared to three baseline classification methods, namely cosine, Naïve Bayes, and the Collins-Thompson and Callan method. When trained and tested using the full text of resources, all three of the baseline machine learning methods experienced F-measures under 0.81 for general audience levels and under 0.68 for specific audience levels. SVMAUD exceeded the performance of these three methods with F-measures of 0.87 for general, and 0.78 for specific audience levels. SVMAUD could not only have been used to predict the audience level of digital library resources, but also could have been used to predict the audience level of any written work.

## **CHAPTER 5**

### **ADJUSTING TERM WEIGHT BASED ON HTML TAGS**

In the current weighting scheme, all terms are weighted based on their frequencies in the current class as well as the number of classes in which the individual term appears. All terms are assigned the same weight independent of their location in the document or the HTML tags in which they occur. However, terms that appear in the title tag or H1 tag should be assigned higher weight than terms that appear in a paragraph tag. This part of the study seeks to optimize the prediction performance of the different machine learning methods by giving additional weight to terms that have a greater degree of importance in describing the document content. Most of the current research focuses on developing a modified term weighting scheme appropriate for search engines or information retrieval systems; however, these methods should be adopted to improve the prediction performance of the cosine, Naïve Bayes, Collins-Thompson & Callan method, and SVMAUD classification methods.

#### **5.1 Previous Studies**

This section reviews a number of studies that seek to improve information retrieval performance for search engines by giving additional weight to terms appearing in certain HTML tags. In one such study, the text between HTML tags in the document is grouped into six different categories, according to Table 5.1 on the next page (Cutler, Shih, & Meng, 1997):

**Table 5.1** Six Categories and Associated HTML Tags

Category Name	HTML Tags
Title	Title
H1-H2	H1, H2
H3-H6	H3, H4, H5, H6
Strong	Strong, B, EM, I, U, DL, OL, UL
Anchor	A (anchor tags from other documents that link to the current document)
Plain Text	All terms not appearing in one of the above classes

The terms appearing in different HTML tags were grouped into various categories to reduce the work required to determine the importance of the terms appearing in each of the HTML tag categories. The title, H1-H2, and H3-H6 tags contain important information describing the topic of the document, and descriptions about the content in different parts of the document. The A category refers to the text present in the anchor tag in another document that contain a hyperlink to the current document; the text in the anchor tag should represent the main ideas of the current document. If a term appears in more than one of the categories, the tag is assigned only to the class that appears earlier in the table. For example, if a term appears in both the title tag and plain text, then the term is counted only for the title tag, and is removed from all other parts of the resource. After all terms in the HTML page are placed into a single category as shown in Table 5.1, the terms and their number of occurrences in each category are counted. The terms that appear in the anchor text of other documents that contain a hyperlink to the current document are also stored, along with their individual frequencies. The Class Importance Vector (CIV) is also calculated. Then, the weight of each of the terms that appear in each HTML tag class is adjusted according to the following formula.

$$CIV = (civ_1, civ_2, civ_3, civ_4, civ_5, civ_6) \quad (5.1)$$

In this formula, each *civ* corresponds to one of the categories in Table 5.1. After all terms and their occurrences in each document are counted, then the term weight with respect to each document is calculated by using the following formula:

$$w = (TFV \bullet CIV) \cdot idf, \text{ where } idf = \ln(N/df) \quad (5.2)$$

In this formula, the inner product of the two vectors  $\bullet TFV$  and  $CIV$ , represents the importance of term  $t$  to the individual document  $d$ , and  $idf$  represents the inverse document frequency. In the  $idf$  calculation,  $N$  represents the number of documents in the collection, and  $df$  represents the number of documents that contain the term. After all of the individual term weights for each document are calculated and stored, the retrieval system uses cosine similarity to calculate the similarity between each document and a query; a smaller cosine between the document and query indicates higher similarity.

One study using this method optimizes the 5-point average precision and 11-point average precision across ten different queries by adjusting the importance of different HTML tag categories in the  $CIV$ . By using this formula, the optimal retrieval performance is found when plain text and H3-H6 category term weights are not adjusted. The term weight for terms appearing in the anchor and strong categories are increased by a factor of eight. In addition, the term weights for words appearing in the H1-H2 category are increased by a factor of six. Finally, the terms appearing in the title category are increased by a factor of four. The best  $CIV$  is (181684), where the 11-point average precision improves by 26% over weighting all terms equally, and the 5-point average precision improves by 44% over weighting all terms equally (Cutler, Shih, & Meng, 1997). This study shows that adjusting the importance of different HTML elements should improve retrieval performance over weighting all terms equally in a digital library

resource. However, this study only considers adjusting weights of the text contained in a few HTML tags.

Another study adjusts the importance of terms by reorganizing HTML tags commonly appearing in web documents into six main categories; then, the optimal weighting for the terms appearing in each of the six categories is obtained by using the same methodology as the previous study, with the categories shown in the next table.

**Table 5.2** Six Categories and Associated HTML Tags

<b>Category Name</b>	<b>HTML Tags</b>
Title	Title
Header	H1, H2, H3, H4, H5, H6
List	DL, OL, UL
Strong	Strong, B, EM, I, U
Anchor	A (anchor tags from other documents that link to the current document)
Plain Text	All terms not appearing in one of the above classes

By using a different grouping of tags, the Class Importance Vector with the highest retrieval performance is found to be (181782). This vector indicates that the weight for the terms appearing in the plain text and list categories should not be adjusted, the weight for terms appearing in the title class should be increased by a factor of two, the weight for terms appearing in the header category should be increased by a factor of seven, and the terms appearing in the strong and anchor classes should be increased by a factor of eight. By adjusting the weight of the terms appearing in different classes, the 11-point average precision retrieval performance is improved by 48.3% over weighting all terms equally, regardless of their appearance in various HTML elements (Cutler, Deng, Maniccam, & Weng, 1999). This study adjusts the importance weight of terms

appearing in more descriptive HTML elements; by boosting the importance of text appearing in certain HTML elements, the retrieval performance is improved over the baseline where all terms are weighted equally. While this study considers a different grouping of HTML elements than the previously described study, a number of HTML elements, particularly the META information, are not considered in this study.

A different approach considers the importance of usage of the HTML META tag to improve retrieval performance, by creating twenty HTML pages in five different subject areas, namely agricultural trade, farm business statistics, poultry statistics, vegetable statistics, and cotton statistics. Four pages are created in each subject area, with one page containing no META tag information, another page containing the META keywords attribute, a third page containing a META description attribute, and a final page containing both META description and META keywords information. After publishing these pages on the web, searches are performed using both AltaVista and Infoseek search engines to find terms appearing in all test pages as well for each keyword appearing in the META tag. By entering keywords in the META attribute, the retrieval performance substantially improves versus neglecting to include this attribute, while the inclusion of the description META attribute does not materially impacted retrieval performance (Turner & Brackbill, 1998). By incorporating the keywords META attribute into the HTML page, the page should be ranked higher in the search results versus neglecting to include this information. Similarly, if the importance of the terms appearing in the META keywords element could be boosted over the remaining text in a digital library resource, the audience level prediction performance should increase over weighting all terms equally.

In order to improve the retrieval performance of a text search system when querying a database containing HTML pages, terms contained in certain HTML tags are given increased importance over all other tags. Rather than simply increasing the importance of terms appearing in certain HTML tags by counting the term multiple times, this model assigns the weight based on a non-linear contextual model. In addition, terms are grouped into twelve different categories, also called *ctags* (Pereira, Molinari, & Pasi, 2005). The terms appearing in each of the HTML elements are placed into one of the twelve tag classes as shown in the following table:

**Table 5.3** HTML Tag Classes

Rank	Class Name ( <i>ctags</i> )	Classified Tags / Parameters
1	Title	Title and META Keywords
2	Header 1	H1, Font Size=7
3	Header 2	H2, Font Size=6
4	Header 3	H3, Font Size=5
5	Linking	A HREF
6	Emphasized	EM, Strong, B, I, U, Strike, S, Blink, Alt
7	Lists	UL, OL, DL, Menu, Dir
8	Emphasized 2	Blockquote, Cite, Big, Pre, Center, TH, TT
9	Header 4	H4, Caption, Center, Font Size=4
10	Header 5	H5, Font Size=3
11	Header 6	H6, Font Size=2
12	Delimiters	P, TD, text not in another tag, Font Size=1

This model does not consider such tags as HR, BR, or Frame that do not hold information that describes the content of the HTML page. After compiling a list of all terms on the HTML page and the tags in which these terms appear, the significance of a single term  $t$  in a single document  $d$  is calculated by following the procedure:

- 1) Since the text appearing in the delimiters group contain a much greater number of terms than those appearing in the title and heading tags, the importance of the



terms appearing in the delimiters class should be reduced over the terms appearing higher in the hierarchy. Since the terms appearing in the title and META keyword tags are typically short and appear once in the HTML page, the base term weight is counted as one if the term appears in this class. For the terms appearing in all other classes in the hierarchy, another function is needed to represent the importance of the term within the respective class, as follows:

$$F_{ctag_i}(d, t) = ti = \frac{T_i}{Z_i} \quad (5.3)$$

In this formula,  $T_i$  represents the occurrences of term  $t$  in  $ctag_i$  in document  $d$  and  $Z_i$  represents the number of occurrences of the most frequent term within  $ctag_i$ . In this way, the importance of the terms appearing in HTML tag classes that contain a high number of terms is reduced over classes that contain fewer terms.

- 2) The numerical importance weight of each  $ctag$  in each document must now be calculated, subject to the following conditions:

$$\sum_{i=1}^n w_i = 1 \text{ and } w_i > 0 \quad (5.4)$$

The  $ctag$  weights are calculated based on the premise that the number of terms appearing in each  $ctag$  higher in the hierarchy should be much lower than the term counts appearing lower in the hierarchy. The degree of importance of the terms appearing in each  $ctag$  is calculated by identifying the count of the terms that appear in each  $ctag$ , or  $S_i$ ; if the  $ctag$  does not contain any terms, then it is not been used in the calculation. The normalized  $ctag$  length, or  $s_i$ , is calculated by the following formula, where  $n$  represents the number of  $ctags$  out of the twelve in the hierarchy containing at least one term:

$$s_i = S_i / \sum_{j=1}^n S_j \quad (5.5)$$

The importance degree  $v_i > 0$  of each *ctag* is now calculated based on the normalized *ctag* lengths, according to the following formula:

$$v_i = e^{-\beta \tilde{s}_i} \text{ where } \tilde{s}_i = s_1 + \dots + s_i, \text{ where } s_i = \sum_{j=1}^i s_j \quad (5.6)$$

If a document contains terms appearing in four different *ctags*, the importance degree of the term appearing in each *ctag*  $v_i$  is calculated by the following formula:

$$v_1 = e^{-\beta \tilde{s}_1} \quad v_2 = e^{-\beta \tilde{s}_2} \quad v_3 = e^{-\beta \tilde{s}_3} \quad v_4 = e^{-\beta \tilde{s}_4} = e^{-\beta} \quad (5.7)$$

$$\tilde{s}_1 = s_1 \quad \tilde{s}_2 = s_1 + s_2 \quad \tilde{s}_3 = s_1 + s_2 + s_3 \quad \tilde{s}_4 = s_1 + s_2 + s_3 + s_4$$

After all of the individual weights for each *ctag* are calculated and stored, they are then normalized so that all *ctag* weights  $w_i$  total to one, according to the following formula:

$$w_i = \frac{e^{-\beta \tilde{s}_i}}{W}, \text{ where } W = \sum_{i=1}^n e^{-\beta \tilde{s}_i} \quad (5.8)$$

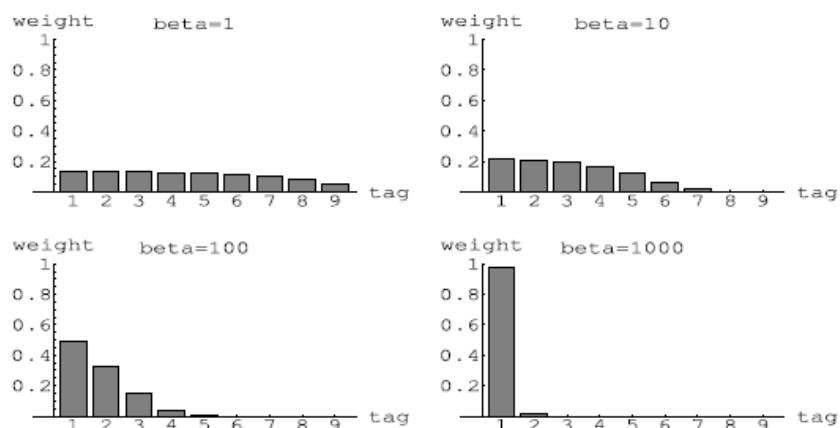
The free parameter  $\beta$  controls the entropy in the formula, or the variation in weight between adjacent *ctags*. As the value of  $\beta$  approaches zero, the entropy is maximized and the weight distribution of *ctags* is constant; in other words, all terms are weighted equally independent of the HTML tag in which it appears. As the value of  $\beta$  increases, the difference in weight between adjacent *ctags* increases, meaning that the importance of terms appearing in the title and META keywords tag increases while the weight for all other *ctags* decreases. As the value of  $\beta$  approaches infinity, the highest *ctag* in the hierarchy that contains text in the document is given all of the weight while the remaining *ctags* are given weights of near zero.

- 3) To calculate the term significance of term  $t$  in the document  $d$ , the weighted average of the normalized term frequencies contained in each of the individual  $ctags$  is calculated by using the following formula:

$$F(d, t) = (\sum_{i=1}^n w_i t_i) \log \left( \frac{N}{NDOC_t} \right) \quad (5.9)$$

In this formula,  $N$  is the number of documents contained in the collection and  $NDOC$  represents the number of documents that contain the term.

This formula modifies the TF-IDF calculation by taking into account the terms contained in different  $ctags$  in the document, by giving more weight to terms that describe the document content. Since the optimal value of  $\beta$  is not known for every collection in advance, the value of  $\beta$  is adjusted to determine the optimal weight distribution among  $ctags$  that results in the highest prediction performance. As an example, consider a document containing nine different  $ctags$   $\vec{S}=(2, 4, 8, 16, 32, 64, 128, 256, 512)$ , in which each  $ctag$  length is twice as long as the preceding one. In this example, the number of terms present in HTML tags with less importance are significantly higher than the preceding one.



**Figure 5.1** Weight distribution among  $ctags$  for different values of  $\beta$ .

When the value of  $\beta$  is set to one, the weight distribution among different *ctags* tends to be fairly constant; however, if the value of  $\beta$  is set to 100, the majority of the *ctag* weight shifts to the title and heading tags. In fact, if  $\beta$  is set to 1000, almost all weight is assigned to the terms appearing in the title tag. While this is a proof-of-concept study, the precision and recall of the non-linear model is found to be much higher than a traditional indexing model, where all terms are given the same weight independent of the HTML tag in which they appear (Pereira, Molinari, & Pasi, 2005).

All of these models rely on the same basic idea, whereby terms that occur in HTML tags that are more important to describing the content of the resource should be given higher weight over terms that appear in the other parts of the resource. Even though the audience level prediction problem seeks to label a resource with an unknown audience level with the most appropriate audience level, a web-based text search system and the audience level prediction problem utilize similar methods to determine the importance of terms appearing in different areas of the resource. For example, terms in the title element in an HTML tag generally describe the main content of the resource and should be given higher weight in both problems than terms appearing in the paragraph tag. Terms that appear in other tags, such as headers and captions, provide information describing the content found in different parts of the resource; these terms should be given more importance than tags in table data or paragraphs. Some models rely on doubling or quadrupling the importance of terms found in certain HTML tags, while other models use mathematical formulas to adjust the weight for terms occurring in different HTML tags. Rather than developing a one-size-fits-all model, the method that

holds the most promise is the non-linear weighting model that includes an additional parameter  $\beta$  that adjusts the weight of terms appearing in different HTML tags.

## 5.2 Document Processing

In the digital library collection, the majority of the collection consisted of HTML pages containing a variety of terms placed inside HTML tags. In the HTML source code, the terms were placed inside explicit tags, such as the terms representing the title tag were placed into the <title> element while the terms representing the top level heading tags were placed inside the <H1> element. The digital library collection under evaluation not only contained HTML pages but also required the processing PDF files that generally followed the same format. By developing a set of rules, the importance of terms in PDF files should have followed the same weighting scheme as found in HTML resources. In the few PDF resources that did not follow these rules, the documents were processed manually. Table 5.4 on the next page shows the grouping of terms that have been found in the different HTML tags and PDF resources into the classes representing different term importance; in this table, # represents any positive whole number, namely 0, 1, 2, 3, 4, 5, 6, 7, 8, or 9.

After the individual terms were extracted from the document and placed into the appropriate HTML class, then the weight for each *ctag* would be calculated and stored for each resource in the collection. The weight for terms appearing in different *ctags* was adjusted according to the non-linear model as proposed by Pereira, Molinari, & Pasi (2005).

**Table 5.4** HTML Tag Classes with PDF Equivalent

<b>Rank</b>	<b>Class Name</b>	<b>Terms in HTML Tag</b>	<b>PDF Equivalent</b>
1	Title	Title and META Keywords	Text at top of page to first blank line containing no text; “Keywords” at start of paragraph until line break
2	Header 1	H1, Font Size=7	All text following # after a line break until the next line break
3	Header 2	H2, Font Size=6	All text following ## after a line break until the next line break
4	Header 3	H3, Font Size=5	All text following ### after a line break until the next line break
5	Linking	A HREF	http:// until the first space
6	Emphasized	EM, Strong, B, I, U, Strike, S, Blink, Alt	All formatting information lost when converting PDF to text format; all words entirely in capital letters
7	Lists	UL, OL, DL, Menu, Dir	# followed by # on the next line
8	Emphasized 2	Blockquote, Cite, Big, Pre, Center, TH, TT	Any text in quotes
9	Header 4	H4, Caption, Center, Font Size=4	Text in a line starting with “Table” or “Figure” until the end of the line; all text following #### after a line break until the next line break
10	Header 5	H5, Font Size=3	All text following ##### after a line break until the next line break
11	Header 6	H6, Font Size=2	All text following ##### after a line break until the next line break
12	Delimiters	P, TD, text not in another tag, Font Size=1	All text not appearing in any of the tag classes stated above

In a text retrieval system, the TF-IDF values for each term appearing in each resource were stored. In the audience level prediction problem, all terms contained in the training document collection for each audience level were extracted and stored along with their TF-IDF values; rather than using the number of documents in which the term appeared to calculate the IDF value, the number of audience levels in which the term appeared versus the total number of possible audience levels was used. When the audience level for an unlabeled resource was predicted, the same term weighting scheme was used for the unlabeled resource by only using the terms appear in different HTML tag classes in that individual resource. The cosine and SVMAUD methods under evaluation relied on the TF-IDF calculation to represent the importance of each term, while the Naïve Bayes and Collins-Thompson & Callan methods considered only the TF part of the calculation. All documents in the training and testing parts of the machine learning methods were processed in the same way.

### 5.3 Evaluation

This study determined whether adjusting the weight for terms appearing in different HTML tags improved the audience level prediction performance for the machine learning methods under evaluation. When the weight for each term was adjusted based solely on their frequency in the class without accounting for the HTML tags in which the term appeared, the specific audience level prediction performance was found to be 0.57, 0.61, 0.68, and 0.78 for cosine, Naïve Bayes, the Collins-Thompson & Callan method, and SVMAUD, respectively. On the other hand, when the general audience level prediction performance was considered, the performance was found to be 0.62, 0.70, 0.81, and 0.87

for cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMMAUD, respectively. This study should improve upon this performance by assigning additional weight to the terms in the HTML tags that describe the document content.

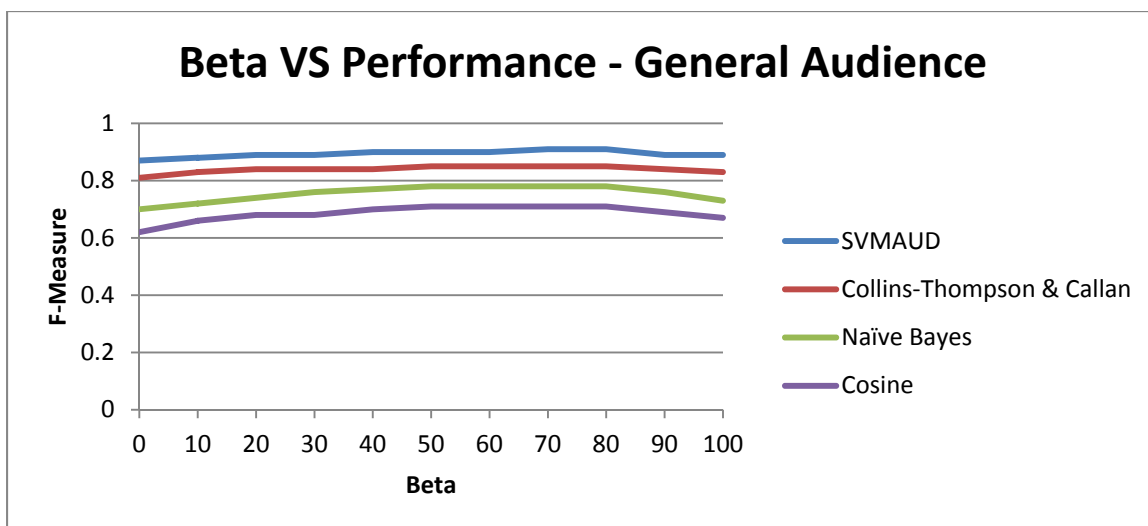
Since the value of  $\beta$  could have been changed to adjust the importance of terms appearing in different *ctags*, ten different studies were conducted as part of this evaluation in order to approximate the optimal value of  $\beta$  that resulted in the highest prediction performance. At the limit of  $\beta$  approaching zero, all terms were assigned the same weight independent of the class in which they have appeared, resulting in the performance that was described in the previous paragraph. Ten additional studies were conducted, by adjusting the value of  $\beta$  to 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100, and then measuring the general and specific audience level prediction performance at each of these ten values of  $\beta$ . The performance was expected to increase as the value of  $\beta$  increased, but, at some point, the performance would have decreased when  $\beta$  became sufficiently large. Since the value of  $\beta$  that resulted in the highest performance would have been different for the general and specific audience level studies and could have varied among the different among different machine learning methods, the performance of each machine learning method with respect to either general or specific audience level prediction was measured. Each evaluation charted the overall F-measure performance of each method against the value of  $\beta$  to identify the optimal value of  $\beta$  that resulted in the highest overall F-measure across all audience levels under evaluation. At the optimal value of  $\beta$ , the precision, recall, and F-measure values are displayed in the table, along with the correlation between human-expert entered and machine-learning suggested audience levels. The evaluation is divided into two different sections; the first measures



the general audience level prediction performance while the second measures the specific audience level prediction performance.

### 5.3.1 General Audience Levels

This study measured the general audience level prediction performance across the four machine learning methods for varying values of  $\beta$ . When  $\beta=0$ , all terms were weighted equally independent of the HTML tag in which it appeared; at the value of  $\beta=100$ , nearly all weight was assigned to the title and heading tags. The following chart plots the value of  $\beta$  versus the overall F-measure performance for general audience levels.



**Figure 5.2**  $\beta$  versus F-measure for general audience levels.

For every value of  $\beta$ , SVMAUD outperformed cosine, Naïve Bayes, the Collins-Thompson and Callan method. The largest increase in prediction performance was between the value of  $\beta=0$  to  $\beta=10$ , with roughly a 0.04 increase in overall F-measure for all machine learning methods. As the value of  $\beta$  increased, the prediction performance

increased at a lower rate until slightly increasing between values of 50 and 80. After the value of  $\beta=90$ , the performance decreased; at this point, most of the term weight was assigned to the text appearing in the title and heading tags. Even though the title was unique for every resource in the collection, the heading tags, especially in the PDF files, was the same for every resource, including such sections as title, related work, system design, evaluation, discussion, and conclusion. The following table shows the highest F-measure prediction performance for each machine learning method, regardless of the value of  $\beta$ .

**Table 5.5** Cosine vs. Naïve Bayes vs. SVMAUD – General Audience Levels

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	2,023	0.70	0.74	0.72	0.73	0.79	0.76	0.90	0.92	0.91
Late Elementary	1,478	0.76	0.45	0.56	0.77	0.67	0.71	0.95	0.83	0.88
Middle School	1,958	0.58	0.66	0.62	0.61	0.78	0.68	0.86	0.89	0.88
High School	2,529	0.65	0.75	0.69	0.89	0.72	0.80	0.87	0.92	0.89
College (Sampled)	2,250	0.95	0.87	0.91	0.93	0.90	0.92	0.98	0.96	0.97
<b>Overall</b>	<b>10,238</b>	<b>0.71</b>	<b>0.71</b>	<b>0.71</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>	<b>0.91</b>	<b>0.91</b>	<b>0.91</b>

Cosine predicted the human-expert entered specific audience level with an overall F-measure of 0.71 at the value of 70 for  $\beta$ , as compared to an overall F-measure of 0.62 when all terms were weighted equally. Naïve Bayes experienced the highest F-measure performance at the value of  $\beta=80$ , by improving from 0.70 when all terms were weighted equally to 0.78 when the term weight was adjusted. SVMAUD similarly experienced the

highest overall F-measure prediction performance when  $\beta=80$  and improved from 0.87, when all terms were weighted equally, to 0.91 when the term weight was adjusted based on the HTML tag in which it appeared. The correlation between human-expert entered and machine learning predicted audience levels was found to be 0.80 for cosine, 0.82 for Naïve Bayes, and 0.94 for SVMAUD. In fact, SVMAUD outperformed both Naïve Bayes and cosine machine learning methods at the 0.0135 level of significance. The next table presents the F-measure prediction performance across all general audience levels for the Collins-Thompson and Callan method versus SVMAUD.

**Table 5.6** Collins-Thompson and Callan vs. SVMAUD – General Audience Levels

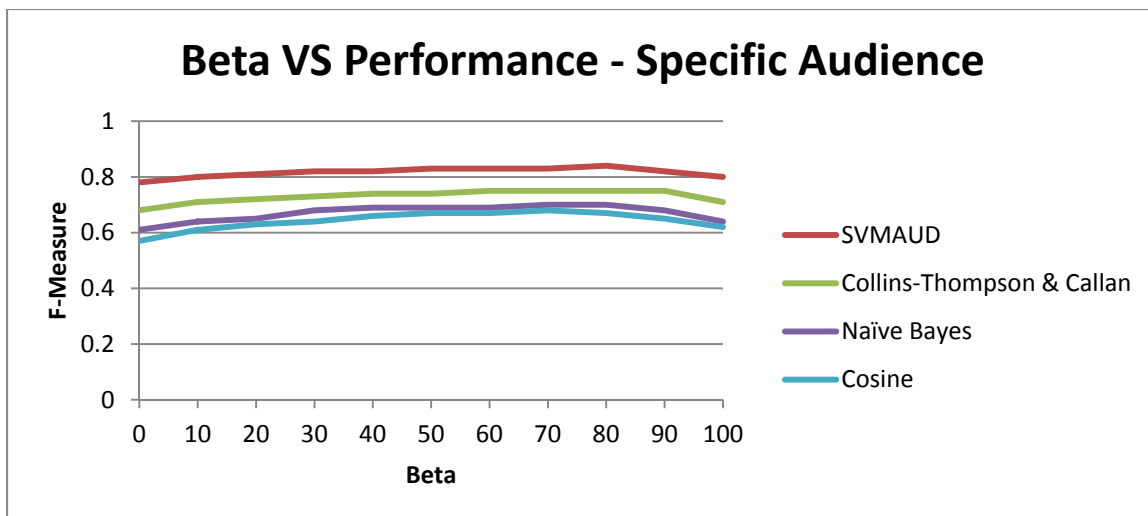
General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	2,023	0.83	0.87	0.85	0.90	0.92	0.91
Late Elementary	1,478	0.91	0.71	0.80	0.95	0.83	0.88
Middle School	1,958	0.77	0.82	0.80	0.86	0.89	0.88
High School	2,529	0.81	0.86	0.84	0.87	0.92	0.89
College (Sampled)	2,250	0.97	0.94	0.95	0.98	0.96	0.97
<b>Overall</b>	<b>10,238</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.91</b>	<b>0.91</b>	<b>0.91</b>

The Collins-Thompson and Callan method experienced the highest general audience level prediction performance at the value of  $\beta=80$ , by improving from an overall F-measure of 0.81 when all terms were weighted equally to an overall F-measure of 0.85 by adjusting the weights based on the HTML tags in which the term occurred. SVMAUD experienced the highest overall F-measure prediction performance when  $\beta=80$

and by improving from 0.87, when all terms were weighted equally, to 0.91 when the term weight was adjusted based on HTML tags. SVMAUD also outperformed the Collins-Thompson and Callan method based on the correlation between human-expert entered and machine-learning suggested values; this correlation was found to be 0.90 for Collins-Thompson and Callan versus 0.94 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0604 level of significance. By adjusting the importance of terms based on the HTML tags in which they occurred, the prediction performance improved over weighting all terms equally.

### **5.3.2 Specific Audience Levels**

The cosine, Naïve Bayes, Collins-Thompson and Callan method, and SVMAUD were compared based on their abilities to correctly predict the specific human-expert entered audience level. As the value of  $\beta$  increased from 0 to 100, more weight was assigned to terms appearing in the title and heading tags, and less weight was assigned to terms appearing in the plain text. Figure 5.3 on the next page plots the value of  $\beta$  versus overall F-measure prediction performance, comparing the prediction performance of the four machine learning methods under evaluation.



**Figure 5.3**  $\beta$  versus F-measure for specific audience levels.

The four machine learning methods were compared on their abilities to correctly predict the human-expert entered specific audience level. SVMAUD again outperformed all other machine learning methods under evaluation in this study. The largest increase in audience level prediction performance was between the values of  $\beta=0$  to  $\beta=10$ , with an increase in overall F-measure of approximately 0.03 across all methods. The highest prediction performance for cosine occurred when  $\beta=70$  for cosine, and when  $\beta=80$  for Naïve Bayes, Collins-Thompson and Callan, and SVMAUD. At this point, most of the weight was assigned to the terms that appeared in the title and heading tags, while a small amount of weight was given to the terms that appear in the plain text portion of the resource. Table 5.7 on the following page displays the specific audience level prediction performance for cosine, Naïve Bayes, and SVMAUD.

