computers

MDPI

Article

Teaching an Algorithm How to Catalog a Book

Ernesto William De Luca [1,2,3,†], Francesca Fallucchi [1,2,*,†] and Roberto Morelato [1]

[1] Department of Engineering Sciences, Guglielmo Marconi University, 00193 Roma, Italy; ew.deluca@unimarconi.it (E.W.D.L.); roberto.morelato@gmail.com (R.M.)
[2] Leibniz Institute for Educational Media | Georg Eckert Institute, Freisestraße 1, 38118 Braunschweig, Germany; deluca@gei.de (E.W.D.L.); fallucchi@gei.de (F.F.)
[3] Faculty of Computer Science, Otto von Guericke Universität Magdeburg, 39106 Magdeburg, Germany
* Correspondence: f.fallucchi@unimarconi.it
† These authors contributed equally to this work.

Abstract: This paper presents a study of a strategy for automated cataloging within an OPAC or for online bibliographic catalogs generally. The aim of the analysis is to offer a set of results, while searching in library catalogs, that goes further than the expected one-to-one term correspondence. The goal is to understand how ontological structures can affect query search results. This analysis can also be applied to search functions other than in the library context, but in this case, cataloging relies on predefined rules and noncontrolled dictionary terms, which means that the results are meaningful in terms of knowledge organization. The approach was tested on an Edisco database, and we measured the system's ability to detect whether a new incoming record belonged to a specific set of textbooks.

Keywords: automated classification; library catalog; Dewey Decimal Classification

check for updates

Citation: De Luca, E.W.; Fallucchi, F.; Morelato, R. Teaching an Algorithm How to Catalog a Book. *Computers* **2021**, *10*, 155. https://doi.org/10.3390/computers10110155

Academic Editor: George Angelos Papadopoulos

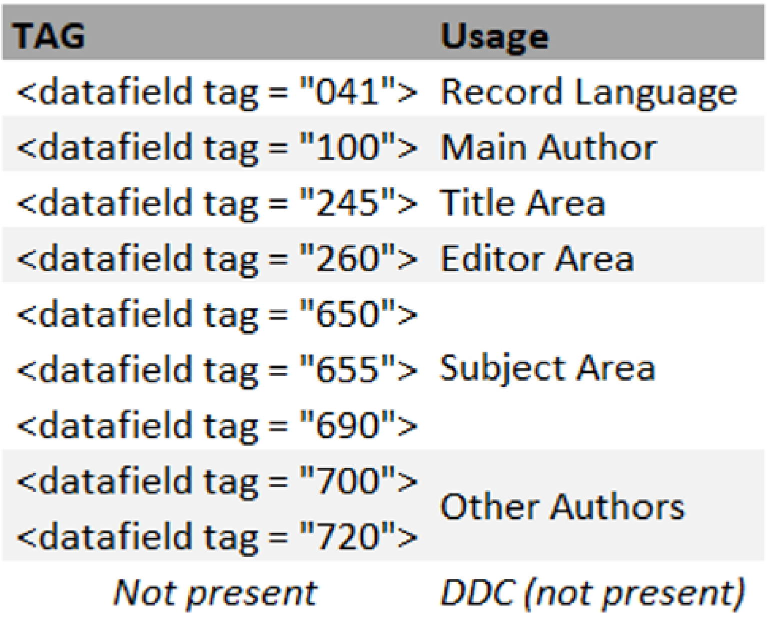
Received: 5 October 2021
Accepted: 15 November 2021
Published: 18 November 2021

1. Introduction

Libraries are increasingly transitioning from manual retrieval systems such as traditional card catalogs to the use of online public access catalogs (OPACs) as information retrieval systems, which are characterized by short bibliographic records of the books, journals, and audio–visual materials available in a particular library [1]. The use of OPACs has greatly changed the nature of libraries and how users access information resources by opening up a variety of portals through which they can access and retrieve information. Currently, the convergence of computers and telecommunications technology has made activities possible that were considered impossible in the past [2], but still with some limitations. If one hears somebody saying, "I bought a new flat" or "I bought a new house", all of us would understand that they are speaking about the place in which they are going to live and that the two terms are used as synonyms. We have the ability to understand the text and abstraction. It is not so simple for a machine, and specifically for a library catalog (OPAC), to understand how house and apartment/flat are used as variants of one another. In general, an OPAC search function performs a linear scan of all the words in a catalog, and if it does not find the exact term sought, no results are returned. The output is based on the insertion of metadata generated by codified operations related to the description of the cataloging material. Many library classification schemes have been developed since the 19th Century [3]. Some even recently have been provided by Common Language Resources and Technology Infrastructure (CLARIN) (https://www.clarin.eu, accessed on 9 November 2021), which has initiated the Component MetaData Infrastructure (CMDI) to overcome the dispersion of data formats and first examples of the creation process for a CMDI metadata profile [4–6]. The most important international classification systems are the classification of the Library of Congress (LCC) (https://www.loc.gov/catdir/cpso/lcco/, accessed on 9 November 2021), the Dewey Decimal Classification (DDC) (https://www.oclc.org/dewey/

