Improving Access to Digital Library Resources by Automatically Generating Complete Reading Level Metadata

Todd Will
Information Systems, New Jersey Institute of Technology, Newark, NJ, United States., todd.will@njit.edu

Yi-Fang Wu
Information Systems Department, New Jersey Institute of Technology, Newark, NJ, United States., wu@njit.edu

Improving Access to Digital Library Resources by Automatically Generating Complete Reading Level Metadata

Todd Will
New Jersey Institute of Technology
todd.will@njit.edu

Yi-Fang Brook Wu
New Jersey Institute of Technology
wu@njit.edu

ABSTRACT

Digital library collections usually hold resources describing a limited set of topics spanning a wide range of reading levels, requiring complete reading level metadata to filter relevant resources from the collection. In order to suggest the reading level for all resources in the test collection, we propose an SVM-based classification tool which predicts the specific reading level with an F-Measure of 0.70 for all resources, outperforming other classification methods and readability formulas under evaluation. To measure the impact of reading level metadata completeness on retrieval performance, a knowledge based system retrieves documents from three collections containing different reading level completeness: one with complete reading level information generated by the proposed SVM method, one missing all reading level information, and the final collection containing limited, human-expert provided metadata. The dataset with automatically identified complete reading level exceeds the performance of collection-provided reading level metadata for all five sample tasks.

Keywords

Automatic Metadata Generation, Reading Level, Digital Libraries, Knowledge Based Filtering

INTRODUCTION

As digital library users do not have access to reference librarians, the retrieval system must take advantage of complete and consistent reading level metadata to identify resources that both challenge and inform the reader. Elementary school students require simple terms while college students require more complex terms to maintain interest. To decrease the effort required by users to identify relevant resources, knowledge based filtering techniques drawn from the e-commerce domain can further refine the results from a digital library retrieval system but only if metadata is complete and consistent.

However, in a 212,000 document collection from the National Science Digital Library (NSDL), we have found that less than 5% of the resources include any reading level metadata. Even though resources may contain human expert-provided reading level, this information is sometimes incompatible among different collections. For example, Teacher's Domain provides the reading level as Kindergarten through Second Grade, while the Digital Library for Earth Systems Education (DLESE) enters more specific reading levels, including First Grade, Second Grade, etc. Rather than manually entering reading level information for all resources in the collection, we propose using a Support Vector Machines (SVM) tool to automatically predict the reading level for all digital library resources in the test collection.

The remainder of the paper is organized as follows. First, the SVM algorithm is reviewed, followed by a review of knowledge based systems. Next, the performance of the SVM-based tool is compared with four other reading level identification methods. Third, the experimental design and results of the user study measuring the impact of reading level metadata consistency and completeness on retrieval performance are described. Finally, the discussion and conclusion section is provided.

LITERATURE REVIEW

This study seeks to measure the impact of varying degrees of reading level metadata completeness on the performance of a digital library retrieval system. After the algorithm of the proposed SVM reading level identification tool is presented, knowledge based systems that can better match users with relevant resources are reviewed.

Support Vector Machines (SVM)

SVM has been originally developed by Joachims (1998) and used in a wide range of classification tasks (Page, 1994; Yang & Liu, 1999). For example, an SVM-based SPAM detection system separates SPAM messages from non-SPAM ones with 94.54% accuracy (Kyriakopoulou & Kalamboukis, 2008). SVM is a binary classification problem where the document falls

into one class or the other (Siolas & d'Alche-Buc, 2000), but can also be used to solve multiclass problems as well (Dumais, et. al., 1998).

In the SVM algorithm, the input training data can be modeled as follows, where k is the number of positive samples, 1-k is the number of negative samples, and $y \in \{+1, -1\}$ and each x_i is an N-dimensional vector in the real-valued space as follows:

$$(x_1^+, y_1), ..., (x_k^+, y_k), (x_{k+1}^-, y_{k+1}) ..., (x_l^-, y_l)$$

After the model is trained with all input resources, the decision boundary hyperplane (H_d) can be represented as follows:

$$H_d : \langle w \bullet x \rangle + b = 0$$

Two additional hyperplanes are defined that pass through the nearest positive and negative normal vectors. As the margin between these hyperplanes increases, the confidence that the new resource is correctly classified also increases. Figure 1 displays a graphical representation of the SVM model:

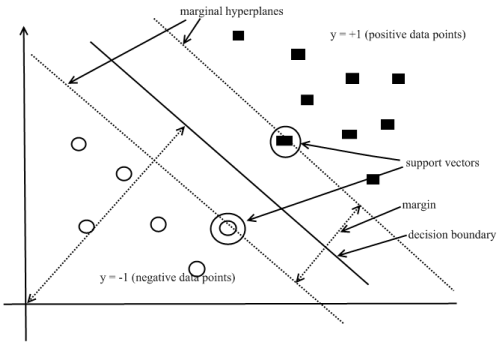

Figure 1: SVM Graphical Representation

SVM performs well for document classification. The complexity of the model remains independent of the number of resources that need to be classified. Simple terms tend to occur in lower reading levels while complex terms usually occur at higher reading levels.

Knowledge Based Filtering

Knowledge based (KB) filtering systems must understand the user's task to identify relevant resources in the collection (Callan et al, 2003). These systems have knowledge about both resources and users to identify relevant resources in the collection, as shown in Table 1 (Burke, 2002); examples of the knowledge type and the digital library equivalent are also provided:

Type	Description	Example	Digital Library Equivalent
Catalog Knowledge	Features of items in the catalog	Type of food a restaurant serves	Document descriptions or metadata
Functional Knowledge	Map user needs to item descriptions	Cool-looking computer means shiny case and bright colors	Matching metadata to user needs that are entered in detailed profile
User Knowledge	Knowledge of users	Educational level, areas of expertise	User enters information about themselves into the profile

Table 1: Required Knowledge for Effective KB Recommendation

Ghani and Fano (2002) have created a product catalog and associated semantic attributes, including age, price point, and degree of formality, from descriptions of different products. Users browse and purchase items on the site to create a database

of preferences based on viewed items. These preferences are applied to products in other categories, such as a buyer purchasing vintage clothes should also like vintage furniture. The system reasons the relationship between the user's preferences and product attributes.

However, the application of KB recommendation in digital libraries is limited, as most studies are E-commerce applications. Some work attempts to relate the current user information need to the current task but ignores other aspects of KB implementations (McNee et al, 2002). Another KB design applies the previous browsing history and user preferences to the current information search, but only with respect to learning object types (Tsai et al, 2006). A more recent knowledge based system for digital libraries uses the metadata elements of subject category, description, coverage, relation, and reading level to refine a text search query; this system outperforms Lucene, content-based, and collaborative filtering systems used in this evaluation (Will, et. al. 2009). An elementary school teacher typically searches for simple concepts with lower reading levels when preparing teaching materials, whereas a college professor would need peer-reviewed articles for a research paper. Complete and consistent specific reading level and other metadata should be stored with every resource in the collection to enable the KB to best match users with appropriate resources.

READING LEVEL PREDICTION PERFORMANCE COMPARISON

To fairly measure the performance of the proposed SVM-based tool, several baseline methods were used in the evaluation. Readability formulas relied on easily-calculated numerical representations of full-text characteristics along with constants derived from regression analysis to predict reading difficulty for any document independent of the domain. The Flesch-Kincaid Reading Age, perhaps the most popular formula, relied on word and sentence characteristics to suggest the specific grade level according to the American educational system (Kincaid, et. al., 1975):

FKRA=-15.59+11.8*(Average Syllables per Word)+0.39*(Average Words per Sentence)

On the other hand, the Dale-Chall Reading Ease Score compared terms in the text with a set of terms understood by a fourth grade student in addition to average sentence length:

R=0.1579*(Percentage of Words Not in Dale Common Word List)+0.0496*(Average Sentence Length)

After adding the number 3.6365 if more than 5% of words did not appear on the Dale Common Word List, the score was converted into grade level; scores of 0 to 4.9 represented less than fourth grade, scores 5.0 to 8.9 represented grades fifth through twelfth, and higher scores represented college resources (Chall & Dale, 1995).