**Table 5.7** Cosine vs. Naïve Bayes vs. SVMAUD – Specific Audience Levels

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	698	0.49	0.70	0.58	0.52	0.70	0.60	0.69	0.83	0.75
First	719	0.88	0.74	0.80	0.86	0.72	0.79	0.93	0.86	0.90
Second	606	0.50	0.88	0.64	0.50	0.90	0.64	0.65	0.94	0.77
Third	418	0.98	0.51	0.67	1.00	0.54	0.70	0.99	0.76	0.86
Fourth	528	0.70	0.63	0.66	0.73	0.65	0.69	0.87	0.79	0.83
Fifth	532	0.99	0.51	0.67	0.99	0.54	0.70	0.99	0.74	0.85
Sixth	664	0.98	0.52	0.68	0.99	0.53	0.69	1.00	0.77	0.87
Seventh	663	0.81	0.54	0.65	0.82	0.56	0.67	0.92	0.76	0.84
Eighth	631	0.32	0.73	0.45	0.36	0.78	0.49	0.54	0.88	0.67
Ninth	693	0.69	0.76	0.72	0.73	0.79	0.76	0.84	0.87	0.85
Tenth	640	0.93	0.67	0.78	0.94	0.72	0.81	0.97	0.86	0.91
Eleventh	552	0.84	0.67	0.75	0.83	0.70	0.76	0.94	0.82	0.87
Twelfth	644	0.66	0.69	0.67	0.72	0.73	0.73	0.84	0.85	0.85
UG Lower (Sampled)	750	0.63	0.87	0.73	0.64	0.89	0.74	0.80	0.94	0.86
UG Upper (Sampled)	750	1.00	0.61	0.76	0.99	0.62	0.76	1.00	0.81	0.89
Graduate (Sampled)	750	0.91	0.68	0.78	0.91	0.70	0.79	0.95	0.85	0.90
<b>Overall</b>	<b>10,238</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>

\*\* UG = Undergraduate

Cosine improved from an overall F-measure prediction performance of 0.57 when all terms were weighted equally to an overall F-measure prediction performance of 0.68 when term weight was adjusted based on the HTML tags. Naïve Bayes also experienced increased overall F-measure performance, from 0.61 when all terms were weighted equally, to 0.70 when the term weight was adjusted based on HTML tags. SVMAUD experienced the highest overall F-measure prediction performance of 0.84 in this study versus 0.78 when all terms were weighted equally. SVMAUD experienced the highest correlation between human-expert entered and machine-learning suggested specific

audience levels of 0.91 versus 0.84 for Naïve Bayes and 0.82 for cosine. In fact, SVMAUD was found to outperform both cosine and Naïve Bayes at the 0.0001 level of significance. The next table compares the specific audience level prediction performance of SVMAUD and the Collins-Thompson and Callan method.

**Table 5.8** Collins-Thompson and Callan vs. SVMAUD – Specific Audience Levels

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	698	0.59	0.78	0.67	0.69	0.83	0.75
First	719	0.90	0.78	0.84	0.93	0.86	0.90
Second	606	0.57	0.91	0.70	0.65	0.94	0.77
Third	418	0.99	0.68	0.81	0.99	0.76	0.86
Fourth	528	0.76	0.69	0.72	0.87	0.79	0.83
Fifth	532	1.00	0.62	0.76	0.99	0.74	0.85
Sixth	664	0.99	0.64	0.78	1.00	0.77	0.87
Seventh	663	0.87	0.64	0.74	0.92	0.76	0.84
Eighth	631	0.41	0.81	0.55	0.54	0.88	0.67
Ninth	693	0.76	0.81	0.79	0.84	0.87	0.85
Tenth	640	0.95	0.76	0.84	0.97	0.86	0.91
Eleventh	552	0.89	0.73	0.80	0.94	0.82	0.87
Twelfth	644	0.73	0.77	0.75	0.84	0.85	0.85
UG Lower (Sampled)	750	0.70	0.90	0.79	0.80	0.94	0.86
UG Upper (Sampled)	750	0.99	0.69	0.82	1.00	0.81	0.89
Graduate (Sampled)	750	0.94	0.77	0.84	0.95	0.85	0.90
<b>Overall</b>	<b>10,238</b>	<b>0.75</b>	<b>0.75</b>	<b>0.75</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>

\*\* UG = Undergraduate

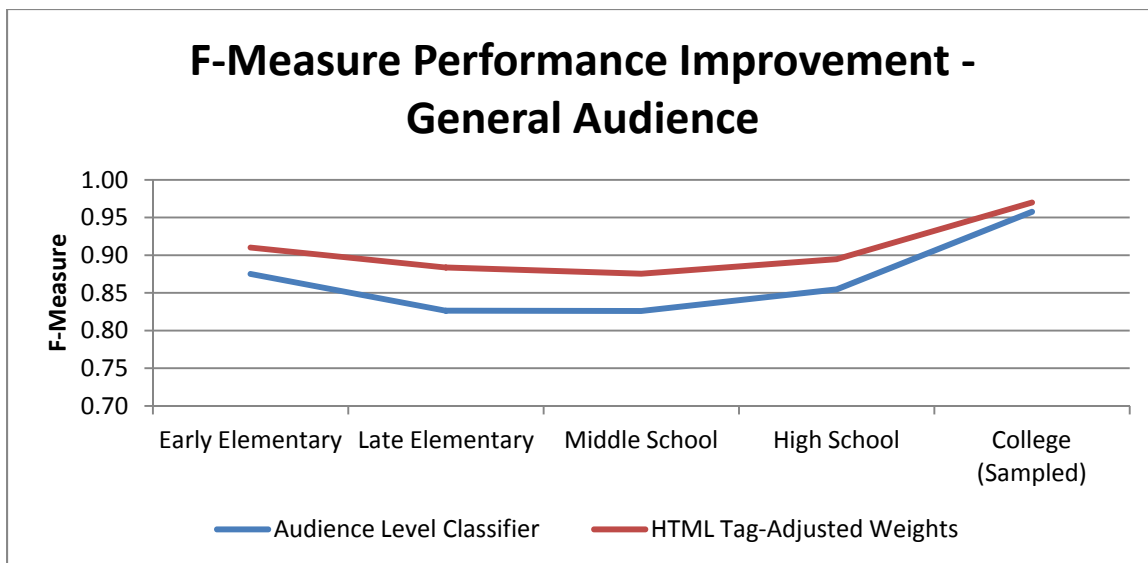
SVMAUD again outperformed the specific audience level prediction performance with an overall F-measure of 0.84 versus 0.75 for the Collins-Thompson and Callan method. The Collins-Thompson and Callan method improved from an overall F-measure

specific audience level prediction performance of 0.68 when all terms were weighted equally to 0.75 when the term weight was adjusted based on HTML tags. SVMAUD again experienced the highest overall F-measure prediction performance of 0.84 in this study, versus 0.78 when all terms were weighted equally. The correlation between human-expert entered and machine-learning suggested values was 0.87 for the Collins-Thompson and Callan method versus 0.91 for SVMAUD. In this study, SVMAUD outperformed the specific audience level prediction performance of the Collins-Thompson and Callan method at the 0.0016 level of significance. Therefore, SVMAUD outperformed the specific audience level prediction performance of all other machine learning methods under evaluation in this study and should have been used to predict the audience level for resources with missing or incompatible audience level metadata.

#### **5.4 Performance Improvement**

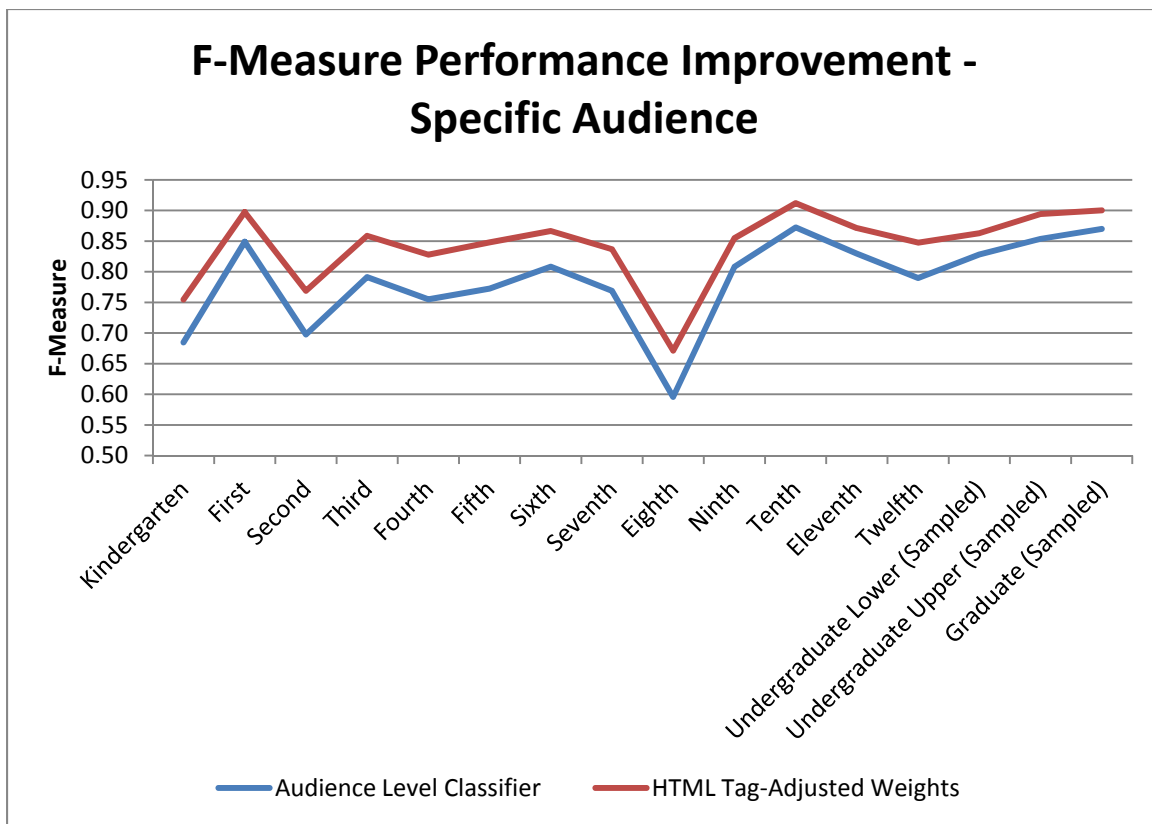
By adjusting the term weight based on the HTML tags in which it appeared, the prediction performance of SVMAUD improved over weighting all terms equally. By providing additional weight to the terms that appeared in the title and header tags and reducing the importance of terms appearing in the plain text, SVMAUD performance improved across all audience levels. The following figure compares the prediction performance for specific audience levels when all terms were weighted equally with the performance when the term weight was adjusted.





**Figure 5.4** SVMAUD general audience level performance improvement.

The general audience level F-measure experienced improved performance over using the full text for training and testing and weighting all terms equally. The largest increase was in the late elementary audience level, while the smallest increase occurred in the college audience level. Since the PDF files drawn from Springerlink contained different vocabulary from the digital library resources and were not structured similarly to HTML pages, little improvement could have been gained for this audience level. The next figure presents the performance improvement for specific audience levels when comparing the original term weights with term weights that were determined by surrounding HTML tags.



**Figure 5.5** SVMAUD specific audience level performance improvement.

The largest performance increase occurred at the eighth grade audience level, while the smallest performance was at the graduate audience level. By using a classifier that accounted for the tags in which the terms appeared, the prediction performance increased over all audience levels. If the HTML tag information could have been available for all resources in the collection, rather than using a set of rules to process PDF files, this performance should have increased even further.

## 5.5 Summary and Conclusion

Rather than simply extracting the full text from HTML pages and conducting training and testing on this dataset, this part of the study adjusted the weight assigned to the terms appearing in different HTML tags. Since the terms appearing in the title and heading tags described the document content, the terms appearing in these tags should have been given more weight than those appearing in table data or the plain text. This study used a nonlinear approach to adjust the importance of terms appearing in different HTML tag classes. By adjusting the term weight based on the HTML tags in which they appeared, information retrieval performance improved; similarly, this weighting scheme also improved the audience level prediction performance of the different classifiers under evaluation.

The terms appearing in title, heading, body, caption, and anchor tags could have been easily extracted from HTML pages by simply viewing the source code. However, the document collection in this study also contained PDF files that, when text was extracted, all formatting information was lost. Therefore, a set of rules was developed to place the terms into tag classes matching the tags present in the HTML source code. This set of rules was applied to simplify the extraction of text; however, some PDF files, particularly those containing a lot of images, could not have been processed by following this set of rules. These PDF files needed to be processed manually in order to apply the same weighting scheme as applied to the HTML pages. If the digital library collection consisted solely of HTML-based web pages, then this method could have more accurately extracted all text belonging in each of the HTML tag classes.

This part of the study considered adjusting the value of  $\beta$  in increments of ten to find the optimal value when the prediction performance of each method was maximized. While the value of  $\beta=70$  for cosine and  $\beta=80$  for Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD resulted in the highest performance, the values such as 65, 75, and 85 were not considered in order to reduce the complexity of the study. Even though these midrange values were not considered, the prediction performance at higher values of  $\beta$  changed slightly between adjacent values of  $\beta$ .

By adjusting the weight of terms appearing in different HTML tags in a web page, the overall F-measure prediction performance for both general and specific audience levels over all machine learning methods improved over weighting all terms equally, independent of the HTML tag in which the term occurred. The general audience level prediction performance increased from 0.62 to 0.71 for cosine, 0.70 to 0.78 for Naïve Bayes, 0.81 to 0.85 for the Collins-Thompson and Callan method, and 0.87 to 0.91 for SVMAUD. On the other hand, the specific audience level prediction performance increased from 0.57 to 0.68 for cosine, 0.61 to 0.70 for Naïve Bayes, 0.68 to 0.75 for the Collins-Thompson and Callan method, and 0.78 to 0.84 for SVMAUD. SVMAUD outperformed all other machine learning methods under evaluation in this HTML tag processing study and should have been used to predict the audience level for all digital library resources with missing or incompatible audience level metadata.

## CHAPTER 6

### REDUCING NOISE IN THE TRAINING DATASET

Even though classification methods can predict the audience level of digital library resources found in the test collection with high performance, the prediction performance can be further improved by reducing the amount of noise present in the training dataset. The training documents, in addition to the full text, contain headers, footers, scripts, and hyperlinks that are present in every document in the digital library collection, independent of the audience level. For example, DLESE places common menu items across the top of the page, links on the left side, and information in the footer that appear in all resources from this collection. In most cases, the abstract summarizes the main ideas presented in the web page. However, additional metadata items, such as the title of the page and keywords, can hold important clues to suggest the audience level of the resource. A title is provided for every resource in the collection; when the additional metadata items of keywords and abstract are provided by the collection manager, they are also used to train the classifiers. Since the number of common words across all audience levels in the digital library collection are reduced, the classification performance should improve for cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMMAUD.

#### **6.1 General Audience Level Noise-Reduced Classification Performance**

In this part of the study, the title, keywords, and abstract metadata were outputted to a text file representing the resource content; if one or more of these elements were missing,

then the remaining elements were placed in the text file and used as the training dataset. The full text of the resources, including HTML tags and script information, were used for testing, since not all resources included abstract or keyword information. The following table summarizes the results from this part of the study, comparing the performance of cosine, Naïve Bayes, and SVMAUD:

**Table 6.1** Noise-Reduced Classification Performance – General Audience Levels

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	2,023	0.60	0.69	0.64	0.70	0.76	0.73	0.89	0.92	0.91
Late Elementary	1,478	0.60	0.48	0.54	0.74	0.62	0.67	0.93	0.88	0.90
Middle School	1,958	0.46	0.65	0.54	0.56	0.75	0.64	0.83	0.92	0.87
High School	2,529	0.80	0.57	0.66	0.87	0.68	0.76	0.97	0.89	0.93
College (Sampled)	2,250	0.89	0.86	0.87	0.91	0.90	0.91	0.97	0.97	0.97
<b>Overall</b>	<b>10,238</b>	<b>0.66</b>	<b>0.66</b>	<b>0.66</b>	<b>0.75</b>	<b>0.75</b>	<b>0.75</b>	<b>0.92</b>	<b>0.92</b>	<b>0.92</b>

In this study, the audience level prediction performance for all three classification methods improved over using full text for training. The F-measures for cosine and Naïve Bayes classification methods increased by 0.04 and 0.05, respectively, over using the full text for training and testing. The prediction performance of SVMAUD increased from 0.87 to 0.92, an increase of 0.05 with respect to overall F-measure. The correlation between human-expert entered and machine-learning suggested values increased for both Naïve Bayes, from 0.77 using full text to 0.80 using the cleaned training dataset, and SVMAUD, from 0.91 using full text for training and testing to 0.94 when the cleaned

dataset had been used for training; on the other hand, the correlation for cosine slightly decreased, from 0.74 to 0.73. SVMAUD exceeded the performance of cosine at the 0.0051 level of significance and Naïve Bayes at the 0.0077 level of significance.

The following table compares the performance of cosine, Naïve Bayes, and SVMAUD when predicting general audience levels:

**Table 6.2** Collins-Thompson and Callan vs. SVMAUD – General Audience Levels

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	2,023	0.80	0.86	0.83	0.89	0.92	0.91
Late Elementary	1,478	0.84	0.75	0.79	0.93	0.88	0.90
Middle School	1,958	0.70	0.84	0.77	0.83	0.92	0.87
High School	2,529	0.93	0.80	0.86	0.97	0.89	0.93
College (Sampled)	2,250	0.95	0.94	0.94	0.97	0.97	0.97
<b>Overall</b>	<b>10,238</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>	<b>0.92</b>	<b>0.92</b>	<b>0.92</b>

SVMAUD once again outperformed the Collins-Thompson and Callan method in its ability to correctly predict the human-expert entered audience level for all resources in the collection, with an overall F-measure of 0.92 versus 0.84 for the Collins-Thompson and Callan method. The correlation between human-expert entered and machine-learning suggested values also improved slightly for both methods, from 0.87 to 0.88 for Collins-Thompson and Callan and 0.91 to 0.94 for SVMAUD. Thus, once again, SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0313 level of significance.

Since the abstract, title, and keywords elements contained text that was appropriate for the individual resource, SVMAUD could have made fine grained distinctions between adjacent general audience levels. In the previous study using the full text of digital library resources, the documents contained headers, footers, scripts, menus, and other portions of the text that were common to all resources in the collection independent of audience level.

## **6.2 Specific Audience Level Noise-Reduced Classification Performance**

Similar to the first study using full text for training and testing, this part of the study compared the prediction performance of cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD in their ability to correctly predict the human-expert entered specific audience level for each resource, using a cleaned training dataset. Table 6.3 on the next page displays the cosine, Naïve Bayes, and SVMAUD classification performance when labeling resources with specific audience levels; once again, the training data consisted of title, keywords, and abstract for each resource, while the testing dataset consisted of the full text.

In this study, the prediction performance increased for cosine, Naïve Bayes, and SVMAUD methods over using full text for training and testing. The cosine and Naïve Bayes F-measures increased by 0.04 and 0.07, respectively, over using full text for training and testing. However, the performance of SVMAUD increased by 0.08, from 0.78 when using the full text for training and testing to 0.86 when the title, keywords, and abstract were used for training.



**Table 6.3** Noise Reduced Classification Performance – Specific Audience Levels

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	698	0.82	0.60	0.69	0.87	0.66	0.75	0.93	0.84	0.88
First	719	0.59	0.65	0.62	0.65	0.71	0.68	0.83	0.85	0.84
Second	606	0.75	0.58	0.65	0.80	0.65	0.72	0.91	0.81	0.86
Third	418	0.60	0.61	0.60	0.66	0.67	0.66	0.86	0.84	0.85
Fourth	528	0.62	0.65	0.64	0.68	0.73	0.70	0.85	0.85	0.85
Fifth	532	0.36	0.64	0.46	0.42	0.70	0.53	0.66	0.86	0.75
Sixth	664	0.77	0.62	0.68	0.82	0.68	0.74	0.92	0.84	0.88
Seventh	663	0.84	0.61	0.71	0.87	0.68	0.77	0.96	0.84	0.89
Eighth	631	0.49	0.69	0.57	0.56	0.74	0.63	0.78	0.89	0.83
Ninth	693	0.71	0.63	0.67	0.76	0.69	0.73	0.92	0.86	0.89
Tenth	640	0.40	0.55	0.47	0.48	0.62	0.54	0.76	0.83	0.80
Eleventh	552	0.85	0.59	0.70	0.90	0.66	0.76	0.95	0.82	0.88
Twelfth	644	0.61	0.59	0.60	0.67	0.65	0.66	0.87	0.84	0.85
UG Lower (Sampled)	750	0.59	0.59	0.59	0.67	0.65	0.66	0.84	0.84	0.84
UG Upper (Sampled)	750	0.86	0.58	0.69	0.90	0.65	0.75	0.96	0.85	0.90
Graduate (Sampled)	750	0.54	0.66	0.59	0.61	0.72	0.66	0.81	1.00	0.89
<b>Overall</b>	<b>10,238</b>	<b>0.61</b>	<b>0.61</b>	<b>0.61</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>

\*\* UG = Undergraduate

Even though a higher proportion of resources were labeled with the human-expert entered audience level, the correlation between human-expert entered and machine-learning suggested values decreased, from 0.77 to 0.64 for cosine, 0.79 to 0.70 for Naïve Bayes, and 0.88 to 0.87 for SVMAUD when comparing the full text training dataset with the cleaned training dataset. Since the titles, abstracts, and keywords contained fewer terms than found in the full text and represented the essence of the terms found in the full text of the resource, the ability of the classifier to suggest an audience level near to the human-expert entered audience level was reduced. However, SVMAUD outperformed

both cosine and Naïve Bayes at the 0.0001 level of significance, indicating that SVMAUD should have been used to predict the audience level of resources with missing or inconsistent audience level metadata.

The next part of this study compared the prediction performance of the Collins-Thompson and Callan method with SVMAUD. Table 6.4 on the next page shows the precision, recall, and F-measure predictions by these two methods when titles, abstracts, and keywords have been used for training and the full text of the resource used for testing. In this study, the Collins-Thompson and Callan method experienced improved prediction performance over using full text for training, with an increase from 0.68 when using full text to 0.75 when titles, abstracts, and keywords were used for training. The correlation between human-expert entered and machine-learning suggested values decreased slightly, from 0.83 when full text was used for training to 0.77 when using titles, abstracts, and keywords for training. Since the titles, abstracts, and keywords summarized the text found in the resource but have not included all of the terms in the full text of the resource, the classifier correctly predicted the human-expert entered audience level with higher performance, but the incorrect predictions were farther away from the human-expert entered values than using full text for training. The prediction performance of SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0001 level of significance. SVMAUD should have been used to predict the audience level for all unlabeled resources in the digital library collection.

**Table 6.4** Collins-Thompson and Callan vs. SVMAUD - Specific Audience Levels

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	698	0.91	0.75	0.82	0.93	0.84	0.88
First	719	0.73	0.79	0.76	0.83	0.85	0.84
Second	606	0.85	0.74	0.79	0.91	0.81	0.86
Third	418	0.73	0.74	0.74	0.86	0.84	0.85
Fourth	528	0.76	0.76	0.76	0.85	0.85	0.85
Fifth	532	0.51	0.78	0.62	0.66	0.86	0.75
Sixth	664	0.86	0.74	0.80	0.92	0.84	0.88
Seventh	663	0.92	0.74	0.82	0.96	0.84	0.89
Eighth	631	0.64	0.79	0.71	0.78	0.89	0.83
Ninth	693	0.83	0.78	0.80	0.92	0.86	0.89
Tenth	640	0.57	0.71	0.64	0.76	0.83	0.80
Eleventh	552	0.93	0.76	0.83	0.95	0.82	0.88
Twelfth	644	0.75	0.73	0.74	0.87	0.84	0.85
UG Lower (Sampled)	750	0.73	0.74	0.74	0.84	0.84	0.84
UG Upper (Sampled)	750	0.92	0.72	0.81	0.96	0.85	0.90
Graduate (Sampled)	750	0.69	0.78	0.73	0.81	1.00	0.89
<b>Overall</b>	<b>10,238</b>	<b>0.75</b>	<b>0.75</b>	<b>0.75</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>

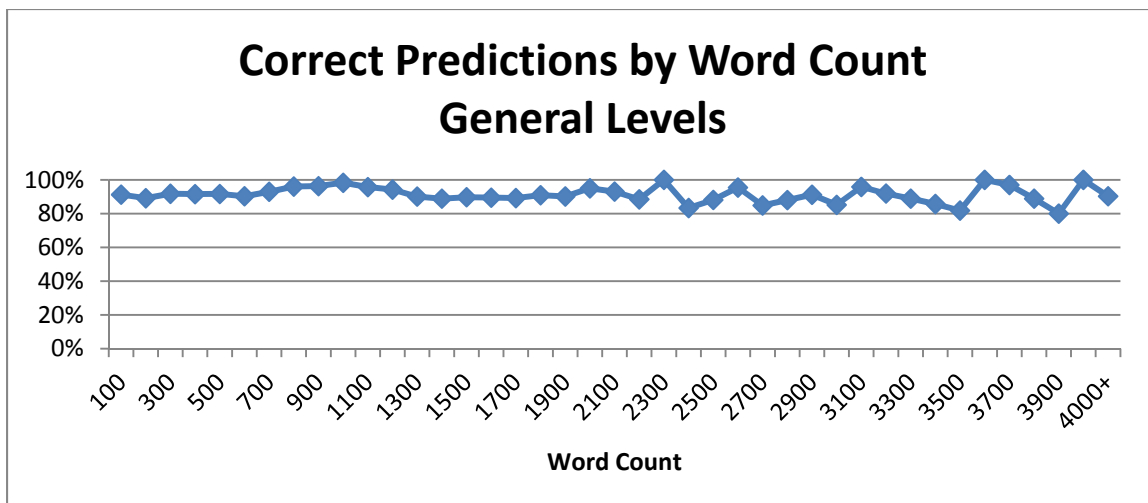
\*\* UG = Undergraduate

By reducing the noise in the training dataset, the classification performance increased for both general and specific audience levels over using the full text for training and testing. SVMAUD experienced the highest performance increase of 0.05 for general audience levels and 0.08 for specific audience levels. The other classification methods under evaluation, the cosine, Naïve Bayes, and the Collins-Thompson and Callan methods, also experienced increased performance, although the increase was smaller. SVMAUD successfully predicted the specific human-expert entered audience level with higher performance than the three other methods under evaluation at the 0.0001 level of

significance. As the digital library retrieval system should have presented the most appropriate information to users, resources should have been labeled with the most specific audience level. If the resources were stored with specific audience levels, such as first grade or second grade, a teacher searching for elementary school resources could have requested resources with audience levels ranging between first grade and fifth grade. However, if the audience level was entered at the general level, such as elementary school, then the retrieval system could have identified resources appropriate for elementary school and not first grade through third grade. By reducing the amount of noise in the training dataset, the prediction performance improved; therefore, this cleaned dataset should have been used to train SVMAUD to suggest the specific audience level for all resources in the collection.

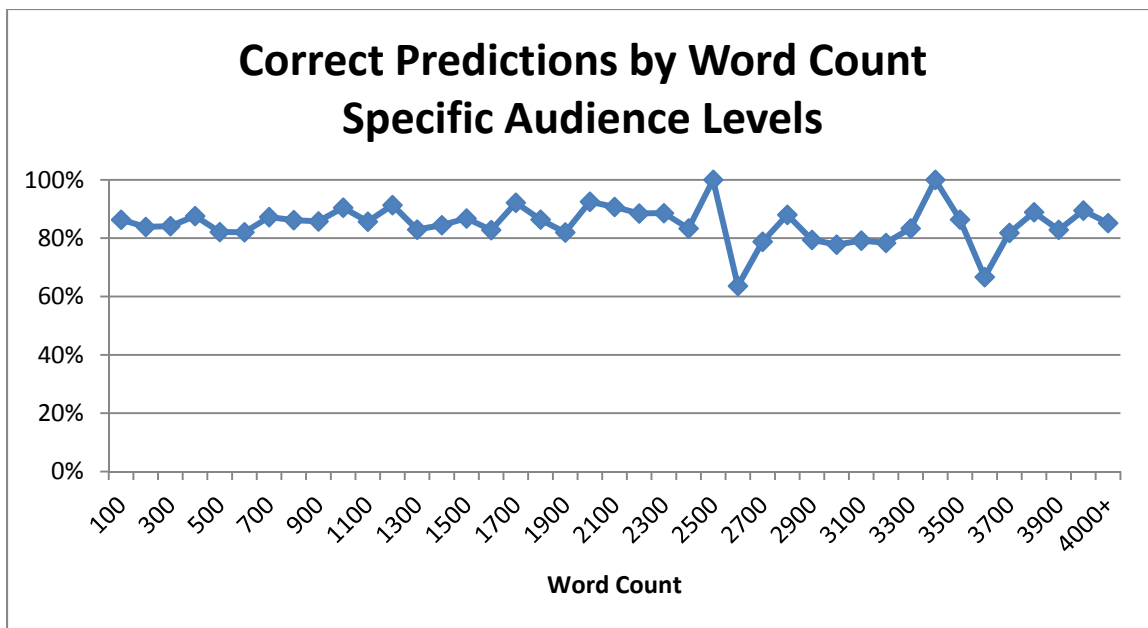
### **6.3 Effect of Resource Length on SVMAUD Classification Performance**

This part of the study sought to measure the prediction performance as a function of the number of terms in the document. As the number of terms in the document increased, the overall prediction performance should have improved as more words were available for comparison. The proxy for performance was the proportion of correct predictions for each category divided by total resources expert-labeled with the audience level. This part of the study considered the ability of SVMAUD to correctly predict the audience level when the title, abstract, and keywords were used for training, and the full text was used for testing. The first chart shows the performance based on the ability to correctly identify the general audience level for different resource lengths.



**Figure 6.1** Effect of resource length on performance – general audience levels.

The chart shown in the previous figure measured the performance of the classifier as a function of the number of words found in the document. SVMAUD suggested the audience level for each resource with performance over 80% for word counts ranging from 100 or fewer, up to 4,000 or more words. Even when the number of terms in the document was small, SVMAUD was able to predict the most appropriate audience level with high performance. The next chart on the following page displays the performance of SVMAUD as the word count increased for specific audience levels.



