A text mining methodology to discovery syllabi similarities among Higher Education Institutions

Gerardo Orellana
Universidad del Azuay
Cuenca, Ecuador
gorellana@uazuay.edu.ec

Marcos Orellana
Universidad del Azuay
Cuenca, Ecuador
marore@uazuay.edu.ec

Victor Saquicela
Universidad de Cuenca
Cuenca, Ecuador
victor.saquicela@ucuenca.edu.ec

Fernando Baculima
Universidad de Cuenca
Cuenca, Ecuador
fernando.baculima@ucuenca.edu.ec

Nelson Piedra
Universidad Técnica Particular de Loja
Loja, Ecuador
nopiedra@utpl.edu.ec

***Abstract*—Students' mobility and credit validation has been a concern for several years among higher education institutions in Ecuador, this process involves a huge amount of manual work due to the absence of an automatic system to measure the similarity between different course contents. In order to tackle this problem, we propose an approach to semantically compare the syllabi contents through text similarity methods. Such methods have been widely used in different domains, in this work we take the higher education institutions syllabi to the Text mining world and develop a method to compare their semantic contents. We propose an approach that uses pre-processing techniques, Latent Semantic Analysis for dimensionality reduction, text enrichment through the Wikipedia API and Google Engine, Support Vector Machine as classifier, and cosine similarity as similarity metric. Our results show that our method successfully measures similarity among higher education institutions syllabi and can be generalized to most Ecuadorian institutions.**

***Index Terms*—Text similarity, syllabus similarity, text mining, education, students' mobility.**

I. Introduction

Students' mobility is a large phenomena across countries and regions [1], [2], thus, credit validation among higher education institutions (HEIs) can bring several difficulties. HEIs of Ecuador are no exception, career coordinators who receive students requests to validate a course which was previously taken in another HEI lack an automatic system to measure the similarity between contents. Moreover, they require a vast amount of manual work to analyze and compare the syllabi contents. Therefore, in order to facilitate students' mobility and credit validation, we address syllabi comparison through text similarity methods.

The construction of a common framework to compare all the different variations in HEIs' syllabi, careers, and courses is not a trivial task. There is large variability among contents [3] and the language ambiguity increases the problem size. Language is found in every single domain of human knowledge, for a human, understanding the difference or similarity between two texts is not difficult, we are able to identify the words' context and picture the semantic meaning in our mind. For instance, it is easy to identify similarity between the words *dog* and *cane*, and dissimilarity between *phone* and *tree*. However, identifying such semantic meaning in order to produce similarity judgments with the help of machine processes is not trivial [4]. Several studies have used text similarity algorithms in many science domains such as education[5], [6], [7], bio-informatics [8], software development [9], [10], [11], medicine [12] plagiarism detection [13] and in classic information retrieval systems [14] where retrieving the right documents for a query is considered imperative.

There are several approaches to measure text similarity which can be used, [15] defines string-based, corpus-based, knowledge-based and hybrid similarity measures. String-based approaches measure distances between two different strings, this measures can be character-based or term-based, the first approach aligns strings and measures similarity based on the alignment precision, the second measures similarity by processing the strings as vectors. Corpus-based approaches focus on finding the semantic meaning of words and sentences as a result of the information provided by a large labeled corpora. Knowledge-based approaches use word networks in the form of synsets, words in the network relate to each other based on their cognitive meaning, the most popular example is Wordnet. Hybrid measures combine two or more of the described approaches, this kind of approach has proven to achieve the best performance as it is shown in [15].

For this work we use a hybrid approach by using Latent Semantic Analysis which is a popular feature transformation method along with classification algorithms, text enrichment, and distance metrics to achieve similarity. This work is part of the CEPRA project in which we proposed a methodology which goes from the extraction of syllabi data from raw sources to the process and creation of a semantic database defined by our model of syllabus ontology [16]. We use this semantic database to obtain the syllabi to test our similarity approach.

As a result of our method we have achieved to measure similarity among contents of different HEIs. In order to measure scope of our success in the task; we compared several courses in the Computer Science syllabi of two Ecuadorian HEIs. As a result of this process our method presents similar

978-1-5386-7612-7/18/$31.00 ©2018 IEEE
DOI 10.1109/INCISCOS.2018.00045

results to a human measurement. We can summarize the main contributions of our work to the following points:

- We present a methodology to measure the similarity from syllabi from different HEIs.
- We develop a model to classify syllabi in areas and subareas of knowledge.
- We develop a method to enrich educational contents using popular tools such as Google Engine and Wikipedia.
- We measure the similarity among different Computer Science courses of two HEIs and compare them to a human measurement.

This work is organized as follows: Section II presents the studies related to text mining and higher education curricula, Section III presents the proposed methodology to solve the discussed problem, Section IV presents our results and discussion, and Section V shows our conclusions and future work.

