Statistical Profiles of Users' Interactions with Videos in Large Repositories: Mining of Khan Academy Repository

Sahar Yassine[1], Seifedine Kadry[2*] and Miguel Angel Sicilia[1]
[1] University of Alcalá, Alcalá de Henares, Madrid, Spain
[E-mail: sahar.yassine@edu.uah.es; msicilia@uah.es]
[2] Department of Mathematics and Computer Science, Faculty of Science,
Beirut Arab University, Lebanon
[e-mail: s.kadry@bau.edu.lb]
*Corresponding author: Seifedine Kadry

Received October 22, 2019; revised February 12, 2020; accepted March 8, 2020; published May 31, 2020

Abstract

The rapid growth of instructional videos repositories and their widespread use as a tool to support education have raised the need of studies to assess the quality of those educational resources and their impact on the quality of learning process that depends on them. Khan Academy (KA) repository is one of the prominent educational videos' repositories. It is famous and widely used by different types of learners, students and teachers. To better understand its characteristics and the impact of such repositories on education, we gathered a huge amount of KA data using its API and different web scraping techniques, then we analyzed them. This paper reports the first quantitative and descriptive analysis of Khan Academy repository (KA repository) of open video lessons. First, we described the structure of repository. Then, we demonstrated some analyses highlighting content-based growth and evolution. Those descriptive analyses spotted the main important findings in KA repository. Finally, we focused on users' interactions with video lessons. Those interactions consisted of questions and answers posted on videos. We developed interaction profiles for those videos based on the number of users' interactions. We conducted regression analysis and statistical tests to mine the relation between those profiles and some quality related proposed metrics. The results of analysis showed that all interaction profiles are highly affected by video length and reuse rate in different subjects. We believe that our study demonstrated in this paper provides valuable information in understanding the logic and the learning mechanism inside learning repositories, which can have major impacts on the education field in general, and particularly on the informal learning process and the instructional design process. This study can be considered as one of the first quantitative studies to shed the light on Khan Academy as an open educational resources (OER) repository. The results presented in this paper are crucial in understanding KA videos repository, its characteristics and its impact on education.

Keywords: Open Educational Resources, Quantitative Analysis, Instructional Video, Users' Interactions, Learning Repository, Online Learning.

http://doi.org/10.3837/tiis.2020.05.013 **ISSN : 1976-7277**

1. Introduction

Communication and exchanging information are major goals of the internet which align with the major concerns of education and online learning. Therefore, a massive amount of learning objects is stored on the internet, so they can be accessed by different stakeholders. Those learning objects are organized in educational repositories and learning environments and they are available to be used, reused and shared over the internet [1]. A widely used open course-ware initiative is Khan Academy which provides open educational resources (OERs) in different subjects. Khan Academy's repository can be considered as m-learning environment because it is accessible anywhere, anytime and by everyone [2]. The instructional videos are the most usable OERs in Khan Academy as they are shared on YouTube and used in different educational sets as well.

Globally, the usage of videos as instructional learning objects is rapidly increasing. Many universities are offering video lectures in their subjects. On the other hand, flipped classroom is a common pedagogical approach practiced widely in education where students can watch online lectures anywhere and benefit from the rich OER's repositories such as Khan Academy [3], then evaluate their knowledge in class. In 2012, Kay [4] reviewed the history and growth of video podcasts in education. He demonstrated their positive impacts on students' behavior and learning performance. Cargile and Harkness [5] examined the use of Khan Academy videos as a tool of instruction in teaching mathematics in schools. They concluded that flipping instruction has become a popular component of school improvement plans. Hwang [6] proposed seamless flipped learning model to facilitate learning in different contexts through watching instructional videos anywhere, collecting relevant information and evaluating in class. In 2018, Karabulut Ilgu [7] highlighted the popularity of the flipped learning approach in engineering education in the period after 2012. He discussed its trends and demonstrated its effectiveness, benefits and challenges.

The variety of video-based learning environments includes general repositories (e.g. YouTube and iTunes) and other video repositories in smart, interactive and open educational platforms (e.g. MOOCs, LORs and OCWs. Those videos are posted in two or more different platforms (e.g. YouTube and Khan Academy) according to the philosophy of "sharing, reusing, improving and sharing again" [8]. Open Educational Resources (OERs) stored in those repositories include course materials, modules, instructional videos, tests, and other techniques used to support knowledge. Those OERs are "used and reused by a community of users such as educators, students and self-learners for teaching, learning and research purposes." [9]. The rapid growth of OERs has raised the concern about their quality and how to measure their effectiveness using accurate quality indicators and measurements. Ochoa & Duval [8] highlighted that open educational resources could be "mined" and quantitative measures of 'good' and 'not good' resources could be compared in order to discover essential attributes related to quality which will help in creating statistical profiles to be used in quality prediction.

Khan Academy is a famous open courseware initiative. It provides OERs in different educational disciplines. Those resources include instructional videos, exercises and other techniques. Khan Academy started in 2008 by its founder Salman Khan [10]. The goal of this initiative was to create a set of online tools to help in educating students. It started by posting short video lessons on YouTube then it converted to a full-time job in 2009 by creating a dedicated platform that provides full different educational lessons. In this paper, we collected

Khan Academy's data using their API and different web scraping techniques as an attempt to understand its characteristics and the impact of such learning repositories on education. This study presents one of the first quantitative studies that sheds the light on Khan Academy (KA) as an open educational resources (OERs) repository. We investigated inside the repository trying to study its structure and its users' interactions to find some quantitative insights that can be utilized in making efficient decisions for improving the quality of both educational resources and learning process. We analyzed different metrics related to its instructional videos and to their users such as video length, users' geographical distribution and users' interactions with those videos. We proposed three profiles to categorize those videos based on users' interactions. Then we analyzed the relation between the proposed profiles and some different metrics. Our analysis provides crucial information in understanding the logic and the learning mechanism inside learning repositories which has major impacts on education in general and on informal learning process and instructional design process in particular.

2. Related Work

2.1 Size and evolution of educational repositories

Ochoa & Duval [8] focused on quantitative analysis of learning objects repositories rather than qualitative ones. They highlighted the importance to measure the progress of learning objects economy through providing empirical answers to questions related to typical repository's size, number of used OERs, repository's growth over time, and the average number of OER's published by contributors. Santos Hermosa [11] assessed OERs' reuse through analyzing content related indicators. The study concluded that the most used repositories in higher education are institutional ones dedicated for educational purposes.

2.2 Previous studies on profiles of educational repositories

Cechinel [12] classified OERs into three statistical profiles (good, average and poor) after considering their quantitative measures, peer reviews and users' ratings. 35 metrics were extracted from 6470 OERs existed inside MERLOT repository. One limitation of this study was considering the resources that received low ratings as poor resources and correlate them to the effort of reviewers because those ratings were given by the community of evaluators and not the community of users. He used those metrics to implement experiments [13] to automatically generate quality information about learning resources based on their essential features and evaluative metadata. The results could be used to provide internal quality information for any newly added learning resources.

2.3 Previous studies on quality of educational learning resources

Due to the increase of open educational resources, evaluating their quality is critical, challenging and involves several dimensions and different stakeholders. Researchers proposed different evaluation approaches to enhance quality measurements. Some of them counted on users to evaluate the quality by rating and commenting such as peer review methods. Others use different techniques by developing set of metrics for ranking OERs' search results inside repositories. Clements [14] created a LOR quality assurance framework to be used in building or updating repositories. Robinson [15] proposed quality guidelines to create online courses with a focus on both content and delivery system. Holland [16] identified how informal online learning can be effectively designed outside a formal online course.