features/summaries.en.html, accessed on 9 November 2021), the Universal Decimal Classification (UDC) (http://www.udcc.org/index.php/site/page?view=about, accessed on 9 November 2021), and the Information Coding Classification (ICC) [7]. A variety of classification schemes have also been generated by public and university libraries. Until now, categorization of library books has mostly been undertaken manually using one of these standards. We had, in a previous work [8], also created a complete suite for analysis in digital humanities that allows managing metadata. However, the accuracy of manual classification depends on the librarian's vast knowledge base of various subjects and disciplines. It is not unusual to find errors in such classification, for example a book on neural networks may be wrongly classified under computer networking. Automatic book classification, i.e., a system that helps construct class numbers using the electronic version of a book classification scheme, is more useful for managing print collections of books and other similar forms of documents. Unfortunately, the nature and quantity of the literature available mean that automatic classification has yet to gain the desired momentum. Nevertheless, research on automatic book classification is being undertaken, and literature on this topic does exist, including literature that discusses the applicability of the principles and practices of bibliographic classification to automatic text classification. For example, Yi [9] reviewed the tools, methods, and models developed for automatic text classification, which are based on bibliographic classification schemes, and discussed the issues and challenges in the adoption of bibliographic classification schemes in automatic text classification. A classifier tries to evaluate whether or not a book can be included in a particular grouping. The most common data used to describe a work are title, author, and publisher. There are also subjects and the DDC fields, which are not always present in all catalogs. The aim of this work is to understand whether it is possible to identify a strategy to automate the classification operations, allowing decisions about the inclusion or exclusion of a book, in a given category, based on a few descriptive elements. This work focuses on the content of the Edisco catalog, an electronic database that aims to register the books for school and education published in Italy between 1800 and 1900. Using an existing and proven database is particularly important as the records are not subject to further classification and their inclusion has already been assessed. The records constitute the information that the classifier will use when evaluating new elements in order to produce positive feedback to a search query. The machine should compare the incoming metadata with those in the database and be able to provide reliable answers. The records are accompanied by numerically identified tags. The most relevant are shown in Figure 1. Note the remarkable absence of the DDC field.

The fields dedicated to the subjects do not follow the Italian Cataloging Rules, especially those related to the New Subject Heading System, which includes fields that allow you to organize cataloging records taxonomically.



TAG	Usage
<datafield tag = "041">	Record Language
<datafield tag = "100">	Main Author
<datafield tag = "245">	Title Area
<datafield tag = "260">	Editor Area
<datafield tag = "650">	Subject Area
<datafield tag = "655">	
<datafield tag = "690">	
<datafield tag = "700">	Other Authors
<datafield tag = "720">	
Not present	*DDC (not present)*

Figure 1. Metadata used for Record description. DDC field is missing.

2. Related Work

Measuring the effectiveness of the use of computer catalogs (OPAC) according to [10] has been a constant area of study for some decades; this has led to ideas about how data extraction systems might be improved in order to better satisfy the informational needs of users. There is, however, another approach to automatic book classification attributed to the library science community, which has been less closely investigated [9,11,12]. This approach focuses less on algorithms and more on taking advantage of comprehensive controlled vocabularies, such as library classification schemes and thesauri, which have been developed and used for the manual classification of holdings in conventional libraries. A library classification system is a coding system for organizing library materials according to their subjects and aims to simplify subject browsing. Library classification systems are used by expert library catalogers to classify books and other materials (e.g., serials, audio–visual materials, computer files, maps, manuscripts, realia) in conventional libraries. The two most widely used classification systems in libraries around the world today are the DDC and the LCC which since their introduction in the late 18th Century, have undergone numerous revisions and updates. A promising avenue for the application of this approach is the automatic classification of resources archived in digital libraries, where using standard library classification schemes is a natural and usually the most suitable choice because of the similarities between conventional and digital libraries. Another application of this approach is in the classification of web pages, where due to their subject diversity, proper and accurate labeling requires a comprehensive classification scheme that covers a wide range of disciplines. In such applications, using library classification schemes can provide fine-grained classes that cover virtually all categories and branches of human knowledge. In general, Automatic Text Classification (ATC) systems that have been developed based on the above library science approach can be divided into two main categories: string-matching systems and ML-based systems. The string-matching systems do not rely on Machine-Learning (ML) algorithms to perform the classification task. Instead, they use a method that involves string-to-string matching between words in a term list extracted from library thesauri and classification schemes and words in the text to be classified. Here, the unlabeled incoming document can be thought of as a search query out to the library classification schemes and thesauri, and the result of this search includes the class(es) of the unlabeled document. One of the most well-known examples of such a system is the Scorpion project [13] by the Online Computer Library Centre (OCLC) [14]. Scorpion is an ATC system for classifying e-documents according to the DDC scheme. It uses a clustering method based on term frequency to find the classes most relevant to the document to be classified. A similar experiment was conducted in the early 1990s by Larson [15], who built normalized clusters for 8435 classes in the LCC scheme from manually classified records of 30,471 library holdings and experimented with a variety of term representation and matching methods. For another example of these systems, see [16]. The ML-based systems utilize ML algorithms to classify e-documents according to library classification schemes such as the DDC and the LCC. They represent a relatively unexplored trend, which aims to combine the power of ML-based ATC algorithms with the enormous intellectual effort that has already been put into developing library classification systems over the last century. Chung and Noh [17] built a specialized web directory for the field of economics by classifying web pages into 757 subcategories of economics listed in the DDC scheme using a k-NN algorithm. Pong et al. [18] developed an ATC system for classifying web pages and digital library holdings based on the LCC scheme. They used both k-NN and Naive Bayes (NB) algorithms and compared the results. Frank and Paynter [19] used the linear SVM algorithm to classify over 20,000 scholarly Internet resources based on the LCC scheme. Wang [20] used both NB and SVM algorithms to classify a bibliographic dataset according to the DDC scheme and compared the results.