II. Related Work

Text mining approaches have been widely used to find interesting knowledge in HEIs' curricula [17], [18] and to develop new methods to automate processes such as academic advising [7] and credit validation [19]. These methods take syllabi and use them as a rich source of information to build upon.

Methods such as [7], [6] use syllabi for academic advising; [7] compares students' transcripts to the different major curricula to build an approach which extracts the features from the courses previously taken by the students and those from the different majors. The major which achieves the highest similarity is advised to students. However, this method is a totally dependent to a supervised approach in which a focus group of experts weights and classifies courses; other approach ti to build a collection of documents for advising as it is done in [6], they present an approach to scrape higher education syllabus from the web and extract structured data from plain html based in tags and keywords.

There are studies which compare and contrast syllabi from education institutions [17], [18], [19]. Secondary school syllabi are compared in [17], they study physical education among Australians with a content analysis done by a software called Leximancer[1]. They use this tool to extract high level concepts and themes from the texts. However, these results were manually calculated without the help of an automatic process to achieve a similarity measurement. [18] analyses the curricula of 5 top universities in the area of computer science, they use TF/IDF as a keyword extraction method. They considered that syllabi are not a complete source of information to analyze curricula. Therefore, they studied further through the course materials of the area they picked. Although, this could be a reasonable solution, we consider not feasible to collect all the courses' material from the different Ecuadorian universities. [19] measures the semantic similarity of Computer Science courses in Thai universities, they use a knowledge-based approach which normalizes the bodies of knowledge through Wordnet and calculate similarity with an extended version of the Wu & Palmer's algorithm. [20] propose a robust text similarity approach in which they use dimensionality reduction techniques such as LSA, a supervised and unsupervised model to classify text, Wordnet as a semantic database of knowledge, urban dictionaries to improve Wordnet accuracy and a transtaltion method to measure similarity across different languages.

Approaches such as [19], [20] are not feasible to apply to our projects, given that a body of knowledge with Wordnet quality is not available for Spanish. In this study we develop a text similarity approach for Spanish which does not depend on knowledge resources like Wordnet. However, we take [20] as inspiration of a robust method and apply a feature transformation technique before our classification methods.

III. Methodology

This methodology shows a classification and similarity measurement of syllabi among HEIs of Ecuador. The process starts by extracting the text from a semantic database which is described in our previous work[16].

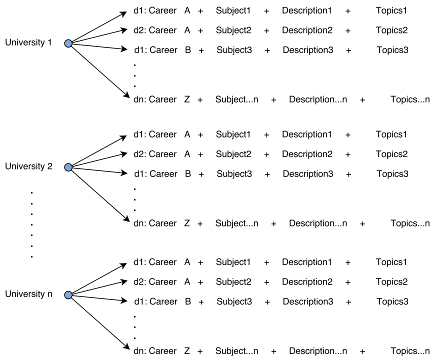

University 1
d1: Career A + Subject1 + Description1 + Topics1
d2: Career A + Subject2 + Description2 + Topics2
d1: Career B + Subject3 + Description3 + Topics3
dn: Career Z + Subject...n + Description...n + Topics...n
University 2
d1: Career A + Subject1 + Description1 + Topics1
d2: Career A + Subject2 + Description2 + Topics2
d1: Career B + Subject3 + Description3 + Topics3
dn: Career Z + Subject...n + Description...n + Topics...n
University n
d1: Career A + Subject1 + Description1 + Topics1
d2: Career A + Subject2 + Description2 + Topics2
d1: Career B + Subject3 + Description3 + Topics3
dn: Career Z + Subject...n + Description...n + Topics...n

Fig. 1. Syllabus Decomposition

Syllabi contain a large amount of information, usually these resources contains information such as description of the course, tittle, goals, learning objectives, topics, assessment criteria and others. However, not all of these information is relevant to take into account for a similarity approach, for instance, the assessment policies and teaching methodology are irrelevant to compare among universities. Therefore, this work focuses in what we define as relevant for a similarity approach which is the career name, course name, course description and topics as it can be seen from figure 1. This basic structure is the seed for the upcoming similarity processes and algorithms.