2.4 Previous studies on analyzing users' interactions with educational resources

Analyzing users' interactions with OERs opens the door to a deeper understanding of how users are utilizing those resources. This will reflect on the quality of learning objects by encouraging their developers to follow a data driven approach in designing learning objects. Many studies agreed on the impact of analyzing users' interactions on improving the next generation of learning videos. Kim [17] analyzed click-level interactions and combined them with video content analysis to study peaks in viewership and student activity. In another study [18], authors clustered users' interactions with MOOCs into different video behavior patterns. While other researchers [19] analyzed a course content with the users' interaction patterns to help in improving the design of learning environments.

The above studies were built on analyzing data inside educational repositories. Research in this area is still limited. This highlights the possibility for more research efforts and opens the door to the need of more descriptive and quantitative analysis for such repositories to discover the mechanism of their learning process.

3. Research Objectives

This research is an attempt to mine an open educational resources repository to find some quantitative measures and insights that help in defining the quality of those resources and to contribute in improving the learning process. The study investigates the repository to understand its characteristics and its impact on learning process and instructional design. It provides descriptive and quantitative analysis on instructional videos and their audiences as well. The study can act as a model for other repositories to search for empirical insights that guide for the quality of learning objects and to enhance the understanding of learning process. Those insights can be utilized in making efficient decisions for improving quality of both educational resources and learning process. Our analysis would have extensive use in the education in general and particularly in improving informal learning environments.

4. Materials and Methods

This section provides an overview of data collection from Khan Academy's repository. It shows methods used to build the database and how we gathered data. It also describes KA website structure and its components.

4.1 Khan Academy's Repository and Data Collection

A database from KA repository was created. The structure of this database was built based on KA-API. Khan Academy's ER diagram **Fig. 1** shows the structure of this database and the relationships between main tables.

Using a designed scraping tool that navigates KA website we managed to go deep in the structure and collect non authenticated data related to KA OERs. We gathered users' interactions with those OERs. Those interactions are accessible for any public user without any kind of authentication.

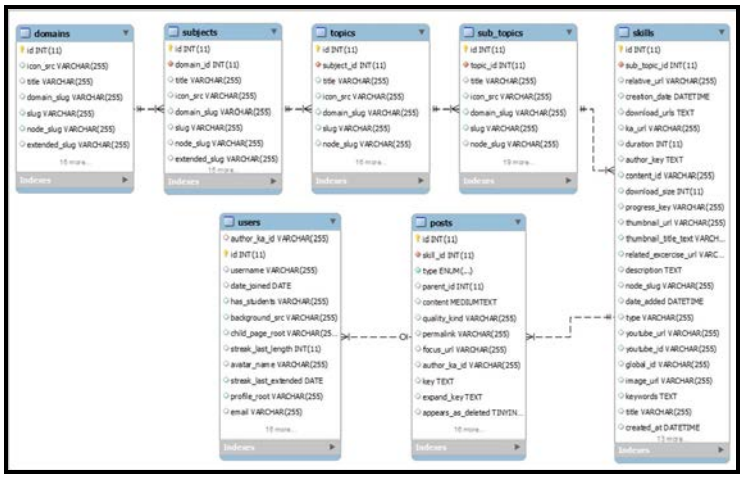

Fig. 1. KA Database ER Diagram

The designed scraping tool use a PHP script to mine inside KA's repository. This tool was designed to traverse KA website and gather all non-authenticated data. The scraping technique starts from the top level (domains) to reach the down level (skills. Those skills are the different OERs including instructional videos. We managed to scrape more than 9,000 videos related to more than 700 different topics from all of Khan Academy's domains. A lot of users interact with those videos by posting questions and answers. Those users have different profiles either private or public ones. We managed to obtain 2,797,382 users' interactions posted by a lot of different users. **Table 1** shows the number of gathered topics, skills and users' interactions from each domain.

Table 1. Number of the scraped topics, videos and interactions

Domain_Name	Number of topics	Number of Users' Interactions
Math	258	2,114,070
Science	146	350,462
Arts and humanities	90	95,895
Test prep	35	73,063
Partner content	102	57,530
Economics and finance	31	46,390
Computing	17	40,677
College- careers- and more	30	19,295

Around 11,000 of those users have public profiles with different non-authenticated details such as user ID, username, bio, country, joining date and count of videos completed. 3,278 of them have a public joining date, 4,777 have a public country information and around 5,000 showed the number of their completed videos. The collected dataset belongs to the period from February 2011 to December 2018. For interpreting the findings and analyzing the results, we used some analyzing and visualizing techniques such as MATLAB, Tableau software and R Studio.

4.2 KA Website Structure

The top level in KA structure is Domains. There are 8 different domains. Main domains are

related to study fields which are Math, Science, Computing, Humanities and Arts, Economics. New domains were added in later stages which are not related to study fields but are used for different purposes such as supporting study fields by cooperating with external partners and preparing students for college admission tests and assisting them in the college admission process. Those domains are 'Partner content', 'Test Prep' and 'College, careers, more'.

The second level is Subjects included in each domain. Most of the subjects are related to math which is the largest domain. Total subjects in all KA domains reached 127 subjects. The third level is the Topics included ineach subject. Total topics presented in KA website reached to 984 topics. In each topic there are sub-topics which can be considered the lessons. Total lessons collected from KA website is 6,376 lessons. Each lesson contains different skills which are categorized to videos and exercises. **Table 2** shows total number of contents in each KA domain. **Table 3** shows examples of subjects, topics and skills.

Table 2. KA Domains and Number of their Contents

Title	Subjects	Topics	Sub-Topics	Skills	Skills: Video	Skills: Exercise
Math	57	446	3350	18010	11354	6656
Science	10	146	619	3234	2944	290
Computing	3	17	144	102	73	29
Humanities	11	90	633	1969	1531	438
Economics-finance-domain	5	40	206	842	750	92
Partner-content	30	169	863	1889	1659	230
Test-prep	7	44	406	2636	2028	608
College-careers-more	4	32	155	498	498	0

Table 3. Examples of Subjects, Topics, Sub-topics and Skills in KA domains

Domain	Subjects (Sample)	Topics (Sample)	Sub-Topics (Sample)	Skills
Math	• Early math • 1st grade • 2nd grade… • Basic geometry • Pre-algebra • Algebra basics • Trigonometry	In Early math Subject: – Counting – Place value – Addition & subtraction In Pre-Algebra Subject: – Arithmetic properties – Fractions & Decimals	In Counting Topic: – Counting – Numbers 0 to 120 – Counting objects – Comparing small numbers	In Counting Sub-Topic: 3-Skills: Videos 2-Skills: Exercises
Science	• Physics • Chemistry • Organic chemistry • Biology • Electrical engineering • Health & medicine	In Physics Subject: – Work and energy – Fluids – Magnetic forces In Chemistry Subject: – Atoms & compounds – Chemical reaction – Periodic table	In Fluids topic: – Density and Pressure – Buoyant Force and Archimedes Principle – Fluid Dynamics	In Newton\'s law of gravitation Sub-Topic: 6-Skills: Videos 3-Skills: Exercises
Computing	• Computer programming • Computer science • Hour of Code	In Hour of Code Subject: – Drawing with code – Creating webpages – Creating SQL databases	In Algorithms Topic: – Intro to algorithms – Binary search – Asymptotic notation – Selection sort	In Make your webpages interactive Sub-topic: 1-Skills: Videos 2-Skills: Exercises
Arts and humanities	• US history • Grammar • World history • Art history • Music	In Art History Subject: – Art history basics – Prehistoric art – Medieval Europe & Byzantine	In Prehistoric art Topic: – Paleolithic art – Neolithic art – Quiz: prehistoric art	In Early colonization projects Sub-topic: 5-Skills: Videos 2-Skills: Exercises

5. Results and Discussion

5.1 Evolution of KA repository and its user base

Although Khan Academy's organization started in 2008 through Yahoo Doodle Images and YouTube, the repository started to gain popularity in 2011. To identify the popularity of KA repository we used two measurements. First one is the number of active users who joined KA repository over years. The dataset included 3,278 users with public joining date **Fig. 2**. Only 4 of them were joined in 2010 and still active. The rapid increase in joined users was during 2013, 2014 and 2015. After that numbers started to decrease until 2018 when only 7 users only joined and still active. This sharp drop might be due to the increase of alternative repositories that gain the users' interest.

