Learning to Identify Educational Materials

Samer Hassan and Rada Mihalcea
University of North Texas
samer@unt.edu, *rada@cs.unt.edu*

Abstract

In this paper, we explore the task of automatically identifying educational materials, by classifying documents with respect to their educative value. Through experiments carried out on a data set of manually annotated documents, we show that the generally accepted notion of a learning object's "educativeness" is indeed a property that can be reliably assigned through automatic classification.

Keywords

learning objects, educational applications, text classification

1 Introduction

With the rapid growth of the amount of information available online and elsewhere, it becomes increasingly difficult to identify documents that satisfy the user needs. Current search engines target broad coverage of information, at the cost of providing limited support for well defined verticals.

In particular, an increasingly large number of users, consisting primarily of students, instructors and self-taught learners, are often seeking educational materials online, to use as standalone instructional materials or to supplement existing class resources. The typical solution is to either refer to existing collections of learning materials, which often lack breadth of coverage, or to search the Web using one of the current search engines, which frequently lead to many irrelevant results. For example, as shown later in Section 3, from the top 50 documents returned by a search performed on a major search engine[1] for the query "tree data structure," only four were found to be strongly educative, while as many as 29 documents were found to be non-educative.

In this paper, we address the task of automatically identifying educational materials. We formulate the task as a text categorization problem, and try to automatically classify the "educativeness" of a document (defined as a property that reflects the educative value of a document). Through annotation experiments carried out on a data set of materials from the domain of computer science, we show that the educativeness of a document is a property that can be reliably assigned by human judges. We also identify several features characteristic to educational resources, which can be used to identify the educativeness of a document. We perform a number of classification experiments, and show that the document educativeness can be learned and automatically assigned.

[1] Throughout our experiments, we conduct our searches using the Google search engine.

2 Background

A learning object is formally defined as "any entity, digital or non-digital, that may be used for learning, education or training" [2], or "any digital resource that can be reused to support learning" [12].

The idea that a document can have an educative property is widely accepted in the growing body of work dedicated to learning objects. Learning object repositories (e.g., [6, 8]) target improved access to learning materials through "sharing and reuse," by providing a common interface to entire collections of learning materials that can be shared among students and instructors and can be reused across courses and disciplines. These definitions are representative for the notion of "educativeness" as used in this paper.

While there has been a large body of work focused toward Learning Object Metadata harnessing [7, 4, 1], we are not aware of any work that has tried to harness the power of the Web as an educational resource through the automatic identification of learning assets on the Web. The work closest to ours is perhaps [9], where the authors addressed the problem of finding educational resources on the Web. However, the focus of their work was limited to metadata extraction for a limited set of fine grained properties. Instead, in this paper, we introduce a method to automatically annotate the educativeness property of a document, which can be used to assist learners in their search for educational materials.

It is important to note that the classification of the educativeness of a document cannot be modeled as a genre classification task. While recognizing the educativeness of a document is relatively easy to do with accomplished readers, different educational materials can have major stylistic inconsistencies, which invalidate their membership to a unified genre [3]. For example, a diagram, textbook, and a blog could all serve as useful and educative resources despite their obvious stylistic differences.

3 Building A Data Set for the Classification of Educational Materials

What is an educational material? The purpose of educational materials is primarily decided by the author

International Conference RANLP 2009 - Borovets, Bulgaria, pages 123–127

or the presenter of the resource, who furthermore decides the target audience and the delivery style (e.g., textbooks, presentation, diagram). While the purpose of the resource is a property that is mainly determined by its author, the strength of the educative resource ("educativeness") is a property evaluated cumulatively by the target audience of the resource (e.g., students or educational experts) . Hence, in the construction of our data set and in the evaluations we run, we focus on the educativeness property of a learning resource as determined by the agreement of their potential users (students).

Educational materials can be located in a variety of sources and formats, including lectures, tutorials, online books, blog articles, publications, even technical forums or expert networks. Most of these learning objects typically include several of the following components: definitions, examples, questions and answers, diagrams, and illustrations.

In order to build a data set for the classification of educational materials, we mimic a hypothetical learner who tries to locate and identify learning assets using current online resources. We use a typical search scenario, which involves the use of a search engine with a disambiguated query to identify candidate materials, followed by a filtering step that selects only those materials that have educational relevance.

