J Knowl Econ (2016) 7:461–487
DOI 10.1007/s13132-014-0219-2

How MOOCs Link with Social Media

Dimitrios Kravvaris · Katia Lida Kermanidis · Georgios Ntanis

Received: 1 February 2014 / Accepted: 16 June 2014 /
Published online: 23 July 2014

Abstract The purpose of this paper is to present a survey conducted on massively open online courses (MOOCs) from Coursera and how they are linked with social media. It examines the web data that have been retrieved from Coursera's MOOCs information pages that can be recommended by the users of the social networks and, in turn, be shared by them. What should be stressed, however, is that our focus is on the study of those data that are open and accessible to everyone and not only to registered users of MOOCs. The aim of our study, therefore, is to find out the attributes of these information pages that can characterize a course popular and those that are considered to be the most important for the users' recommendation procedure. It is shown that the courses providing information about the assignments and the exams of the course are mostly recommended in the social media. Furthermore, we proved the correlation among the three largest social networks: Facebook, Google+, and Twitter, based on the information pages' data, using statistical and machine learning methods. Finally, statistical experiments were carried out concerning the MOOCs users' shares to social media.

Keywords MOOC · Social media · Clustering · Attribute selection · Correlation · Regression

Introduction

The rapid growth of the Internet, as well as the increase in the number of its users, have made imperative the use of new methods of processing the enormous amount of Web data, i.e., data mining and machine learning methods. The aforementioned methods of processing Web data have created new study areas such as educational data mining and social data mining. Many of our everyday activities take place on the Web, such as communicating with others through social media and learning from online courses. The human interaction with the Web pages is a very interesting area of study. It can reveal

D. Kravvaris (✉) · K. L. Kermanidis · G. Ntanis
Department of Informatics, Ionian University, Tsirigoti Square 7, 41900 Corfu, Greece
e-mail: jkravv@gmail.com

useful information concerning the context of the Web pages that users prefer to read and the hyperlinks that they like to follow or share with others.

In this research, we focus on the study of content data that are retrieved from the information pages of massively open online courses (MOOCs) from Coursera (Coursera 2014) and how those are linked with social media. More specifically, we examine if there are certain characteristics in the information pages which can make a lesson likable to the page visitors, who, in their turn, will share it through their social network, and finally, what these characteristics are. Furthermore, we examine the relation between the social networks concerning the promotion of MOOCs by their users and we study the possibility to extract conclusions from a social network, knowing what has happened to the other social network. Finally, we analyze the university popularity and the course categories in the social media based on shares of these courses. All the above, apply to the field of MOOCs finances, e.g., not only when the number of registered users increases, but also to the scientific field of open data, e.g., examining the sharing of open data in the social media.

The Massachusetts Institute of Technology, with the introduction of OpenCourseWare in 2004, made open source education known to the public (Cecil-Reed 2013), while in the past year, MOOCs' popularity rapidly increased according to studies (Pappano 2012; Anderson 2012). The MOOCs can enroll all the students that want to attend a lesson, as opposed to traditional online courses that enroll a small number of students. The MOOCs do not carry credits and they are usually free, whereas, traditional online courses carry credits and in many cases cost money.

The largest MOOCs provider worldwide is Coursera. It provides very popular courses and over 5 million users from 190 countries have registered so far (Coursera 2013a; Waldrop 2013). Up to the time this study was carried out, Coursera had offered 333 courses from 62 universities in 17 countries. The course categories cover a wide range of subjects such as Computer Science, Humanities, Economics and Finance, Mathematics, Information Tech and Design, Medicine, and Physics.

Coursera's site visitors can read the information page of the course and get a variety of information about it, in order to decide if they will enroll or not to the course. The site's visitors (registered or not) can publish the MOOC information page on their favorite social network (Facebook, Google+, Twitter), just by clicking "Like," "G+ 1," or "Tweet" buttons that appear in the MOOC's information page. Clicking a button sends a recommendation notice to the user's social network, informing all the user's friends/contacts/followers about it. The number of those clicks makes the MOOC less or more recommended, through social networks, depending on the number of those clicks.

Previous relevant research about MOOCs focuses mainly on the courses' format and content (Rodriguez 2012; Koutropoulos et al. 2012) and their pedagogical role (Mak et al. 2010; de Waard et al. 2011). More analytically, what was studied were the structure of the popular courses, the students' comments, and the communication tools that are used while new methods of collaborative learning and knowledge construction were suggested.

Our research concerns the exploration of data that come from MOOCs' information pages using data mining methods. The focus of our study is (a) the extraction of

information from open data that concern education and, more specifically, MOOCs and (b) the comparison of the three largest social networks and, more specifically, their differences and similarities focusing on their users' shares for a specific research field (MOOCs). Our study's innovations regarding MOOC mainly concern:

- Data
  The data used come from MOOCs' information pages and are available to the users without enrolling to the course. In all the previous research studies, the users had to register in order to obtain their data of interest. These data represent the exact information the users have at their disposal in order to decide whether they will attend the course or whether they will share it through their social network.
- Human action
  This research focuses on this set of human actions that have an impact on social media. This action is recorded by the press of a button. More specifically, a user can click a button on Facebook, Google+ or Twitter, and in this way, they can share the MOOC's information page with their friends in social media.
- Social media
  We study the popularity of MOOCs in the three most popular social networks so as to find the features of the information pages that play a determining role in the recommendation procedure in the social media.

Furthermore, we study these information pages from both of the following aspects:

- Statistics
  We conduct a further detailed statistic study than the one available from Coursera and we link it with the social networks.
- Economy
  We refer to relevant bibliography that concern MOOCs and we relate it to the findings of our research.

The first section of our paper concerns the theoretical background of our study. More analytically, we present (a) an overview of the MOOCs' economy, since it is the driving force in the following detailed analysis; (b) a presentation of related work in the fields of educational and social mining; and (c) the research questions of our study and how it differs from other studies in the field. The second section focuses on the methodology of our experiment. At the beginning, we analyze the web data retrieved from the MOOCs information pages, we explain their representation as a set of feature-value vectors, necessary for the machine learning experiments that have been conducted. We also present in detail the attributes that this dataset is structured of and then we analyze the three major axes of clustering, of attribute selection, and of correlation–regression analysis concerning our data. The results of these experiments are shown in the third section. The fourth section shows statistical results and an analysis between the number of enrolled users and the number of shares in social media that have been made per course category. Finally, the last section of our study, presents interesting conclusions concerning the MOOCs economy.

Theoretical Background

Economy

MOOCs created a new market for online education. The fact that world-famous universities provide free online courses to all students that want to attend has motivated the interest of many professionals in the educational industry. Bersin (2013) estimated that there are more than 2 billion potential students around the world today. This number attracted the interest of the biggest social networks. Perault, Facebook's head of global policy development, says that a future collaboration with the providers of massive open online courses will help the courses to spread effectively worldwide (Collins 2013).

Are the courses honestly free? The attendance, the assignments, and the final exams are free. However, the certification of attendance of the courses is not free. The prices of the certificates in Coursera vary from US$30 to US$100 depending on the course and the university that provides it[1]. For the time being, there is no correlation between courses and academic credits, but new information shows that this will happen soon[2], increasing, thereby, the certification cost.

Another financial model that is being promoted is "take a lesson here and get certified somewhere else." In this model, the students attend courses through MOOC's but get certified by a different educational institution, other than the one offering the course, paying fewer fees. Thereby, students have the chance to attend a high-standard course, and pay a reduced certification cost (Rafter 2013).

Employees constitute the majority of the registered users in MOOCs[3]. They have the opportunity to attend a high-standard course that will help promote their existing career or turn them to a new occupational direction. Moreover, MOOCs content analysis can improve the recruiting process by giving companies and employers access to information about their students' qualifications (Dellarocas and Van Alstyne 2013).

Lastly, the advent of MOOCs created new professions, like companies that provide online help for the completion and certification of the courses (MOOCs Mentor 2013). Such companies sell a wide range of services, like one-to-one mentoring and dedicated helplines, in order to assist aspirants of MOOC courses.

Related Work

As was also mentioned in the "Introduction", MOOCs have created a new study field, which researchers examine from different views/aspects. Thus, research topics related to MOOCs vary. There is, for example, some research that concerns the acceptance of MOOCs by the Internet users and the participants' characteristics (Rayyan et al. 2013; DeBoer et al. 2013), there is also research concerning MOOCs design and their adaption to the various learning styles (Grünewald et al. 2013; McAndrew 2013), there is some research that examines the users' contribution in the lesson forums (Huang

[1] Information retrieved from Coursera's blog article "Introducing Signature Track" (Coursera 2013b)
[2] According to American Council on Education (2014)
[3] Information retrieved from Coursera's blog article "A Triple Milestone" (Coursera 2013a)

et al. 2014), and finally, there is some other interesting research regarding peer assessment grading and the way of delivering accurate results (Piech et al. 2013).

This innovative research can help the spread of MOOCs through the social media as well as their financial analysis. Moreover, it can complete, reinforce, and move on the already existing research as follows.

The research titled *Research Issues in Web Data Mining* (Madria et al. 1999) suggested that Web data mining can be divided into three main areas relating to content, structure, and usage. The advancement of the Internet and the study of human behavior on it urge us to focus on the specific areas. Our research constitutes a study of Web content mining specialized in the field of MOOCs, adding some very interesting findings concerning the specific research area.