**Figure 6.2** Effect of resource length on performance – specific audience levels.

This part of the study measured the performance of SVMAUD as a function of word count when suggesting specific audience levels. The performance, for the most part, remained constant across all word counts, with performance exceeding 80% correct predictions. However, the performance with respect to documents containing 2,600 words and 3,600 words was far lower, around 60% correct predictions, since approximately twenty documents appeared in each of these categories while the remaining categories consisted of a hundred or more documents. Since a small number of documents contained these word counts, one incorrect prediction could have severely impacted the audience level prediction performance.

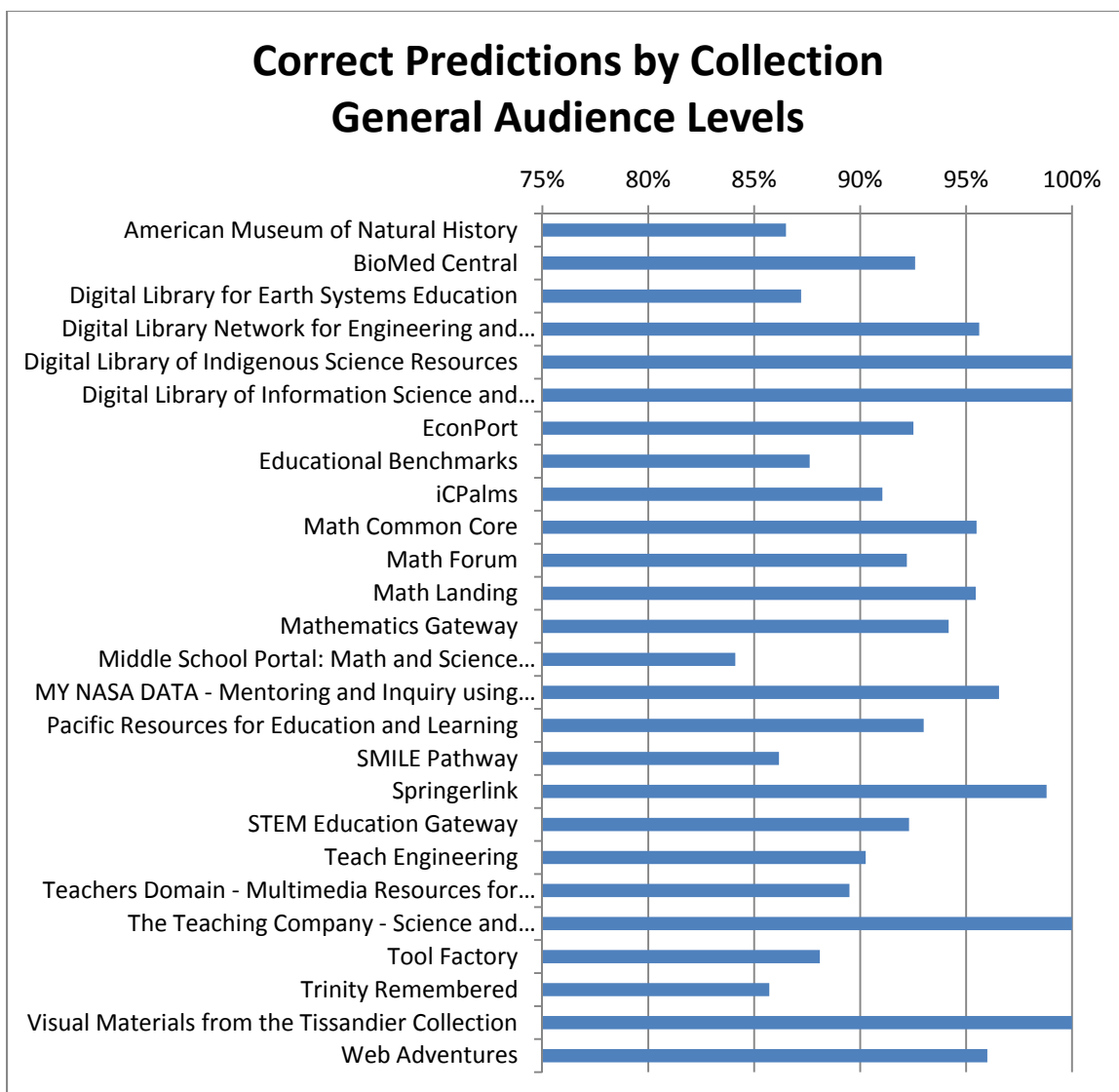
By reducing the amount of noise in the training dataset, SVMAUD's performance improved by 0.05 for general and 0.08 for specific audience levels. In addition, SVMAUD was able to predict the audience level with higher performance across all word counts. Only 802 out of the 10,238 resources in this collection contained more than

4,000 words. In fact, the longest resource used in the test dataset was held by the DLNET digital library named *Multicultural Pathways to Ocean Sciences Education*, and contained 30,413 words. Even though a resource could have contained a small number of words, SVMAUD was still able to make fine-grained distinctions between adjacent audience levels.

#### 6.4 SVMAUD Performance by Digital Library Collection

This part of the evaluation seeks to measure the performance of SVMAUD with respect to the individual digital library collection. Again, the performance is measured using the title, keywords, and abstract for training, and the full text for testing; the proxy for performance is the number of resources correctly labeled with audience level by SVMAUD divided by the total number of resources provided by the digital library collection. College and graduate level libraries typically hold resources with more specialized content than school libraries that cater to a population with lower audience levels. This analysis measures the prediction performance of SVMAUD for each collection; Figure 6.3 on the next page displays the performance by collection for general audience levels.

This chart showed that SVMAUD performance was much higher for collections holding resources appropriate for college level, including Springerlink, Digital Library of Indigenous Science Resources, and Digital Library of Information Science and Technology, since these resources generally contained a higher percentage of unique terms found at this audience level. On the other hand, the Middle School Portal and SMILE Pathway held resources for younger readers within a few audience levels.

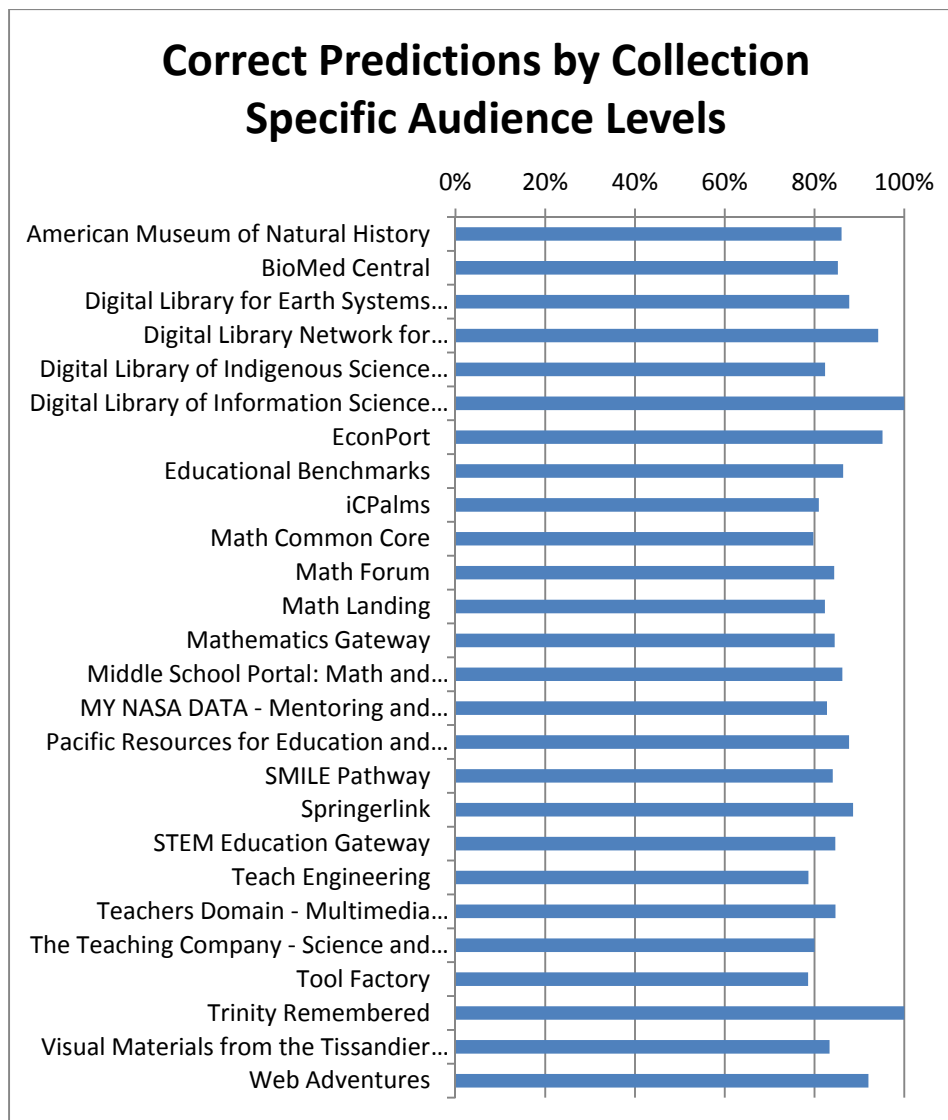


**Figure 6.3** SVMAUD performance by collection – general audience levels.

In addition, all of the resources held by these high-performing collections contained a low proportion of words common to all resources in the individual collection.

This next evaluation compares the performance of SVMAUD with respect to predicting specific audience levels. This performance changes substantially from the general audience level chart, as shown in Figure 6.4 on the next page.





**Figure 6.4** SVMAUD performance by collection – specific audience levels.

In this evaluation, the Digital Library of Information Science and Technology (DLIST) performed well, with nearly 100% correct predictions. The Trinity Remembered collection cataloged resources pertaining to the Trinity atomic test site, consisting of pictures, videos, and historical documents, targeted mainly toward higher audience levels. The Teach Engineering collection experienced the lowest performance at 79% correct predictions; this collection mainly held engineering resources spanning

kindergarten through twelfth grade levels. If the collection held resources appropriate for a small number of adjacent higher audience levels, its performance was generally higher than a collection that spans a wide range of audience levels.

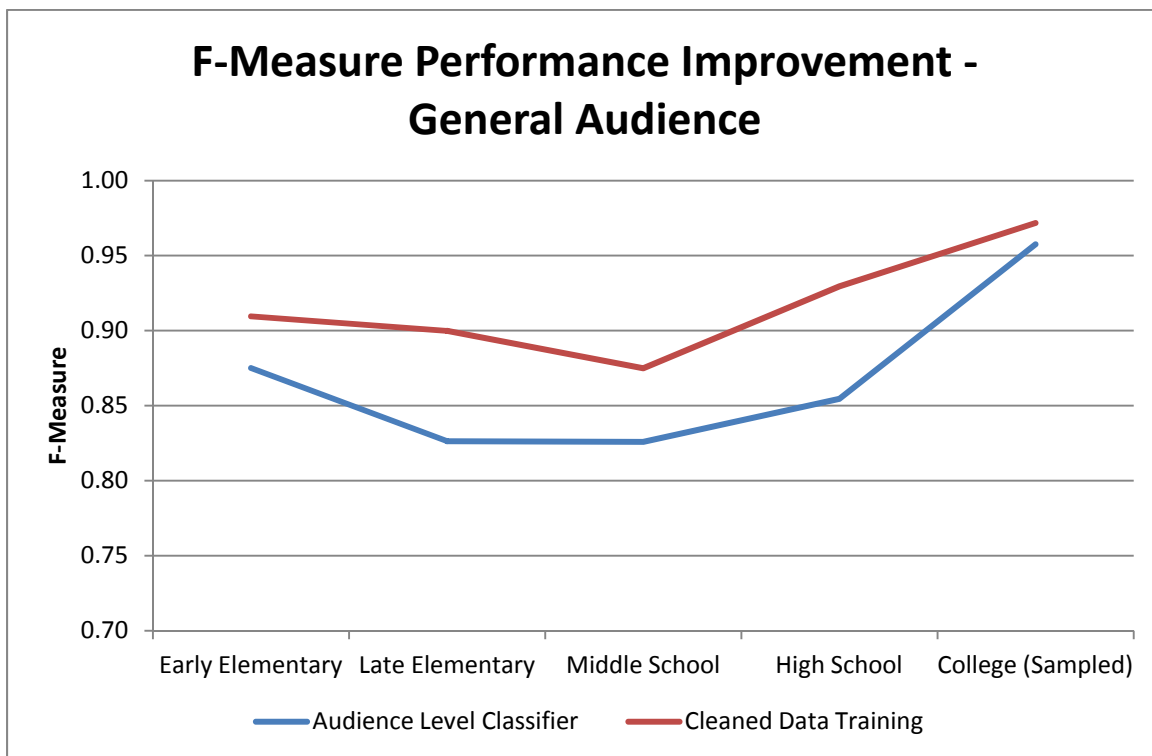
Overall, the digital libraries that catered to higher audience levels experienced higher performance over those that catered to lower audience levels. Since resources targeted toward younger readers required more common words to appeal to this audience, the prediction performance was lower for these collections; all elementary school students attending a school district would have generally followed the same curriculum. On the other hand, collections that held resources appropriate for college level students needed to be more specialized; college students could have chosen from a variety of different majors and the resources should have been written to target the topics taught by a particular course rather than the general student body.

### **6.5 SVMAUD Performance Improvement**

Since SVMAUD was trained using titles, keywords, and abstracts to predict the audience for resources in the collection, the performance had improved over using full text for training and testing. Figure 6.5 on the following page plots the general audience level prediction performance versus F-measure comparing full text for training and testing versus cleaned data being used for training.

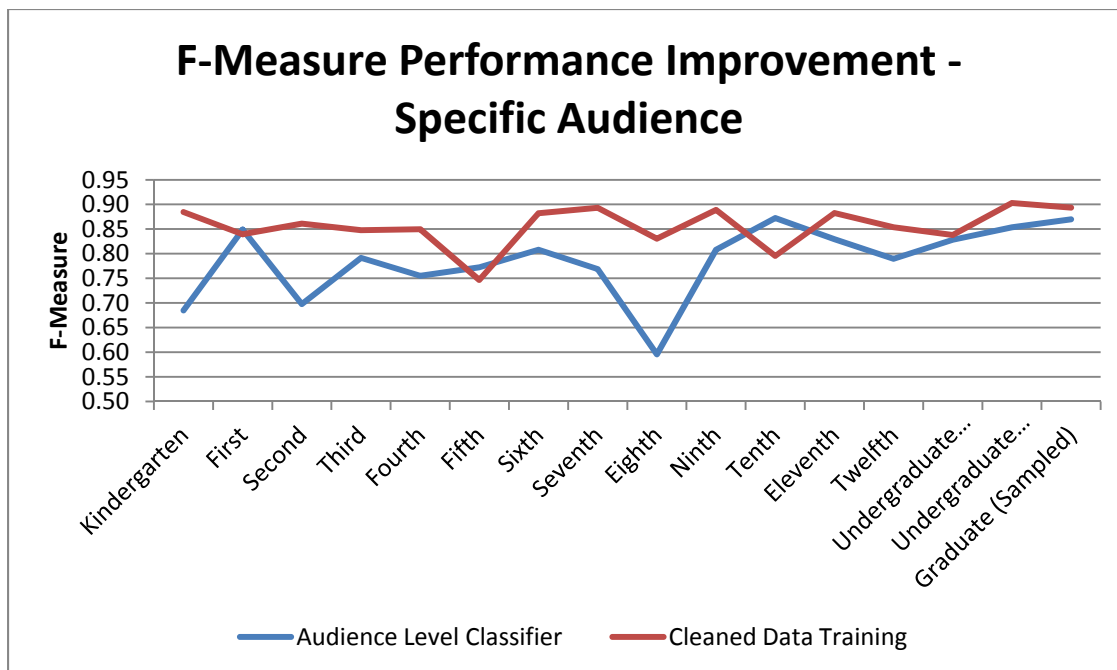
The performance improvement was highest at the high school audience level, while the lowest performance improvement occurred at the college audience level. Since PDF files in the college audience level contained formatting following grammatical and sentence conventions of written English and only contained text appropriate for the

audience level of the resource, the performance improvement was expected to be minimal for this level.



**Figure 6.5** SVMAUD general audience level performance improvement.

The following figure plots the prediction performance improvement across all audience levels for specific audience levels when considering the cleaned training dataset versus full text for training and testing.



**Figure 6.6** SVMAUD specific audience level performance improvement.

The specific audience level prediction performance substantially decreased for the tenth grade audience level. However, these performance decreases were balanced by the resources in the eighth grade audience level, which resulted in an overall F-measure improvement of almost 0.25 over the full text training and testing study. In total, SVMAUD prediction performance increased by approximately 0.08 across all specific audience levels.

## 6.6 Performance Summary and Conclusion

After removing the noise from the input documents by using abstract, keywords, and title to create the training dataset, the audience level prediction performance of SVMAUD increased. In the NSDL digital library collections, only title and URL were explicitly required to be entered when cataloging a new resource, while all other metadata elements

were optional. If the cataloger for the collection had not considered this additional metadata to be important, then a smaller amount of text was available to train the classification methods. Even though a smaller amount of text was available for training for each audience level, the prediction performance for all four machine learning methods increased by approximately 0.05 for general and 0.07 for specific audience levels. Since a small number of resources were cataloged with keywords and abstract versus the availability of full text for all resources, the classifier could not have used the abstract or keywords to predict the audience level of unlabeled resources. Therefore, the classifiers were trained using abstracts, titles, and keywords for all resources that included this information and then predicted the audience level by using the full text of the resource.

In addition to the performance of the noise reduced training dataset, another part of this study sought to compare SVMAUD performance based on the number of words in the resource as well as by collection. SVMAUD was found to experience high performance across all document lengths, ranging from fewer than 100 words to well over 4,000 words. In addition, collections containing college level resources typically contained a much higher proportion of unique words than those with lower audience levels. If all resources included complete title, keywords, and abstract metadata, then the noise-reduced classification performance should have further increased over using full text for training and testing.

After reducing the noise in the training documents by using title, keywords, and abstract, the classification performance of cosine, Naïve Bayes, and the Collins-Thompson and Callan method increased by about 0.05 for general audience levels, and by about 0.06 for specific audience levels. However, SVMAUD outperformed these

measures by correctly predicting the specific audience level with an F-measure of 0.86, an increase of 0.08 over using the full text for training and testing. When labeling resources with the most appropriate audience level, the noise-reduced training dataset should have been used to train the classification models.

This study also measured performance by the number of words that appeared in the document as well as by the collection providing the resource. SVMAUD, when using title, keywords, and abstract as training data, performed well across all word counts, ranging from fewer than 100 words to 4,000 words and above, indicating that a small amount of text was necessary for the classifier to correctly suggest the human-expert provided audience level. In general, SVMAUD performance, with respect to collections targeting higher audience levels, exceeded the performance of collections targeting lower audience levels, since the topics discussed in college level resources were generally more specialized.

SVMAUD was found to outperform the three other machine learning methods under evaluation. In addition, by reducing the amount of noise in the training dataset, its performance would have further increased over using the full text for each resource in the collection. SVMAUD, due to its high prediction performance, could have been used to predict the audience level for all resources held in a digital library collection containing missing or incompatible audience level metadata.

## **CHAPTER 7**

### **SUBJECT-SPECIFIC CLASSIFICATION**

In the previous performance tuning evaluations, the resources drawn from the NSDL collection are used to train and test a one-size-fits-all classifier, where the subject category is not considered. This evaluation seeks to develop a series of subject-specific classifiers, where the resources from one subject category are used to train the classifier to predict the audience level for other resources discussing the same subject. Since all of the NSDL resources in the previous studies contain an entry for the subject category metadata, this information is used to split the collection into the different subject categories. The six subject categories commonly taught in school consist of reading and writing, history and geography, health sciences, science, technology and engineering, and mathematics.

#### **7.1 Digital Library Collection Overview by Subject**

The digital library collection consisted of 10,238 resources drawn from NSDL collections and Springerlink to represent college level resources. Table 7.1 presents the distribution of documents across all subject categories commonly taught in grades kindergarten through college.

**Table 7.1** Digital Library Subject Category Coverage

<b>Subject Category</b>	<b>Audience Levels Covered</b>	<b>Docs</b>
Health Sciences	15	335
History & Geography	16	602
Mathematics	16	3,133
Reading & Writing	10	22
Science	16	4,301
Technology & Engineering	16	1,845
<b>Total</b>		<b>10,238</b>

The first column showed the subject category, the second column showed the number of audience levels that contained training resources, and the last column displayed the number of documents labeled with each subject category. Since the NSDL mainly held STEM, or science, technology, engineering, and mathematics resources, these subject categories contained a much higher number of resources than the other subject categories. If no resource was expert-labeled with a specific audience level in the training dataset, then no unlabeled resources would have been placed into that audience level by the subject-specific classifier. Reading and writing, with a total of twenty-two resources, only contain training resources in ten out of the sixteen specific audience levels used in this study; similarly, the health sciences category did not contain any resources for one out of the sixteen possible audience levels. Even though the history and geography subject category contained resources spanning all audience levels, the resources were not evenly distributed among all possible audience levels, ranging from a low of nine resources in the kindergarten audience level to a high of ninety-six resources in the fourth grade audience level. Since all classifiers required an approximately equal number of resources for each audience level in the training dataset in order to perform



well, additional resources needed to be collected to represent all subjects commonly taught in grades K-college.

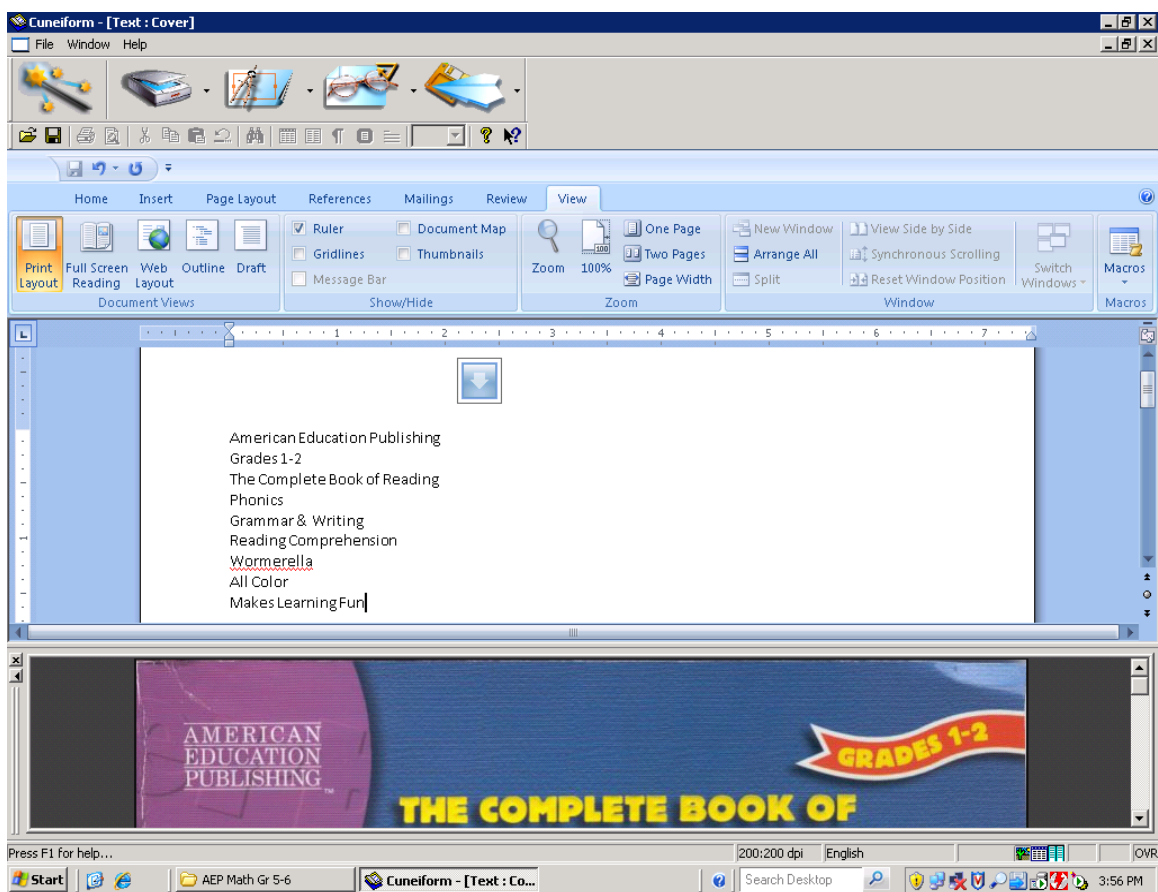
## **7.2 Home School Resource Collection Overview**

If a parent would rather teach his or her child at home rather than sending them to the local school district, the home school collection provides such parents with a large number of resources to educate their homeschooled children. Three main publishers provide resources to students and educators in home schooling education programs, namely A Beka Book, American Education Publishing (AEP), and the Teacher's Syndicate. A Beka Book seeks to provide Christian and home schools with the best academic resources available. AEP is a part of Carson-Dellosa publishing that seeks to provide innovative solutions and resources to students in grades kindergarten through eighth. The Teaching Syndicate provides fun and educational resources to home schooling parents and educators in grades kindergarten through twelve; after registering with the site, the educator could browse a wide variety of resources to create lesson plans. Since a home school student could not receive a college degree by his or her parents, additional resources from the Springerlink collection, cataloging journal articles and conference papers appropriate for college students, are used to represent the vocabulary found in higher levels of education.

These home school collections consist of books, pamphlets, short exercises, activities, and other educational materials that a teacher or parent can use to develop lesson plans. Since all of these resources, with the exception of the Springerlink collection, consist of scanned book pages that the parent can print and hand to his or her

child, the text in these scanned images is extracted and converted into text files for use by the classification programs.

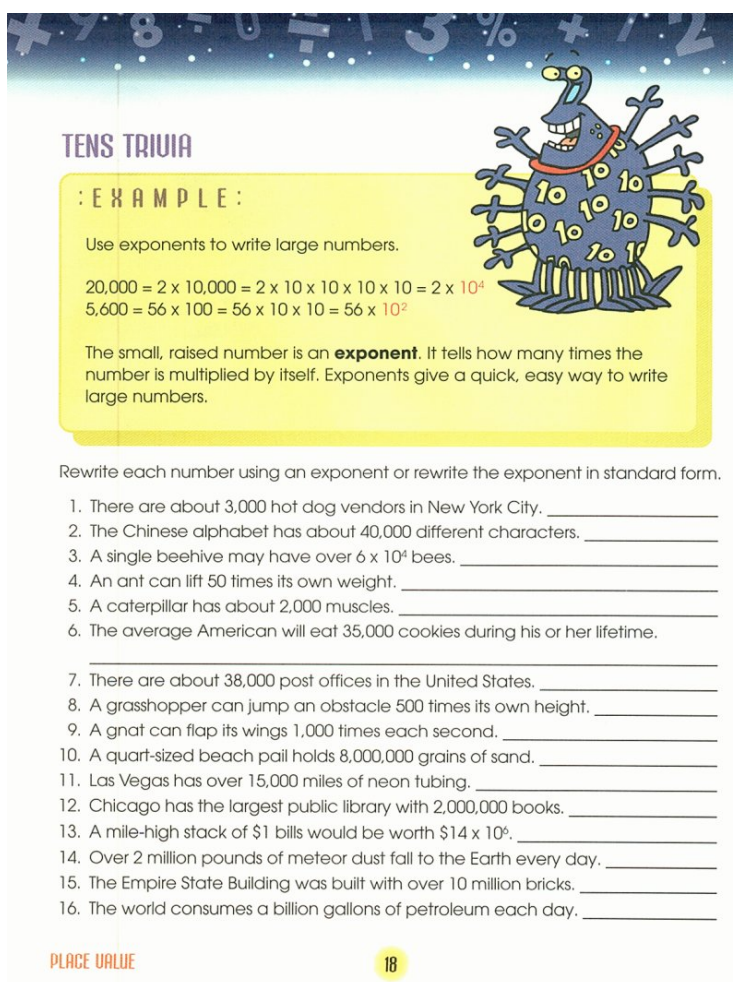
To automatically extract the text, the Cuneiform Optical Character Recognition (OCR) program is chosen due to its ability to accurately extract text from a set of images. After loading the image in the bottom of the screen, the extracted text is shown at the top of the screen. The following figure shows the results of extracting the text from a printable textbook (Cognitive Technologies, 2012):



**Figure 7.1** Screen capture of Cuneiform program.

This program also supports batch conversion of image files to text files by replacing the image file with the associated text file containing extracted text. After running this program to extract the text from each image, the text from each page is combined to create one text file representing a single resource.

The following two figures show a sample book page and the extracted text from the same page from the Complete Book of Math, Grades 5-6. This book is written for students in grades five and six and consists of a variety of different math problems that the student can solve at his or her own pace (McGraw-Hill Children's Publishing, 2002).



**TENS TRIVIA**

**: EXAMPLE :**

Use exponents to write large numbers.

$$20,000 = 2 \times 10,000 = 2 \times 10 \times 10 \times 10 \times 10 = 2 \times 10^4$$

$$5,600 = 56 \times 100 = 56 \times 10 \times 10 = 56 \times 10^2$$

The small, raised number is an **exponent**. It tells how many times the number is multiplied by itself. Exponents give a quick, easy way to write large numbers.

Rewrite each number using an exponent or rewrite the exponent in standard form.

