International Review of Research in Open and Distributed Learning

Social Web Content Enhancement in a Distance Learning Environment: Intelligent Metadata Generation for Resources

Andrés García-Floriano, Ángel Ferreira-Santiago, Cornelio Yáñez-Márquez, Oscar Camacho-Nieto, Mario Aldape-Pérez et Yenny Villuendas-Rey

Volume 18, numéro 1, février 2017

Special Issue: Advances in Research on Social Networking in Open and Distributed Learning

URI : https://id.erudit.org/iderudit/1066183ar
DOI : https://doi.org/10.19173/irrodl.v18i1.2646

Aller au sommaire du numéro

Éditeur(s)

Athabasca University Press (AU Press)

ISSN

1492-3831 (numérique)

Découvrir la revue

Citer cet article

García-Floriano, A., Ferreira-Santiago, Á., Yáñez-Márquez, C., Camacho-Nieto, O., Aldape-Pérez, M. & Villuendas-Rey, Y. (2017). Social Web Content Enhancement in a Distance Learning Environment: Intelligent Metadata Generation for Resources. *International Review of Research in Open and Distributed Learning*, *18*(1), 161–176. https://doi.org/10.19173/irrodl.v18i1.2646

Résumé de l'article

Social networking potentially offers improved distance learning environments by enabling the exchange of resources between learners. The existence of properly classified content results in an enhanced distance learning experience in which appropriate materials can be retrieved efficiently; however, for this to happen, metadata needs to be present. As manual metadata generation is time-costly and often eschewed by the authors of the social web resources, automatic generation is a fertile area for research as several kinds of metadata, such as author or topic, can be generated or extracted from the contents of a document. In this paper we propose a novel metadata generation system aimed at automatically tagging distance learning resources. This system is based on a recently-created intelligent pattern classifier; specifically, it trains on a corpus of example documents and then predicts the topic of a new document based on its text content. Metadata is generated in order to achieve a better integration of the web resources with the social networks. Experimental results for a two-class problem are promising and encourage research geared towards applying this method to multiple topics.

International **R**eview of **R**esearch in **O**pen and **D**istributed **L**earning
Volume 18, Number 1

February – 2017

Social Web Content Enhancement in a Distance Learning Environment: Intelligent Metadata Generation for Resources

Andrés García-Floriano[1], Angel Ferreira-Santiago[1], Cornelio Yáñez-Márquez[1], Oscar Camacho-Nieto[2], Mario Aldape-Pérez[2], and Yenny Villuendas-Rey[2]
[1]Centro de Investigación en Computación, Instituto Politécnico Nacional, México, [2]Centro de Innovación y Desarrollo Tecnológico en Cómputo, Instituto Politécnico Nacional, México

Abstract

Social networking potentially offers improved distance learning environments by enabling the exchange of resources between learners. The existence of properly classified content results in an enhanced distance learning experience in which appropriate materials can be retrieved efficiently; however, for this to happen, metadata needs to be present. As manual metadata generation is time-costly and often eschewed by the authors of the social web resources, automatic generation is a fertile area for research as several kinds of metadata, such as author or topic, can be generated or extracted from the contents of a document. In this paper we propose a novel metadata generation system aimed at automatically tagging distance learning resources. This system is based on a recently-created intelligent pattern classifier; specifically, it trains on a corpus of example documents and then predicts the topic of a new document based on its text content. Metadata is generated in order to achieve a better integration of the web resources with the social networks. Experimental results for a two-class problem are promising and encourage research geared towards applying this method to multiple topics.

Keywords: social networking, distance learning, social web content, metadata generation, intelligent classification

Introduction

In recent years social networks have grown considerably. By establishing relationships within educational environments, these networks become a much more effective medium of communication among learning

peers. Emerging social networks bear social capital as inherent value, manifested in the common interests that help foster near instant contact between affine members. Users demonstrate a tendency to interact more closely and with higher frequency with people whom they share interests or opinions (Lytras, Mathkour, Abdalla, Yáñez-Márquez, & De Pablos , 2014).

According to Willis, Szabo-Reed, Ptomey, Steger, Honas, Al-Hihi, Lee, Vansaghi, Washburn, and Donnelly (2016), the dramatic growth in technology and online social networks has generated great interest in open and distance learning. In the context of social networks, distance learning environments allow for unlimited access to materials that can be completed at a work rate that is comfortable for the individual and the ability for constant group feedback; behavior change is reinforced by social group support and distance learning strategies.

