Search model of educational trends based on Data Mining techniques

Rosario Huanca-Gonza[1][0000−0002−1437−5829], Julio Vera-Sancho[2][0000−0001−5526−5223], Carlos Eduardo Arbieto-Batallanos[3][0000−0002−7094−4272], and María del Carmen Córdova-Martínez[4][0000−0002−5186−6598]

Universidad Nacional de San Agustín de Arequipa
{rhuancag,jveras,carbieto,mcordovam}@unsa.edu.pe

Abstract. Internet is the broadest means of communication that has existed and is a highly effective means for the dissemination of information that allows access to millions of pages of textual and multimedia content, this leads to an information overload and a problem called infoxication, and Researchers and / or teachers are not the exception when searching for information on educational trends in research. For this reason, we propose a model to search for educational trends using Data Mining techniques, which will allow us to capture, analyze, disseminate and exploit the main topics that are currently being developed on educational trends.

Keywords: Data mining · educational trends · Machine learning

1 Introduction

At present, we live in an era where information is easily accessible and due to the large amount of information, and that this information that exists on the web, is increasing, according to an IDC report (International Data Corporation), that only 33% of the information is valuable, if it is analyzed, and that by 2020 this information will increase about 5GB [8]. Currently, as part of this great information, it is that infoxication appears, which is the excess of information that creates confusion in the users of ICT. It is also known as info-saturation in relation to the cognitive effects produced by access to large amounts of information that the individual fails to appropriate [14]. Based on this great information, there is a need among researchers and / or educators, the search for educational trends, which allow improving the teaching and / or learning process, both by teachers and students, there is a large number of repositories specialized in research on education such as: ERIC, which is a bibliographic database of international coverage in the field of education, includes indexes and summaries of journal articles and reports, known as the documents of *Education Resources Information Center (ERIC)*, from 1966 to the present, it has a monthly update frequency and has more than 1,341,146 records [15]. The search for educational trends in research, has been carried out in recent years manually, with the ability to filter

information that is related to search, assessment and synthesis of information. For which the individual in an environment of abundant information is able to critically select the information and give meaning and meaning [14] In the advances of artificial intelligence and data processing, there are investigations and techniques that allow us to perform this entire process automatically, based on the fact that this massive information that exists is known as *Big Data*, which needs to be processed and thus generate value. For which the algorithms of of *Data mining* It allows us to solve these types of problems. The types of learning are Supervised, Unsupervised and Semi supervised [5]; Supervised learning takes a known set of input data and known responses, which are labeled, and then make an algorithm that will generate a prediction to respond to new data, this type of learning uses classification or regression algorithms. The unsupervised learning unlike the supervised, does not have tagged data, its objective is to find the regularities at the entrance, so that certain patterns can be found. The phases used in*Data Mining* are, data filtering, variable selection, knowledge extraction, interpretation and evaluation [17] which in our proposal will help us discover the knowledge of educational trends in research.

2 State of the art

Slamet in 2018 in his research "Web Scraping and Naïve Bayes Classification for Job Search Engine" proposes that many organizations use websites to share information about new hires for workers and that this information is overflowed in thousands of sites with different attributes and criteria. However, this availability of information is very complex in the selection process and leads to inefficient execution time, which is why it proposes a simple method to simplify the job search through a construction and development of web techniques scraping and sorting using Naive Bayes in the search engine. In 2016, Meschenmoser in his research "Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction" proposes strategies to programmatically access data in scientific web repositories. We demonstrate the strategies as part of an open source tool (MIT license) that allows comparisons of research performance based on Google Scholar data, emphasizing that the Scraping included in the tool should only be used if the operator of a repository gives its consent [11]. Klochikhin2016 in his project "Collaborative innovation beyond science: exploring new data and new methods with computer science" mentions that bibiometry and patent analysis have laid an important basis for a better understanding of the dynamics of innovation; The new computational methods and tools can take this analysis one step further while providing additional information on the mechanisms of collaborative innovation. Web Scraping, record linking algorithms and computational linguistics provide a wide range of approaches to facilitate, enrich and replace traditional data sources and analytical tools. proposes to use new techniques to study the mechanisms of scientific collaboration and the composition of research teams; analyze innovation networks; following knowledge and ideas while communicating between scientists, engineers and entrepreneurs; and study the intri-