3. Understanding the Bibliographic Elements

The idea is to consider the contribution that all the fields that describe the cataloging record can give, with respect to the need for automated classification. It is useful to understand how they can be treated, transforming them from a descriptive element to a Boolean or numerical type. It is therefore necessary to establish how the system should behave when information is lacking. Some fields, such as series or publisher, are less significant. Definitely significant however are metadata relating to the subject, which consist of the attribution of an index item (a descriptor) to a document that summarizes its content. The DDC is an enumerative indexing system that allows you to optimize the location, but also to carry out an ontological grouping. The summary or abstract is an encoded field in the catalog that contains extracts or descriptions of the record. By simulating a new insertion in the catalog, we can view two examples taken from the Florence National Central Library (BNCF, Biblioteca Nazionale Centrale di Firenze) catalog (https://www.bncf.firenze.sbn.it/cataloghi/, accessed on 9 November 2021) reported in Table 1.

Table 1. Cataloging records for two books in the BNCF.

Book #1
Matematica : per il biennio della scuola superiore (*Mathematics for the first two years of high school*) (https://opac.bncf.firenze.sbn.it/bncf-prod/resource?uri=CFI0531816&found=1, accessed on 9 November 2021) Titti Alvino
Book #2
Parole di scuola(*Words from school*) (https://opac.bncf.firenze.sbn.it/bncf-prod/resource?uri=TO01901965&found=1, accessed on 9 November 2021) Mariapia Veladiano - I mattoncini – Erickson

Book #1 defines very clearly the content and target audience for which it was designed. Book #2 was chosen for the title that contains the word school and for the brevity of the title. There are no further publications of the author in the starting database; it is ambiguous with respect to the end of the classification in the DB Edisco. This allows you to simulate how the classification algorithm should proceed to evaluate a new publication

Reading only the title of Book #2, this seems interesting for the collection, but it necessarily requires a series of further investigations. The abstract on the website of another publishing house, Guanda, clarifies that the volume actually is about the school, but it is not a textbook that could be adopted for teaching. "Mariapia Veladiano, after more than twenty years in the school, first as a teacher and then as a principal, knows the school well. She knows the boys, the energy that runs between the benches, the adolescence made of fear and desire, the future she promises, and at the same time she threatens". Being a novel, no value is attributed to the field of the subject, so it is not possible to take advantage of this form of grouping of the content. Not even in the BNCF is there an indication of the DDC, so it is not possible to make further considerations. At this point, it is necessary to evaluate other works by the same author present in the catalog. A manual search on the BNCF OPAC returned four records:

- Veladiano, Mariapia. *La vita accanto*(Life next door). Einaudi, 2012. 853.92 (ed. 22) - NARRATIVA ITALIANA, 2000;
- Veladiano, Mariapia. *Il tempo è un dio breve*(Time is a short god). Einaudi, 2012. 853.92 (ed. 22) - NARRATIVA ITALIANA, 2000;
- Veladiano, Mariapia. *Una storia quasi perfetta: [romanzo]*(An almost perfect story: [novel]). Guanda, 2016. 853.92 (ed. 22) - NARRATIVA ITALIANA, 2000;
- Veladiano, Mariapia. *Ma come tu resisti, vita*(But how you resist, lif). Einaudi, 2013. (DDC value not reported).

If we were to consider inserting the second book into the catalog based only on its title, we would probably make a mistake because it is a novel and not a school textbook. A possibility for an algorithm that does not have access to a subject or the DDC category would be to evaluate other publications by the same author. In the BNCF, the DDC values related to the four other books by the same author define them as fiction.

To understand how an automated procedure can work, it is necessary to assess how the operation carried out by a library staff member would work. This operation is described in the following steps. Consider again the catalogs extracted from the catalog of BNCF, the Biblioteca Nazionale Centrale di Firenze, and reported in Table 2. Book #1 defines very clearly the content and target audience for which it was designed. Book #2 was chosen for the title that contains the word school and for the brevity of the title. Five steps are then followed:

- Step 1: Evaluation of the words contained in the titles;
- Step 2: A manual search of the OPAC of BNCF returned four records by the same author. In all cases, they were novels;
- Step 3: Evaluating the indexing, i.e., the field of the Dewey Decimal Classification (DDC), of all the volumes returned for the author Veladiano, we noticed that 75% of the items present in the set cannot be considered textbooks; they are storytelling. For the remaining 25%. the DDC was not indicated;
- Step 4: Evaluation of the subjects. In this case, as it is storytelling, they are not indicated, a further indication that it is not an essay or textbook;
- Step 5: Once the abstract has been evaluated, it is immediately clear that this is not a school text: "... in the classroom you learn the right words to understand yourself, others, the world and life". (https://www.ibs.it/parole-di-scuola-libro-mariapia-veladiano/e/9788823513259, accessed on 9 November 2021).

For the sake of brevity, the operations carried out on the text of the author Alvino are summarized in a table comparing them with those just carried out for the author Veladiano.

If the operation of an appointee were to end here, there would be no particular difficulty in considering Book #1 as acceptable for classification, while for Book #2, although there are various elements of ambiguity, it could not be classified as a school textbook.

Table 2. Results of each step of the cataloging operation on Book #1 and Book #2.