However, determining the relevancy of a piece of social web content in a distance learning environment is an issue that has not been fully resolved (Min, Shi, Grishman, & Lin, 2012). In order to solve this problem, it is necessary to have vital information about a document's content at one's disposal. Metadata is information about a document that satisfies this need; it describes an e-learning document's properties in an organized way (Kovačević, Ivanović, Milosavljević, Konjović, & Surla 2011).

The term *metadata* is used to refer to all information that describes the properties of objects stored in repositories such as websites. Metadata is the basis of initiatives such as the Semantic Web and has been applied to real-world problems (Castellanos-Nieves, Fernández-Breis, Valencia-García, Martínez-Béjar, & Iniesta-Moreno, 2011). Employing metadata is undoubtedly useful for organizing and searching documents and resources. In spite of the advantages it offers, one of the greatest problems with it is that authors seldom spare time for tagging their creations with appropriate metadata. Even though supplying metadata once offers many advantages in the long run for professors and students, authors more often than not consider it as tedious and non-compulsory extra work. However, having properly tagged resources available in social networks is important for distance learning systems, as users would benefit from being able to find and share relevant, quality material easily and efficiently. The emergence of systems for automatic metadata generation could help solve this problem, minimizing the need for human intervention (Bauer, Maier, & Thalmann, 2010).

Two main approaches exist for automatic metadata generation: *harvesting* and *extraction* (Paynter, 2005). The former refers to pulling metadata out of the document itself, and the latter to assigning or predicting the content of metadata fields based on the document's contents, aided by cues such as text positioning and formatting. The usual approaches to tackling metadata generation are contained in the fields of natural language processing, statistical methods, rule-based systems, and machine learning (Bot, Wu, Chen, & Li, 2004).

Specifically regarding the machine learning approach, there exist two main approaches for generating metadata; that is, using a pre-defined, finite vocabulary of keywords (classification) versus designing a system for automatically detecting them (extraction). This is akin to a supervised learning versus an unsupervised learning (clustering) approach. In the keyword classification or supervised learning approach,

a human operator must tag documents beforehand for them to function as training examples for a machine learning method. While this implies tedious work at the beginning, substantial benefits can be reaped later as a supervised approach can speed up the automatic tagging of resources using a well-defined, controlled, and desired vocabulary.

In contrast, keyword extraction has the upper hand regarding pre-processing time, as it does not require an agent to supervise and tag objects prior to applying the algorithm; keywords are generated automatically based on the set of training documents (Medelyan & Witten, 2006). However, keyword extraction may not yield the desired keywords a human agent would devise, leaving all the work to the algorithm with no way to tell which keywords will result in the end or even if said keywords will be suitable for the problem at hand. Additionally, different unsupervised algorithms will yield different tags when acting on the same documents, leaving questions regarding which set of keywords is the most desirable or accurate.

Our proposal aims to achieve fast, efficient, automatic topic extraction while taking care of selecting appropriate and relevant keywords so that the system is able to correctly discriminate between the possible topics. The means used to achieve this goal are rooted in the field of machine learning: a novel associative classifier. This kind of models have been previously applied in the context of distance learning systems (Chrysafiadi & Virvou, 2013); however, rather than enhancing educational material, most of the related works refer to the modeling of the learning styles of students or to the prediction of academic success in a study program. A branch of Data Mining known as Educational Data Mining (Peña-Ayala, 2014) is oriented towards this kind of problems.

The rest of the paper is organized as follows. The Literature Review section presents related work in order to contextualize our proposal and showcase how the problem has been tackled before. Next, the proposed method for automatic metadata generation is outlined and the pattern classifier used for this purpose is explained. The Results section follows, which explains how the data set used for the experiments was collected, details the experimental design we followed and presents the results obtained by our proposal. The paper concludes with a discussion on the results and proposals for related future work.

Literature Review

Automatic generation of metadata for Web resources is a relatively new focus of research. As we mentioned above, there exists a vast number of approaches to tackle the problem. After a comprehensive review of the available literature, we have summarized below some interesting and representative works dealing with automatic metadata generation.

As mentioned in the introduction, two main approaches exist for metadata generation: extraction and harvesting. Greenberg (2004) compared them by evaluating the capabilities of two applications using the top pages of 29 educational resources obtained from repositories of the National Institute of Environmental Health Sciences. Each of the tools was based on a different approach: the *Klarity* tool made use of extraction methods while its counterpart, *DC.dot*, employed harvesting methods. The study concludes that the

obtained results were satisfactory, though they can be improved by considering issues such as extending the level of analysis of web pages, using web resource genre and textual density to determine which algorithm to use, eliminating character limitations, and harvesting metadata created by humans even though the task is related to extraction.