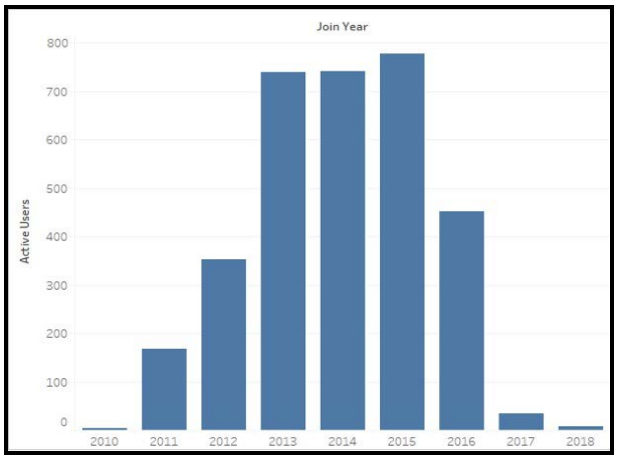

Fig. 2. Measuring Popularity by number of Active users in KA over years

The second used measurement is number of users' posts over years. As mentioned before, we collected more than 2.7M users interactions categorized in questions and answers **Fig. 3**. Users started to interact with KA repository in 2011 (with emerging math domain. Interactions started to increase gradually until they reached to the maximum in 2016 (> 470,000 posts. In 2017 interactions level declined until 2018 when it reached 187,000 posts. This drop can be due to several reasons. One main reason could be the increase in competition and having many alternatives.

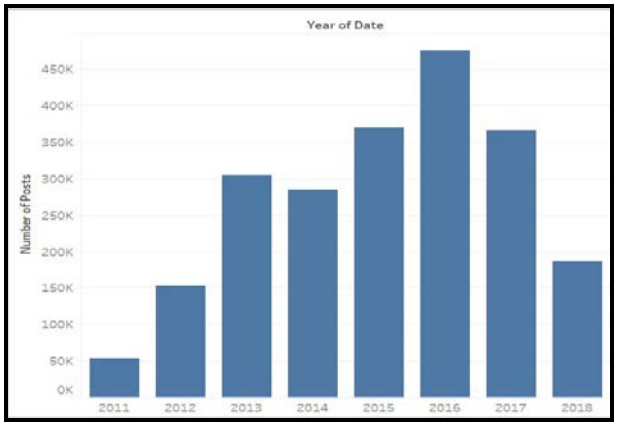

Fig. 3. Measuring Popularity by number of users' interaction over years

Growth rate of KA repository (**Table 4**) differs from one domain to another. If we look to the oldest and largest domain (Math. In 2011 was the first big addition to it. KA team added more than 2,500 videos to the repository. In 2013 they added more than 2,600 math videos. After that KA team maintained an increment of 1,000 to 1,400 videos per year until 2018 when the addition decreased to 289 videos. This growth pattern may indicate that math domain reached to its saturation level.

Table 4. KA Repository Growth over Years

	Year Added							
Domain	2011	2012	2013	2014	2015	2016	2017	2018
College-careers-more	31		26	146	5	17	193	80
Computing		28	10	21	13	1		
Economics-finance-domain	373	218	33	9		9	6	102
Humanities	54	99	164	148	107	277	348	334
Math	2,666	699	2,689	1,083	1,419	1,417	1,092	289
Partner-content	5	15	238	678	365	216	43	99
Science	636	195	440	701	527	368	32	45
Test-prep	125	151	292	772	559	21	80	28

Science domain follows a similar growth pattern. It also started in 2011 by publishing 636 science videos. Another peek point was in 2014 when they added around 700 videos. After that, the addition declined until 2018 when 45 videos only were added. Economics and finance domain started in 2011 with 373 videos. Then it declined until KA stopped publishing in 2015. After that they tried to revive this domain again, so they published 102 videos in 2018. This shows that this domain is less popular than others. Humanities is the most growing domain. It started by publishing 54 videos in 2011. KA team kept increasing the publishing gradually until 2018 when they added 334 videos. Computing is the least focusing domain. In 2012, KA team started publishing computing videos with very humble numbers until they stopped adding in 2017. Due to this very low addition pattern, this domain in KA platform cannot be considered a competitor in its field.

In 2014, KA repository formulated 3 new domains: 'test prep', 'college, careers, more', and 'partner content'. Although KA team started to build those domains earlier than 2014 but official announcements started in 2014. Regarding 'Partner content' domain, NASA announced in May 2014 that a new collaboration was made between NASA and Khan Academy to bring STEM opportunities to online learners through publishing dynamic educational materials in KA repository [20]. In 2015 and in regards to 'Test prep' domain, Khan Academy and College Board Organization announced the creation of free SAT study tools which were published in KA repository [21]. While 'College, careers, more' domain were announced by Khan Academy in 2014. They created college admissions resource for high school students and college counselors. It guides students through their college application processes and helps them in navigating different college options.

The repository's size, as defined by Ochoa is number of objects presented in repository's courses. According to that, the size of KA repository can be identified based on **Table 2**. In KA repository, a total of 29,180 learning objects are related to 984 different topics. That means average course size distribution is 29 objects per course. Comparing this average to other repositories included in Ochoa's study [8] **Table 5**, we can find that the course size of KA repository is bigger than all of others.

Table 5. Course Size Distribution (Ochoa's Study)

Repository	Courses	Objects	Average	Shape	Scale
MIT OCW	1,796	42,527	24	1.03	26.5
OpenLearn OCW	405	10,644	27	0.71	18.1
Connexions LORP	268	5,242	20	0.78	15.4
SIDWeb LMS	1,445	23,370	16	0.65	9.46

5.2 Descriptive analyses of KA Repository

We applied some descriptive analyses to describe KA repository, its characteristics and its users. Those analyses are divided into two categories: General Descriptive analysis which is related to KA repository performance in general and interactions related analysis which is analyzing the users' interactions with KA's videos.

5.2.1 General Descriptive analysis

5.2.1.1 Average Duration of KA Videos:

An important video's characteristic presented in this repository is video length. It is varied from domain to another. **Fig. 4** arranges the videos in each domain according to their durations. Some outliers are found in Arts and Humanities, College and Career and Test Prep domains. In **Table 6**, we calculated the average video duration (in seconds), its maximum and minimum in each domain separately.