Classification methods took a different approach. First, human experts identified representative resources appropriate for each reading level. Then, the classification model was trained by extracting the text from each of these pre-labeled resources and creating a vocabulary for each reading level containing the terms along with the term weight. Finally, the terms from each document with missing reading level information were extracted and compared with the vocabulary compiled for each reading level; the resource was assigned to the class whose terms were most similar to the unlabelled resource terms. Cosine was based on the classic vector space model that represented documents, queries, or other textual information in a vector space (Salton, Wong, & Yang, 1975); the unlabeled document was labeled with the class with the smallest cosine between the document and class vectors. The Naïve Bayes classification model was based on Bayes theorem from statistics, which stated that the presence or absence of a feature was unaffected by the presence or absence of any other feature; in classification, the Bayes decision rule was based on the maximum likelihood that a document was created from the class terms (Collins-Thompson & Callan, 2005). SVM relied on mathematical expressions that divided the vocabulary from positive and negative classes using a linear hyperplane to achieve the highest separation (Joachims, 1998). With respect to the proposed SVM tool, after sorting the classes descending based on the number of training documents contained in each class, the top ranked class was considered positive while all other classes were considered negative. If the unlabeled document fell into the positive class, it was labeled with this reading level and the process stopped; if the document fell into the negative class, the next-ranked class was considered positive while all other classes were negative. This process continued until the document falls into the positive class. These methods required human-expert labeled training samples for each reading level; this training data was only appropriate for this dataset and could not be applied to another collection unless additional samples were collected.

All readability formulas and classification methods were evaluated using the standard classification performance measures of precision, recall, and F-measure. Precision was the proportion of resources assigned a particular reading level by the automated method that matched the human-expert identified reading level. Recall was the proportion of resources with a human-expert identified reading level that the automated method correctly identified. The F-measure was the harmonic mean between precision and recall, calculated by 2*(Precision * Recall) / (Precision + Recall). Tables 2 and 3 summarized the reading level prediction precision, recall, and F-measure for the test collection containing 5,815 web-based digital library

resources distributed among all reading levels; training and testing was performed using the full text of the resources. Since the KB system required precise metadata to match users with resources, all methods suggested the specific rather than general reading level for all resources with missing reading level information. If the user sought elementary school resources, the retrieval system could include all resources labeled with first through fifth reading levels. Table 2 compared the performance of readability formulas with the proposed SVM tool.

Specific Reading Level	Docs Per Reading Level	Flesch-Kinkaid			Dale-Chall			SVM		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Kindergarten	291	0.04	0.02	0.02	0.00	0.00	0.00	0.66	0.52	0.58
First	315	0.00	0.00	0.00	0.00	0.00	0.00	0.55	0.62	0.58
Second	263	0.03	0.00	0.01	0.00	0.00	0.00	0.60	0.49	0.54
Third	258	0.00	0.00	0.00	0.02	0.03	0.03	0.60	0.48	0.54
Fourth	308	0.00	0.00	0.00	0.06	0.93	0.11	0.71	0.76	0.74
Fifth	314	0.00	0.00	0.00	0.02	0.03	0.02	0.86	0.75	0.80
Sixth	532	0.07	0.01	0.01	0.05	0.01	0.01	0.69	0.69	0.69
Seventh	312	0.01	0.00	0.01	0.06	0.01	0.01	0.51	0.77	0.62
Eighth	317	0.04	0.02	0.03	0.06	0.00	0.01	0.58	0.70	0.64
Ninth	420	0.00	0.00	0.00	0.00	0.00	0.00	0.82	0.92	0.87
Tenth	382	0.13	0.08	0.10	0.00	0.00	0.00	0.79	0.71	0.75
Eleventh	388	0.07	0.06	0.06	0.00	0.00	0.00	0.90	0.60	0.72
Twelfth	467	0.04	0.04	0.04	0.00	0.00	0.00	0.64	0.64	0.64
Undergraduate Lower Division	409	0.07	0.25	0.11	0.00	0.00	0.00	0.73	0.74	0.73
Undergraduate Upper Division	439	0.07	0.16	0.10	0.00	0.00	0.00	0.91	0.74	0.82
Graduate	400	0.07	0.30	0.12	0.00	0.00	0.00	0.65	0.86	0.74
Overall Totals	**5,815**	**0.07**	**0.07**	**0.07**	**0.05**	**0.05**	**0.05**	**0.70**	**0.70**	**0.70**

Table 2: SVM Reading Level Prediction Performance Comparison Vs. Readability Formulas

Web pages were much shorter than books, typically comprising one or two printed pages, and used tables and figures to represent information versus sentences in books, leading to extremely poor performance with F-measures under 0.10. On the other hand, since SVM matched terms in the unlabeled document with a set of terms appropriate for each reading level, these issues were minimized since SVM did not require lengthy documents nor were sentences considered.