The research of Veeramachaneni et al. (2013) suggests the developing of data standards for MOOC, and the creation of analytic scripts that extract multiple attributes for a course. The aforementioned research could be completed with our attributes that were extracted from the information pages of the courses, since they constitute substantial criteria for the choice of the courses that the user is to register to.

Gundecha and Liu (2012) studied the hypothesis that people who are connected through social media share the same or similar interests, they prefer their friends' recommendations and they can be easily influenced by them. The objective of this research is to improve the quality of recommendation in social media. With our research, we study the choice of the users to share a MOOC with their friends, showing the attributes of the MOOC information pages that are more important in that procedure. In this way, we contribute to the qualitative comprehension of the recommendation procedure.

The research titled *A Reference Model for Learning Analytics* (Chatti et al. 2012) describes a reference model for learning analytics. It studies the social network analysis methods in different learning analytics tasks, aiming at the quantitative study of the relationships between individuals or organizations. We propose the analysis not only for one social network but for the three largest ones worldwide (Facebook, Google+, Twitter) concerning the MOOC information pages. What is more, we examine the relations between the Coursera and the users of these three social networks.

Finally, Aguillo (2010) studied the universities' Web ranking and proposed that a large number of variables should be taken into account. In the specific research, it is mentioned that there are no reliable sources of data from the universities. We, however, with our research, solve this problem, suggesting that we can use open data provided by third parties, such as Coursera, and, thus, rank not only universities, but also course categories as well, using a social media prospect.

Research

Our research focuses on MOOCs' open data. Our first goal is to locate the appropriate data for our research from MOOCs information pages and to describe them analytically. We focus on the characteristics that have the greatest value for the Web pages' visitors, who, in their turn, will share them through the social media. Contrary to other studies, which were described earlier, we focus on data which do not require users to register in order to have access to them. Furthermore, there is a triple examination of the social

networks since we examine the shares of the courses on Facebook, Google+, and Twitter. Another important part of our study is the scientific examination of the relation between the three social networks under the same field of research. While the usual procedure is to examine the social media separately, we move a step forward and through our research, we answer the question of whether the examination of an issue in any other social network would present the same results irrespective of the social network that is being examined.

Methodology

Data

We created our dataset, retrieving heterogeneous data from the information pages of each course. These data were relevant to the structure, the content per section and the introduction video of each MOOC information page. We have a total of 320 records (examples); this is the number of the courses Coursera has introduced so far (excluding those in the Chinese language). Each record in our dataset represents a MOOC information page. The attributes of each record concern the heterogeneous data of the pages, which we extracted and categorized under 19 attributes. There are no missing values in our dataset, avoiding missing data that may adversely affect the experimental results. The attributes chosen represent the exact information the visitor of the MOOC information page has, at their disposal, according to which they will choose to recommend the specific course on their social network. Below, we present these attributes divided in categories, referring to their contribution to our research:

- Creation
  This category contains the attributes that concern details relevant to the creation of MOOC and have an impact on their choice of the users (Thakur 2007). These are as follows:

– "University" is an attribute containing the name of the university that implements the course. There are 60 different universities.
– "Number of Teachers" is a numeric attribute that contains the number of instructors of each course.

- Content
  This category contains the attributes that concern details relevant to the course and that are of special importance since according to a survey (Murray et al. 2012), the student selectively attends according to the degree that they will acquire knowledge. These are as follows:

– "Category" is a nominal attribute that describes the category of each course. There are 21 different categories.
– "About the Course" is the number of words that constitute the course description text. This is a numeric attribute.

- "Text Complexity" is a numeric attribute denoting the complexity of the "About the Course" section text of the MOOC's information page, i.e., a higher number means more complex text, according to Lexile Analyzer (Lennon and Burdick 2004).
- "Previous Background": several courses are intended for people who already have some background on this subject. In most courses, there is a separate section that describes this background. This attribute relates to whether or not previous background on this subject area is required.
- "Introduction Video": in most courses, there is an introductory video. This attribute mentions the time of this video in minutes. If there is no available video, the value is zero.
- "Number of FAQ" is a numerical attribute that contains the number of frequently asked questions for the course.
- "Section in Page's Structure" is the number of sections that describe a course (e.g., About the Course, Background, FAQ, and more).

• Time

This category contains the attributes that relate to time. These attributes are very important for the choice of the course, since they differ according to the needs of the students (Delfino and Persico 2007). These are as follows:

- "Duration" is the time length (in weeks), which is required for the completion of the course.
- "Workload" is the average time per week that is required for studying for the course. This is a numeric attribute.
- "Time started" is a nominal attribute about the course start date, that consists of three different values (the course has already started, the course has not started yet, or we have no information when the course starts).

• Assessments

The attributes of this category describe the work needed for the completion of the course. Moreover, they help in the consolidation of the course as well as in the active participation of the student in the course (Vrasidas and McIsaac 1999). These are as follows.

- "Online quizzes": in Coursera, there is a possibility for some video courses to include quizzes for the users to answer. This attribute takes one of the following values: (a) Yes (the course has online quizzes), (b) No (the course has no online quizzes), and (c) Don't know (there is no information about it).
- "Offline assignments" is a nominal attribute denoting whether or not a course has offline assignments. More specifically, this attribute takes one of the following values: (a) Yes (the course has offline assignments), (b) No (the course has no offline assignments), and (c) Don't know (there is no information about it).
- "Final exam" this is another nominal attribute denoting whether or not a course has final exams. More specific, this attribute takes one of the following values: (a) Yes (the course has final exams), (b) No (the course has no final exams), and (c) Don't know (there is no information about it).

- Certification

  It refers to the attribute that describes the possibility of certification in the specific course. This is of special importance since, as we mentioned in the "Economy" section, it triggers the interest of the users that want to acquire a certificate and internally motivates the user to continue with the course (Frankola 2001). This attribute is the following:

– "Signature Track" is a Boolean nominal attribute. Anyone enrolling in this track must pay to have extra services, namely to obtain a verified certificate and to be an official Internet student of this course.

- Social Media

  This category includes the findings of the shares of the users in the social media which can lead new users to register to courses (Domingos 2005). These are as follows.

– "Twitter" is a numeric attribute that contains the number of Twitter hits (shares on a twitter account) each course has.
– "Google+" is a numeric attribute that contains the number of Google+ hits (shares on a Google+ account) each course has.
– "Facebook" is a numeric attribute that contains the number of Facebook hits (shares on a Facebook account) each course has.

Datasets

We divide the initial dataset in order to create three new ones; each one has a classification attribute: Facebook, Google+, or Twitter, correspondingly. Because these attributes are numeric, we discretize them into two categories (binary values) depending on whether they have many clicks or not. Since it is subjective to say whether the actual number of clicks is high or not, and it differs in every subject under study, two well-known mathematical methods are implemented to determine the threshold, i.e., the calculation of (a) the mean and (b) the median (Srivastava et al. 2000) value of the attributes. Finally, six datasets were produced with a binary classification attribute as follows in Table 1. If the value of the attribute is higher or equal to mean or median value, respectively, then the MOOC is categorized as *more* (recommended); otherwise, it is categorized as *less* (recommended).

Table 1 Mean and median values of the datasets

Attribute	Dataset name	Mean	Dataset name	Median
Facebook	Dataset 1	2,239	Dataset 2	743
Google+	Dataset 3	412	Dataset 4	88
Twitter	Dataset 5	362	Dataset 6	124

Clustering

In the experimental procedure, we used WEKA version 3.6.6 software (Weka 2013). The clustering experiments used all six datasets, employing unsupervised learning methods. More specifically, a centroid-based clustering algorithm (Kanungo et al. 2000) using SimpleKMeans, with the Manhattan distance function (Han et al. 2006) has been used. We chose SimpleKMeans because it is a simple and flexible algorithm that is easy to understand and explains the clustering outcome (Vora and Oza 2013). Two clusters were chosen for the value of K (in *K-means*), in order to enable "classes to clusters evaluation" (Färber et al. 2010). Thereby, a correlation analysis between the different values of the attributes (i.e., the clusters formed) and the popularity of the courses (i.e., the two class values: more/less popular).

Attribute Selection

In the second part of our experiments, we performed an attribute selection following a wrapper method which uses a subset evaluator, in order to find which attributes are important (Hall and Holmes 2003) for each of the six datasets described above. In our setup, all datasets were normalized, and the Weka ClassifierSubsetEval process (Indra Devi et al. 2008) (with the sequential minimal optimization (SMO) classifier and BestFirst as the search algorithm) was employed for attribute selection (Pitt and Nayak 2007). The SMO classifier refers to John C. Platt's sequential minimal optimization algorithm implementation, for training support vector machines (SVM) (Hearst et al. 1998) classifier in Weka. The SVM algorithm was selected because it suits our data. More specifically (Romero et al. 2011), (a) it can create a general purpose model, (b) it can handle non-linear class boundaries, and (c) it is accurate on small datasets.

