Ranking Educational Videos: The Impact of Social Presence

Dimitrios Kravvaris
Department of Informatics
Ionian University
Corfu, Greece
jkravv@gmail.com

Katia Lida Kermanindis
Department of Informatics
Ionian University
Corfu, Greece
kerman@ionio.gr

Konstaninos Chorianopoulos
Department of Informatics
Ionian University
Corfu, Greece
choko@ionio.gr

***Abstract*—The information conveyed via the social media, in addition to the content data, also contains social characteristics that come from the social network users. A special interesting data category concerns the data that come from the natural language present in the social media mainly in the form of video. Our study focuses on the speech content of the videos in the form of transcript and the opinion of the social network users that have watched them. The representation of content data is made through a vector space model that uses cosine similarity measure for the relevant ranking of the transcripts. In order for the ranking to be more comprehensive we suggest the addition of a new parameter that of social weight during the procedure, which will reflect the users' opinion. There is an analytic presentation of the method being suggested; all the possible cases are being examined and the rules that define the new ranking are put forward. Furthermore, we apply this method to video lectures derived from YouTube. The findings of the experiments show that the addition of the social weight parameter reflects the users' views without changing to great extent the content based ranking of the video lectures. Finally, a user evaluation experiment was carried out and showed that the ranking procedure that includes the social weight parameter is closer to the users' ranking preferences.**

Keywords—cosine similarity; social media; educational video; ranking; mean average precision

I. INTRODUCTION

The search for information through the Internet constitutes an object of continuous research by many research communities since the nature of the Internet is dynamic [3]. The piece of information nowadays spreads rapidly from the moment it is being uploaded via the social networks. Both the information content and its social presence in the network are two characteristics that are of special interest in the process of searching. On the one hand, it is the content of the data that the network user looks for and on the other hand, it is the users' view related to this data which gives it an additional qualitative value. In the present study, we propose a merging of the content and social characteristics when searching for information and we study this merging in real data.

The types of data published in the social networks vary. The data that come from the human natural language constitutes a special data category since it can be represented either as speech or as written text [9]. In our experiments we used data from the YouTube, which constitutes the largest provider of the human natural language data and at the same time it is the largest social video content network [22]. More specifically, we chose to study educational video lectures among numerous theme units. Through YouTube API, it is possible for us to extract both the speech content of the videos as text (transcripts), which represents the natural language, and the social characteristics such as the number of *likes* and *dislikes* of the users that watched them [17].

We also chose to rank the search findings based on the content, that is, the transcript of the natural language of the speaker. Thus, we suggest the application of the vector space model [20] that represents the words of the transcript as a vector and can be used for relevancy ranking [7]. Each word corresponds to one dimension of the vector and its value is estimated by the *tf-idf* weighting scheme [18]. More specifically, the vector space model uses cosine similarity as a ranking measure, which estimates the similarity of two vectors as a cosine of the angle between them [15]. Cosine similarity is a popular method, suitable to be applied to high-dimensional text data [24] [2] [12], such as the transcripts of video lectures.

As for our next innovation, i.e. the addition of the social weight value to each video, we chose the *likes* and the *dislikes* of each video for this purpose [16]. More analytically, we chose the formula *like*/(*likes*+*dislikes*), since it can describe how likable a video lecture is. The likes and dislikes are actions realized only by registered users, who are more possible to characterize a video after they have watched it. It should be stressed, though, that through their use it is obvious how dynamic the social media are, since the number of *likes* and *dislikes* change dynamically the users' preferences. We want the social weight parameter to function positively on the video lectures under examination and to add value to them. Any video lecture a) with only *dislikes* or without any *like* or b) without any *like* or *dislike* is not given any additional value besides the one it has from its content.

With our aforementioned choice we avoid the problem which would be created if the number of views was used as social weight, since the number of the viewers of a video is being recorded without knowing if they liked it or not. If, on the other hand, we chose the comments on the videos as social weight we would face two problems: the first one would be the complicated and time consuming procedure required in order to

978-1-4673-6630-4/15/$31.00 ©2015 IEEE

characterize the users' opinion [14] [6] and the second would concern the ability to comment the video lecture, which could be deactivated by the creator and, thus, we would have no relevant comments.

A user evaluation procedure was necessary so as to compare the ranking of the videos suggested by the users to the content based ranking and the content and social based ranking that is created by adding the social weight. This comparison is possible through mean average precision measure [25], which has been shown to have especially good discrimination and stability [28]. Furthermore, through the user evaluation procedure we can record the qualitative features that are related to the uses' choice concerning video ranking. Thus, it is possible to have a complete picture of the way the users use the most appropriate educational videos.

Taking, thus, all the above into consideration, in the first part of the paper, we present works related to our study, while in the second part we provide a theoretical analysis of the ranking algorithm we suggest. More specifically, we study all the possible cases of content similarity and social weight that take part in our ranking proposal. In the next part, we apply the algorithm to real data. More analytically, we present the methodology of the experiment on real data from YouTube and we present the findings which are then being discussed. Furthermore, a part of user evaluation is included in our paper that compares the users' preferences in ranking to the ones of our proposal. In the end, we present the conclusions extracted from our study.