A. Data Query

Different HEIs have different ways to storage and structure syllabi information in their systems, thus, in order to access all this information from Ecuadorian HEIs [16] proposes an approach based on the Semantic web; the result of this work is a unified semantic database containing all syllabi information of the participating HEIs. This study queries the data form such source through the use of SPARQL; this is a language to query RDF information. The data query is the first step of the

[1] https://info.leximancer.com/

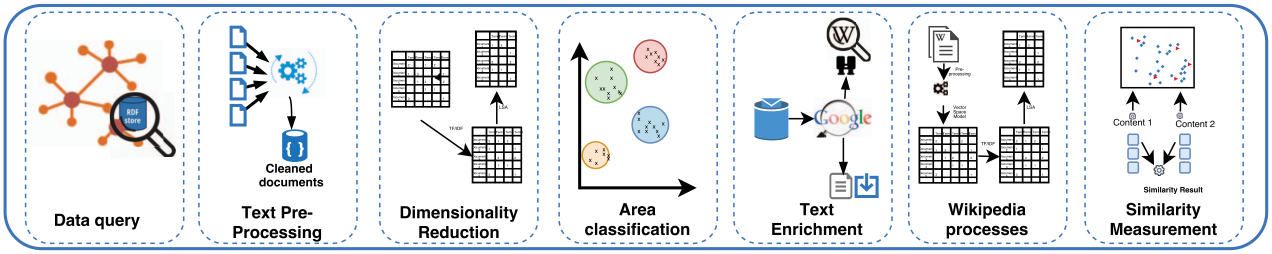


Fig. 2. Methodology Process

process as it is shown in Figure 2. The resulting triplets from such queries follows the structure shown in Figure 1, meaning than only triplets which corresponds to career name, course name, course description and course contents are retrieved from the semantic database.

B. Text pre-processing

A common practice in text mining is to pre-process the text before the application of a classification or feature transformation algorithm[21], [22]. Pre-processing techniques aim to reduce the text to a simpler representation which keeps the main characteristics of the original text. The most common techniques are stopwords removal, punctuation removal, lower case transformation, stemming, and lemmatization. The last two, transform words into a their root or base form while the others remove unnecessary characters and words and uniform the case.

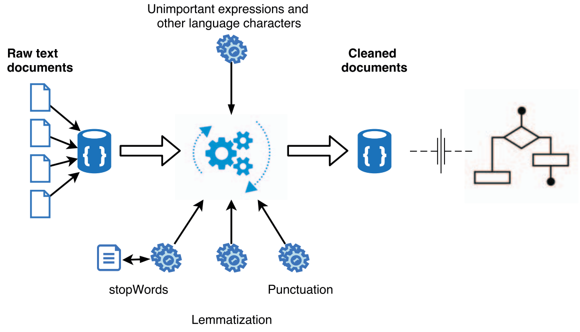


Fig. 3. Cleaning Process

Figure 3 shows the text pre-processing process where the techniques are applied to arrive to a cleaned dataset ready to be processed by any algorithm. The techniques we apply in this work are lowercase conversion, stopwords and punctuation removal and lemmatization.

C. Dimentionality Reduction

In [23] it is shown that using a unsupervised dimensionality reduction technique such as Principal Component Analysis (PCA) before applying a clustering algorithm improve performance. For this work we apply a similar approach by using TF/IDF and Latent Semantic Analysis (LSA) as dimensionality reduction techniques before applying a classification algorithm. We use these given that according to [24], the combination of numerical selection and feature transformation techniques achieve superior results .

1) Feature Selection Technique: One technique that has been widely used is TF/IDF which is named for (Term Frequency / Inverse Document Frequency). This technique measures the importance of a term in a document as well as its importance among all documents. Thus, this measure obtains a combined value which provides a richer metric to filter in the most important documents and to filter out those which are not important enough. In this study, we selected 75% of the terms which have the best TF/IDF score in the syllabus dataset, meaning that 25% of infrequent and not relevant terms are left behind for the next steps of the study.

2) Feature Transformation techniques: In this work we use a similar approach, we use Latent Semantic Analysis (LSA) technique to capture the most relevant features in our dataset and by reducing its dimensionality, an example of this can be seen in [25], where Singular Value Decomposition (SVD) is applied to filter out the least relevant features. This feature transformation technique can be seen as similar to Principal Component Analysis (PCA) which is a non-parametric method. PCA analyzes the co-variance of the eigenvectors of the term-document matrix. On the other hand LSA is a parametric method which takes as input the number of features to use them as a projection matrix [26], [20]. We chose LSA to perform dimensionality reduction because it is a better method for large text as it is the case with syllabus of higher education institutions, also PCA is not a good method when there are too many dimensions and the data is too sparse [27].

Dimensionality reduction has been widely applied in several text similarity works to obtain better measurements in a smaller dimensional space, some applications of LSA can be seen in: [20], [28], [29].

D. Area of Knowledge Classification

The amount of careers, courses and topics for each course can result in an overwhelming amount of dimensions which are difficult to handle by any semantic algorithm. Thus, we develop a two phase similarity analysis in which first, the syllabi is classified into a the areas of knowledge defined by UNESCO[30]. UNESCO defines 9 main areas and 24 subareas of knowledge.