We collect a data set covering the domain of computer science. We select fourteen topics frequently addressed in data structures and algorithms courses, as shown in Table 2. Starting with each of the fourteen topics, a query is constructed and run against the Google search engine, and the top 60 ranked search results are collected.[2] Note that the meaning of some terms can be ambiguous, e.g., "tree" or "list," and thus we explicitly disambiguate the query by adding the phrase "data structure." By performing this explicit disambiguation, we can focus on the educativeness property of the documents returned by the search, rather than on the differences that could arise from ambiguities of meaning.

3.1 Properties of Educational Materials

We define a set of features largely based on the properties associated with learning objects, as defined in standards such as IEEE LOM [2]. Some of the features are also motivated by previous work on educational metadata [11]. The following features are associated with each document in the data set.

Educativeness
To be able to capture the educativeness of a resource, the annotators had to score each page on its overall educative value. This feature serves as the major class of the documents in the data set. The annotators were instructed to evaluate the resource as a necessary asset for a student to understand the topic, and score each document on a four point scale ranging from "non-educative" to "strongly-educative."

[2] From the top 60 documents, some had to be removed prior to any further processing, because they were either unreachable or they contained non-English characters.

Relevance
We want to measure how human-assigned relevance can contribute to our task, and see if an accurate (manual) measure of relevance can result in a better identification of learning objects, as compared to the search engine ranking. We measure relevance on a four point scale ranging from "non-relevant" to "very relevant."

Content Categories
The content category is a feature that classifies the type of content found on the target page. We assume that the typical content of a learning object can be categorized into one or more of the following types:

Definition: The content presents a textual definition of a concept or any of its associated properties.

Example/Use: The content presents examples that help clarify a concept, demonstrative use of a concept, or the use of operations in that concept. (e.g., the queue data structure push and pop operations)

Questions & Answers: The content presents a question and answer dialogue, as usually found in technical forums and sometime in blog articles.

Illustration: The content presents an illustration of a concept or a process, either through the use of images, or through diagrams.

Other: This group contains all the other types that do not fit in the previous categories.

Resource Type
One of the interesting properties of the learning asset is its source. Under the assumption that the type of the resource can contribute to the document educativeness, the annotators were instructed to choose all the possible types that apply from a pre-compiled list. The list was generated by observing and inspecting the collection of retrieved documents. These types are not mutually exclusive.

Class webpage: A typical class home page where the teacher would provide lecture notes, tips, quizzes and answers for the class homework.

Encyclopedia: A resource for educative materials, representing semi-structured or fully structured knowledge contributed by experts in the field.

Blog: Web log or blog represents an online personal journal. It varies in format and purpose and it is an increasing popular online form of self-expression.

Mailing list/forums: It is a typical example of expert network where users pose their questions to an expert (or group of experts) in the field and receive one or more answers. Usually such content is very technical but not always useful.

Online book: This category represents electronic books in an online format (e.g., HTML, PDF).

Presentation: A demonstrative material that consists of a set of slides or pages, representing the main points to be addressed with respect to a topic.

Publication: This group includes scientific publications, such as journal articles, conference proceedings, article abstracts, and patent descriptions.

How-To article: This source type addresses the use of a specific concept on a step by step basis.

Reference manual: A technical reference or manual, which explains the use and the inner workings of a concept (e.g., Java language documentations).

Other: This category includes all other content (e.g. product catalogs, company homepages)

Expertise
Learning objects are very diverse and are subject to the judgment of the learner. An expert in the field needs little introduction to the topic, and may require a high level of technical insight. Instead, the same information might seem non-educative and irrelevant from the perspective of a novice user, who seeks basic fundamentals. To address this problem, we asked the annotators to indicate their expertise in each of the selected topics on a four point scale.

3.2 Final Set of Features

Taken together, all the features defined above are referred to as "user features," and are listed in Table 1. In addition to these features, for each document in the data set we also collect its search engine ranking and its document type (ppt, pdf, html, doc, etc.). We also calculate the hubness of each page as a ratio of its hyper-linked contents to its original content.