II. Related Work

Document ranking based on the use of the content similarity has been the object of study in many surveys of the past [1] [13] and it still is in newer studies, which have retained the basis of the cosine similarity technique and have applied it both in surveys relevant to clustering [19] and to ones relevant to semantic similarity [4]. Furthermore, the introduction of the social media in education [8] [5] has led the researchers to the suggestion of frameworks that aimed to the combination of data's content and the respective opinions of the users for the search and ranking of data. Some of the studies that led us to the creation of our own proposal are presented below.

The early research [26] constitutes one of the first attempts which uses the vector space model for the representation of the content of a multimedia database. The users of the database provide weight for positive examples. The aim is for the searches to have as similar results as possible. Our research moves a step further, since we apply the whole process to the largest video database in the world, i.e. YouTube. Thus ,we have a more objective system by adding the opinion of the YouTube registered users since it is not limited to one or few users but to the whole social network community.

A more recent study [27] aims at the ranking of the web content by combining the content- similarity with the users-similarity. The suggested user-similarity parameter is based on previous results of a query i.e. if a user uses the same query as the previous users the results they will get will be similar to the ones the previous users chose. Although there is a recording of the visits of the web data, there is no recording of the user's opinion on it, a parameter taken into consideration in our research.

Another study [31] uses a social-textual approach to search the most suitable for the user information in the web. The social part is based on parameters derived from the friends of the user in a social network. Those parameters are focused within the social network and the users' connection to it is required in order to proceed with the search. Our approach is released from this commitment and applies to all users regardless of whether they are enrolled in social networks or not.

Finally, research [29] suggests a new ranking framework which adds the basic *tf-idf* score based on metadata and a social score which is the similarity between the user that searches for a video and the user that owns the video. The choice of these two parameters is present in our research, as well. We, however, give emphasis on the content of natural language of the video and not on the metadata and we also suggest that the users' opinions should be taken into consideration since it is a qualitative parameter, as we mentioned before.

The status of this content/external-parameters (social parameters and others) interface has been studied in the literature, as is described in this section. Studies taking into account this combination report in general a positive impact of these parameters on the content ranking performance.

III. Content and Social Approach

The first part of our ranking scheme is implemented using the transcripts of the video lectures, as vectors. More specifically, we estimated the content similarity, using the cosine similarity measure between the transcripts of the video lectures and a search query. Following this methodology we find the similarity value of every video lecture in relation to the query search and we have a first ranking. The formula below describes the similarity value (s), with d as the vector of the words of the video's transcript and q as the vector of the words of the search query. V parameter shows the dimension of the vector space.

$$s = \cos(\vec{q}, \vec{d}) = \frac{\vec{q} \bullet \vec{d}}{\left|\vec{q}\right|\left|\vec{d}\right|} = \frac{\vec{q}}{\left|\vec{q}\right|} \bullet \frac{\vec{d}}{\left|\vec{d}\right|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2} \sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

The suggested social-content similarity (*scs*) which takes into consideration the users opinions on the video lecture, hereafter social weight (*sw*) which equals to the formula *likes*/(*likes*+*dislikes*) gives social value to the video lecture at a specific time. The social-content similarity formula in its final form is as follows:

$$scs = s(1 + sw) = \cos(\vec{q}, \vec{d})(1 + \frac{likes}{likes + dislikes})$$

The s parameter takes values from 0 to 1, since we study only the video transcripts that are related, so the angle between the vectors d and q is between 0 and 90 degrees. The *sw* parameter takes values from 0 to 1, as well, since the number of *likes* cannot exceed the total number of *likes* and *dislikes*. Finally, in order for the additional values of the *sw* to the s to

be obvious, we multiply it with (1+*sw*), increasing, thus, the *scs* value area and takes values from 0 to 2.

In the second part we analyze the social-content similarity outcomes of all the content similarity (using the cosine similarity measure) and social weight cases that can be observed. Thus, we examine the ranking between two transcript vectors of video lectures (d_1 and d_2) based on a particular query vector (q) calculating the cosine similarity measure and how this ranking order changes adding the social weight parameter for each one of the videos. The cases under examination are as follows.

A. *Equal cosine similarity*

Cases where the two initial cosine similarity values are equal: $\cos(\theta_1)=\cos(\theta_2)$, i.e. $\theta_1=\theta_2$, are described analytically below.

1) Case $s_1=s_2$ and $sw_1=sw_2$

$$sw_1 = sw_2 \Rightarrow 1+sw_1 = 1+sw_2 \quad (1)$$

$$s_1 = s_2 \overset{(1)}{\Rightarrow} s_1(1+sw_1) = s_2(1+sw_2) \Rightarrow scs_1 = scs_2$$

We find out that in this case there is no change in the ranking of the videos under examination.

2) Case $s_1=s_2$ and $sw_1>sw_2$

$$sw_1 > sw_2 \Rightarrow 1+sw_1 > 1+sw_2 \quad (2)$$

$$s_1 = s_2 \overset{(2)}{\Rightarrow} s_1(1+sw_1) > s_2(1+sw_2) \Rightarrow scs_1 > scs_2$$

We find out that in this case there is a change in the ranking of the videos under examination i.e. the one with the highest value of social weight takes a higher ranking position.