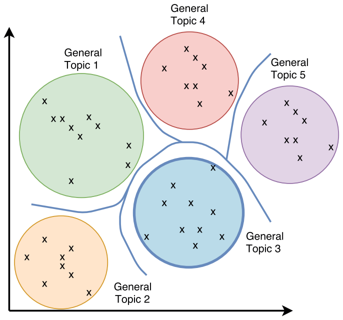


Fig. 4. Syllabus Classification

Figure 4, shows the two-phase approach we use for classifying and measuring similarity, on the left side of the figure we show the syllabi vector space model in which we reduce dimensions and group main features through LSA and cluster through Support Vector Machine (SVM); on the right part of the figure, the enrichment process is shown, more details of this are in sub-section III-E, we reduce dimensionality and group main features through LSA. We use SVM because it is shown in [31] that it performs better when classifying syllabi than other well known supervised classifiers such as Naive Bayes, random Forest, K-Nearest neighbors, and Decision Trees .

E. Text Enrichment

Syllabus are hierarchical documents which detail the contents that are thought in a course, this contents are usually just listed providing no more information. Therefore, to measure the similarity between two syllabus this contents become critical. In order to compare this very short content descriptions, we first enrich them with additional external resources. There are two different text enrichment techniques according to [5], a domain specific one selecting just topic related Wikipedia documents and a general one in which they use the whole body of knowledge of Wikipedia. Conclusions show that a domain specific method performs better. We choose Wikipedia[2], a free encyclopedia which is known for containing a large corpora of articles written in different languages for its academic focus. The contents are enriched with Wikipedia articles which are approximate in meaning to them, this enriches the vector space by keeping coherence with the target topic defined before in our classification. Wikipedia has been used before to construct bodies of knowledge as it can be seen in [32], [5]. In order to find the Wikipedia articles to enrich our syllabi topics, we use Google Search Engine API. This approach have been used before in [33] where they build a general purpose corpora from Google Search Engine query results. We use a similar although narrowed approach to build a topic specific corpus which contains only Wikipedia results in Spanish.

Figure 5, shows our method to enrich contents; we start by querying the syllabus contents belonging to a classified general topic, afterwords, we use Google Search Engine as a help for finding the 10 most related relevant Wikipedia articles for each topic. The Wikipedia API returns two things, the summary and the complete article.

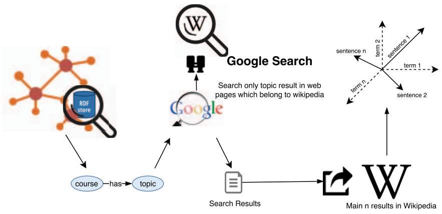
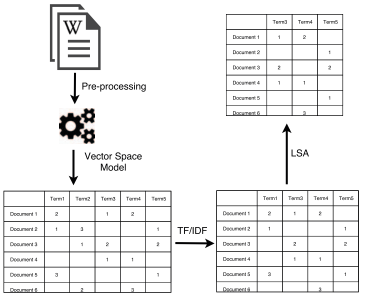


Fig. 5. Topic enrichment process

F. Wikipedia Articles process

The sub-section III-E described a text enrichment process which takes the articles from Wikipedia. However, these articles have to go through pre-processing to reduce the text to a simpler representation and to locate terms closer to each other according to their context as it was described in sub-sections III-B and III-C. Figure 6, shows the process where the set of Wikipedia articles are pre-processed through the same methods described in sub-section III-B. The Wikipedia articles first are parsed to a vector space model, here they are processed through TF/IDF to save the most relevant terms, and finally they are processed through LSA, where they are located closed to each other according to their context.



	Term1	Term2	Term3	Term4	Term5
Document 1	2		1	2	
Document 2	1	3			1
Document 3		1	2		2
Document 4			1	1	
Document 5	3				1
Document 6		2		3	

	Term1	Term3	Term4	Term5
Document 1	2	1	2	
Document 2	1			1
Document 3		2		2
Document 4		1	1	
Document 5	3			1
Document 6			3	

	Term3	Term4	Term5
Document 1	1	2	
Document 2			1
Document 3	2		2
Document 4	1	1	
Document 5			1
Document 6		3	

Fig. 6. Similarity Measurements

G. Similarity measurement

In section II, the text similarity approaches were covered, we use a hybrid method with a corpus-based approach and a string-based one as it is cosine similarity. This is a standard method to measure the similarity between documents represented as vectors in the vector space model [14]. This measure evaluates the cosine of the angle between two vectors and give us a measure within 1 and -1 which shows the degree of similarity with 1 representing equal vectors and -1 orthogonal ones.

Figure 7, shows the bases of the cosine similarity measure. It stars with all the documents located in the vector space, where the terms and documents are distributed according the processes done in sub-section III-F.