HasDefinition	IsForum	HasExamples
HasQA	IsManual	HasIllustrations
HasOther	IsBook	IsOther
IsHowTo	IsClassWebpage	IsPublication
Rank	Relevance	Hubness
IsBlog	IsPresentation	IsEncyclopedia
DocType	Expertise	Educativeness

Table 1: *User features*

3.3 Agreement Study

Two judges individually annotated the collected documents based on a set of annotation guidelines. The annotators were required to identify the value associated with all the document features described above, along with the educativeness property of a document. The annotators were instructed to evaluate the resource from a college student perspective, therefore discarding highly technical and specific resources as non-educative or marginally-educative.

We measure the inter-annotator agreement by calculating the kappa statistic for the annotations made by the two human annotators. The inter-annotator agreement and the kappa statistic for all the features are shown in Figure 1.

The final data set is created by asking a third annotator to arbitrate the disagreements among the first two annotators. The final distribution across the educativeness class labels is shown in Table 2. As seen in the table, the distribution across educative and non-educative classes is relatively balanced with a few exceptions. Topics such as "queue" and "tree" tend to have more non-educative pages, unlike topics such as "binary search," which tend to have more educative pages.

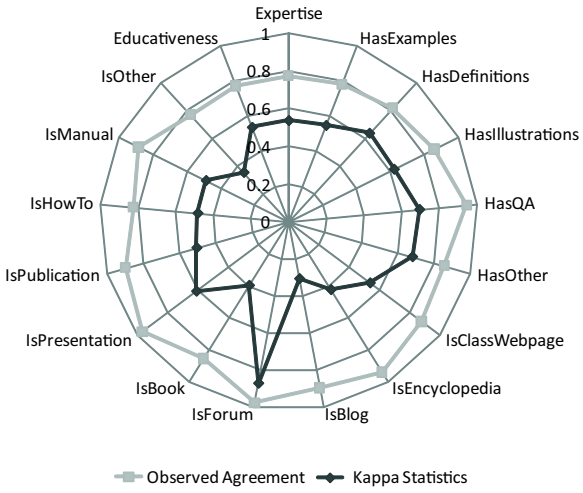

Fig. 1: *Kappa statistic and inter-annotator agreement*

Topic	NE	ME	E	SE	Total
Array	22	13	17	6	48
Queue	23	15	18	4	50
Stack	20	13	20	7	60
Tree	29	11	11	4	55
Linked list	22	10	19	9	60
Skip list	18	3	18	10	49
Heap	21	10	15	6	52
Priority queue	18	7	21	5	51
Hash table	22	9	17	7	55
Dictionary	28	6	17	3	54
Graph	25	9	14	6	54
Sorting algorithms	20	8	22	10	60
Binary search	12	6	28	14	60

Table 2: *Distribution of classes across the topics. Number of non-educational (NE), marginally educational (ME), educational (E) and strongly educational (SE) materials.*

4 Experiments

Using the data set described in the previous section, we experiment with automatic classifiers to annotate the educativeness of a given document. Through these evaluations, we measure the ability of a system to automatically detect and classify documents according to their educative value.

The four-point scale used for the educativeness annotation allows us to perform both a fine grained and a coarse grained evaluation. In the fine grained evaluation, all four dimensions are considered, and thus we run a four-way classification. In the coarse grained evaluation, we combine the non-educative and marginally educative documents into one class (non-educative), and the educative and strongly educative pages into another class (educative), and run a two-way classification. All the evaluations are conducted using a ten-fold cross validation.

Through our experiments, we seek answers to the following questions:

1. Can the content of a document be used to classify its educativeness? We evaluate the use of the doc-

ument content to learn and detect its educativeness. The content is used to construct a feature representation of each document. The terms appearing in the learning objects serve as features in the learning algorithm, with a weight indicating their frequency in the learning object.

2. Are the user-features useful for the classification of a document educativeness? We evaluate the selected user features as possible dimensions to learn and detect the educativeness of target examples. We use all the user features summarized in Table 1 to construct a feature vector representation for each learning asset. Since these features were manually assigned by the annotators, these annotations serve as an upper bound on the accuracy that can be achieved by using such features.

3. Can the content of a document be used to automatically predict the user-features? We run an evaluation where each of the selected user-features serves as its own class. The learning assets in which this feature has been selected by the annotators serve as positive examples, while the documents in which the feature was not encountered serve as negative examples. The content of the documents is used to build the feature vectors. The examples are then used to train a classifier to classify each of the features automatically.