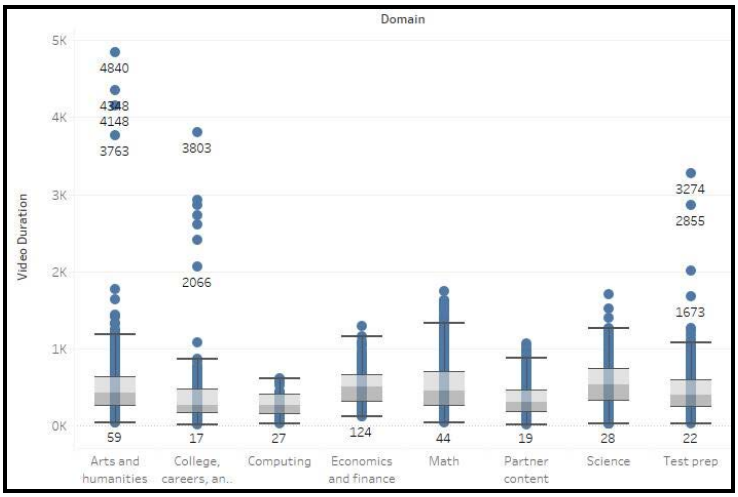

Fig. 4. Videos' Duration in Each Domain

The longest video in the whole repository is (Art Making Programs for Individuals with Dementia) in partner content domain (partner: The Museum of Modern Art. It lasts for 1 hour, 20 minutes, 40 seconds (4,840 seconds. While the shortest video in the whole repository is called (Student story: Work study as a study hall) in college careers more domain and it lasts for 17 seconds only. The typical average video length in the repository is around 412 seconds (7 minutes).

Table 6. Average Video Duration in Each Domain

Domain	Avg. Duration	Max. Duration	Min. Duration
College-careers-more	361	3,803	17
Computing	290	616	27
Economics-finance-domain	484	1,287	123
Humanities	474	4,348	36
Math	334	1,739	44
Partner-content	318	4,840	19
Science	551	1,706	27
Test-prep	486	3,274	22

5.2.1.2 Geographical Distribution of KA Users:

4,777 users of the scraped ones mentioned their location in their profiles publicly **Fig. 5**. the United States, we have 82% of them (3,935 users. This indicates that KA is widely used and very popular in US. Most of those users are math users. The second country where KA repository is widely used is Canada (290 users. The third one is India with 197 users. Khan Academy is also popular in UK with 3.6%. Australia and New Zeeland also are good potential regions for KA repository with 4.3% of users.

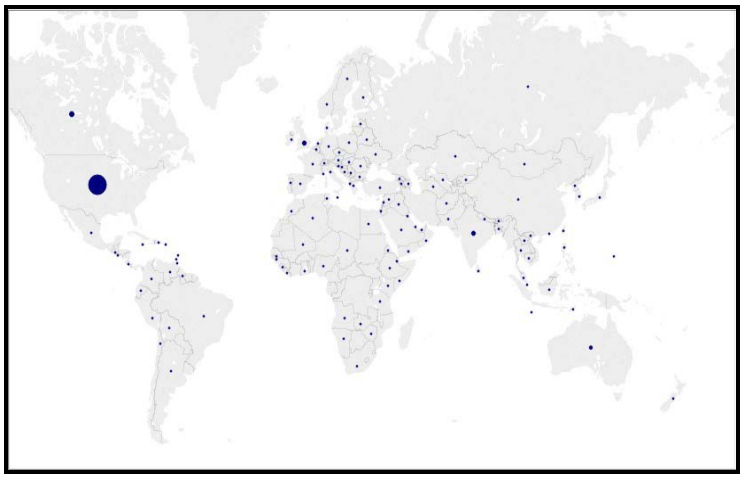

Fig. 5. Geographical Distribution of KA Users

5.2.1.3 Number of Videos Completed by Users:

Another important indicator for the use of KA videos is number of videos completed by each user **Fig. 6** which means how many videos were watched completely (till the end) by the user. We collected this information for more than 10,000 users. Almost 7,000 of them never completed any video (count of video completed is 0. 120 users completed up to 3 videos. 652 users completed (4 to 27 videos) with average of 27 users. Around 350 users completed (28 to 50 videos. 690 users completed (51 to 120 videos. Around 600 users completed (120 to 300 videos) with average of 3 users. For the higher counters, number of users tends to reach one user for 500 completed videos and above. This means that most of users are either got bored quickly or maybe not interested in watching the whole video. They maybe watch only the part that they are interested in.

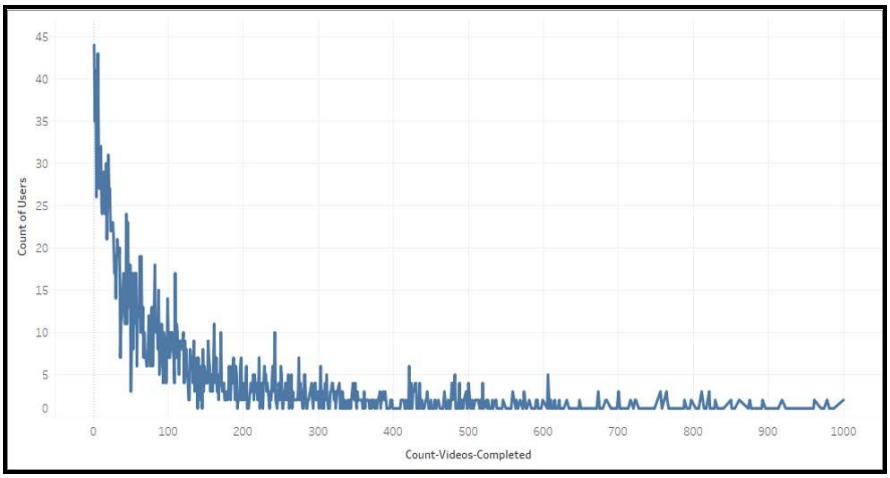

Fig. 6. Number of Videos Completed by Users

5.2.2 Interactions related analysis

5.2.2.1 Number of Users' Interactions per Video:

More than 2.7M users' interactions were gathered. Those interactions are in questions and answers posted on videos from different domains after watching them. 2.1M were posted on 3,134 math videos while only 571,000 interactions were posted on 5,749 videos belong to other domains. This indicates that the most watchable and popular videos are related to math. Most of users subscribe for KA repository to get help in math, watch it videos and interact with them. Those interactions help in figuring out to what extent the video is interested and how much users' attention it gained. The most interested video that gained the maximum number of interactions is (Radius, diameter, circumference, Pi) with more than 20K posts. The first non-Math video that gained users' attention is (What is Programming?) with around 9K posts and it is ranked the 21st video based on number of interactions **Fig. 7**.

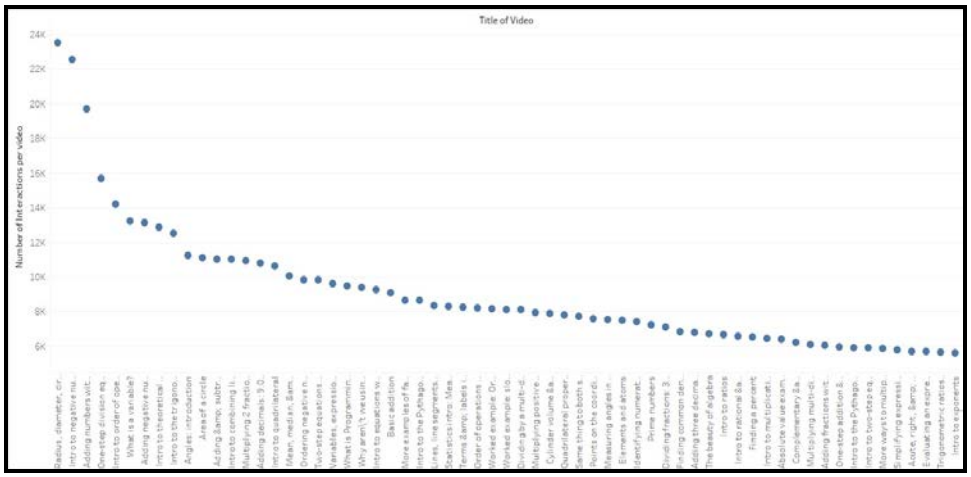

Fig. 7. Number of Users Interactions per Video

5.2.2.2 Evolution and distribution of interactions around KA contents:

Users interaction is a very crucial indicator for the importance of video content and shows how much the video attracts users' attention. **Fig. 8** shows number of users' interactions (questions and answers) that gathered by scraping each domain. More than 2.1M interactions related to math domain while 571,709 interactions related to other domains. The second highest domain

is science which attracted many users to interact with a total of 350K posts. The third domain is Arts and Humanities with more than 90K interactions. The fourth is Economics and Finance with 46K interactions.