[2] www.wikipedia.org

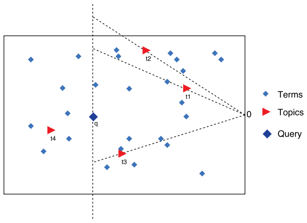


Fig. 7. Measuring topic-query distance

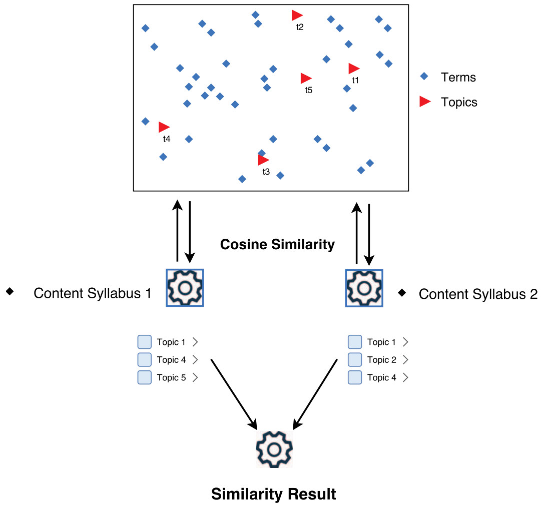


Fig. 8. Similarity Approach

In a enriched space which has been already processed by LSA, the terms and documents are located within their context as it can be seen in Figure 7. However, a further process is needed to compare similarity between two syllabuses. For such process we take as main object of study the syllabi contents. Figure 8, illustrates the complete measurement process in which we compare the different syllabi in a query fashion way. A content from a syllabus becomes the query, and with cosine similarity we discover the most relevant Wikipedia articles and store them as a vector. We determine the most relevant documents by defining a similarity threshold.

$$c_i \sim a_j \leq \theta_1 \tag{1}$$

Equation 1, shows that a content c_i is similar to a Wikipedia article a_j if the result of its cosine similarity is less or equal than the threshold θ_1. Therefore, after a complete execution of the similarity process, we end up with all contents and their most relevant Wikipedia articles for all the classes. However, this still does not represent a measure of similarity among syllabi of different HEIs, thus, we compare the contents between syllabi and define a measure of similarity according to the percentage Wikipedia articles in which the two contents are present.

$$\frac{W_i \cap W_j}{W_i \cup W_j} \geq \theta_2 \tag{2}$$

Equation 2 shows the similarity measurement among contents where W_i and W_j are the most relevant Wikipedia articles of two different contents i and j, and θ_2 is a similarity threshold. Measurements above or equal to this threshold indicate that two contents are similar.

$$\frac{C_{s1} \sim C_{s2}}{C_{s1} \cup C_{s2}} \geq \theta_3 \tag{3}$$

As stated earlier in this section, a syllabus is composed by several contents. Therefore, the syllabi similarity measurement is given by equation 3 where C_{s1} and C_{s2} are all the contents of two different syllabi and θ_3 is the syllabi threshold. The equation shows that the final similarity measurement is given by the number of contents which are similar in two syllabi by union of the contents from the two syllabi.

IV. Results and Discussion

As described in Section III, we start from data collected in a semantic database. As pre-processing techniques we executed the following: lower case transformation, stopwords and punctuation removal, and lemmatization for Spanish texts. We started with syllabi containing a total of 17340 different terms also called dimensions and after the execution of LSA over the syllabi, we ended up with 500 dimensions in each component of the vector space model. For the classification process we took the 1442 syllabi from a university and manually classify them with the UNESCO areas and subareas of knowledge[30].

TABLE I
Classification Results

	Classification Results			
	Accuracy	Precision	Recall	F-measure
Area	0.7993	0.7934	0.7557	0.7711
Subarea	0.7391	0.7324	0.6823	0.6882

Table I, shows the classification results with the SVM algorithm. As it can be seen in the table, the results of the areas of knowledge are better than those of the subareas. We conclude this difference is due to the amount of classes for each category. UNESCO classifies 9 areas and 24 subareas. The classification is set as a prediction model to classify the syllabi of other Ecuadorian HEIs. For this work, we had availability of all the syllabi from a second HEI which we use to test our methods.

We tested our method with the information from the syllabi corresponding to the Computer Science career for which we performed all the methodology. First we enriched the contents with all the Computer Science syllabi from one University which has a total of 57 courses. Moreover, there are 1955 contents corresponding to these courses. As a result of the enrichment, we obtained a total of 9088 Wikipedia articles enriching the contents in the different areas of knowledge described before in this Section.