- There are about 3,000 hot dog vendors in New York City. \_\_\_\_\_
- The Chinese alphabet has about 40,000 different characters. \_\_\_\_\_
- A single beehive may have over  $6 \times 10^4$  bees. \_\_\_\_\_
- An ant can lift 50 times its own weight. \_\_\_\_\_
- A caterpillar has about 2,000 muscles. \_\_\_\_\_
- The average American will eat 35,000 cookies during his or her lifetime. \_\_\_\_\_
- There are about 38,000 post offices in the United States. \_\_\_\_\_
- A grasshopper can jump an obstacle 500 times its own height. \_\_\_\_\_
- A gnat can flap its wings 1,000 times each second. \_\_\_\_\_
- A quart-sized beach pail holds 8,000,000 grains of sand. \_\_\_\_\_
- Las Vegas has over 15,000 miles of neon tubing. \_\_\_\_\_
- Chicago has the largest public library with 2,000,000 books. \_\_\_\_\_
- A mile-high stack of \$1 bills would be worth  $\$14 \times 10^6$ . \_\_\_\_\_
- Over 2 million pounds of meteor dust fall to the Earth every day. \_\_\_\_\_
- The Empire State Building was built with over 10 million bricks. \_\_\_\_\_
- The world consumes a billion gallons of petroleum each day. \_\_\_\_\_

PLACE VALUE 18

**Figure 7.2** Page 18 image from The Complete Book of Math, Grades 5-6.

## Tens Trivia

:Example:

Use exponents to write large numbers

$$20,000=2 \times 10,000=2 \times 10 \times 10 \times 10 \times 10=2 \times 10^4$$

$$5,600=56 \times 100=56 \times 10 \times 10=56 \times 10^2$$

The small raised number is an exponent. It tells you how many times the number is multiplied by itself Exponents give a quick easy way to write large numbers

Rewrite each number using an exponent or rewrite the exponent in standard form

- 1 There are about 3,000 hot dog vendors in New York City.
- 2 The Chinese alphabet had about 40,000 different characters.
- 3 A single beehive may have over  $6 \times 10^4$  bees.
- 4 An ant can lift 50 times its own weight.
- 5 A caterpillar had about 2,000 muscles.
- 6 The average American will eat 35,000 cookies during his or her lifetime.
- 7 There are about 38,000 post offices in the United States.
- 8 A grasshopper can jump an obstacle 500 times its own height.
- 9 A gnat can flap its wings 1,000 times each second.
- 10 A quart sized beach pail holds 8,000,000 grains of sand.
- 11 Las Vegas had over 15,000 miles of neon tubing.
- 12 Chicago had the largest public library with 2,000,000 books.
- 13 A mile-high stack of 1 bills would be worth  $14 \times 10^6$ .
- 14 Over 2 million pounds of meteor dust fall to the Earth every day.
- 15 The Empire State Building was built with over 10 million bricks.
- 16 The world consumes a billion gallons of petroleum each day.

**Figure 7.3** Page 18 text extracted from The Complete Book of Math, Grades 5-6.

The second sample resource is called the Complete Book of Reading and published by American Education Publishing. This resource is appropriate for students in first and second grade who are beginning readers, so the emphasis is on identifying different words and sounds rather than complete sentences. The following two figures on the next two pages present the original image file representing the page in the book and the extracted text from that page (American Education Publishing, 2000).


A B C D E F G H I J K L

## Color the Letter Partners

Name \_\_\_\_\_

Letter partners are capital and small letters that go together. These pairs of letters are letter partners: **Aa, Bb, Cc, Dd, Ee, Ff, Gg, Hh, Ii, Jj, Kk, Ll, Mm, Nn, Oo, Pp, Qq, Rr, Ss, Tt, Uu, Vv, Ww, Xx, Yy, Zz.**

◆ **Directions:** Use a different color to color each pair of letter partners.

Letter Recognition  © 2000 Tribune Education. All Rights Reserved.

**Figure 7.4** Page 8 image from The Complete Book of Reading.

Color the Letter Partners Name

Letter partners are capital and small letters that go together. These pairs of letters are letter partners: Aa, Bb, Cc, Dd, Ee, Ff, Gg, Hh, Ii, Jj, Kk, Ll, Mm, Nn, Oo, Pp, Qq, Rr, Ss, Tt, Uu, Vv, Ww, Xx, Yy, Zz.

Directions use a different color to color each pair of letter partners.

M q B

M n

B G D

N d Q g

Letter Recognition 8 O 2000 Tribune Education. All Rights Reserved.

**Figure 7.5** Page 8 text extracted from The Complete Book of Reading.

The text from each of these individual pages is combined into one text file representing all of the text extracted from the book page images. These text files are used to augment the resources from the digital library collection to cover a wider variety of topics taught to students. By developing a series of six subject-specific audience level classifiers, the prediction performance of SVM AUD and other machine learning methods should improve over a one-size-fits-all classifier covering all subject categories.

Resources from the Springerlink collection were again included to represent college level vocabulary, since home school resources were generally targeted toward kindergarten through twelfth grade students. A total of 4,039 resources were added to the 10,238 digital library resources already in the test collection, for a total of 14,277 resources spread among all subjects commonly taught in elementary school through college. The following table summarizes the collections that have provided resources used in this evaluation.

**Table 7.2** Home School Resource Collection Summary

<b>Collection</b>	<b>Collection URL</b>	<b>Documents</b>
A Beka Books	<a href="http://www.abeka.com">http://www.abeka.com</a>	1,024
Carson Dellosa Publishing	<a href="http://www.carsondellosa.com">http://www.carsondellosa.com</a>	963
Springerlink	<a href="http://www.springerlink.com">http://www.springerlink.com</a>	900
Teaching Syndicate	<a href="http://www.teachersyndicate.com">http://www.teachersyndicate.com</a>	1,152
<b>Total Documents</b>		<b>4,039</b>

These home school resources spanned a wide range of subjects commonly found in kindergarten through high school grades. The subjects included reading and writing, mathematics, health, and geography, and other subjects, in addition to the STEM topics held by the digital library collection. The next table shows the subjects discussed by the resources in the collection, along with the number of documents in each subject category for both home school and digital library collections.

**Table 7.3** Subject-Specific Classifier Document Collection Summary

<b>Subject Category</b>	<b>Home School Collection</b>	<b>Digital Library Collection</b>	<b>Total Documents</b>
Health Sciences	443	335	778
History & Geography	742	602	1,344
Mathematics	950	3,133	4,083
Reading & Writing	722	22	744
Science	310	4,301	4,611
Technology & Engineering	872	1,845	2,717
<b>Total</b>	<b>4,039</b>	<b>10,238</b>	<b>14,277</b>

Even though the home school resources covered a wider variety of topics, there were a smaller number of resources for each subject category. Since the NSDL mainly focused on cataloging STEM topics, the Science, Technology & Engineering, and

Mathematics subject categories contained a much higher number of resources than the health sciences, history and geography, and reading and writing categories. However, the subject category with the lowest number of resources, reading and writing, contained approximately 750 resources spread across all audience levels, so all classifiers should have performed well. Since the titles, abstracts, and keywords used for training SVMAUD outperformed the full text and HTML tag processing for training SVMAUD, the titles, abstracts, and keywords, when available, were used for training the classifier, while the full text was used for testing. However, with regard to the home school resources, there was little duplication between different books, all text was appropriate for the resource and did not include menus and headers common to every resource, and abstracts were not available for this collection, the full text of each resource was used for training and testing. Precision (P), Recall (R), F-Measure (F), correlation, and the t-test were used to evaluate the performance of cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVM with respect to the six subject category classifiers.

### **7.3 Health Sciences Subject-Specific Classifier Performance**

Health sciences covered a wide variety of topics, ranging from physical fitness exercises and eating habits in elementary school to medical literature appropriate for doctors and other medical professionals. This study trained and tested the cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD classifiers with respect to the health sciences subject category. The results from this study are shown in the following two tables. The first table shows the results of the specific audience level prediction study for cosine, Naïve Bayes, and SVMAUD.



**Table 7.4** Health Sciences Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	100	0.60	0.72	0.65	0.72	0.81	0.76	0.82	0.90	0.86
First	7	0.07	0.43	0.12	0.08	0.43	0.14	0.28	0.71	0.40
Second	30	0.52	0.83	0.64	0.60	0.83	0.69	0.70	0.93	0.80
Third	25	0.79	0.60	0.68	0.84	0.64	0.73	0.95	0.80	0.87
Fourth	23	0.38	0.52	0.44	0.47	0.74	0.58	0.76	0.83	0.79
Fifth	8	0.11	1.00	0.20	0.16	1.00	0.28	0.22	1.00	0.36
Sixth	44	0.91	0.70	0.79	0.97	0.77	0.86	1.00	0.91	0.95
Seventh	2	0.00	0.00	N/A	0.33	0.50	0.40	1.00	1.00	1.00
Eighth	70	0.94	0.66	0.77	0.94	0.73	0.82	1.00	0.87	0.93
Ninth	38	0.47	0.24	0.32	0.75	0.47	0.58	0.89	0.66	0.76
Tenth	70	0.23	0.31	0.27	0.40	0.60	0.48	0.75	0.69	0.72
Eleventh	79	0.71	0.52	0.60	0.80	0.59	0.68	0.96	0.81	0.88
Twelfth	36	0.91	0.56	0.69	0.95	0.58	0.72	1.00	0.94	0.97
Undergraduate Lower (Sampled)	70	0.96	0.63	0.76	0.96	0.67	0.79	0.98	0.89	0.93
Undergraduate Upper (Sampled)	82	0.98	0.61	0.75	1.00	0.66	0.79	1.00	0.85	0.92
Graduate (Sampled)	94	0.84	0.62	0.71	0.88	0.71	0.79	0.95	1.00	0.97
<b>Overall</b>	<b>778</b>	<b>0.59</b>	<b>0.59</b>	<b>0.59</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>

\*\* N/A = Not Available; F-Measure cannot be calculated due to division by zero.

The first, fifth, and seventh grades contained the lowest number of documents, with a total count under ten for each of these grades, leading to poor prediction performance in the health science subject category. The remaining audience levels contained a higher number of resources and performed well. Overall, cosine performance decreased slightly from 0.61 for the single audience level classifier to 0.59 for this subject-specific classifier; the performance of the Naïve Bayes and SVMAUD classifiers remained the same. With respect to the correlation between human-expert entered and machine-learning suggested specific audience levels, SVMAUD experienced the highest

correlation at 0.89, Naïve Bayes experienced a correlation at 0.76, and cosine experienced the lowest correlation at 0.70. SVMAUD outperformed cosine at the 0.0010 level of significance and Naïve Bayes at the 0.0060 level of significance.

The next part of this study considers the prediction performance of the Collins-Thompson and Callan method versus SVMAUD, as presented in the next table.

**Table 7.5** Health Sciences Specific Audience Level Prediction–Thompson & Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	100	0.79	0.86	0.82	0.82	0.90	0.86
First	7	0.11	0.43	0.17	0.28	0.71	0.40
Second	30	0.70	0.87	0.78	0.70	0.93	0.80
Third	25	0.86	0.72	0.78	0.95	0.80	0.87
Fourth	23	0.58	0.78	0.67	0.76	0.83	0.79
Fifth	8	0.21	1.00	0.34	0.22	1.00	0.36
Sixth	44	0.97	0.82	0.89	1.00	0.91	0.95
Seventh	2	1.00	0.50	0.67	1.00	1.00	1.00
Eighth	70	0.95	0.81	0.88	1.00	0.87	0.93
Ninth	38	0.83	0.63	0.72	0.89	0.66	0.76
Tenth	70	0.51	0.69	0.59	0.75	0.69	0.72
Eleventh	79	0.85	0.72	0.78	0.96	0.81	0.88
Twelfth	36	0.96	0.67	0.79	1.00	0.94	0.97
Undergraduate Lower (Sampled)	70	0.96	0.76	0.85	0.98	0.89	0.93
Undergraduate Upper (Sampled)	82	1.00	0.77	0.87	1.00	0.85	0.92
Graduate (Sampled)	94	0.90	0.79	0.84	0.95	1.00	0.97
<b>Overall</b>	<b>778</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>

SVMAUD outperformed the Collins-Thompson and Callan method with an overall F-measure of 0.86 versus an F-measure of 0.77 for the Collins-Thompson and Callan method across all specific audience levels. SVMAUD also experienced a higher

correlation between human-expert entered and machine-learning suggested audience levels of 0.89 versus 0.82 for the Collins-Thompson and Callan method. In fact, SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0673 level of significance. SVMAUD was found to outperform the three other machine learning methods under evaluation and should have been used to predict the specific audience level for unlabeled health sciences resources.

This evaluation considered the general audience level prediction performance among the machine learning methods. The next table shows the results for the health sciences classifier for cosine, Naïve Bayes, and SVMAUD.

**Table 7.6** Health Sciences General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	137	0.47	0.63	0.54	0.58	0.74	0.65	0.90	0.95	0.93
Late Elementary	56	0.54	0.27	0.36	0.78	0.50	0.61	0.96	0.91	0.94
Middle School	116	0.63	0.82	0.71	0.71	0.86	0.78	0.92	0.97	0.95
High School	223	0.66	0.45	0.54	0.79	0.60	0.68	0.97	0.91	0.94
College (Sampled)	246	0.92	0.98	0.95	0.95	0.99	0.97	0.99	1.00	0.99
<b>Overall</b>	<b>778</b>	<b>0.69</b>	<b>0.69</b>	<b>0.69</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>

This portion of the study trained and tested the cosine, Naïve Bayes, and SVMAUD classifiers using general audience levels. In this evaluation, the prediction performance improved slightly over the general digital library audience level classifier, with F-measures increasing from 0.66 to 0.69 for cosine, from 0.75 to 0.78 for Naïve Bayes, and from 0.92 to 0.95 for SVMAUD. The correlation between human-expert entered and machine-learning suggested values was found to be 0.67 for cosine, 0.76 for Naïve Bayes, and 0.95 for SVMAUD. SVMAUD was found to outperform cosine at the

0.0149 level of significance and Naïve Bayes at the 0.0154 level of significance. SVMAUD outperformed the cosine and Naïve Bayes machine learning methods under evaluation in this part of the study.

The next part of this study measured the performance of SVMAUD versus the Collins-Thompson and Callan method. The results from this part of the study are shown in the following table.

**Table 7.7** Health Sciences General Audience Level Prediction – Thompson & Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	137	0.71	0.83	0.77	0.90	0.95	0.93
Late Elementary	56	0.87	0.70	0.77	0.96	0.91	0.94
Middle School	116	0.77	0.90	0.83	0.92	0.97	0.95
High School	223	0.87	0.72	0.79	0.97	0.91	0.94
College (Sampled)	246	0.96	1.00	0.98	0.99	1.00	0.99
<b>Overall</b>	<b>778</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.95</b>	<b>0.95</b>	<b>0.95</b>

SVMAUD outperformed the Collins-Thompson and Callan method when predicting the human-expert entered audience level for unlabeled resources, with an overall F-measure of 0.95 versus 0.85 for the Collins-Thompson and Callan method. The correlation between human-expert entered and machine-learning suggested general audience levels was found to be 0.84 for the Collins-Thompson and Callan method versus 0.95 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0174 level of significance.

For all evaluations in this section, the prediction performance for both general and specific audience levels was close to the prediction performance of the subject-combined

classifier discussed in the earlier chapters. SVMAUD outperformed all other methods at the 0.0673 level of significance for specific audience level prediction and the 0.0174 level of significance for general audience level prediction. SVMAUD should have been used to predict the audience level for all unlabeled resources discussing health sciences, since it far outperformed all other methods under evaluation.

#### **7.4 History and Geography Subject-Specific Classifier Performance**

The history and geography subject category contained documents ranging from state capitals in the United States in elementary school geography to the study of ancient cultures and archaeology taught at the college level. This study extracted all documents that discuss history and geography from the home school and digital library collections to train and test the four different classifiers. Table 7.8 on the next page shows the results from the specific audience level prediction study for the history & geography subject category for cosine, Naïve Bayes, and SVMAUD.

The performance for all three classifiers under evaluation improved over the general subject category classifier. The cosine classifier improved from an F-measure of 0.61 to 0.66, the Naïve Bayes classifier improved from an F-measure of 0.68 to 0.74, and SVMAUD improved from an F-measure of 0.86 to 0.90. The correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.72 for cosine, 0.77 for Naïve Bayes, and 0.94 for SVMAUD. SVMAUD was found to outperform both cosine and Naïve Bayes at the 0.0001 level of significance.

**Table 7.8** History & Geography Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	14	0.58	0.79	0.67	0.61	0.79	0.69	0.75	0.86	0.80
First	129	0.61	0.60	0.60	0.72	0.71	0.71	0.92	0.84	0.88
Second	60	0.65	0.85	0.74	0.70	0.87	0.78	0.87	0.97	0.91
Third	92	0.72	0.83	0.77	0.78	0.93	0.85	0.87	0.96	0.91
Fourth	183	0.64	0.77	0.70	0.71	0.83	0.76	0.87	0.92	0.90
Fifth	87	0.32	0.79	0.45	0.39	0.83	0.53	0.64	0.90	0.75
Sixth	113	0.75	0.48	0.58	0.80	0.58	0.67	0.92	0.82	0.87
Seventh	95	0.94	0.52	0.67	0.97	0.66	0.79	0.99	0.84	0.91
Eighth	107	0.66	0.78	0.71	0.70	0.82	0.76	0.88	0.96	0.92
Ninth	91	0.98	0.59	0.74	0.98	0.66	0.79	1.00	0.87	0.93
Tenth	97	0.96	0.68	0.80	0.96	0.71	0.82	0.99	0.89	0.93
Eleventh	68	0.82	0.60	0.69	0.89	0.62	0.73	0.94	0.90	0.92
Twelfth	85	0.90	0.54	0.68	0.93	0.61	0.74	0.99	0.91	0.94
Undergraduate Lower (Sampled)	26	0.68	0.65	0.67	0.86	0.73	0.79	0.96	0.88	0.92
Undergraduate Upper (Sampled)	36	0.84	0.72	0.78	0.90	0.78	0.84	0.97	0.89	0.93
Graduate (Sampled)	61	0.67	0.52	0.59	0.75	0.66	0.70	0.98	0.98	0.98
<b>Overall</b>	<b>1,344</b>	<b>0.66</b>	<b>0.66</b>	<b>0.66</b>	<b>0.74</b>	<b>0.74</b>	<b>0.74</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>

The next table on the following page presents the specific audience level prediction performance for the history and geography subject category.

**Table 7.9** History & Geography Specific Audience Prediction–Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	14	0.61	0.79	0.69	0.75	0.86	0.80
First	129	0.73	0.74	0.73	0.92	0.84	0.88
Second	60	0.77	0.88	0.82	0.87	0.97	0.91
Third	92	0.79	0.89	0.84	0.87	0.96	0.91
Fourth	183	0.77	0.85	0.81	0.87	0.92	0.90
Fifth	87	0.44	0.89	0.59	0.64	0.90	0.75
Sixth	113	0.83	0.63	0.71	0.92	0.82	0.87
Seventh	95	0.96	0.69	0.80	0.99	0.84	0.91
Eighth	107	0.75	0.83	0.79	0.88	0.96	0.92
Ninth	91	0.98	0.71	0.83	1.00	0.87	0.93
Tenth	97	0.97	0.75	0.85	0.99	0.89	0.93
Eleventh	68	0.89	0.75	0.82	0.94	0.90	0.92
Twelfth	85	0.93	0.67	0.78	0.99	0.91	0.94
Undergraduate Lower (Sampled)	26	0.81	0.65	0.72	0.96	0.88	0.92
Undergraduate Upper (Sampled)	36	0.91	0.81	0.85	0.97	0.89	0.93
Graduate (Sampled)	61	0.78	0.77	0.78	0.98	0.98	0.98
<b>Overall</b>	<b>1,344</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>

SVMAUD was again found to outperform the Collins-Thompson and Callan method, with an overall F-measure of 0.90 versus 0.77 for the Collins-Thompson and Callan method. SVMAUD also outperformed the Collins-Thompson and Callan method in regards to the correlation between human-expert entered and machine-learning suggested values, with a correlation of 0.81 for the Collins-Thompson and Callan method and 0.94 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0001 level of significance.

Since a specific subject category was used for training and testing, the proportion of unique terms in each category increased, leading to an increased ability to discriminate

between adjacent audience levels. In addition, the number of documents in this subject category was spread more evenly among audience levels, leading to higher performance over the health sciences classifier.

The next table displays the results from the general audience level prediction portion of this study for the history & geography classifier.

**Table 7.10** History & Geography General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	203	0.59	0.62	0.60	0.71	0.69	0.70	0.92	0.91	0.92
Late Elementary	362	0.71	0.74	0.72	0.79	0.81	0.80	0.95	0.94	0.95
Middle School	315	0.64	0.64	0.64	0.74	0.75	0.74	0.94	0.94	0.94
High School	341	0.73	0.76	0.74	0.78	0.81	0.79	0.93	0.97	0.95
College (Sampled)	123	0.83	0.54	0.66	0.87	0.69	0.77	0.99	0.90	0.94
<b>Overall</b>	<b>1,344</b>	<b>0.68</b>	<b>0.68</b>	<b>0.68</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

The history and geography subject-specific classifiers also experienced increased performance over the classifier that predicted the audience level for all subject categories. The cosine classifier experienced increased performance as measured by the overall F-measure, from 0.66 to 0.68, the Naïve Bayes classifier increased the F-measure performance from 0.75 to 0.77, and SVMAUD increased the F-measure performance from 0.92 to 0.94. The correlation between human-expert entered and machine-learning suggested values was found to be 0.65 for cosine, 0.74 for Naïve Bayes, and 0.92 for



SVMAUD. In fact, SVMAUD outperformed both Naïve Bayes and cosine at the 0.0002 level of significance.

Similar results were found by using the Collins-Thompson and Callan method compared to SVMAUD. The results are presented in the following table.

**Table 7.11** History & Geography General Audience Prediction – Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	203	0.77	0.79	0.78	0.92	0.91	0.92
Late Elementary	362	0.82	0.87	0.84	0.95	0.94	0.95
Middle School	315	0.81	0.76	0.78	0.94	0.94	0.94
High School	341	0.84	0.87	0.85	0.93	0.97	0.95
College (Sampled)	123	0.91	0.76	0.83	0.99	0.90	0.94
<b>Overall</b>	<b>1,344</b>	<b>0.82</b>	<b>0.82</b>	<b>0.82</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

In this study, the Collins-Thompson and Callan method was compared against SVMAUD in predicting the general audience level for resources labeled with the history and geography subject category, with an overall F-measure of 0.82 for the Collins-Thompson and Callan method and 0.94 for SVMAUD. The correlation between human-expert entered and machine-learning suggested general audience levels was found to be 0.81 for the Collins-Thompson and Callan method and 0.92 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0003 level of significance.

By creating a separate classifier to suggest the audience level for history and geography resources, the prediction performance increased slightly across most classifiers, with the exception of the Collins-Thompson and Callan method decreasing

slightly by 0.02 over the single subject category classifier with respect to general audience levels. SVMAUD was found to outperform all other methods under evaluation in this study and should have been used to predict the audience level for all unlabeled resources in the history and geography subject category.

### **7.5 Mathematics Subject-Specific Classifier Performance**

The resources for the mathematics subject covered the entire range of audience levels, ranging from simple addition and subtraction problems taught in kindergarten to calculus and trigonometry taught at the college level. This part of the study trained and tested the four classifiers using documents labeled with the mathematics subject category. Table 7.12 on the following page displays the specific audience level prediction performance for cosine, Naïve Bayes, and SVMAUD for documents discussing mathematics.

The mathematics subject category performance approximately followed the one-size-fits-all classifier performance, where all documents were used for training and testing rather than developing a classifier for each subject category. The cosine classifier performance improved from 0.61 to 0.62, the Naïve Bayes classifier performance improved from 0.68 to 0.70, and SVMAUD performance slightly improved from 0.86 to 0.87. The correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.69 for cosine, 0.75 for Naïve Bayes, and 0.89 for SMVAUD. SVMAUD was found to outperform both cosine and Naïve Bayes at the 0.0001 level of significance.

**Table 7.12** Mathematics Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	507	0.82	0.64	0.72	0.86	0.71	0.78	0.93	0.88	0.90
First	388	0.63	0.63	0.63	0.71	0.70	0.70	0.87	0.84	0.85
Second	285	0.55	0.66	0.60	0.64	0.74	0.69	0.81	0.87	0.84
Third	95	0.65	0.48	0.55	0.74	0.53	0.61	0.91	0.79	0.85
Fourth	241	0.55	0.67	0.61	0.62	0.77	0.69	0.82	0.90	0.85
Fifth	159	0.34	0.77	0.47	0.42	0.86	0.57	0.66	0.92	0.77
Sixth	107	0.80	0.55	0.65	0.84	0.64	0.72	0.95	0.80	0.87
Seventh	330	0.96	0.54	0.69	0.97	0.62	0.76	0.99	0.82	0.90
Eighth	212	0.35	0.67	0.46	0.43	0.74	0.54	0.69	0.91	0.78
Ninth	430	0.89	0.65	0.75	0.91	0.71	0.80	0.97	0.89	0.93
Tenth	298	0.72	0.52	0.61	0.78	0.61	0.69	0.92	0.87	0.89
Eleventh	326	0.90	0.61	0.73	0.94	0.69	0.80	0.99	0.85	0.91
Twelfth	397	0.79	0.63	0.70	0.82	0.70	0.76	0.93	0.86	0.89
Undergraduate Lower (Sampled)	136	0.38	0.65	0.48	0.46	0.72	0.56	0.66	0.86	0.75
Undergraduate Upper (Sampled)	57	0.54	0.56	0.55	0.67	0.68	0.68	0.83	0.88	0.85
Graduate (Sampled)	115	0.36	0.63	0.46	0.42	0.67	0.52	0.71	0.98	0.82
<b>Overall</b>	<b>4,083</b>	<b>0.62</b>	<b>0.62</b>	<b>0.62</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The next part of this study considered the ability of the Collins-Thompson and Callan method and SVMAUD to correctly predict the specific audience level for resources labeled with the mathematics subject category. The results from this study are summarized in table 7.13 on the next page.

**Table 7.13** Mathematics Specific Audience Level Prediction Results–Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	507	0.91	0.81	0.86	0.93	0.88	0.90
First	388	0.80	0.82	0.81	0.87	0.84	0.85
Second	285	0.78	0.83	0.81	0.81	0.87	0.84
Third	95	0.83	0.73	0.78	0.91	0.79	0.85
Fourth	241	0.75	0.86	0.80	0.82	0.90	0.85
Fifth	159	0.57	0.92	0.70	0.66	0.92	0.77
Sixth	107	0.93	0.73	0.82	0.95	0.80	0.87
Seventh	330	0.98	0.73	0.84	0.99	0.82	0.90
Eighth	212	0.59	0.81	0.68	0.69	0.91	0.78
Ninth	430	0.96	0.84	0.90	0.97	0.89	0.93
Tenth	298	0.87	0.75	0.81	0.92	0.87	0.89
Eleventh	326	0.96	0.82	0.88	0.99	0.85	0.91
Twelfth	397	0.90	0.82	0.86	0.93	0.86	0.89
Undergraduate Lower (Sampled)	136	0.60	0.84	0.70	0.66	0.86	0.75
Undergraduate Upper (Sampled)	57	0.77	0.75	0.76	0.83	0.88	0.85
Graduate (Sampled)	115	0.55	0.79	0.65	0.71	0.98	0.82
<b>Overall</b>	<b>4,083</b>	<b>0.81</b>	<b>0.81</b>	<b>0.81</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

In this study, SVMAUD outperformed the specific audience level prediction performance for the mathematics subject category, with an overall F-measure of 0.81 for the Collins-Thompson and Callan method versus 0.87 for SVMAUD. In addition, the correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.83 for the Collins-Thompson and Callan method and 0.89 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0043 level of significance.

Table 7.14 shows the results when using cosine, Naïve Bayes, and SVMAUD to predict the general audience level for mathematics resources.

**Table 7.14** Mathematics General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	1,180	0.63	0.84	0.72	0.73	0.91	0.81	0.91	0.97	0.94
Late Elementary	495	0.66	0.65	0.65	0.78	0.78	0.78	0.93	0.92	0.93
Middle School	649	0.63	0.35	0.45	0.82	0.56	0.66	0.96	0.89	0.92
High School	1,451	0.80	0.79	0.79	0.87	0.84	0.86	0.96	0.96	0.96
College (Sampled)	308	0.73	0.56	0.63	0.79	0.71	0.75	0.96	0.93	0.94
<b>Overall</b>	<b>4,083</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.80</b>	<b>0.80</b>	<b>0.80</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

The cosine classifier performance improved from an overall F-measure of 0.66 to 0.70, the Naïve Bayes classifier improved from 0.75 to 0.80, and SVMAUD performance increased from 0.92 to 0.94 when considering the single subject category versus the mathematics subject category general audience level prediction. The correlation between human-expert entered and machine-learning suggested general audience levels for the mathematics subject category classifier was found to be 0.65 for cosine, 0.77 for Naïve Bayes, and 0.94 for SVMAUD. SVMAUD outperformed both the cosine and Naïve Bayes general audience level prediction methods at the 0.0035 level of significance.

The next part of this study measures the ability of SVMAUD and the Collins-Thompson and Callan method to correctly predict the human-expert entered general audience level for the mathematics subject category, with the results shown in Table 7.15.

**Table 7.15** Mathematics General Audience Level Prediction Results–Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	1,180	0.80	0.92	0.85	0.91	0.97	0.94
Late Elementary	495	0.83	0.82	0.82	0.93	0.92	0.93
Middle School	649	0.88	0.71	0.79	0.96	0.89	0.92
High School	1,451	0.91	0.91	0.91	0.96	0.96	0.96
College (Sampled)	308	0.90	0.78	0.83	0.96	0.93	0.94
<b>Overall</b>	<b>4,083</b>	<b>0.86</b>	<b>0.86</b>	<b>0.86</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

SVMAUD again outperformed the Collins-Thompson and Callan method by predicting the general audience level with an overall F-measure of 0.94 versus 0.86 for the Collins-Thompson and Callan method; both SVMAUD and the Collins-Thompson and Callan method improved their overall F-measure performance by 0.02 over using a single subject category classifier. The correlation between the human-expert entered and the machine-learning suggested general audience levels was found to be 0.82 for the Collins-Thompson and Callan method and 0.94 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0029 level of significance.

By developing a subject-specific training and testing dataset for the four classifiers, the prediction performance improved with respect to both general and specific audience levels over the single subject category classifier. SVMAUD was found to significantly outperform the cosine, Naïve Bayes, and the Collins-Thompson and Callan methods and should have been used to predict the audience level for resources labeled with the mathematics subject category.

## 7.6 Reading and Writing Subject-Specific Classifier Performance

The reading and writing subject category covered all audience levels, ranging from the formation of letters in elementary school to research papers in college. This set of classifiers was trained and tested using documents associated with the reading and writing subject category. The specific audience level prediction results are displayed in the following table.