An alternative to automatically generating metadata, collaborative tagging has also been considered for metadata generation. Bauer, Maier, and Thalmann (2010) analyzed automated and collaborative metadata generators for learning objects. After an exhaustive content suitability analysis and using context-based methods for generating metadata compatible with IEEE LOM, the authors concluded that an approach based on collaborative tagging for metadata generation makes it easier to design a system with a large number of users.

Metadata is helpful not only in the context of distance learning environments, but also in several areas such as enterprise applications. Şah and Wade (2012) proposed a framework for automatic metadata extraction in multilingual enterprise content. Their rationale is that personalization is an increasingly important aspect for enterprises in order to properly reach and satisfy their customers, along with a marked increase in non-English language Web content. To achieve full personalization support, the availability of high-quality metadata is paramount; such metadata describes structural relationships, subject relations between documents, and cognitive metadata about the documents themselves. However, there are several challenges to this goal, such as the inability to scale the manual annotating of large databases. This is compounded with the fact that it is very complex to automatically understand the semantics of the data, as well as the spike in non-English language publishing making previous English-focused metadata extraction systems obsolete. The proposed framework is based on a fuzzy method which employs different document parsing algorithms to extract metadata from multilingual documents. The authors report an average precision above 89% on the knowledge base of Symantec Norton 360.

Some works are related to the modification of well-known standards. Dharinya and Jayanthi (2013) formulated an extension of the IEEE LOM standard for effective retrieval of learning objects; they add elements such as domain ontology and automatic annotation based on a semantically-oriented approach. Their method uses specific indexing for parsing the documents and standard classification algorithms for extraction. Their method further employs concept identification and significance to achieve personalization. The authors found that their proposed indexing method outperformed traditional methods in terms of precision and recall.

Metadata extraction has also been used to identify useful resources for educators. Atkinson, Gonzalez, Munoz, and Astudillo (2014) proposed a metadata extraction method with the goal of identifying helpful educational content. Their model was based on concepts of natural language processing like corpus-based methods and a Naïve Bayes classifier to assign extracted metadata to a topic. This tool combination enabled the proposed model to process semi-structured information from which metadata is extracted. The experimental results showed that the proposed approach is capable of providing quality results to teachers who are looking for rich educational resources. Although the obtained results were satisfactory, it is

important to consider that there exist many formats of Web documents that are difficult to process and interpret, a situation which limits the amount of information available to users.

While Natural Language Processing and Machine Learning models are undoubtedly the most used methods for metadata extraction, Miranda and Ritrovato (2015) showed that it is possible to use alternative tools in order to automatically generate metadata to enhance learning object repositories. Their proposed method extracts metadata directly from the resource file itself and is able to produce the learning object itself. This method makes use of Shannon's information theory, learning models, statistical analysis, and heuristics to generate basic rules about the content of the learning objects and their interactivity type and level; this combination of elements leads to the extraction of the set of metadata. The authors reported very promising results in their experiments, which were performed using 2600 learning objects related to mathematics and computer science. They concluded that the use of metadata extractors is suited for learning object repositories and environments that need to manage large amounts of users and data. Finally, the authors also claimed that their approach could be used as a plug-in for systems that require an improvement in search engine performance.

Automatic metadata generation represents a solution to the scalability problem inherent in metadata: ideally, human agents should tag all content according to their expert knowledge; however, with an ever increasing amount of resources the production of metadata surpasses the human capabilities of quality control. Automatically generating metadata solves the scalability problem but the quality control issue remains. Thus, research has been devoted to assess the quality of generated metadata; for instance, Ochoa and Duval (2009) presented a set of scalable metrics based on the Bruce and Hillman (2004) framework for metadata quality control. Through statistical analysis, the authors found that many of their metrics correlate well with human evaluation.

Method

To take on the problem of metadata generation, we propose the application of a novel pattern classifier to automatically predict the topic of social web content. From the reviewed approaches in the state of the art, metadata generation using a supervised learning or classification approach was chosen. The chosen algorithm for achieving the goal is a novel supervised pattern classifier: the Heaviside Classifier.

To more concisely demonstrate the differentiation potential of the proposed system, the scope of this work has been narrowed down to selected topics that are typically studied during an undergraduate program in Computer Engineering in a distance learning environment. Some of the topics the first test of the method will consider are:

- Programming.
- Electronics.

- Databases.
- Computer Networks.
- Artificial Intelligence.