Step	Book #1	Book #2
Step 1	*Matematica: per il biennio della scuola superiore*	*Parole di scuola*
Step 2	4	not a textbook
Step 3	510	not present
Step 4	Mathematics	not present
Step 5	not present	not a textbook

For the study of a classifier, the behavior being analyzed could be represented in a decision tree path schema in a first overview. However, a true/false, discrete variable process may be too simplistic to represent such a complex universe. An automated classification system might consider the record from the following point of view:

1. DDC;
2. Subject;
3. Author;
4. Title;
5. Abstract.

By executing the five phases in sequence (see Figure 2), the classifier should associate a value with each operation, in some cases a Boolean value, whether accepted or not, and in others, a quantitative value. The goal is to transform the cataloging operation into a classification system based on quantifiable metrics. At the end of the sequence of operations, it could assign a confidence value, used to consider whether a record belongs to the desired category.

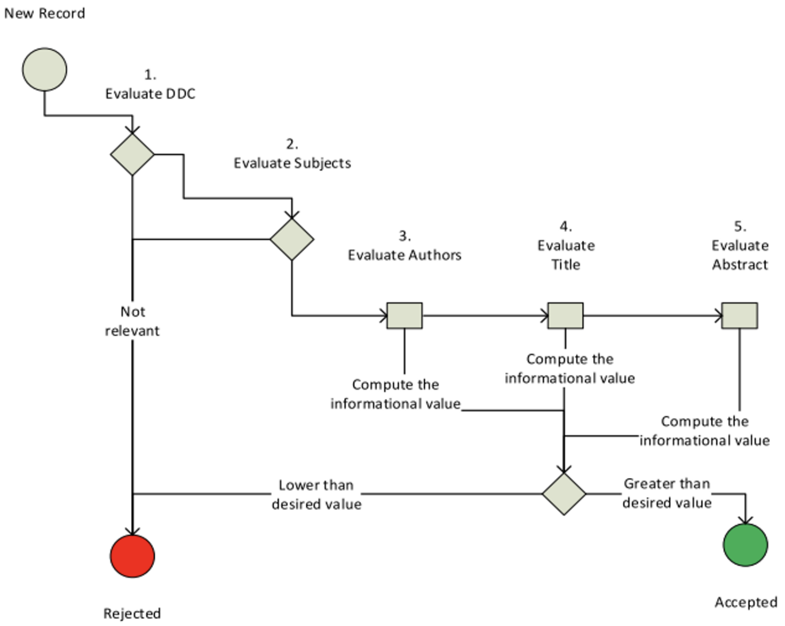


Figure 2. Schematic approach for evaluating a new record.

Currently, there are no OPACs that interpret the information contribution of each cataloging element as proposed here. This process could be applied both when inserting a new record and during the search operations performed by a user.

4. Integration of External Data

To integrate external data, it was essential to find a set of records to generate a training set and a test set to pass to the classifier. The strategy adopted is to carry out a series of searches on certain online catalogs that can be likened to existing content. By grouping titles and subjects of the Edisco DB (https://www.edisco.unito.it, accessed on 9 November 2021) together, a set of words was returned that could be used as the starting point to run a search in other catalogs. By analyzing the n-grams, a threshold value was determined that would ignore words such as names of people. The study of n-grams, which are schematized models of fundamental recurrent architectures in language, consists of assigning a certain probability to a word occurring in combination with other words. Given a dictionary, or a set of words, it is therefore a question of the system assigning a certain probability to an n-gram and considering it as the probability that the last word would appear after the other n-1 words (in that order). The idea is to derive some series of possible n-grams starting from the strings offered by the DB Edisco, in particular from titles and topics related to the works. Once the set of words was refined, it was possible to submit a series of queries to Italian book collections that would allow queries according to machine languages. The set of identified words was used as a search key in the subject field. A rather heterogeneous catalog that allows remote querying is that of the Linked Open Data project of the Coordination of Special and Specialist Libraries of Turin (CoBiS), which contains 438,942 records. Records with language tags not corresponding to Italian publications were ignored. Records with titles shorter than 11 characters were also discounted. A limit was set for the sample analysis so that only works were shown that were connected to others according to an FRBR hierarchical structure. An additional filtering process of valid records was implemented. The strategy was to consider only those records that included a linked subject descriptor. This choice was due to extracting the relevant queries, searching for new records that have subject descriptors. In the evaluation phase of the records generated by the CoBiS import, the grouping in digraphs, n-grams composed of two graphemes were used. This type of operation was carried out both individually on the Edisco and CoBiS records and then again by combining the two data sources. In the set of documents containing all the records of the two catalogs, the two-grams obtained are filtered according to a minimum frequency rule according to which documents with a “document frequency” lower than the desired value were not considered. This part of the work was particularly useful to understand the composition of CoBiS records, without having to analyze them individually. Bringing out the most important n-grams allowed easily evaluating the type of records available. By creating lists of words to ignore, it was possible to quickly filter records that were not relevant, improving the quality of the set of titles to be kept. At the end of all the operations, it was possible to obtain a set of consistent records equal to 55,256 units, books that largely deal with topics relating to mountain excursions, the local history of Northern Italy, congresses and conferences, and the history of music and musical scores. In total, the Edisco database contains 25,343 records, of which 24,374 are in Italian.