Excluding Classification Attributes

The attributes Facebook, Google+, and Twitter were excluded from the first part of the attribute selection procedure, since they have already been used as classification attributes for the datasets. In this way, datasets 1 and 2 include all the attributes apart from the Google+ and the Twitter attribute, datasets 3 and 4 include all the attributes apart from the Facebook and the Twitter attribute, and datasets 5 and 6 include all the attributes apart from the Facebook and the Google+ attribute.

Including Classification Attributes

In the second part of our experiment, Facebook, Google+, and Twitter are included in the attribute selection procedure as classification attributes, so the object of our study is whether there is an interrelation among them and whether the findings of the previous procedure change in any way.

Correlation and Regression Analysis

The first aim of those experiments is to find out whether the classification attributes correlate, examining the following three pairs: (a) Facebook and Twitter, (b) Facebook

and Google+, and (c) Google+ and Twitter. Thus, a new dataset was created that consisted of the Facebook, Google+, and Twitter attributes. One of the most common measures of correlation in statistics is the Pearson correlation, which shows the linear relationship between two variables (Norusis 2008). Correlation between variables (classification attributes) is a measure of how well the variables are related. Furthermore, a linear regression analysis will be executed in the three cases in order to examine whether it is possible to predict (with relative accuracy) the value of one of the attributes of social networks given the value of the other attribute of another social network (Montgomery et al. 2010).

Additional experiments were carried out aiming to study the correlation among all three social media. In those experiments, we proceeded with the classical linear regression analysis in Weka (Norusis 2008) and then used the random forest algorithm (Liaw and Wiener 2002), since it can give us great results according to research (Caruana and Niculescu-Mizil 2006).

The random forest algorithm is combined with the RegressionByDiscretization method (Robnik-Šikonja 2004). The RegressionByDiscretization can parameterize the number of categories which can be used in our experiment. We chose to have equal number of categories to the number of district values of Facebook, Google+, and Twitter hits, respectively. The parameterization of random forest was set to default, using the number of trees in the forest (numTrees) parameter equal to 10. For all the experiments we used the 10-fold cross-validation technique.

The last aim of our experiments is to find out whether the classification attributes correlate with all the other attributes of our dataset. We followed the same approach as described above implementing the random forest algorithm. Those experiments are very interesting since we try to predict the outcome of a human action (represented by the classification attributes) based on the MOOCs' information pages data.

Experimental Setup

Clustering

In the first set of clustering experiments, we used *Facebook* as a clustering evaluator. The attributes that differ considerably between the two clusters along with their values appear in Table 2. These attributes describe the way each course is assessed (*Online*

Table 2 Clustering using Facebook attribute as class to cluster evaluator

	Dataset 1		Dataset 2	
Attributes	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Online quizzes	Yes	Don't know	Yes	Don't know
Offline assignments	Yes	Don't know	Yes	Don't know
Final exam	Yes	Don't know	Yes	Don't know
Number of FAQ	3	1	3	1
Sections in page's structure	6	4	6	4

Table 3 Clustering using Google+ attribute as class to cluster evaluator

	Dataset 3		Dataset 4	
Attributes	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Online quizzes	Yes	Don't know	Yes	Don't know
Offline assignments	Yes	Don't know	Yes	Don't know
Final exam	Yes	Don't know	Yes	Don't know
Number of FAQ	3	1	3	1
Sections in page's structure	6	4	6	4

Quizzes, *Offline Assignments*, *Final Exams*) and the amount of information related to the course (*Number of FAQ*, *Sections in Page's Structure*). The first cluster that includes the "more" popular courses is the one where courses contain specific information concerning the course assessment and more details than courses contained in the second, "less" popular courses, cluster. As far as dataset 1 is concerned, the percentage of the correctly classified instances is 59.06 %, while for dataset 2, it is 53.75 %.

Another clustering process was performed using *Google+* as a clustering evaluator. The attributes that differ considerably between the two clusters along with their values appear in Table 3. It is found that the results are exactly the same as the ones in the previous experiment. The only difference is that the percentage of the correctly classified instances for dataset 3 is 61.56 %, while for dataset 4, it is 52.50 %.

Finally, a clustering procedure was carried out using *Twitter* as a clustering evaluator. The attributes that differ considerably between the two clusters along with their values appear in Table 4. In this case it was also observed that the cluster with the "more" popular courses has more useful information for the visitors of the MOOC Information Page than the cluster of the "less" popular courses. As far as dataset 5 is concerned, the percentage of the correctly classified instances is 57.81 %, while for dataset 6, it is 52.19 %.

Comparing the findings of the three sets of experiments, we conclude that the second set used as clustering evaluator the *Google+* attribute has better results. Studying the findings of the experiments, we find out that the *Workload* attribute differs by one (working hour per week) in the cluster that has less recommended compared to the corresponding clusters of the other two sets of the experiments. Thus, we can conclude

Table 4 Clustering using Twitter attribute as class to cluster evaluator

	Dataset 5		Dataset 6	
Attributes	Cluster 0	Cluster 1	Cluster 0	Cluster 1
Online quizzes	Yes	Don't know	Yes	Don't know
Offline assignments	Yes	Don't know	Yes	Don't know
Final exam	No	Don't know	No	Don't know
Number of FAQ	3	1	3	1
Sections in page's structure	6	4	6	4

that besides the attributes of the tables above, the workload attribute also affect the popularity of MOOC.

Valuable Attributes

The findings of the attribute selection procedure are presented below under two categories: (a) excluding classification attributes and (b) including classification attributes. As far as the search algorithm is concerned, in addition to BestFirst, GreedyStepwise and LinearForwardSelection (Kirkby et al. 2006) were used in the tests, but there was no difference in the results.

Excluding Classification Attributes

The aforementioned procedure has been conducted for all six datasets excluding the classification attributes and a graph was created showing the frequency appearance of the attributes in all six cases, as shown in Fig. 1. The most important attributes in each case are *Category* and *University*. We can see the attributes that gives to users' information about the content and the creator of the MOOC playing an important role for the information pages analysis.

Including Classification Attributes

The same procedure was followed for all six datasets, this time, including the classification attributes and the graph shown in Fig. 2 were created, showing the frequency appearance of each attribute in all six cases.

The most important attributes in each case are: *Category*, *Facebook*, *University*, *Google+* and *Twitter*. We should stress the fact, however, that the Facebook, Google+ and Twitter attributes altered the previous findings and there seems to be a correlation

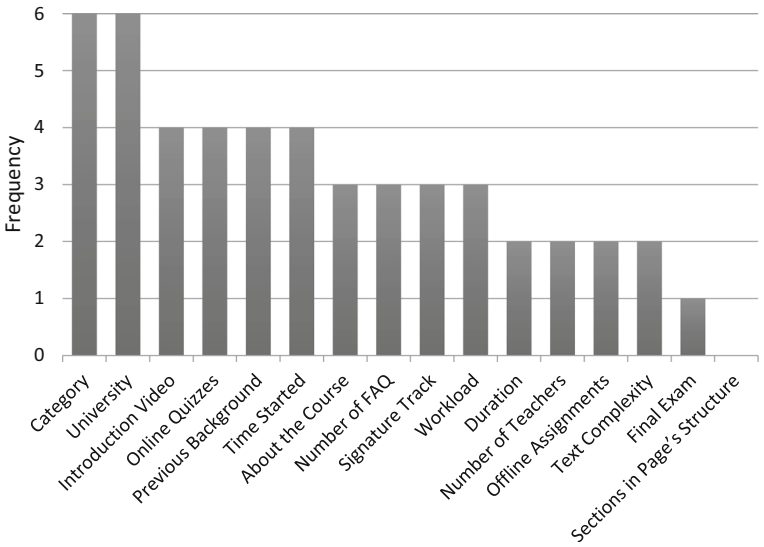

Fig. 1 Excluding classification attributes and attribute's frequency appearance

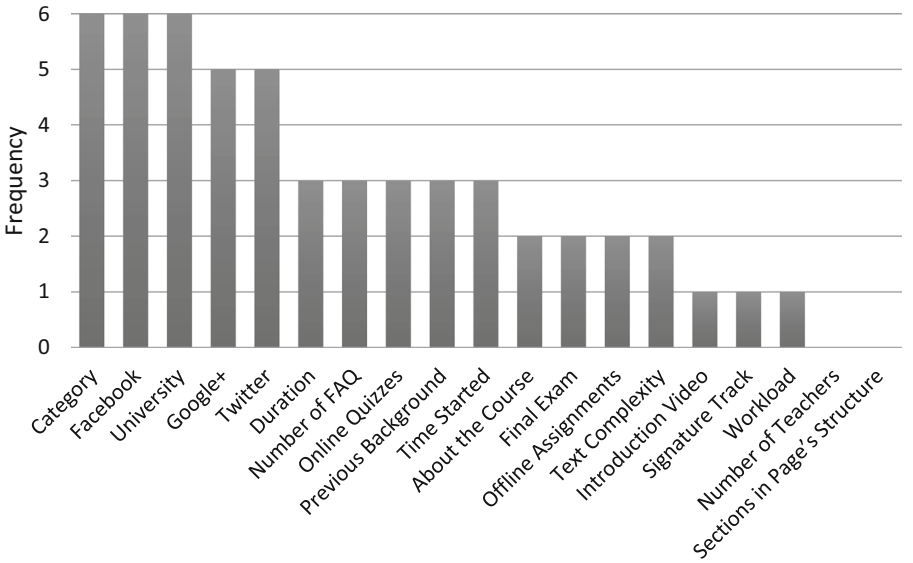

Fig. 2 Including classification attributes and attribute's frequency appearance

among them. The attributes concerning social media give information about the popularity of the course and are of high importance in relation to other categories of attributes of the information pages, such as Assessment and Certification.