**Table 7.16** Reading & Writing Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	56	0.48	0.55	0.52	0.62	0.66	0.64	0.80	0.88	0.84
First	106	0.47	0.58	0.52	0.58	0.66	0.62	0.76	0.79	0.78
Second	100	0.39	0.95	0.55	0.45	0.94	0.61	0.66	0.98	0.79
Third	63	0.41	0.38	0.40	0.61	0.60	0.61	0.84	0.78	0.81
Fourth	75	0.32	0.08	0.13	0.70	0.31	0.43	0.88	0.77	0.82
Fifth	91	0.65	0.68	0.67	0.73	0.80	0.76	0.89	0.87	0.88
Sixth	15	0.87	0.87	0.87	0.93	0.93	0.93	1.00	1.00	1.00
Seventh	16	0.00	0.00	N/A	0.88	0.44	0.58	0.88	0.44	0.58
Eighth	28	0.70	0.82	0.75	0.77	0.86	0.81	0.84	0.93	0.88
Ninth	63	0.71	0.08	0.14	0.85	0.17	0.29	0.97	0.57	0.72
Tenth	57	1.00	0.26	0.42	1.00	0.44	0.61	1.00	0.84	0.91
Eleventh	19	0.65	0.68	0.67	0.75	0.79	0.77	0.89	0.89	0.89
Twelfth	47	0.97	0.68	0.80	0.97	0.81	0.88	1.00	0.94	0.97
Undergraduate Lower (Sampled)	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Undergraduate Upper (Sampled)	1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Graduate (Sampled)	6	0.63	0.83	0.71	0.63	0.83	0.71	0.86	1.00	0.92
<b>Overall</b>	<b>744</b>	<b>0.52</b>	<b>0.52</b>	<b>0.52</b>	<b>0.64</b>	<b>0.64</b>	<b>0.64</b>	<b>0.83</b>	<b>0.83</b>	<b>0.83</b>

\*\* N/A = Not Available; the F-Measure calculation results in a division by zero.

As measured by the overall F-measure, the cosine classifier performance decreased from 0.61 in the single audience level classifier to 0.52 in the reading and

writing classifier, the Naïve Bayes classifier performance decreased from 0.68 to 0.64, and SVMAUD performance decreased from 0.86 to 0.83. The correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.57 for cosine, 0.68 for Naïve Bayes, and 0.85 for SVMAUD. SVMAUD outperformed both Naïve Bayes and cosine at the 0.0051 level of significance. Since first grade and second grade resources generally covered the same topics depending on the local school district, all classifiers performed poorly with respect to these audience levels. The uneven distribution of resources across all subject categories, with reading and writing resources generally covering elementary school grades and few resources in the college audience level also reduced performance over classifiers that used a more even distribution of resources across all audience levels.

The next part of this study considered the abilities of the Collins-Thompson and Callan method versus SVMAUD when predicting the specific audience level for resources in the reading and writing subject category. The results from this study are displayed in table 7.17 on the next page. SVMAUD again outperformed the Collins-Thompson and Callan method by correctly predicting the specific human-entered audience level with an overall F-measure of 0.83 versus 0.76 for the Collins-Thompson and Callan method. The correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.76 for the Collins-Thompson and Callan method and 0.85 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0746 level of significance.



**Table 7.17** Reading & Writing Specific Audience Level Results-Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	56	0.75	0.86	0.80	0.80	0.88	0.84
First	106	0.72	0.78	0.75	0.76	0.79	0.78
Second	100	0.57	0.96	0.72	0.66	0.98	0.79
Third	63	0.73	0.76	0.74	0.84	0.78	0.81
Fourth	75	0.90	0.59	0.71	0.88	0.77	0.82
Fifth	91	0.79	0.86	0.82	0.89	0.87	0.88
Sixth	15	0.94	1.00	0.97	1.00	1.00	1.00
Seventh	16	0.90	0.56	0.69	0.88	0.44	0.58
Eighth	28	0.80	0.86	0.83	0.84	0.93	0.88
Ninth	63	0.92	0.38	0.54	0.97	0.57	0.72
Tenth	57	1.00	0.54	0.70	1.00	0.84	0.91
Eleventh	19	0.75	0.79	0.77	0.89	0.89	0.89
Twelfth	47	0.98	0.87	0.92	1.00	0.94	0.97
Undergraduate Lower (Sampled)	1	1.00	1.00	1.00	1.00	1.00	1.00
Undergraduate Upper (Sampled)	1	1.00	1.00	1.00	1.00	1.00	1.00
Graduate (Sampled)	6	0.83	0.83	0.83	0.86	1.00	0.92
<b>Overall</b>	<b>744</b>	<b>0.76</b>	<b>0.76</b>	<b>0.76</b>	<b>0.83</b>	<b>0.83</b>	<b>0.83</b>

Since the digital library collection mainly provided STEM resources, the home school collection provided the majority of documents. However, since the home school collection mainly covered grades K-12, few resources were placed into the college audience levels, leading to high F-measures since the same documents were used for training and testing; one document could not have been divided into five different folds. SVMAUD was found to significantly outperform cosine, Naïve Bayes, and the Collins-Thompson and Callan method with respect to specific audience level prediction in the reading and writing subject category.

The next part of this study considered the abilities of the four audience level prediction methods to correctly predict the general audience level for resources labeled

with the reading and writing subject category. The following table presents the prediction performance for cosine, Naïve Bayes, and SVMAUD.

**Table 7.18** Reading & Writing General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	262	0.61	0.87	0.72	0.71	0.92	0.80	0.88	0.98	0.93
Late Elementary	229	0.75	0.36	0.49	0.88	0.55	0.67	0.98	0.83	0.90
Middle School	59	0.25	0.31	0.28	0.46	0.56	0.50	0.81	0.95	0.88
High School	186	0.68	0.65	0.66	0.81	0.78	0.80	0.96	0.94	0.95
College (Sampled)	8	0.55	0.75	0.63	0.67	0.75	0.71	0.88	0.88	0.88
<b>Overall</b>	<b>744</b>	<b>0.61</b>	<b>0.61</b>	<b>0.61</b>	<b>0.74</b>	<b>0.74</b>	<b>0.74</b>	<b>0.92</b>	<b>0.92</b>	<b>0.92</b>

Similar to the specific audience level prediction results, the performance of the cosine and Naïve Bayes classifiers decreased over the single classifier representing all subject categories. In this part of the study, as measured by the overall F-measure, the cosine classifier performance decreased from 0.66 to 0.61, the Naïve Bayes classifier performance decreased from 0.75 to 0.74, and SVMAUD performance remained the same with an overall F-measure of 0.92. The correlation between the human-expert entered and the machine-learning predicted audience levels was found to be 0.57 for cosine, 0.71 for Naïve Bayes, and 0.91 for SVMAUD. In addition, SVMAUD outperformed both cosine and Naïve Bayes at the 0.0076 level of significance.

The next part of this study considered the abilities of the Collins-Thompson and Callan method and SVMAUD to correctly predict the human-entered general audience level for the reading and writing subject category. The results from this study are shown in Table 7.19 on the next page.

**Table 7.19** Reading & Writing General Audience Level Prediction – Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	262	0.79	0.95	0.86	0.88	0.98	0.93
Late Elementary	229	0.94	0.73	0.83	0.98	0.83	0.90
Middle School	59	0.68	0.76	0.72	0.81	0.95	0.88
High School	186	0.88	0.85	0.87	0.96	0.94	0.95
College (Sampled)	8	1.00	0.75	0.86	0.88	0.88	0.88
<b>Overall</b>	<b>744</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>	<b>0.92</b>	<b>0.92</b>	<b>0.92</b>

SVMAUD again outperformed the prediction performance of the Collins-Thompson and Callan method, with an overall F-measure of 0.92 versus 0.84 for the Collins-Thompson and Callan method; both SVMAUD and the Collins-Thompson and Callan method experienced the same prediction performance between the single and the reading and writing subject category classifiers at 0.92 and 0.84, respectively. The correlation between human-expert entered and machine-learning suggested general audience levels was found to be 0.81 for the Collins-Thompson and Callan method and 0.91 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0221 level of significance.

Due to the extremely uneven distribution of resources across all audience levels, with most resources labeled with elementary school level and few resources in the college level, the prediction performance had not increased over the baseline single subject category classifier. In addition, the high number of resources in the first and second grades, where many reading and writing topics overlap, such as learning to write letters and numbers, led to a decreased ability by the different methods to discriminate between these two similar audience levels. Few resources were labeled with the college

audience level, since both NSDL and home school resources generally covered grades K-12, contributing to poor performance by all classifiers.

Even though the overall prediction performance across all audience levels decreased when compared to other subject categories, SVMAUD again outperformed all other methods under evaluation when predicting the human-expert entered audience level for the reading and writing subject category. Therefore, SVMAUD should have been used to predict the general and / or specific audience level for all resources in the reading and writing subject category that were already labeled with the audience level.

### **7.7 Science Subject-Specific Classifier Performance**

The science subject category spanned all audience levels from simple science experiments taught in elementary school to astronomy and physics taught in college courses. The four classifiers were trained and tested using resources that were labeled with the science subject category. Similar to the other studies, the first part of this study considered the specific audience level prediction performance while the second part considered the general audience level prediction performance.

This part of the study compared the performance of cosine, Naïve Bayes, and SVMAUD when predicting the human-expert entered audience level for resources in the science subject category. The following table displays the specific audience level prediction results for the science subject-specific classifier.

**Table 7.20** Science Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	225	0.81	0.62	0.70	0.86	0.68	0.76	0.95	0.87	0.90
First	267	0.74	0.58	0.65	0.79	0.65	0.71	0.93	0.89	0.91
Second	201	0.67	0.70	0.68	0.71	0.73	0.72	0.89	0.91	0.90
Third	224	0.46	0.57	0.51	0.54	0.67	0.60	0.78	0.87	0.82
Fourth	229	0.75	0.57	0.65	0.80	0.67	0.73	0.94	0.84	0.89
Fifth	177	0.25	0.73	0.37	0.28	0.76	0.41	0.52	0.90	0.66
Sixth	360	0.72	0.66	0.69	0.78	0.73	0.75	0.90	0.87	0.89
Seventh	371	0.81	0.66	0.73	0.86	0.71	0.78	0.94	0.86	0.90
Eighth	356	0.71	0.72	0.71	0.75	0.74	0.74	0.90	0.91	0.90
Ninth	263	0.69	0.70	0.69	0.75	0.74	0.74	0.93	0.89	0.91
Tenth	282	0.34	0.61	0.44	0.42	0.67	0.52	0.72	0.85	0.78
Eleventh	194	0.93	0.59	0.72	0.96	0.66	0.78	0.98	0.85	0.91
Twelfth	199	0.48	0.58	0.53	0.55	0.64	0.59	0.84	0.84	0.84
Undergraduate Lower (Sampled)	481	0.85	0.60	0.70	0.89	0.66	0.76	0.97	0.86	0.91
Undergraduate Upper (Sampled)	384	0.96	0.60	0.74	0.96	0.68	0.80	0.99	0.86	0.92
Graduate (Sampled)	398	0.79	0.70	0.74	0.84	0.77	0.80	0.93	1.00	0.96
<b>Overall</b>	<b>4,611</b>	<b>0.64</b>	<b>0.64</b>	<b>0.64</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>

By training the three classifiers using only science resources to predict the audience level of other science resources, the prediction performance improved over the one-size-fits-all classifier covering all subject categories. The cosine classifier improved from 0.61 to 0.64, the Naïve Bayes classifier performance improved from 0.68 to 0.70, and SVMAUD performance improved from 0.86 to 0.88. The correlation between human-expert entered and machine-learning suggested audience levels was found to be 0.75 for cosine, 0.79 for Naïve Bayes, and 0.92 for SVMAUD. In addition, SVMAUD was found to outperform both cosine and Naïve Bayes at the 0.0001 level of significance.

The next part of this study considered the ability of the Collins-Thompson and Callan method and SVMAUD to correctly predict the human-expert entered specific audience level. The results from this study are shown in the next table.

**Table 7.21** Science Specific Audience Level Prediction Results – Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	225	0.88	0.72	0.79	0.95	0.87	0.90
First	267	0.83	0.69	0.75	0.93	0.89	0.91
Second	201	0.73	0.77	0.75	0.89	0.91	0.90
Third	224	0.60	0.71	0.65	0.78	0.87	0.82
Fourth	229	0.85	0.72	0.78	0.94	0.84	0.89
Fifth	177	0.33	0.79	0.47	0.52	0.90	0.66
Sixth	360	0.81	0.76	0.78	0.90	0.87	0.89
Seventh	371	0.88	0.77	0.82	0.94	0.86	0.90
Eighth	356	0.79	0.80	0.79	0.90	0.91	0.90
Ninth	263	0.77	0.77	0.77	0.93	0.89	0.91
Tenth	282	0.49	0.72	0.58	0.72	0.85	0.78
Eleventh	194	0.97	0.72	0.83	0.98	0.85	0.91
Twelfth	199	0.61	0.70	0.65	0.84	0.84	0.84
Undergraduate Lower (Sampled)	481	0.92	0.72	0.80	0.97	0.86	0.91
Undergraduate Upper (Sampled)	384	0.97	0.72	0.83	0.99	0.86	0.92
Graduate (Sampled)	398	0.86	0.81	0.84	0.93	1.00	0.96
<b>Overall</b>	<b>4,611</b>	<b>0.75</b>	<b>0.75</b>	<b>0.75</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>

SVMAUD again outperformed the Collins-Thompson and Callan method by correctly predicting the human-expert entered audience level with an overall F-measure of 0.88 versus 0.75 for the Collins-Thompson and Callan method. The SVMAUD performance slightly improved over the single subject category classifier, increasing from an overall F-measure of 0.86 in the single subject classifier to 0.88 in the science subject

category classifier; however, the Collins-Thompson and Callan method experienced roughly the same performance with an overall F-measure of 0.75. The correlation between human-expert entered and machine-learning suggested values was found to be 0.82 for the Collins-Thompson and Callan method versus 0.92 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0002 level of significance.

The next part of this study measured the performance of the four classifiers with respect to predicting the general audience level for science resources. The next table shows the performance comparison between cosine, Naïve Bayes, and SVMAUD.

**Table 7.22** Science General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	693	0.78	0.48	0.60	0.87	0.65	0.74	0.98	0.90	0.94
Late Elementary	630	0.68	0.37	0.48	0.85	0.57	0.68	0.96	0.89	0.93
Middle School	1,087	0.45	0.92	0.61	0.56	0.95	0.70	0.83	0.98	0.90
High School	938	0.80	0.35	0.49	0.92	0.58	0.71	0.98	0.87	0.92
College (Sampled)	1,263	0.95	0.92	0.94	0.97	0.95	0.96	0.99	0.99	0.99
<b>Overall</b>	<b>4,611</b>	<b>0.66</b>	<b>0.66</b>	<b>0.66</b>	<b>0.78</b>	<b>0.78</b>	<b>0.78</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

With respect to the science subject category and general audience level prediction performance, SVMAUD and Naïve Bayes experienced improved F-measure performance from 0.92 and 0.75 to 0.94 and 0.78, respectively. The cosine classifier performance remained the same with an overall F-measure of 0.66. The correlation between human-expert entered audience level and machine-learning suggested audience level was found to be 0.77 for cosine, 0.85 for Naïve Bayes, and 0.96 for SVMAUD. SVMAUD outperformed both cosine and Naïve Bayes at the 0.0118 level of significance.

The next table presents the results of the study comparing the performance of the Collins-Thompson and Callan method versus SVMAUD.

**Table 7.23** Science General Audience Level Prediction Results – Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	693	0.91	0.73	0.81	0.98	0.90	0.94
Late Elementary	630	0.90	0.69	0.78	0.96	0.89	0.93
Middle School	1,087	0.64	0.96	0.77	0.83	0.98	0.90
High School	938	0.95	0.72	0.82	0.98	0.87	0.92
College (Sampled)	1,263	0.98	0.96	0.97	0.99	0.99	0.99
<b>Overall</b>	<b>4,611</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

The Collins-Thompson and Callan method was able to predict the general audience level with approximately the same performance as the one-size-fits-all subject category with an overall F-measure of 0.84; on the other hand, SVMAUD slightly increased its performance, improving from an overall F-measure of 0.92 to 0.94. The correlation between human-expert entered and machine-learning suggested general audience levels was found to be 0.89 for the Collins-Thompson and Callan method and 0.96 for SVMAUD. SVMAUD was found to outperform the Collins-Thompson and Callan method at the 0.0200 level of significance.

SVMAUD significantly outperformed the audience level prediction performance for cosine, Naïve Bayes, and the Collins-Thompson and Callan method for the science subject category. SVMAUD experienced improved performance over using a single subject category covering all documents in the training dataset and, when possible, resources labeled with both the science subject category and the audience level should



have been used to train the classifiers to suggest the audience level for all other resources in the science subject category.

### **7.8 Technology and Engineering Subject-Specific Classifier Performance**

The technology and engineering subject category spanned all audience levels, ranging from computer games in elementary school to the construction of buildings and tunnels in civil engineering in college. Table 7.24 on the following page displays the results from the specific audience level prediction study with respect to the technology and engineering subject category for cosine, Naïve Bayes, and SVMAUD.

In this study, the prediction performance again improved over the single baseline classifier covering all subject categories. The cosine classifier F-measure performance increased from 0.61 to 0.65, the Naïve Bayes classifier performance increased from 0.68 to 0.71, and SVMAUD experienced increased performance with the F-measure increasing from 0.86 to 0.88. The correlation between human-expert entered and machine-learning suggested specific audience level was found to be 0.70 for cosine, 0.75 for Naïve Bayes, and 0.90 for SVMAUD. In addition, SVMAUD outperformed both cosine and Naïve Bayes at the 0.0005 level of significance.

**Table 7.24** Technology & Engineering Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	83	0.84	0.49	0.62	0.94	0.58	0.72	0.99	0.86	0.92
First	258	0.57	0.70	0.63	0.63	0.74	0.68	0.84	0.90	0.87
Second	250	0.86	0.58	0.70	0.90	0.66	0.76	0.96	0.82	0.89
Third	228	0.87	0.61	0.71	0.90	0.65	0.76	0.98	0.84	0.91
Fourth	176	0.74	0.61	0.67	0.80	0.69	0.74	0.94	0.81	0.87
Fifth	232	0.76	0.57	0.66	0.81	0.64	0.72	0.93	0.88	0.90
Sixth	185	0.95	0.58	0.72	0.97	0.66	0.78	0.99	0.86	0.92
Seventh	19	0.45	0.47	0.46	0.59	0.68	0.63	0.93	0.74	0.82
Eighth	23	0.45	0.87	0.60	0.44	0.87	0.59	0.85	0.96	0.90
Ninth	18	0.16	0.56	0.25	0.21	0.67	0.32	0.48	0.89	0.63
Tenth	18	0.16	0.39	0.23	0.19	0.39	0.25	0.60	0.83	0.70
Eleventh	11	0.18	0.64	0.28	0.23	0.64	0.34	0.42	0.91	0.57
Twelfth	14	0.19	0.57	0.28	0.24	0.71	0.36	0.48	0.86	0.62
Undergrad Lower (Sampled)	335	0.56	0.78	0.65	0.63	0.82	0.71	0.80	0.92	0.86
Undergrad Upper (Sampled)	491	0.81	0.70	0.75	0.85	0.75	0.80	0.93	0.92	0.93
Graduate (Sampled)	376	0.53	0.64	0.58	0.61	0.72	0.66	0.81	0.89	0.85
<b>Overall</b>	<b>2,717</b>	<b>0.65</b>	<b>0.65</b>	<b>0.65</b>	<b>0.71</b>	<b>0.71</b>	<b>0.71</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>

The second part of this study considered the abilities of the Collins-Thompson and Callan method and SVMAUD to correctly predict the specific audience level for resources in the technology and engineering subject category. The results from this study are shown in the following table.

**Table 7.25** Tech & Eng. Specific Audience Level Prediction Results–Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	83	0.92	0.73	0.82	0.99	0.86	0.92
First	258	0.71	0.81	0.76	0.84	0.90	0.87
Second	250	0.92	0.71	0.80	0.96	0.82	0.89
Third	228	0.92	0.74	0.82	0.98	0.84	0.91
Fourth	176	0.83	0.77	0.80	0.94	0.81	0.87
Fifth	232	0.85	0.71	0.77	0.93	0.88	0.90
Sixth	185	0.96	0.74	0.84	0.99	0.86	0.92
Seventh	19	0.72	0.68	0.70	0.93	0.74	0.82
Eighth	23	0.59	0.96	0.73	0.85	0.96	0.90
Ninth	18	0.22	0.56	0.32	0.48	0.89	0.63
Tenth	18	0.38	0.72	0.50	0.60	0.83	0.70
Eleventh	11	0.30	0.82	0.44	0.42	0.91	0.57
Twelfth	14	0.33	0.71	0.45	0.48	0.86	0.62
Undergrad Lower (Sampled)	335	0.67	0.82	0.74	0.80	0.92	0.86
Undergrad Upper (Sampled)	491	0.87	0.78	0.82	0.93	0.92	0.93
Graduate (Sampled)	376	0.65	0.77	0.71	0.81	0.89	0.85
<b>Overall</b>	<b>2,717</b>	<b>0.76</b>	<b>0.76</b>	<b>0.76</b>	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>

SVMAUD again outperformed the Collins-Thompson and Callan method by predicting the human-entered specific audience level with an overall F-measure of 0.88 versus 0.76 for the Collins-Thompson and Callan method. Both the Collins-Thompson and Callan method and SVMAUD experienced slightly higher performance over using a single subject category classifier, with the Collins-Thompson and Callan method increasing from an overall F-measure of 0.75 to 0.76, while SVMAUD increased from an overall F-measure of 0.86 to 0.88. The correlation between human-expert entered and machine-learning suggested audience levels was found to be 0.80 for the Collins-

Thompson and Callan method versus a higher correlation of 0.90 for SVMAUD. In fact, SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0074 level of significance.

The second part of this evaluation measured the prediction performance when all four classifiers were used to predict the general audience level of technology and engineering resources. The first part of this study compared the abilities of cosine, Naïve Bayes, and SVMAUD to correctly predict the human-expert entered general audience level; the results from this study are shown in the next table

**Table 7.26** Technology & Engineering General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	591	0.66	0.72	0.69	0.78	0.81	0.79	0.92	0.95	0.94
Late Elementary	636	0.69	0.57	0.62	0.81	0.70	0.75	0.95	0.93	0.94
Middle School	227	0.41	0.46	0.43	0.57	0.68	0.62	0.90	0.87	0.88
High School	61	0.40	0.46	0.43	0.54	0.57	0.56	0.92	0.89	0.90
College (Sampled)	1,202	0.95	0.96	0.96	0.97	0.97	0.97	0.99	0.99	0.99
<b>Overall</b>	<b>2,717</b>	<b>0.76</b>	<b>0.76</b>	<b>0.76</b>	<b>0.84</b>	<b>0.84</b>	<b>0.84</b>	<b>0.96</b>	<b>0.96</b>	<b>0.96</b>

The subject-specific technology and engineering classifier performance again improved over the classifier where all subjects have been grouped together. The cosine classifier F-measure performance increased from 0.66 to 0.76, the Naïve Bayes classifier F-measure increased from 0.75 to 0.84, and SVMAUD performance increased from 0.92 to 0.96. The correlation between human-expert entered and machine-learning suggested

general audience levels was found to be 0.89 for cosine, 0.93 for Naïve Bayes, and 0.98 for SVMAUD. SVMAUD outperformed both cosine and Naïve Bayes at the 0.0272 level of significance.

This part of the study considered the ability of the Collins-Thompson and Callan method and SVMAUD in their abilities to correctly predict the human-expert entered audience level. The results from this study are shown in the next table.

**Table 7.27** Tech & Eng. General Audience Level Prediction Results-Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	591	0.77	0.82	0.79	0.92	0.95	0.94
Late Elementary	636	0.82	0.72	0.77	0.95	0.93	0.94
Middle School	227	0.62	0.67	0.64	0.90	0.87	0.88
High School	61	0.59	0.66	0.62	0.92	0.89	0.90
College (Sampled)	1,202	0.97	0.98	0.97	0.99	0.99	0.99
<b>Overall</b>	<b>2,717</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.96</b>	<b>0.96</b>	<b>0.96</b>

In this part of the study, SVMAUD again outperformed the Collins-Thompson and Callan method by correctly predicting the human-expert entered audience level with an overall F-measure of 0.96 versus 0.85 for the Collins-Thompson and Callan method. The correlation between human-expert entered and machine-learning suggested values was found to be 0.93 for the Collins-Thompson and Callan method and 0.98 for SVMAUD. SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0261 level of significance.

In all four evaluations conducted during this study, SVMAUD significantly outperformed the audience level prediction performance of cosine, Naïve Bayes, and the

Collins-Thompson and Callan method. Since technology and engineering tended to contain more specific terms than reading or writing that could have been taught at all grade levels, its performance improved substantially over using a single classifier for all subject categories.

### **7.9 Overall Subject-Specific Classifier Performance**

When training and testing a classifier using a set of documents that belonged to the same general subject category, the performance, as measured by calculating the F-measure across all audience levels and correlation between human-expert and machine-learning suggested values, generally increased over developing one audience level prediction program, SVMAUD, for all subject categories. Since a higher proportion of unique terms were available in the training dataset, SVMAUD and the other classifiers were better able to make fine-grained distinctions between adjacent audience levels. SVMAUD significantly outperformed all other classifiers under evaluation at the 0.0272 level of significance for general audience level prediction for technology and engineering resources and the 0.0673 level of significance for specific audience level prediction for health sciences resources; the significance level at which SVMAUD outperformed all other classifiers for both general and specific audience level prediction was found to be higher for all other subject-specific classifiers.

After each of the six subject-specific classifiers were used to suggest the audience level for other resources discussing the same subject, the predicted audience level for each resource was compared with the human-expert suggested audience level to measure the overall performance across all subject categories. This study sought to quantify the

performance improvement of using a set of six subject-specific classification methods over using one classifier to predict the audience level for all resources in a collection.

The following table displays the prediction performance of the classifier across all specific audience levels for cosine, Naïve Bayes, and SVMAUD.

**Table 7.28** Overall Specific Audience Level Prediction Results

Specific Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Kindergarten	985	0.75	0.63	0.69	0.82	0.70	0.76	0.92	0.87	0.89
First	1,155	0.59	0.62	0.61	0.67	0.69	0.68	0.86	0.86	0.86
Second	926	0.59	0.70	0.64	0.67	0.75	0.71	0.84	0.89	0.86
Third	727	0.62	0.59	0.60	0.70	0.67	0.68	0.88	0.85	0.86
Fourth	927	0.63	0.60	0.62	0.70	0.70	0.70	0.88	0.86	0.87
Fifth	754	0.36	0.69	0.48	0.44	0.76	0.55	0.68	0.89	0.77
Sixth	824	0.79	0.61	0.69	0.84	0.69	0.75	0.94	0.86	0.90
Seventh	833	0.86	0.58	0.69	0.90	0.67	0.76	0.97	0.83	0.89
Eighth	796	0.56	0.72	0.63	0.62	0.76	0.68	0.83	0.91	0.87
Ninth	903	0.75	0.60	0.67	0.80	0.66	0.73	0.94	0.85	0.89
Tenth	822	0.47	0.53	0.50	0.56	0.63	0.59	0.83	0.84	0.84
Eleventh	697	0.81	0.60	0.69	0.88	0.67	0.76	0.95	0.85	0.90
Twelfth	778	0.67	0.61	0.64	0.72	0.67	0.70	0.90	0.87	0.89
Undergraduate Lower (Sampled)	1,050	0.63	0.67	0.65	0.70	0.72	0.71	0.86	0.88	0.87
Undergraduate Upper (Sampled)	1,050	0.85	0.65	0.74	0.88	0.72	0.79	0.95	0.89	0.92
Graduate (Sampled)	1,050	0.61	0.65	0.63	0.68	0.73	0.70	0.86	0.96	0.91
<b>Overall</b>	<b>14,277</b>	<b>0.63</b>	<b>0.63</b>	<b>0.63</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The specific audience level prediction performance by using six subject-specific classifiers improved over using a single classifier for all subject categories. The cosine classifier F-measure improved from 0.61 to 0.63, while the Naïve Bayes classifier F-measure had improved from 0.68 to 0.70. However, SVMAUD again experienced the

highest performance, with an overall F-measure of 0.87, an increase of 0.01 over using one classifier for all subject categories. The correlation between human-expert entered and machine-learning suggested specific audience levels was found to be 0.72 for cosine, 0.77 for Naïve Bayes, and 0.91 for SVMAUD. SMVAUD outperformed both cosine and Naïve Bayes at the 0.0001 level of significance.

The next part of this study considered the ability of the Collins-Thompson and Callan method versus SVMAUD in correctly predicting the human-expert entered specific audience level. The results from this part of the study are shown in table 7.29 on the next page. The Collins-Thompson and Callan method correctly predicted the human-expert entered specific audience level with an overall F-measure of 0.77, an improvement of 0.02 over the one-size-fits-all single subject category classifier, versus the higher F-measure of 0.87 for SVMAUD. The correlation between human-expert entered and machine-learning suggested audience levels was found to be 0.82 for the Collins-Thompson and Callan method versus 0.91 for SVMAUD. SVMAUD was found to outperform the prediction performance of the Collins-Thompson and Callan method at the 0.0001 level of significance.



**Table 7.29** Overall Specific Audience Level Prediction Results – Thompson&Callan

Specific Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Kindergarten	985	0.87	0.79	0.83	0.92	0.87	0.89
First	1,155	0.75	0.77	0.76	0.86	0.86	0.86
Second	926	0.76	0.80	0.78	0.84	0.89	0.86
Third	727	0.75	0.75	0.75	0.88	0.85	0.86
Fourth	927	0.79	0.78	0.79	0.88	0.86	0.87
Fifth	754	0.52	0.81	0.63	0.68	0.89	0.77
Sixth	824	0.87	0.74	0.80	0.94	0.86	0.90
Seventh	833	0.92	0.74	0.82	0.97	0.83	0.89
Eighth	796	0.72	0.81	0.77	0.83	0.91	0.87
Ninth	903	0.85	0.76	0.80	0.94	0.85	0.89
Tenth	822	0.65	0.72	0.68	0.83	0.84	0.84
Eleventh	697	0.90	0.77	0.83	0.95	0.85	0.90
Twelfth	778	0.80	0.77	0.78	0.90	0.87	0.89
Undergraduate Lower (Sampled)	1,050	0.77	0.77	0.77	0.86	0.88	0.87
Undergraduate Upper (Sampled)	1,050	0.91	0.76	0.83	0.95	0.89	0.92
Graduate (Sampled)	1,050	0.73	0.79	0.76	0.86	0.96	0.91
<b>Overall</b>	<b>14,277</b>	<b>0.77</b>	<b>0.77</b>	<b>0.77</b>	<b>0.87</b>	<b>0.87</b>	<b>0.87</b>

The second half of this study considered the general audience level prediction performance of the four classifiers. The following table shows the overall performance measurements of cosine, Naïve Bayes, and SVMAUD with respect to general audience levels.