Proposed System

To date, the first phase of the system's development has been completed. In this phase the problem is addressed as a binary classification problem to assess the viability of the model; that is, two topics are selected and resources are then classified as belonging to one of two possible classes. Once the system is able to attain satisfactory results for a binary classification problem it can be scaled to a multi-topic problem. Two classes were considered in the experiments: documents corresponding to Programming and to Electronics courses. Only documents written in English were considered.

Prior to operating the system, a collection of documents related to the desired topics is gathered to function as training data. Each training document is manually labeled according to its topic. These documents can be of any kind (lecture notes, books, summaries, exercises, or quizzes) as long as they contain plain text. The union of the text content of all the documents in the collection is then analyzed and every unique word is counted, excluding the most common grammatical particles, articles, conjunctions, and prepositions.

A cutoff number n is defined and the n most frequently occurring words in the whole of the collection's text are considered as the training set's dictionary. Then, for each training document, the frequency of each of the words in the dictionary relative to the total number of words in the document is computed. Therefore, a document is represented in vector space by an n-dimensional vector whose i-th component represents the relative frequency of the i-th word in the dictionary for that text sample. This vector has a final component representing the class label; that is, the topic of the document that each vector represents. The classifier is then trained using this information; namely, it learns to distinguish how frequently certain words appear in one kind of documents compared to the other. This information is what allows the classifier to correctly predict the topic of a document.

Once the training set is complete, one can proceed to the classification of new documents. An uncategorized document is presented to the system, which first automatically extracts the text within and then computes the relative frequencies of the words against the same dictionary and using the same methodology as in the training process.

When the new sample's word frequency vector is generated, it is presented to the Heaviside classifier. The classifier then assigns a topic label to the unclassified document based on the prior training data. Depending on this assigned label, a metadata file in the Resource Description Framework (RDF) format is generated. Figure 1 details the classification process.

The crux of this system is the classification stage. While almost any classifier, such as Neural Networks, SVM, statistical methods, or any of the approaches reviewed above could be plugged into this stage, a recently developed associative classifier was implemented.

This model is capable of classifying a pattern in a single step, avoiding convergence issues as is the case with iterative methods. This classifier is based on the Heaviside function as well as on lattice theory, algebraic structures and numerical systems (García-Floriano, Camacho-Nieto, & Yáñez-Márquez, 2015). This differentiates it conceptually from other classifiers in the field of pattern recognition.

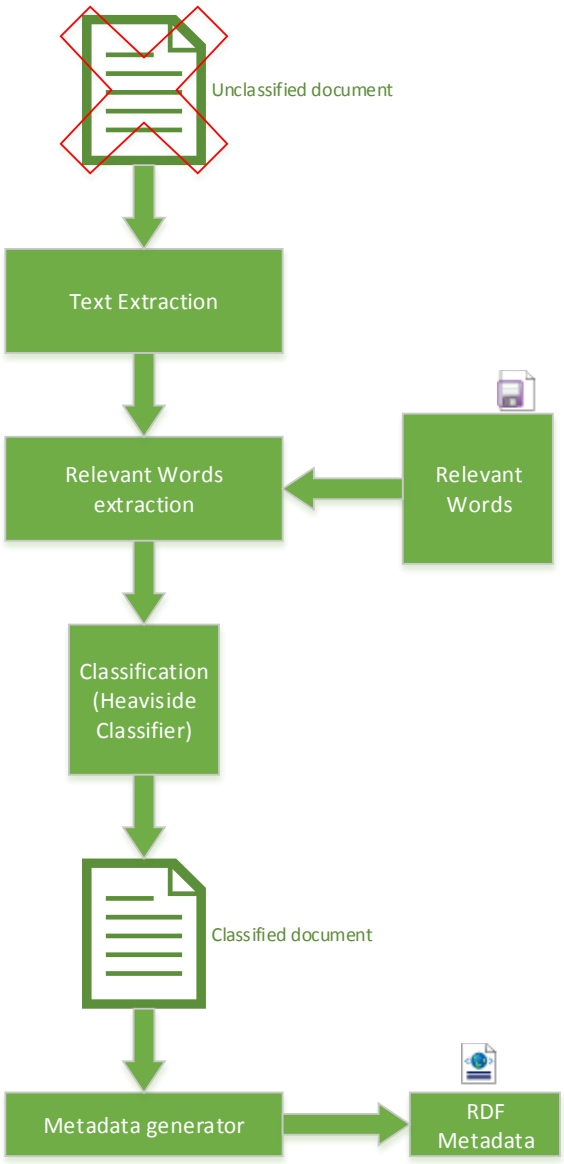

Figure 1. Proposed system.