3) Case $s_1=s_2$ and $sw_1<sw_2$

$$sw_1 < sw_2 \Rightarrow 1+sw_1 < 1+sw_2 \quad (3)$$

$$s_1 = s_2 \overset{(3)}{\Rightarrow} s_1(1+sw_1) < s_2(1+sw_2) \Rightarrow scs_1 < scs_2$$

We find out that in this case there is a change in the ranking of the videos under examination i.e. the one with the highest value of social weight takes a higher ranking position.

B. *Unequal Cosine Similarity*

Cases where the two initial cosine similarity values are not equal: $\cos(\theta_1)>\cos(\theta_2)$, i.e $\theta_1<\theta_2$, are described below.

1) Case $s_1>s_2$ and $sw_1=sw_2$:

$$sw_1 = sw_2 \Rightarrow 1+sw_1 = 1+sw_2 \quad (4)$$

$$s_1 > s_2 \overset{(4)}{\Rightarrow} s_1(1+sw_1) > s_2(1+sw_2) \Rightarrow scs_1 > scs_2$$

We find out that in this case there is no change in the ranking of videos under examination.

2) Case $s_1>s_2$ and $sw_1>sw_2$:

$$sw_1 > sw_2 \Rightarrow 1+sw_1 > 1+sw_2 \quad (5)$$

$$s_1 > s_2 \overset{(5)}{\Rightarrow} s_1(1+sw_1) > s_2(1+sw_2) \Rightarrow scs_1 > scs_2$$

We find out that in this case there is no change in the ranking of videos under examination.

3) Case $s_1>s_2$ and $sw_1<sw_2$:

In this case the *scs* has a different outcome, as it is shown in graphs in Fig. 1, Fig. 2 and Fig. 3, of the formula $scs=s(1+sw)$ of the values under examination. So, sometimes there is a change and some others there is not in the ranking of the videos. We analyse below when these changes occur and in what conditions.

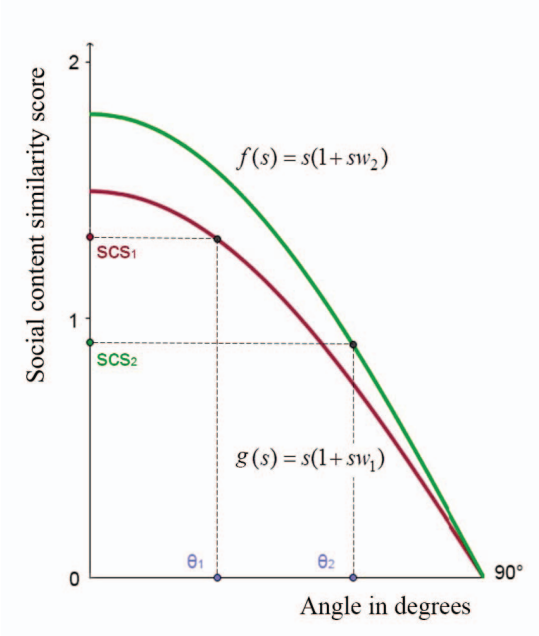

Fig. 1. Case where scs_1 is greater than scs_2.

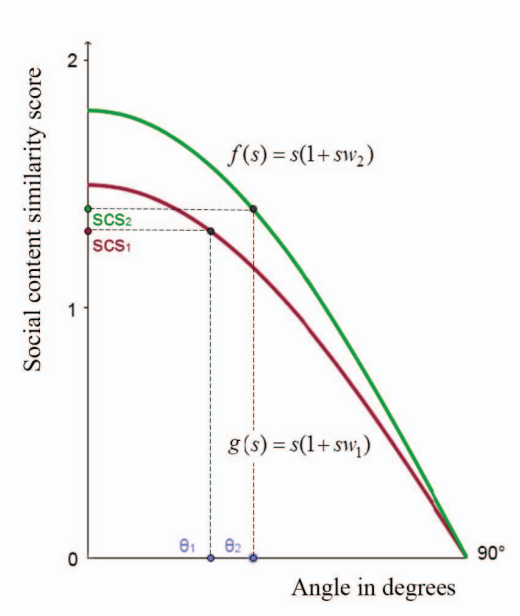

Fig. 2. Case where scs_2 is greater than scs_1.

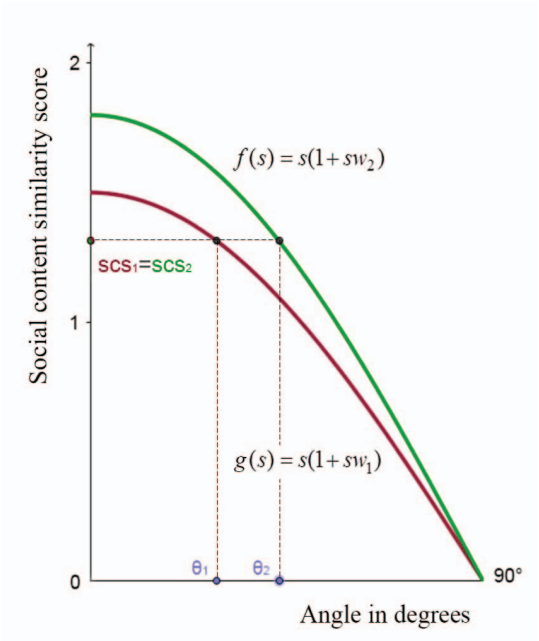

Fig. 3. Case where scs_1 is equal to scs_2.