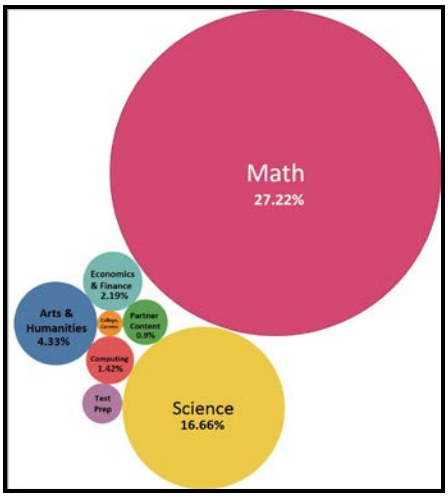

Fig. 8. Number of Users Interactions per Domain

Then computing domain with approximately 30K interactions. The rest of domains have the smallest shares of interactions with less than 27K each. This indicates that the popularity for the subject related domains is still much higher than the service-related ones. This distribution can be due to variations in targeted users' segments or audience. For example, the targeted users' segments for math includes all school students from all levels, all math teachers and college students. While the targeted users' segments for Arts and humanities may include only high school students and Arts college students. In **Fig. 9**, we explored users interactions taking in consideration the different subjects across domains.

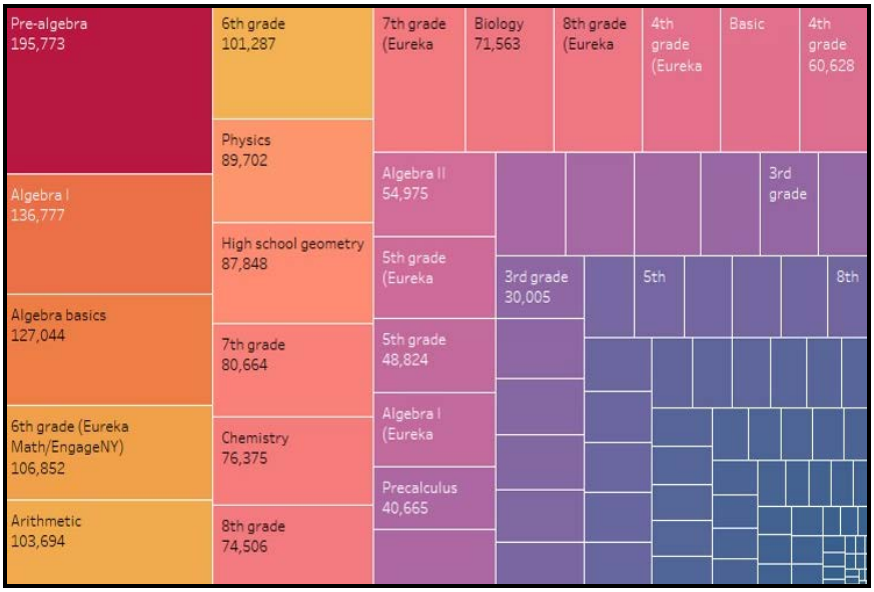

Fig. 9. Number of Users Interactions per Subject

The most popular subject is Pre-Algebra. It gained the highest number of interactions (195,773. The second one is Algebra I (136,777 interactions) while the third one is Algebra basics (127,044. The first non-Math subject with the highest interactions is Physics which is ranked the 7th with (89,702 interactions. Another distribution in **Fig. 10** shows the fluctuation of number of users' interactions over years for each domain. This confirms again that math domain and its subjects, specifically Algebra related topics are the ones that attract most of the users' attention.

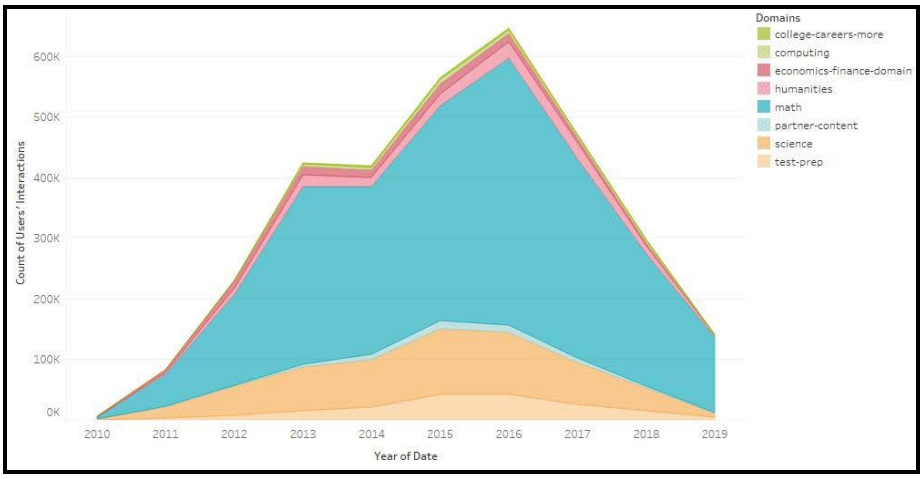

Fig. 10. Number of Users' Interactions per Domain over Years

5.2.2.3 Interaction profiles of videos:

User data for different 30 metrics were extracted from the 9076 videos in the dataset. We categorized those metrics to three classes based on the related collected data. Those classes are: Video content related measures, interaction related measure and user related measures **Table 7.**

Table 7. Metrics Collected for the study

Class of Measure	Metrics
Video-content-related measures	Publishing date, Video Duration, Author's Key, Download Size, Download Link, Keywords, YouTube ID, Subtopic's Related, Subject's Related, Domain's Related and Number of Reusing the video in different subjects
Interaction-related measures	Number of Questions Posted per Video, Number of Answers Posted per Video, Type of Interaction (Question or Answer), Date Posted, Content of The Interaction
User-related measures	User Name, Join Date, Streak Length, Energy Points, Avatar Name, City, Country, Number of Completed Videos

We develop interaction profiles for videos in the dataset (9076 videos) to classify them according to number of interactions associated with them. We used Tableau software to visualize the big number of users' interactions which have been posted over years in Box–Whiskers plots. **Fig. 11** shows different plots that represent distributional patterns of interactions in each domain separately over years. Math and Science related plots show that the interactions' distribution over years follows power law but with many outliers in 2011.