We selected six courses from the Computer Science degree which by name should be similar to show the scope of our approach. We compared these six courses to each other as

TABLE II
RESULTS OVER DIFFERENT COMPUTER SCIENCE COURSES WITH SEVERAL ARTICLE SIMILARITY THRESHOLDS

Courses Universities		Similarity Results By Article Thresholds				
University A	University B	0.75	0.80	0.85	0.90	Human Measure
Databases I	Databases I	00.00	28.57	55.95	82.14	54.11
Artificial Intelligence	Artificial Intelligence	00.00	09.09	29.09	36.36	10.90
Databases II	Databases II	00.00	25.92	85.18	93.51	40.18
Systems analysis I	Analysis and design of Software I	00.00	00.00	19.20	09.60	25.00
Entrepreneurship	Entrepreneurship	00.00	00.00	09.52	03.17	06.35
Software Engineering	Software Engineering	00.00	00.00	34.02	94.84	64.23
Artificial Intelligence	Data Bases I	00.00	00.00	17.33	00.00	00.00
Entrepreneurship	Databases I	00.00	00.00	00.00	00.00	00.00
Artificial Intelligence	Entrepreneurship	00.00	00.00	00.00	00.00	00.00
Entrepreneurship	Analysis and design of Software I	00.00	00.00	00.00	00.00	00.00
Databases I	Databases II	00.00	32.00	61.00	80.00	13.13
Databases II	Databases I	00.00	23.91	65.21	75.00	22.58

TABLE III
RESULTS OVER DIFFERENT COMPUTER SCIENCE COURSES WITH SEVERAL CONTENT THRESHOLDS

Courses Universities		Similarity Results By Article Thresholds				
University A	University B	0.20	0.30	0.40	0.50	Human Measure
Databases I	Databases I	86.90	55.95	23.80	08.33	54.11
Artificial Intelligence	Artificial Intelligence	52.72	29.09	25.45	20.00	10.90
Databases II	Databases II	87.96	85.18	82.40	74.07	40.18
Systems analysis I	Analysis and design of Software I	34.40	19.20	07.20	00.00	25.00
Entrepreneurship	Entrepreneurship	12.69	09.52	04.76	04.76	06.35
Software Engineering	Software Engineering	39.18	34.02	28.86	27.83	64.23
Artificial Intelligence	Data Bases I	28.00	17.33	13.33	08.00	00.00
Entrepreneurship	Databases I	00.00	00.00	00.00	00.00	00.00
Artificial Intelligence	Entrepreneurship	00.00	00.00	00.00	00.00	00.00
Entrepreneurship	Analysis and design of Software I	00.00	00.00	00.00	00.00	00.00
Databases I	Databases II	89.00	61.00	32.00	12.00	13.13
Databases II	Databases I	77.17	65.21	43.48	29.34	22.58

it is shown in table II. However, in order to exemplify the correctness of the approach, we also compare courses that should be different to each other and show the results. In order to test our method, we manually measured the similarity between the syllabi contents of the selected courses and later compare them to those automatically classified.

Table II, shows in the two first columns the courses which are compared to each other and in the next columns measurements of similarity with different article similarity thresholds and the last columns is the measure done by a human. The threshold in the table is given by equation 1. Thus, table II shows that our approach is very sensible to this threshold. We can visually see that the results that seem more similar to what a human would measure are those with 0.80 or 0.85. However, in order to determine the best threshold numerically, we calculate an error by equation 4 where R_i represents every result for a column i in the table, M_i represents the corresponding human measurement, and n represents the total number or compared elements.

$$\sum_{i=1}^{n} \frac{(R_i - M_i)^2}{n} \tag{4}$$

We determine that the best article threshold is given by a threshold of 0.85 since Table IV shows that it has the lowest error. This threshold value represents the distance of the content as query term to the Wikipedia articles. Measurements lower than this threshold filter out too many articles and therefore the similarity measurements tend to go lower until they all reach zero at 0.75. In the other hand, when the threshold is bigger than 0.85, too many articles are included as similar and the measurements tend to be higher.

Table II shows at the top results of courses which by name appear to be similar and the results seem to be similar to reality. However, we compare every course to each other and thus at the bottom of the table we show the results of comparing unrelated courses. We can see in the table that courses that are unrelated or very distant in topics such as Databases and Artificial Intelligence have similarities of zero with almost every threshold which show a perfect concordance with the reality. However, courses that are different in content such as Databases I and Databases II but talk about the same topic which in this case is databases show large similarity measurements. This shows that our approach presents failures in such scenario, we assume that this is caused because our method has proven to be effective filtering in the articles related to a specific topic without content discrimination. Therefore, further improvements should be done to overcome this limitation.

Table III, shows the similarity results of our approach with different values of contents' threshold shown by equation 2 with an article threshold of 0.85. It can be seen in the table that the higher the content threshold the more strict the approach results showing lower similarity measures. A visual inspection

TABLE IV
ARTICLES AND CONTENTS THRESHOLD ERRORS

	Article Thresholds				Content Thresholds			
	0.75	0.80	0.85	0.90	0.20	0.30	0.40	0.50
Errors	2927.95	2144.22	1292.77	1808.93	1313.70	1292.70	1571.87	1917.36

of the data shows that the 0.3 and 0.4 thresholds may be the closer to the measurement done by a human. However, by the calculation of the error shown by equation 4 and shown in Table IV, we determine that the results with the lowest difference with the human measurement are those with a threshold of 0.3.

