[MHRS07] *Marek Meyer, Alexander Hannappel, Christoph Rensing, Ralf Steinmetz;* **Automatic Classification of Didactic Functions of e-Learning Resources**; Proceedings of the 15th international conference on Multimedia, ACM Press, September 2007, S. 513-516

Automatic Classification of Didactic Functions of e-Learning Resources

Marek Meyer[1,2], Alexander Hannappel[2], Christoph Rensing[2], Ralf Steinmetz[2]

[1)] SAP AG
SAP Research CEC Darmstadt
Bleichstraße 8
64283 Darmstadt, Germany
+49 6227 7 68822
marek.meyer@sap.com

[2)] KOM Multimedia Communications Lab
Technische Universität Darmstadt
Merckstraße 25
64283 Darmstadt, Germany
+49 6151 16 6888
{rensing, steinmetz}@kom.tu-darmstadt.de

ABSTRACT

Re-use of digital resources is an important issue in e-Learning scenarios, because only intensive re-use can make e-Learning cost efficient. Besides reusing whole courses, authors often desire to re-use fine grained parts of courses for creating new Learning Resources. The granularity which appears to be most promising for this kind of re-use is the level of information objects. Information objects each have a dedicated didactic function; a set of information objects with different didactic functions are combined into Learning Objects. This paper analyzes how didactic functions of existing information objects can be automatically classified using machine learning technology. The results of such classification methods on a set of Learning Resources from medical science are discussed.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Indexing methods*

K.3.1 [**Computers and Education**]: Computer Uses in Education – *Distance Learning*

General Terms

Algorithms, Measurement

Keywords

E-Learning, Didactic Classification, Learning Object, Information Object, Metadata

1. INTRODUCTION

E-Learning and especially Web-Based Trainings (WBT) have become an important instrument for improving efficiency in some learning scenarios. WBTs provide the possibility to learn anywhere and anytime the learner wishes. Companies can save money by providing WBTs to their employees instead of sending them to external training courses. A common interchange format for WBTs is the Sharable Content Object Reference Model (SCORM) [1].

But the efficiency of such e-Learning activities depends on the costs of the necessary materials. High-quality educational resources can require up to several thousand Euros per learning hour. The production costs can be significantly reduced if existing materials are re-used as building blocks for the new content. However, there is a granularity paradox in re-use: Producing large units (e.g. whole courses) provides the best usability in the first place, but decreases the probability that this large unit be re-used in another context. Small modular content would be best for re-use but is less suited for the original learning context.

E-Learning courses are commonly hierarchically structured, consisting of multiple granularity levels. Hence there is a chance that some parts of a course at a finer granularity could also be re-used in other courses. The granularity level that is most promising for re-use is the level of so called information objects. Information objects are small elements (typically one or a few screen pages) that each have a dedicated didactic function, such as an overview, a theorem, an example or a test. A large number of possible didactic functions are described by Meder's didactic ontologies [2]. Multiple information objects are aggregated to form a learning object, which is suited to achieve a particular educational objective.

Reusing an information object requires that it can first be found. An author should be able to search selectively for particular types of information objects. Hence it is necessary that each information object is labeled with its didactic function type. Unfortunately, authors tend to maintain metadata very sparsely; didactic function types are rarely available. Automatic metadata generation is an umbrella term for different methods to automatically create missing metadata, e.g. by means of machine learning technology. Up to now, classification of didactic functions has not been addressed by existing metadata generation methods.

This paper presents recent work concerning automatic classification of the didactic functions of information objects. The requirements for performing such a classification task using machine learning technology are analyzed in Section 2. Section 3 presents the setup of our classification system and Section 4 discusses the achieved experimental results. An outlook for future work on didactical classification is given in Section 5.

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

2. CLASSIFICATION AND FEATURES

Classification is the task of assigning objects to predefined categories. Many researchers have studied how classification can be performed automatically by computers. Most automatic classification approaches are found in the area of machine learning [5]. Machine learning describes all algorithms that learn behavior (e.g. how to classify an object) based on training information [4]. Typical methods are: support vector machines (SVM), decision tree learners, Bayes classifiers or artificial neural networks. All these machine learning algorithms have in common that they need a training corpus of objects with known category membership. After a training phase new objects can be classified as well. Classifiers do not take complete objects as input but require mapping the objects to a set of features. Typical features of text documents, for example, could be occurring words; but other attributes, such as document size or average length of sentences are also imaginable. In the case of multimedia content (e.g. images or videos), more sophisticated features are needed – for instance color histograms, thickness of lines or detected objects.