5. Defining the Ideal Classifier

In order to classify a record, it is necessary to structure a measurement system that allows the definition of metrics to be applied to the data that constitute the record. If you consider the two books in Table 1, Book #1, by Titti Alvino, should pass the measurement test, but not the other. All the values related to Book #2 are too low to generate a final acceptable result, mostly those related to the DDC, subject abstracts, and other publications of the author Mariapia Veladiano. The general idea is to assign a weighting to a set of descriptors (features). Each record should be valorized, producing a result that is generically identified with R, which could fluctuate between zero and one-hundred. This value can be a vector of attributes composed of other vectors. Let us assume R is a vector describing a generic record composed of the weighting values A0 by the main author, (A0,A1,. . .,An), which represent possible co-authors, E the editor, C editorial series, D the physical dimensions of the cataloging object, and T the semantic analysis of the title. A hypothetical vector would therefore be composed as follows:

$$Rx = (A0 : 46, E : 24, C : 12, D : 80, T : 49) \tag{1}$$

An example of building the vector for the following example taken from BNCF:
New course in analytical geometry and algebra complements: for scientific high schools: with 560 examples and over 4600 exercises / N. Dodero, P. Baroncini, R. Manfredi | Analytical geometry - School texts | 516.3 (ed. 21) - ANALYTICAL GEOMETRIES

In this case, the value taken as the subject (Analytical geometry - School texts) S and the indexing DDC (516.3 - ANALYTICAL GEOMETRIES) I would be particularly significant. The author attribute A, measured with respect to the calculated values of the subject and index, would be expressed according to the following formula:

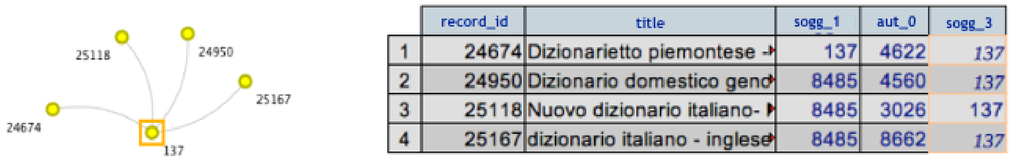
$$A = (S : 100, I : 100) \tag{2}$$

In order to understand whether the definition of the ideal classifier is viable or not, it is necessary to conduct a whole series of tests. Investigations were carried out using author names, subjects, and the words of the titles to verify the distribution of each element with respect to the total number of records available. A relatively common issue is that the same author name can be registered in the system in different ways. It is therefore necessary to first standardize the names of the authors, trying to iron out the differences present due to errors or different entry styles. A basic structure for an algorithm could verify the dates of the works present in the system and calculate the average and distribution over the period. To support the decision, the machine could therefore consider the subjects and the DDC of the various works, evaluating similarities and differences each time. Any outliers, that is to say publications very far from the central hub, could be evaluated as less important than those closest to the hub. The use of dates deserves further study as cataloging practice dictates that the year indicated in the metadata is the one relating to the edition. This means that it might not closely relate to the actual publication dateof the work. In addition to the possible analogies among names, there are also several cases in which the names of the authors are not particularly significant, as for the following: A and C, or A. S., and many more.

Subject descriptors are certainly useful, but must also be preprocessed. Edisco contains 293, which is not many given the size of the catalog.

After excluding the subjects that are duplicated in the three categories, only 178 significant subjects remained. This is important in order to understand if, and how, the subject field within the classifier can be used. Overall, no more than one third of the records have one associated subject. Different subject descriptors are present several times in the Edisco DB with similar forms to each other. For example, in the case of French Grammar, it is also listed as: French Grammar forms (two spaces between the headwords), or Frnch Grammar (without an "e") or Ferench Grammar (one letter "e" too many). By calculating the Levenshtein distance, which is the minimum number of replacements, deletions, or insertions that must be made to obtain one string from another, analogous strings can be grouped into clusters, further reducing the number of subjects useful for research purposes that are related to the adoption of the ideal classifier. It is useful to investigate whether there are any relationships between subjects and authors, relating to the Edisco DB. When looking for how many subjects are connected to at least one author, 166 unique strings emerged. These have a relationship with 1852 different authors in total. By reversing the relationships that link these subjects to the records in the database, a total of 4048 items could be reached.

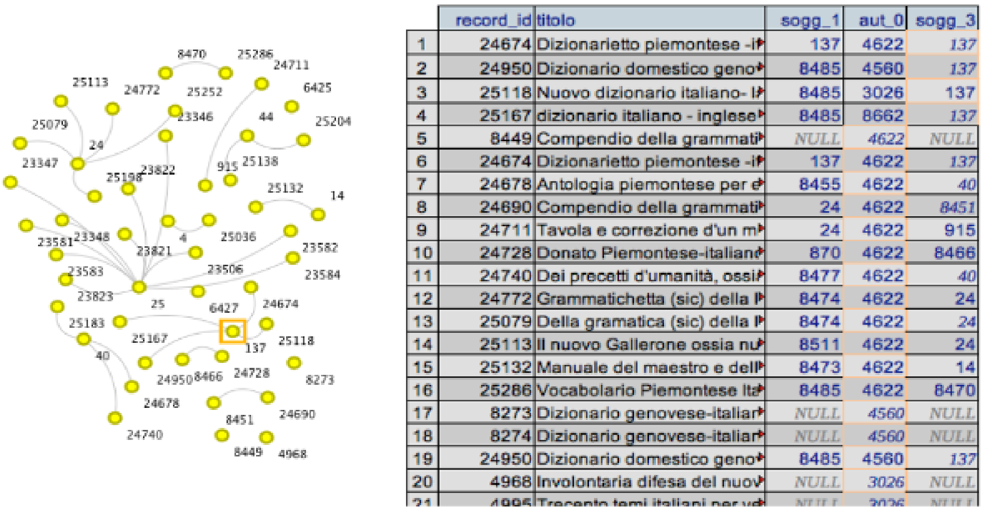
Figure 3 presents an example of using the term "dictionary" (ID 137) as a query term. The search results based on the term "dictionary" were four records. Each record is composed of an Identifier (ID), a title, two subjects (sogg_1, sogg_3), and the reference to each specific author (aut_0). They correspond to the aforementioned results of Figure 1, in the subject area. sogg_1 stands for tag 650 and sogg_3 for tag 690.



	record_id	title	sogg_1	aut_0	sogg_3
1	24674	Dizionarietto piemontese -	137	4622	137
2	24950	Dizionario domestico geno	8485	4560	137
3	25118	Nuovo dizionario italiano- I	8485	3026	137
4	25167	dizionario italiano - inglese	8485	8662	137

Figure 3. Records returned searching the term "dictionary".