To fairly measure the performance of SVM, two additional automated reading level identification tools were developed, based on cosine and Naïve Bayes. As shown in Table 3, the proposed SVM-based reading level prediction tool performance was an overall F-measure of 0.70, outperforming Naïve Bayesian and cosine methods. Our SVM-based tool was then used to complete the reading level for all resources with missing or incompatible metadata. If the reading level metadata did not follow the same standard or was missing, the retrieval system could not use this information to refine the retrieved documents beyond a text search query.

Specific Reading Level	Docs Per Reading Level	Cosine			Naïve Bayes			SVM		
		Precision	Recall	F-Measure	Precision	Recall	F-measure	Precision	Recall	F-Measure
Kindergarten	291	0.35	0.26	0.30	0.52	0.38	0.44	0.66	0.52	0.58
First	315	0.36	0.43	0.39	0.46	0.53	0.49	0.55	0.62	0.58
Second	263	0.35	0.27	0.31	0.46	0.39	0.42	0.60	0.49	0.54
Third	258	0.41	0.25	0.31	0.51	0.34	0.41	0.60	0.48	0.54
Fourth	308	0.57	0.56	0.56	0.67	0.71	0.69	0.71	0.76	0.74
Fifth	314	0.73	0.47	0.58	0.84	0.64	0.73	0.86	0.75	0.80
Sixth	532	0.49	0.41	0.45	0.63	0.61	0.62	0.69	0.69	0.69
Seventh	312	0.41	0.56	0.48	0.47	0.64	0.54	0.51	0.77	0.62
Eighth	317	0.33	0.51	0.40	0.45	0.64	0.53	0.58	0.70	0.64
Ninth	420	0.71	0.76	0.74	0.81	0.85	0.83	0.82	0.92	0.87
Tenth	382	0.73	0.62	0.67	0.78	0.65	0.71	0.79	0.71	0.75
Eleventh	388	0.82	0.44	0.58	0.88	0.56	0.69	0.90	0.60	0.72
Twelfth	467	0.55	0.50	0.53	0.60	0.53	0.56	0.64	0.64	0.64
Undergraduate Lower Division	409	0.47	0.48	0.47	0.53	0.56	0.54	0.73	0.74	0.73
Undergraduate Upper Division	439	0.85	0.59	0.70	0.39	0.63	0.48	0.91	0.74	0.82
Graduate	400	0.36	0.77	0.49	1.03	0.81	0.91	0.65	0.86	0.74
Overall Totals	**5,815**	**0.51**	**0.51**	**0.51**	**0.61**	**0.61**	**0.61**	**0.70**	**0.70**	**0.70**

Table 3: SVM Reading Level Prediction Performance Vs. Other Classification Methods

USER STUDY DESIGN

The user study measured the performance of a knowledge-based filtering system with varying degrees of completeness and consistency of reading level metadata. Five different task profiles were created to represent a variety of topics, namely data mining, ethnomathematics, symmetry, web spider, and water cycle. Each of these task profiles were entered into the KB system that retrieved resources based on keywords and reading level information (Will, et. al., 2009). After creating a KB profile for each task, the same task profile retrieved resources from three different collections containing the same set of resources: one containing complete and consistent reading level metadata suggested by the SVM tool, another containing metadata entered by the digital librarian, and the third missing all reading level information. In the case of missing reading level metadata, text search was used for the baseline. For example, a college student pursing a bachelor's degree could seek resources describing data mining; in this case, the keywords would be "data mining" and the reading level would be undergraduate college. The same KB task profile retrieved resources from three different collections with varying degrees of reading level completeness.