cate nature of university-industry links and collaborative innovation [9].Moscoso in 2016, brings together a wide range of techniques and algorithms that allow the extraction of knowledge from databases for decision making using data mining. Which have been applied to different fields of study. Focusing on an important research field such as education. The application of data mining in education is known as educational data mining (EDM). The main objective of EDM is to analyze data from educational institutions using different techniques such as: prediction, grouping, time series analysis, classification, among others. This paper presents a holistic view of EDM that includes the classification of algorithms, methods and tools used in data mining processes. In addition, the processes and indicators that could be improved are analyzed in educational institutions. This study covers papers submitted from 2005 to 2015 [12].

3 Theoretical Background

3.1 Web Scraping

Web Scraping is the practice of collecting data through any means other than a person, which is usually a program that interacts with an API. This is generally achieved by writing an automated program that consults a web server, requests data and analyzes that data to extract the necessary information.[3]. For data extraction, there are a set of libraries that help us in this process, among them are: Jsoup, Scrapy, etc. Scrapy is an open source library that is developed and works with Python, which generates a structured project, and which is optimized for Scraping tasks. It can be used for a wide range of purposes, from data mining to automated monitoring and testing [16].

3.2 Data Mining

Data mining is a discipline that has emerged at the confluence of several other disciplines, driven primarily by the growth of large databases. The basic motivating stimulus behind data mining is that one looks for surprising, novel, unexpected or valuable information, and the goal is to extract this information. This means that the subject is closely related to the exploratory data analysis. However, problems arising from the size of databases, as well as ideas and tools imported from other areas, mean that data mining is more than just an exploratory data analysis. [4].

Data filtering From the set of data collected and already defined the objectives that we want to achieve, we proceed to choose available data to carry out the study and integrate them into one that can favor reaching the objectives of the analysis. Many times this information can be found in the same source (centralized) or can be distributed [5].

Variable selection The selection of variables is a very important part, even after having been preprocessed, in most cases there is a large amount of data. The selection of characteristics reduces the size of the data staying with a vector of k-dimensions, choosing the most influential variables, without sacrificing the quality of the knowledge model obtained from the mining process[5]. The methods for variable selection are [5]:

- Those based on the choice of the best attributes of the problem.
- Those looking for independent variables through sensitivity tests, distance or heuristic algorithms.

Knowledge Extraction Algorithms Knowledge extraction in databases (KDD) is "the non-trivial process of identifying valid, novel, potentially useful and, ultimately, understandable patterns from the data" Data mining only constitutes A stage of this process whose objective is to obtain patterns and models by applying statistical methods and machine learning techniques. Finally, the process of knowledge extraction also involves the evaluation and interpretation of the patterns or models obtained in the data mining stage [10]. Within the knowledge extraction algorithms, within Machine Learning, we have the following types of learning:

- **Supervised Learning** Supervised learning is a learning model created to make predictions, where given a set of input data, your output responses are known. [5].
- **Unsupervised Learning** Unlike supervised learning, unsupervised learning finds certain patterns that exist in the input data, so there is no information on the category of the input data [5].
- **Semi Supervised Learning** This learning technique is the combination of supervised and unsupervised learning. The objective of semi-supervised learning is to classify some of the unlabeled data using the set of labeled information.

Interpretation and evaluation In this phase of *Data mining* it is verified if the results are consistent. Once the learning model is obtained, it must be validated, checking that the conclusions it produces are valid and sufficiently satisfactory. If several models are obtained by using different techniques, the models should be compared in search of the one that best fits the problem [1].

3.3 Dimensionality Reduction

Dimensionality reduction refers to the process of mapping an n-dimensional point, in a lower k-dimensional space. This operation reduces the size to represent and store an object or a set of data in general [5]. The dimensionality reduction is divided into two categories, Selection and Extraction of Characteristics, where the first one chooses a subset of characteristics with one criterion,

and the second one transforms the data of high dimension into data of low dimension. The reduction is very important since having a large amount of data and examining text strings, these can become k-dimensional that can cause processing to delay.