Heaviside Classifier

The Heaviside Classifier is sustained on the following four hypotheses:

Hypothesis 1: The Heaviside classifier is a supervised learning method. It is assumed that a previously labeled dataset with a finite number of classes and patterns is available. This is achieved with the process explained above, which generates labeled vectors for a two-class problem.

Hypothesis 2: Patterns can be represented as finite vectors whose components are real numbers. If negative numbers exist in the dataset, it is possible to apply a transformation for obtaining only non-negative real numbers. Additionally, it is possible to round or truncate those real, non-negative numbers in order to obtain finite vectors with rational, non-negative components. These numbers have a finite amount of decimal places. This is achieved, as the generated vectors represent word frequencies which are real, non-negative and have a finite number of decimal places.

Hypothesis 3: It is possible to find a scaling factor that allows us to transform these rational numbers into non-negative integers. Since the numerical components of the vectors possess finite decimals, this is possible.

These three hypotheses allow us to transform the original dataset into a set of finite, non-negative integer-valued vectors. From this, the fourth hypothesis is derived.

Hypothesis 4: It is possible to represent each vector component in terms of a positional base-*b* numeral system, where *b* is an integer greater than one. To achieve this, simply compute the component's expansion in base *b*.

Therefore, after transforming the data set to comply with the four hypotheses, the resulting dataset consists of non-negative integer-valued vectors represented in a base-*b* positional system. Thus, the keyword frequencies of the documents are transformed into this representation.

It is important to note that the definition of the Heaviside function used in this work is the one found in the book by Abramowitz and Stegun (1972).

Definition 1: Let x be a real number. The Heaviside function evaluated in x is defined as:

$$H(x) = \begin{cases} 1 & \text{if } x>0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

After an analysis of lattice theory and ordered algebraic structures, we arrived at the conclusion that the Heaviside classifier is operate in two different modes: the *HI* mode, corresponding to the infimum and the *HS* mode, corresponding to the supremum. In this work an application of the Heaviside classifier working in *HI* mode is presented. A further publication shall make use of the *HS* mode of the classifier.

Fundamental operations on which the Heaviside Classifier is based. The proposed model is based on two fundamental operations: L, which will be used in the learning phase, and C, which is used for the classification phase.

It should be remembered at all times that the Heaviside classifier assumes that the patterns are represented as non-negative, integer-valued vectors in a positional base *b* system. Therefore, patterns can be expressed in terms of a base 2 (binary), base 3 (ternary), base 8 (octal), base 16 (hexadecimal) system; any other larger base can be also used.

The *L* operation for the *HI* mode.

Definition 2: Let $A^k{}_i$ and $A^d{}_j$ be the *i*-th and *j*-th components of the *k*-th pattern and the *d*-th pattern of the training set, respectively. Then, for both modes of operation the *L* operation is defined as:

$$L(A_i^k, A^d{}_j) = [H(A_i^k + A^d{}_j + b)] \cdot [(A^k{}_i + b) - (A^d{}_j + 1)]$$

The *C* operation for the HI mode.

Definition 3: Let M_{ij} and $P^k{}_j$ the *ij*-th component of the model of the Heaviside classifier *M* and the *j*-th component of the *k*-th pattern of the test set *P*, respectively. Then, for the *HI* mode of the classifier, the *C* operation is defined as:

$$C(M_{ij}P^k{}_j) = H[(M_{ij} + 2) - (b - P^k{}_j)] \cdot H[(2b - 1) - (M_{ij} + P^k{}_j)] \cdot [(M_{ij} + 1) - (b - P^k{}_j)]$$

Heaviside Classifier algorithm for the HI mode. The following are definitions that constitute the theoretical foundations of the training and classification phases of this model for the HI operation mode.

Definition 4: Outer product under L operation. Let A^k and A^d two patterns in the learning set A in the Heaviside classifier. The outer product under L of both patterns is defined as a matrix whose ij-th component is calculated as follows:

$$L(A^k{}_i, A^d{}_j)$$

Note 1. If the dimensionality of the patterns in A is n, then the outer product under L will be a $n \times n$ matrix.

Note 2. The value of k=d is valid.

Definition 5: Learning Phase. The ij-th component of M, while in HI mode with a training set A is defined as:

$$M_{ij} = \min[L(A^k{}_i, A^k{}_j)]$$

Note. The minimum value is taken with respect to the index k, which has a range from 1 to |A|.

Definition 6: Classification Phase. Let *M* be a Heaviside classifier operating in *HI* mode and let P^d be a pattern from the testing set *P* with dimensionality *n*. The class assigned to this pattern is the class of the pattern recovered upon performing the following operation. This operation obtains the *i*-th component of the recovered pattern *R*.

$$R_i = \max[C(M_{ij}, P^d{}_j)]$$

Note. The maximum value is taken with respect to the index *j*, which has a range from 1 to *n*.