After five task profiles were created in the KB and ran against the three collections, the results from these collections were combined into one list for consideration by the subject. All subjects were drawn from the student body at a technical university in New Jersey pursuing doctoral degrees in information systems. All five tasks were related to this degree or discussed general knowledge topics familiar to all people, allowing subjects to identify relevant documents that matched the task description. After the five subjects completed a consent form, each subject identified the relevant documents for two KB task profiles; the relevance of each document with respect to the task was judged by two subjects. After the relevance judgments were completed, the subjects completed a questionnaire regarding their confidence level to correctly identify relevant documents. The resource was counted as relevant if both subjects selected the resource as relevant; if one indicated that the resource was relevant while the other did not, it was counted as not relevant. This tie-handling method produced more conservative performance measures.

USER STUDY RESULTS

The impact of reading level metadata completeness and consistency on retrieval performance was evaluated using precision, recall, and F-measure, as described in the prediction performance comparison section. In addition to these measures, recall effort was the number of documents that must be browsed in order to find a given number of relevant documents (in this case, five relevant documents). Since text search engines considered user-entered keywords without regard to reading levels, two recall effort numbers were shown for each task – one where the reading level for the marked relevant document matched the reading level of the task profile while the other showed the relevant documents without considering reading level. The total KB results obtained for each task ranged between 30, where all documents overlap, and 90, where none of the documents overlap.

The first task asked subjects to identify resources that discussed data mining appropriate for an undergraduate college class. Data mining was a branch of computer science that extracted patterns from large data sets mainly through the use of statistics. Mining also referred to the extraction of precious metals from the earth. Data mining would be appropriate to discuss at the college level while mining would be more appropriate for high school students. As shown in Table 4, the precision, recall, recall effort, and F-measure did not increase between the baseline text search and the text search plus complete reading level metadata. When collection provided reading level metadata was used to find resources, zero relevant documents were found in the top ten, indicating that the relevant resources did not include reading level metadata that followed NSDL guidelines.

Database Used	Reading Level Used	Number of Results	Precision	Recall	F-Measure	Recall Effort	
Baseline Text Search	None	Top 10	0.80	0.20	0.32	RL Not Considered	RL Matches Task
		Top 20	0.70	0.35	0.47		
		Top 30	0.60	0.45	0.51	6	6
Collection Provided Metadata	Undergraduate	Top 10	0.00	0.00	N/A	>30	
		Top 20	0.15	0.075	0.10		
		Top 30	0.13	0.10	0.11		
SVM Predicted Reading Level Metadata	Undergraduate	Top 10	0.80	0.20	0.32	6	
		Top 20	0.70	0.35	0.47		
		Top 30	0.60	0.45	0.51		

Table 4: Data Mining Task Results

At a college level, web spiders could refer to downloading all pages from a website in order to create a search engine database but could also be taught in elementary school biology referencing a spider creating a web to trap flies to eat. This task asked subjects to identify resources appropriate for an undergraduate college course that discussed web spiders. As shown in Table 5, the collection provided metadata improved retrieval performance over the baseline text search, except in the case of recall effort that required eight documents to be browsed to find five relevant ones. The complete SVM-predicted reading level metadata retrieval performance improved over the text search except for the recall effort that remained the same as text search.

Database Used	Reading Level Used	Number of Results	Precision	Recall	F-Measure	Recall Effort	
Baseline Text Search	None	Top 10	0.50	0.33	0.40	RL Not Considered	RL Matches Task
		Top 20	0.45	0.60	0.51		
		Top 30	0.43	0.87	0.58	5	5
Collection Provided Metadata	Undergraduate	Top 10	0.60	0.40	0.48	8	
		Top 20	0.55	0.73	0.63		
		Top 30	0.47	0.93	0.62		
SVM Predicted Reading Level Metadata	Undergraduate	Top 10	0.60	0.40	0.48	5	
		Top 20	0.55	0.73	0.63		
		Top 30	0.47	0.93	0.62		

Table 5: Web Spider Task Results

The third task asked subjects to find relevant documents that discussed symmetry for a high school class. Symmetry referred to the ability to draw a line through an object such that both halves have the same appearance. However, symmetry also described the harmonious motion of a spring in physics, a topic more appropriate for a college class. As shown in Table 6, the inclusion of complete reading level resulted in 0.50 precision, where five of the top ten documents were relevant to the task, far exceeding the text search precision of 0.10. The recall effort decreased from well over twenty for text search and collection provided metadata to nine for complete, consistent reading level metadata; if the resource reading level must match the task description, then the recall effort increased to over thirty for the text search.