Ant Colony Optimization Algorithm The reduction is very important since having a large amount of data. In 1992, Marco Dorigo, in his PhD thesis proposes an algorithm based on the behavior of ants, in search of food, being its first application in the problem of the traveler [2]. Ants in the real world wander randomly in search of food, they are almost blind, so the way to communicate with each other is through pheromones. By randomly wandering from their nest to the food source, they leave their pheromone trail until they find their food, and return to the nest. Since other ants are found around them, they persist in places that are most traveled by the ant that has found its way to food.

3.4 Support Vector Machine

An SVM (*Super Vector machine*) is a discriminative classifier formally defined by a separation plane. In other words, given the training data labeled (supervised learning), the algorithm generates an optimal hyperplane that categorizes new examples. In two dimensional spaces, this hyper-plane is a line that divides a plane into two parts where each class is on each side [13]

3.5 Display

Data visualization is the presentation of data in illustrated or graphic formats. Allowing people to see the analytics presented visually, so that they can capture complicated concepts or identify new patterns. With interactive visualization, you can take the concept one step further using technology to deepen diagrams and graphs to observe more detail, interactively changing what data you see and how it is processed [7].

4 Proposal

This section describes the proposal to search for educational trends in research based on Data mining techniques, below in Fig.1, The whole procedure is shown.

4.1 Data collection

For data collection, the ERIC database has been selected, which provides us with scientific articles related to the area of Education, this first stage is divided into three parts:

- *Web Crawling*: The website of the following website is inspected [6]
- *Web Scrapping*: The information is extracted according to the website.
- Save information: the extracted information is stored in the database, with the following fields: “title” ,“category” ,“year” ,“authors” ,“urlsource” ,“description”

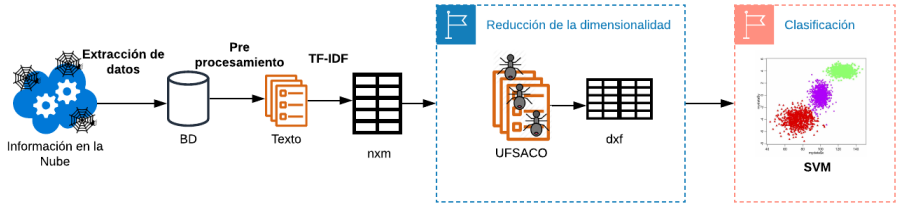

Fig. 1. Proposed Architecture

4.2 Pre-processing Text

To perform the pre-processing of the Text, the methods of:

- Tokenization: is the process of segmenting text into called words *tokens* and at the same time also discard punctuation marks.
- *Stop-Words*: they are common words of a language like, "the", "a", "is", etc. These words are irrelevant in word processing.
- *Semming*: This process reduces the words to their root form.

After performing these steps, we will proceed to apply the algorithm of *Term Frequency - Inverse document Frequency* (TF-IDF), with which we will obtain how relevant each word is in the document, where "t" is the term, "d" the document and "D" is the set of documents. Applying the multiplication of these two values will give us a score, the higher the score is then the more relevant is that word in the document.

$$tf * idf(t, d, D) = log(1 + freq(t, d)) * log(\frac{N}{count(d \in D : t \in d)}) \quad (1)$$

- *Term Frequency*: the frequency of a term is denoted tf(t, d), it is how frequent a term "t" is in the document "d".
- *Inverse document Frequency*: Indicates how common a word is in a whole set of documents. It is calculated by taking the total number of documents ("N") and divided by the number of documents that contain a word.

The pre-processing output will be the ant colony algorithm input, this entry is of the following form:

[('computer', 0.0651), ('learn', 0.1789), ('web', 0.0601)]

Selection of Unsupervised Features For the selection of characteristics, the bioinspired ant colony algorithm is applied, for this, before carrying out the characteristic selection process, we will create an unguided graph, denoted by $G = (F, E)$, where F are the characteristics and E are the edges, to find the value of the edges the similarity of the cosine between characteristics is used (2).

$$S_{A,B} = |\frac{\sum_{i=1}^{p}(a_i b_i)}{(\sqrt{\sum_{i=p}^{p} a_i^2})(\sqrt{\sum_{i=p}^{p} b_i^2})}| \quad (2) \qquad \frac{1}{S_{A,B}} \quad (3)$$

Where A and B are two characteristics of dimensionality "p", according to the equation the value of similarity ranges between 0 and 1 if the characteristics are similar, 1 is obtained, otherwise 0. After having the graph, the Ant Colony Optimization algorithm is applied, this algorithm has two important characteristics, the first is its "Heuristic Information" and the second is its "desirability". The Heuristic Information is defined as the inverse of the similarity between characteristics, that is (3) and the desirability is the amount of pheromones, this desirability is denoted as τ.