Correlation and Regression

Experiments have been conducted on IBM's Superior Performance Software System (SPSS), studying the correlation and regression among the three classification attributes. We examined them in pairs: (a) Facebook and Twitter, (b) Facebook and Google+, and (c) Google+ and Twitter. The results of each correlation experiment are shown below.

Facebook and Twitter

The Pearson correlation value between Facebook and Twitter is 0.799, as shown in Table 5; this means that there is a strong correlation between the two attributes (the correlation value is close to 1). Also, it is a positive correlation, meaning that as one attribute increases in value, the second attribute also increases in value. Similarly, as one attribute decreases in value, the second attribute also decreases in value. The Sig. (two-tailed) value shows if there is a statistically significant correlation between the two attributes. Our Sig. (two-tailed) value is 0.000, the value is less than or equal to 0.05, so we can conclude that there is a statistically significant correlation between the two attributes. This means that the increases or decreases of one attribute do significantly relate to increases or decreases of the second attribute.

The results of our linear regression analysis of those two attributes are shown in Fig. 3. The middle line represents the regression line (the relationship between the

Table 5 Pearson correlation matrix

		Facebook	Google+	Twitter
Facebook	Pearson correlation	1	0.670**	0.799**
	Sig. (2-tailed)		0.000	0.000
	N	320	320	320
Google+	Pearson correlation	0.670**	1	0.812**
	Sig. (2-tailed)	0.000		0.000
	N	320	320	320
Twitter	Pearson correlation	0.799**	0.812**	1
	Sig. (2-tailed)	0.000	0.000	
	N	320	320	320

**Correlation is significant at the 0.01 level (two-tailed)

attributes Facebook and Twitter) and the other two lines represent the 95 % confidence interval for the mean of the data. The *R*-square value of our experiment is equal to 0.639, meaning that it is possible to predict with a relative accuracy of 63.9 % the value of the Twitter attribute given the value of the Facebook attribute and vice versa.

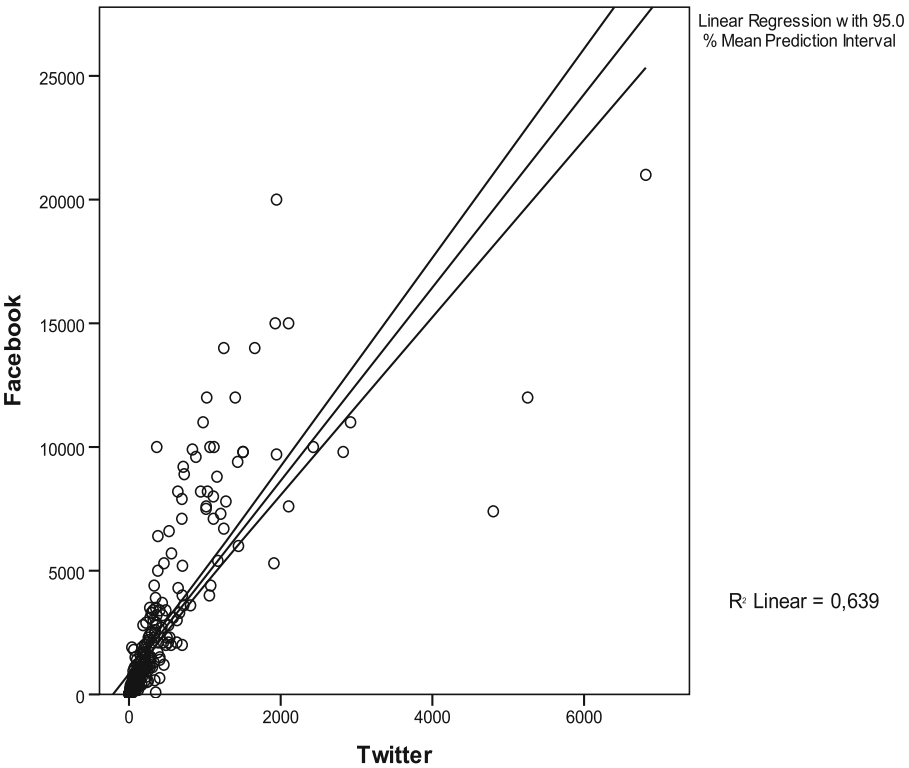

Fig. 3 Scatter plot for Facebook and Twitter

Facebook and Google+

The Pearson correlation value between Facebook and Google+ is 0.670 as shown in Table 5; this means that there is a moderate correlation between the two attributes, as the correlation value is not very close to 1. Also, it is a positive correlation, like in the previous feature pair. Our Sig. (two-tailed) value is 0.000, the value is again less than or equal to 0.05, so we can conclude that there is a statistically significant correlation between the Facebook and Google+ attributes. The results of our linear regression analysis of those two attributes are shown in Fig. 4. The middle line again represents the regression line and the other two lines, the 95 % confidence interval for the mean of the data. The *R*-square value of our experiment is equal to 0.45.

Google+ and Twitter

The Pearson correlation value between Google+ and Twitter is 0.812, as shown in Table 5, denoting a strong correlation between the two attributes. Also, it is a positive correlation, like in the previous two cases. Our Sig. (two-tailed) value is again 0.000, so we can conclude that there is a statistically significant correlation between the Google+ and Twitter attributes. The results of our linear regression analysis of those two attributes are shown in Fig. 5. The *R*-square value of our experiment is equal to 0.66.