a) In the case of $scs_1 > scs_2$:

$$scs_1 > scs_2 \Rightarrow s_1(1+sw_1) > s_2(1+sw_2) \Rightarrow \frac{s_1}{s_2} > \frac{1+sw_2}{1+sw_1}$$

The $scs_1 > scs_2$ is valid when the above equation is valid for the values of s_1, sw_1 and s_2, sw_2 of the two videos under examination. We find that in this case there is no change in the ranking of the videos.

b) In the case of $scs_1 < scs_2$:

$$scs_1 < scs_2 \Rightarrow s_1(1+sw_1) < s_2(1+sw_2) \Rightarrow \frac{s_1}{s_2} < \frac{1+sw_2}{1+sw_1}$$

The $scs_1 < scs_2$ is valid when the above equation is valid for the values of s_1, sw_1 and s_2, sw_2 of the two videos under examination. We find that in this case there is a change in the ranking of the videos.

c) In the case of $scs_1 = scs_2$:

$$scs_1 = scs_2 \Rightarrow s_1(1+sw_1) = s_2(1+sw_2) \Rightarrow \frac{s_1}{s_2} = \frac{1+sw_2}{1+sw_1}$$

The $scs_1 = scs_2$ is valid when the above equation is valid for the values of s_1, sw_1 and s_2, sw_2 of the two videos under examination. We find that in this case there is a change in the ranking of the videos and more specifically their difference decreases and as a result they have the same ranking position. In this case, we suggest that the initial ranking be kept. So d_1 vector should be ranked before d_2 since $s_1 > s_2$.

IV. Real Data Experiments

Next we conduct experiments on real data. In this way, we can study our proposal for the addition of the social weight parameter to the content similarity in order to find out from the results if there are changes in the ranking and to what extent these are represented in the above theoretical framework.

A. Data

Our data were video lectures collected from YouTube. Searching through the category of Education of YouTube by inserting 40 unique keywords from different scientific fields as shown in Table I, 20.830 video lectures were collected among which 1.116 (5.4%) had English transcript. The number of the videos that we study is quite large considering that the majority of the YouTube videos does not have transcripts that are necessary for our research. From each video we collected the transcript as well as the numbers of *likes* and *dislikes* using the YouTube API v2.

TABLE I. Keywords

Keywords			
computer science	data mining	machine learning	biology
medicine	web	c++	statistics
theory	art	php	social
physics	health	java	analysis
space	class	network	geography
human	programming	public	mathematics
teaching	economics	ted	internet
philosophy	financial	comptia	laboratory
chemistry	learn	algorithms	experiment
database	lecture	biology	energy

B. Methodology

Initially, we clustered all the transcripts based on the *tf-idf* schema for the word vector creation. All the transcripts are processed in Rapidminer v5.3 [10] [11] according to the following procedures in order to keep the most valuable unique words of each transcript:

- Tokenize (non letters). This procedure splits the texts of a transcript into a series of single word tokens.
- Filter Stopwords. It filters words that do not really add meaning to the transcript, such as the words a, and, the, of, etc.
- Stem English words (Porter algorithm). Porter algorithm removes the most common morphological and inflexional endings from English words [23].
- Transform cases. This operator transforms all characters to lower case.

We used K-means where K equals 40, which is the number of entries (keywords) used initially as search queries to collect our data from YouTube, and cosine similarity measure in order to calculate the distance between the objects in our clustering procedure. We chose K-means because it is a simple and flexible algorithm that is easy to understand and explains the clustering outcome [21]. The choice of cosine similarity

measure, as we mentioned in the beginning of this paper, is a suitable technique to be applied to high-dimensional text data, such as the transcripts of video lectures.

In every cluster we add a query as a text document and follow the "*Data to Similarity*" procedure in Rapidminer. What we will get as a result is a table with values from all the pair combinations of transcripts (documents) including the query based on cosine similarity. We keep the content similarity values of those pairs and by ranking them we can see the similarity rank of the transcripts in relation to the search query.

Then we add social weight which corresponds to every video lecture and we re-rank them. The new ranking has this time a social character and we can study the differences between the previous and the new one, moreover, we can examine the theoretical cases that appear according to the first part of our study.

C. *Experimental procedure*

Forty clusters were created, which on average contain 29 video lectures each. In Fig.4 below, we can observe the distribution of videos per cluster in ascending order. Furthermore, in Fig. 5 we present boxplots of the *likes*, *dislikes* and the social weight of the total values of the video lectures. From the boxplots we find out that the number of *likes* is higher than the number of *dislikes* on the video lectures since the majority of the social weight values is on average equal to 0.96. Moreover, there are some extreme cases where there is no *like* but there are *dislikes*. Cases, however, where there are only likes on a video lecture are more frequent (13%).

When we added social weight we observed some changes in the ranking of the video lectures compared to the ranking estimated only with the cosine similarity. In Fig. 6 below, we present how the ranking sequence changes.

We find out that the ranking of a great 43% of the video lectures has not changed. But 57% of the video lectures under examination have changed, which shows the dynamics that social weight has on the ranking of the results. In most cases a 41% of social weight had a positive effect on the video lectures while only in 16% of the cases the video lectures went down to a lower position in the ranking. This means that the addition of social weight can give different results from the ones we get when we simply use cosine similarity. Thus, the social network users have the opportunity to play a major role in this change by simply pressing the *like* or *dislike* button on the video lecture they watched.