Asking the system to generate the graph of relations dependent on the four authors connected in column aut_0, a network of 40 records was obtained. These in turn had a total of 13 connected subjects (see Figure 4).



	record_id	titolo	sogg_1	aut_0	sogg_3
1	24674	Dizionarietto piemontese -i	137	4622	137
2	24950	Dizionario domestico geno	8485	4560	137
3	25118	Nuovo dizionario italiano- I	8485	3026	137
4	25167	dizionario italiano - inglese	8485	8662	137
5	8449	Compendio della grammati	NULL	4622	NULL
6	24674	Dizionarietto piemontese -i	137	4622	137
7	24678	Antologia piemontese per e	8455	4622	40
8	24690	Compendio della grammati	24	4622	8451
9	24711	Tavola e correzione d'un m	24	4622	915
10	24728	Donato Piemontese-italiano	870	4622	8466
11	24740	Dei precetti d'umanità, ossi	8477	4622	40
12	24772	Grammatichetta (sic) della I	8474	4622	24
13	25079	Della gramatica (sic) della I	8474	4622	24
14	25113	Il nuovo Gallerone ossia nu	8511	4622	24
15	25132	Manuale del maestro e dell	8473	4622	14
16	25286	Vocabolario Piemontese Ita	8485	4622	8470
17	8273	Dizionario genovese-italiar	NULL	4560	NULL
18	8274	Dizionario genovese-italiar	NULL	4560	NULL
19	24950	Dizionario domestico geno	8485	4560	137
20	4968	Involontaria difesa del nuov	NULL	3026	NULL
21	4995	Trecento temi italiani per v	NULL	3026	NULL

Figure 4. The list of the first 20 over 40 records, related to the 4 authors in Figure 3.

6. Semantic Analysis

The two datasets, CoBiS and EDISCO, must be comparable. The goal was to produce a single set of data from which to extract training and test sets. For each set, the following operations were carried out: (a) The first was the creation of a document vector where scores were assigned to all the words present in order to transform free text into something understandable for a machine-learning model. A Bag Of Words (BOW) was created, which led to the following: (b) First was a study of the TF-IDF frequency; the vectorization function considered a word less important, even if it appeared many times in a text, when it detected the same word in other texts as well. The absolute TF, DF, and IDF frequencies were calculated, for the entire set of Edisco and CoBiS words. (c) The second was a topic extraction via parallel LDA looking for 10 topics, a probabilistic model of the unsupervised type, which allowed the natural language to be analyzed by evaluating the similarity between the distribution of the terms in the document and another of a specific topic. This allows you to enter a new document into the system and evaluate the classifier's goodness-of-fit. The classification process was based on the measurement, by the machine, of the text contained in the various titles. The classifier was developed according to the scheme in Figure 5.

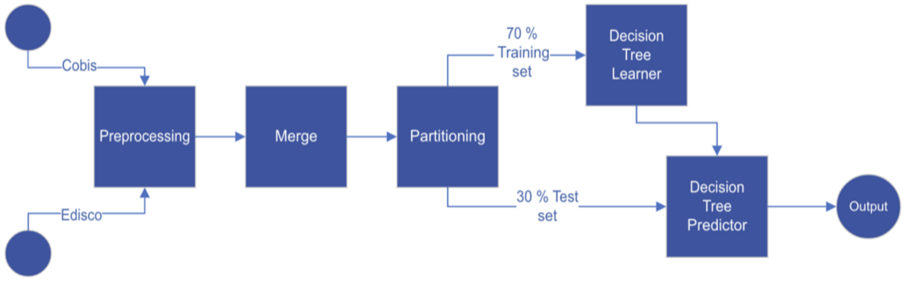
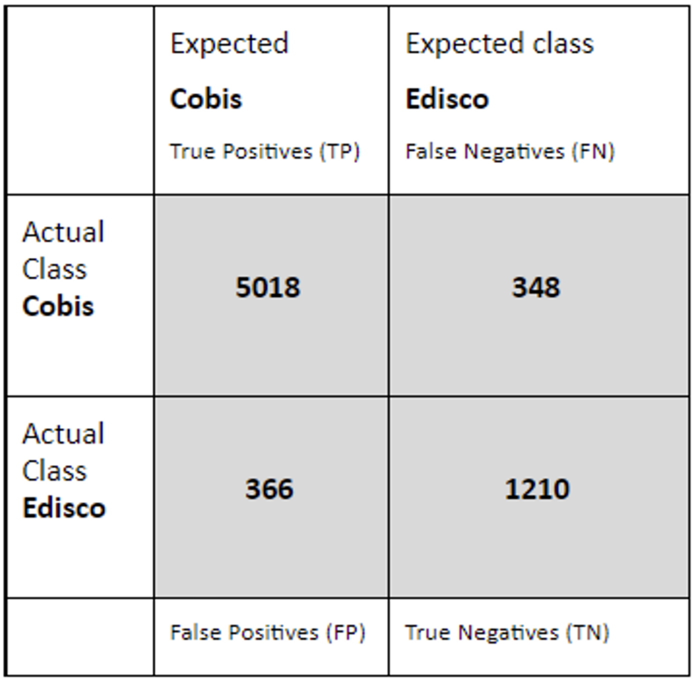


Figure 5. Structure of the classifier.