Many classification systems for text-based documents rely solely on textual features. Textual features can be divided into simple statistical information (such as word occurrences) and natural language analysis. Examples of the latter would be lexical chains, word sense disambiguation or grammatical mood.

Classification of information objects is comparable to the task of genre detection [7] – both classify not the subject but rather another dimension of the document. Hence, using only textual features overrates the subject dimension of a document. For didactic classification additional features are potentially useful. Besides textual features, Web Based Trainings also contain multimedia aspects, which are likely to differ between different didactic functions. For instance, the presence of interactive media, such as flash animations or usage of scripting languages (JavaScript) could be an indicator for assessments or demonstrations, whereas they are less likely to appear in other information object types.

Several possible features have been identified for the intended classification task. They have been categorized into linguistic features, recurring structural patterns and hypertext features.

Linguistic features could be the total text length, occurrence of key terms, headlines and sentence types. Recurring structural patterns are the position of an information object within the tree structure of a SCORM package or special knowledge about patterns in other courses from the same author or authoring tool. Hypertext features are structural similarity of HTML documents, referenced style sheets, in-link analysis and embedding of interactive media contents or scripts.

3. EXPERIMENT SETUP

An experiment was set up to evaluate whether multimedia features can be used for classification of didactic functions. E-Learning courses from two sources have been used. One source was the kMed project, which is a joint project of several medical university chairs in Germany that has produced courses for medical students [6]. Other samples were taken from the Content Sharing project [3]. Thus, the samples have been created from two different authoring environments and by multiple authors. The sample courses have been split into information objects. In total, 166 information objects were used for training and 207 samples for performance measurement.

Each information object was manually labeled with its didactic function. The available didactic function types were taken from Meder's didactic ontologies [2]. According to the didactic ontologies, the function types are hierarchically ordered on three levels of detail. The first level of detail differentiates between receptive knowledge types and interactive assessments. Receptive knowledge types are further subdivided into source knowledge, orientation knowledge (facts or overview), explanation knowledge (what-explanation or example) and action knowledge (checklist or principle). Interactive assessments are either multiple choice tests or assignment tests.

For the implementation of the classification task the free classification framework weka [8] was used. Four different classifiers were evaluated: a Bayes network classifier (Bayes), a support vector machine (SVM), a rule based learner (JRip) and a decision tree learner (C4.5). Human judgment was chosen as the baseline for comparing the automatic classifiers against. Six people were asked to manually classify the given samples.

Nine different features were selected for the experiment. These features take into account not the pure text but rather multimedia aspects. Furthermore, most of the features are independent of the particular course language. The only textual feature is the headline keyword class. For this feature, particular decisive words that may occur in a headline are mapped to one of a set of keyword classes. The features are listed in Table 1.

Table 1. Selected features for classification.

Feature Name	Description
WORD_COUNTER	Length of the text
JS_COUNTER	Number of JavaScript functions
CONTAINS_LIST	HTML code contains at least one list
CONTAINS_FORM	HTML code contains forms
CONTAINS_INPUT	HTML code contains input elements
CONTAINS_CHOICE	HTML code contains choice elements
CONTAINS_INT	HTML code contains interaction elements
CONTAINS_SWF	Flash animations are embedded
HEADLINE_KW	Significant keywords that have been found in the page headline

The didactic ontology has three levels of detail. The performance of a classifier can be measured for each of the levels. Furthermore, it is also possible to classify hierarchically. A first classifier decides only which top-level category an information object belongs to, the second classifier decides at the middle level and a third classifier categorizes only on the highest level of detail. Each classifier uses the result of the previous classifier as additional feature of the object being classified. The experimental setup was arranged to allow both flat and hierarchical classification.

4. RESULTS AND DISCUSSION

As described in the previous section the chosen classifiers were trained with a training corpus of 166 information objects. 207 further information objects have been available as test corpus for classification performance evaluation. Most of the experiments were set up as single-label classification; that is, each information object is assigned to exactly one category. Six people were asked for their judgment in order to obtain a baseline. The classifiers are evaluated on three levels of detail according to the three levels from Meder's didactic ontologies.