More analytically, the changes in the ranking are shown in Fig. 7. As we can see, the changes in the ranking of the video lectures are usually 1 to 4 positions up or down, which covers about 81.74% of the video lectures that have changed order. There is the case, however, where a video lecture has gone down 23 positions. In this case we find out that the *sw*=0 and more specifically there is 1 *dislike* and no *likes*, The average *sw*, as we mentioned, is quite high 0.96 which means that the extreme cases where the *sw* is quite low are very few and cannot give any social value to them so as to change their position in the ranking of the video lectures.

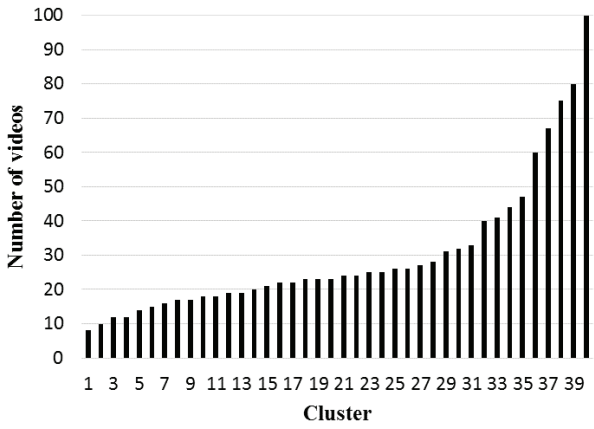

Fig. 4. Distribution of videos per cluster.

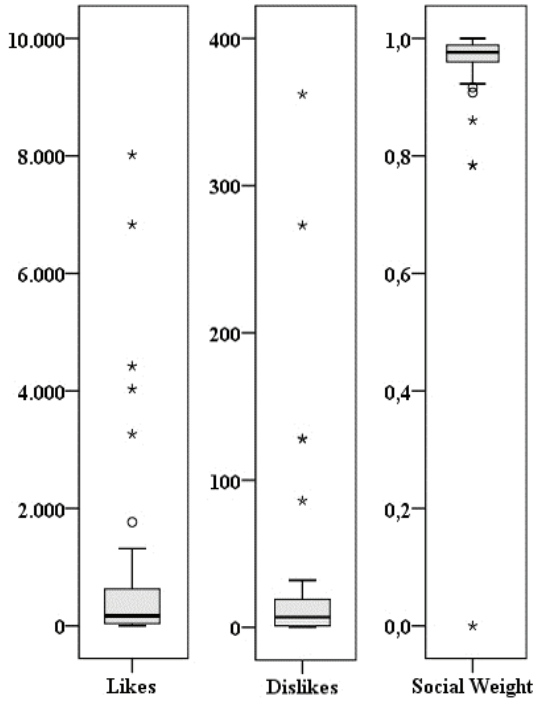

Fig. 5. Boxplot of Likes, Dislikes and Social Weight.

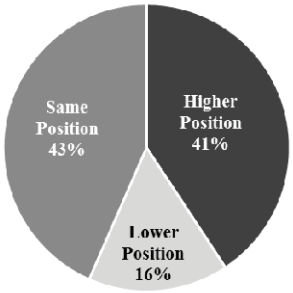

Fig. 6. Distribution of changes in the ranking sequence.

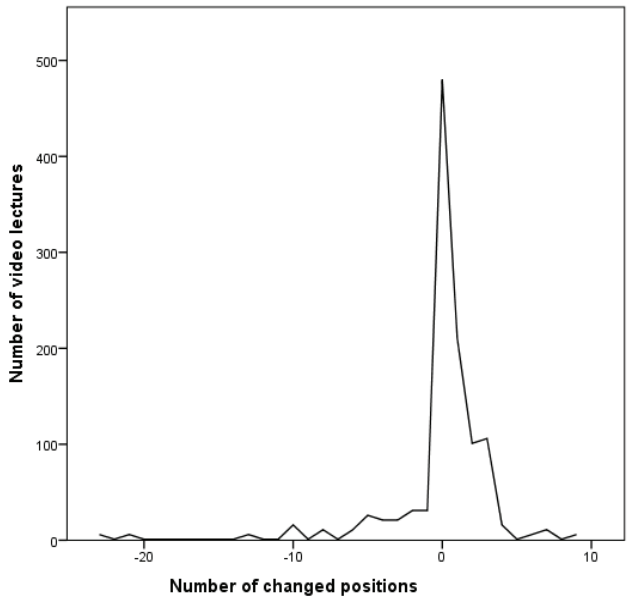

Fig. 7. Distribution of the number of changed positions.