Once the learning and classification phases are implemented, the classifier can be plugged into the system and function as a sub-system that receives a pre-processed document represented by a numerical vector and outputs a document topic.

Results

In this section we present the results of the preliminary tests performed with the proposed system. First, we will outline the experimental design used for evaluating the performance of the proposed model. Then, we detail the construction of the dataset used for training and validating the Heaviside classifier for document topic classification. Finally, this section concludes with the presentation of the results.

Experimental Design

The problem at hand is automatic topic prediction for distance learning content, and the proposed approach to solve this is to implement a pattern classifier which is trained on a corpus of example texts, which renders it capable to predict the topic of an unknown document. The pattern classification model chosen for this task is the Heaviside classifier (García-Floriano, Camacho-Nieto & Yáñez-Márquez, 2015). Our hypothesis is that the Heaviside classifier is capable of a high accuracy in predicting the topic of a document.

The Heaviside classifier has one main tunable parameter which becomes an independent variable for this experiment: the numeral base *b* used to encode the patterns. Different values of *b* were tested in this experiment. The second independent variable stems from the choice between a large and comprehensive or a short but relevant dictionary: the parameter *n* that defines how many of the most frequent words among the training documents are considered for classification. As discussed above, the choice of *n* may impact the performance of the system by introducing several words that give no information about the topic of the document. This parameter was also varied in the experiment to explore this effect.

The dependent variable in the experiment is thus the performance of the classifier over a testing set of documents. Rather than evaluating how well the system predicts the topic of a previously-learned document, it is of interest to examine the generalization power of the proposed model; that is, how well it performs when faced with a new, unknown document and how accurately it predicts the topics of such

samples. Techniques such as K-Fold Cross-Validation or Leave-One-Out (LOO) Cross-Validation (itself a specific case of KFCV) are commonly used in pattern recognition to validate the generalization ability of a classifier. In LOO cross-validation, all but one of the patterns are used to train the classifier and the remaining pattern is then classified. This process is repeated until every single pattern has been used as test data. Translated to the domain of topic prediction, out of the chosen dataset, one document is held out and the rest of the documents are used to teach the system to predict their topics based on the keywords within. The holdout document is then presented to the trained system as an unknown sample which has its topic predicted. The system then compares whether the predicted topic is the same as the actual topic and tallies a hit or a miss; repeating the process until all documents have been held out and tested. Since the size of our dataset (60) is relatively small, employing the usually expensive LOO cross-validation becomes feasible and desirable, as it allows for the assessment of every pattern as testing data on its own and removes the variance inherent in other validation methods.

In several of the reviewed works regarding automatic metadata generation, the F-Measure was employed to indicate the performance of the tested models. In keeping with this trend, we have also considered the F-Measure alongside the raw classification performance reported by LOO in our evaluation.

The training material used for this experiment was a dataset of our own compilation containing distance learning documents taken from dedicated repositories. The design process of this dataset is discussed in the following subsection.

The procedure to follow for the experiment is thus outlined:

1. Construct a two-topic dataset from reliable distance learning repositories.
2. Select a value for b and n.
3. Process the dataset and assemble the vector representation of the documents according to the algorithm presented in the Method section.
4. Obtain the classification performance and F-Measure of the Heaviside classifier using the vectors obtained in the previous steps and Leave-One-Out cross-validation
5. Repeat Steps 2-5 with new parameters

Data Collection

After a thorough search we were unable to find a pre-made dataset that had the specifications we desired; namely, a set of vectors describing various kinds of distance learning resources written in English that dealt with Computer Engineering topics. For this reason we resorted to generating our own dataset.

To generate a quality dataset it is necessary to have several real distance learning resources. We decided to seek these resources in public repositories. After considering several sources, we decided to use documents available in Rice University's OpenStax CNX repository (http://cnx.org). This repository contains a great

amount of documents grouped into books and pages. In the case of pages, the kinds of content are diverse; some are study programs for a certain topic, others are lessons and some others are exercises and quizzes. These features represent the desired ones for the training data set, as we were also able to find resources related to our desired main topic: Computer Engineering studies. As stated earlier, in this first experiment the problem will be treated as a binary class problem. Therefore, two topics will be considered in this work; namely, Programming and Electronics materials.

In the case of programming resources, 4991 search hits were returned of which 58 were books and the rest were pages. For electronics, 2156 hits were returned of which 14 were books.

