AUTOMATIC LEARNING OBJECT CATEGORIZATION FOR INSTRUCTION USING AN ENHANCED LINEAR TEXT CLASSIFIER

THOMAS GEORGE KANNAMPALLIL

*School of Information Sciences and Technology, Pennsylvania State University, University Park,
Pa 16802, USA*
thomasg@psu.edu

ROBERT G. FARRELL

Next Generation Web Dept, IBM T.J. Watson Research Center, Hawthorne, NY 10532, USA
robfgarr@us.ibm.com

This paper explores the use of a machine learning algorithm to automate the task of classifying learning materials into categories useful for instruction. A collection of documents was segmented manually into independent learning objects. A regularized linear text classifier was trained to recognize four topic categories and eleven instructional use categories using manual category labels as training data. The classifier was able to categorize text-based learning objects into topic categories with high accuracy, but initial performance for instructional use classification was poor. An enhanced classifier was able to distinguish between conceptual and procedural categories of instructional use with high accuracy.

1 Introduction

The World Wide Web (WWW) is increasingly used as a learning tool because of the wealth of information available on it. While search engines have been developed for the Web (Maudlin, 1998) and produce good results, there is an increasing number of specialized search engines for learning materials. These search engines improve browsing and retrieve through the use of learning object metadata. The Institute for Electronic and Electrical Engineers (IEEE) has defined an international standard for learning object metadata (IEEE LOM, 2002) that is gaining acceptance in both corporate and university training and education communities. Increasingly, authoring tools allow content developers to enter metadata that fits this standard. However, manually entering metadata is a time-consuming process. Automated techniques are required for wider adoption of the standard.

The IEEE learning object metadata standard defines several attributes that may be assigned to each learning object. We investigated whether two of these attributes could be derived automatically: classification and learning resource type. In many applications, the classification is by topic, so we concentrated on how to assign learning objects to one of several distinct topics (for e.g. topics like “java” or “web services”). The learning resource type attribute describes the kind of learning object, such as an “exercise” or a “simulation”. We used a specialized vocabulary of eleven learning resource types focused on the “instructional use” of the learning object in web-based courseware (Farrell et. al., 2004). These categories include “introduction”, “concepts”, and “procedures”.

The topic and instructional use for a learning object are central to the operation of learning tools and applications, such as the second author's larger effort directed at dynamic assembly of learning objects (Farrell et. al, 2004). Automating the process of assigning topics and instructional use categories for new text-based learning materials could aid in building a large repository of learning objects for retrieval and assembly. Unfortunately, standard metadata editing tools do not automatically supply a value for these critical attributes.

2 Procedure

In this work, we attempted to automatically infer topic and instructional use categories using machine learning techniques, based upon existing manual assignments made by experts. We chose a high performance text classification system developed at IBM (Zhang & Oles, 2001).

We defined a small set of information technology topic categories based on interviews with a domain expert and two instructional experts on our team defined a set of instructional use categories. A team of independent subject-matter experts then assigned learning objects to both sets of categories. We then used these category assignments to train a linear text classifier. We chose three instructional use categories (procedures, concepts, and introduction) and four topic categories (j2ee, web services, websphere, and security) for our experiments, based upon the availability of sufficient training data assigned to these categories.

The basic experimental set up is shown in figure 1. Technical reference documents were collected and stored in the DocBook XML format (Docbook). PowerPoint™ presentation documents were collected and saved as HTML.

Figure 1: Experimental Setup for Automatic Learning Object Text Classification

First, text-based learning objects were manually extracted from reference books and informational presentations and tagged with topic and instructional use categories (Farrell et. al. 2003). Next, the resulting category labels were saved in a learning object metadata (LOM) XML file. A random subset of the learning objects with associated text content

and category labels were then passed to a regularized linear text classifier for training. The trained classifier was then used to automatically categorize the remaining learning objects.

3 Results from Topic and Instructional Use Classification

Sixty-eight learning object documents in each of the four categories (“websphere”, “j2ee”, “web services” and “security”) were used for this experiment. The linear classifier assigned 83 percent of the documents to one of these four categories correctly. When the percentage of training documents was increased to 80 percent, the classification accuracy was 94 percent, as would be expected. This level of accuracy is similar to that reported in the literature for subject-oriented text classification of news articles and other types of documents.

For instructional use classification, the entire collection was utilized. When training with 20 percent of the documents in each category, 52.5 percent of the remaining documents were classified correctly. The experiment was repeated using 80 percent of the documents with a resulting classification precision of 57 percent. The resulting improvement in precision is to be expected, given the additional training data. However, the low classification accuracy, even with 80 percent of the data used for training, was not expected. We attributed this result to an insufficient number of training documents in several of the instructional use categories. A second experiment was conducted for just “procedures” and “concepts” categories using 60 documents from each group. Training with 20 percent of the learning objects yielded 68 percent classification accuracy and this improved to 74 percent when using 80 percent of the documents for training. The result from this experiment is shown below in figure 3(a). While this is a large improvement, it is still a lower classification precision than desired.

4 Enhancing Instructional Use Classification

We next investigated if we could boost the precision of the text classifier for instructional use categories by improving the weights of certain features. In particular, we artificially boosted the frequency of “important” words in the document text. We computed important words by three different techniques: highest frequency within a document corpus, words appearing in document titles, and the words appearing as synonyms of words in the category name from a predefined list appearing in WordNet, a widely available research thesaurus (Miller, 1985; WordNet). The experimental set up for the enhanced classifier is shown in figure 2. The basic experimental set up is similar to figure 1. We identified two mechanisms for enhancing the instructional use classification of the different categories:

1. To improve the accuracy of classifying learning objects as procedures vs. concepts, we boosted salient words from the text corpus within each of these two categories.