TABLE II. ALL THE POSSIBLE CASES WHEN $s_1>s_2$

Case	Percent of Appearance
$s_1>s_2$ and $sw_1=sw_2$	9%
$s_1>s_2$ and $sw_1>sw_2$	47%
$s_1>s_2$ and $sw_1<sw_2$	44%

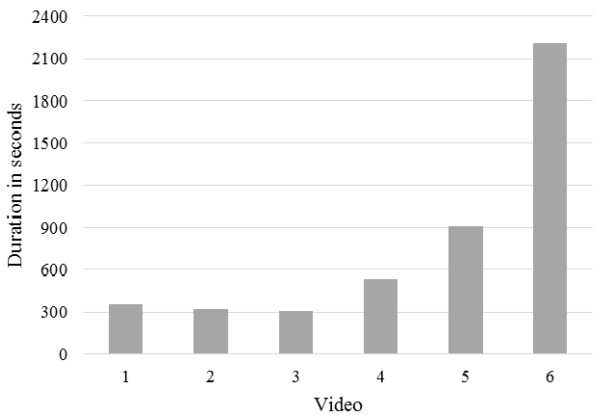

Fig. 8. Video duration.

D. Statistics

Finally, we present a statistic analysis which concerns the appearance of the theoretical cases in the experimental procedure. In our experiment we did not record any similar s, which means that the first theoretical case $s_1=s_2$ did not appear. We have, however, recorded similar *sw* and specifically in the case of *sw*=1, which seems to concern video lectures that have only *likes* and constitute the 16.28% of the total *sw*. All the cases of our experiment follow the second part of our theoretical approach. The percentage of appearance of each cases is presented in Table II.

The cases where the $sw_1=sw_2$ are only 9%, whereas the cases where the $sw_1>sw_2$ are only 3% more that the cases where $sw_1<sw_2$. The second part of the theoretical analysis (subsection III.B) that we present covers all the cases of the experiment we conducted, confirming, thus, the number of the possible cases that can occur with the addition of the social weight.

E. Results

The conclusions that we reach by applying our proposal to real data are really interesting. First of all, as far as the change of ranking is concerned, we find out that there is a change in 57% of the video lectures and 87% of them either go up or go down by 1 to 4 positions. That is, the change in the rank does not usually cause changes in the search results. More specifically, the neighboring video lectures after the first ranking with the content similarity do not differ much in score, so their change by 1 to 4 positions after the addition of the social weight parameter does not have a negative effect on the initial result. On the contrary, the ranking is made based on the opinion of the YouTube users on similar video lectures. As we observed during the experimental procedure the extreme cases of *sw* affect the order of ranking in a negative way. This means that we can find a relevant video lecture at a lower position. The opposite however, does not seem to happen. That is, a video lecture that was initially ranked with a low score cannot go up by many positions with the addition of *sw*.

What is also interesting is the confirmation of the cases of the model we suggested. The case where we would have two transcripts with the same content similarity (subsection III.A) in relation to the same query did not appear in our experimental procedure. However, all the other cases, as we described in the second part of our theoretical analysis (subsection III.B) did appear. The new ranking order of the video lectures, as was presented in the third case of subsection III.B ($s_1>s_2$ and $sw_1<sw_2$), depend both on their own social weight and on the social weight of the neighboring video lectures for the corresponding content similarity values of the specific search query under examination.

V. USER EVALUATION

In this part of the paper we present the ranking of the videos as it was suggested by the users and we compare the results to the ones of the content similarity method which uses the cosine similarity measure, and the ones of the social-content similarity method suggested in our paper. More analytically, we describe the collection of users' data the methodology we followed and we analyze the results both quantitatively and qualitatively.

A. Data

The data were collected from fifteen users-rankers who belong to different age groups and have different interests. We used online questionnaires which contained six videos from the keyword "*database*" clusters. We chose to study six videos

because that is the number of videos that appear on the users' screen after a search in YouTube and before they scroll down. The total duration of the videos was 77 minutes and the duration of each of the six videos appears in Fig. 8.

B. Methodology

The videos were presented to each ranker in a random order. Each user had to rank the videos in the order they would like them to be presented after a possible search of the word "database" on the YouTube. Furthermore, they were asked if they watched the whole videos or stopped watching them at some point and which type of video they preferred. They were also asked to describe both the positive and the negative features each video had, in their opinion, which helped to rank them. Finally, there was an interview with five of the rankers [30] which aimed at recording their attitude towards the videos they liked the most or not at all.

Our research question is to find out which of the two aforementioned ranking methods is closer to the users' ranking. We assume, thus, that every user's ranking is correct. We compare the ranking of every ranker to the content similarity video ranking and the social-content similarity video ranking respectively. We chose to use the Mean Average Precision (MAP) measure for this quantitative comparison.

MAP provides a single-figure measure of quality across recall levels [28]. The MAP's score in the case of the content similarity and social-content similarity rankings is defined as the mean of average precisions of all the correctly ranked videos based on the users' choice. Average precision for each case is defined as the mean of the precision at six values computed after each video is ranked in the right position.

C. Experiment

The suggested ranking of the six videos by each one of the fifteen users-rankers is presented in Table IV. Numbers 1 to 6 correspond to the six videos under examination. Table III presents the ranking based on the content similarity and the social-content similarity methods for the same videos. We observe that the two methods differ in the ranking only in the fourth and fifth position of the videos

Table IV shows that only user4 ranked the video number 4 above the video number 5, while the rest fourteen users ranked the video number 5 above the video number 4. Five out of these fourteen users have ranked those two videos exactly as they are ranked by the social-content similarity method. Moreover, all users preferred video 2 to video 1 in their ranking, which both the content similarity and the social-content similarity methods failed to rank in this order.