A random sample of 30 documents of various formats from each topic was selected. The text content of all the documents was then extracted and concatenated into a single file. This file was then analyzed and its word counts were obtained. As mentioned before, the size of the dictionary, n, is critical for balancing between a comprehensive and a relevant list of terms. The n most frequently appearing words in the combination of the 60 documents, excluding the most common words of the English language, were selected as the dynamically-generated but controlled dictionary of relevant terms.

Once the reference word list was obtained, the feature vector representation of each document was generated by obtaining the number of instances of each reference word and dividing it by the total word count. In summary, the feature vector representation of a document is the relative frequency of every word contained in the controlled dictionary. The generated data set thus contains 60 rows of n real-valued columns, plus a class label for each pattern. The two classes consisted of 30 elements each. The data set was stored as a comma-separated value file.

Experimental Results

The first test was performed using $n = 500$; that is, generating the dataset using the 500 most used words. The most important tunable parameter of the classifier, the numerical base, was left at its default value, the maximum integer in the dataset. The results after applying Leave-One-Out cross-validation are shown in Table 1.

Of note is the fact that, apart from excluding common English grammatical connectors, we employed a naïve approach and did not filter out potentially problematic words beforehand. Inside the list of 500 words, terms such as *two*, *will,* and *therefore* appeared; these words do not contribute useful information for discriminating between topics.

In addition, the more words that are considered in the dictionary the more likely it is that irrelevant terms will appear in the list. In spite of the naïveté of the approach, the classifier is successful in correctly categorizing most documents. However, further refining of the word list results in improved performance.

Table 1

500-*Feature Dataset Results*

	Numerical base	Correctly classified patterns	F-measure
Heaviside Classifier	Maximum value in the dataset	86.66%	0.86667

After the removal of words commonly occurring in both categories of documents and restricting the dictionary to a lesser number of words, performance increases were noticed. Peak performance was reached when using only the 35 most frequent terms in the documents in addition to a base-4 numerical system for the classifier. The results of this test are shown in Table 2.

Table 2

35-*Feature Dataset Results*

	Numerical base	Correctly classified patterns	F-measure
Heaviside Classifier	4	95%	0.9508

As can be observed in Table 2, the performance spike is quite significant. This is due to two factors: first, restricting the dictionary to fewer words means that, in general, the judging of a document's content is stricter and the selected features better differentiate one class from the other. The second factor is that the selected numerical base is a better choice for this specific problem.

Conclusion

Authors of social web content in distance learning environments seldom tag their works with appropriate metadata. This situation greatly limits the potential to build systems in which teachers and learners can quickly and efficiently retrieve helpful and relevant material, as untagged content complicates the cataloging of material and causes search engines to offer sub-optimal results. This, in turn, can potentially result in a subpar learning experience. Automatic metadata generation could alleviate these problems. One of the many possible kinds of metadata that can be generated, and quite possibly the most important one, is the main topic of a document. Having this information allows students to find material on a subject of their interest more easily.

In this work we proposed a system for automatically obtaining the topic of a learning object based on training examples using a supervised learning approach. We focused our work on the field of Computer Engineering and as a first approach we aimed to automatically classify learning material pertaining to the subjects of programming and electronics.

We constructed an original data set from actual learning objects of several different kinds such as lessons, exercises, and study programs. The content of the documents is represented as a feature vector of keyword frequencies. We then applied an original supervised learning method, the Heaviside classifier, to predict the topic of a document based on previously trained data.

The Heaviside classifier, even though it is a model still in development as of the writing of this article, has been able to obtain very satisfactory results with this dataset. We conjecture that this model can be further fine-tuned to obtain improved results.

This paper contributes to the state of the art in automatic metadata generation or prediction by applying a novel classification model that achieves good results with relatively minor preprocessing of the text content of the resources.

Future work will focus on increasing the possible topics for classification; that is, generalizing the problem from binary classification to multiple-class discrimination. If work is successful in discriminating between several classes, this method could be employed to automatically tag a wide range of learning objects pertaining to several different areas of knowledge; resulting in a very useful tool for an e-learning platform.

The automatic metadata generation approach proposed in this work would enable the creation of an e-Learning system capable of offering enhanced resources to its users by intelligently tagging educational material created by educators or found in sources like the Web. This could be achieved by considering a greater number of topics and generating a larger dataset from the analysis of more kinds of resources such as notes or exams.

It is also possible to make some slight modifications to the proposed method in order to generate useful metadata geared toward the Social Web. For this purpose, it would be necessary to create a training sample from the contents of at least one social network of interest. Further work will also follow the improvements achieved with the Heaviside classifier as it is fully developed.