In our methodology in sub-section III-C1, we denote the use of three thresholds; two of them were discussed and shown in tables II and III. The last similarity threshold indicates whether a syllabus in general terms is similar to another. We have set this threshold to 0.4, thus, we determine that the similar courses are Databases I and Databases II. We have discovered similarity in four contents with our method while the human measurement show similarity in just two. Furthermore, in the similar syllabi our method made two mistakes, thus, we can summarize our results in the following confusion matrix:

TABLE V
CONFUSION MATRIX - SIMILARITY RESULTS

	similar	not similar	Totals
similar	2	2	4
not similar	0	8	8
Totals	2	10	12

According to table V, we can calculate the accuracy, precision and recall of our method. We achieve an accuracy of 83.33 %, a precision of 50.00% and a recall of 100%. These values show promising results. However, values such as the precision and recall are strongly influenced by the short number of comparisons in our sample. Even though we have tested and validated our method with the information provided here, the values of precision and recall would be greatly influenced with more data samples.

V. CONCLUSIONS AND FUTURE WORK

The following are the conclusions we draw from the application of our methodology to several Computer science courses of two different HEIs.

- Our work show a methodology of seven different steps that has proven to be useful and close to the measurement done by a human in most of the cases.
- In cases when two very unrelated topics are compared, our results show perfect alignment with the reality.
- In courses formed as succession of others such as Databases I and Databases II, our approach shows failures in the comparisons among two HEIs. We perform a topic related comparison and when two courses may have different contents but belong to the same topic our methodology gives a high similarity measure. Some further work needs to be done to overcome this issue.
- Our approach is very sensible to the threshold values for similarity. This may indicate that different areas of knowledge may need different thresholds to successfully predict similarity. The threshold defined for the Computer science courses may not perform as good for other courses.
- The classification results such as accuracy, precision, recall and f-measure need to be improved to make the method more reliable and to give an assurance that it will work for every course. Some further work with deep networks can be done to improve the results.
- The methodology should be further validated by involving experts in every area to measure similarity of the contents.
- Methods to graphically show the similarity measurements of contents and syllabi should be developed to present them to a final user.
- We achieved an accuracy of 83.33%, 50% precision, and 100% recall for our method with the selected course values for comparison. This values show promising results and have tested our methodology concept with several Computer Science courses. However, further measurements should be done to arrive to a more general conclusion

VI. ACKNOWLEDGMENTS

This work is part of the ”Detección de similitudes entre contenidos académicos de carrera a través de la aplicación de tecnologías semánticas y minería de datos” project, supported by the Ecuadorian Corporation for Research and Academia (CEDIA for its Spanish acronym).

REFERENCES

[1] B. Rico and R. Emilia, “La movilidad internacional por razones de estudio: Geografía de un fenómeno global,” *Migraciones internacionales*, vol. 8, no. 1, pp. 95–125, 2015.

[2] C. V. Solís, C. G. Martín, and A. Correa, “Circularidad migratoria entre ecuador y españa. transformación educativa y estrategias de movilidad,” *Migraciones. Publicación del Instituto Universitario de Estudios sobre Migraciones*, no. 39, pp. 183–210, 2016.

[3] A. E. García Muñoz, “Estudio del conocimiento de las carreras que se ofertan en el país y la demanda de ellos por parte de los futuros bachilleres de la ciudad de guayaquil,” B.S. thesis, 2003.

[4] D. Croft, S. Coupland, J. Shell, and S. Brown, “A fast and efficient semantic short text similarity metric,” in *Computational Intelligence (UKCI), 2013 13th UK Workshop on*. IEEE, 2013, pp. 221–227.

[5] M. Mohler and R. Mihalcea, “Text-to-text semantic similarity for automatic short answer grading,” in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 567–575.

[6] Y. Biletskiy, J. A. Brown, and G. Ranganathan, “Information extraction from syllabi for academic e-advising,” *Expert Systems with Applications*, vol. 36, no. 3, pp. 4508–4516, 2009.

[7] L. Mostafa, G. Oately, N. Khalifa, and W. Rabie, “A case based reasoning system for academic advising in egyptian educational institutions,” in *2nd International Conference on Research in Science, Engineering and Technology (ICRSET2014) March*, 2014, pp. 21–22.

[8] J. Lewis, S. Ossowski, J. Hicks, M. Errami, and H. R. Garner, "Text similarity: an alternative way to search medline," *Bioinformatics*, vol. 22, no. 18, pp. 2298–2304, 2006.

[9] A. Lazar, S. Ritchey, and B. Sharif, "Improving the accuracy of duplicate bug report detection using textual similarity measures," in *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM, 2014, pp. 308–311.