The MAP between each ranker and the content similarity and social-content similarity methods is shown in Table V below. We notice that the social-content similarity method is most of the times close to the users' ranking at satisfying percentages, considering that we have to measure human behavior. Thus, the social-content similarity method predicts the users preference by 27.39% while the content similarity by 19.23%.

As for the fourth and fifth position in the case of content similarity video number 4 and number 5 have been ranked as fourth and fifth respectively, however, none of the users ranked them so. In the case of social-content similarity in the fourth and fifth position videos number 5 and number 4 have been ranked respectively. In this case there are five rankers (user1, user2, user3, user13, user14) that made this ranking with an average percentage in the MAP measure that reaches 52.5%.

D. Results

From the quantitative outcomes of the experiments it is obvious that the social-content similarity method is more successful concerning the prediction of the correct ranking of the videos as regards the users' choice by 8.11%, compared to the content similarity method. The difference is statistically significant taking into account that the data comes from human actions [32]. Moreover, there are also qualitative features that were extracted both from the open-type questions that the users answered regarding the video ranking and personal interviews.

TABLE III. RANKING ORDER BASED ON SIMILARITY METHODS

	Ranking Order					
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
Content Similarity	1	2	3	4	5	6
Social-Content Similarity	1	2	3	5	4	6

TABLE IV. USERS RANKING ORDER

	Ranking Order					
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**
user1	2	1	3	5	4	6
user2	2	1	3	5	4	6
user3	2	1	3	5	4	6
user4	2	1	4	3	5	6
user5	2	1	3	5	6	4
user6	2	1	5	3	4	6
user7	2	1	5	6	4	3
user8	2	1	5	3	4	6
user9	2	1	5	3	6	4
user10	2	1	5	3	6	4
user11	2	1	5	3	6	4
user12	2	1	5	4	3	6
user13	2	1	3	5	4	6
user14	2	1	3	5	4	6
user15	2	1	5	3	6	4

TABLE V. MEAN AVERGAGE PRECISION

	Mean Average Precision	
	Content Similarity	**Social-Content Similarity**
user1	33.33%	52.50%
user2	33.33%	52.50%
user3	33.33%	52.50%
user4	26.67%	16.67%
user5	33.33%	41.67%
user6	16.67%	26.67%
user7	0.00%	20.00%
user8	16.67%	26.67%
user9	0.00%	0.00%
user10	0.00%	0.00%
user11	0.00%	0.00%
user12	29.17%	16.67%
user13	33.33%	52.50%
user14	33.33%	52.50%
user15	0.00%	0.00%

From the open type questions we find out that the videos were chosen by the users based on their simplicity, understating and directness of the speaker. On the other hand, they did not choose the videos that contained tiring details and the speaker was cold, in the opinion of each user. This shows that the content of a video plays a major role in its ranking. Furthermore, from the personal interviews we find out that the users choose to press the "*like*" option in a YouTube video more easily when the like it without having second thoughts. On the other hand, they would choose to press the "*dislike*" option more hesitantly and only in cases where the title of the video would be irrelevant to its content, or it would be a waste of time for someone to watch it. Most of the users that are not satisfied with a video just simply stop watching it. Thus, we understand that the positive opinion (*likes*) of the users as a measure is far more reliable that the corresponding *dislikes*, which reinforces the social weight formula we suggested *likes*/(*likes*+*dislikes*), having emphasized the positive opinions of the users.

VI. CONCLUSION

The ranking of information in the social media for data that come from natural language constituted the research fields of the present study. The representation of the natural language in transcripts helped us to use a vector space model approach. Our proposal to add the social weight parameter that measures the popularity of each piece of information was presented both in theoretical terms, where there was a thorough analysis and in practical terms by conducting experiments on real data derived from YouTube.

The theoretical analysis of our proposal was based on the content and the social presentation of the transcripts of the video lectures. The combination of these two factors define when there is a change in the ranking or not in relation to the simple ranking that exclusively concerns the content and is made using the cosine similarity. The dynamics of social weight that is added can reverse the initial ranking, can maintain it or reinforce it. The whole procedure affects the final score of each video under examination in a positive way and the race to the top of the ranking is continual and dynamic, since it is defined by the users' acts. Thus, the order of ranking can change depending on the users' preferences.

The experiment on real data showed interesting results. One of these is the fact that the majority of the changes in position does not affect the new ranking to such a degree that it would bring a less relevant to the content video lecture to the top positions, irrespectively of how high its social weight is. What is also interesting is that a low social weight can cause a video lecture to go down to a much lower position, which indicates the effect of the users' opinions, even on video lectures that are at the top positions based on their content relevance. Thus, the final ranking will contain the video lectures with the most relevant content and the most favorable views on the part of the users.

The user evaluation experiment confirmed that the social-content similarity method is more reliable than content similarity method, with a percentage that reaches 52.5% that the videos will be ranked in the same order as they will be ranked by the users. It is also confirmed that the content of video is considered to be more important than other characteristics. Furthermore, it was found out that the users express their positive opinion (*like*) for the videos more easily than their negative opinion (*dislike*). Thus, the choice of using the social content similarity method in the ranking of videos from social media is reinforced.

From all the above, it is suggested that the addition of this special social weight parameter to the ranking based on the content (in our case the natural language transcript) constitutes an appropriate technique for the utilization of the users' opinions on the content they have watched. We suggest further research be conducted in other types of information and more social networks separately or in combinations be utilized since each of them have its special characteristics.