Different measurement methods for performance exist. Typical indicators are precision, recall, f-measure and accuracy values. These values have varying relevance depending on the number and size of classes and other parameters. Because most of the experiments in this section involve multiple categories, accuracy has been chosen as main performance measure. Accuracy is calculated as

$$Accuracy = \frac{|correctly\_classified\_samples|}{|number\_of\_samples|}$$

The first experiment was classification at the lowest level of detail; meaning the classifiers must decide only if a given information object is either a knowledge type or an assessment. The experiment was performed first with all nine features and afterwards with a selection of six features: JS_COUNTER, CONTAINS_FORM, CONTAINS_INPUT, CONTAINS_INT, CONTAINS_SWF and HEADLINE_KW. All four classifiers performed the task with an accuracy of 100 % after feature selection. Without feature selection, the Bayes and JRip classifiers achieved only 99 %. This high accuracy was surprising, but became clear after a closer look at the information objects: all assessments contain markup elements that enable user interaction, whereas most knowledge types do not have these elements. Thus, the markup-based features chosen are very decisive for distinguishing between knowledge types and assessments.

The next experiment was classification at the second level of detail. On this level, there are two different types of assessments and three different knowledge types. These five level-two types are used as categories. It is assumed that no information about the lowest level of detail is known. First, the classifiers were trained with all nine features. In a second run, only three selected features were used as input: CONTAINS_FORM, CONTAINS_CHOICE, HEADLINE_KW. The evaluation results are compared in Table 2.

Table 2. Classification of second level of detail.

	Bayes	SVM	JRip	C4.5	Human Judgment
Accuracy (all feat.)	0.579	0.613	0.585	0.618	0.787
Accuracy (selected features)	0.609	0.613	0.604	0.614	

First of all, the performance of human judgment is noteworthy. Apparently, the sample information objects could not be assigned as clearly to one of the knowledge types as theory suggests. Human judgment achieved an accuracy of only 78 %. This value also has another implication: Retrieval systems should consider that different users, who are looking for the same information object, may search by different attribute values.

The C4.5 classifier showed the best performance compared to the other classifiers, both with all features and with a reduced feature set. The Bayes and JRip classifiers improved by reducing the number of features, whereas the C4.5 classifier slightly degraded. The differences between the four classifiers shrunk after feature selection.

The results of the feature selection indicate that the two types of assessments can be distinguished by different types of interactive HTML markup. But markup is not significant for differentiating different knowledge types; of **all** examined features headline keywords are most expressive. Future experiments should find out whether linguistic features may result in a better performance.

The next experiment evaluated the performance of classification for the highest level of detail. This level of detail consists of eight classes. A flat classification is assumed; meaning no classification information from the other levels of detail is known. The performance results are denoted in Table 3.

Table 3. Flat classification of highest level of detail.

	Bayes	SVM	JRip	C4.5	Human Judgment
Accuracy (all feat.)	0.396	0.382	0.367	0.430	0.618
Accuracy (selected features)	0.391	0.381	0.353	0.454	

These results are much worse than those of the second level of detail. Even the best classifier C4.5 has achieved only **43 %** using all features, which is almost 20 % below the human baseline.

Feature selection slightly improves the performance of the tree learner: using only the features WORD_COUNTER, CONTAINS_FORM, CONTAINS_CHOICE and HEADLINE_KW raises the accuracy value to **45** %. This accuracy is still too low for practical application and needs to be improved.

An approach for increasing the performance is hierarchical classification, which means that there is a separate classifier for each level of detail. Each classifier uses the category information from the lower level of detail as additional feature.

Table 4. Hierarchical classification of highest level of detail.

	Bayes	SVM	JRip	C4.5
Accuracy (all features)	0.700	0.680	0.667	0.660
Accuracy (select. features)	0.705	0.686	0.686	0.676

Hierarchical classification was tested at the highest level of detail having the category from the second level available as known input. First, all features were used. Afterwards, only the features WORD_COUNTER, CONTAINS_LIST, HEADLINE_KW and CLASS2 (known second-level category) were selected for determining the third level of detail.

This time, the Bayes network showed the best performance both for all features and for just the selected features. In both cases an accuracy of 70 % has been reached.

All the above measurements were performed using single-label classification. However, the human judgment demonstrates that information objects often can not be clearly assigned to a single category. Assigning an object to more than one category is called multi-label classification.