The decision tree algorithm operated by splitting the training set each time features with a value greater than specified occurred. The reference target variable was of a discrete type and was set during the preprocessing phase. The quality measure used was the Gini index, which minimizes the variance. The split operations were performed once the average value of the two partitions had been calculated. The size of the tree was limited during the training operations, by dictating that the minimum value for each node be equal to two and subsequently performing reduced error pruning operations. This was in order to reduce classification errors, minimizing the risk of overfitting in the model. Algorithms that start from the leaves replace the nodes that have derived them with the most suitable class, but only if the level of accuracy of the prediction remains stable. For nodes that did not return a result (no true child problem), the decision tree algorithm was set to return a value of NULL. To study the progress of the classifier and carry out an evaluation of the model, the contingency table (confusion matrix) was generated (see Figure 6). It presented different combinations of predicted and actual values taking into consideration the two datasets used: Edisco and CoBiS.



	Expected **Cobis** True Positives (TP)	Expected class **Edisco** False Negatives (FN)
Actual Class **Cobis**	**5018**	**348**
Actual Class **Edisco**	**366**	**1210**
	False Positives (FP)	True Negatives (TN)

Figure 6. Confusion matrix for the combined results of the two datasets.

Table 3 shows the promising results obtained for the two dataset considered above. Here, the correct classification of the results achieved approximately 93.5% precision. On the other side, the recall in the whole datasets was 93.2%.

Table 3. Results obtained for the Edisco and CoBiS datasets.

Sensitivity: measures how the classifier behaves in predicting events belonging to the class (also called recall, which measures the model's goodness-of-fit with respect to its ability to capture positive events in classifying textbooks):	$\frac{TP}{(TP+FN)} = \frac{5018}{(5018+348)} = 0.9351$
Specificity: measures the accuracy of class assignments:	$\frac{TN}{(TN+FP)} = \frac{1210}{(1210+366)} = 0.7678$
Precision: measures the model's ability to classify documents that do not belong to the school book class:	$\frac{TP}{(TP+FP)} = \frac{5018}{(5018+366)} = 0.9320$
F-measure: the harmonic mean between recall and precision:	$2 * \frac{recall * precision}{recall + precision} = 0.8844$

In Figure 7, we show the performance of our classification model, and the points above the diagonal represent the good classification results we obtained.

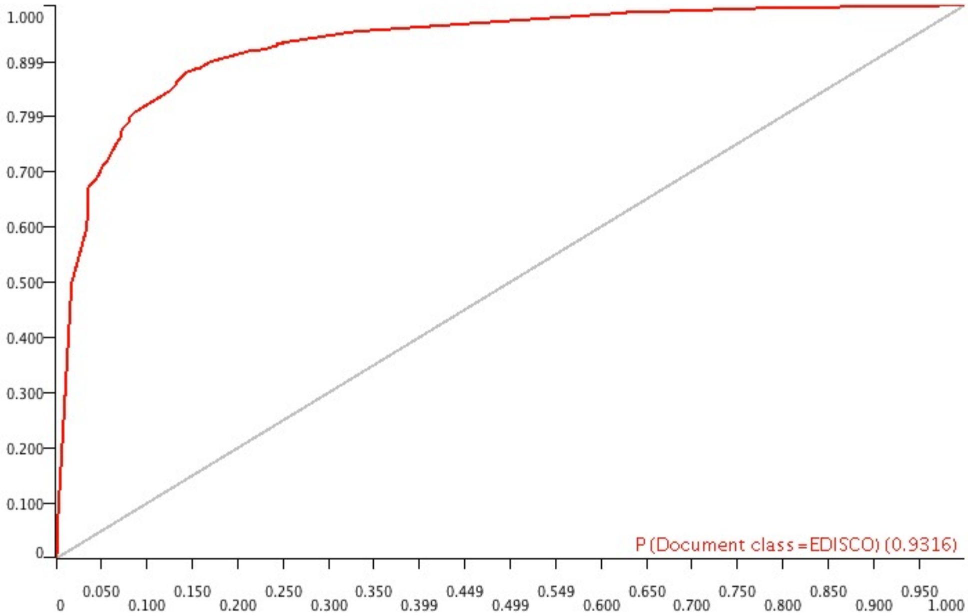


Figure 7. ROC curve for the Edisco class.

7. Conclusions

The description in Section 5 represents the basis of the classifier, which was obtained by exclusively evaluating the text contained in the titles of the records. Although this already allowed us to achieve good results, it could be optimized by extending the work to include the process described in Section 4, where the operation for the ideal classifier is outlined. This assigns imaginary values to vector R and in particular A0 as the lead author, E as the editor, C as the series, D as the dimensions, and T for semantic analysis of the title. As hypothesizing a deterministic link is rarely plausible, a random variable error must be added that summarizes the uncertainty about the true relationship among values contained in brackets. The vector may now be expressed as a function:

$$n = f(A0[46], E[24], C[12], D[80], T[49]) + \varepsilon \tag{3}$$

It would also be interesting to determine whether there is an average dependency among different phenomena. This is possible through a linear regression evaluation. We set the value of the author as fixed and randomly varied the value relating to the editor. These two elements should not have a particular relationship.