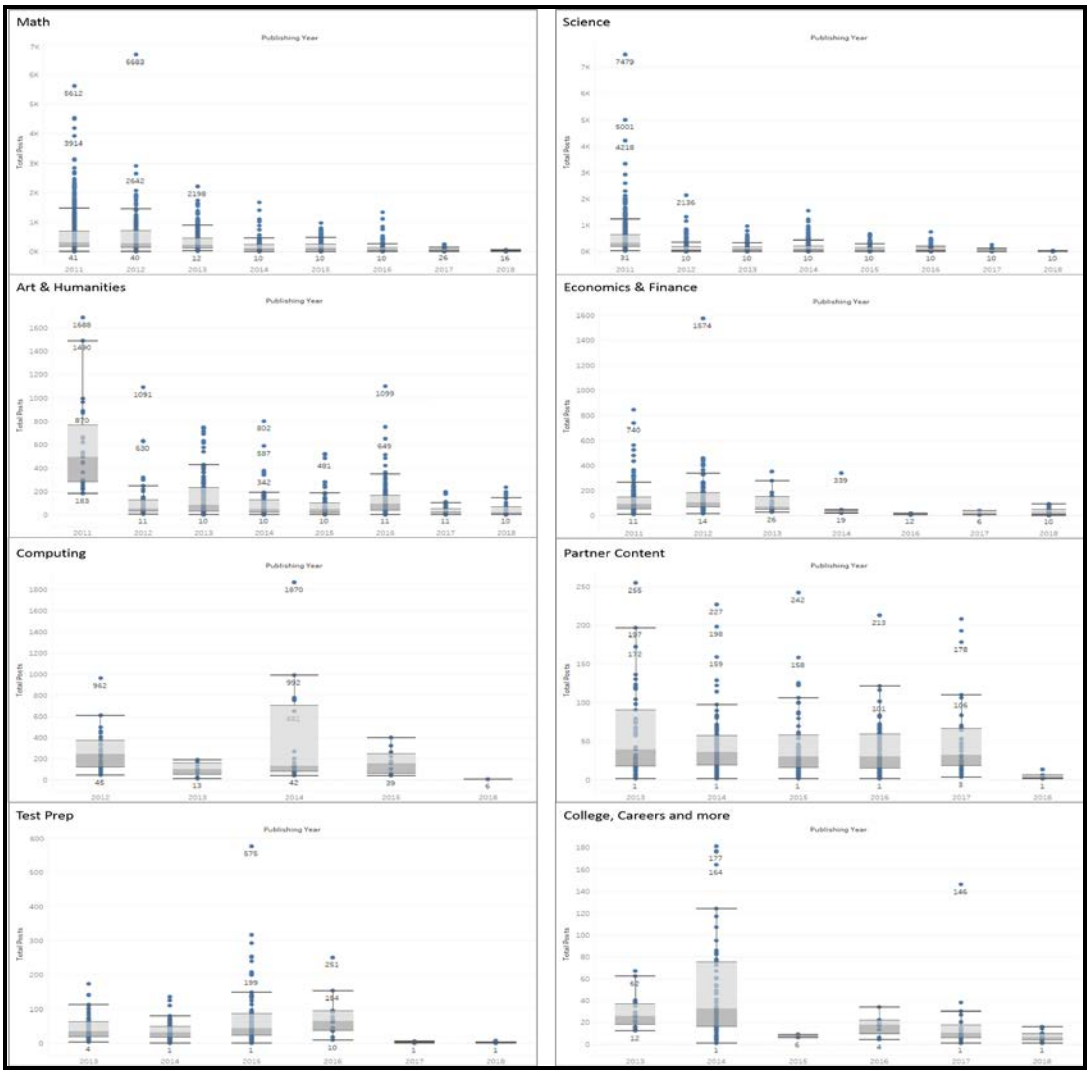

Fig. 11. Box Whiskers Plots showing Distributional Patterns of Users Interactions in each Domain

Arts & humanities and Economics & finance plots also follow power law but with some fluctuations in distribution. In Arts & humanities, the median of 2016 has raised to 94 interactions per video which is almost double of 2015's median then it dropped down to 28 interactions in 2017. Computing plot shows that videos posted in 2014 gained the highest number of interactions with a median of 163 interactions. It shows also that no new videos were published during 2017 and only one is published in 2018 which gives indication that computing domain is not a popular one anymore. This may be due to the high competition in different computing specialized repositories. Test prep and college careers more plots show that total interactions follow Poisson distribution over years. Test prep interactions' distribution reached its peak with videos published in 2016 with a median of 65 interactions per video while college careers more reached its peak in 2014 with a median of 32 interactions. Another Box Whiskers plot was conducted to present the overall distributional patterns of those interactions with all videos from all domains together **Fig. 12.**

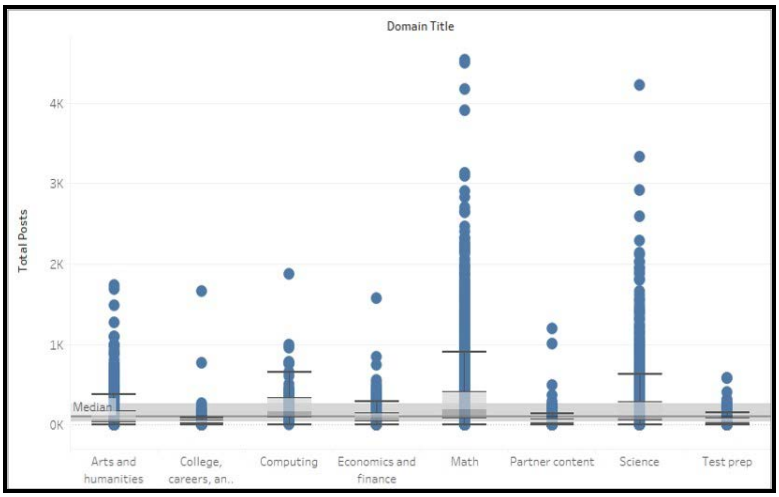

Fig. 12. Box Whiskers Plot showing total Users Interactions

Fig. 12 illustrates the overall median for total interactions across all domains in the dataset which is 104.5 interactions per video. The upper quartile is 266 interactions while the lower quartile is 39. According to that, we classified videos into three profiles based on number of interactions. Those three profiles are: high interaction profile, medium interaction profile and low interaction profile **Table 8**.

Table 8. Number of KA Learning Objects per Domain per Statistical Profile

Domain Title	Low-interaction Profile	Medium-interaction Profile	High-interaction Profile
Math	884	1548	730
Science	1009	1051	314
Arts and humanities	676	343	75
Economics and finance	125	310	28
Computing	3	43	24
Partner content	839	142	5
Test prep	445	156	5
College, careers, and more	285	33	3

Low interaction profile considers learning objects with number of interactions that falls below the interquartile range. This includes videos with (1 to 38) interactions. This category was found to be the largest one by including most of learning objects in the dataset (4266 videos. 1009 learning objects of them are related to science. The second-high domain is math with 884 videos while the lowest contribution is for computing with only 3 videos.

Medium interaction profiles include learning objects with number of interactions belongs to the interquartile range (39 and 266 included. This category contains 3626 videos. The highest contribution is for math domain with 1548 videos while the second is science with 1051 videos. The lowest contribution is for college careers more domain with 33 videos.

To consider the learning object as a high interaction profile, number of interactions must be above interquartile range which means above 266. This profile includes 1184 videos only. The highest number of interactions was for “What is Programming?” learning object from computing domain which reached 9454 interactions.

5.3 Analyzing the reactions of Videos' Profiles' toward different metrics:

Analysis was conducted to test relationship between some metrics and number of interactions used to develop interaction profiles. The analysis was conducted on each profile separately to compare their patterns.

5.3.1 Relationship between Domain and Number of Interactions:

Pearson Chi Square test and Phi & Cramer's V test was applied in **Table 9** to measure the strength of association between the type of video's domain and number of interactions posted on the same video. Chi Square test showed a significant association between them in low interaction profile. Having a Phi value of (0.417, 0.000) indicated that this relation is a strong positive one. On the other hand, Chi Square test for medium interaction and high interaction profiles showed that there is no statistically significant association between domain's type and number of interactions.