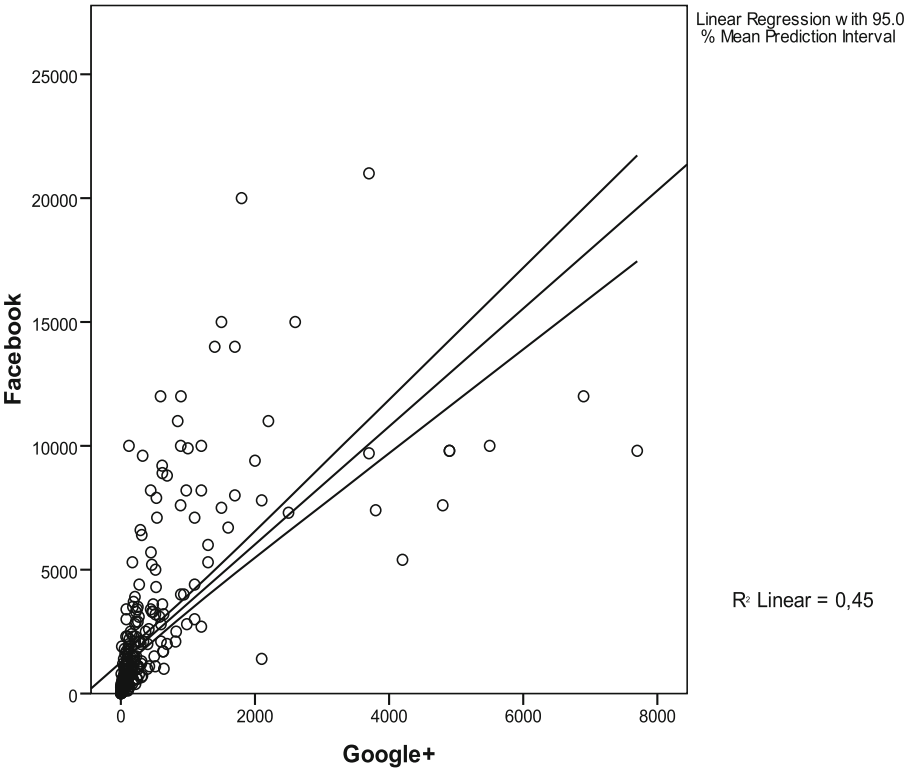

Fig. 4 Scatter plot for Facebook and Google+

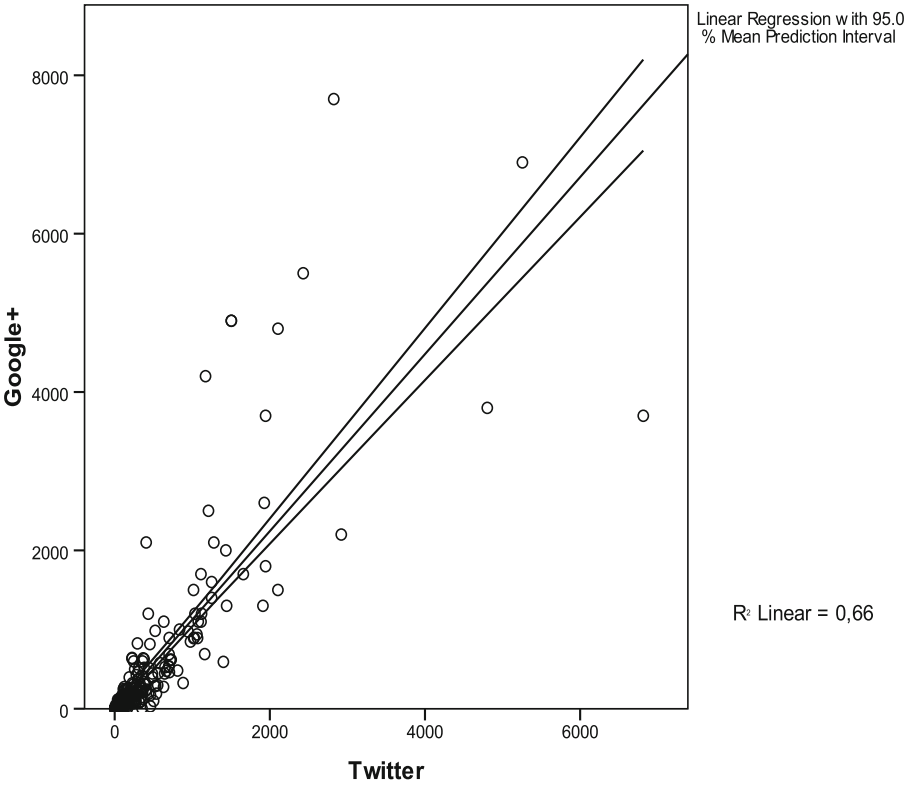

Fig. 5 Scatter plot for Google+ and Twitter

Facebook, Google+, and Twitter

Some experiments have also been conducted, studying the correlation and regression among all three classification attributes. First, we examined them using the linear regression method using Weka. The equations in Table 6 show the linear regression model for each classification attribute examined and the results of our experiments are shown in Table 7.

More specifically, we can understand from the correlation coefficient values in Table 7 that there is a strong correlation among the three social networks, examining the data from MOOCs information pages. The relation is described as strong because those data have social characteristics. They represent a human action (press of a button) that interacts in a social environment (Facebook, Google+, or Twitter) and, according to

Table 6 Linear regression model equations

Attribute	Equation
Facebook	= 3.4078 Twitter+0.4446 GooglePlus+823.5526
GooglePlus	= 0.9881 Twitter+0.0332 Facebook−19.9585
Twitter	= 0.3616 GooglePlus+0.0931 Facebook+4.2464

Table 7 Linear regression

	Facebook	Google+	Twitter
Correlation coefficient	0.7701	0.7676	0.8483
Mean absolute error	1,261.6017	218.6917	130.9558
Root mean squared error	2,203.7297	621.9705	370.7863
Relative absolute error	53.4094 %	43.2618 %	34.3911 %
Root relative squared error	64.4039 %	64.5967 %	53.0124 %

previous research (Shortell 2001), the interpretation of the coefficient depends, in part, on the area of study. When we are experimenting in areas that involve social or human data, we should expect the correlation coefficients to be lower than those in surveys studying demographic data.

Moreover, the mean absolute error and the root mean squared error values show that equations in Table 6 cannot estimate correctly the values of Facebook, Google+, and Twitter, respectively.

Secondly, we proceeded our study using RegressionByDiscretization with the Random Forest method in order to examine the correlation and the regression among the three classification attributes. The results are shown in Table 8. We can understand from the correlation coefficient values in Table 8 that there is an even stronger correlation among the three examined social networks (Shortell 2001) compared to linear regression. More analytically, the Facebook correlation increased by 5.04 %, the Google+ correlation increased by 6.85 %, and the Twitter correlation increased by 0.01 %. This proves that there is a connection between them since there is an increase in percentage of correlation in all the three social media under the same experiment conditions.

All Other Attributes

As was mentioned in the methodology, in the specific experiment, we plan to investigate the correlation and the regression of the classification attributes based on all the other attributes of the MOOCs information page. The same method was followed as in the experiment of the previous subsection. The results are shown in Table 9.

Table 8 RegressionByDiscretization with Random Forest

	Facebook	Google+	Twitter
Correlation coefficient	0.8205	0.8361	0.8484
Mean absolute error	997.8446	215.9734	207.5215
Root mean squared error	1,991.7736	557.0484	531.9431
Relative absolute error (%)	42.2433	42.7241	41.0521
Root relative squared error (%)	58.2095	57.8540	55.2466

Table 9 RegressionByDiscretization with Random Forest

	Facebook	Google+	Twitter
Correlation coefficient	0.5339	0.4511	0.3682
Mean absolute error	1939.7772	420.7755	331.4012
Root mean squared error	2908.2463	857.6465	649.2145
Relative absolute error	82.1196 %	83.2383 %	87.0314 %
Root relative squared error	84.9934 %	89.0736 %	92.82 %

We can also find out that there is a correlation between each classification attribute and the other attributes of the MOOCs information page with values 0.53339, 0.4511, and 0.3682 corresponding for Facebook, Google+, and Twitter shares. However, it has been impossible to predict the number of shares because of the mean absolute error and of root mean squared error values. These values are large, because they are a result of both of the great number attributes studied and the variety of their data type.

As was mentioned previously, both the findings of the previous subsection and the finding of this subsection derive from the study of a human action which means that we should expect the correlation coefficients to be lower than those in surveys studying demographic data (Shortell 2001).

Statistics

The statistical analysis is of special importance in our research. We can extract useful information about the attributes that trigger the interest of the users in the social media. The attributes that we study in this section are Categories and University since in the previous experiments, we proved that they are of vital importance for the shares of the MOOC information pages in the social media.

Categories

From Fig. 6, which shows the names of the course categories and the corresponding percentage of recommendation clicks (shares) in social networks, we can see that the categories of MOOCs that are most recommended in all social networks are: *Computer Science*, *Information Tech and Design*, *Humanities*, and *Economics and Finance*. In Facebook, the category with the most shares is *Humanities* with 1.08 % more shares than the second in order category, namely *Computer Science*. In Google+, the most famous category is *Computer Science* with 21.76 % more clicks than the second in order category, namely *Information Tech and Design*. In Twitter, the category with the most recommendation clicks is also *Computer Science* with 1.83 % more clicks than the second in order category, i.e., *Information Tech and Design*. The number of the clicks actually is the number of the users that are interested in the specific course category. In this way, we become aware of the interests of a great number of MOOCs' users. This information is of great value for the MOOCs creators since they can estimate the number of possible users in a new course.

Moreover, we noticed that the course categories do not offer the same number of courses. For that reason, we proceeded with the normalization of our data. In Fig. 7, we present the average percentage of recommendation clicks per category for all the social networks in question. These percentage values represent the users' choices for the MOOC categories in the social media. Analytically, Facebook users prefer the most to recommend courses of the *Economics and Finance* category, Google+ users prefer to share the *Statistics and Data Analysis* category MOOCs, and the Twitter users prefer to tweet courses of the *Education* category. These results are very important since it is possible to specify the users' choices per social network and that way, to obtain interesting information about the users' social profile.

Universities

The following graphs show the names of the universities and the corresponding percentage of shares in the social networks. More specifically, Fig. 8 shows that the top seven universities with the most recommendation clicks represent 63 % of all the shares that have been made on Facebook. Figure 9 shows that Stanford University represents 33 % of all shares in Google+, and also that the top eight universities with the most recommendation clicks represent 75 % of all the shares. In Fig. 10, we can observe that the top eight Universities with the most recommendation clicks represents