**Table 7.30** Overall General Audience Level Prediction Results

General Audience Level	Docs	Cosine			Naïve Bayes			SVMAUD		
		P	R	F	P	R	F	P	R	F
Early Elementary	3,066	0.64	0.71	0.68	0.75	0.81	0.78	0.92	0.95	0.94
Late Elementary	2,408	0.68	0.53	0.60	0.81	0.68	0.74	0.95	0.91	0.93
Middle School	2,453	0.49	0.67	0.57	0.62	0.78	0.69	0.88	0.94	0.91
High School	3,200	0.76	0.63	0.69	0.85	0.74	0.79	0.96	0.93	0.94
College (Sampled)	3,150	0.93	0.89	0.91	0.95	0.93	0.94	0.99	0.98	0.98
<b>Overall</b>	<b>14,277</b>	<b>0.70</b>	<b>0.70</b>	<b>0.70</b>	<b>0.79</b>	<b>0.79</b>	<b>0.79</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

The general audience level prediction performance improved for all classifiers. The cosine classifier increased from an F-measure of 0.66 to 0.70 while the Naïve Bayes classifier improved from 0.75 to 0.79. However, once again, SVMAUD predicted the general audience level with an F-measure of 0.94, an increase from 0.92 when using one classifier to predict the audience level for all subject categories. The correlation between human-expert entered and machine-learning predicted general audience levels was found to be 0.76 for cosine, 0.84 for Naïve Bayes, and 0.96 for SVMAUD. SVMAUD outperformed both cosine and Naïve Bayes at the 0.0089 level of significance.

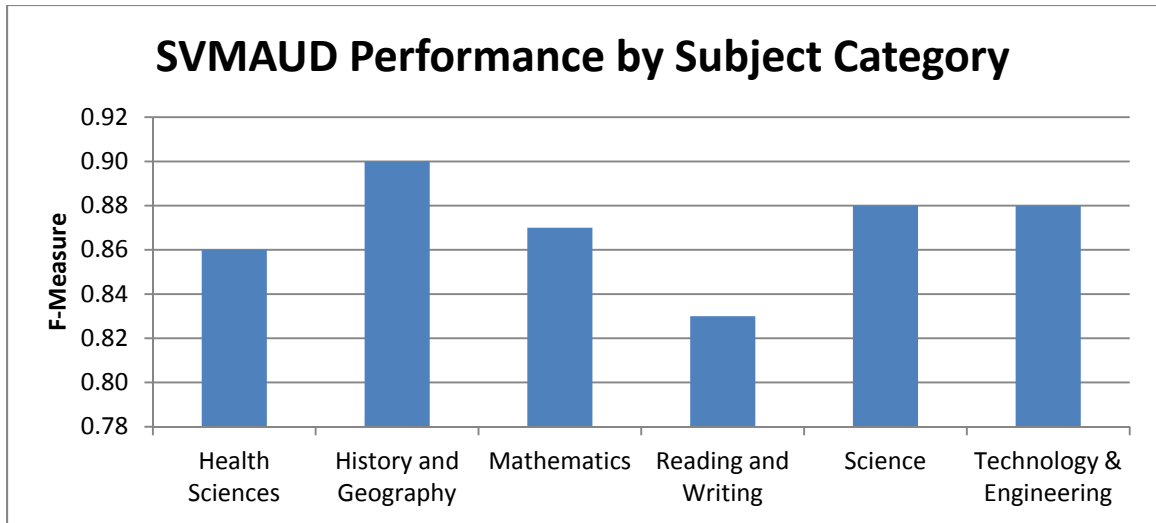
This part of the study considers the ability of SVMAUD and the Collins-Thompson and Callan method to correctly predict the human-expert entered general audience level; the results from this study are summarized in Table 7.31 on the following page.

**Table 7.31** Overall General Audience Level Prediction Results – Thompson&Callan

General Audience Level	Docs	Collins-Thompson & Callan			SVMAUD		
		P	R	F	P	R	F
Early Elementary	3,066	0.81	0.85	0.83	0.92	0.95	0.94
Late Elementary	2,408	0.85	0.76	0.80	0.95	0.91	0.93
Middle School	2,453	0.71	0.83	0.77	0.88	0.94	0.91
High School	3,200	0.90	0.83	0.86	0.96	0.93	0.94
College (Sampled)	3,150	0.96	0.94	0.95	0.99	0.98	0.98
<b>Overall</b>	<b>14,277</b>	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>0.94</b>	<b>0.94</b>	<b>0.94</b>

The Collins-Thompson and Callan method improved slightly over using a single classifier for all subject categories, increasing from 0.84 for the one-size-fits-all classifier to 0.85 for a subject-specific classifier. The SVMAUD performance slightly increased as well over the single classifier representing all subject categories, from 0.92 to 0.94. The correlation between human-expert entered and machine-learning suggested general audience level was found to be 0.88 for the Collins-Thompson and Callan method and 0.96 for SVMAUD. SVMAUD outperformed the Collins-Thompson and Callan method at the 0.0162 level of significance.

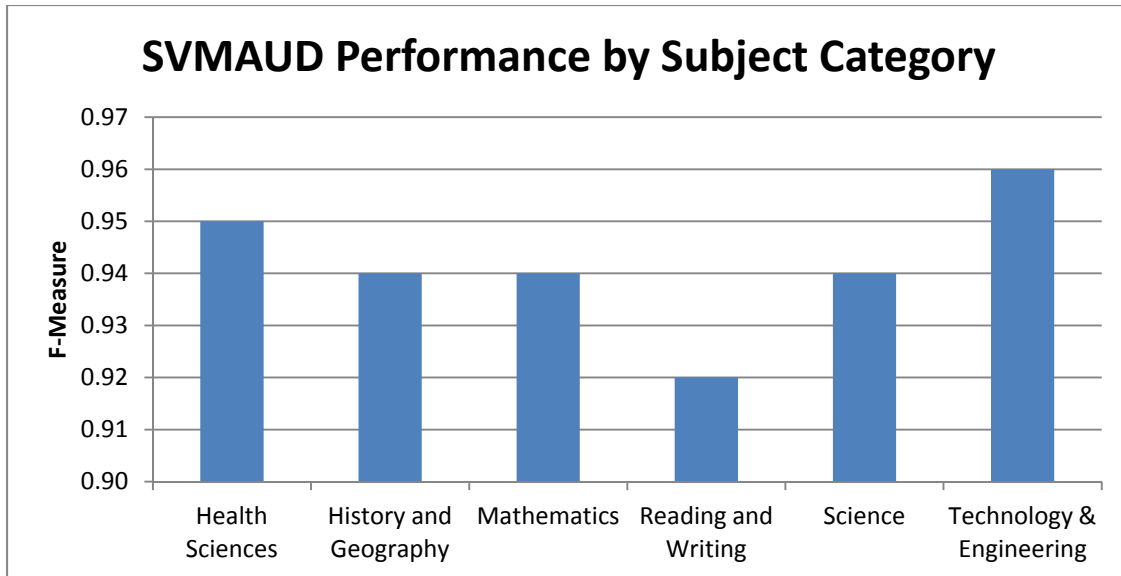
Figure 7.6 on the next page compares the specific audience level prediction performance by SVMAUD among all six subject categories.



**Figure 7.6** SVMAUD specific audience level performance by subject category.

The history and geography subject category experienced the highest performance, with an F-measure of 0.90, while the reading and writing subject category experienced the lowest performance with an F-measure of 0.83. Since the reading and writing subject category could have included other topics, such as writing a research paper on the planets in the solar system, the prediction performance was found to be lowest. History and geography tended to be more specialized, discussing different time periods and locations, rather than including topics from other categories, leading to higher performance.

Figure 7.7 on the next page presents the SVMAUD general audience level prediction performance by subject category.



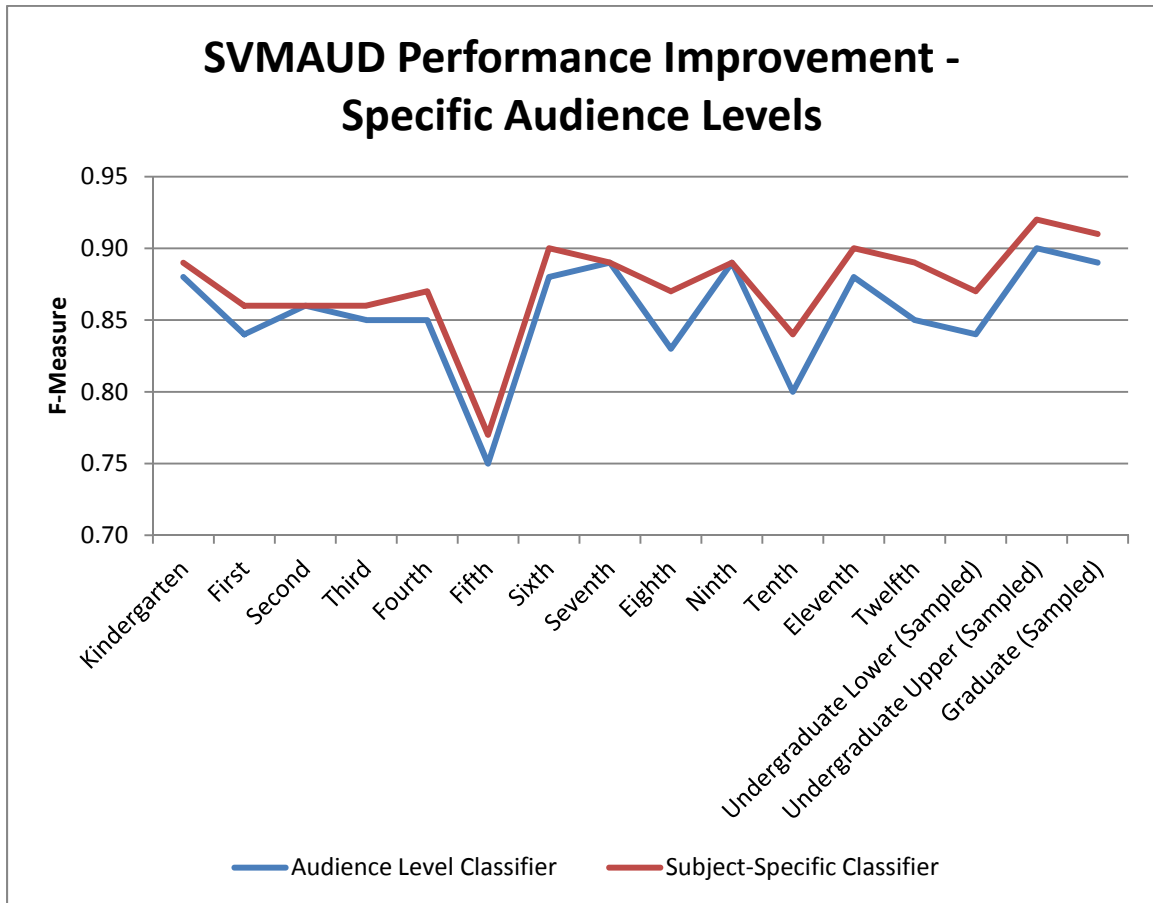
**Figure 7.7** SVMAUD general audience level performance by subject category.

In this study, SVMAUD experienced the highest prediction performance in the technology and engineering subject category by predicting the general human-expert assigned audience level with the highest performance, with an overall F-measure around 0.96. Technology and engineering tended to be another more specialized area like history and geography with little overlap between different subject categories. The reading and writing subject category again experienced the lowest performance with an overall F-measure of 0.92, since reading and writing would have included a high proportion of terms from other subject categories.

### 7.10 SVMAUD Subject-Specific Classifier Improvement

The next part of this analysis compared the performance of the one-size-fits-all classifier against the performance of the subject-specific classifier for both specific and general audience levels. In both figures, the red line shows the performance of the subject-

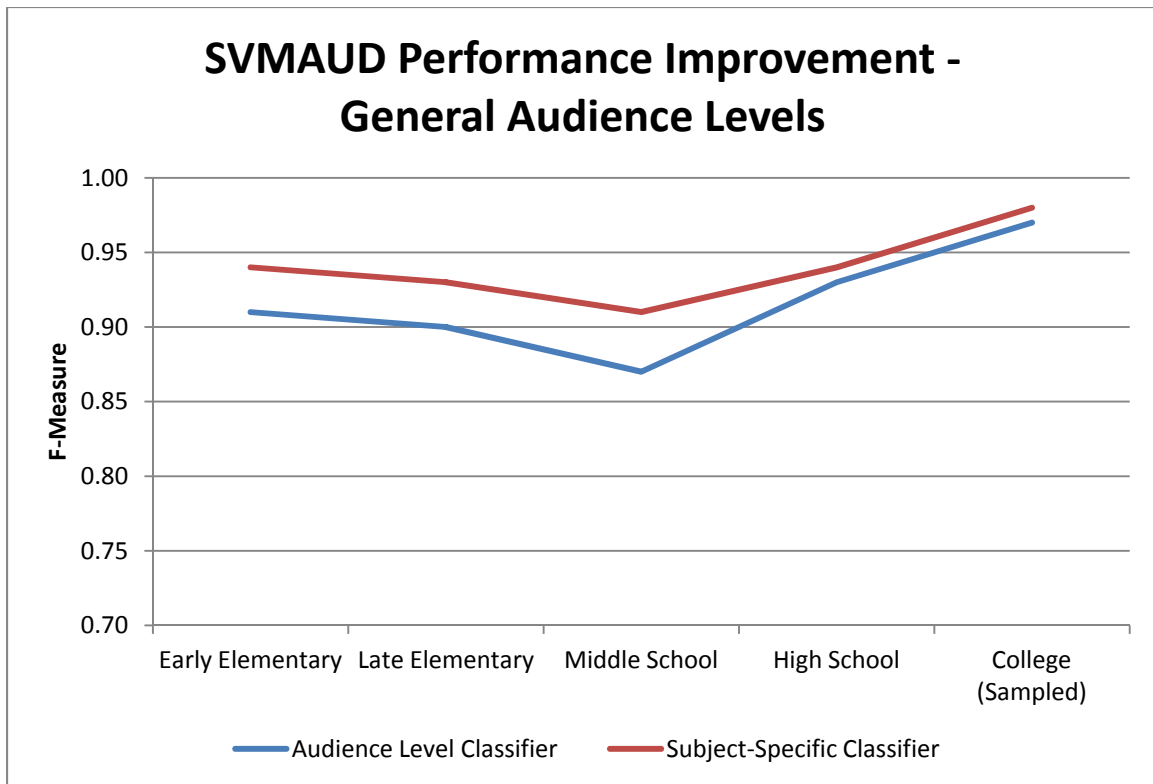
specific classifier while the blue line shows the performance of the general audience level classifier.



**Figure 7.8** SVMAUD specific audience level performance improvement.

With respect to specific audience levels, the performance increase was highest at the undergraduate lower division audience level while the performance increase was lowest at the ninth grade audience level. However, the SVMAUD subject-specific classifier performance improved over the entire spectrum of specific audience levels.

The next figure shows the SVMAUD performance improvement by general audience levels.



**Figure 7.9** SVMAUD general audience level performance improvement.

In this study, the performance improvement of using a subject-specific classifier over a general classifier with training and testing over the entire collection remained fairly constant across all general audience levels, with a lower performance improvement at the higher audience levels. By using a subject-specific classifier to predict the audience level for both home school and digital library resources, the performance for all four classifiers improved over using one classifier for all subject categories. With respect to SVMAUD, the F-measure performance increased by 0.02 for general audience levels and 0.01 for specific audience levels.

### 7.11 Machine Learning Using Subject-Specific Classifiers Discussion

This study sought to improve the already high prediction performance experienced by SVMAUD for both general and specific audience levels by developing a series of subject-specific audience level classifiers. While a single individual subject category, namely the reading and writing category, experienced decreased performance over using a single training and testing dataset for the entire collection, the overall performance across all other subject categories increased.

Some of the subject categories, namely, the Science, Technology and Engineering, and Mathematics category, contained a much higher number of documents spread across all audience levels since the digital library catalogers focused on these areas. As the digital library collection covered limited subject categories, resources from home school collections were used to augment the STEM collection to include additional subject categories commonly found in K-12 education. However, these additional collections provided a much smaller number of resources than found in the digital library collection, resulting in fewer resources available for training and testing in non-STEM subjects. If the documents available for each subject category could have been more evenly spread among all subject categories so that each subject category contained approximately the same number of documents, SVMAUD performance should have improved over the current findings.

Most states in the United States of America, in general, provided curriculum standards to educators that described the topics to be covered in each grade level. While these curriculum standards could have varied from one state to another, most of the topics taught in each grade level should have been similar. The human experts that cataloged



resources in the NSDL collection would have generally followed the national teaching standard. However, the home school collection tasked a different set of experts to catalog resources and those experts could have followed a different teaching standard. If all resources in both the home school and digital library collections were cataloged by the same group of human experts or followed the same teaching standards, then the subject-based audience level prediction performance should have experienced greater improvement than found in this study.

Even though six different subject-specific SVMAUD classifiers were developed to predict the general and specific audience levels for digital library and home school resources, the performance increase was fairly small, with an overall F-measure increase of 0.02 for general audience levels and an increase of 0.01 for specific audience levels. SVMAUD was able to outperform the three other methods, namely cosine, Naïve Bayes, and the Collins-Thompson and Callan method, under evaluation at the 0.0001 level of significance with respect to specific audience level prediction and the 0.0162 level of significance for general audience level prediction. SVMAUD would have been more useful as a one-size-fits-all classifier, where a single classifier predicted the audience level for all resources held by a collection. Since not all documents were cataloged with subject category metadata that followed the same coding scheme, the most appropriate subject category classifier could not have been selected for all unlabeled resources in the collection.

### **7.12 Machine Learning Using Subject-Specific Classifiers Conclusion**

This study sought to improve the performance of SVMAUD by developing a classifier that could have predicted the general and specific audience levels for a single subject category. The technology and engineering subject category experienced the highest performance when predicting the general audience level, with an F-measure of 0.96; on the other hand, the reading and writing subject category experienced the lowest performance when predicting the general audience level, with an F-measure of 0.92. When predicting the specific audience level for resources in a single subject category, the reading and writing category again experienced the lowest performance with an F-measure of 0.83, while the highest performance was found to be in the history and geography subject category, with an F-measure of 0.90. SVMAUD was found to outperform the cosine, Naïve Bayes, and the Collins-Thompson and Callan methods under evaluation at the 0.0001 level of significance for specific audience levels and the 0.0162 level of significance for general audience levels. If the subject category was cataloged with each resource in the collection, the benefits of developing a series of subject-specific classifiers would have outweighed the initial upfront cost to label an unlabeled resource with a single subject category.

## **CHAPTER 8**

### **CONCLUSION**

In order to sell books and newspapers, authors of written works needed a method to verify that their chosen vocabulary is appropriate for their readers. Similarly, librarians required an effective way to identify the audience level for all resources to match users with resources that both challenged and informed readers. This dissertation proposed an SVM-based audience level prediction program, called SVMAUD, which identified the audience level for all resources held in a collection by asking a human expert to identify a small number of training samples appropriate for each audience level. A number of different methods to improve the performance of SVMAUD and the other machine learning methods when predicting the audience level of digital library resources were also presented. Since the NSDL collection mainly covered STEM topics, an additional collection containing home school resources was used to augment this collection to cover a wider range of subject categories. This chapter summarizes the results from these different studies and provides contributions, implications, and future research directions that arose during the course of these studies.

#### **8.1 Completion of Study Objectives**

This study was conceived and carried out with the notion that authors, educators, information consumers, and librarians required a computer program that could have automatically identified the audience level of written works with high performance in order to verify that the vocabulary contained in the resource was appropriate for the

audience. Authors needed an accurate and consistent method to suggest the audience level for their works to both challenge and inform readers, while information consumers needed a way to find resources that would have been appropriate for their reading abilities.

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 predicted the specific audience level with F-measures of 0.10 and 0.05, respectively. Due to the random values of the inputs as the audience level had increased, the readability formulas experienced extremely poor performance. On the other hand, cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD improved the specific audience level prediction F-measures to 0.57, 0.61, 0.68, and 0.78, respectively. Machine learning methods were found to far outperform readability formulas when predicting the human expert audience level for digital library resources.

The next part of this research sought to improve the prediction performance of the machine learning methods by holding the method constant and modifying the training and testing data. Since digital library resources mainly consisted of web pages that used HTML tags for displaying data, the term weight was adjusted based on the HTML tag in which it appeared, resulting in overall F-measure specific audience level prediction performance of 0.68 for cosine, 0.70 for Naïve Bayes, 0.75 for the Collins-Thompson & Callan method, and 0.84 for SVMAUD. Since the title and header information summarized the content on the page, the weight of the terms appearing in these HTML tags was increased over terms appearing in HTML tags lower in the hierarchy, leading to increased audience level prediction performance by all machine learning methods.

When titles, keywords, and abstracts were used for training and the full text was used for testing, the specific audience level prediction F-measures for cosine, Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD improved to 0.61, 0.68, 0.75, and 0.86, respectively. Since all terms in the training samples were appropriate for the audience level by removing titles, menu headings, footers, and other text common to all resources in the collection, the performance improved over using full text for training and testing. SVMAUD outperformed all other machine learning methods under evaluation when training using cleaned data and testing using the full text of the resource.

Since the NSDL collection mainly held STEM (Science, Technology, Engineering, and Mathematics) topics, the NSDL audience level prediction program could have only predicted the audience level for unlabeled resources discussing these subject categories. In order to train the classifier for a wider variety of subject categories, resources from home school collections, covering all topics commonly taught in grades K-12, were used to augment the NSDL resources in the training and testing collections, resulting in an overall F-measure specific audience level prediction performance of 0.63, 0.70, 0.77, and 0.87, respectively. By adjusting the training and testing datasets based on the subject categories covered by the documents, the performance improved over using a one-size-fits-all prediction approach.

## **8.2 Answers to All Research Questions**

The SVM-based audience level prediction program, called SVMAUD, was proposed to predict the general and specific audience levels of resources with missing or incompatible information. The first study measured the performance of a variety of readability

formulas and machine learning methods in their abilities to correctly predict the human-expert entered audience level. The second study sought to improve the performance of the machine learning methods by adjusting the weight of terms based on the HTML tags in which they occurred; using title, abstract, and keyword metadata to train the classifiers; and developing a series of subject-specific classifiers. Since the NSDL collection mainly held resources covering the STEM topics, resources from home school collections were used to augment the NSDL collection and cover a wider range of subject categories. The five research questions that were posed prior to these studies are answered in this section.

RQ1: Could SVMAUD be used to predict the audience level for digital library resources with performance exceeding readability formulas?

The metadata element of audience level could have been populated using SVMAUD, with a prediction performance F-measure of 0.87 for general and 0.78 for specific audience levels for digital library resources. Since readability formulas have not considered the vocabulary chosen by the author, simpler words could have contained more syllables than more complex ones, artificially increasing the audience level of the resource. In addition, digital library resources typically contained headers, footers, images, and scripts in addition to the full text displayed on the page; this additional information could have distorted the inputs to the readability formulas, in particular due to the sentence length parameter. Both the Flesch-Kincaid Reading Age and Dale-Chall Reading Ease Score performed extremely poorly with respect to specific audience level prediction for the digital library collection, with overall F-measures under 0.10. In fact, SVMAUD was found to outperform readability formulas at the 0.0001 level of

significance for specific audience levels and the 0.0004 level of significance for general audience levels. Therefore, the audience level for digital library resources could have been suggested by using SVMAUD with much higher performance than readability formulas such as the Dale-Chall Reading Ease Score and Flesch-Kincaid Reading Age.

RQ2: Which machine learning method, among cosine, SVMAUD, Naïve Bayes, and the Collins-Thompson and Callan method, would result in the highest performance when suggesting the audience level for all documents in a collection?

The content of online resources differed from traditional books that contained a large number of words placed into sentences, required for input to the readability formulas. Cosine experienced a specific audience level prediction performance of an overall F-measure of 0.57 for digital library resources. Naïve Bayes correctly predicted the specific audience level for digital library resources with an F-measure of 0.61. By using the language modeling approach described by Collins-Thompson and Callan, this specific audience level prediction performance improved to an F-measure of 0.68. However, SVMAUD outperformed all of these methods by correctly predicting the specific audience level with an overall F-measure of 0.78 for digital library resources. SVMAUD outperformed the next best performing machine learning method, the Collins-Thompson and Callan method, at the 0.0013 level of significance for specific and the 0.0931 level of significance for general audience levels. In this evaluation, SVMAUD outperformed all other machine learning methods and readability formulas under consideration and SVMAUD should have been used to automatically predict the audience

level for all resources held by these collections with missing or incompatible audience level metadata.

RQ3: Since digital library resources have been predominantly web pages, could the machine learning audience level prediction performance be improved if term weights have been adjusted according to the HTML tags in which they have appeared?

Digital library resources typically consisted of web pages that displayed content using HTML tags. For this reason, the term weight could have been adjusted based on the HTML tag in which it appeared. For example, terms that appeared in the title and heading tags summarized the content as well as providing information about the various sections of the resource. Text that appeared in the plain text or table data should have been given less weight since these terms have not been given the same level of importance by the author. By adjusting the term weights based on the HTML tags in which they appeared, the specific audience level prediction performance improved from 0.57 to 0.68 for cosine, 0.61 to 0.70 for Naïve Bayes, 0.68 to 0.75 for the Collins-Thompson and Callan method, and 0.78 to 0.84 for SVMAUD. Even though the prediction performance improved across all methods, SVMAUD outperformed the next-best performing specific audience level prediction method, the Collins-Thompson and Callan method, at the 0.0016 level of significance. The prediction performance by all machine learning methods substantially improved over weighting all terms equally independent of the tags in which they appeared. When predicting the audience level for digital library resources or web pages in general, the term weight should have been adjusted based on the HTML tags in which they appeared.



RQ4: By training the machine learning methods using metadata associated with each resource in a digital library collection, could the audience level classification performance be improved?

The machine learning methods were trained using only title, abstract, and keywords that contained vocabulary appropriate for the specific audience level, rather than the full text that contained headers, footers, and menus that were common to all resources in the collection independent of the audience level. In this study, the specific audience level prediction performance F-measure for cosine increased from 0.57 to 0.61, an increase of 0.04, when trained using the cleaned dataset. When using title, abstract, and keywords as training samples rather than full text, the Naïve Bayes method experienced a specific audience level prediction performance increase from 0.61 to 0.68, a difference of 0.07. The Collins-Thompson and Callan language modeling method also improved, from a specific audience level prediction performance F-measure of 0.68 to 0.75, an improvement of 0.07. SVMAUD experienced the largest increase of 0.08 when using cleaned training samples, by increasing the specific audience level prediction performance from 0.78 to 0.86. In fact, SVMAUD was found to outperform the next-best machine learning method, the Collins-Thompson and Callan method, at the 0.0001 level of significance for specific audience levels. Therefore, by using a cleaned training dataset, the prediction performance of machine learning methods improved over using the full text of each resource, even though fewer words were available for training for each audience level.

RQ5: Could the audience level classification performance be improved if the machine learning methods have been trained and tested using resources discussing the same subject?

Since the NSDL mainly provided resources covering STEM topics, this collection contained sparse coverage of other topics commonly taught in grades K-college, especially in the case of reading and writing. When augmenting the NSDL collection with home school resources to develop a subject-specific classifier covering a wider range of subjects, the overall F-measure specific audience level prediction performance slightly improved from 0.61 to 0.63 for cosine, 0.68 to 0.70 for Naïve Bayes, 0.75 to 0.77 for the Collins-Thompson and Callan method, and 0.86 to 0.87 for SVMAUD over using a one-size-fits-all audience level prediction program. This performance only slightly improved since a different group of human experts had entered the audience level for home school resources versus NSDL resources and could have followed a different set of teaching standards. If the same human experts were called upon to identify the specific and general audience levels for all resources in the test collection, the subject-based prediction performance should have improved significantly over the one-size-fits-all audience level prediction performance.

### **8.3 Theoretical and Practical Implications**

This section provides the theoretical and practical implications that have grown out of this work. This section first describes the theoretical implications, while the second part describes the practical implications.

### 8.3.1 Theoretical Implications

First, this study advanced the state of the art in automatic audience level identification while simultaneously reducing the effort required by humans to manually enter the audience level metadata for each resource in a collection. This part of the study identified and evaluated several different readability formulas and machine learning methods with respect to their performance when predicting the audience level for digital library resources. In particular, SVMAUD was used to automatically identify the specific audience level with an overall F-measure of 0.78 for digital library resources when trained and tested using all resources independent of subject category, outperforming cosine, Naïve Bayes, and Collins-Thompson and Callan methods that experienced F-measures of 0.57, 0.61, and 0.68, respectively. On the other hand, the readability formulas of Flesch-Kincaid Reading Age and Dale-Chall Reading Ease Score predicted the specific human-expert audience level with extremely poor performance, with overall F-measures less than 0.10 for digital library resources. By using the SVMAUD program, complete and consistent audience level metadata could have been stored with every resource in a collection, reducing the effort required by a human expert to manually enter this information. After this complete audience level information was imported into a database, it could then be used by the search system to allow additional refinement of the search queries, by matching users with resources that fit their current reading abilities.

Second, by adjusting the term weights based on the HTML tags in which those terms appeared, the prediction performance for all machine learning methods improved over assigning all terms the same weight. Since terms appearing in title and heading tags summarized the resource content and divided the resource into different sections, these

tags should have been given more weight than the terms that appeared in the plain text of the resource. The terms appearing in table data and plain text should have been assigned lower weight. By adjusting the value of  $\beta$  between 0 and 100, the weights assigned to the various HTML tags appearing in each group were adjusted. At the value of  $\beta=0$ , all terms were weighted equally independent of the tags in which they appeared; on the other hand, when  $\beta=100$ , the vast majority of the weight was assigned to the terms appearing in the title and header tags with virtually no weight for other tags. When the value of  $\beta=70$ , the prediction performance for cosine was maximized, with an overall F-measure prediction performance of 0.71 for general and 0.68 for specific audience levels. When the value of  $\beta=80$ , the prediction performance for Naïve Bayes, the Collins-Thompson and Callan method, and SVMAUD was maximized. Naïve Bayes correctly predicted the human-expert entered audience level with F-measures of 0.78 for general and 0.70 for specific audience levels. The Collins-Thompson and Callan method experienced higher performance by predicting the human-expert entered audience level with F-measures of 0.85 for general and 0.75 for specific audience levels. However, SVMAUD outperformed these three methods by correctly predicting the human-expert entered audience level with an F-measure of 0.91 for general and 0.84 for specific audience levels. By adjusting the term weights based on the HTML tags in which they appeared, the prediction performance was improved over assigning the same weight to all terms.