Table 9. Testing Domain vs. Total Number of interactions

Chi-Square Tests									
		Low Interaction Profiles			Medium Interaction Profiles			High Interaction Profiles		
		Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square		741.382[a]	259	.000	1564.395[a]	1589	.665	4203.365[a]	4333	.919
Likelihood Ratio		733.839	259	.000	1481.854	1589	.973	1583.111	4333	1.000
N of Valid Cases		4266			3626			1184		

Symmetric Measures							
		Low Interaction Profiles		Medium Interaction Profiles		High Interaction Profiles	
		Value	Approximate Significance	Value	Approximate Significance	Value	Approximate Significance
Nominal by Nominal	Phi	.417	.000	.657	.665	1.884	.919
	Cramer's V	.158	.000	.248	.665	.712	.919
N of Valid Cases		4266		3626		1184	

5.3.2 Relationship between OER's Publishing Year and Number of Interactions:

We applied Pearson Chi Square test and Phi & Cramer's V test in **Table 10** to measure the strength of association between video's publishing year and number of interactions posted on the video. Chi Square test showed that there is a strong positive significant relation between them for the low interaction profile videos. Although it is less significant than low interaction profiles, but still medium interaction profile showed a statistically significant association with publishing year. Chi square test for high interaction profile showed that there is no significant relationship between publishing year and number of interactions.

Table 10. Testing Publishing Year vs. Number of interactions

Chi-Square Tests									
	Low Interaction Profiles			Medium Interaction Profiles			High Interaction Profiles		
	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	1197.363	259	.000	1689.763	1589	.039	3129.799	3714	1.000
Likelihood Ratio	1130.017	259	.000	1753.494	1589	.002	2038.576	3714	1.000
Linear-by-Linear Association	541.100	1	.000	164.162	1	.000	27.984	1	.000
N of Valid Cases	4266			3626			1184		

Symmetric Measures							
		Low Interaction Profiles		Medium Interaction Profiles		High Interaction Profiles	
		Value	Approximate Significance	Value	Approximate Significance	Value	Approximate Significance
Nominal by Nominal	Phi	.530	.000	.683	.039	1.626	1.000
	Cramer's V	.200	.000	.258	.039	.664	1.000
N of Valid Cases		4266		3626		1184	

5.3.3 Relationship between Number of Interactions, video length and reuse rate:

A Regression analysis (**Table 12** and **Table 13**) was applied to find relationship between number of interactions as a dependent variable and video length and number of video reuses in different subjects, those two variables were considered as predictors. ANOVA test showed significant associations for all interaction profiles. The resulted regression models for each profile have the following appearance **Table 11**.

Table 11. Results of regression models for each profile

Low Interaction Profile:	$\hat{y} = 11.928 + 0.003x_0 + 0.684x_1$
Medium Interaction Profile:	$\hat{y} = 82.159 + 0.014x_0 + 7.601x_1$
High Interaction Profile:	$\hat{y} = 406.498 + 0.315x_0 + 37.198x_1$

Where $\hat{y}$ represents the estimated value of number of interactions related to videos in a specific profile and x_0 represents video length. Different coefficients indicate that there is a weak positive correlation. The high interaction profile specifically can be ranked as the highest one in terms of the correlation between video length as independent variable and number of interactions associated with the video. x_1 represents number of video reuses in different subjects. The coefficients show a significant positive association between reuse rate and number of interactions. This association is weak in low interaction profile while it is strong in the high interaction profile.

Table 12. Testing Video Duration and Number of Reuses vs. Number of interactions ANOVA Test

ANOVA[a]						
Low Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3237.718	2	1618.859	14.765	.000[b]
	Residual	467410.742	4263	109.644		
	Total	470648.460	4265			
Medium Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	383789.462	2	191894.731	56.815	.000[b]
	Residual	12236919.450	3623	3377.565		
	Total	12620708.910	3625			
High Interaction Profiles						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10514728.470	2	5257364.237	11.991	.000[b]
	Residual	517789342.100	1181	438432.974		
	Total	528304070.600	1183			

a. Dependent Variable: Total Posts

b. Predictors: (Constant), Reuses (for the video), Duration

Table 13. Testing Video Duration and Number of Reuses vs. Total Number of interactions Coefficients

Coefficients						
Low Interaction Profiles						
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	11.928	.528		22.579	.000
	Duration	.003	.001	.067	4.368	.000
	Reuses (for the video)	.684	.194	.054	3.528	.000
Medium Interaction Profiles						
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	82.159	2.939		27.953	.000
	Duration	.014	.004	.062	3.678	.000
	Reuses (for the video)	7.601	.716	.179	10.611	.000
High Interaction Profiles						
Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	406.498	60.560		6.712	.000
	Duration	.315	.074	.130	4.237	.000
	Reuses (for the video)	37.198	9.853	.116	3.775	.000

6. Conclusions and outlook

In this paper, we presented a large scale of data driven analyses on evolution of Khan Academy's repository and its characteristics. We analyzed data that has been gathered using KA's API and web scraping techniques. We studied a collected dataset of KA's instructional videos, their characteristics and their users' interactions. Our descriptive analyses include general analysis related to the repository itself such as: videos duration, number of completed videos by users and users' geographical distribution and they include interactions related analysis, such as: number of interactions per video, their distribution around KA contents and we clustered the videos according to those interactions. We developed interaction profiles based on the total number of questions and answers posted by users on those videos. Those profiles were categorized into three categories based on number of interactions associated with them. We tested their relationship with different characteristics such as domain which reflects the type of content, publishing year which reflects the age of video, video length and reuse rate which shows how many times the video is used in different subjects. The low interaction profile videos showed a strong significant correlation with all those characteristics. Medium interaction profile showed a correlation with publishing year, video length and reuse rate but there is no significant relation with domain's type. High interaction profile showed no correlation with domain's type and publishing year but there is still a significant relation with video length and reuse rate.

Our results show that the most significant variables that affect number of interactions in all profiles are reuse rate and video length. This is because most of the users prefer short videos although they may attend to somewhat longer videos if it is justified [22], but they still get bored of the long videos especially if they are without any segmentation. Reusing the instructional video in different places, sites and learning repositories increases its watching rate which will definitely raise the users' interaction rate as well. The developed interaction profiles method can play a good role in creating an automated model for videos' quality evaluation especially if we know that Khan Academy doesn't offer rating options to its instructional videos and they don't show any method to measure the quality of those objects.

The presented work has limitations, but it can still be an introduction to further research attempts. One limitation is that the implemented study covered only videos which are the most common type of learning objects in KA repository, but it didn't include other types of objects such as exercises. Another limitation is the analyzed interactions include posted questions and answers only. There are many types of users' interactions and all of them can present different indications regarding the quality of learning process. A third limitation is that we gathered users' interactions regardless the quality of their content. Content of the posts might be core concern related to the content of video or it can be irrelevant.

Interacting with instructional videos is an interesting activity for millions of users. A massive number of interactions can be turned in to valuable information that informs instructional designers and learners. A dynamic analysis process for those rich interactions' datasets implemented in regular basis [23] will offer opportunities for enhancing the whole learning process in general and the quality of learning objects specifically. Studying users interactions and understanding their behavior definitely will help in making better educational content and learning technologies [24]. More advanced future analysis can include the whole interactive environment provided by Khan Academy as a virtual classroom which includes not only instructional videos but also slides, lecture notes, exercises and interactive buttons. Linking this whole picture to the impact on education will contribute in enhancing students' learning performance, efficiency and the whole experience [25]. Other promising results can

be found through analyzing the content of users' interactions to define their behavioral patterns and study their engagement. As we know that KA videos are used widely and reposted in many different repositories, more promising results can be associated in analyzing KA videos that are reused outside the repository. KA repository is a huge repository connecting thousands of users around the world who are producing millions of interactions. This rich environment can be useful ground for fruitful research projects that attracts researchers from different backgrounds.