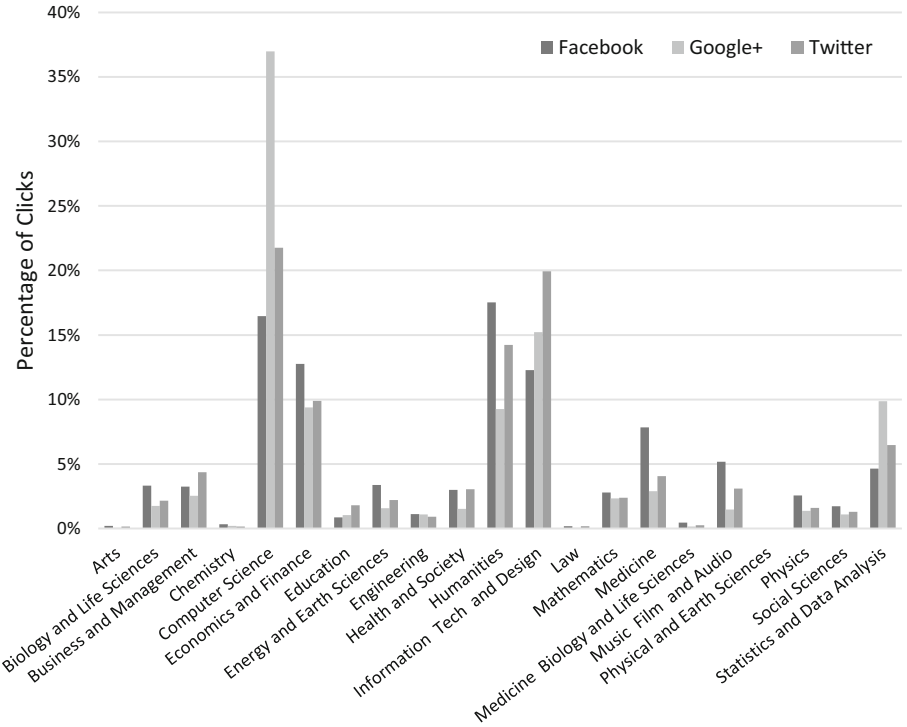

Fig. 6 Percentage of clicks per category

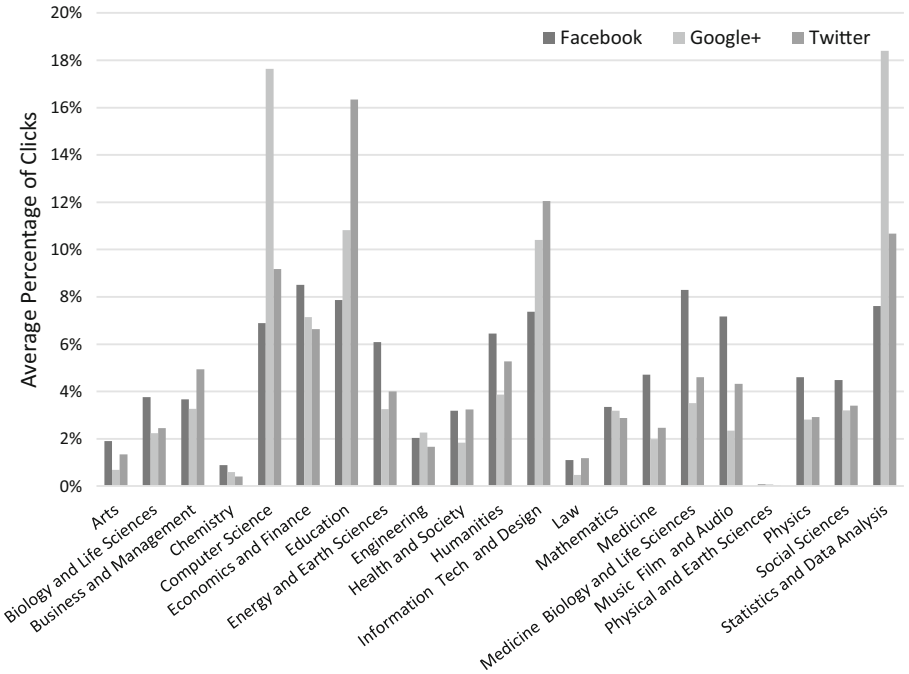

Fig. 7 Average percentage of clicks per category

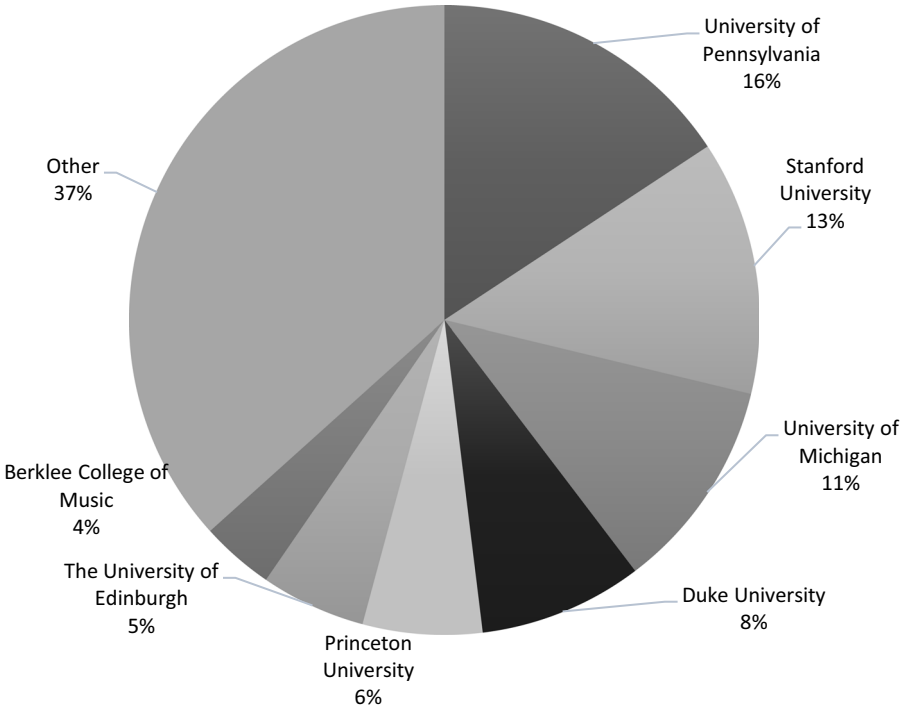

Fig. 8 Percentage of shares in Facebook per university

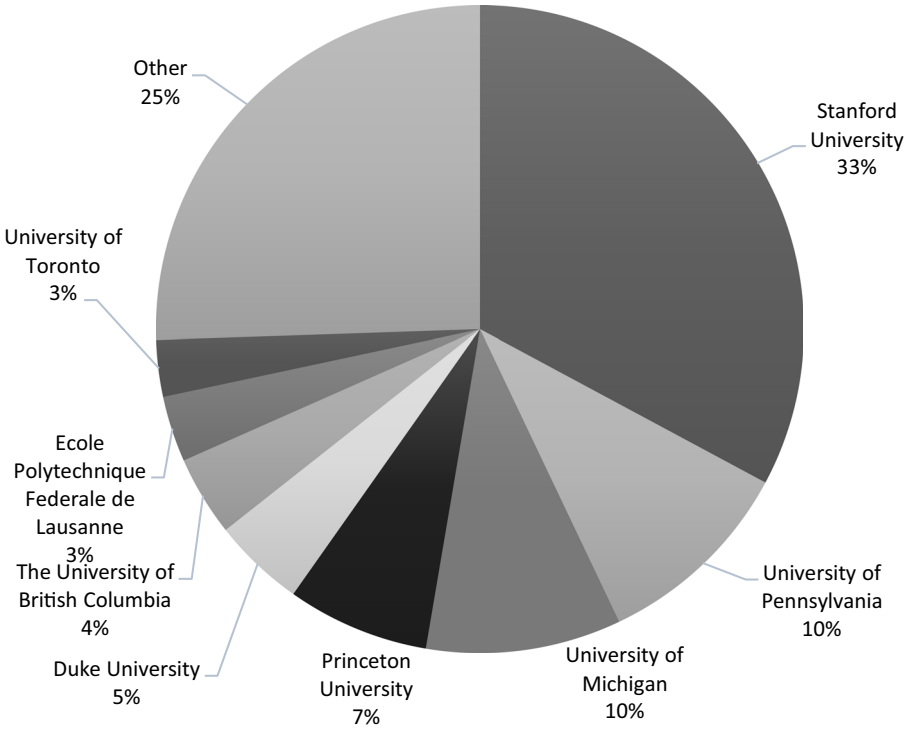

Fig. 9 Percentage of shares in Google+ per university

67 % of all the shares that have been made in Twitter. Also, Fig. 11 presents the five universities with the greater number of the courses offered.

Combining these results with the information presented in Fig. 11, we come up with some interesting results. The University of Pennsylvania offers more courses than any other university, but is not the most popular university for Google+ and Twitter social networks. Furthermore, the third in order Georgia Institute of Technology does not appear in the most popular MOOC creators as shown in Figs. 8, 9, and 10. We can conclude that the popularity of the universities does not relate exclusively with the number of their offering courses and that the users' recommendation procedure is more qualitative than quantitative.

Categories Enrollment

Recently Coursera has presented some statistic results of its own (Coursera 2013a) in October 2013. One of these concerned the enrolled users per course category. More specifically, it presented the five most popular categories along with the corresponding number of registered users. We, on the other hand, have the shares in the social media of the course categories at our disposal with data until May 2013. Thus, we can assume that the shares in the social media affect the enrollment of users in the courses in a positive way and conduce to the fact that users choose to enroll to specific courses.

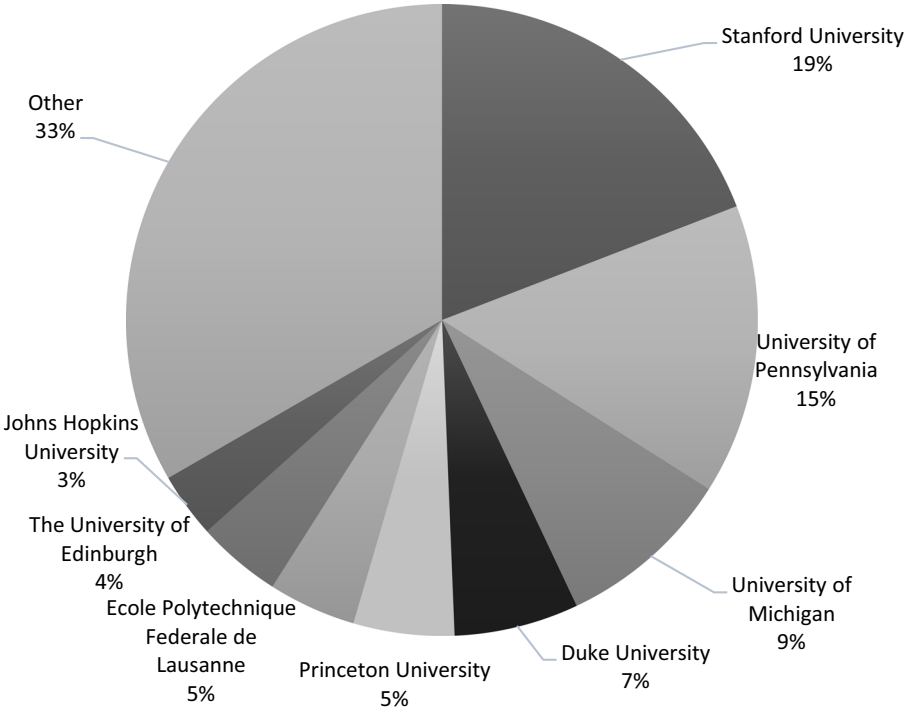

Fig. 10 Percentage of shares in Twitter per university

Figure 12 shows the course categories and the enrolled users in millions, as well as the average percentage of shares $P(c)$ in the social media calculated with the following formula:

$$P(c) = (PF\ (c) + PG(c) + PT(c))\Big/3, \quad \mathrm{c} \ = \ \text{category}$$