4. Can the automatically predicted user-features be used to learn and detect the educativeness of a document? Finally, given the set of classifiers generated in the previous experiment, we use their output to construct a machine weighted user-feature representation of the given document. This evaluation is similar to the one relying on manually assigned user-features. However, instead of using the user annotations, we use the output automatically predicted by the classifiers.

For the experiments, we used two classifiers: Naïve Bayes[5] and SVM [10], selected based on their performance and diversity of learning methodologies.

5 Results

We run a first experiment where we use the content of the documents, with minimal pre-processing (tokenization, stopword removal), and classify them with respect to the fine-grained and coarse-grained educativeness class. We use a 10-fold cross validation on the entire data set. The rows labeled with "document content" in Tables 3 show the results of this experiment. To answer the first question, these experiments show that the use of raw content is useful and can be effectively used to classify the educativeness of a document. In fact, compared to the baseline of selecting the most common class across all the documents, the content-based classification results in a 22-23% absolute increase in F-measure.

Next, we use the manual annotations for the user-features to classify the educativeness of a document. The results obtained in this experiment are shown in Table 3 in the rows labeled with "user-features (manual)." The results are clearly superior, which answers the second question and suggests the usefulness of these features for the classification of educativeness. Note that these results represent an upper bound for our evaluations, since they rely on manually annotated features.

Features	NB	SVM
Fine-grained		
Document content	53.88	61.25
User-features (manual)	74.33	76.24
User-features (predicted)	58.70	62.38
Baseline	38.63	38.63
Coarse-grained		
Document content	77.00	78.65
User-features (manual)	87.80	88.56
User-features (predicted)	78.78	77.38
Baseline	55.02	55.02

Table 3: *Classification results*

Since the user-features seem to exhibit the best performance, next we evaluate the ability of automatically labeling these features using the content of the documents. The accuracy of the automatic classification of the user-features is shown in Figure 2. Both SVM and Naive Bayes seem to be able to label these features with relatively high accuracy. The lowest performance is achieved for Relevance (50-56% F-measure) and the highest for IsEncyclopedia (86-95% F-measure). This experiment provides an answer to the third question: all the user-features that proved useful for the classification of educativeness can be predicted based on the document content.

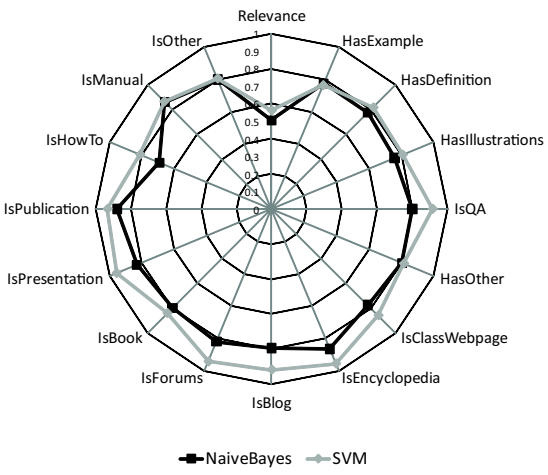

Fig. 2: *Classification results for user-features*

Finally, we answer the fourth question by running an experiment where the automatically predicted user-features are used as input to a classifier to annotate the educativeness of a document. The results obtained in this experiment are shown in Table 3, under the rows labeled "user-features (predicted)." The performance obtained by this classifier shows slight advantage (1-5% absolute increase in F-measure) over the one obtained by using the raw content alone. This indicates that a prediction of high accuracy might help in closing the gap with the upper-bound obtained with

the manually annotated user-features. This result can be the basis for future improvements, by seeking improvements in the classification of the individual features prediction (e.g., by using syntactic or semantic features in addition to lexical features).

6 Discussion

Based on our experiments, we found that the educativeness of a document is a property that can be automatically identified. Not surprisingly, the classification with respect to a set of coarse-grained classes is significantly higher than the fine-grained classification. In terms of features, the raw content of a document was found useful, as were other properties associated with a document (referred to as "user-features").

To evaluate how each of the user-features contribute to the accuracy of the classification, we measured the information gain associated with each feature based on the manual annotations. Figure 3 shows the feature weights. Not surprisingly, the content categories (e.g., HasDefinition, HasExample, HasIllustration) score the highest, indicating their significant discriminative power. Interestingly, the Relevance feature has a higher discriminative power than the Rank feature, which indicates that the relevance of a document might be a good feature to consider when modeling its educativeness. Other intuitive features such as resource types (e.g., IsHowTo, IsPresentation) seem to also contribute to the classification. Note however that the degree of their contribution might be affected by the implicit dependency on content categories (e.g., pages classified as IsEncyclopedia often include definitions, which also activate the HasDefinition feature).