Calculating the value of the coefficient of determination R^2, which is the relationship between the variability of the data and the correctness of the statistical model used, records in Column #1 have a value of $R^2 = 0.8832$, which means that the model obtained does not interpret the relationship between the variables as very strong. The second record highlights a weak relationship between author and publisher with a value of $R^2 = 0.0038$. By identifying relationships between pairs of cataloging elements, such as author/subject or author/DDC, you can ascertain whether there are recurring connections. This provides the classifier with additional elements with which to evaluate new, incoming data. Furthermore, as stated in Section 4, where the neighborhood relations between subjects and authors were investigated, the results may be different if the order in which the operations are performed is changed. In the example above, the decision was made to proceed incrementally: starting from a single subject (Step A), then looking for the linked authors (Step B), and grouping (Step C). Different possibilities arise from linking the order of operations in a different way. Once Step B is finished, you could choose to proceed by extracting all the subjects derived from this grouping of records and considering them in one new Step E, running a different path along the structure.

A next step to optimize the process would be to conduct a series of measurements on catalogs that expose all the metadata, in particular the DDC. It would be important to understand how to calibrate the information supplied by each element of the metadata. Finally, it would be interesting to extend the study to the area of search engines in order to improve the results provided by the OPACs.

Author Contributions: Conceptualization, R.M. and F.F.; methodology, R.M. and F.F.; software, R.M.; validation, R.M., F.F., E.W.D.L.; writing—original draft preparation, R.M., F.F.; writing—review and editing, R.M., F.F., E.W.D.L.; supervision, E.W.D.L. All authors have read and agreed to the published version of the manuscript.

Funding: Leibniz Institute for Educational Media | Georg Eckert Institute.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data has been present in main text.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Thanuskodi, S. Use of Online Public Access Catalogue at Annamalai University Library. *Int. J. Intell. Syst.* **2012**, *2*, 70–74. [CrossRef]
2. Adeogun, M. The digital divide and university education systems in sub-Saharan Africa. *Afr. J. Libr. Arch. Inf. Sci.* **2003**, *13*, 11–20.
3. Lorenz, B. Systematische Aufstellung in Vergangenheit und Gegenwart. In *Beiträge zum Buch- und Bibliothekswesen 45*; Harrassowitz Verlag: Wiesbaden, Germany, 2003; p. 365.
4. Fallucchi, F.; Steffen, H.; De Luca, E. *Creating CMDI-Profiles for Textbook Resources*; Springer: Berlin, Germany, 2019; Volume 846. [CrossRef]
5. Fallucchi, F.; De Luca, E. *Connecting and Mapping LOD and CMDI Through Knowledge Organization*; Springer: Berlin, Germany, 2019; Volume 846. [CrossRef]
6. Fallucchi, F.; de Luca, E.W. CMDIfication process for textbook resources. *Int. J. Metadata Semant. Ontol.* **2020**, *14*, 135–148. [CrossRef]

7. Dahlberg, I. The information coding classification (ICC): A modern, theory-based fullyfaceted, universal system of knowledge fields. *Axiomathes* **2008**, *18*, 161–176. [CrossRef]
8. De Luca, E.W.; Fallucchi, F.; Ligi, A.; Tarquini, M. A Research Toolbox: A Complete Suite for Analysis in Digital Humanities. In *Research Conference on Metadata and Semantics Research*; Springer: Cham, Switzerland, 2019; Volume 1057.
9. Yi, K. Automated Text Classification Using Library Classification Schemes: Trends, Issues, and Challenges. *Int. Cat. Bibliogr. Control* **2007**, *36*, 78–82.
10. Villen-Rueda, L.; Senso, J.A.; de Moya-Anegón, F. The use of OPAC in a large academic library: A transactional log analysis study of subject searching. *J. Acad. Librariansh.* **2007**, *33*, 327–337. [CrossRef]
11. Markey, K. Forty Years of Classification Online: Final Chapter or Future Unlimited? *Cat. Classif. Q.* **2006**, *42*, 1–63. [CrossRef]
12. Golub, K. Automated Subject Classification of Textual Documents in the Context of Web-Based Hierarchical Browsing. *Knowl. Organ.* **2011**, *38*, 230–244. [CrossRef]
13. OCLC Research. Available online: https://www.oclc.org/research/activities/scorpion.html (accessed on 9 November 2021).
14. OCLC (Online Computer Library Center). Available online: https://www.oclc.org/en/home.html?redirect=true (accessed on 9 November 2021).
15. Larson, R.R. Experiments in automatic Library of Congress Classification. *J. Am. Soc. Inf. Sci.* **1992**, *43*, 130–148. [CrossRef]
16. Jenkins, C.E.A. Automatic Classification of Web Resources Using Java and Dewey Decimal Classification. *Comput. Netw. ISDN Syst.* **1998**, *30*, 646–648. [CrossRef]
17. Chung, Y.M.; Noh, Y.H. Developing a specialized directory system by automatically classifying Web documents. *J. Inf. Sci.* **2003**, *29*, 117–126. [CrossRef]
18. Pong, J.Y.H.; Kwok, R.C.W.; Lau, R.Y.K.; Hao, J.X.; Wong, P.C.C. A comparative study of two automatic document classification methods in a library setting. *J. Inf. Sci.* **2008**, *34*, 1–18. [CrossRef]
19. Frank, E.; Paynter, G.W. Predicting library of congress classifications from library of congress subject headings. *J. Am. Soc. Inf. Sci. Technol.* **2004**, *55*, 214–227. [CrossRef]
20. Wang, J. An extensive study on automated dewey decimal classification. *J. Am. Soc. Inf. Sci. Technol.* **2009**, *60*, 2269–2286. [CrossRef]