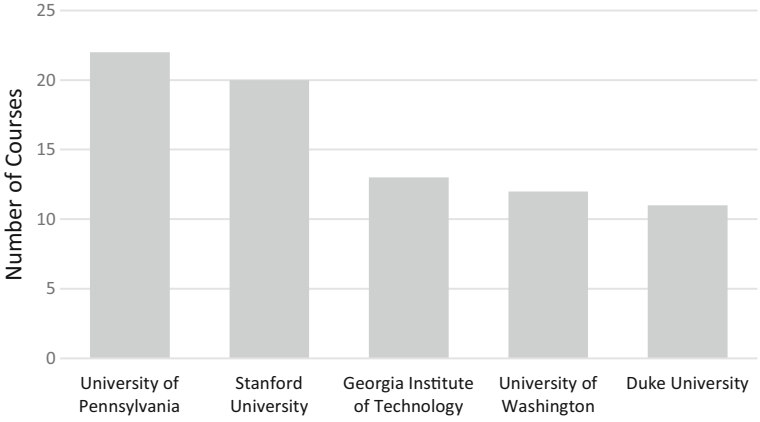

Fig. 11 Number of courses offered per university

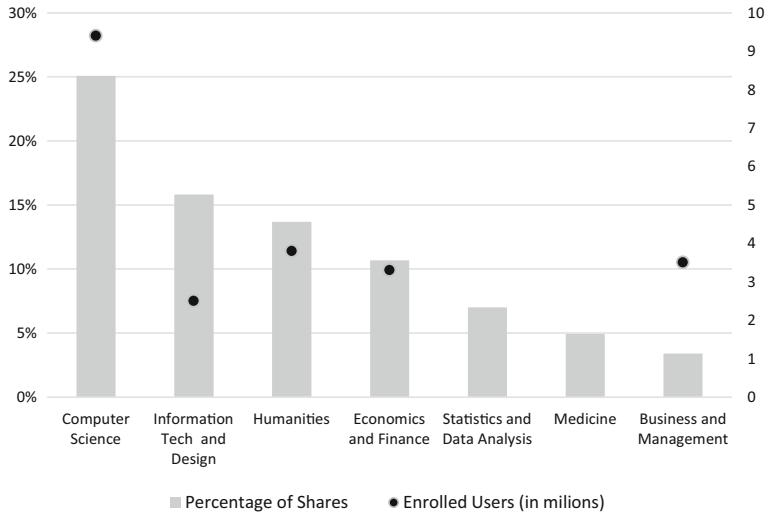

Fig. 12 Enrolled users and percentage of shares per category

where *PF*, *PG*, and *PT* are the corresponding share percentages of the category in Facebook, Google+ and Twitter.

The five categories to which most users enrolled according to Coursera are among the seven most popular categories according to our own data. Indeed, the category with the most registered users is Computer Science which is the most widespread category in social media. Unfortunately, Coursera does not provide more information about the other course categories, which would prove useful to extract further interesting conclusions. However, the general picture shows the positive effect the shares in social media have on the choice of the users to enroll to specific courses.

Discussion

MOOCs seem to be the hottest issue in educational industry. The financial figures and the number of people that come along with them are enormous. Coursera is still the largest MOOC provider with 50 % of the courses offered worldwide (Shah 2013). Thus, the present study is considered to be important since there is a detailed study of the information of MOOCs and their relation to the social media.

The present research is not just a case study but a complete research on the free data provided by Coursera, through the courses' information pages. The evaluation of the data and the knowledge extraction were made using human action, i.e., the sharing of the MOOCs information pages on social networks. The first part of the research concerned the identification of the attributes that cause some courses to be "more" recommended than others. It was found that the users prefer to share the courses that give specific information about the course itself. What is interesting, some of those course attributes concern the assignments and the course exams, and in this way, the prospective registered users know the way they will be assessed beforehand.

The second part of our experiments shows that we can define the features that are of the greatest value for the Information Pages of the MOOCs. Furthermore, it shows that

among the most significant attributes were all classification attributes which means that there is a correlation among them.

In the third part of our research, we studied the correlation between (a) Facebook and Google+, (b) Facebook and Twitter, and (c) Twitter and Google+ recommendation clicks using the Pearson correlation method. The results showed that there are strong correlations between these pairs and a linear regression analysis was carried out. The findings have shown that in the best case, it is possible to predict the value of one of the attributes of social networks given the value of the other attribute of another social network with relative accuracy of 66 %. Also, we studied the correlation and the regression between all three social media under study using Linear Regression and RegressionByDiscretization with the Random Forest method in Weka. The results showed that there are very strong correlations between those attributes representing the recommendation clicks for the three social networks. Also, we concluded that it is not possible to produce a model to estimate with precision the user's choice of recommendation clicks for the Facebook, Google+, and Twitter social networks. Finally, some experiments were conducted to find out the correlation and the regression of the classification attributes based on the other attributes of the MOOCs information pages, which showed that there is a strong correlation but is impossible to create a prediction model.

In the last part of the study, some very important statistics were presented that show the most popular universities and the most popular course categories. It was also presented that social media have a positive effect on the users that intend to enroll to the courses.

These research findings are of special importance in the MOOCs market. Being aware of the attributes that can motivate a user to register to a course or to share it through their social network, the MOOC providers will be able to increase their registered uses and, consequently, to increase their income. From the shares on the Internet, we can extract the trends concerning the choice of the courses so that the MOOC creators can save time and money aiming at the most popular course categories. Finally, for the companies that want to train their personnel, by knowing the details of the courses, they can decide more easily about the course and the model of attendance that suits better for their employees.

Furthermore, the findings of our research make a great contribution to the field of open data and social networks. More specifically, it proves that open data have to be well defined and to-the-point depending on the field of study. Open data that derive from human actions such as the sharing in the social media can show out the general tendency of the users towards a specific subject, such as the promotion of a certain course category in the social media. As for the social media, we find out that every users' community gives a different value to the subject under examination. However, all of them show more or less the same tendency. Thus, we can suggest the methodology that we followed as a complete study framework of research that relates to information sharing in social media from Web pages.

Conclusion

In conclusion, it can be said that the present study can prove very useful for a number of various reasons. Firstly, the universities can increase the enrollment in their courses,

while companies that already exist in the field or want to get involved in it, can increase their popularity. Furthermore, those who offer parallel tutoring to the MOOCs students can target the proper course categories so that they can increase the number of customers. In addition, advertisers and social media experts can target their viewers/users with greater precision. Also, the field experts of education and social media can have a more comprehensive approach to the MOOCs issue. Finally, the findings can be exploited by the three social networks that have been examined in order to personalize the profiles of the users that may be interested in attending a MOOC.

The examination of open data in new research fields such as MOOCs are of special scientific interest. The new methodologies followed can be confirmed in other studies and a new study framework can be defined in relation to the new field. The successful definition of the attributes that make MOOCs more shared in social media sets the scientific foundation of the methodology we followed. The social media continue to constitute a hot research field and what is of great interest is the result of our research, i.e., that the users of all the three social networks that were examined showed similar tendencies concerning information sharing in relation to MOOCS information pages. Moreover, the examination of the correlation between the social media concerning the same issue constitutes an important addition to the study framework of future research. The fact that there is some, though very little, difference in our research findings in the social media examined leads us to the conclusion that there should be a separate examination of each social network so as to extract more comprehensive results in relation to our research subject.

The present research can extend to other MOOCs providers confirming our findings. Moreover, it can become extensive by adding more attributes from the content of the MOOCs lectures and the results of their attendance. The research could also extend on the part of the social media, by examining the choice of the users to share a MOOC in their social network and more specifically by studying their friends' response to the sharing.

References

Aguillo, I. (2010). Web, webometrics and the ranking of universities. In Proceedings of the 3rd European Network of Indicators Designers Conference on STI Indicators for Policymaking and strategic decision, CNAM, Paris (to appear, 2010).

American Council on Education. (2014). College credit recommendation service. http://www2.acenet.edu/credit/?fuseaction=browse.getOrganizationDetail&FICE=1007444. Accessed 21 Jan 2014.

Anderson, N. (2012). Elite education for the masses. Washington Post. http://www.washington post.com/local/education/elite-education-for-the-masses/2012/11/03/c2ac8144-121b-11e2-ba83-a7a396e6b2a7_story.html. Accessed 20 Jan 2014.