Database Used	Reading Level Used	Number of Results	Precision	Recall	F-Measure	Recall Effort	
Baseline Text Search	None	Top 10	0.10	0.06	0.07	RL Not Considered	RL Matches Task
		Top 20	0.20	0.22	0.21		
		Top 30	0.33	0.56	0.42	23	>30
Collection Provided Metadata	High School	Top 10	0.30	0.17	0.21	25	
		Top 20	0.15	0.17	0.16		
		Top 30	0.30	0.50	0.38		
SVM Predicted Reading Level Metadata	High School	Top 10	0.50	0.28	0.36	9	
		Top 20	0.45	0.50	0.47		
		Top 30	0.57	0.94	0.71		

Table 6: Symmetry Task Results

The fourth task asked subjects to find documents discussing ethnomathematics appropriate for an undergraduate college class. Ethnomathematics studied the relationship between math and culture in order to appreciate the connection between the

two. Ethnomathematics spanned all reading levels, from addition in elementary school to cultural development at college level. As shown in Table 7, the performance between the baseline text search and collection provided metadata decreased; the performance of the baseline text search was poor, requiring fourteen documents to be browsed before five relevant ones were identified. The system with complete reading level metadata outperformed all others by cutting the recall effort nearly in half over the baseline text search while doubling the precision and recall for the Top 10 results.

Database Used	Reading Level Used	Number of Results	Precision	Recall	F-Measure	Recall Effort	
Baseline Text Search	None	Top 10	0.30	0.09	0.14	RL Not Considered	RL Matches Task
		Top 20	0.50	0.30	0.38		
		Top 30	0.47	0.42	0.44	14	14
Collection Provided Metadata	Undergraduate	Top 10	0.30	0.09	0.14	>30	
		Top 20	0.15	0.09	0.11		
		Top 30	0.10	0.09	0.10		
SVM Predicted Reading Level Metadata	Undergraduate	Top 10	0.60	0.18	0.28	8	
		Top 20	0.65	0.39	0.49		
		Top 30	0.63	0.58	0.60		

Table 7: Ethnomathematics Task Results

The final task asked subjects to identify resources describing the water cycle appropriate for elementary school students. The water cycle referred to the process of water evaporating into the atmosphere, returning to the earth as rain, and then flowing to large water bodies through rivers and streams. Elementary school students would conduct simple experiments that demonstrated the water cycle while college students would be more interested in why the process occurred. As shown in Table 8, the collection manager did not provide reading level metadata for the relevant documents in the collection, causing relevant documents to be ranked lower. While the performance for SVM predicted metadata decreased slightly for the Top 10 results, the retrieval performance improved for the Top 20 and Top 30 results. If the reading level metadata for the document must match the task description, then the recall effort increased from six to fifteen for the text search.

Database Used	Reading Level Used	Number of Results	Precision	Recall	F-Measure	Recall Effort	
Baseline Text Search	None	Top 10	0.60	0.33	0.43	RL Not Considered	RL Matches Task
		Top 20	0.35	0.39	0.37		
		Top 30	0.27	0.44	0.33	6	15
Collection Provided Metadata	Elementary school	Top 10	0.20	0.11	0.14	16	
		Top 20	0.25	0.28	0.26		
		Top 30	0.17	0.28	0.21		
SVM Predicted Reading Level Metadata	Elementary School	Top 10	0.50	0.28	0.36	7	
		Top 20	0.40	0.44	0.42		
		Top 30	0.33	0.56	0.42		

Table 8: Water Cycle Task Results

The inclusion of complete and consistent reading level metadata maintained or improved the retrieval performance in four out of the five tasks; while the performance decreased for the Top 10 results for the water cycle task, the precision and recall increased for the Top 20 and Top 30 results. In all five tasks, the text search system outperformed the collection-provided reading level metadata system, indicating that the KB retrieval system removed many relevant resources due to missing or incompatible reading level metadata. When considering complete reading level metadata versus collection-provided metadata as a whole, the SVM-predicted reading level metadata experienced higher retrieval performance for all five tasks.