2. To improve the accuracy of classifying learning objects into the “introduction:” category, we repeated the salient words from the learning object’s title and added words from the WordNet thesaurus.

Figure 2: Enhanced Instructional Use Classification

4.1 Investigating “procedures” vs. “concepts” Instructional Use Classification

For the “procedures” vs. “concepts” classification, a training file was first created using hand-coded data. Then, the top salient words were extracted for each category. First, all the words of the documents assigned to the category were extracted after removing stop words and performing stemming. Next, the number of occurrences of each resulting term was computed per category as a percentage of the total number of terms in that category. For the same term, the term frequency was calculated based on the entire corpus of terms in all categories. Thus for each term in a category, a salience value can be computed as follows:

$$S(t) = (TF_c(t) / n) - (TF_d(t) / N)$$

In this equation, $S(t)$ = salience value of a term t ; $TF_c(t)$ – Term Frequency of a term t in the whole corpus; $TF_d(t)$ – Term Frequency of the term t in a category; N – Number of terms in the entire corpus; n – Number of terms in the category

The salience value, $S(t)$, was computed for the top hundred words in each category and these terms were then arranged in a decreasing order of salience. A positive salience value indicates that the particular word has a value greater than the normalized frequency in the corpus. The top 100 words of the category procedure were extracted. Words from

this list appearing in the training data were boosted by repeating words multiple times (up to $n = 10$ times). Example set of salient words from the “procedures” category included - *Task, execute, manage, develop, enter, click, next, resource, process, select*. Example set of salient words from the “concepts” category included - *Application, Service, Theme, Role, Provide, System, Contain, Specifications*.

4.2 Investigating using title words for “introduction” category

For the “introduction” category, salient words were extracted from the text based on the frequency of the words appearing just in titles. The top 10 words were included in the testing file multiple times (up to a maximum of $n = 10$ times). A sample set of words in this category included words like “start”, “overview”, and “begin”. Along with this, the WordNet thesaurus was used to retrieve words similar to the word “introduction”. For example, WordNet words for “introduction” include “start, begin, launch, debut”. A combination of the title words and WordNet words were included in the training document to boost these words. Each word that appeared in the WordNet list or the title was repeated five times. WordNet gave different “senses” of each word, and these were each repeated ($n = 5$ times) in the testing file created.

4.3 Results and Evaluation

Based on the enhanced feature set and feature boosting, classification of procedures vs concepts for 66 documents using 20 percent training yielded 74 percent classification accuracy. Increasing the training percentage to 80 percent resulted in 84 percent classification accuracy. By enhancing the feature set by repeating the salient words in a document (up to $n = 10$ times), the precision (classification accuracy) was also increased. The results are plotted in Figure 3 (b).

Figure 3: Instructional Use Classification (66 documents each of Procedures and Concepts) (a) before enhancements (b) after enhancements

The enhancement using the title words and WordNet words did not result in any change in the performance efficiency for “procedures” vs. “concepts” and also did not improve classification for the “introduction” category. Upon further investigation, we determined that the WordNet thesaurus terms were too infrequent in our corpus. In addition, orthographic and grammatical variations may have created an artificially low term overlap. We also found that repeating terms in the section or slide titles also did not

boost classification accuracy. One explanation is that words like “overview” and “introduction” appeared in titles, but they also appeared with some frequency in the rest of the document text.

5 Conclusions

This paper described the development of an enhanced linear classifier for performing topic and instructional use classification of learning objects. Though instructional use is important for learners, it is an extremely difficult classification problem. We were able to improve the precision of instructional use classification by weighting salient category-specific corpus terms. Our experiments indicate that the repeated use of “stylized verbs” in procedural texts can be used to accurately differentiate conceptual vs. procedural learning materials. Our technique is an initial step toward fully automatic and accurate tagging of learning materials with standardized learning object metadata.

References

- Docbook, Available at: <http://www.oasis-open.org/docbook/>, Accessed on March 14, 2005.
- Farrell, R. et.al. Learner-driven Assembly of Web-based Courseware”. Proceedings of E-Learn 2003 (Phoenix AZ, Nov 2003).
- Farrell, R. G., Liburd, S. D., and Thomas J. C. (2004). Dynamic assembly of learning objects. WWW 2004 (Alternate Track Papers & Posters) 2004, 162-169.
- IEEE LOM 2002, IEEE Standard for Learning Object Metadata, IEEE Standards Department, Institute of Electrical and Electronic Engineers, Inc. (2002).
- Maudlin, M. (1998). A history of search engines. Available at: <http://www.wiley.com/compbooks/sonnenreich/history.html>, last accessed on March 14, 2005.
- Miller, G. A. (1985). WordNet: a dictionary browser, In Proceedings of the First International Conference on Information in Data, University of Waterloo, Waterloo.
- WordNet, Available at: <http://wordnet.princeton.edu/cgi-bin/webwn>, last accessed on March 14, 2005
- Zhang, T., Oles, F.J., (2001). Text categorization based on regularized linear classification methods, Information Retrieval, 4, 5-31.