4.3 Learning model

Once the dimensionality of the feature vector has been reduced, it serves as input to our algorithm of *Clustering*, which in our case we are using *K-Means*, this in order to find common characteristics, and that can be grouped, to be able to visualize and interpret the results of the algorithm. And as part of the verification of the results obtained, we have applied the SVM supervised learning algorithm, with them we verify that the labels generated for each document have coherence and their corresponding classification.

5 Results

This section describes the experiments performed applying the clustering algorithm *K-Means*, to see the grouping of scientific articles related to educational trends, and the application of the algorithm of *SVM*, to validate the learning model.

Database The database used in this work is a compendium of ERIC - Education [6], where a taxonomy in education has been proposed, based on educational trends in the year 2019 [18]. To create our database is that we create a pivot of search start from 20 kinds of Educational Trends, which allows us to obtain a large number of scientific articles related to education, this was done because ERIC, can not be performed blank searches.
Parameter Settings: The following proposed parameters have a maximum number of cycles numCi = 10, the amount of ants will be equal to the number of threads numHor = numHeb, the initial amount of pheromone for each characteristic is $\tau_i = 0.2$, in the same way the evaporation coefficient will be $\delta = 0.2$, the parameter q_{ini} it will be equal to 0.7 with which the exploration and exploitation value will be controlled, the value of β Indicates the importance of pheromone. According to the database collected, a maximum number of 50 features will be available.

Results: The results obtained by applying the parallelization of the PUFSACO algorithm (*Parallelization unsupervised future selection based on Ant Colony Optimization*). The tests were performed on an HP computer, Intel Core i7. The methods were written in c ++ which runs on Ubuntu 18.04.2 LTS. In the experiments, the ROC curve is used to measure performance. In this work, one third of the database was used for the test stage (41,160 data). WEKA software is used to classify the text, for which the SVM algorithm is applied (*Support Vector Machine*), the kernel used is polykernel, this algorithm is only used as validation.

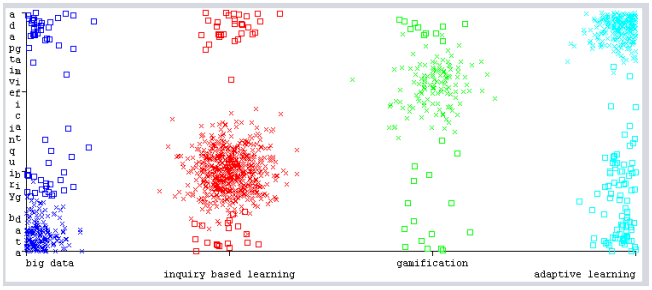

Fig. 2. Classification applying the SVM algorithm in Weka

The ROC curve shows the balance between sensitivity (or True positives) and specificity (1 - False positives). The classifiers that give curves closer to the upper left corner indicate a better performance, in the tests carried out there is a ROC of 0.9. As part of the results is that we use the visualization of information in order to graph in a faster and more concise way the quantity of labels that were collected referring to a single topic, the visualization technique is known as word cloud or cloud of labels, you can enter for review in the following link http://tendenciaseducativas.rf.gd/. Below is the graphic representation of the educational taxonomy and the tags and their categories, in tagging clouds or *tag cloud*, in the same measure the labels, related to education, are displayed.

[a] 

[b] 

Fig. 3. a)Behavior of the labels. b)Behavior of educational categories

and finally the visualization of the information referring to the years that were published and their quantity, and linked to the category they belong to.