DISCUSSION AND CONCLUSION

The KB system required complete and consistent specific reading level metadata to identify resources that both challenged and informed users. Even though the readability formulas were simpler to use, did not require any training samples, and were applicable to any domain, the SVM-based classification tool far outperformed these formulas with an overall F-measure of 0.70 with respect to predicting the specific reading level of digital library resources. While this tool only predicted reading level, this tool could also suggest the values for other metadata elements, including subject category and coverage, if a representative set of training documents for each class could be gathered.

The second and main part of this study focused on the performance increase of a system that drew upon complete reading level metadata versus one that used the collection-provided metadata and a third with text search only. All subjects judged document relevance with confidence levels ranging between medium and high. The data mining task, querying a collection with complete metadata, did not show any improvement in performance over a text search only system, indicating that most of the resources in the collection were appropriate for a college class. The water cycle complete reading level metadata retrieval performance, when compared with text search only, decreased for the Top 10 results but improved for the Top 20 and Top 30 results. With respect to the three remaining tasks, the performance improved over text search and collection-provided metadata, indicating that consistent and complete reading level metadata was important to identify relevant resources. However, in all five tasks, the retrieval performance of the KB with complete reading level metadata improved over the collection-provided metadata, indicating that incomplete and inconsistent reading level metadata was more detrimental to retrieval performance than not entering any reading level information. Collection managers should seek to enter complete and consistent reading level metadata for all resources to maximize retrieval performance.

Thus, this study demonstrated that consistent and complete reading level metadata was important to filter relevant resources from a digital library collection. In all five tasks, the retrieval performance using complete reading level metadata exceeded the human-expert provided metadata; in four out of five tasks, the performance increased over text search. This study showed that complete and consistent reading level metadata was necessary to retrieve relevant resources. Digital librarians should enter complete and consistent metadata with each new resource to improve KB retrieval performance.

ACKNOWLEDGMENTS

We would like to thank the curators of the NSDL collections that provided the evaluation dataset.

REFERENCES

1. Burke, R. (2002) Hybrid Recommendation Systems: Survey & Experiments. *User Modeling and User-Adapted Interaction*.
2. Callan, J., Smeaton, M., Beaulieu, P., & Brusilovsky, M. (2003) Personalization and Recommender Systems in Digital Libraries. *Joint NSF-EU DELOS Working Group Report*, August, 2003.
3. Chall, J. & Dale, E. (1995) *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.
4. 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.
5. Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998) Inductive Learning Algorithms and Representations for Text Categorization. *7th ACM International Conference on Information and Knowledge Management*, 148-155.
6. Ghani, R. & Fano, A. (2002) Building Recommender Systems using a Knowledge Base of Product Semantics. *Workshop on Recommender Systems and Personalization in Ecommerce at the 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*.
7. Joachims, T. (1998) Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *10th European Conference on Machine Learning*, 137-142.

8. Kincaid, J. P. et. al. (1975) Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel.
9. Kyriakopoulou, A. & Kalamboukis, T. (2008) Combining Clustering with Classification for Spam Detection in Social Bookmarking Systems. *ECML 2008*.
10. McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Rashid, A., Konstan, J., & Riedl, J. (2002) On the Recommending of Citations for Research Papers. *ACM 2002 Conference on Computer Supported Cooperative Work (CSCW 2002)*, 116-125.
11. Page, E. (1994) New Computer Grading of Student Prose Using Modern Concepts and Software. *Journal of Experimental Education* 62(2), 127-142.
12. Salton, G., Wong, A., & Yang, C. (1975) A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 18(11), 613-620.
13. Siolas, G. & d'Alche-Buc, F. (2000) Support Vector Machines Based on a Semantic Kernel for Text Categorization. *IEEE-INNS-ENNS International Joint Conference on Neural Networks*, 205-209.
14. Tsai, K.H., Chiu, T.K., Lee, M.C., & Wang, T.I. (2006) A Learning Objects Recommendation Model based on the Preference and Ontological Approaches. *Sixth IEEE International Conference on Advanced Learning Technologies (ICALT)*, 36-40.
15. Will, T., Srinivasan, A., Im, I., & Wu, Y-F. (2009) Search Personalization: Knowledge-Based Recommendation in Digital Libraries. *Americas Conference on Information Systems*.
16. Yang, Y. & Liu, X. (1999) A Re-Examination of Text Categorization Methods. *22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, 42-49.