Bersin, J. (2013). The MOOC marketplace takes off. http://www.forbes.com/sites/joshbersin/ 2013/11/30/the-mooc-marketplace-takes-off/. Accessed 20 Jan 2014.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning (ICML 2006).

Cecil-Reed. P. (2013). The MOOC takeover Pt 1: founders and innovators. http://www.onlinecolleges.com/educational-trends/ocw/moocs-founders-and-innovators.html. Accessed 20 Jan 2014.

Chatti, M. A., Dyckhoff, A. L., Schroeder, U., & Thüs, H. (2012). A reference model for learning analytics. *International Journal of Technology Enhanced Learning, 4*(5), 318–331.

Collins, K. (2013). Facebook could become a distribution vehicle for MOOCs, says global policy chief. http://www.wired.co.uk/news/archive/2013-11/01/facebook-moocs. Accessed 20 Jan 2014.

Coursera. (2013a). A triple milestone: 107 partners, 532 courses, 5.2 million students. http://blog.coursera.org/post/64907189712/a-triple-milestone-107-partners-532-courses-5-2. Accessed 20 Jan 2014.
Coursera. (2013b). Introducing signature track. http://blog.coursera.org/post/40080531667/signaturetrack. Accessed 20 Jan 2014.
Coursera. (2014). Company's website. https://www.coursera.org. Accessed 20 Jan 2014.
de Waard, I., Koutropoulos, A., Keskin, N., Abajian, S. C., Hogue, R., Rodriguez, C. O., et al. (2011). *Exploring the MOOC format as a pedagogical approach for mLearning*. Beijing: Proceedings from mLearn 2011.
DeBoer, J., Stump, G. S., Seaton, D., & Breslow, L. (2013). Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x. In Proceedings of the Sixth Learning International Networks Consortium Conference.
Delfino, M., & Persico, D. (2007). Online or face–to–face? Experimenting with different techniques in teacher training. *Journal of Computer Assisted Learning, 23*(5), 351–365.
Dellarocas, C., & Van Alstyne, M. (2013). Money models for MOOCs. *Communications of the ACM, 56*(8), 25–28. doi:10.1145/2492007.2492017.
Domingos, P. (2005). Mining social networks for viral marketing. *IEEE Intelligent Systems, 20*(1), 80–82.
Färber, I., Günnemann, S., Kriegel, H. P., Kröger, P., Müller, E., Schubert, E., et al. (2010). On using class-labels in evaluation of clusterings. In MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD.
Frankola, K. (2001). Why online learners drop out. *WORKFORCE-COSTA MESA, 80*(10), 52–61.
Grünewald, F., Meinel, C., Totschnig, M., & Willems, C. (2013). Designing MOOCs for the support of multiple learning styles. In Scaling up Learning for Sustained Impact (pp. 371-382). Springer Berlin Heidelberg.
Gundecha, P., & Liu, H. (2012). Mining social media: a brief introduction. Tutorials in Operations Research, 1(4).
Hall, M. A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *Knowledge and Data Engineering, IEEE Transactions on, 15*(6), 1437–1447.
Han, J., Kamber, M., & Pei, J. (2006). Data mining: concepts and techniques. Morgan Kaufmann.
Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support vector machines, intelligent systems and their applications. *IEEE, 13*(4), 18–28.
Huang, J., Dasgupta, A., Ghosh, A., Manning, J., & Sanders, M. (2014). Superposter behavior in MOOC forums.
Indra Devi, M., Rajaram, R., & Selvakuberan, K. (2008). Generating best features for web page classification. *Webology, 5*(1), 52.
Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R., & Wu, A. Y. (2000). The analysis of a simple k-means clustering algorithm. In Proceedings of the sixteenth annual symposium on computational geometry, (pp. 100-109).
Kirkby, R., Frank, E., & Reutemann, P. (2006). WEKA explore user guide. Retrieved from http://www.cse.yorku.ca/course_archive/2006-07/W/4412/doc/weka/ExplorerGuide-3.5.5.pdf. Accessed 20 Jan 2014.
Koutropoulos, A., Gallagher, M., Abajian, S., de Waard, I., Hogue, R., Keskin, N., & Rodriguez, C. (2012). Emotive vocabulary in MOOCs: context and participant retention. The European Journal of Open, Distance and E-Learning, vol. 1.
Lennon, C., & Burdick, H. (2004). The Lexile framework as an approach for reading measurement and success. http://www.lexile.com/research/1/. Accessed 20 Jan 2014.
Liaw, A., &Wiener, M. (2002). Classification and regression by randomForest. R News, 2/3:18–22. ISNN 1609-3631.
Madria, S. K., Bhowmick, S. S., Ng, W. K., & Lim, E. P. (1999). Research issues in web data mining. In data warehousing and knowledge discovery (pp. 303-312). Springer Berlin Heidelberg.
Mak, S., Williams, R., & Mackness, J. (2010). Blogs and forums as communication and learning tools in a MOOC. In: Proceedings of the 7th International Conference on Networked Learning. pp. 275-285. ISBN 9781862202252.
McAndrew, P. (2013). Learning from open design: running a learning design MOOC. eLearning Papers, (33).
Montgomery, G. C., Peck, E., &Vining, G. G. (2010). Introduction to linear regression analysis. Published by John Wiley and Sons, Inc. ISBN: 978-0-470-54281-1.
MOOCs Mentor. (2013). MOOCs mentor releases dedicated MOOCs helpline, creates revolution. http://www.prlog.org/12261228-moocs-mentor-releases-dedicated-moocs-helpline-creates-revolution.html. Accessed 20 Jan 2014.
Murray, M., Pérez, J., Geist, D., & Hedrick, A. (2012). Student interaction with online course content: build it and they might come. *Journal of Information Technology Education: Research, 11*(1), 125–140.

Norusis, M. (2008). *SPSS 16.0 Guide to Data Analysis*. NJ: Published by Prentice Hall Press Upper Saddle River. ISBN: 0136061362 9780136061366.

Pappano, L. (2012). The year of the MOOC. New York Times. http://www.nytimes.com/2012/11/04/education/edlife/massive-open-online-courses-are-multiplying-at-a-rapid-pace.html?pagewanted=all&_r=0. Accessed 20 Jan 2014.

Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). Tuned models of peer assessment in MOOCs. arXiv preprint arXiv:1307.2579.

Pitt, E., & Nayak, R. (2007). The use of various data mining and feature selection methods in the analysis of a population survey dataset. *AIDM '07 Proceedings of the 2nd international workshop on Integrating artificial intelligence and data mining, 84*, 83–93.

Rafter, M.V. (2013). MOOCs could lower cost of college. http://money.msn.com/personal-finance/moocs-could-lower-cost-of-college. Accessed 20 Jan 2014.

Rayyan, S., Seaton, D. T., Belcher, J., Pritchard, D. E., & Chuang, I. (2013). Participation and performance In 8.02 x Electricity And Magnetism: The First Physics MOOC From MITx. arXiv preprint arXiv: 1310.3173.

Robnik-Šikonja, M. (2004). Improving random forests. In Machine Learning: ECML 2004, pp. 359-370. Springer Berlin Heidelberg.

Rodriguez, O. (2012). MOOCs and the AI-Stanford like courses: two successful and distinct course formats for massive open online courses. European Journal of Open and Distance Learning. http://www.eurodl.org/index.php?p=current&article=516. Accessed 20 Jan 2014.

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. (2011). Handbook of educational data mining. *Published by CRC Press, 2011*.

Shah, D. (2013). MOOCs in 2013: breaking down the numbers. https://www.edsurge.com/n/2013-12-22-moocs-in-2013-breaking-down-the-numbers. Accessed 21 Jan 2014.

Shortell, T. (2001). Online textbook: an introduction to data analysis & presentation, Chapter 18: Correlations. http://www.shortell.org/book/chap18.html. Accessed 20 Jan 2014.

Srivastava, J., Cooley, R., Deshpande, M. & Tan, P. N., (2000). Web usage mining: discovery and applications of usage patterns from Web data, ACM SIGKDD Explorations Newsletter, v.1 n.2. doi:10.1145/846183.846188.

Thakur, M. (2007). The impact of ranking systems on higher education and its stakeholders. *Journal of Institutional Research, 13*(1), 83–96.

Veeramachaneni, K., Dernoncourt, F., Taylor, C., Pardos, Z., & O'Reilly, U. M. (2013). Moocdb: Developing data standards for MOOC data science. In AIED 2013 Workshops Proceedings Volume (p. 17).

Vora, P., & Oza, B. (2013). A survey on K-mean clustering and particle swarm optimization. *International Journal of Science and Modern Engineering (IJISME), 1*, 24–26.

Vrasidas, C., & McIsaac, M. S. (1999). Factors influencing interaction in an online course. *American Journal of Distance Education, 13*(3), 22–36.

Waldrop, M. M. (2013). Massive open online courses are transforming higher education—and providing fodder for scientific research. *Nature, 495*, 160–163.

Weka. (2013). Weka 3: Data Mining Software in Java. University of Waikato. http://www.cs.waikato.ac.nz/ml/weka/. Accessed 20 Jan 2014.