Acknowledgments

The authors would like to thank the Instituto Politécnico Nacional (Secretaría Académica, COFAA, SIP, and CIC), the CONACyT and SNI. The authors also want to thank the anonymous reviewers whose valuable comments help improving this paper.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions with formulas, graphs and mathematical tables*. New York: Dover Publications.

Atkinson, J., Gonzalez, A., Munoz, M., & Astudillo, H. (2014). Web metadata extraction and semantic indexing for learning objects extraction. *Applied Intelligence, 41(2)*, 649-664.

Bauer, M., Maier, R., & Thalmann, S. (2010). Metadata generation for learning objects: An experimental comparison of automatic and collaborative solutions. In M. H. Breitner, F. Lehner, J. Staff, & U. Winand (Eds.), *E-Learning 2010* (pp. 181–195). Berlin Heidelberg: Springer.

Bot, R. S., Wu, Y. B., Chen, X., & Li, Q. (2004). A hybrid classifier approach for Web retrieved documents classification. In A. Spink (Ed.), *Proceedings of the International Conference on Information Technology: Coding and Computing, 2004* (pp. 326-330). Las Vegas, Nevada: IEEE Computer Society.

Bruce, T., & Hillman, D. (2004). The continuum of metadata quality: Defining, expressing, exploiting. In D. Hillman & E. Westbrooks (Eds.), *Metadata in practice* (pp. 238-256). Chicago: ALA.

Castellanos-Nieves, D., Fernández-Breis, J. T., Valencia-García, R., Martínez-Béjar, R., & Iniesta-Moreno, M. (2011). Semantic web technologies for supporting learning assessment. *Information Sciences*, *181*(9), 1517–1537.

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729.

Dharinya, V. S., & Jayanthi, M. K. (2013). Effective retrieval of text and media learning objects using automatic annotation. *World Applied Sciences Journal, Volume*, *27*(1), 123–129.

García-Floriano, A., Camacho-Nieto, O., & Yáñez-Márquez, C. (2015). Clasificador de Heaviside. *Nova Scientia, 7(14),* 365-397.

Greenberg, J. (2004). Metadata extraction and harvesting: A comparison of two automatic metadata generation applications. *Journal of Internet Cataloging, 6(4)*, 59-82.

Kovačević, A., Ivanović, D., Milosavljević, B., Konjović, Z., & Surla, D. (2011). Automatic extraction of metadata from scientific publications for CRIS systems. *Program*, *45*(4), 376–396.

Lytras, M. D., Mathkour, H., Abdalla, H. I., Yáñez-Márquez, C., & De Pablos, P. O. (2014). The social media in academia and education research r-evolutions and a paradox: Advanced next generation social learning innovation. *J. UCS, 20*(15), 1987-1994.

Medelyan, O., & Witten, I. H. (2006). Thesaurus based automatic keyword indexing. In G. Marchionini (Ed.), *Proceedings of the 2006 Joint Conference on Digital Libraries* (pp. 296–297). Chapel Hill, NC: IEEE Computer Society.

Min, B., Shi, S., Grishman, R., & Lin, C. (2012). Towards large-scale unsupervised relation extraction from the web. *International Journal on Semantic Web and Information Systems*, *8*(3), 1–23.

Miranda, S., & Ritrovato, P. (2015). Supporting learning object repository by automatic extraction of metadata. *Journal of e-Learning and Knowledge Society,* 11(1).

Ochoa, X., & Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, *10*(2-3), 67–91. Doi: http://doi.org/10.1007/s00799-009-0054-4

Paynter, G. W. (2005). Developing practical automatic metadata assignment and evaluation tools for internet resources. In M. Marlino (Ed.), *Proceedings of the 2005 Joint Conference on Digital Libraries* (pp. 291–300). Denver, CO: IEEE Computer Society.

Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4, Part 1), 1432–1462.

Şah, M., & Wade, V. (2012). Automatic metadata mining from multilingual enterprise content. *Web Semantics: Science, Services and Agents on the World Wide Web*, *11*, 41–62. doi: http://doi.org/10.1016/j.websem.2011.11.001

Willis, E. A., Szabo-Reed, A. N., Ptomey, L. T., Steger, F. L., Honas, J. J., Al-Hihi, E. M., ... Donnelly, J, E. (2016). Distance learning strategies for weight management utilizing social media: A comparison of phone conference call verses social media platform. Rational and design for a randomized study. *Contemporary Clinical Trials, 47,* 282-288. doi: 10.1016/j.cct.2016.02.005