References

[1] Wiley, D., "Connecting Learning Objects to Instructional Design Theory: A definition, a metaphor, and a taxonomy," *Instr. use Learn. objects*, 10, 1–35, 2000. Article (CrossRef Link)

[2] Lee, B.G., Kim, S.J., Park, K.C., Kim, S.J., Jeong, E.S., "Empirical analysis of learning effectiveness in u-learning environment with digital textbook," *KSII Trans. Internet Inf. Syst*., 6, 869–885, 2012. Article (CrossRef Link)

[3] Zengin, Y., "Investigating the use of the Khan Academy and mathematics software with a flipped classroom approach in mathematics teaching," *Educ. Technol. Soc*., 20, 89–100, 2017. Retrieved from Article (CrossRef Link)

[4] Kay, R.H., "Exploring the use of video podcasts in education: A comprehensive review of the literature," *Comput. Human Behav*., 28, 820–831, 2012. Article (CrossRef Link)

[5] Cargile, L.A., Harkness, S.S., "Flip or Flop: Are Math Teachers Using Khan Academy as Envisionedby Sal Khan?," *TechTrends*, 59, 21–28, 2015. Article (CrossRef Link)

[6] Hwang, G.-J., Lai, C.-L., Wang, S.-Y., "Seamless flipped learning: a mobile technology-enhanced flipped classroom with effective learning strategies," *J. Comput. Educ.*, 2, 449–473, 2015. Article (CrossRef Link)

[7] Karabulut-Ilgu, A., Jaramillo Cherrez, N., Jahren, C.T., "A systematic review of research on the flipped learning method in engineering education," *Br. J. Educ. Technol*, 49, 398–411, 2018. Article (CrossRef Link)

[8] Ochoa, X., Duval, E., "Quantitative analysis of learning object repositories," *TLT*, pp. 226–238, 2009. Article (CrossRef Link)

[9] Hylén, J., "Open Educational Resources : Opportunities and Challenges," *Proc. Open Educ.*, 4963, 49–63, 2006. Article (CrossRef Link)

[10] Thompson, C., "How Khan Academy is changing the rules of education," *Wired Mag.*, 126, 1–5, 2011. Retrieved from Article (CrossRef Link)

[11] Santos-Hermosa, G., Ferran-Ferrer, N., Abadal, E., "Repositories of open educational resources: An assessment of reuse and educational aspects," *Int. Rev. Res. Open Distance Learn*, 18, 84–120, 2017. Article (CrossRef Link)

[12] Cechinel, C., Sánchez-Alonso, S., García-Barriocanal, E., "Statistical profiles of highly-rated learning objects," *Comput. Educ.*, 57, 1255–1269, 2011. Article (CrossRef Link)

[13] Cechinel, C., Camargo, S.D.S., Sànchez-Alonso, S., Sicilia, M.Á., "Towards automated evaluation of learning resources inside repositories," *Recommender Systems for Technology Enhanced Learning: Research Trends and Applications, Springer*, pp. 25–46, 2014. Article (CrossRef Link)

[14] Clements, K., Pawlowski, J., Manouselis, N., "Open educational resources repositories literature review – Towards a comprehensive quality approaches framework," *Comput. Human Behav*., 51, 1098–1106, 2015. Article (CrossRef Link)

[15] Robinson, D.E., Wizer, D.R., "Universal Design for Learning and the Quality Matters Guidelines for the Design and Implementation of Online Learning Events," *Int. J. Technol. Teach. Learn*, 12, 17–32, 2016. Article (CrossRef Link)

[16] Holland, A.A., "Effective principles of informal online learning design: A theory-building metasynthesis of qualitative research. Comput," *Educ*., 128, 214–226, 2019. Article (CrossRef Link)

[17] Kim, J., Li, S.-W. (Daniel), Cai, C.J., Gajos, K.Z., Miller, R.C., “Leveraging Video Interaction Data and Content Analysis to Improve Video Learning," in *Proc. of CHI ’14 Workshop on Learning Innovation at Scale*, pp. 1–6, 2014.Article (CrossRef Link)
[18] Li, N., Kidziński, L., Jermann, P., Dillenbourg, P., “MOOC video interaction patterns: What do they tell us?," Design for Teaching and Learning in a Networked World, pp. 197–210, 2015. Article (CrossRef Link)
[19] Haya, P.A., Daems, O., Malzahn, N., Castellanos, J., Hoppe, H.U., “Analysing content and patterns of interaction for improving the learning design of networked learning environments," *Br. J. Educ. Technol*, 46, 300–316, 2015. Article (CrossRef Link)
[20] Loff, S., “NASA, Khan Academy Collaborate to Bring STEM Opportunities to Online Learners, https://www.nasa.gov/content/nasa-khan-academy-collaborate-to-bring-stem-opportunities-to-online-learners
[21] Khan Academy Practice | SAT Suite of Assessments – The College Board, https://collegereadiness.collegeboard.org/about/benefits/khan-academy-practice
[22] Alpert, F., Hodkinson, C.S., “Video use in lecture classes: current practices, student perceptions and preferences," *Educ. Train*, 61, 31–45, 2019. Article (CrossRef Link)
[23] Giannakos, M.N., Chorianopoulos, K., Chrisochoides, N., “Making sense of video analytics: Lessons learned from clickstream interactions, attitudes, and learning outcome in a video-assisted course," *Int. Rev. Res. Open Distance Learn*, 16, 260–283, 2015. Article (CrossRef Link)
[24] Saurabh, S., Gautam, S., “Modelling and statistical analysis of YouTube’s educational videos: A channel Owner’s perspective," *Comput. Educ*., 128, 145–158, 2019. Article (CrossRef Link)
[25] Yip, J., Wong, S.H., Yick, K.L., Chan, K., Wong, K.H., “Improving quality of teaching and learning in classes by using augmented reality video," *Comput. Educ*., 128, 88–101, 2019. Article (CrossRef Link)

Sahar Yassine is a PhD student in the Computer Science Department at Alcalá University. Her Doctoral research investigates interactions with online learning and focuses on detecting community’s in e-learning environments. She has interests in educational repositories, learning-technology and social network analysis techniques.

Seifedine Kadry has a Bachelor degree in Applied Mathematics in 1999 from Lebanese University, MS degree in Computation in 2002 from Reims University (France) and EPFL (Lausanne), PhD in 2007 from Blaise Pascal University (France), HDR degree in Engineering Science in 2017 from Rouen University. At present his research focuses on education using technology, smart cities, system prognostics, stochastic systems, and probability and reliability analysis. He is a fellow of IET, fellow of ACSIT and ABET program evaluator.

Miguel-Angel Sicilia is currently full professor at the Computer Science Department of the University of Alcalá (Madrid, Spain. He holds degrees in Computer Science (Pontifical University of Salamanca) and in Information Science (University of Alcalá) and a PhD in Computer Science from Carlos III University. Before joining academia, Miguel Angel was part of the R&D and e-commerce architecture staff of iSOCO. Miguel Angel has developed his research activity in the fields of Artificial Intelligence, machine learning and analytics applied to different fields, including learning, health, computational science, command and control systems and information security.