REFERENCES

[1] J.A. Aslam, and M. Frost, "An information-theoretic measure for document similarity," in SIGIR, vol. 3, pp. 449-450, July 2003 .

[2] K. Bache, D. Newman, and P. Smyth, "Text-based measures of document diversity," in Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 23-31, August 2013.

[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, "The role of social networks in information diffusion," in Proceedings of the 21st international conference on World Wide Web, pp. 519-528, April 2012.

[4] C. Banea, Y. Choi, L. Deng, S. Hassan, M. Mohler, B. Yang, ... and J. Wiebe, "CPN-CORE: A Text Semantic Similarity System Infused with Opinion Knowledge," in Atlanta, Georgia, USA, p. 221, 2013.

[5] M. Blankenship, "How social media can and should impact higher education," in Education Digest, 76(7), pp. 39-42, 2011.

[6] H. Chen, and D. Zimbra, "AI and opinion mining," in Intelligent Systems, IEEE, 25(3), pp. 74-80, 2010.

[7] W.B. Croft, D. Metzler, and T. Strohman, "Search engines: Information retrieval in practice," in Reading: Addison-Wesley, p. 283, 2010

[8] K. Fernandez, M. d'Aquin, and E. Motta, "Linking data across universities: an integrated video lectures dataset," in The Semantic Web–ISWC 2011, pp. 49-64, Springer Berlin Heidelberg, 2011.

[9] E. Finegan, "Language: Its structure and use," in Cengage Learning, 2011.

[10] M. Hofmann, and R. Klinkenberg, "RapidMiner: Data Mining Use Cases and Business Analytics Applications," CRC Press, 2013.

[11] F. Jungermann, "Information extraction with rapidminer," in Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities', pp. 50-61, February 2009.

[12] H. Kim, P. Howland, H. Park and N. Christianini, "Dimension Reduction in Text Classification with Support Vector Machines," in Journal of Machine Learning Research, 6(1), 2005.

[13] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H.R. Garner, "Text similarity: an alternative way to search MEDLINE," in Bioinformatics, 22(18), pp. 2298-2304, 2006.

[14] B. Liu, "Sentiment analysis and opinion mining," in Synthesis Lectures on Human Language Technologies, 5(1), pp. 1-167, 2012.

[15] C.D. Manning, "Foundations of statistical natural language processing," H. Schütze (Ed.), MIT press, 1999.

[16] A. Oghina, M. Breuss, M. Tsagkias, and M. de Rijke, "Predicting imdb movie ratings using social media," in Advances in information retrieval, pp. 503-507, Springer Berlin Heidelberg, 2012.

[17] A. Padilla, and A. DeFields, "Beginning Zend Framework," Apress, 2009.

[18] J. Ramos, "Using tf-idf to determine word relevance in document queries," in Proceedings of the First Instructional Conference on Machine Learning, December 2003.

[19] A.R.M. Reddy, "Implementation of Multi View point method for similarity Measure in clustering the documents," in International Journal, 2(1), 2014.

[20] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing," in Communications of the ACM, 18(11), pp. 613-620, 1975.

[21] P. Vora, and B. Oza, B. "A Survey on K-mean Clustering and Particle Swarm Optimization," in International Journal of Science and Modern Engineering (IJISME), pp. 24-26, 2013.

[22] M. Wattenhofer, R. Wattenhofer, and Z. Zhu, "The YouTube Social Network," in ICWSM, June 2012.

[23] P. Willett, "The Porter stemming algorithm: then and now," Program: electronic library and information systems, 40(3), pp. 219-223, 2006.

[24] Y. Xu, Z. Zhang, P. Yu, and B. Long, "Pattern change discovery between high dimensional data sets," in Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1097-1106, October 2011.

[25] A. Dengel, M. Junker, and A. Weisbecker, (Eds.), "Reading and learning: adaptive content recognition," vol. 2956, Springer Science & Business Media, 2004.

[26] L. Wu, C. Faloutsos, K. Sycara, and T.R. Payne, "Falcon: Feedback adaptive loop for content-based retrieval," No. CMU-CS-00-142, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2000.

[27] A. Telang, C. Li, and S. Chakravarthy, "One Size Does Not Fit All: Toward User-and Query-Dependent Ranking for Web Databases. Knowledge and Data Engineering," IEEE Transactions on, 24(9), pp. 1671-1685, 2012.

[28] C.D. Manning, P. Raghavan, and H. Schütze, "Introduction to information retrieval," vol. 1, p. 496, Cambridge university press, 2008.

[29] L. Gou, X.L. Zhang, H.H Chen, J.H. Kim, and C.L. Giles, "Social network document ranking," in Proceedings of the 10th annual joint conference on Digital libraries, pp. 313-322, June 2010

[30] J. Yew, D.A. Shamma, and E.F. Churchill, "Knowing funny: genre perception and categorization in social video sharing," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 297-306, May 2011.

[31] A. Khodaei, and C. Shahabi, "Social-Textual Search and Ranking," CrowdSearch, 37(5), pp. 3-8, 2012.

[32] T. Shortell, "Online textbook: An introduction to data analysis & presentation," http://www.shortell.org/book/chap18.html accessed 20 Dec 2014, 2010.