Multi-label classification requires different approaches for both the classification algorithm and the performance measurement [9]. A common approach for multi-label classification is to employ classifiers, which calculate probabilities for each class, such as Bayesian networks. All categories are ranked according to the probability that a given object belongs to that category. A number of categories is then chosen using a few strategies. Three different selection strategies were evaluated: *second-best*, *p-ratio* and *p-difference*: *second-best* selects both the best category and also the second best category, as long as its probability is above a certain threshold; *p-ratio* selects all categories above a threshold relative to the best result; and *p-difference* selects all categories whose probability is not more than p worse than the best match.

Multi-label classification was applied as a final experiment on the second level of detail to increase the recall values at the expense of precision. The best selection strategies, *second-best* and *p-ratio*, achieved an accuracy rate of 85 %, with macro-average precision decreasing to 50.5 % (micro-average: 49.5 %). If the classification result is used only for searching, the multi-label approach is a reasonable approach: the higher recall value implies that more than four-fifth of all information objects can correctly be found by their didactic function, whereas the lower precision only adds some inaccurate objects to the result list.

5. CONCLUSIONS

Re-use of e-Learning resources for authoring new contents is often desired at the granularity level of so called information objects. In contrast to larger units, these information objects have dedicated didactical functions. Current automatic metadata generation methods are not able to classify an information object according to its didactic function.

This paper has made a contribution towards didactical classification of information objects using machine learning technology. First, different types of features, which might be relevant for this task, have been identified. Then a series of experiments have been presented, where, in particular, multimedia features – such as markup or embedded interactive contents – have been used for automatic classification.

According to the chosen categories from a didactic ontology, the classification performance has been evaluated at different levels of detail. The coarsest level of detail only differentiates between receptive knowledge types and interactive assessments. At this level a classification accuracy of 100 % could be achieved. However, this performance was partially due to the special characteristics of the evaluated data sets.

On finer levels of detail the accuracy decreased. On the second level of detail the accuracy amounted to about 61 %. On the third and finest level of detail only 45 % accuracy could be achieved. This value could be raised to 70 % by applying hierarchical classification.

Besides the single-label classification, multi-label classification has been identified as an alternative to improve the retrieval of information objects. The evaluated multi-label classification methods provide more than one category assignment for some of the information objects. This leads to improved recall values at the expense of lower precision values. Using this method, accuracy rose to 85 % at the second level of detail.

These experiments have proved that automatic didactic classification of information objects is possible and that multimedia features – especially markup information – are suitable. However, the performance requires significant improvement for use in practical applications. This might be achieved by taking into account further types of features. Two types of features, which were not used in the discussed experiments, appear to be especially promising. The first additional feature type is the position of an information object within the learning object or course it belongs to. An argument for using the position as a feature is that the arrangement of information objects is often influenced by an author's intended learning strategy. The second promising feature type is linguistic information; the style of speech of didactic texts varies depending on the particular didactic intention. Thus, linguistic features may complement the feature set to achieve a higher performance.

6. REFERENCES

[1] Advanced Distributed Learning. *Sharable Content Object Reference Model (SCORM) 2004*, http://www.adlnet.org/.

[2] Meder, N.: Didaktische Ontologien. *Globalisierung und Wissensorganisation: Neue Aspekte für Wissen, Wissenschaft und Informationssysteme*, 2000, 401-416.

[3] Content Sharing project. http://www.contentsharing.com.

[4] Mitchell, T. M.: Machine Learning, McGraw-Hill Higher Education, 1997.

[5] Sebastiani, F.: Machine Learning in Automated Text Categorization. In *ACM Computing Surveys*, Vol. 34, No.1, March 2002, pp 1-47.

[6] k-MED Knowledge in Medical Education, http://www.k-med.org (last accessed: 05/2007)

[7] Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Text genre detection using common word frequencies. In *Proceedings of the 18th International Conference on Computational Linguistics*, 2000, 808 - 814.

[8] Weka Machine Learning Project, http://www.cs.waikato.ac.nz/~ml/ (last accessed: 05-2007)

[9] Hsu, C.W., Lin, C.J.: A comparison of methods for multiclass support vector machines. In IEEE Transactions on Neural Networks. Vol. 13, No. 2, 2002, 415-425.