Third, this study sought to improve the performance of SVMAUD and other machine learning methods by reducing the amount of noise in the digital library training dataset. Since digital library web pages contained information common to all resources, including headers, footers, tables, figures, and menu items that distorted the true audience

level calculation, the use of keywords, title, and abstract unique to each resource was used to reduce the text overlap between different resources in adjacent audience levels. By using this noise reduced training dataset, SVMAUD improved the specific audience level prediction performance to an overall F-measure of 0.86 for cleaned versus 0.78 for full text training data. Cosine experienced an increase as well by using the cleaned training data, by improving the overall F-measure by 0.04 to 0.61. In addition, the performance of the Naïve Bayes classifier experienced increased F-measure performance, escalating from 0.61 to 0.68 for specific audience levels. The Collins-Thompson and Callan method also improved the specific audience level prediction performance by increasing from 0.68 to 0.75 when using the cleaned training dataset. This study also showed that SVMAUD performed well on all resource lengths, ranging from fewer than one hundred words to over 4,000 words. If title, abstract, and keyword metadata could have been available for all resources in the collection, this text could have been used to create the testing resources and further improved this performance.

The fourth part of this study developed a set of six subject-specific classifiers, where resources from one subject were used to predict the audience level for other resources discussing the same subject. Since the NSDL resources mainly discussed STEM topics, additional resources from a collection of home school resources were used to augment the NSDL collection to cover reading and writing, history and geography, and health sciences topics. When using this augmented collection for training and testing versus a one-size-fits-all prediction method covering all subject categories, the overall F-measure specific audience level prediction performance slightly improved from 0.61 to 0.63 for cosine, 0.68 to 0.70 for Naïve Bayes, 0.75 to 0.77 for the Collins-Thompson and

Callan method, and 0.86 to 0.87 for SVMAUD. By developing a set of six subject-specific audience level classification methods, the prediction performance was maximized; this set of subject-specific classifiers should have been used to predict the audience level for all unlabeled resources in the digital library collection if each resource was stored with the subject category.

### **8.3.2 Practical Implications**

SVMAUD would have reduced the effort required by librarians to label all resources in a collection with the most appropriate audience level. This study then sought to improve its performance by adjusting the term weight based on the HTML tag in which it appeared, using cleaned data to train the different machine learning methods, and developing a series of subject-specific classifiers.

First, librarians should have sought to enter audience level metadata as completely and as specifically as possible, based on the same set of teaching standards, in order to guide users to the most relevant documents in the collection. Since the NSDL did not require all collections to use the same coding scheme, some collections entered audience level as “grade 1,” others entered audience level as “first grade,” another collection entered audience level as “elementary school,” and yet another collection did not include any audience level information. By matching a user with resources that were appropriate for his or her reading ability, the user would have been able to be challenged yet also informed by the resource content. By labeling a small set of training samples appropriate for each specific audience level, SVMAUD could have been used to automatically predict the audience level for all other resources in the collection following a standard coding scheme. This coding scheme for audience level should have followed the same

set of teaching standards that describe the topics to be covered in each grade level in preschool through high school; for example, if simple addition and subtraction would have been taught to first grade students, all resources discussing this topic should have been labeled with the “first grade” audience level. In different parts of the country, different states enacted different curricula standards that could have been more specific than the national teaching standards; one set of standards should have been selected and then all librarians tasked with entering resources into the collection should have followed the same standard. By cataloging all resources with the most appropriate specific audience level that followed a standard coding scheme, the retrieval system could have used this information to reduce the number of resources that would have been needed to be browsed by the user to identify relevant resources.

Second, in addition to suggesting the appropriate audience level for all resources in a collection, SVMAUD could have been extended to other metadata elements that required a pre-defined set of categories. For example, the subject category was typically limited to a few categories consisting of STEM topics; by developing a series of subject-specific classifiers, the prediction performance was found to be highest in this study. The coverage metadata element described the time period or location where the resource was applicable, such as a resource describing the construction of pyramids would be appropriate for the country of Egypt. As another example, the metadata element of Type described the resource by stating whether the resource would have been best used as a demonstration, experiment, or informational piece; this metadata element could have been predicted by using the machine learning methods. By placing a small sample of pre-labeled documents into known categories, SVMAUD could have completed this

metadata information for all resources previously entered in the collection following a standard coding scheme.

Third, the authors of the resources held by digital library collections should have taken care to structure the page by using caption and heading tags rather than simply bolding the text and employing descriptive anchor text not only to improve the audience level prediction performance but also to increase the code readability. For example, some collections used bold font to denote the captions that appeared below pictures rather than using the caption tag. Other resources used bold or italic text and increased the font size to denote different headers on the page. Still other pages used tables to format the page, rather than using tables only to display numbers with row and column headings. As another example, some collections used the words “click here” to denote a hyperlink; the link should have used text to describe the linked page. If all pages were structured by using the proper HTML elements to denote the elements on the page, a Cascading Style Sheet could have been developed to hold the formatting information for every tag used in every resource in the collection. In this way, the HTML tag weighting scheme would have been more accurate and a text color or font face could have been changed in one place rather than requiring the HTML code in every resource to be changed individually.

Fourth, the resource content should have been separated from the menus, headings, and footers that have appeared on every page regardless of the topic covered by that page. Since the same information appeared on every page in a collection, it could have influenced the audience level prediction if a large proportion of these common terms appeared in a single audience level; the prediction of a particular audience level could have been assigned based on header and footer text, rather than the content of the



resource. If the complete full text of the resource, exclusive of terms that appeared on every page in the collection, could have been used for training and testing, the prediction performance should have improved over only using title, abstract, and keyword text for training.

Fifth, most collections used the meta keywords HTML tag to describe the subject category for the page and not included this information in the metadata database that had described the resource. This information followed an inconsistent coding scheme even within the same collection, but most terms were entered by the author of the page without access to a controlled vocabulary. Since a number of collections simply linked to resources held by other sites on the World Wide Web, the individual collection exercised little control over the content or structure of the page. Even if the page did not contain the keywords information, SVMAUD could have been used to complete the meta tag keywords by asking a human expert librarian to identify a small set of resources for each subject. As an alternative, SVMAUD could have been used to complete the subject category information for all resources in the collection; in this way, it would not matter what terms the author chose, since the terms in the collection database would have followed a standard coding scheme. Then, this information could have been used to improve the prediction performance for specific and general audience levels or even aided in the completion and consistency of other metadata elements.

In conclusion, a computer-based program was developed to aid librarians in cataloging written works, and users in their quest to find relevant resources in a collection to learn new information. While SVMAUD was tested with respect to audience level prediction performance, it could have also been used to suggest metadata values for other

elements that have used a controlled vocabulary; since a small sample of representative resources were required for training, the upfront effort required to identify these resources would have been more than balanced since the metadata values for remaining resources could have been suggested automatically. By using SVMAUD to automatically predict the audience level for an unlabeled resource in the digital library catalog entry system, users could have been better matched to resources that both challenged and informed the reader.

#### **8.4 Contributions**

Digital librarians required an automated program to automatically generate audience level metadata for all resources in the collection with missing or incompatible audience level information. On the other hand, users could have used this complete and consistent metadata information to enhance a text search only system by asking for resources targeted toward his or her individual reading ability. Several different evaluations were conducted not only to show the feasibility of employing machine learning methods to generate audience level metadata but also to improve its performance by adjusting the training and testing datasets.

The first study advanced the state of the art of audience level identification by evaluating several different machine learning methods and readability formulas in their ability to correctly predict the human-expert entered audience level for digital library resources when using the full text for training and testing. Digital library resources contained headers, footers, menus, scripts, and other attributes common to all resources in the collection independent of audience level. Similarly, images, tables, and hyperlinks

did not follow sentence conventions of written English, distorting the true audience level far upward. Since machine learning methods compared vocabulary in an unlabeled resource with a predefined vocabulary appropriate for each audience level, they experienced far higher performance with respect to general and specific audience level prediction. SVMAUD far outperformed traditional readability formulas and other machine learning methods with respect to predicting the audience level for these resources.

The second part of the study improved the audience level prediction performance by holding the machine learning methods constant and modifying the training and testing datasets. By adjusting the term weight according to the HTML tag in which it appeared, the performance for all machine learning methods increased over weighting all terms equally. When a cleaned dataset, consisting of titles, keywords, and abstracts, was used for training and the full text for testing, the specific audience level prediction F-measures increased for all machine learning methods. By developing a series of subject-specific classifiers, whereby the resources from a single subject category were used to predict the audience level for other resources in the same subject category, the specific audience level prediction performance further improved. By keeping the machine learning technique constant and modifying the training and testing data, the audience level prediction performance improved over weighting all terms equally, independent of their location in the resource.

SVMAUD was found to outperform the readability formulas and other machine learning methods under evaluation in both the digital library and home school evaluations. SVMAUD could have not only predicted the audience level for all resources

held in a collection by being trained using a small sample of human-expert labeled resources, but could have also verified that an author chosen vocabulary appropriate for his or her audience. If the resource was targeted toward a higher or lower audience level than the author desired, he or she could have used SVMAUD in conjunction with a thesaurus to replace words in the document with words that would have been more appropriate for the audience.

### **8.5 Future Work**

Even though SVMAUD, cosine, Naïve Bayes, and the Collins-Thompson and Callan methods suggested the most appropriate audience level with high F-measures, more work could have been done to further improve the performance of these methods. When adjusting term weights based on HTML tags, SVMAUD performance when predicting the specific audience level improved from 0.78 to 0.84. When removing noise by training SVMAUD using title, keywords, and abstract, SVMAUD performance improved to 0.86. Since the NSDL collection only held resources discussing STEM topics, resources from a home school collection were collected to cover additional topics commonly taught in grades K-college; however, since these resources were cataloged by a different set of human experts that could have followed different teaching standards, the overall specific audience level prediction performance only improved by 0.02 or less across all machine learning methods. While this study conducted several different experiments to measure the performance of readability formulas and machine learning methods across a number of different conditions, more work could have been done to

further improve the prediction performance of these methods by grouping resources into other categories, such as by type or coverage.

The focus of this dissertation is on automatic audience level identification for one of the metadata elements provided by the NSDL. Other elements, such as format, type, and language, could also hold important clues to better filter relevant results from the collection. After a human expert selects a representative set of resources for each possible metadata value, SVMAUD could use these training samples to automatically suggest the values for additional elements, rather than requiring the human expert to manually identify the value for each resource in the collection. By including complete and consistent metadata for all resources in the collection, retrieval algorithms could use this information to better target resources to the user beyond the ability of text search only systems.

SVMAUD could potentially be applied to other areas outside of the scope of digital libraries to determine the most appropriate audience level of documents. In particular, some newspapers and magazines seek to appeal to a particular reader base while others seek to appeal to all audience levels. This program could be used by the editor of a newspaper, or integrated into a word processing program, in order to ensure that all articles written by the staff contained the appropriate vocabulary to appeal to the correct audience. Particularly in the case of medical literature written for an audience of doctors and nurses, children and adult patients probably would not be able to understand the information presented, or will misinterpret symptoms, if the language is too advanced, leading to complications or increased hospital stays. Video game manuals need to use vocabulary appropriate for the target users who play the game; manuals for

games targeted toward younger users should employ simpler vocabulary and shorter sentences than those games targeted toward higher audience levels. By ensuring that the vocabulary is appropriate for readers of different literature sources, the readers could be informed, yet not challenged, by the resource content. SVMAUD could be applied to just about any domain to suggest audience level, as long as sufficient training samples existed to train the classifier.

SVMAUD, the three other machine learning methods, and the two readability formulas under evaluation, only consider the textual information on the page. However, HTML pages contain additional information beyond just words, ranging from images to applets and multimedia files that could also hold important clues to suggest the audience level of the resource. For some resources, particularly with respect to the Teacher's Domain collection, that hold multimedia resources viewed in an embedded media player, only the caption and title information is available to predict the audience level, while all other information in the multimedia file is ignored. College level students view detailed formulas, charts, and graphs, while an elementary school student would only learn simple addition and subtraction or read picture books; by incorporating similar image structures with audience level, the performance could be further improved beyond only using the text on the page.

After complete audience level information is generated by SVMAUD or another method, this information could be associated with each resource in a digital library collection. The digital library user could then search not only by keyword but also for resources targeted toward his or her reading ability, reducing the time and effort required to identify resources that described the needed information, and understood by the user

without frequent trips to a dictionary. If users could easily find the required information in the collection, these users would be more likely to return to that collection to find additional resources in the future.

A number of future research directions were described that could have not only improved the audience level prediction performance of SVMAUD, but could also have applied SVMAUD to other areas outside of the scope of digital libraries. In all evaluations, SVMAUD outperformed all readability formulas and other machine learning methods under evaluation. SVMAUD proved its abilities to correctly predict the human-expert entered audience levels and could have been further studied to further improve its audience level prediction performance.

## 8.6 Conclusion

This dissertation completed several objectives revolving around using different classification techniques and readability formulas to automatically suggest the human-expert assigned audience level for all resources in digital library collections with missing or incompatible audience level metadata. With respect to specific audience levels, SVMAUD was found to outperform common readability formulas as well as other machine learning methods with an overall specific audience level prediction F-measure of 0.78 for digital library resources. When the term weights were adjusted based on the HTML tags in which those terms occurred, SVMAUD correctly predicted the human-expert entered specific audience level with an overall F-measure of 0.84. When training using title, abstract, and keywords metadata elements, the SVMAUD specific audience level prediction performance improved to 0.86. When a set of six subject-specific

classifiers were developed to cover all topics commonly taught in grades K-college, this specific audience level prediction performance F-measure was found to be 0.87. By using SVMAUD to generate complete and consistent audience level metadata for resources held by digital library collections, the user could have drawn upon this additional information to reduce the time and effort required to find relevant resources in the collection that matched his or her reading ability.

This chapter discussed in great detail the answers to the five research questions, the theoretical and practical implications, and contributions of this work. By applying SVMAUD to automatically predict complete and consistent audience level metadata for all resources held in a digital library collection, the effort required by users to find relevant resources in a collection would have been reduced. In addition, if a controlled vocabulary could have been developed to represent all possible values for any other metadata element, SVMAUD could have been used to automatically assign labels from the controlled vocabulary to each resource after a human expert identifies a small set of resources appropriate for each category. Then, retrieval systems could have called upon this complete and consistent metadata to reduce the effort required by users to identify relevant resources by allowing for more than full text searches. Even though machine learning methods were more complicated and required a human expert to identify a set of samples for each class, the performance improvement more than balanced this upfront cost. SVMAUD could be used not only to predict the audience level of resources held by digital library collections but also the audience level of any written work.



## APPENDIX A

### DALE COMMON WORD LIST

Appendix A contains the 3,000 words found in the Dale Common Word List (Chall & Dale, 1995)

a	able	aboard	about	above
absent	accept	accident	account	ache
aching	acorn	acre	across	act
acts	add	address	admire	adventure
afar	afraid	after	afternoon	afterward
afterwards	again	against	age	aged
ago	agree	ah	ahead	aid
aim	air	airfield	airplane	airport
airship	airy	alarm	alike	alive
all	alley	alligator	allow	almost
alone	along	aloud	already	also
always	am	America	American	among
amount	an	and	angel	anger
angry	animal	another	answer	ant
any	anybody	anyhow	anyone	anything
anyway	anywhere	apart	apartment	ape
apiece	appear	apple	April	apron
are	aren't	arise	arithmetic	arm
armful	army	arose	around	arrange
arrive	arrived	arrow	art	artist
as	ash	ashes	aside	ask
asleep	at	ate	attack	attend
attention	August	aunt	author	auto
automobile	autumn	avenue	awake	awaken
away	awful	awfully	awhile	ax
axe	baa	babe	babies	back
background	backward	backwards	bacon	bad
badge	badly	bag	bake	baker
bakery	baking	ball	balloon	banana
band	bandage	bang	banjo	bank
banker	bar	barber	bare	barefoot

barely	bark	barn	barrel	base
baseball	basement	basket	bat	batch
bath	bathe	bathing	bathroom	bathtub
battle	battleship	bay	be	beach
bead	beam	bean	bear	beard
beast	beat	beating	beautiful	beautify
beauty	became	because	become	becoming
bed	bedbug	bedroom	bedspread	bedtime
bee	beech	beef	beefsteak	beehive
been	beer	beet	before	beg
began	beggar	begged	begin	beginning
begun	behave	behind	being	believe
bell	belong	below	belt	bench
bend	beneath	bent	berries	berry
beside	besides	best	bet	better
between	bib	bible	bicycle	bid
big	bigger	bill	billboard	bin
bind	bird	birth	birthday	biscuit
bit	bite	biting	bitter	black
blackberry	blackbird	blackboard	blackness	blacksmith
blame	blank	blanket	blast	blaze
bleed	bless	blessing	blew	blind
blindfold	blinds	block	blood	bloom
blossom	blot	blow	blue	blueberry
bluebird	blush	board	boast	boat
bob	bobwhite	bodies	body	boil
boiler	bold	bone	bonnet	boo
book	bookcase	bookkeeper	boom	boot
born	borrow	boss	both	bother
bottle	bottom	bought	bounce	bow
bowl	bow-wow	box	boxcar	boxer
boxes	boy	boyhood	bracelet	brain
brake	bran	branch	brass	brave
bread	break	breakfast	breast	breath
breathe	breeze	brick	bride	bridge
bright	brightness	bring	broad	broadcast
broke	broken	brook	broom	brother
brought	brown	brush	bubble	bucket
buckle	bud	buffalo	bug	buggy
build	building	built	bulb	bull
bullet	bum	bumblebee	bump	bun
bunch	bundle	bunny	burn	burst

bury	bus	bush	bushel	business
busy	but	butcher	butt	butter
buttercup	butterfly	buttermilk	butterscotch	button
buttonhole	buy	buzz	by	bye
cab	cabbage	cabin	cabinet	cackle
cage	cake	calendar	calf	call
caller	calling	came	camel	camp
campfire	can	canal	canary	candle
candlestick	candy	cane	cannon	cannot
canoe	can't	canyon	cap	cape
capital	captain	car	card	cardboard
care	careful	careless	carelessness	carload
carpenter	carpet	carriage	carrot	carry
cart	carve	case	cash	cashier
castle	cat	catbird	catch	catcher
caterpillar	catfish	catsup	cattle	caught
cause	cave	ceiling	cell	cellar
cent	center	cereal	certain	certainly
chain	chair	chalk	champion	chance
change	chap	charge	charm	chart
chase	chatter	cheap	cheat	check
checkers	cheek	cheer	cheese	cherry
chest	chew	chick	chicken	chief
child	childhood	children	chill	chilly
chimney	chin	china	chip	chipmunk
chocolate	choice	choose	chop	chorus
chose	chosen	christen	Christmas	church
churn	cigarette	circle	circus	citizen
city	clang	clap	class	classmate
classroom	claw	clay	clean	cleaner
clear	clerk	clever	click	cliff
climb	clip	cloak	clock	close
closet	cloth	clothes	clothing	cloud
cloudy	clover	clown	club	cluck
clump	coach	coal	coast	coat
cob	cobbler	cocoa	coconut	cocoon
cod	codfish	coffee	coffeepot	coin
cold	collar	college	color	colored
colt	column	comb	come	comfort
comic	coming	company	compare	conductor
cone	connect	coo	cook	cooked
cookie	cookies	cooking	cool	cooler

coop	copper	copy	cord	cork
corn	corner	correct	cost	cot
cottage	cotton	couch	cough	could
couldn't	count	counter	country	county
course	court	cousin	cover	cow
coward	cowardly	cowboy	cozy	crab
crack	cracker	cradle	cramps	cranberry
crank	cranky	crash	crawl	crazy
cream	creamy	creek	creep	crept
cried	cries	croak	crook	crooked
crop	cross	cross-eyed	crossing	crow
crowd	crowded	crown	cruel	crumb
crumble	crush	crust	cry	cub
cuff	cuff	cup	cup	cupboard
cupful	cure	curl	curly	curtain
curve	cushion	custard	customer	cut
cute	cutting	dab	dad	daddy
daily	dairy	daisy	dam	damage
dame	damp	dance	dancer	dancing
dandy	danger	dangerous	dare	dark
darkness	darling	darn	dart	dash
date	daughter	dawn	day	daybreak
daytime	dead	deaf	deal	dear
death	December	decide	deck	deed
deep	deer	defeat	defend	defense
delight	den	dentist	depend	deposit
describe	desert	deserve	desire	desk
destroy	devil	dew	diamond	did
didn't	die	died	dies	difference
different	dig	dim	dime	dine
ding-dong	dinner	dip	direct	direction
dirt	dirty	discover	dish	dislike
dismiss	ditch	dive	diver	divide
do	dock	doctor	does	doesn't
dog	doll	dollar	dolly	done
donkey	don't	door	doorbell	doorknob
doorstep	dope	dot	double	dough
dove	down	downstairs	downtown	dozen
drag	drain	drank	draw	draw
drawer	drawing	dream	dress	dresser
dressmaker	drew	dried	drift	drill
drink	drip	drive	driven	driver

drop	drove	drown	drowsy	drub
drum	drunk	dry	duck	due
dug	dull	dumb	dump	during
dust	dusty	duty	dwarf	dwelt
dwelt	dying	each	eager	eagle
ear	early	earn	earth	east
eastern	easy	eat	eaten	edge
egg	eh	eight	eighteen	eighth
eighty	either	elbow	elder	eldest
electric	electricity	elephant	eleven	elf
elm	else	elsewhere	empty	end
ending	enemy	engine	engineer	English
enjoy	enough	enter	envelope	equal
erase	eraser	errand	escape	eve
even	evening	ever	every	everybody
everyday	everyone	everything	everywhere	evil
exact	except	exchange	excited	exciting
excuse	exit	expect	explain	extra
eye	eyebrow	fable	face	facing
fact	factory	fail	faint	fair
fairy	faith	fake	fall	family
fan	fancy	far	faraway	fare
farm	farmer	farming	far-off	farther
fashion	fast	fasten	fat	father
fault	favor	favorite	fear	feast
feather	February	fed	feed	feel
feet	fell	fellow	felt	fence
fever	few	fib	fiddle	field
fife	fifteen	fifth	fifty	fig
fight	figure	file	fill	film
finally	find	fine	finger	finish
fire	firearm	firecracker	fireplace	fireworks
firing	first	fish	fisherman	fist
fit	fits	five	fix	flag
flake	flame	flap	flash	flashlight
flat	flea	flesh	flew	flies
flight	flip	flip-flop	float	flock
flood	floor	flop	flour	flow
flower	flowery	flutter	fly	foam
fog	foggy	fold	folks	follow
following	fond	food	fool	foolish
foot	football	footprint	for	forehead

forest	forget	forgive	forgot	forgotten
fork	form	fort	forth	fortune
forty	forward	fought	found	fountain
four	fourteen	fourth	fox	frame
free	freedom	freeze	freight	French
fresh	fret	Friday	fried	friend
friendly	friendship	frighten	frog	from
front	frost	frown	froze	fruit
fry	fudge	fuel	full	fully
fun	funny	fur	furniture	further
fuzzy	gain	gallon	gallop	game
gang	garage	garbage	garden	gas
gasoline	gate	gather	gave	gay
gear	geese	general	gentle	gentleman
gentlemen	geography	get	getting	giant
gift	gingerbread	girl	give	given
giving	glad	gladly	glance	glass
glasses	gleam	glide	glory	glove
glow	glue	go	goal	goat
gobble	God	god	godmother	goes
going	gold	golden	goldfish	golf
gone	good	good-by	goodbye	goodbye
good-bye	good-looking	goodness	goods	goody
goose	gooseberry	got	govern	government
gown	grab	gracious	grade	grain
grand	grandchild	grandchildren	granddaughter	grandfather
grandma	grandmother	grandpa	grandson	grandstand
grape	grapefruit	grapes	grass	grasshopper
grateful	grave	gravel	graveyard	gravy
gray	graze	grease	great	green
greet	grew	grind	groan	grocery
ground	group	grove	grow	guard
guess	guest	guide	gulf	gum
gun	gunpowder	guy	ha	habit
had	hadn't	hail	hair	haircut
hairpin	half	hall	halt	ham
hammer	hand	handful	handkerchief	handle
handwriting	hang	happen	happily	happiness
happy	harbor	hard	hardly	hardship
hardware	hare	hark	harm	harness
harp	harvest	has	hasn't	haste
hasten	hasty	hat	hatch	hatchet

hate	haul	have	haven't	having
hawk	hay	hayfield	haystack	he
head	headache	heal	health	healthy
heap	hear	heard	hearing	heart
heat	heater	heaven	heavy	he'd
heel	height	held	hell	he'll
hello	helmet	help	helper	helpful
hem	hen	henhouse	her	herd
here	here's	hero	hers	herself
he's	hey	hickory	hid	hidden
hide	high	highway	hill	hillside
hilltop	hilly	him	himself	hind
hint	hip	hire	his	hiss
history	hit	hitch	hive	ho
hoe	hog	hold	holder	hole
holiday	hollow	holy	home	homely
homesick	honest	honey	honeybee	honeymoon
honk	honor	hood	hoof	hook
hoop	hop	hope	hopeful	hopeless
horn	horse	horseback	horseshoe	hose
hospital	host	hot	hotel	hound
hour	house	housetop	housewife	housework
how	however	howl	hug	huge
hum	humble	hump	hundred	hung
hunger	hungry	hunk	hunt	hunter
hurrah	hurried	hurry	hurt	husband
hush	hut	hymn	I	ice
icy	I'd	idea	ideal	if
ill	I'll	I'm	important	impossible
improve	in	inch	inches	income
indeed	Indian	indoors	ink	inn
insect	inside	instant	instead	insult
intend	interested	interesting	into	invite
iron	is	island	isn't	it
its	it's	itself	I've	ivory
ivy	jacket	jacks	jail	jam
January	jar	jaw	jay	jelly
jellyfish	jerk	jig	job	jockey
join	joke	joking	jolly	journey
joy	joyful	joyous	judge	jug
juice	juicy	July	jump	June
junior	junk	just	keen	keep

kept	kettle	key	kick	kid
kill	killed	kind	kindly	kindness
king	kingdom	kiss	kitchen	kite
kitten	kitty	knee	kneel	knew
knife	knit	knives	knob	knock
knot	know	known	lace	lad
ladder	ladies	lady	laid	lake
lamb	lame	lamp	land	lane
language	lantern	lap	lard	large
lash	lass	last	late	laugh
laundry	law	lawn	lawyer	lay
lazy	lead	leader	leaf	leak
lean	leap	learn	learned	least
leather	leave	leaving	led	left
leg	lemon	lemonade	lend	length
less	lesson	let	let's	letter
letting	lettuce	level	liberty	library
lice	lick	lid	lie	life
lift	light	lightness	lightning	like
likely	liking	lily	limb	lime
limp	line	linen	lion	lip
list	listen	lit	little	live
lively	liver	lives	living	lizard
load	loaf	loan	loaves	lock
locomotive	log	lone	lonely	lonesome
long	look	lookout	loop	loose
lord	lose	loser	loss	lost
lot	loud	love	lovely	lover
low	luck	lucky	lumber	lump
lunch	lying	ma	machine	machinery
mad	made	magazine	magic	maid
mail	mailbox	mailman	major	make
making	male	mama	mamma	man
manager	mane	manger	many	map
maple	marble	march	March	mare
mark	market	marriage	married	marry
mask	mast	master	mat	match
matter	mattress	may	May	maybe
mayor	maypole	me	meadow	meal
mean	means	meant	measure	meat
medicine	meet	meeting	melt	member
men	mend	meow	merry	mess



message	met	metal	mew	mice
middle	midnight	might	mighty	mile
miler	milk	milkman	mill	million
mind	mine	miner	mint	minute
mirror	mischief	miss	Miss	misspell
mistake	misty	mitt	mitten	mix
moment	Monday	money	monkey	month
moo	moon	moonlight	moose	mop
more	morning	morrow	moss	most
mostly	mother	motor	mount	mountain
mouse	mouth	move	movie	movies
moving	mow	Mr.	Mrs.	much
mud	muddy	mug	mule	multiply
murder	music	must	my	myself
nail	name	nap	napkin	narrow
nasty	naughty	navy	near	nearby
nearly	neat	neck	necktie	need
needle	needn't	Negro	neighbor	neighborhood
neither	nerve	nest	net	never
nevermore	new	news	newspaper	next
nibble	nice	nickel	night	nightgown
nine	nineteen	ninety	no	nobody
nod	noise	noisy	none	noon
nor	north	northern	nose	not
note	nothing	notice	November	now
nowhere	number	nurse	nut	oak
oar	oatmeal	oats	obey	ocean
o'clock	October	odd	of	off
offer	office	officer	often	oh
oil	old	old-fashioned	on	once
one	onion	only	onward	open
or	orange	orchard	order	ore
organ	other	otherwise	ouch	ought
our	ours	ourselves	out	outdoors
outfit	outlaw	outline	outside	outward
oven	over	overalls	overcoat	overeat
overhead	overhear	overnight	overturn	owe
owing	owl	own	owner	ox
pa	pace	pack	package	pad
page	paid	pail	pain	painful
paint	painter	painting	pair	pal
palace	pale	pan	pancake	pane

pansy	pants	papa	paper	parade
pardon	parent	park	part	partly
partner	party	pass	passenger	past
paste	pasture	pat	patch	path
patter	pave	pavement	paw	pay
payment	pea	peace	peaceful	peach
peaches	peak	peanut	pear	pearl
peas	peck	peek	peel	peep
peg	pen	pencil	penny	people
pepper	peppermint	perfume	perhaps	person
pet	phone	piano	pick	pickle
picnic	picture	pie	piece	pig
pigeon	piggy	pile	pill	pillow
pin	pine	pineapple	pink	pint
pipe	pistol	pit	pitch	pitcher
pity	place	plain	plan	plane
plant	plate	platform	platter	play
player	playground	playhouse	playmate	plaything
pleasant	please	pleasure	plenty	plow
plug	plum	pocket	pocketbook	poem
point	poison	poke	pole	police
policeman	polish	polite	pond	ponies
pony	pool	poor	pop	popcorn
popped	porch	pork	possible	post
postage	postman	pot	potato	potatoes
pound	pour	powder	power	powerful
praise	pray	prayer	prepare	present
pretty	price	prick	prince	princess
print	prison	prize	promise	proper
protect	proud	prove	prune	public
puddle	puff	pull	pump	pumpkin
punch	punish	pup	pupil	puppy
pure	purple	purse	push	puss
pussy	pussycat	put	putting	puzzle
quack	quart	quarter	queen	queer
question	quick	quickly	quiet	quilt
quit	quite	rabbit	race	rack
radio	radish	rag	rail	railroad
railway	rain	rainbow	rainy	raise
raisin	rake	ram	ran	ranch
rang	rap	rapidly	rat	rate
rather	rattle	raw	ray	reach