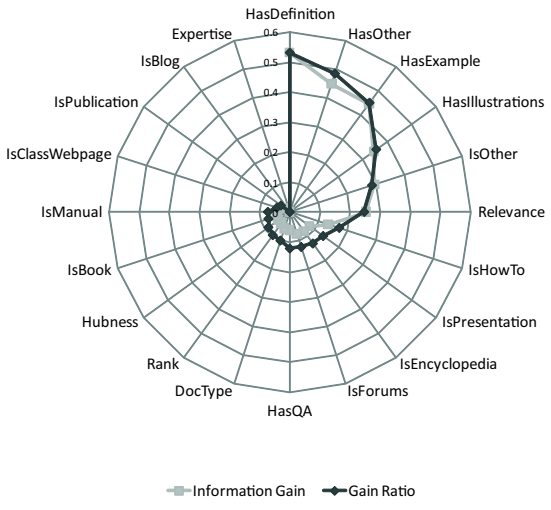

Fig. 3: *Information gain for user-features*

7 Conclusion

In this paper, we addressed the task of automatically identifying learning materials. We constructed a data set by manually annotating the educativeness of the documents retrieved for fourteen topics in computer science. An annotation experiment carried out on this data set showed that the educativeness of a document is a property that can be reliably assigned by human judges. Moreover, through a number of classification experiments, we showed that the educativeness property can also be automatically assigned, with up to 23% absolute increase in F-measure as compared to the most common class baseline.

Through our experiments, we identified several promising lines for future research. First, we plan to explore ways of improving the classification accuracy for the individual user-features, as well as ways of combining them with the features extracted from the content of a document, in order to improve the overall accuracy of the classification of educativeness. Second, we plan to carry out larger-scale experiments to explore the portability across different domains.

The data set introduced in the paper can be downloaded from http://lit.csci.unt.edu/index.php/Downloads

Acknowledgments

The authors are grateful to Carmen Banea and Ravi Sinha for their help with the data annotations.

References

[1] J. Greenberg. Metadata extraction and harvesting. *Journal of Library Metadata*, 6(4):59–82, 2004.

[2] W. Hodgins and E. Duval. Draft standard for learning technology - learning object metadata - iso/iec 11404. Technical report, 2002.

[3] J. Karlgren. The wheres and whyfores for studying textual genre computationally. In *In Proceedings of the AAAI Fall Symposium of Style and Meaning in Language, Art and Music.*, Washington D.C., 2004.

[4] Marek. Categorizing learning objects based on wikipedia as substitute corpus.

[5] A. McCallum and K. Nigam. A comparison of event models for Naive Bayes text classification. In *Proceedings of AAAI Workshop on Learning for Text Categorization*, 1998.

[6] F. Neven and E. Duval. Reusable learning objects: a survey of LOM-based repositories. In *Proceedings of the ACM International Conference on Multimedia*, France, 2002.

[7] L. T. E. Pansanato and R. P. M. Fortes. Strategies for automatic lom metadata generating in a web-based cscl tool. In *WebMedia '05: Proceedings of the 11th Brazilian Symposium on Multimedia and the web*, pages 1–8, New York, NY, USA, 2005. ACM.

[8] S. Smith Nash. Learning objects, learning object repositories and learning theory: Preliminary best practices for online courses. *Interdisciplinary Journal of Knowledge and Learning Objects*, 1, 2005.

[9] C. Thompson, J. Smarr, H. Nguyen, and C. Manning. Finding educational resources on the web: Exploiting automatic extraction of metadata. In *ECML Workshop on Adaptive Text Extraction and Mining*, 2003.

[10] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[11] E. Westerhout and P. Monachesi. Creating glossaries using pattern-based and machine learning techniques. In *Proceedings of the Sixth International Language Resources and Evaluation*, Marrakech, Morocco, 2008.

[12] D. Wiley. *Learning Object Design and Sequencing Theory.* PhD thesis, Department of Instructional Psychology and Technology Brigham Young University., 2000.