Fig. 4. Behavior of the information with the relation quantity and time, own elaboration

6 Conclusions

The results of this investigation that in its first stage of Data Collection, Web Crawling was used to inspect websites and Web Scraping to extract the information, can be visualized in a graphic repository of educational trends of type word cloud or tags, which They allow us to better understand how they were grouped by categories and what relationship they have in the number of searches with the search year, you can enter for review in the following link http://tendenciaseducativas.rf.gd/. An important part of the research is also aimed at reducing dimensionality, using bioinspired algorithms, helping to reduce the large amount of information, leaving us with relevant information.The unsupervised learning application shows us that it can help us discover information that does not stand out with the naked eye, but when processed by this type of algorithm, it allows us to notice more relevant information, and that it can be applied to the taking of decisions.

7 Acknowledgment

The present research work was carried out within the framework of the research project IBA-0029-2016 " Technological Surveillance Services for research centers and Technological Innovation Classroom, Oriented to the Development of R + D + I Projects in ICTs and Education " , we express our deepest gratitude to the Universidad Nacional de San Agustín, for making this study possible.

References

1. Aparna U.R., Paul, S.: Feature selection and extraction in data mining. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET). pp. 1–3 (Nov 2016). https://doi.org/10.1109/GET.2016.7916845
2. Colorni, A., Dorigo, M., Maniezzo, V.: An investigation of some properties of an ant algorithm. In: Proc. Parallel Problem Solving from Nature Conference. pp. 509–520 (1992)
3. Haddaway, N.R.: The use of web-scraping software in searching for grey literature. Grey J **11**(3), 186–90 (2015)
4. Hand, D.J.: Data Mining Based in part on the article "Data mining" by David Hand, which appeared in the Encyclopedia of Environmetrics. American Cancer Society (2013). https://doi.org/10.1002/9780470057339.vad002.pub2, https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470057339.vad002.pub2
5. Herrera, F., Charte, F., Rivera, A.J., Del Jesus, M.J.: Multilabel classification. In: Multilabel Classification, pp. 17–31. Springer (2016)
6. Institute of Education Sciences: Eric, https://eric.ed.gov/
7. Jain, A.K.: Data clustering: 50 years beyond k-means. Pattern recognition letters **31**(8), 651–666 (2010)
8. KDnuggets: Idc study: Digital universe in 2020, https://www.kdnuggets.com/2012/12/idc-digital-universe-2020.html
9. Klochikhin, E.: Collaborative innovation beyond science: Exploring new data and new methods with computer science (2016)
10. Mariñelarena-Dondena, L., Errecalde, M.L., Solano, A.C.: Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. Revista Argentina de Ciencias del Comportamiento **9**(2), 65–76 (2017)
11. Meschenmoser, P., Meuschke, N., Hotz, M., Gipp, B.: Scraping scientific web repositories: Challenges and solutions for automated content extraction. D-Lib Magazine **22**(9/10) (2016)
12. Moscoso-Zea, O., Luján-Mora, S.: Educational data mining: An holistic view. In: 2016 11th Iberian Conference on Information Systems and Technologies (CISTI). pp. 1–6. IEEE (2016)
13. Ramli, M.A., Twaha, S., Al-Turki, Y.A.: Investigating the performance of support vector machine and artificial neural networks in predicting solar radiation on a tilted surface: Saudi arabia case study. Energy conversion and management **105**, 442–452 (2015)
14. Santos, A.R.P., Carreño, J.D., Pinto, Y.A.S.: Infoxicación y capacidad de filtrado: Desafíos en el desarrollo de competencias digitales. Etic@ net **18**(1), 102–117 (2018)
15. Schindler, L., Puls-Elvidge, S., Welzant, H., Crawford, L.: Definitions of quality in higher education: A synthesis of the literature. Higher Learning Research Communications **5**(3), 3–13 (2015)
16. Scrapy: Scrapy, https://scrapy.org/
17. Srivastava, M., Garg, R., Mishra, P.K.: Analysis of data extraction and data cleaning in web usage mining. In: Proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology (ICARCSET 2015). pp. 13:1–13:6. ICARCSET '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2743065.2743078, http://doi.acm.org/10.1145/2743065.2743078
18. teachthought: 30 of the most popular trends in education, https://www.teachthought.com/the-future-of-learning/most-popular-trends-in-education/