read	reader	reading	ready	real
really	reap	rear	reason	rebuild
receive	recess	record	red	redbird
redbreast	refuse	reindeer	rejoice	remain
remember	remind	remove	rent	repair
repay	repeat	report	rest	return
review	reward	rib	ribbon	rice
rich	rid	riddle	ride	rider
riding	right	rim	ring	rip
ripe	rise	rising	river	road
roadside	roar	roast	rob	robber
robe	robin	rock	rocket	rocky
rode	roll	roller	roof	room
rooster	root	rope	rose	rosebud
rot	rotten	rough	round	route
row	rowboat	royal	rub	rubbed
rubber	rubbish	rug	rule	ruler
rumble	run	rung	runner	running
rush	rust	rusty	rye	sack
sad	saddle	sadness	safe	safety
said	sail	sailboat	sailor	saint
salad	sale	salt	same	sand
sandwich	sandy	sang	sank	sap
sash	sat	satin	satisfactory	Saturday
sausage	savage	save	savings	saw
say	scab	scales	scare	scarf
school	schoolboy	schoolhouse	schoolmaster	schoolroom
scorch	score	scrap	scrape	scratch
scream	screen	screw	scrub	sea
seal	seam	search	season	seat
second	secret	see	seed	seeing
seek	seem	seen	seesaw	select
self	selfish	sell	send	sense
sent	sentence	separate	September	servant
serve	service	set	setting	settle
settlement	seven	seventeen	seventh	seventy
several	sew	shade	shadow	shady
shake	shaker	shaking	shall	shame
shan't	shape	share	sharp	shave
she	shear	shears	shed	she'd
sheep	sheet	shelf	shell	she'll
shepherd	she's	shine	shining	shiny

ship	shirt	shock	shoe	shoemaker
shone	shook	shoot	shop	shopping
shore	short	shot	should	shoulder
shouldn't	shout	shovel	show	shower
shut	shy	sick	sickness	side
sidewalk	sideways	sigh	sight	sign
silence	silent	silk	sill	silly
silver	simple	sin	since	sing
singer	single	sink	sip	sir
sis	sissy	sister	sit	sitting
six	sixteen	sixth	sixty	size
skate	skater	ski	skin	skip
skirt	sky	slam	slap	slate
slave	sled	sleep	sleepy	sleeve
sleigh	slept	slice	slid	slide
sling	slip	slipped	slipper	slippery
slit	slow	slowly	sly	smack
small	smart	smell	smile	smoke
smooth	snail	snake	snap	snapping
sneeze	snow	snowball	snowflake	snowy
snuff	snug	so	soak	soap
sob	socks	sod	soda	sofa
soft	soil	sold	soldier	sole
some	somebody	somehow	someone	something
sometime	sometimes	somewhere	son	song
soon	sore	sorrow	sorry	sort
soul	sound	soup	sour	south
southern	space	spade	spank	sparrow
speak	speaker	spear	speech	speed
spell	spelling	spend	spent	spider
spike	spill	spin	spinach	spirit
spit	splash	spoil	spoke	spook
spoon	sport	spot	spread	spring
springtime	sprinkle	square	squash	squeak
squeeze	squirrel	stable	stack	stage
stair	stall	stamp	stand	star
stare	start	starve	state	States
station	stay	steak	steal	steam
steamboat	steamer	steel	steep	steeple
steer	stem	step	stepping	stick
sticky	stiff	still	stillness	sting
stir	stitch	stock	stocking	stole

stone	stood	stool	stoop	stop
stopped	stopping	store	stories	stork
storm	stormy	story	stove	straight
strange	stranger	strap	straw	strawberry
stream	street	stretch	string	strip
stripes	strong	stuck	study	stuff
stump	stung	subject	such	suck
sudden	suffer	sugar	suit	sum
summer	sun	Sunday	sunflower	sung
sunk	sunlight	sunny	sunrise	sunset
sunshine	supper	suppose	sure	surely
surface	surprise	swallow	swam	swamp
swan	swat	swear	sweat	sweater
sweep	sweet	sweetheart	sweetness	swell
swept	swift	swim	swimming	swing
switch	sword	swore	table	tablecloth
tablespoon	tablet	tack	tag	tail
tailor	take	taken	taking	tale
talk	talker	tall	tame	tan
tank	tap	tape	tar	tardy
task	taste	taught	tax	tea
teach	teacher	team	tear	tease
teaspoon	teeth	telephone	tell	temper
ten	tennis	tent	term	terrible
test	than	thank	thankful	thanks
Thanksgiving	that	that's	the	theater
thee	their	them	then	there
these	they	they'd	they'll	they're
they've	thick	thief	thimble	thin
thing	think	third	thirsty	thirteen
thirty	this	thorn	those	though
thought	thousand	thread	three	threw
throat	throne	through	throw	thrown
thumb	thunder	Thursday	thy	tick
ticket	tickle	tie	tiger	tight
till	time	tin	tinkle	tiny
tip	tiptoe	tire	tired	title
to	toad	toadstool	toast	tobacco
today	toe	together	toilet	told
tomato	tomorrow	ton	tone	tongue
tonight	too	took	tool	toot
tooth	toothbrush	toothpick	top	tore

torn	toss	touch	tow	toward
towards	towel	tower	town	toy
trace	track	trade	train	tramp
trap	tray	treasure	treat	tree
trick	tricycle	tried	trim	trip
trolley	trouble	truck	truly	trunk
trust	truth	try	tub	Tuesday
tug	tulip	tumble	tune	tunnel
turkey	turn	turtle	twelve	twenty
twice	twig	twin	two	ugly
umbrella	uncle	under	understand	underwear
undress	unfair	unfinished	unfold	unfriendly
unhappy	unhurt	uniform	United	unkind
unknown	unless	unpleasant	until	unwilling
up	upon	upper	upset	upside
upstairs	uptown	upward	us	use
used	useful	valentine	valley	valuable
value	vase	vegetable	velvet	very
vessel	victory	view	village	vine
violet	visit	visitor	voice	vote
wag	wagon	waist	wait	wake
waken	walk	wall	walnut	want
war	warm	warn	was	wash
washer	washtub	wasn't	waste	watch
watchman	water	watermelon	waterproof	wave
wax	way	wayside	we	weak
weaken	weakness	wealth	weapon	wear
weary	weather	weave	web	we'd
wedding	Wednesday	wee	weed	week
weep	weigh	welcome	well	we'll
went	were	we're	west	western
wet	we've	whale	what	what's
wheat	wheel	when	whenever	where
which	while	whip	whipped	whirl
whiskey	whisky	whisper	whistle	white
who	who'd	whole	who'll	whom
who's	whose	why	wicked	wide
wife	wiggle	wild	wildcat	will
willing	willow	win	wind	windmill
window	windy	wine	wing	wink
winner	winter	wipe	wire	wise
wish	wit	witch	with	without

woke	wolf	woman	women	won
wonder	wonderful	won't	wood	wooden
woodpecker	woods	wool	woolen	word
wore	work	worker	workman	world
worm	worn	worry	worse	worst
worth	would	wouldn't	wound	wove
wrap	wrapped	wreck	wren	wring
write	writing	written	wrong	wrote
wrung	yard	yarn	year	yell
yellow	yes	yesterday	yet	yolk
yonder	you	you'd	you'll	young
youngster	your	you're	yours	yourself
yourselves	youth	you've	FALSE	TRUE

## APPENDIX B

### NSDL METADATA GUIDELINES

Appendix B contains descriptions of the metadata elements that NSDL member collections should use when cataloging new resources; only title and URL are required with all other elements being optional. This research seeks to predict the audience level for all resources in the digital library collection; the audience level is known as the education level in the NSDL metadata (National Science Digital Library (NSDL), 2013).

Element	Recommended Usage	Definition	Sample XML tags
Title	Required	The name by which the resource or collection of resources is formally known.	<dc:title>... </dc:title>
Alternative Title	Recommended if applicable	A refinement of the Title element used to express varying form(s) of a title [e.g., <i>Journal of polymer science</i> (title); <i>Polymer symposia</i> (Alternative Title)].	<dc:alternative>... </dc:alternative>
Identifier	Required	URL to the resource	<dc:identifier>... </dc:identifier>
Subject	Strongly recommended	Populate each Subject field with only one subject term (or phrase) that describes the topics, concepts or content of the resource; repeat as needed.	<dc:subject>... </dc:subject>
Education Level	Strongly recommended	Use to describe the appropriate learning level or range associated with a resource. A refinement of the audience element. NSDL controlled vocabulary available.	<dc:educationLevel>... </dc:educationLevel>



Element	Recommended Usage	Definition	Sample XML tags
Audience	Recommended	A broad category that best describes the recipient or user for whom the resource is primarily intended. NSDL controlled vocabulary available.	< dct:audience>... </ dct:audience>
Mediator	Optional	A class of entity that mediates access to the resource and for whom the resource is intended or useful.	< dct:mediator>... </ dct:mediator>
Description	Strongly recommended	A free-text account of a resource. May include abstracts or table of contents. Used as primary search field and display field.	< dc:description>... </ dc:description>
Type	Strongly recommended	The nature, function or typical use of a resource. NSDL controlled vocabulary and DCMI type list available. To describe the file format, physical medium, or dimensions of the resource, use Format element.	< dc:type>... </ dc:type>
Rights	Recommended	Rights information typically includes a free-text statement about various property rights associated with the resource, including intellectual property rights. May be populated with a URL that links to specific rights language in the resource.	< dc:rights>... </ dc:rights>
Access Rights	Optional	Information describing conditions or requirements for viewing and/or downloading NSDL material. NSDL controlled vocabulary available; a refinement of the Rights element.	< dct:accessRights>... </ dct:accessRights>

Element	Recommended Usage	Definition	Sample XML tags
License	Optional	A legal document giving official permission to do something with the resource. A refinement of the Rights element.	< dct:license>... </ dct:license>
Contributor	Recommended	Entity responsible for making contributions to the resource. Populate each Contributor field with only one contributor term; repeat as needed.	< dc:contributor>... </ dc:contributor>
Creator	Recommended	Entity primarily responsible for making the resource.	< dc:creator>... </ dc:creator>
Publisher	Recommended	Entity responsible for making the resource available.	< dc:publisher>... </ dc:publisher>
Language	Recommended	Primary language of the resource. NSDL_DC recommends use of LOC's <i>ISO 639-2</i> controlled vocabulary.	< dc:language>... </ dc:language>
Coverage	Optional	Statement of resource's spatial/geographic and/or temporal coverage. Named places (countries, cities, etc.) or time periods (epochs, date ranges, etc.) are typical Coverage values.	< dc:coverage>... </ dc:coverage>
Spatial	Optional	Spatial characteristics of the intellectual content of the resource.	< dct:spatial>... </ dct:spatial>
Temporal	Optional	Temporal characteristics of the intellectual content of the resource.	< dct:temporal>... </ dct:temporal>
Date	Recommended	A point or period of time associated with an event in the lifecycle of the resource. Employ W3CDTF encoding scheme that looks like YYYY-MM-DD.	< dc:date>... </ dc:date>
Created	Recommended	A refinement of the Date element	< dct:created>... </ dct:created>

Element	Recommended Usage	Definition	Sample XML tags
Available	Optional	A refinement of the Date element	< dct:available>... </ dct:available>
date Accepted	Optional	A refinement of the Date element	< dct:dateAccepted>... </ dct:dateAccepted>
date Copyrighted	Optional	A refinement of the Date element	< dct:dateCopyrighted>... </ dct:dateCopyrighted>
date Submitted	Optional	A refinement of the Date element	< dct:dateSubmitted>... </ dct:dateSubmitted>
Issued	Optional	A refinement of the Date element	< dct:issued>... </ dct:issued>
Modified	Optional	A refinement of the Date element	< dct:modified>... </ dct:modified>
Valid	Optional	A refinement of the Date element	< dct:valid>... </ dct:valid>
Interactivity Type	Recommended if applicable	The type of interactions supported by a resource (active, expositive, mixed, undefined)	< ieee:interactivityType>... </ ieee:interactivityType>
Interactivity Level	Recommended if applicable	The level of interaction between a resource and end user; that is the degree to which the learner can influence the behavior of the resource (very high, high, medium, low, very low)	< ieee:interactivityLevel>... </ ieee:interactivityLevel>
Typical Learning Time	Optional	The typical amount of time for a particular education level to interact with the resource.	< ieee:typicalLearningTime>... </ ieee:typicalLearningTime>
Format	Optional	Physical medium and/or file/MIME format	< dc:format>... </ dc:format>
Extent	Optional	The size or duration of the resource.	< dct:extent>... </ dct:extent>
Medium	Optional	The material or physical carrier of the resource.	< dct:medium>... </ dct:medium>

Element	Recommended Usage	Definition	Sample XML tags
Relation	Recommended if applicable	A related resource. Best practice to express relationships to related resources and the item being cataloged is to employ the applicable refinements below. Enter either the title and/or URL of the related resource.	<dc:relation>... </dc:relation>
• conformsTo		A refinement of the Relation element. Also used to provide educational standard via a URI (e.g. as ASN URIs).	< dct:conformsTo>... </dct:conformsTo>
• isFormatOf		A refinement of the Relation element	< dct:isFormatOf>... </dct:isFormatOf>
• hasFormat		A refinement of the Relation element	< dct:hasFormat>... </dct:hasFormat>
• isPartOf		A refinement of the Relation element	< dct:isPartOf>... </dct:isPartOf>
• hasPart		A refinement of the Relation element	< dct:hasPart>... </dct:hasPart>
• isReferencedBy		A refinement of the Relation element	< dct:isReferencedBy>... </dct:isReferencedBy>
• references		A refinement of the Relation element	< dct:References>... </dct:References>
• isReplacedBy		A refinement of the Relation element	< dct:isReplacedBy>... </dct:isReplacedBy>
• replaces		A refinement of the Relation element	< dct:replaces>... </dct:replaces>
• isRequiredBy		A refinement of the Relation element	< dct:isRequiredBy>... </dct:isRequiredBy>
• requires		A refinement of the Relation element	< dct:requires>... </dct:requires>
• isVersionOf		A refinement of the Relation element	< dct:isVersionOf>... </dct:isVersionOf>
• hasVersion		A refinement of the Relation element	< dct:hasVersion>... </dct:hasVersion>
Abstract	Optional	A summary of the content of the resource. A refinement of the Description element	< dct:abstract>... </dct:abstract>

<b>Element</b>	<b>Recommended Usage</b>	<b>Definition</b>	<b>Sample XML tags</b>
Table of Contents	Optional	A list of subunits of the content of the resource. A refinement of the Description element	<dct:tableOfContents>... </dct:tableOfContents>
Bibliographic citation	Optional	A bibliographic reference for the resource. A refinement of the Identifier element	<dct:bibliographicCitation>.. </dct:bibliographicCitation>
Instructional method	Optional	Describes process by which knowledge, attitudes, and/or skills are instilled.	<dct:instructionalMethod>... </dct:instructionalMethod>
Provenance	Optional	Statement of ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation.	<dct:provenance>... </dct:provenance>
Accrual method	Optional	Method by which items are added to a collection; rarely used in NSDL.	<dct:accrualMethod>... </dct:accrualMethod>
Accrual periodicity	Optional	Frequency with which items are added to a collection; rarely used in NSDL.	<dct:accrualPeriodicity>... </dct:accrualPeriodicity>
Accrual policy	Optional	Policy governing the addition of items to a collection.	<dct:accrualPolicy>... </dct:accrualPolicy>

## REFERENCES

- Adage.com (2010). *Wall Street Journal Platforms – 360 Media Guide*. Retrieved June 30, 2010 from <http://brandedcontent.adage.com/360/details.php?brand=17>.
- American Education Publishing (2000). *The Complete Book of Reading*. Greensboro, NC.
- Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G., & Spyropoulos, C. (2000). An Evaluation of Naive Bayesian Anti-Spam Filtering. *Machine Learning in the New Information Age*, 9-17.
- Bernstein, J.H. (2006). From the Ubiquitous to the Nonexistent: A Demographic Study of OCLC WorldCat. *Library Resources and Technical Services*, 50(2), 79-90.
- Berry, M. & Browne, M. (2005). Understanding Search Engines: Mathematical Modeling and Text Retrieval. *Society for Industrial and Applied Mathematics*.
- Berry, M., Dumais, S., & O'Brien, G. (1995). Using Linear Algebra for Intelligent Information Retrieval. *Society for Industrial and Applied Mathematics Review*, 37(4), 573-595.
- Bezdek, J. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, NY: Plenum Press.
- Bo, P., Lillian, L., & Shivakumar, V. (2002). Thumbs Up? Sentiment Classification using Machine Learning Techniques. *ACL Conference on Empirical Methods in Natural Language Processing*.
- Bradford, R. (2008). An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. *Proceedings of the 17<sup>th</sup> ACM Conference on Information and Knowledge Management*, 153-162.
- Burstein, J., Kukich, K., Wolff, S., Chi, L., & Chodorow, M. (1998). Enriching Automated Essay Scoring Using Discourse Marking. *Workshop on Discourse Relations and Discourse Marking, Annual Meeting of the Association of Computational Linguistics*.
- Burstein, J., Leacock, C., & Swartz, R. (2001). Automated Evaluation of Essay and Short Answers. *Proceedings of the Sixth International Computer Assisted Assessment Conference*.
- Buse, R. & Weimer, W. (2008). A Metric for Software Readability. *Proceedings of the International Symposium in Software Testing and Analysis (ISSTA)*.
- Carroll, J.B., Davies, P., & Richmond, B. (1971). *The Word Frequency Book*. Boston, MA: Houghton Mifflin.
- Chall, J. & Dale, E. (1995). *Readability Revisited: The New Dale-Chall Readability Formula*. Cambridge, MA: Brookline Books.

- Chami.com. (2012). Windows Application – URL2File. Retrieved April 20, 2012 from [http://www.chami.com/free/url2file\\_wincon.html](http://www.chami.com/free/url2file_wincon.html).
- Chessman, P., & Stutz, J. (1996). Bayesian Classification (Auto Class): Theory and Results. *Advances in Knowledge Discovery and Data Mining*. Boston, MA: MIT Press, 153-180.
- Cognitive Technologies. (2012). Open OCR Cuneiform. Retrieved April 2, 2012 from <http://en.openocr.org/>.
- Collins-Thompson, K. & Callan, J. (2004). A Language Modeling Approach to Predicting Reading Difficulty. *Proceedings of the 2004 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Collins-Thompson, K. & Callan, J. (2005). Predicting Reading Difficulty with Statistical Language Models. *Journal of the American Society for Information Science and Technology*, 56(13), 1448-1462.
- Cutler, M., Deng, H., Maniccam, S. S., & Meng, W. (1999). A New Study on Using HTML Structures to Improve Retrieval. *Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence*.
- Cutler, M., Shih, Y., & Meng, W. (1997). Using the Structure of HTML Documents to Improve Retrieval. *Proceedings of the USENIX Symposium on Internet Technologies and Systems*.
- D'Alessandro, D., Kingsley, P., & Johnson-West, J. (2001). The Readability of Pediatric Patient Education Materials on the World Wide Web. *Archives of Pediatric and Adolescent Medicine*, 155(7), 807-812.
- Deerwester, S., Dumais, S., & Harshman, R. (1988). Improving Information Retrieval Using Latent Semantic Indexing. *Proceedings of the 51st Annual Meeting of the American Society for Information Science*, 36-40.
- Diederich, J., Kindermann, J., Leopold, E., & Paass, G. (2003). Authorship Attribution with Support Vector Machines. *Applied Intelligence*, 19(1/2), 109-123.
- Dikmen, I., Birgonul, T., & Budayan, C. (2009). Strategic Group Analysis in the Construction Industry. *Journal of Construction Engineering and Management*.
- Ding, C. (1999). A Similarity-based Probability Model for Latent Semantic Indexing. *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 59-65.
- Dumais, S.T., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive Learning Algorithms and Representations for Text Categorization. *Proceedings of CIKM, the ACM International Conference on Information and Knowledge Management*, 148-155.
- Dunn, J. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*, 3(3), 32-57.

- Elberrichi1, Z., Rahmoun, A., & Bentaalah, M. (2008). Using WordNet for Text Categorization. *The International Arab Journal of Information Technology*, 5(1), 3-37.
- Elhadi, G.F. & Abbas, M.A. (2010). Clustering DNA Sequences by Self Organizing Map and Similarity Functions. *Proceedings of The 7th International Conference on Informatics and Systems (INFOS)*.
- Flesch, R. (1948). A New Readability Yardstick. *Journal of Applied Psychology*, 32(3), 221-233.
- Fraley, C. & Raftery, A. (2009). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. *Technical Report No 504*. Seattle, WA: Department of Statistics, University of Washington.
- Furnas, G., Landauer, T., Gomez, M., & Dumais, S. (1987). The Vocabulary Problem in Human-System Communication. *Communications of the ACM*, 30(11), 964-971.
- Gale, W. (1995). Good-Turing Smoothing Without Tears. *Journal of Quantitative Linguistics*, 2(3), 217-237.
- Gee, K. (2003). Using Latent Semantic Indexing to Filter Spam. *Proceedings of the ACM Symposium on Applied Computing*, 460-464.
- George, R.K. (2000). The Measurement of Readability: Useful Information for Communicators. *ACM Journal of Computer Documentation*, 24(3), 11-25.
- Gong, Y., & Liu, X. (2001). Creating Generic Text Summaries. *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, 903-907.
- Gordan, M. & Dumais, S. (1998). Using Latent Semantic Indexing for Literature Based Discovery. *Journal of the American Society for Information Science*, 49(8), 674-685.
- Gunning, R. (1952). *The Technique of Clear Writing*. New York, NY: McGraw-Hill International Book Co.
- Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of The 10th European Conference on Machine Learning (ECML)*, 137-142.
- Joachims, T. (1999). Transductive Inference for Text Classification using Support Vector Machines. *Proceedings of the 16th International Conference on Machine Learning (ICML)*, 200-209.
- Karnik, A. Goswami, S., & Guha, R. (2007). Detecting Obfuscated Viruses Using Cosine Similarity Analysis. *Proceedings of the First Asia International Conference on Modeling & Simulation (AMS)*.
- Kidzone.ws. (2012). The Water Cycle. Retrieved April 20, 2012 from <http://www.kidzone.ws/water/>.



- Kincaid J., Fishburne R., Rogers R., & Chissom, B. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count, and Flesch Reading Ease Formula) for Navy Enlisted Personnel. *Research Branch Report 8-75*. Memphis, TN: U. S. Naval Air Station.
- Kyriakopoulou, A. & Kalamboukis, T. (2008). Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems. *Proceedings of the European Conference on Machine Learning (ECML)*.
- Larkey, L. (1998). Automatic Essay Grading Using Text Categorization Techniques. *Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR)*.
- Learning Resource Metadata Initiative (LRMI). (2013). Learning Resource Metadata Initiative Specification. Retrieved March 29, 2013 from <http://www.lrmi.net/the-specification>.
- Leroy, G. Miller, T., Roseblat, G. & Browne, A. (2008). A Balanced Approach to Health Information Evaluation: A Vocabulary-Based Naïve Bayes Classifier and Readability Formulas. *Journal of the American Society for Information Science and Technology*, 59(9), 1409-1419.
- Leslie, C., Eskin, E., & Noble, W. (2002). The Spectrum Kernel: A String Kernel for SVM Protein Classification. *Pacific Symposium on Biocomputing*.
- Lexile.com. (2010). Lexile to Grade Level Correspondence. Retrieved October 20, 2010 from <http://www.lexile.com/about-lexile/grade-equivalent/grade-equivalent-chart/>.
- Lin, H-J., Yen, S-H., Yeh, J-P., & Lin, M-J. (2008). Face Detection Based on Skin Color Segmentation and SVM Classification. *Proceedings of The Second International Conference on Secure System Integration and Reliability Improvement*.
- Liu, X., Croft, W.B., Oh, P., & Hart, D. (2004). Automatic Recognition of Reading Levels from User Queries. *Proceedings of the 27<sup>th</sup> Annual International ACM SIGIR conference on Research and Development in Information Retrieval*.
- Lloyd., S. (1982). "Least squares quantization in PCM." *IEEE Transactions on Information Theory*, 28(2), 129-137.
- McCallum, D. R. & Peterson, J.L. (1982). Computer-Based Readability Indexes. *Proceedings of ACM*, 44-48.
- McCown, F., Johan, B., & Michael, N. (2005). Evaluation of the NSDL and Google for Obtaining Pedagogical Resources. *Proceedings of the European Conference on Digital Libraries*, 344-355.
- McGraw-Hill Children's Publishing. (2002). The Complete Book of Math, Grades 5-6. Columbus, Ohio.
- McLaughlin, G.H. (1969). SMOG Grading – A New Readability Formula. *Journal of Reading*, 12(8), 639-646.

- McNamara, D., Graesser, A., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavior Research Methods, Instruments, & Computers*, 36(2), 193-202.
- MediaMetrics. (2007). *Using the Lexile Analyzer for Educators and Media Specialists*. Cammeray, Australia: MediaMetrics, Inc.
- Miller, T., Leroy, G., Chatterjee, S., Fan, J., & Thoms, B. (2007). A Classifier to Evaluate Language Specificity of Medical Documents. *Proceedings of the 40<sup>th</sup> Annual Hawaii International Conference on System Science (HICSS)*.
- Milone, M. (2010). The Development of ATOS: The Renaissance Readability Formula. *Technical Report*. Wisconsin Rapids, WI: Renaissance Learning.
- National Science Digital Library (NSDL). (2013). NSDL\_DC Metadata Guidelines. Retrieved on January 10, 2013 from [https://wiki.ucar.edu/display/nsdl\\_docs/nsdl\\_dc](https://wiki.ucar.edu/display/nsdl_docs/nsdl_dc).
- NetMBA. (2012). Comparative Advantage. Retrieved April 21, 2012 from <http://www.netmba.com/econ/micro/comparative-advantage/>.
- Onecle. (2010). Florida Laws: FL Statutes - Title XXXVII Insurance Section 627.011 Short title. Retrieved November 30, 2010 from <http://law.onecle.com/florida/insurance/627.4145.html>.
- Ostermeier, E. (2010). Free Republic: "Professor" Obama? President's SOTUS Notches 4th Lowest Flesch-Kincaid Grade Level Score. Retrieved October 10, 2010 from <http://www.freerepublic.com/focus/f-news/2441696/posts>.
- Page, E.B. (1994). New Computer Grading of Student Prose Using Modern Concepts and Software. *Journal of Experimental Education*, 62(2), 127-142.
- Pereira, R.A., Molinari, A., & Pasi, G. (2005). Contextual Weighted Representations and Indexing Models for the Retrieval of HTML Documents. *Soft Computing*, 9(7), 481-492.
- Ramos, V. & Abraham, A. (2005). Self-Organized Ant-based Clustering Model for Intrusion Detection Systems (ANTIDS). *Proceedings of the 4th IEEE International Conference on Soft Computing as Transdisciplinary Science and Technology*, 977-986.
- Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago, IL: University of Chicago Press.
- Salton, G., Wong, A., & Yang, C. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620.
- Si, L. & Callan, J. (2001). A Scientific Model for Scientific Readability. *Proceedings of the Tenth International Conference on Information and Knowledge Management*, 574-576.
- Spache, G. (1953). A New Readability Formula for Primary-Grade Reading Materials. *The Elementary School Journal*, 53(7), 410-413.

- Steinbach, M. Karypis, G., & Kumar, V. (2000). A Comparison of Document Clustering Techniques. *KDD Workshop on Text Mining*.
- Stenner, A.J. (1992). Meaning and Method in Reading Comprehension. *Presented at American Education Research Association, Division D, Rasch Special Interest Group*, San Francisco, CA.
- Stone, C. (1956). Measuring Difficulty of Primary Reading Material: A Constructive Criticism of Spache's Measure. *The Elementary School Journal*, 57(1), 36-41.
- Tay, F. & Cao, L. (2001). Application of Support Vector Machines in Financial Time Series Forecasting. *Omega*, 29(4).
- Theodoridis, S. & Koutroumbas, K. (2006). *Pattern Recognition 3<sup>rd</sup> Edition*. Waltham, MA: Academic Press, 635.
- Turner, T., & Brackbill, L. (1998). Rising to the Top: Evaluating the Use of the HTML META Tag to Improve Retrieval of World Wide Web Documents through Internet Search Engines. *Library Resources & Technical Services*, 42(4), 258-271.
- Turney, P.D. (2001). Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Review. *Proceedings of the 40<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education*, 2, 319-330.
- Weatherley, J. (2012). jOAI: NSDL Technical Documentation. Retrieved April 20, 2012 from <https://wiki.ucar.edu/display/nsdl/docs/jOAI>.
- Wei, C-P., Yang, C. & Lin, C-M. (2008). A Latent Semantic Indexing-Based Approach to Multilingual Document Clustering. *Decision Support Systems*, 45(3), 606-620.
- White, H.D. (2008). Better than Brief Tests: Coverage Power Tests of Collection Strength. *College & Research Libraries*, 155-174.
- Whittington, D., & Hunt, H. (1999). Approaches to the Computerized Assessment of Free Text Responses. *Proceedings of the Sixth International Computer Assisted Assessment Conference*.
- Will, T., Srinivasan, A., Im, I., & Wu, Y.-F. (2009). Search Personalization: Knowledge Based Recommendation in Digital Libraries. *Proceedings of the Americas Conference on Information Systems (AMCIS)*.
- Wright, B. & Stenner, A. (1998). Readability and Reading Ability. *Presentation to the Australian Council on Education Research (ACER)*.
- Yang, Y. & Liu, X. (1999). A Re-Examination of Text Categorization Methods. *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 42-49.
- Zhang, H. (2004). The Optimality of Naïve Bayes. *Proceedings of the International Florida Artificial Intelligence Research Society Conference (FLAIRS)*.