[10] A. De Lucia, M. Di Penta, R. Oliveto, A. Panichella, and S. Panichella, "Labeling source code with information retrieval methods: an empirical study," *Empirical Software Engineering*, vol. 19, no. 5, pp. 1383–1420, 2014.

[11] G. Bavota, B. Dit, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia, "An empirical study on the developers' perception of software coupling," in *Proceedings of the 2013 International Conference on Software Engineering*. IEEE Press, 2013, pp. 692–701.

[12] W. Wang, Q. Jiang, T. Lv, W. Guo, and C. Wang, "An improved text similarity algorithm research for clinical decision support system," in *Cloud Computing and Intelligence Systems (CCIS), 2016 4th International Conference on*. IEEE, 2016, pp. 155–159.

[13] J. Ferrero, F. Agnes, L. Besacier, and D. Schwab, "Compilig at semeval-2017 task 1: Cross-language plagiarism detection methods for semantic textual similarity," *arXiv preprint arXiv:1704.01346*, 2017.

[14] D. M. Christopher, R. Prabhakar, and S. Hinrich, "Introduction to information retrieval," *An Introduction To Information Retrieval*, vol. 151, p. 177, 2008.

[15] W. H. Gomaa and A. A. Fahmy, "A survey of text similarity approaches," *International Journal of Computer Applications*, vol. 68, no. 13, 2013.

[16] V. Saquicela, F. Baculima, G. Orellana, M. Orellana, N. Piedra, and M. Espinoza, "Similarity detection among academic contents through semantic technologies and text mining," in *Proceedings INFOBAE Cuba 2018*, 2018.

[17] B. Hyndman and S. Pill, "The curriculum analysis of senior education in physical education (case-pe) study," *Curriculum Perspectives*, vol. 37, no. 2, pp. 147–160, 2017.

[18] K. Kawintiranon, P. Vateekul, A. Suchato, and P. Punyabukkana, "Understanding knowledge areas in curriculum through text mining from course materials," in *Teaching, Assessment, and Learning for Engineering (TALE), 2016 IEEE International Conference on*. IEEE, 2016, pp. 161–168.

[19] C. Nuntawong, C. S. Namahoot, and M. Brückner, "A semantic similarity assessment tool for computer science subjects using extended wu & palmers algorithm and ontology," *Lecture Notes in Electrical Engineering*, vol. 339, p. 989, 2015.

[20] A. Kashyap, L. Han, R. Yus, J. Sleeman, T. Satyapanich, S. Gandhi, and T. Finin, "Robust semantic text similarity using lsa, machine learning, and linguistic resources," *Language Resources and Evaluation*, vol. 50, no. 1, pp. 125–161, 2016.

[21] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut, "A brief survey of text mining: Classification, clustering and extraction techniques," *arXiv preprint arXiv:1707.02919*, 2017.

[22] M. R. Patel and M. G. Sharma, "A survey on text mining techniques," *International Journal Of Engineering And Computer Science ISSN*, vol. 2319, no. 7242, pp. 5621–5625, 2014.

[23] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 29.

[24] B. Tang, M. Shepherd, E. Milios, and M. I. Heywood, "Comparing and combining dimension reduction techniques for efficient text clustering," in *Proceeding of SIAM International Workshop on Feature Selection for Data Mining*, 2005, pp. 17–26.

[25] G. Lapesa and S. Evert, "A large scale evaluation of distributional semantic models: Parameters, interactions and model selection," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 531–545, 2014.

[26] W.-t. Yih, K. Toutanova, J. C. Platt, and C. Meek, "Learning discriminative projections for text similarity measures," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2011, pp. 247–256.

[27] R. Ramya, K. Venugopal, S. Iyengar, and L. Patnaik, "Feature extraction and duplicate detection for text mining: A survey," *Global Journal of Computer Science and Technology*, vol. 16, no. 5, 2017.

[28] A. kumar Jayapal, M. Emms, and J. Kelleher, "Tcdscss: Dimensionality reduction to evaluate texts of varying lengths-an ir approach," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 2014, pp. 619–623.

[29] H. Kwon, J. Kim, and Y. Park, "Applying lsa text mining technique in envisioning social impacts of emerging technologies: The case of drone technology," *Technovation*, 2017.

[30] M. I. Juan Pastor, Javier Martnez. Unesco nomenclature for fields of science and technologyn. [Online]. Available: http://skos.um.es/unesco6/

[31] N. Rathod and L. Cassel, "Building a search engine for computer science course syllabi," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*. ACM, 2013, pp. 77–86.

[32] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, "Exploiting wikipedia as external knowledge for document clustering," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 389–396.

[33] S. Sharoff, "Creating general-purpose corpora using automated search engine queries," *WaCky*, pp. 63–98, 2006.